

## Contents

Preface.....	2
Solutions to Chapter 2 Problems.....	4
Solutions to Chapter 3 Problems.....	11
Solutions to Chapter 4 Problems.....	15
Solutions to Chapter 5 Problems.....	38
Solutions to Chapter 6 Problems.....	57
Solutions to Chapter 7 Problems.....	80
Solutions to Chapter 8 Problems.....	104
Solutions to Chapter 9 Problems.....	126
Solutions to Chapter 10 Problems.....	151
Solutions to Chapter 11 Problems.....	207
Solutions to Chapter 12 Problems.....	242
Solutions to Chapter 13 Problems.....	270
Solutions to Chapter 14 Problems.....	295
Solutions to Chapter 15 Problems.....	304
Solutions to Chapter 16 Problems.....	341
Solutions to Chapter 17 Problems.....	358
Solutions to Chapter 18 Problems.....	406
Solutions to Chapter 19 Problems.....	445
Solutions to Chapter 20 Problems.....	465
Solutions to Chapter 21 Problems.....	484
Solutions to Chapter 22 Problems.....	538

## Preface

This manual contains the solutions to all of the problems in the second edition of my MIT Press book, *Econometric Analysis of Cross Section and Panel Data*. In addition to the problems printed in the text, I have included some “bonus problems” along with their solutions. Several of these problems I left out due to space constraints and others occurred to me since the book was published. I have a collection of other problems, with solutions, that I have used over the past 10 years for problem sets, takehome exams, and in class exams. I am happy to provide these to instructors who have adopted the book for a course.

I solved the empirical examples using various versions of Stata, ranging from 8.0 through 11.0. I have included the Stata commands and output directly in the text. No doubt there are Stata users and users of other software packages who will, at least in some cases, see more efficient or more elegant ways to compute estimates and test statistics.

Some of the solutions are fairly long. In addition to filling in all or most of the algebraic steps, I have tried to offer commentary about why a particular problem is interesting, why I solved the problem the way I did, or which conclusions would change if we varied some of the assumptions. Several of the problems offer what appear to be novel solutions to situations that can arise in actual empirical work.

My progress in finishing this manual was slowed by a health problem in spring and summer of 2010. Fortunately, several graduate students came to my aid by either working through some problems or organizing the overall effort. I would like to thank Do Won Kwak, Cuicui Lu, Myoung-Jin Keay, Shenwu Sheng, Iraj Rahmani, and Monthien Satimanon for their able assistance.

I would appreciate learning about any mistakes in the solutions and also receiving

suggestions for how to make the answers more transparent. Of course I will gladly entertain suggestions for how the text can be improved, too. I can be reached via email at [wooldri1@msu.edu](mailto:wooldri1@msu.edu).

## Solutions to Chapter 2 Problems

2.1. a. Simple partial differentiation gives

$$\frac{\partial E(y|x_1, x_2)}{\partial x_1} = \beta_1 + \beta_4 x_2$$

and

$$\frac{\partial E(y|x_1, x_2)}{\partial x_2} = \beta_2 + 2\beta_3 x_2 + \beta_4 x_1$$

b. By definition,  $E(u|x_1, x_2) = 0$ . Because  $x_2^2$  and  $x_1 x_2$  are functions of  $(x_1, x_2)$ , it does not matter whether or not we also condition on them:  $E(u|x_1, x_2, x_2^2, x_1 x_2) = 0$ .

c. All we can say about  $\text{Var}(u|x_1, x_2)$  is that it is nonnegative for all  $x_1$  and  $x_2$ :

$E(u|x_1, x_2) = 0$  in no way restricts  $\text{Var}(u|x_1, x_2)$ .

2.2. a. Because  $\partial E(y|x)/\partial x = \delta_1 + 2\delta_2(x - \mu)$ , the marginal effect of  $x$  on  $E(y|x)$  is a linear function of  $x$ . If  $\delta_2$  is negative then the marginal effect is less than  $\delta_1$  when  $x$  is above its mean. If, for example,  $\delta_1 > 0$  and  $\delta_2 < 0$ , the marginal effect will eventually be negative for  $x$  far enough above  $\mu$ . (Whether the values for  $x$  such that  $\partial E(y|x)/\partial x < 0$  represents an interesting segment of the population is a different matter.)

b. Because  $\partial E(y|x)/\partial x$  is a function of  $x$ , we take the expectation of  $\partial E(y|x)/\partial x$  over the distribution of  $x$ :  $E[\partial E(y|x)/\partial x] = E[\delta_1 + 2\delta_2(x - \mu)] = \delta_1 + 2\delta_2 E[(x - \mu)] = \delta_1$ .

c. One way to do this part is to apply Property LP.5 from Appendix 2A. We have

$$\begin{aligned} L(y|1, x) &= L[E(y|x)] = \delta_0 + \delta_1 L[(x - \mu)|1, x] + \delta_2 L[(x - \mu)^2|1, x] \\ &= \delta_0 + \delta_1(x - \mu) + \delta_2(\gamma_0 + \gamma_1 x), \end{aligned}$$

because  $L[(x - \mu)|1, x] = x - \mu$  and  $\gamma_0 + \gamma_1 x$  is the linear projection of  $(x - \mu)^2$  on  $x$ . By assumption,  $(x - \mu)^2$  and  $x$  are uncorrelated, and so  $\gamma_1 = 0$ . It follows that

$$L(y|x) = (\delta_0 - \delta_1\mu + \delta_2\gamma_0) + \delta_1x$$

**2.3.** a.  $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2 + u$ , where  $u$  has a zero mean given  $x_1$  and  $x_2$ :

$E(u|x_1, x_2) = 0$ . We can say nothing further about  $u$ .

b.  $\partial E(y|x_1, x_2)/\partial x_1 = \beta_1 + \beta_3x_2$ . Because  $E(x_2) = 0$ ,  $\beta_1 = E[\partial E(y|x_1, x_2)/\partial x_1]$ , that is,  $\beta_1$  is the average partial effect of  $x_1$  on  $E(y|x_1, x_2)/\partial x_1$ . Similarly,  $\beta_2 = E[\partial E(y|x_1, x_2)/\partial x_2]$ .

c. If  $x_1$  and  $x_2$  are independent with zero mean then  $E(x_1x_2) = E(x_1)E(x_2) = 0$ . Further, the covariance between  $x_1x_2$  and  $x_1$  is  $E(x_1x_2 \cdot x_1) = E(x_1^2x_2) = E(x_1^2)E(x_2)$  (by independence)  $= 0$ . A similar argument shows that the covariance between  $x_1x_2$  and  $x_2$  is zero. But then the linear projection of  $x_1x_2$  onto  $(1, x_1, x_2)$  is identically zero. Now just use the law of iterated projections (Property LP.5 in Appendix 2A):

$$\begin{aligned} L(y|1, x_1, x_2) &= L(\beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2|1, x_1, x_2) \\ &= \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3L(x_1x_2|1, x_1, x_2) \\ &= \beta_0 + \beta_1x_1 + \beta_2x_2. \end{aligned}$$

d. Equation (2.47) is more useful because it allows us to compute the partial effects of  $x_1$  and  $x_2$  at *any* values of  $x_1$  and  $x_2$ . Under the assumptions we have made, the linear projection in (2.48) does have as its slope coefficients on  $x_1$  and  $x_2$  the partial effects at the population average values of  $x_1$  and  $x_2$  – zero in both cases – but it does not allow us to obtain the partial effects at any other values of  $x_1$  and  $x_2$ . Incidentally, the main conclusions of this problem go through if we allow  $x_1$  and  $x_2$  to have nonzero population means.

**2.4.** By assumption,

$$E(u|\mathbf{x}, v) = \delta_0 + \mathbf{x}\boldsymbol{\delta} + \rho_1v$$

for some scalars  $\delta_0, \rho_1$  and a column vector  $\boldsymbol{\delta}$ . Now, it suffices to show that  $\delta_0 = 0$  and  $\boldsymbol{\delta} = \mathbf{0}$ .

One way to do this is to use LP.7 in Appendix 2A, and in particular, equation (2.56). This says

that  $(\delta_0, \delta')'$  can be obtained by first projecting  $(1, \mathbf{x})$  onto  $v$ , and obtaining the population residual,  $\mathbf{r}$ . Then, project  $u$  onto  $\mathbf{r}$ . Now, since  $v$  has zero mean and is uncorrelated with  $\mathbf{x}$ , the first step projection does nothing:  $\mathbf{r} = (1, \mathbf{x})$ . Thus, projecting  $u$  onto  $\mathbf{r}$  is just projecting  $u$  onto  $(1, \mathbf{x})$ . Since  $u$  has zero mean and is uncorrelated with  $\mathbf{x}$ , this projection is identically zero, which means that  $\delta_0 = 0$  and  $\delta = 0$ .

**2.5.** By definition and the zero conditional mean assumptions,  $\text{Var}(u_1|\mathbf{x}, \mathbf{z}) = \text{Var}(y|\mathbf{x}, \mathbf{z})$  and  $\text{Var}(u_2|\mathbf{x}) = \text{Var}(y|\mathbf{x})$ . By assumption, these are constant and necessarily equal to  $\sigma_1^2 \equiv \text{Var}(u_1)$  and  $\sigma_2^2 \equiv \text{Var}(u_2)$ , respectively. But then Property CV.4 implies that  $\sigma_2^2 \geq \sigma_1^2$ . This simple conclusion means that, when error variances are constant, the error variance falls as more explanatory variables are conditioned on.

**2.6. a.** By linearity of the linear projection,

$$L(q|1, \mathbf{x}) = L(q^*|1, \mathbf{x}) + L(e|1, \mathbf{x}) = L(q^*|1, \mathbf{x}),$$

where the last inequality follows because  $L(e|1, \mathbf{x}) = 0$  when  $E(e) = 0$  and  $E(\mathbf{x}'e) = \mathbf{0}$ .

Therefore, the parameters in the linear projection of  $q$  onto  $(1, \mathbf{x})$  are the same as the linear projection of  $q^*$  onto  $(1, \mathbf{x})$ . This fact is useful for studying equations with measurement error in the explained or explanatory variables.

$$\begin{aligned} \text{b. } r &= q - L(q|1, \mathbf{x}) = (q^* + e) - L(q|1, \mathbf{x}) = (q^* + e) - L(q^*|1, \mathbf{x}) \text{ (from part a)} \\ &= [q^* - L(q^*|1, \mathbf{x})] + e = r^* + e. \end{aligned}$$

**2.7.** Write the equation in error form as

$$\begin{aligned} y &= g(\mathbf{x}) + \mathbf{z}\beta + u \\ E(u|\mathbf{x}, \mathbf{z}) &= 0. \end{aligned}$$

Take the expected value of the first equation conditional only on  $\mathbf{x}$ :

$$E(y|\mathbf{x}) = g(\mathbf{x}) + [E(\mathbf{z}|\mathbf{x})]\boldsymbol{\beta}$$

and subtract this from the first equation to get

$$y - E(y|\mathbf{x}) = [\mathbf{z} - E(\mathbf{z}|\mathbf{x})]\boldsymbol{\beta} + u$$

or

$$\tilde{y} = \tilde{\mathbf{z}}\boldsymbol{\beta} + u$$

Because  $\tilde{\mathbf{z}}$  is a function of  $(\mathbf{x}, \mathbf{z})$ ,  $E(u|\tilde{\mathbf{z}}) = 0$  [since  $E(u|\mathbf{x}, \mathbf{z}) = 0$ ], and so  $E(\tilde{y}|\tilde{\mathbf{z}}) = \tilde{\mathbf{z}}\boldsymbol{\beta}$ .

This basic result is fundamental in the literature on estimating *partial linear models*. First, one estimates  $E(y|\mathbf{x})$  and  $E(\mathbf{z}|\mathbf{x})$  using very flexible methods (typically, *nonparametric methods*). Then, after obtaining residuals of the form  $\tilde{y}_i \equiv y_i - \hat{E}(y_i|\mathbf{x}_i)$  and  $\tilde{\mathbf{z}}_i \equiv \mathbf{z}_i - \hat{E}(\mathbf{z}_i|\mathbf{x}_i)$ ,  $\boldsymbol{\beta}$  is estimated from an OLS regression  $\tilde{y}_i$  on  $\tilde{\mathbf{z}}_i, i = 1, \dots, N$ . Under general conditions, this kind of nonparametric partialling-out procedure leads to a  $\sqrt{N}$ -consistent, asymptotically normal estimator of  $\boldsymbol{\beta}$ . See Robinson (1988) and Powell (1994).

In the case where  $E(y|\mathbf{x})$  and the elements of  $E(\mathbf{z}|\mathbf{x})$  are approximated as linear functions of a common set of functions, say  $\{h_1(\mathbf{x}), \dots, h_Q(\mathbf{x})\}$ , the partialling out is equivalent to estimating a linear model

$$y = \alpha_0 + \alpha_1 h_1(\mathbf{x}) + \dots + \alpha_Q h_Q(\mathbf{x}) + \mathbf{x}\boldsymbol{\beta} + \text{error}$$

by OLS.

**2.8. a.** By exponentiation we can write  $y = \exp[g(\mathbf{x}) + u] = \exp[g(\mathbf{x})] \exp(u)$ . It follows that

$$E(y|\mathbf{x}) = \exp[g(\mathbf{x})]E[\exp(u)|\mathbf{x}] = \exp[g(\mathbf{x})]a(\mathbf{x})$$

Using the product rule gives

$$\begin{aligned}\frac{\partial E(y|\mathbf{x})}{\partial x_j} &= \frac{\partial g(\mathbf{x})}{\partial x_j} \exp[g(\mathbf{x})] a(\mathbf{x}) + \exp[g(\mathbf{x})] \frac{\partial a(\mathbf{x})}{\partial x_j} \\ &= \frac{\partial g(\mathbf{x})}{\partial x_j} E(y|\mathbf{x}) + E(y|\mathbf{x}) \frac{\partial a(\mathbf{x})}{\partial x_j} \cdot \frac{1}{a(\mathbf{x})}\end{aligned}$$

Therefore,

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j} \cdot \frac{x_j}{E(y|\mathbf{x})} = \frac{\partial g(\mathbf{x})}{\partial x_j} \cdot x_j + \frac{\partial a(\mathbf{x})}{\partial x_j} \cdot \frac{x_j}{a(\mathbf{x})}$$

We can establish this relationship more simply by assuming  $E(y|\mathbf{x}) > 0$  for all  $\mathbf{x}$  and using equation (2.10).

b. Write  $z_j \equiv \log(x_j)$  so  $x_j = \exp(z_j)$ . Then, using the chain rule,

$$\frac{\partial g(\mathbf{x})}{\partial \log(x_j)} = \frac{\partial g(\mathbf{x})}{\partial z_j} = \frac{\partial g(\mathbf{x})}{\partial x_j} \cdot \frac{\partial x_j}{\partial z_j} = \frac{\partial g(\mathbf{x})}{\partial x_j} \cdot \exp(z_j) = \frac{\partial g(\mathbf{x})}{\partial x_j} \cdot x_j$$

c. From  $\log(y) = g(\mathbf{x}) + u$  and  $E(u|\mathbf{x}) = 0$  we have  $E[\log(y)|\mathbf{x}] = g(\mathbf{x})$ . Therefore, using (2.11), the elasticity would be simply

$$\frac{\partial g(\mathbf{x})}{\partial \log(x_j)} = \frac{\partial g(\mathbf{x})}{\partial x_j} \cdot x_j$$

which, compared with the definition based on  $E(y|\mathbf{x})$ , omits the elasticity of  $a(\mathbf{x})$  with respect to  $x_j$ .

**2.9.** This is easily shown by using iterated expectations:

$$E(\mathbf{x}'y) = E[E(\mathbf{x}'y|\mathbf{x})] = E[\mathbf{x}'E(y|\mathbf{x})] = E[\mathbf{x}'\mu(\mathbf{x})]$$

Therefore,

$$\delta = [E(\mathbf{x}'\mathbf{x})]^{-1} E(\mathbf{x}'y) = [E(\mathbf{x}'\mathbf{x})]^{-1} E[\mathbf{x}'\mu(\mathbf{x})]$$

and the latter equation is the vector of parameters in the linear projection of  $\mu(\mathbf{x})$  on  $\mathbf{x}$ .

**2.10.** a. As given in the hint, we can always write



$$E(y|\mathbf{x}, s) = (1 - s) \cdot \mu_0(\mathbf{x}) + s \cdot \mu_1(\mathbf{x})$$

Now condition only on  $s$  and use iterated expectations:

$$\begin{aligned} E(y|s) &= E[E(y|\mathbf{x}, s)|s] = E[(1 - s) \cdot \mu_0(\mathbf{x}) + s \cdot \mu_1(\mathbf{x})|s] \\ &= (1 - s)E[\mu_0(\mathbf{x})|s] + sE[\mu_1(\mathbf{x})|s] \end{aligned}$$

Therefore,

$$\begin{aligned} E(y|s = 1) &= E[\mu_1(\mathbf{x})|s = 1] \\ E(y|s = 0) &= E[\mu_0(\mathbf{x})|s = 0] \end{aligned}$$

and so, by adding and subtracting  $E[\mu_0(\mathbf{x})|s = 1]$ , we get

$$\begin{aligned} E(y|s = 1) - E(y|s = 0) &= E[\mu_1(\mathbf{x})|s = 1] - E[\mu_0(\mathbf{x})|s = 0] \\ &= \{E[\mu_1(\mathbf{x})|s = 1] - E[\mu_0(\mathbf{x})|s = 1]\} + \{E[\mu_0(\mathbf{x})|s = 1] - E[\mu_0(\mathbf{x})|s = 0]\} \end{aligned}$$

b. Use part a and linearity of the conditional means:

$$\begin{aligned} E(y|s = 1) - E(y|s = 0) &= [E(\mathbf{x}|s = 1)\boldsymbol{\beta}_1 - E(\mathbf{x}|s = 1)\boldsymbol{\beta}_0] + [E(\mathbf{x}|s = 1)\boldsymbol{\beta}_0 - E(\mathbf{x}|s = 0)\boldsymbol{\beta}_0] \\ &= E(\mathbf{x}|s = 1) \cdot (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + [E(\mathbf{x}|s = 1) - E(\mathbf{x}|s = 0)] \cdot \boldsymbol{\beta}_0 \end{aligned}$$

This decomposition attributes the difference in the unconditional means,

$E(y|s = 1) - E(y|s = 0)$ , to two pieces. The first part is due to differences in the regression parameters,  $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$  – where we evaluate the difference at the average of the covariates from the  $s = 1$  subpopulation. The second part is due to a difference in means of the covariates from the two subpopulations – where we apply the regression coefficients from the  $s = 0$  subpopulation. If, for example, the two regression functions are the same – that is,  $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0$  – then any difference in the subpopulation means  $E(y|s = 0)$  and  $E(y|s = 1)$  is due to a difference in averages of the covariates across the subpopulations. If the covariate means are the same – that is,  $E(\mathbf{x}|s = 1) = E(\mathbf{x}|s = 0)$  – then  $E(y|s = 1)$  and  $E(y|s = 0)$  can still differ if

$\beta_1 \neq \beta_0$ . In many applications, both pieces in  $E(y|s = 1) - E(y|s = 0)$  are present.

Incidentally, the approach in this problem is not the only interesting way to decompose  $E(y|s = 1) - E(y|s = 0)$ . See, for example, T.E. Elder, J.H. Goddeeris, and S.J. Haider, “Unexplained Gaps and Oaxaca–Blinder Decompositions,” *Labour Economics*, 2010.

## Solutions to Chapter 3 Problems

**3.1.** To prove Lemma 3.1, we must show that for all  $\varepsilon > 0$ , there exists  $b_\varepsilon < \infty$  and an integer  $N_\varepsilon$  such that  $P[|x_N| \geq b_\varepsilon] < \varepsilon$ , all  $N \geq N_\varepsilon$ . We use the following fact: since  $x_N \xrightarrow{p} a$ , for any  $\varepsilon > 0$  there exists an integer  $N_\varepsilon$  such that  $P[|x_N - a| > 1] < \varepsilon$  for all  $N \geq N_\varepsilon$ . [The existence of  $N_\varepsilon$  is implied by Definition 3.3(1).] But  $|x_N| = |x_N - a + a| \leq |x_N - a| + |a|$  (by the triangle inequality), and so  $|x_N| - |a| \leq |x_N - a|$ . It follows that  $P[|x_N| - |a| > 1] \leq P[|x_N - a| > 1]$ . Therefore, in Definition 3.3(3) we can take  $b_\varepsilon \equiv |a| + 1$  (irrespective of the value of  $\varepsilon$ ) and then the existence of  $N_\varepsilon$  follows from Definition 3.3(1).

**3.2.** Each element of the  $K \times 1$  vector  $\mathbf{Z}_N' \mathbf{x}_N$  is the sum of  $J$  terms of the form  $Z_{Nji} x_{Nj}$ . Because  $Z_{Nji} = o_p(1)$  and  $x_{Nj} = O_p(1)$ , each term in the sum is  $o_p(1)$  from Lemma 3.2(4). By Lemma 3.2(1), the sum of  $o_p(1)$  terms is  $o_p(1)$ .

**3.3.** This follows immediately from Lemma 3.1 because  $\mathbf{g}(\mathbf{x}_N) \xrightarrow{p} \mathbf{g}(\mathbf{c})$ .

**3.4.** Both parts follow from the continuous mapping theorem and basic properties of the normal distribution.

a. The function defined by  $\mathbf{g}(\mathbf{z}) = \mathbf{A}' \mathbf{z}$  is clearly continuous. Further, if  $\mathbf{z} \sim \text{Normal}(\mathbf{0}, \mathbf{V})$  then  $\mathbf{A}' \mathbf{z} \sim \text{Normal}(\mathbf{0}, \mathbf{A}' \mathbf{V} \mathbf{A})$ . By the continuous mapping theorem,

$$\mathbf{A}' \mathbf{z}_N \xrightarrow{d} \mathbf{A}' \mathbf{z} \sim \text{Normal}(\mathbf{0}, \mathbf{A}' \mathbf{V} \mathbf{A}).$$

b. Because  $\mathbf{V}$  is nonsingular, the function  $g(\mathbf{z}) = \mathbf{z}' \mathbf{V}^{-1} \mathbf{z}$  is continuous. But if  $\mathbf{z} \sim \text{Normal}(\mathbf{0}, \mathbf{V})$ ,  $\mathbf{z}' \mathbf{V}^{-1} \mathbf{z} \sim \chi_K^2$ . So  $\mathbf{z}_N' \mathbf{V}^{-1} \mathbf{z}_N \xrightarrow{d} \mathbf{z}' \mathbf{V}^{-1} \mathbf{z} \sim \chi_K^2$ .

**3.5.** a. Because  $\text{Var}(\bar{y}_N) = \sigma^2/N$ ,  $\text{Var}[\sqrt{N}(\bar{y}_N - \mu)] = N(\sigma^2/N) = \sigma^2$ .

b. By the CLT,  $\sqrt{N}(\bar{y}_N - \mu) \overset{a}{\sim} \text{Normal}(0, \sigma^2)$ , and so  $\text{Avar}[\sqrt{N}(\bar{y}_N - \mu)] = \sigma^2$ .

c. We obtain  $\text{Avar}(\bar{y}_N)$  by dividing  $\text{Avar}[\sqrt{N}(\bar{y}_N - \mu)]$  by  $N$ . Therefore,  $\text{Avar}(\bar{y}_N) = \sigma^2/N$ .

As expected, this coincides with the actual variance of  $\bar{y}_N$ .

d. The asymptotic standard deviation of  $\bar{y}_N$  is the square root of its asymptotic variance, or  $\sigma/\sqrt{N}$ .

e. To obtain the asymptotic standard error of  $\bar{y}_N$ , we need a consistent estimator of  $\sigma$ .

Typically, the unbiased estimator of  $\sigma^2$  is used:  $\hat{\sigma}^2 = (N-1)^{-1} \sum_{i=1}^N (y_i - \bar{y}_N)^2$ , and then  $\hat{\sigma}$  is the positive square root. The asymptotic standard error of  $\bar{y}_N$  is simply  $\hat{\sigma}/\sqrt{N}$ .

**3.6.** From Definition 3.4, we need to show that for any  $0 \leq c < 1/2$ ,  $N^c(\hat{\theta}_N - \theta) = o_p(1)$ .

But

$$N^c(\hat{\theta}_N - \theta) = N^{[c-(1/2)]} \sqrt{N}(\hat{\theta}_N - \theta) = N^{[c-(1/2)]} \cdot O_p(1).$$

Because  $c < 1/2$ ,  $N^{[c-(1/2)]} = o(1)$ , and so  $N^c(\hat{\theta}_N - \theta) = o(1) \cdot O_p(1) = o_p(1)$ .

**3.7. a.** For  $\theta > 0$  the natural logarithm is a continuous function, and so

$$\text{plim}[\log(\hat{\theta})] = \log[\text{plim}(\hat{\theta})] = \log(\theta) = \gamma.$$

b. We use the delta method to find  $\text{Avar}[\sqrt{N}(\hat{\gamma} - \gamma)]$ . In the scalar case, if  $\hat{\gamma} = g(\hat{\theta})$  then  $\text{Avar}[\sqrt{N}(\hat{\gamma} - \gamma)] = [dg(\theta)/d\theta]^2 \text{Avar}[\sqrt{N}(\hat{\theta} - \theta)]$ . When  $g(\theta) = \log(\theta)$  – which is, of course, continuously differentiable –  $\text{Avar}[\sqrt{N}(\hat{\gamma} - \gamma)] = (1/\theta)^2 \text{Avar}[\sqrt{N}(\hat{\theta} - \theta)]$ .

c. In the scalar case, the asymptotic standard error of  $\hat{\gamma}$  is generally  $|dg(\hat{\theta})/d\theta| \cdot \text{se}(\hat{\theta})$ .

Therefore, for  $g(\theta) = \log(\theta)$ ,  $\text{se}(\hat{\gamma}) = \text{se}(\hat{\theta})/(\hat{\theta})$ . When  $\hat{\theta} = 4$  and

$$\text{se}(\hat{\theta}) = 2, \hat{\gamma} = \log(4) \approx 1.39 \text{ and } \text{se}(\hat{\gamma}) = 1/2.$$

d. The asymptotic  $t$  statistic for testing  $H_0 : \theta = 1$  is  $(\hat{\theta} - 1)/\text{se}(\hat{\theta}) = 3/2 = 1.5$ .

e. Because  $\gamma = \log(\theta)$ , the null of interest can also be stated as  $H_0 : \gamma = 0$ . The  $t$  statistic based on  $\hat{\gamma}$  is about  $1.39/(.5) = 2.78$ . This leads to a very strong rejection of  $H_0$ , whereas the  $t$  statistic based on  $\hat{\theta}$  is, at best, marginally significant. The lesson is that, using the Wald test,

we can change the outcome of hypotheses tests by using nonlinear transformations.

**3.8 a.** This follows by Slutsky's Theorem since the function  $g(\theta_1, \theta_2) \equiv \theta_1/\theta_2$  is continuous at all points in  $\mathbb{R}^2$  where  $\theta_2 \neq 0$ :  $\text{plim}(\hat{\theta}_1/\hat{\theta}_2) = [\text{plim}(\hat{\theta}_1)/\text{plim}(\hat{\theta}_2)] = \theta_1/\theta_2$ .

b. To find  $\text{Avar}(\hat{\gamma})$  we need to find  $\nabla_{\theta}g(\theta)$ , where  $g(\theta_1, \theta_2) = \theta_1/\theta_2$ . But  $\nabla_{\theta}g(\theta) = (1/\theta_2, -\theta_1/\theta_2^2)$ , and so  $\text{Avar}(\hat{\gamma}) = (1/\theta_2 - \theta_1/\theta_2^2)[\text{Avar}(\hat{\theta})](1/\theta_2 - \theta_1/\theta_2^2)'$ .

c. If  $\hat{\theta} = (-1.5, .5)'$  then  $\nabla_{\theta}g(\hat{\theta}) = (2, 6)$ . Therefore,  $\widehat{\text{Avar}}(\hat{\gamma}) = (2, 6)[\widehat{\text{Avar}}(\hat{\theta})](2, 6)' = 66.4$ . Taking the square root gives  $\text{se}(\hat{\gamma}) \approx 8.15$ .

**3.9.** By the delta method,

$$\text{Avar}[\sqrt{N}(\hat{\gamma} - \gamma)] = \mathbf{G}(\theta)\mathbf{V}_1\mathbf{G}(\theta)', \quad \text{Avar}[\sqrt{N}(\hat{\gamma} - \gamma)] = \mathbf{G}(\theta)\mathbf{V}_2\mathbf{G}(\theta)',$$

where  $\mathbf{G}(\theta) = \nabla_{\theta} \mathbf{g}(\theta)$  is  $Q \times P$ . Therefore,

$$\text{Avar}[\sqrt{N}(\hat{\gamma} - \gamma)] - \text{Avar}[\sqrt{N}(\hat{\gamma} - \gamma)] = \mathbf{G}(\theta)(\mathbf{V}_2 - \mathbf{V}_1)\mathbf{G}(\theta)'$$

By assumption,  $\mathbf{V}_2 - \mathbf{V}_1$  is positive semi-definite, and therefore  $\mathbf{G}(\theta)(\mathbf{V}_2 - \mathbf{V}_1)\mathbf{G}(\theta)'$  is p.s.d.

This complete the proof.

**3.10.** By assumption,  $\sigma^2 = E(w_i^2) = \text{Var}(w_i) < \infty$ . Because of the i.i.d. assumption,

$$\text{Var}(x_N) = (N^{-1/2})^2 N \sigma^2 = \sigma^2.$$

Now, Chebyshev's inequality gives that for any  $b_{\varepsilon} > 0$ ,

$$P[|x_N| \geq b_{\varepsilon}] \leq \frac{\text{Var}(x_N)}{b_{\varepsilon}^2} = \frac{\sigma^2}{b_{\varepsilon}^2}$$

Therefore, in the definition of  $O_p(1)$ , for any  $\varepsilon > 0$  choose  $b_{\varepsilon} = \sigma/\sqrt{\varepsilon}$  and  $N_{\varepsilon} = 1$  and then

$P[|x_N| \geq b_{\varepsilon}] \leq \varepsilon$  for all  $N \geq N_{\varepsilon}$ .

**3.11. a.** Let  $x_N = N^{-1} \sum_{i=1}^N (w_i - \mu_i)$  so that

$$\text{Var}(x_N) = N^{-2} \sum_{i=1}^N \text{Var}(w_i) = N^{-2} \sum_{i=1}^N \sigma_i^2$$

By Chebyshev's inequality, for any  $\varepsilon > 0$ ,

$$\mathbb{P}[|x_N| > \varepsilon] \leq \frac{\text{Var}(x_N)}{\varepsilon^2} = \frac{N^{-2} \sum_{i=1}^N \sigma_i^2}{\varepsilon^2}$$

It follows that  $\mathbb{P}[|x_N| > \varepsilon] \rightarrow 0$  as  $N \rightarrow \infty$  if  $N^{-2} \sum_{i=1}^N \sigma_i^2 \rightarrow 0$  as  $N \rightarrow \infty$ .

b. If  $\sigma_i^2 < b < \infty$  for all  $i$  – that is, the sequence of variances is bounded – then

$$N^{-2} \sum_{i=1}^N \sigma_i^2 \leq b/N \rightarrow 0 \text{ as } N \rightarrow \infty.$$

Thus, uniformly bounded variances is sufficient for i.n.i.d. sequences to satisfy the WLLN.

## Solutions to Chapter 4 Problems

4.1. a. Exponentiating equation (4.49) gives

$$\begin{aligned} wage &= \exp(\beta_0 + \beta_1 married + \beta_2 educ + \mathbf{z}\boldsymbol{\gamma} + u) \\ &= \exp(u) \exp(\beta_0 + \beta_1 married + \beta_2 educ + \mathbf{z}\boldsymbol{\gamma}). \end{aligned}$$

Therefore,

$$E(wage|\mathbf{x}) = E[\exp(u)|\mathbf{x}] \exp(\beta_0 + \beta_1 married + \beta_2 educ + \mathbf{z}\boldsymbol{\gamma}),$$

where  $\mathbf{x}$  denotes all explanatory variables. Now, if  $u$  and  $\mathbf{x}$  are independent

then  $E[\exp(u)|\mathbf{x}] = E[\exp(u)] = \delta_0$ , say. Therefore

$$E(wage|\mathbf{x}) = \delta_0 \exp(\beta_0 + \beta_1 married + \beta_2 educ + \mathbf{z}\boldsymbol{\gamma}).$$

If we set  $married = 1$  and  $married = 0$  in this expectation (keeping all else equal) and find the proportionate increase we get

$$\frac{\delta_0 \exp(\beta_0 + \beta_1 + \beta_2 educ + \mathbf{z}\boldsymbol{\gamma}) - \delta_0 \exp(\beta_0 + \beta_2 educ + \mathbf{z}\boldsymbol{\gamma})}{\delta_0 \exp(\beta_0 + \beta_2 educ + \mathbf{z}\boldsymbol{\gamma})} = \exp(\beta_1) - 1.$$

Thus, the percentage difference is  $100 \cdot [\exp(\beta_1) - 1]$ .

b. Since  $\theta_1 = 100 \cdot [\exp(\beta_1) - 1] = g(\beta_1)$ , we need the derivative of  $g$  with respect to  $\beta_1$ :

$dg/d\beta_1 = 100 \cdot \exp(\beta_1)$ . The asymptotic standard error of  $\hat{\theta}_1$  using the delta method is

obtained as the absolute value of  $d\hat{g}/d\beta_1$  times  $se(\hat{\beta}_1)$ :

$$se(\hat{\theta}_1) = 100 \cdot [\exp(\hat{\beta}_1)] \cdot se(\hat{\beta}_1).$$

c. We can evaluate the conditional expectation in part a at two levels of education, say  $educ_0$  and  $educ_1$ , all else fixed. The proportionate change in expected wage from  $educ_0$  to  $educ_1$  is

$$[\exp(\beta_2 educ_1) - \exp(\beta_2 educ_0)] / \exp(\beta_2 educ_0) = \exp[\beta_2(educ_1 - educ_0)] - 1 = \exp(\beta_2 \Delta educ) - 1.$$

Using the same arguments in part b,  $\hat{\theta}_2 = 100 \cdot [\exp(\beta_2 \Delta educ) - 1]$  and

$$se(\hat{\theta}_2) = 100 \cdot |\Delta educ| \exp(\hat{\beta}_2 \Delta educ) se(\hat{\beta}_2).$$

d. For the estimated version of equation (4.29),  $\hat{\beta}_1 = .199$ ,  $se(\hat{\beta}_1) = .039$ ,  $\hat{\beta}_2 = .065$ , and  $se(\hat{\beta}_2) = .006$ . Therefore,  $\hat{\theta}_1 = 22.01$  and  $se(\hat{\theta}_1) = 4.76$ . For  $\hat{\theta}_2$  we set  $\Delta educ = 4$ . Then  $\hat{\theta}_2 = 29.7$  and  $se(\hat{\theta}_2) = 3.11$ .

**4.2.** a. For each  $i$  we have, by OLS.2,  $E(u_i|\mathbf{X}) = 0$ . By independence across  $i$  and Property CE.5,  $E(u_i|\mathbf{X}) = E(u_i|\mathbf{x}_i)$  because  $(u_i, \mathbf{x}_i)$  is independent of the explanatory variables for all other observations. Letting  $\mathbf{U}$  be the  $N \times 1$  vector of all errors, this implies  $E(\mathbf{U}|\mathbf{X}) = \mathbf{0}$ . But  $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}$  and so

$$E(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{U}|\mathbf{X}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot \mathbf{0} = \boldsymbol{\beta}.$$

b. From the expression for  $\hat{\boldsymbol{\beta}}$  in part a we have

$$\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}(\mathbf{U}|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

Now, because  $E(\mathbf{U}|\mathbf{X}) = \mathbf{0}$ ,  $\text{Var}(\mathbf{U}|\mathbf{X}) = E(\mathbf{U}\mathbf{U}'|\mathbf{X})$ . For the diagonal terms,

$$E(u_i^2|\mathbf{X}) = E(u_i^2|\mathbf{x}_i) = \text{Var}(u_i|\mathbf{x}_i) = \sigma^2, \text{ where the last equality is the homoskedasticity}$$

assumption. For the covariance terms, we must show that  $E(u_i u_h|\mathbf{X}) = 0$  for all

$i \neq h, i, h = 1, \dots, N$ . Again using Property CE.5,  $E(u_i u_h|\mathbf{X}) = E(u_i u_h|\mathbf{x}_i, \mathbf{x}_h)$  and

$$E(u_i|\mathbf{x}_i, u_h, \mathbf{x}_h) = E(u_i|\mathbf{x}_i) = 0. \text{ But then } E(u_i u_h|\mathbf{x}_i, u_h, \mathbf{x}_h) = E(u_i|\mathbf{x}_i, u_h, \mathbf{x}_h) u_h = 0. \text{ It follows}$$

immediately by iterated expectations that conditioning on the smaller set also yields a zero

conditional mean:  $E(u_i u_h|\mathbf{x}_i, \mathbf{x}_h) = 0$ . This completes the proof.

**4.3.** a. Not in general. The conditional variance can always be written as

$$\text{Var}(u|\mathbf{x}) = E(u^2|\mathbf{x}) - [E(u|\mathbf{x})]^2; \text{ if } E(u|\mathbf{x}) \neq 0, \text{ then } E(u^2|\mathbf{x}) \neq \text{Var}(u|\mathbf{x}).$$

b. It could be that  $E(\mathbf{x}'u) = 0$ , in which case OLS is consistent, and  $\text{Var}(u|\mathbf{x})$  is constant.



But, generally, the usual standard errors would not be valid unless  $E(u|\mathbf{x}) = 0$  because it is  $E(u^2|\mathbf{x})$  that should be constant.

**4.4.** For each  $i$ ,  $\hat{u}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}} = u_i - \mathbf{x}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ , and so

$\hat{u}_i^2 = u_i^2 - 2u_i\mathbf{x}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + [\mathbf{x}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]^2$ . Therefore, we can write

$$N^{-1} \sum_{i=1}^N \hat{u}_i^2 \mathbf{x}_i' \mathbf{x}_i = N^{-1} \sum_{i=1}^N u_i^2 \mathbf{x}_i' \mathbf{x}_i - 2N^{-1} \sum_{i=1}^N [u_i \mathbf{x}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \mathbf{x}_i' \mathbf{x}_i + N^{-1} \sum_{i=1}^N [\mathbf{x}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]^2 \mathbf{x}_i' \mathbf{x}_i.$$

Dropping the "-2", the second term can be written as the sum of  $K$  terms of the form

$$N^{-1} \sum_{i=1}^N [u_i x_{ij}(\hat{\beta}_j - \beta_j)] \mathbf{x}_i' \mathbf{x}_i = (\hat{\beta}_j - \beta_j) N^{-1} \sum_{i=1}^N (u_i x_{ij}) \mathbf{x}_i' \mathbf{x}_i = o_p(1) \cdot O_p(1),$$

where we have used  $\hat{\beta}_j - \beta_j = o_p(1)$  and  $N^{-1} \sum_{i=1}^N (u_i x_{ij}) \mathbf{x}_i' \mathbf{x}_i = O_p(1)$  whenever

$E[|u_i x_{ij} x_{ih} x_{ik}|] < \infty$  for all  $j, h$ , and  $k$  (as would just be assumed). Similarly, the third term can be written as the sum of  $K^2$  terms of the form

$$(\hat{\beta}_j - \beta_j)(\hat{\beta}_h - \beta_h) N^{-1} \sum_{i=1}^N (x_{ij} x_{ih}) \mathbf{x}_i' \mathbf{x}_i = o_p(1) \cdot o_p(1) \cdot O_p(1) = o_p(1),$$

where we have used  $N^{-1} \sum_{i=1}^N (x_{ij} x_{ih}) \mathbf{x}_i' \mathbf{x}_i = O_p(1)$  whenever  $E[|x_{ij} x_{ih} x_{ik} x_{im}|] < \infty$  for all  $j, h$ ,

$k$ , and  $m$ . We have shown that  $N^{-1} \sum_{i=1}^N \hat{u}_i^2 \mathbf{x}_i' \mathbf{x}_i = N^{-1} \sum_{i=1}^N u_i^2 \mathbf{x}_i' \mathbf{x}_i + o_p(1)$ , which is what we wanted to show.

**4.5.** Write equation (4.50) as  $E(y|\mathbf{w}) = \mathbf{w}\boldsymbol{\delta}$ , where  $\mathbf{w} = (\mathbf{x}, z)$ . Since  $\text{Var}(y|\mathbf{w}) = \sigma^2$ , it follows by Theorem 4.2 that  $\text{Avar} \sqrt{N}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})$  is  $\sigma^2[E(\mathbf{w}'\mathbf{w})]^{-1}$ , where  $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\beta}}', \hat{\gamma})'$ . Importantly, because  $E(\mathbf{x}'z) = \mathbf{0}$ ,  $E(\mathbf{w}'\mathbf{w})$  is block diagonal, with upper block  $E(\mathbf{x}'\mathbf{x})$  and lower block  $E(z^2)$ . Inverting  $E(\mathbf{w}'\mathbf{w})$  and focusing on the upper  $K \times K$  block gives

$$\text{Avar} \sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sigma^2[E(\mathbf{x}'\mathbf{x})]^{-1}.$$

Next, we need to find  $\text{Avar}\sqrt{N}(\tilde{\beta} - \beta)$ . It is helpful to write  $y = \mathbf{x}\beta + v$  where  $v = \gamma z + u$  and  $u \equiv y - E(y|\mathbf{x}, z)$ . Because  $E(\mathbf{x}'z) = 0$  and  $E(\mathbf{x}'u) = 0$ ,  $E(\mathbf{x}'v) = 0$ . Further,  $E(v^2|\mathbf{x}) = \gamma^2 E(z^2|\mathbf{x}) + E(u^2|\mathbf{x}) + 2\gamma E(zu|\mathbf{x}) = \gamma^2 E(z^2|\mathbf{x}) + \sigma^2$ , where we use  $E(zu|\mathbf{x}, z) = zE(u|\mathbf{x}, z) = 0$  and  $E(u^2|\mathbf{x}, z) = \text{Var}(y|\mathbf{x}, z) = \sigma^2$ . Unless  $E(z^2|\mathbf{x})$  is constant, the equation  $y = \mathbf{x}\beta + v$  generally violates the homoskedasticity assumption OLS.3. So, without further assumptions,

$$\text{Avar}\sqrt{N}(\tilde{\beta} - \beta) = [E(\mathbf{x}'\mathbf{x})]^{-1}E(v^2\mathbf{x}'\mathbf{x})[E(\mathbf{x}'\mathbf{x})]^{-1}.$$

Now we can show  $\text{Avar}\sqrt{N}(\tilde{\beta} - \beta) - \text{Avar}\sqrt{N}(\hat{\beta} - \beta)$  is positive semi-definite by writing

$$\begin{aligned} \text{Avar}\sqrt{N}(\tilde{\beta} - \beta) - \text{Avar}\sqrt{N}(\hat{\beta} - \beta) &= [E(\mathbf{x}'\mathbf{x})]^{-1}E(v^2\mathbf{x}'\mathbf{x})[E(\mathbf{x}'\mathbf{x})]^{-1} - \sigma^2[E(\mathbf{x}'\mathbf{x})]^{-1} \\ &= [E(\mathbf{x}'\mathbf{x})]^{-1}E(v^2\mathbf{x}'\mathbf{x})[E(\mathbf{x}'\mathbf{x})]^{-1} - \sigma^2[E(\mathbf{x}'\mathbf{x})]^{-1}E(\mathbf{x}'\mathbf{x})[E(\mathbf{x}'\mathbf{x})]^{-1} \\ &= [E(\mathbf{x}'\mathbf{x})]^{-1}[E(v^2\mathbf{x}'\mathbf{x}) - \sigma^2E(\mathbf{x}'\mathbf{x})][E(\mathbf{x}'\mathbf{x})]^{-1} \end{aligned}$$

Because  $[E(\mathbf{x}'\mathbf{x})]^{-1}$  is positive definite, it suffices to show that  $E(v^2\mathbf{x}'\mathbf{x}) - \sigma^2E(\mathbf{x}'\mathbf{x})$  is p.s.d.

To this end, let  $h(\mathbf{x}) \equiv E(z^2|\mathbf{x})$ . Then by the law of iterated expectations,

$$E(v^2\mathbf{x}'\mathbf{x}) = E[E(v^2|\mathbf{x})\mathbf{x}'\mathbf{x}] = \gamma^2 E[h(\mathbf{x})\mathbf{x}'\mathbf{x}] + \sigma^2 E(\mathbf{x}'\mathbf{x}). \text{ Therefore,}$$

$E(v^2\mathbf{x}'\mathbf{x}) - \sigma^2 E(\mathbf{x}'\mathbf{x}) = \gamma^2 E[h(\mathbf{x})\mathbf{x}'\mathbf{x}]$ , which, when  $\gamma \neq 0$ , is actually a positive definite matrix except by fluke. In particular, if  $E(z^2|\mathbf{x}) = E(z^2) = \eta^2 > 0$  (in which case  $y = \mathbf{x}\beta + v$  satisfies the homoskedasticity assumption OLS.3),  $E(v^2\mathbf{x}'\mathbf{x}) - \sigma^2 E(\mathbf{x}'\mathbf{x}) = \gamma^2 \eta^2 E(\mathbf{x}'\mathbf{x})$ , which is positive definite.

**4.6.** Because *nonwhite* is determined at birth, we do not have to worry about *nonwhite* being determined simultaneously with any kind of response variable. Measurement error is certainly a possibility, as a binary indicator for being Caucasian is a very crude way to measure race. Still, many studies hope to isolate systematic differences between those classified as white versus other races, in which case a binary indicator might be a good proxy. Of course, it

is always possible that people are misclassified in survey data. But an important point is that measurement error in *nonwhite* would not follow the classical errors-in-variables assumption. For example, if the issue is simply recording the incorrect entry, then the true indicator, *nonwhite\**, is also binary. Then, there are four possible outcomes: *nonwhite\** = 1 and *nonwhite* = 1; *nonwhite\** = 0 and *nonwhite* = 1; *nonwhite\** = 1 and *nonwhite* = 0; *nonwhite\** = 0 and *nonwhite* = 0. In the first and last cases, no error is made. Generally, it makes no sense to write *nonwhite* = *nonwhite\** + *e*, where *e* is a mean-zero measurement error that is independent of *nonwhite\**.

Probably in applications that seek to estimate a race effect, we would be most concerned about omitted variables. While race is determined at birth, it is not independent of other factors that generally affect economic and social outcomes. For example, we would want to include family income and wealth in an equation to test for discrimination in loan applications. If we cannot, and race is correlated with income and wealth, then an attempt to test for discrimination can fail. Many other applications could suffer from endogeneity caused by omitted variables. In looking at crime rates by race, we also need to control for family background characteristics.

4.7. a. One important omitted factor in *u* is family income: students that come from wealthier families tend to do better in school, other things equal. Family income and PC ownership are positively correlated because the probability of owning a PC increases with family income. Another factor in *u* is quality of high school. This may also be correlated with *PC*: a student who had more exposure with computers in high school may be more likely to own a computer.

b.  $\hat{\beta}_3$  is *likely* to have an upward bias because of the positive correlation between *u* and *PC*,

but it is not clear cut because of the other explanatory variables in the equation. If we write the linear projection

$$u = \delta_0 + \delta_1 hsGPA + \delta_2 SAT + \delta_3 PC + r$$

then the bias is upward if  $\delta_3$  is greater than zero. This measures the partial correlation between  $u$  (say, family income) and  $PC$ , and it is likely to be positive.

c. If data on family income can be collected then it can be included in the equation. If family income is not available sometimes level of parents' education is. Another possibility is to use average house value in each student's home zip code, as zip code is often part of school records. Proxies for high school quality might be faculty–student ratios, expenditure per student, average teacher salary, and so on.

**4.8.** a.  $\partial E(y|x_1, x_2)/\partial x_1 = \beta_1 + \beta_3 x_2$ . Taking the expected value of this equation with respect to the distribution of  $x_2$  gives  $\alpha_1 \equiv \beta_1 + \beta_3 \mu_2$ . Similarly,

$\partial E(y|x_1, x_2)/\partial x_2 = \beta_2 + \beta_3 x_1 + 2\beta_4 x_2$ , and its expected value is  $\alpha_2 \equiv \beta_2 + \beta_3 \mu_1 + 2\beta_4 \mu_2$ .

b. One way to write  $E(y|x_1, x_2)$  is

$$E(y|x_1, x_2) = \delta_0 + \alpha_1 x_1 + \alpha_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + \beta_4 (x_2 - \mu_2)^2,$$

where  $\delta_0 = \beta_0 + \beta_3 \mu_1 \mu_2 - \beta_4 \mu_2^2$  (as can be verified by matching the intercepts in the two equations).

c. Regress  $y_i$  on  $1, x_{i1}, x_{i2}, (x_{i1} - \mu_1)(x_{i2} - \mu_2), (x_{i2} - \mu_2)^2, i = 1, 2, \dots, N$ . If we do not know  $\mu_1$  and  $\mu_2$ , we can estimate these using the sample averages,  $\bar{x}_1$  and  $\bar{x}_2$ .

d. The following Stata session can be used to answer this part:

```
. sum educ exper
```

Variable	Obs	Mean	Std. Dev.	Min	Max
educ	935	13.46845	2.196654	9	18
exper	935	11.56364	4.374586	1	23

```
. gen educ0exper0 = (educ - 13.47)*(exper - 11.56)
. gen exper0sq = (exper - 11.56)^2
. reg lwage educ exper educ0exper0 exper0sq
```

Source	SS	df	MS	Number of obs =	935
Model	22.7093743	4	5.67734357	F( 4, 930) =	36.94
Residual	142.946909	930	.153706354	Prob > F =	0.0000
				R-squared =	0.1371
				Adj R-squared =	0.1334
Total	165.656283	934	.177362188	Root MSE =	.39205

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0837981	.0069787	12.01	0.000	.0701022	.097494
exper	.0223954	.0034481	6.49	0.000	.0156284	.0291624
educ0exper0	.0045485	.0017652	2.58	0.010	.0010843	.0080127
exper0sq	.0009943	.000653	1.52	0.128	-.0002872	.0022758
_cons	5.392285	.1207342	44.66	0.000	5.155342	5.629228

```
. gen educexper = educ*exper
. gen expersq = exper^2
. reg lwage educ exper educexper expersq
```

Source	SS	df	MS	Number of obs =	935
Model	22.7093743	4	5.67734357	F( 4, 930) =	36.94
Residual	142.946909	930	.153706354	Prob > F =	0.0000
				R-squared =	0.1371
				Adj R-squared =	0.1334
Total	165.656283	934	.177362188	Root MSE =	.39205

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0312176	.0193142	1.62	0.106	-.0066869	.0691221
exper	-.0618608	.0331851	-1.86	0.063	-.1269872	.0032656
educexper	.0045485	.0017652	2.58	0.010	.0010843	.0080127
expersq	.0009943	.000653	1.52	0.128	-.0002872	.0022758
_cons	6.233415	.3044512	20.47	0.000	5.635924	6.830906

In the equation where *educ* and *exper* are both demeaned before creating the interaction and the squared terms, the coefficients on *educ* and *exper* seem reasonable. For example, the coefficient on *educ* means that, at the average level of experience, the return to another year of education is about 8.4%. As experience increases above its average value, the return to

education also increases (by .45 percentage points for each year of experience above 11.56). In the model containing  $educ \cdot exper$  and  $exper^2$ , the coefficient on  $educ$  is the return to education when  $exper = 0$  – not an especially interesting segment of the population, and certainly not representative of the men in the sample. (Notice that the standard error of  $\hat{\beta}_{educ}$  in the second regression is almost three times the standard error in the first regression. This difference illustrates that we can estimate the marginal effect at the average values of the covariates much more precisely than at extreme values of the covariates.) The coefficient on  $exper$  in the first regression is the return to another year of experience at the average values of both  $educ$  and  $exper$ . So, for a man with about 13.5 years of education and 11.6 years of experience, another year of experience is estimated to be worth about 2.2%. In the second regression, where  $educ$  and  $exper$  are not first demeaned, the coefficient on  $exper$  is the return to the first year of experience for a man with no schooling. This is not an interesting part of the U.S. population, and, in a sample where the lowest completed grade is ninth, we have no hope of estimating such an effect, anyway. The negative, large coefficient on  $exper$  in the second regression is puzzling only when we forget what it actually estimates. Note that the standard error on  $\hat{\beta}_{exper}$  in the second regression is about 10 times as large as the standard error in the first regression.

**4.9.** a. Just subtract  $\log(y_{-1})$  from both sides and define  $\Delta \log(y) = \log(y) - \log(y_{-1})$ :

$$\Delta \log(y) = \beta_0 + \mathbf{x}\boldsymbol{\beta} + (\alpha_1 - 1)\log(y_{-1}) + u.$$

Clearly, the intercept and slope estimates on  $\mathbf{x}$  will be the same. The coefficient on  $\log(y_{-1})$  becomes  $\alpha_1 - 1$ .

b. For simplicity, let  $w = \log(y)$  and  $w_{-1} = \log(y_{-1})$ . Then the population slope coefficient in a simple regression is always  $\alpha_1 = \text{Cov}(w_{-1}, w) / \text{Var}(w_{-1})$ . By assumption,

$\text{Var}(w) = \text{Var}(w_{-1})$ , which means we can write  $\alpha_1 = \text{Cov}(w_{-1}, w)/(\sigma_{w_{-1}}\sigma_w)$ , where  $\sigma_{w_{-1}} = \text{sd}(w_{-1})$  and  $\sigma_w = \text{sd}(w)$ . But  $\text{Corr}(w_{-1}, w) = \text{Cov}(w_{-1}, w)/(\sigma_{w_{-1}}\sigma_w)$ , and since a correlation coefficient is always between  $-1$  and  $1$ , the result follows.

**4.10.** Write the linear projection of  $x_K^*$  onto the other explanatory variables as

$x_K^* = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \dots + \delta_{K-1} x_{K-1} + r_K^*$ . Now, because  $x_K = x_K^* + e_K$ ,

$$\begin{aligned} L(x_K|1, x_1, \dots, x_{K-1}) &= L(x_K^*|1, x_1, \dots, x_{K-1}) + L(e_K|1, x_1, \dots, x_{K-1}) \\ &= L(x_K^*|1, x_1, \dots, x_{K-1}) \end{aligned}$$

because  $e_K$  has zero mean and is uncorrelated with  $x_1, \dots, x_{K-1}$  [and so

$L(e_K|1, x_1, \dots, x_{K-1}) = 0$ ]. But the linear projection error  $r_K$  is

$$r_K \equiv x_K - L(x_K|1, x_1, \dots, x_{K-1}) = [x_K^* - L(x_K^*|1, x_1, \dots, x_{K-1})] + e_K = r_K^* + e_K.$$

Now we can use the two-step projection formula: the coefficient on  $x_K$  in  $L(y|1, x_1, \dots, x_K)$  is the coefficient in  $L(y|r_K)$ , say  $\pi_1$ . But

$$\pi_1 = \text{Cov}(r_K, y)/\text{Var}(r_K) = \beta_K \text{Cov}(r_K^*, x_K^*)/\text{Var}(r_K)$$

since  $e_K$  is uncorrelated with  $x_1, \dots, x_{K-1}, x_K^*$ , and  $v$  by assumption and  $r_K^*$  is uncorrelated with

$x_1, \dots, x_{K-1}$ , by definition. Now  $\text{Cov}(r_K^*, x_K^*) = \text{Var}(r_K^*)$  and  $\text{Var}(r_K) = \text{Var}(r_K^*) + \text{Var}(e_K)$

[because  $\text{Cov}(r_K^*, e_K) = 0$ ]. Therefore  $\pi_1$  is given by equation (4.47), which is what we wanted to show.

**4.11.** Here is some Stata output obtained to answer this question:

```
. reg lwage exper tenure married south urban black educ iq kww
```

Source	SS	df	MS	Number of obs =	935
Model	44.0967944	9	4.89964382	F( 9, 925) =	37.28
Residual	121.559489	925	.131415664	Prob > F =	0.0000
Total	165.656283	934	.177362188	R-squared =	0.2662
				Adj R-squared =	0.2591
				Root MSE =	.36251

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-------	-------	-----------	---	------	----------------------

exper	.0127522	.0032308	3.95	0.000	.0064117	.0190927
tenure	.0109248	.0024457	4.47	0.000	.006125	.0157246
married	.1921449	.0389094	4.94	0.000	.1157839	.2685059
south	-.0820295	.0262222	-3.13	0.002	-.1334913	-.0305676
urban	.1758226	.0269095	6.53	0.000	.1230118	.2286334
black	-.1303995	.0399014	-3.27	0.001	-.2087073	-.0520917
educ	.0498375	.007262	6.86	0.000	.0355856	.0640893
iq	.0031183	.0010128	3.08	0.002	.0011306	.0051059
kww	.003826	.0018521	2.07	0.039	.0001911	.0074608
_cons	5.175644	.127776	40.51	0.000	4.924879	5.426408

```
. test iq kww
```

```
( 1)  iq = 0
( 2)  kww = 0
```

```
F( 2, 925) = 8.59
Prob > F = 0.0002
```

a. The estimated return to education using both *IQ* and *KWW* as proxies for ability is about 5%. When we used no proxy the estimated return was about 6.5%, and with only *IQ* as a proxy it was about 5.4%. Thus, we have an even lower estimated return to education, but it is still practically nontrivial and statistically very significant.

b. We can see from the *t* statistics that these variables are going to be jointly significant. The *F* test verifies this, with *p*-value = .0002.

c. The wage differential between nonblacks and blacks does not disappear. Blacks are estimated to earn about 13% less than nonblacks, holding other factors in the regression fixed.

d. Adding the interaction terms described in the problem gives the following results:

```
. sum iq kww
```

Variable	Obs	Mean	Std. Dev.	Min	Max
iq	935	101.2824	15.05264	50	145
kww	935	35.74439	7.638788	12	56

```
. gen educiq0 = educ*(iq - 100)
```

```
. gen educkww0 = educ*(kww - 35.74)
```

```
. reg lwage exper tenure married south urban black educ iq kww educiq0 educkww0
```

```
Source |          SS          df          MS          Number of obs =          935
```



Model	45.1916886	11	4.10833533	F( 11, 923) = 31.48
Residual	120.464595	923	.130514187	Prob > F = 0.0000
Total	165.656283	934	.177362188	R-squared = 0.2728
				Adj R-squared = 0.2641
				Root MSE = .36127

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
exper	.0121544	.0032358	3.76	0.000	.005804	.0185047
tenure	.0107206	.0024383	4.40	0.000	.0059353	.015506
married	.1978269	.0388272	5.10	0.000	.1216271	.2740267
south	-.0807609	.0261374	-3.09	0.002	-.1320565	-.0294652
urban	.178431	.026871	6.64	0.000	.1256957	.2311664
black	-.1381481	.0399615	-3.46	0.001	-.2165741	-.0597221
educ	.0452316	.0076472	5.91	0.000	.0302235	.0602396
iq	.0048228	.0057333	0.84	0.400	-.006429	.0160745
kww	-.0248007	.0107382	-2.31	0.021	-.0458749	-.0037266
educiq0	-.0001138	.0004228	-0.27	0.788	-.0009436	.0007161
educkww0	.002161	.0007957	2.72	0.007	.0005994	.0037227
_cons	6.080005	.5610875	10.84	0.000	4.978849	7.18116

```
. test educiq0 educkww0
```

```
( 1) educiq0 = 0
( 2) educkww0 = 0
```

```
F( 2, 923) = 4.19
Prob > F = 0.0154
```

The interaction *educkww0* is statistically significant, and the two interactions are jointly significant at the 2% significance level. The estimated return to education at the average values of *IQ* and *KWW* (in the population and sample, respectively) is somewhat smaller now: about 4.5%. Further, as *KWW* increases above its mean, the return to education increases. For example, if *KWW* is about one standard deviation (7.64) above its mean, the return to education is about  $.045 + .0022(7.6) = .06172$ , or about 6.2%. So “knowledge of the world of work” interacts positively with education levels.

**4.12.** Here is the Stata output when *union* is added to both equations:

```
. reg lscrap grant union if d88
```

Source	SS	df	MS	Number of obs = 54
Model	4.59902319	2	2.29951159	F( 2, 51) = 1.16
Residual	100.763637	51	1.97575759	Prob > F = 0.3204
				R-squared = 0.0436

-----					Adj R-squared = 0.0061	
Total		105.36266	53	1.98797472	Root MSE = 1.4056	
-----						
lscrap		Coef.	Std. Err.	t	P> t	[95% Conf. Interval
-----						
grant		-.0276192	.4043649	-0.07	0.946	-.8394156 .7841772
union		.6222888	.4096347	1.52	0.135	-.2000873 1.444665
_cons		.2307292	.2648551	0.87	0.388	-.3009896 .762448
-----						

```
. reg lscrap grant union lscrap_1 if d88
```

Source	SS	df	MS	Number of obs = 54		
-----				F( 3, 50) = 122.33		
Model	92.7289733	3	30.9096578	Prob > F = 0.0000		
Residual	12.6336868	50	.252673735	R-squared = 0.8801		
-----				Adj R-squared = 0.8729		
Total	105.36266	53	1.98797472	Root MSE = .50267		
-----						
lscrap	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
-----						
grant	-.2851103	.1452619	-1.96	0.055	-.5768775	.0066568
union	.2580653	.1477832	1.75	0.087	-.0387659	.5548965
lscrap_1	.8210298	.043962	18.68	0.000	.7327295	.90933
_cons	-.0477754	.0958824	-0.50	0.620	-.2403608	.14481
-----						

The basic story does not change: initially, the grant is estimated to have essentially no effect, but adding  $\log(\text{scrap}_{-1})$  gives the grant a strong effect that is marginally statistically significant. Interestingly, unionized firms are estimated to have larger scrap rates; over 25% more in the second equation. The effect is significant at the 10% level.

#### 4.13. a. Using the 90 counties for 1987 gives

```
. reg lcrmte lprbarr lprbconv lprbpris lavgsen if d87
```

Source	SS	df	MS	Number of obs = 90		
-----				F( 4, 85) = 15.15		
Model	11.1549601	4	2.78874002	Prob > F = 0.0000		
Residual	15.6447379	85	.18405574	R-squared = 0.4162		
-----				Adj R-squared = 0.3888		
Total	26.799698	89	.301120202	Root MSE = .42902		
-----						
lcrmte	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
-----						
lprbarr	-.7239696	.1153163	-6.28	0.000	-.9532493	-.4946898
lprbconv	-.4725112	.0831078	-5.69	0.000	-.6377519	-.3072706
lprbpris	.1596698	.2064441	0.77	0.441	-.2507964	.570136
lavgsen	.0764213	.1634732	0.47	0.641	-.2486073	.4014499
-----						

_cons		-4.867922	.4315307	-11.28	0.000	-5.725921	-4.009923
-------	--	-----------	----------	--------	-------	-----------	-----------

Because of the log-log functional form, all coefficients are elasticities. The elasticities of crime with respect to the arrest and conviction probabilities are the sign we expect, and both are practically and statistically significant. The elasticities with respect to the probability of serving a prison term and the average sentence length are positive but are statistically insignificant.

b. To add the previous year's crime rate we first generate the first lag of *lcrmrte*:

```
. xtset county year
      panel variable:  county (strongly balanced)
      time variable:  year, 81 to 87
      delta:  1 unit
```

```
. gen lcrmrte_1 = L.lcrmrte
(90 missing values generated)
```

```
. reg lcrmrte lprbarr lprbconv lprbpris lavgsen lcrmrte_1 if d87
```

Source		SS	df	MS	Number of obs =	90
Model		23.3549731	5	4.67099462	F( 5, 84) =	113.90
Residual		3.4447249	84	.04100863	Prob > F =	0.0000
Total		26.799698	89	.301120202	R-squared =	0.8715
					Adj R-squared =	0.8638
					Root MSE =	.20251

lcrmrte		Coef.	Std. Err.	t	P> t	[95% Conf. Interval
lprbarr		-.1850424	.0627624	-2.95	0.004	-.3098523 -.0602325
lprbconv		-.0386768	.0465999	-0.83	0.409	-.1313457 .0539921
lprbpris		-.1266874	.0988505	-1.28	0.204	-.3232625 .0698876
lavgsen		-.1520228	.0782915	-1.94	0.056	-.3077141 .0036684
lcrmrte_1		.7798129	.0452114	17.25	0.000	.6899051 .8697208
_cons		-.7666256	.3130986	-2.45	0.016	-1.389257 -.1439946

There are some notable changes in the coefficients on the original variables. The elasticities with respect to *prbarr* and *prbconv* are much smaller now, but still have signs predicted by a deterrent-effect story. The conviction probability is no longer statistically significant. Adding the lagged crime rate changes the signs of the elasticities with respect to *prbpris* and *avgsen*, and the latter is almost statistically significant at the 5% level against a

two-sided alternative ( $p$ -value = .056). Not surprisingly, the elasticity with respect to the lagged crime rate is large and very statistically significant. (The elasticity is also statistically less than unity.)

c. Adding the logs of the nine wage variables gives the following:

```
. reg lcrmte lprbarr lprbconv lprbpris lavgsen lcrmte_1 lwcon- lwloc if d87
```

Source	SS	df	MS	Number of obs =	90
Model	23.8798774	14	1.70570553	F( 14, 75) =	43.81
Residual	2.91982063	75	.038930942	Prob > F =	0.0000
Total	26.799698	89	.301120202	R-squared =	0.8911
				Adj R-squared =	0.8707
				Root MSE =	.19731

lcrmte	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
lprbarr	-.1725122	.0659533	-2.62	0.011	-.3038978	-.0411265
lprbconv	-.0683639	.049728	-1.37	0.173	-.1674273	.0306994
lprbpris	-.2155553	.1024014	-2.11	0.039	-.4195493	-.0115614
lavgsen	-.1960546	.0844647	-2.32	0.023	-.364317	-.0277923
lcrmte_1	.7453414	.0530331	14.05	0.000	.6396942	.8509887
lwcon	-.2850008	.1775178	-1.61	0.113	-.6386344	.0686327
lwtuc	.0641312	.134327	0.48	0.634	-.2034619	.3317244
lwtrd	.253707	.2317449	1.09	0.277	-.2079525	.7153665
lwfir	-.0835258	.1964974	-0.43	0.672	-.4749687	.3079171
lwser	.1127542	.0847427	1.33	0.187	-.0560619	.2815703
lwmfg	.0987371	.1186099	0.83	0.408	-.1375459	.3350201
lwfed	.3361278	.2453134	1.37	0.175	-.1525615	.8248172
lwsta	.0395089	.2072112	0.19	0.849	-.3732769	.4522947
lwloc	-.0369855	.3291546	-0.11	0.911	-.6926951	.6187241
_cons	-3.792525	1.957472	-1.94	0.056	-7.692009	.1069593

```
. testparm lwcon-lwloc
```

- ( 1) lwcon = 0
- ( 2) lwtuc = 0
- ( 3) lwtrd = 0
- ( 4) lwfir = 0
- ( 5) lwser = 0
- ( 6) lwmfg = 0
- ( 7) lwfed = 0
- ( 8) lwsta = 0
- ( 9) lwloc = 0

```
F( 9, 75) = 1.50
Prob > F = 0.1643
```

The nine wage variables are jointly insignificant even at the 15% level. Plus, the elasticities

are not consistently positive or negative. The two largest elasticities – which also have the largest absolute  $t$  statistics – have the opposite sign. These are with respect to the wage in construction (−.285) and the wage for federal employees (.336).

d. The following Stata output gives the heteroskedasticity-robust  $F$  statistic:

```
. qui reg lcrmte lprbarr lprbconv lprbpris lavgsen lcrmte_1 lwcon- lwloc if
. testparm lwcon-lwloc

( 1)  lwcon = 0
( 2)  lwtuc = 0
( 3)  lwtrd = 0
( 4)  lwfir = 0
( 5)  lwser = 0
( 6)  lwmfg = 0
( 7)  lwfed = 0
( 8)  lwsta = 0
( 9)  lwloc = 0

      F(  9,      75) =      2.19
      Prob > F =      0.0319
```

Therefore, we would reject the null at the 5% significance level. But we might hesitate to rely on asymptotic theory – which the heteroskedasticity-robust test requires – with  $N = 90$  and  $K = 15$  parameters to estimate. (This heteroskedasticity-robust  $F$  statistic is the heteroskedasticity-robust Wald statistic divided by the number of restrictions being tested, which is nine in this example. The division by the number of restrictions turns the asymptotic chi-square statistic into one that can be treated as having roughly an  $F$  distribution.)

**4.14. a.** Before doing the regression, it is helpful to know some summary statistics for the variables of primary interest:

```
. sum stndfnl atndrte
```

Variable	Obs	Mean	Std. Dev.	Min	Max
stndfnl	680	.0296589	.9894611	-3.308824	2.783613
atndrte	680	81.70956	17.04699	6.25	100

Because the final exam score has been standardized, it has close to a zero mean and its

standard deviation is close to one. The values are not closer to zero and one, respectively, because the standardization was done with a larger data set that included students with missing values on other key variables. It might make sense to redefine the standardized test score using the mean and standard deviation in the sample of 680, but the effect should be minor.

The regression that controls only for year in school in addition to attendance rate is as follows:

```
. reg stndfnl atndrte frosh soph
```

Source	SS	df	MS	Number of obs = 680		
Model	19.3023776	3	6.43412588	F( 3, 676) = 6.74		
Residual	645.46119	676	.954824246	Prob > F = 0.0002		
Total	664.763568	679	.979033237	R-squared = 0.0290		
				Adj R-squared = 0.0247		
				Root MSE = .97715		

stndfnl	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
atndrte	.0081634	.0022031	3.71	0.000	.0038376	.0124892
frosh	-.2898943	.1157244	-2.51	0.012	-.5171168	-.0626719
soph	-.1184456	.0990267	-1.20	0.232	-.3128824	.0759913
_cons	-.5017308	.196314	-2.56	0.011	-.8871893	-.1162724

If *atndrte* increases by 10 percentage points (say, from 75 to 85), the standardized test score is estimated to increase by about .082 standard deviations.

b. Certainly there is a potential for self-selection. The better students may also be the ones attending lecture more regularly. So the positive effect of the attendance rate simply might capture the fact that better students tend to do better on exams. It is unlikely that controlling just for year in college (*frosh* and *soph*) solves the endogeneity of *atndrete*.

c. Adding *priGPA* and *ACT* gives

```
reg stndfnl atndrte frosh soph priGPA ACT
```

Source	SS	df	MS	Number of obs = 680		
Model	136.801957	5	27.3603913	F( 5, 674) = 34.93		
Residual	527.961611	674	.783325833	Prob > F = 0.0000		
				R-squared = 0.2058		
				Adj R-squared = 0.1999		

Total | 664.763568 679 .979033237 Root MSE = .88506

stndfnl	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
atndrte	.0052248	.0023844	2.19	0.029	.000543	.0099065
frosh	-.0494692	.1078903	-0.46	0.647	-.2613108	.1623724
soph	-.1596475	.0897716	-1.78	0.076	-.3359132	.0166181
priGPA	.4265845	.0819203	5.21	0.000	.2657348	.5874343
ACT	.0844119	.0111677	7.56	0.000	.0624843	.1063395
_cons	-3.297342	.308831	-10.68	0.000	-3.903729	-2.690956

The effect of *atndrte* has fallen, which is what we expect if we think better, smarter students also attend lectures more frequently. The estimate now is that a 10 percentage point increase in *atndrte* increases the standardized test score by .052 standard deviations; the effect is statistically significant at the usual 5% level against a two-sided alternative, but the *t* statistic is much lower than in part a. The strong positive effects of prior GPA and ACT score are also expected.

d. Controlling for *priGPA* and *ACT* causes the sophomore effect (relative to students in year three and beyond) to get slightly larger in magnitude and more statistically significant. These data are for a course taught in the second term, so each *frosh* student does have a prior GPA – his or her GPA for the first semester in college. Adding *priGPA* in particular causes the “freshman effect” to essentially disappear. This is not too surprising because the average prior GPA for first-year students is notably less than the overall average *priGPA*.

e. Here is the Stata session for adding squares in the proxy variables. Because we are not interested in the effects of the proxies, we do not demean them before creating the squared terms:

```
. gen priGPAsq = priGPA^2
```

```
. gen ACTsq = ACT^2
```

```
. reg stndfnl atndrte frosh soph priGPA ACT priGPAsq ACTsq
```

Source	SS	df	MS	Number of obs =	680
				F( 7, 672) =	28.94

Model	153.974309	7	21.9963299	Prob > F	=	0.0000
Residual	510.789259	672	.760103064	R-squared	=	0.2316
						Adj R-squared = 0.2236
Total	664.763568	679	.979033237	Root MSE	=	.87184

stndfnl	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
atndrte	.0062317	.0023583	2.64	0.008	.0016011	.0108623
frosh	-.1053368	.1069747	-0.98	0.325	-.3153817	.1047081
soph	-.1807289	.0886354	-2.04	0.042	-.3547647	-.0066932
priGPA	-1.52614	.4739715	-3.22	0.001	-2.456783	-.5954966
ACT	-.1124331	.098172	-1.15	0.253	-.3051938	.0803276
priGPAsq	.3682176	.0889847	4.14	0.000	.1934961	.5429391
ACTsq	.0041821	.0021689	1.93	0.054	-.0000766	.0084408
_cons	1.384812	1.239361	1.12	0.264	-1.048674	3.818298

Adding the squared terms – one of which is very significant, the other of which is marginally significant – actually increases the attendance rate effect. And it does so while slightly reducing the standard error on *atndrte*, resulting in a *t* statistic that is notably more significant than in part c.

f. Adding the squared attendance rate is not warranted, as it is very insignificant:

```
. gen atndrtesq = atndrte^2
. reg stndfnl atndrte frosh soph priGPA ACT priGPAsq ACTsq atndrtesq
```

Source	SS	df	MS	Number of obs = 680		
Model	153.975323	8	19.2469154	F( 8, 671) = 25.28		
Residual	510.788245	671	.761234344	Prob > F = 0.0000		
				R-squared = 0.2316		
				Adj R-squared = 0.2225		
Total	664.763568	679	.979033237	Root MSE = .87249		

stndfnl	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
atndrte	.0058425	.0109203	0.54	0.593	-.0155996	.0272847
frosh	-.1053656	.1070572	-0.98	0.325	-.3155729	.1048418
soph	-.1808403	.0887539	-2.04	0.042	-.355109	-.0065716
priGPA	-1.524803	.475737	-3.21	0.001	-2.458915	-.5906902
ACT	-.1123423	.0982764	-1.14	0.253	-.3053087	.080624
priGPAsq	.3679124	.0894427	4.11	0.000	.192291	.5435337
ACTsq	.0041802	.0021712	1.93	0.055	-.0000829	.0084433
atndrtesq	2.87e-06	.0000787	0.04	0.971	-.0001517	.0001574
_cons	1.394292	1.267186	1.10	0.272	-1.093835	3.88242

The very large increase in the standard error on *atndrte* suggest that *atndrte* and *atndrte*<sup>2</sup>



are highly collinear. In fact, their sample correlation is about .983. Importantly, the coefficient on *atndrte* now has an uninteresting interpretation: it measures the partial effect of *atndrte* starting from *atndrte* = 0. The lowest attendance rate in the sample is 6.25, with the vast majority of students (94.3%) attending 50 percent or more of the lectures. If the quadratic term were significant, we might want to center *atndrte* about its mean or median before creating the square. Or, a more sophisticated functional form might be called for. It may be better to define several intervals for *atndrrte* and include dummy variables for those intervals.

**4.15.** a. Because each  $x_j$  has finite second moment,  $\text{Var}(\mathbf{x}\boldsymbol{\beta}) < \infty$ . Since  $\text{Var}(u) < \infty$ ,  $\text{Cov}(\mathbf{x}\boldsymbol{\beta}, u)$  is well-defined. But each  $x_j$  is uncorrelated with  $u$ , so  $\text{Cov}(\mathbf{x}\boldsymbol{\beta}, u) = 0$ . Therefore,  $\text{Var}(y) = \text{Var}(\mathbf{x}\boldsymbol{\beta}) + \text{Var}(u)$  or  $\sigma_y^2 = \text{Var}(\mathbf{x}\boldsymbol{\beta}) + \sigma_u^2$ .

b. This is nonsense when we view  $\mathbf{x}_i$  as a random draw along with  $y_i$ . The statement “ $\text{Var}(u_i) = \sigma^2 = \text{Var}(y_i)$  for all  $i$ ” assumes that the regressors are nonrandom (or  $\boldsymbol{\beta} = \mathbf{0}$ , which is not a very interesting case). This is another example of how the assumption of nonrandom regressors can lead to counterintuitive conclusions. Suppose that an element of the error term, say  $z$ , which is uncorrelated with each  $x_j$ , suddenly becomes observed. When we add  $z$  to the regressor list, the error changes, and so does the error variance. In the vast majority of economic applications, it makes no sense to think we have access to the entire set of factors that one would ever want to control for, and so we should allow for error variances to change across different sets of explanatory variables that we might use for the same response variable. We avoid trouble by focusing on joint distributions in the population.

c. Write  $R^2 = 1 - \text{SSR}/\text{SST} = 1 - (\text{SSR}/N)/(\text{SST}/N)$ . Therefore,  $\text{plim}(R^2) = 1 - \text{plim}[(\text{SSR}/N)/(\text{SST}/N)] = 1 - [\text{plim}(\text{SSR}/N)]/[\text{plim}(\text{SST}/N)] = 1 - \sigma_u^2/\sigma_y^2 =$  where we use the fact that  $\text{SSR}/N$  is a consistent estimator of  $\sigma_u^2$  and  $\text{SST}/N$  is a consistent

estimator of  $\sigma_y^2$ .

d. The derivation in part c assumed nothing about  $\text{Var}(u|\mathbf{x})$ . The population  $R$ -squared depends on only the *unconditional* variances of  $u$  and  $y$ . Therefore, regardless of the nature of heteroskedasticity in  $\text{Var}(u|\mathbf{x})$ , the usual  $R$ -squared consistently estimates the population  $R$ -squared. Neither  $R$ -squared nor the adjusted  $R$ -squared has desirable finite-sample properties, such as unbiasedness, so the only analysis we can give in any generality involves asymptotics. The statement in the problem is simply wrong.

**4.16.** a. The proof is fairly similar to that for random sampling. First, note that the assumptions  $N^{-1} \sum_{i=1}^N [\mathbf{x}_i' \mathbf{x}_i - E(\mathbf{x}_i' \mathbf{x}_i)] \xrightarrow{p} \mathbf{0}$  – which is how the WLLN is stated for i.n.i.d. sequences – and  $N^{-1} \sum_{i=1}^N E(\mathbf{x}_i' \mathbf{x}_i) \rightarrow \mathbf{A}$  – which is not crucial but is pretty harmless and simplifies the proof – imply

$$N^{-1} \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \xrightarrow{p} \mathbf{A}$$

In addition,  $E(\mathbf{x}_i' u_i) = \mathbf{0}$  and the assumption that  $N^{-1} \sum_{i=1}^N \mathbf{x}_i' u_i$  satisfies the law of large numbers imply

$$N^{-1} \sum_{i=1}^N \mathbf{x}_i' u_i \xrightarrow{p} \mathbf{0}.$$

We are also given that  $\mathbf{A}$  is positive definite, which means  $\mathbf{X}'\mathbf{X}/N$  is invertible with probability approaching one and  $(\mathbf{X}'\mathbf{X}/N)^{-1} \xrightarrow{p} \mathbf{A}^{-1}$ . Therefore,

$$\begin{aligned}
\text{plim}_{N \rightarrow \infty}(\hat{\boldsymbol{\beta}}) &= \boldsymbol{\beta} + \text{plim} \left[ \left( N^{-1} \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \right)^{-1} N^{-1} \sum_{i=1}^N \mathbf{x}_i' u_i \right] \\
&= \boldsymbol{\beta} + \text{plim} \left[ \left( N^{-1} \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \right] \text{plim} \left( N^{-1} \sum_{i=1}^N \mathbf{x}_i' u_i \right) \\
&= \boldsymbol{\beta} + \mathbf{A}^{-1} \cdot \mathbf{0} = \boldsymbol{\beta}.
\end{aligned}$$

b. Because  $N^{-1/2} \sum_{i=1}^N \mathbf{x}_i' u_i \xrightarrow{d} \text{Normal}(\mathbf{0}, \mathbf{B})$ , the sequence  $N^{-1/2} \sum_{i=1}^N \mathbf{x}_i' u_i$  is  $O_p(1)$ . We already used in part a that

$$\left( N^{-1} \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \right)^{-1} - \mathbf{A}^{-1} = o_p(1)$$

Now, as in the i.i.d. case, write

$$\begin{aligned}
\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left( N^{-1} \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left( N^{-1/2} \sum_{i=1}^N \mathbf{x}_i' u_i \right) \\
&= \left[ \left( N^{-1} \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \right)^{-1} - \mathbf{A}^{-1} \right] \left( N^{-1/2} \sum_{i=1}^N \mathbf{x}_i' u_i \right) + \mathbf{A}^{-1} \left( N^{-1/2} \sum_{i=1}^N \mathbf{x}_i' u_i \right) \\
&= o_p(1) \cdot O_p(1) + \mathbf{A}^{-1} \left( N^{-1/2} \sum_{i=1}^N \mathbf{x}_i' u_i \right) \\
&\xrightarrow{d} \text{Normal}(\mathbf{0}, \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1})
\end{aligned}$$

where we use the assumption  $N^{-1/2} \sum_{i=1}^N \mathbf{x}_i' u_i \xrightarrow{d} \text{Normal}(\mathbf{0}, \mathbf{B})$ . The asymptotic variance of  $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$  has the usual sandwich form,  $\mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}$ .

c. We already know that

$$N^{-1} \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \xrightarrow{p} \mathbf{A}.$$

Further, by the WLLN and the assumption that  $\mathbf{B}_N \rightarrow \mathbf{B}$ ,

$$N^{-1} \sum_{i=1}^N u_i^2 \mathbf{x}_i' \mathbf{x}_i \xrightarrow{p} \mathbf{B}$$

The hard part – just as with the i.i.d. case – is to show that replacing the  $u_i$  with the OLS residuals,  $\hat{u}_i$ , does not affect consistency. Nevertheless, under general assumptions it follows that

$$N^{-1} \sum_{i=1}^N \hat{u}_i^2 \mathbf{x}_i' \mathbf{x}_i \xrightarrow{p} \mathbf{B}$$

Naturally, we can use the same degrees-of-freedom adjustment as in the i.i.d. case: replace  $N^{-1}$  with  $(N - K)^{-1}$ .

d. The point of this exercise is that we are led to exactly the same heteroskedasticity-robust estimator whether we assume i.i.d. observations or i.n.i.d. observations. In particular, even if unconditional variances are constant – as they must be in the i.i.d. case – we still might need heteroskedasticity-robust standard errors. In the i.n.i.d. case, the robust variance matrix estimator allows for changing unconditional variances as well as conditional variances that depend on  $\mathbf{x}_i$ .

**4.17.** We know that, in general,

$$\text{Avar} \sqrt{N} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = [\text{E}(\mathbf{x}'\mathbf{x})]^{-1} \text{E}(u^2 \mathbf{x}'\mathbf{x}) [\text{E}(\mathbf{x}'\mathbf{x})]^{-1}.$$

Now we just apply iterated expectations to the matrix in the middle:

$$\text{E}(u^2 \mathbf{x}'\mathbf{x}) = \text{E}[\text{E}(u^2 \mathbf{x}'\mathbf{x} | \mathbf{x})] = \text{E}[\text{E}(u^2 | \mathbf{x}) \mathbf{x}'\mathbf{x}] = \text{E}[h(\mathbf{x}) \mathbf{x}'\mathbf{x}]$$

**4.18.** a. This is a fairly common misconception – or at least misstatement. Recall that the distribution of any random draw,  $u_i$ , is the population distribution of  $u$ . But, of course, the population distribution of  $u$  is what it is; it does not change with the sample size. In fact, it has

nothing to do with the sample size. Therefore, the random draws on  $u_i$  have the same distribution regardless of  $N$ . A correct statement is that the standardized average of the errors,  $N^{-1/2} \sum_{i=1}^N u_i = \sqrt{N} \bar{u}$  approaches normality as  $N \rightarrow \infty$ . This is a much different statement. (In regression analysis, we use the fact that  $N^{-1/2} \sum_{i=1}^N \mathbf{x}_i' u_i$  generally converges to a multivariate normal distribution, which implies the convergence of  $N^{-1/2} \sum_{i=1}^N u_i$  to normality when  $\mathbf{x}_i$  contains unity.)

b. It is tempting but incorrect to think that a single squared OLS residual can consistently estimate a conditional mean,  $E(u_i^2 | \mathbf{x}_i) \equiv h(\mathbf{x}_i)$ , but there is no sense in which this statement is true. It is not even clear what we would mean by it, but we can make some headway by writing  $\hat{u}_i^2 = u_i^2 - 2u_i \mathbf{x}_i(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + [\mathbf{x}_i(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]^2$ . Now, we can conclude  $\hat{u}_i^2 - u_i^2 \xrightarrow{p} 0$  and  $N \rightarrow \infty$  because  $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ . But remember  $u_i^2 = h(\mathbf{x}_i) + v_i$  where  $E(v_i | \mathbf{x}_i) = 0$ . There is no sense in which  $u_i^2$  is a consistent estimator of  $h(\mathbf{x}_i)$ ; they do not even depend on the sample size  $N$ .

It was the view that we needed  $\hat{u}_i^2$  to be a good estimate of  $E(u_i^2 | \mathbf{x}_i)$  that possibly held up progress on heteroskedasticity-consistent covariance matrices. Fortunately, all we need to consistent estimate is the population mean

$$\mathbf{B} = E(u^2 \mathbf{x}' \mathbf{x}),$$

for which the obvious consistent (and unbiased) estimator is

$$N^{-1} \sum_{i=1}^N u_i^2 \mathbf{x}_i' \mathbf{x}_i.$$

The rest is demonstrating the replacing the implicit  $\boldsymbol{\beta}$  with  $\hat{\boldsymbol{\beta}}$  preserves consistency (not unbiasedness). As we know, this requires some tricky algebra with  $o_p(1)$  and  $O_p(1)$ , but the work is not too onerous.

## Solutions to Chapter 5 Problems

5.1. Define  $\mathbf{x}_1 \equiv (\mathbf{z}_1, y_2)$  and  $x_2 \equiv \hat{v}_2$ , and let  $\hat{\boldsymbol{\beta}} \equiv (\hat{\boldsymbol{\beta}}_1', \hat{\rho}_1')'$  be OLS estimator from (5.52), where  $\hat{\boldsymbol{\beta}}_1 = (\hat{\boldsymbol{\delta}}_1', \hat{\alpha}_1')$ . Using the hint,  $\hat{\boldsymbol{\beta}}_1$  can also be obtained by partitioned regression:

- (i) Regress  $\mathbf{x}_1$  onto  $\hat{v}_2$  and save the residuals, say  $\tilde{\mathbf{x}}_1$ .
- (ii) Regress  $y_1$  onto  $\tilde{\mathbf{x}}_1$ .

But when we regress  $\mathbf{z}_1$  onto  $\hat{v}_2$  the residuals are just  $\mathbf{z}_1$  because  $\hat{v}_2$  is orthogonal in sample to  $\mathbf{z}$ . (More precisely,  $\sum_{i=1}^N \mathbf{z}_{i1}' \hat{v}_{i2} = \mathbf{0}$ .) Further, because we can write  $y_2 = \hat{y}_2 + \hat{v}_2$ , where  $\hat{y}_2$  and  $\hat{v}_2$  are orthogonal in sample, the residuals from regressing  $y_2$  onto  $\hat{v}_2$  are simply the first stage fitted values,  $\hat{y}_2$ . In other words,  $\tilde{\mathbf{x}}_1 = (\mathbf{z}_1, \hat{y}_2)$ . But the 2SLS estimator of  $\boldsymbol{\beta}_1$  is obtained exactly from the OLS regression  $y_1$  on  $\mathbf{z}_1, \hat{y}_2$ .

5.2. a. Unobserved factors that tend to make an individual healthier also tend to make that person exercise more. For example, if *health* is a cardiovascular measure, people with a history of heart problems are probably less likely to exercise. Unobserved factors such as prior health or family history are contained in  $u_1$ , and so we are worried about correlation between *exercise* and  $u_1$ . Self-selection into exercising predicts that the benefits of exercising will be, on average, overestimated. Ideally, the amount of exercise could be randomized across a sample of people, but this can be difficult.

b. If people do *not* systematically choose the location of their homes and jobs relative to health clubs based on unobserved health characteristics, then it is reasonable to believe that *disthome* and *distwork* are uncorrelated with  $u_1$ . But the location of health clubs is not necessarily exogenous. Clubs may tend to be built near neighborhoods where residents have higher income and wealth, on average, and these factors can certainly affect overall health. It

may make sense to choose residents from neighborhoods with very similar characteristics but where one neighborhood is located near a health club.

c. The reduced form for *exercise* is

$$\begin{aligned} \text{exercise} = & \pi_0 + \pi_1 \text{age} + \pi_2 \text{weight} + \pi_3 \text{height} \\ & + \pi_4 \text{male} + \pi_5 \text{work} + \pi_6 \text{disthome} + \pi_7 \text{distwork} + u_1, \end{aligned}$$

For identification we need at least one of  $\pi_6$  and  $\pi_7$  to be different from zero. this assumption can fail if the amount that people exercise is not systematically related to distances to the nearest health club.

d. An  $F$  test of  $H_0 : \pi_6 = 0, \pi_7 = 0$  is the simplest way to test the identification assumption in part c. As usual, it would be a good idea to compute a heteroskedasticity-robust version.

**5.3.** a. There may be unobserved health factors correlated with smoking behavior that affect infant birth weight. For example, women who smoke during pregnancy may, on average, drink more coffee or alcohol, or eat less nutritious meals.

b. Basic economics says that *packs* should be negatively correlated with cigarette price, although the correlation might be small (especially because price is aggregated at the state level). At first glance it seems that cigarette price should be exogenous in equation (5.54), but we must be a little careful. One component of cigarette price is the state tax on cigarettes. States that have lower taxes on cigarettes may also have lower quality of health care, on average. Quality of health care is in  $u$ , and so maybe cigarette price fails the exogeneity requirement for an IV.

c. OLS is followed by 2SLS (IV, in this case):

```
. reg lbwght male parity lfaminc packs
```

Source	SS	df	MS
Model	1.76664363	4	.441660908

Number of obs =	1388
F( 4, 1383) =	12.55
Prob > F =	0.0000

Residual		48.65369	1383	.035179819		R-squared	=	0.0350
-----+								
Total		50.4203336	1387	.036352079		Adj R-squared	=	0.0322
						Root MSE	=	.18756

lbwght	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
male	.0262407	.0100894	2.60	0.009	.0064486	.0460328
parity	.0147292	.0056646	2.60	0.009	.0036171	.0258414
lfaminc	.0180498	.0055837	3.23	0.001	.0070964	.0290032
packs	-.0837281	.0171209	-4.89	0.000	-.1173139	-.0501423
_cons	4.675618	.0218813	213.68	0.000	4.632694	4.718542

```
. ivreg lbwght male parity lfaminc (packs = cigprice)
```

Instrumental variables (2SLS) regression

Source		SS	df	MS		Number of obs	=	1388
-----+								
Model		-91.350027	4	-22.8375067		F( 4, 1383)	=	2.39
Residual		141.770361	1383	.102509299		Prob > F	=	0.0490
-----+								
Total		50.4203336	1387	.036352079		R-squared	=	
						Adj R-squared	=	
						Root MSE	=	.32017

lbwght	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
packs	.7971063	1.086275	0.73	0.463	-1.333819	2.928031
male	.0298205	.017779	1.68	0.094	-.0050562	.0646972
parity	-.0012391	.0219322	-0.06	0.955	-.044263	.0417848
lfaminc	.063646	.0570128	1.12	0.264	-.0481949	.1754869
_cons	4.467861	.2588289	17.26	0.000	3.960122	4.975601

Instrumented: packs

Instruments: male parity lfaminc cigprice

The difference between OLS and IV in the estimated effect of *packs* on *bwght* is huge. With the OLS estimate, one more pack of cigarettes is estimated to reduce *bwght* by about 8.4%, and is statistically significant. The IV estimate has the opposite sign, is huge in magnitude, and is not statistically significant. The sign and size of the smoking effect are not realistic.

d. We can see the problem with IV by estimating the reduced form for *packs*.

```
. reg packs male parity lfaminc cigprice
```

Source		SS	df	MS		Number of obs	=	1388
-----+								
						F( 4, 1383)	=	10.86



Model	3.76705108	4	.94176277	Prob > F	=	0.0000
Residual	119.929078	1383	.086716615	R-squared	=	0.0305
-----				Adj R-squared	=	0.0276
Total	123.696129	1387	.089182501	Root MSE	=	.29448
-----						
packs	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
-----						
male	-.0047261	.0158539	-0.30	0.766	-.0358264	.0263742
parity	.0181491	.0088802	2.04	0.041	.0007291	.0355692
lfaminc	-.0526374	.0086991	-6.05	0.000	-.0697023	-.0355724
cigprice	.000777	.0007763	1.00	0.317	-.0007459	.0022999
_cons	.1374075	.1040005	1.32	0.187	-.0666084	.3414234
-----						

The reduced form estimates show that *cigprice* does not significantly affect *packs*. In fact, the coefficient on *cigprice* does not have the sign we expect. Thus, *cigprice* fails as an IV for *packs* because *cigprice* is not partially correlated with *packs* with a sensible sign for the correlation. This is separate from the problem that *cigprice* may not truly be exogenous in the birth weight equation.

#### 5.4. a. Here are the OLS results:

```
. reg lwage educ exper expersq black south smsa reg661-reg668 smsa66
```

Source	SS	df	MS	Number of obs = 3010		
-----				F( 15, 2994) = 85.48		
Model	177.695591	15	11.8463727	Prob > F = 0.0000		
Residual	414.946054	2994	.138592536	R-squared = 0.2998		
-----				Adj R-squared = 0.2963		
Total	592.641645	3009	.196956346	Root MSE = .37228		
-----						
lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
-----						
educ	.0746933	.0034983	21.35	0.000	.0678339	.0815527
exper	.084832	.0066242	12.81	0.000	.0718435	.0978205
expersq	-.002287	.0003166	-7.22	0.000	-.0029079	-.0016662
black	-.1990123	.0182483	-10.91	0.000	-.2347927	-.1632318
south	-.147955	.0259799	-5.69	0.000	-.1988952	-.0970148
smsa	.1363845	.0201005	6.79	0.000	.0969724	.1757967
reg661	-.1185698	.0388301	-3.05	0.002	-.194706	-.0424335
reg662	-.0222026	.0282575	-0.79	0.432	-.0776088	.0332036
reg663	.0259703	.0273644	0.95	0.343	-.0276846	.0796251
reg664	-.0634942	.0356803	-1.78	0.075	-.1334546	.0064662
reg665	.0094551	.0361174	0.26	0.794	-.0613623	.0802725
reg666	.0219476	.0400984	0.55	0.584	-.0566755	.1005708
reg667	-.0005887	.0393793	-0.01	0.988	-.077802	.0766245
reg668	-.1750058	.0463394	-3.78	0.000	-.265866	-.0841456
smsa66	.0262417	.0194477	1.35	0.177	-.0118905	.0643739
cons	4.739377	.0715282	66.26	0.000	4.599127	4.879626

-----

The estimated return to education is about 7.5%, with a very large  $t$  statistic. These reproduce the estimates from Table 2, Column (2) in Card (1995).

b. The reduced form for *educ* is

```
. reg educ exper expersq black south smsa reg661-reg668 smsa66 nearc4
```

Source	SS	df	MS	Number of obs =	3010
Model	10287.6179	15	685.841194	F( 15, 2994) =	182.13
Residual	11274.4622	2994	3.76568542	Prob > F =	0.0000
				R-squared =	0.4771
				Adj R-squared =	0.4745
Total	21562.0801	3009	7.16586243	Root MSE =	1.9405

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
exper	-.4125334	.0336996	-12.24	0.000	-.4786101	-.3464566
expersq	.0008686	.0016504	0.53	0.599	-.0023674	.0041046
black	-.9355287	.0937348	-9.98	0.000	-1.11932	-.7517377
south	-.0516126	.1354284	-0.38	0.703	-.3171548	.2139296
smsa	.4021825	.1048112	3.84	0.000	.1966732	.6076918
reg661	-.210271	.2024568	-1.04	0.299	-.6072395	.1866975
reg662	-.2889073	.1473395	-1.96	0.050	-.5778042	-.0000105
reg663	-.2382099	.1426357	-1.67	0.095	-.5178838	.0414639
reg664	-.093089	.1859827	-0.50	0.617	-.4577559	.2715779
reg665	-.4828875	.1881872	-2.57	0.010	-.8518767	-.1138982
reg666	-.5130857	.2096352	-2.45	0.014	-.9241293	-.1020421
reg667	-.4270887	.2056208	-2.08	0.038	-.8302611	-.0239163
reg668	.3136204	.2416739	1.30	0.194	-.1602434	.7874841
smsa66	.0254805	.1057692	0.24	0.810	-.1819071	.2328682
nearc4	.3198989	.0878638	3.64	0.000	.1476194	.4921785
_cons	16.84852	.2111222	79.80	0.000	16.43456	17.26248

The important coefficient is on *nearc4*. Statistically, *educ* and *nearc4* are partially correlated, and in a way that makes sense: holding other factors in the reduced form fixed, someone living near a four-year college at age 16 has, on average, almost one-third a year more education than a person not near a four-year college at age 16. This is not trivial a effect, so *nearc4* passes the requirement that it is partially correlated with *educ*.

c. Here are the IV estimates:

```
. ivreg lwage exper expersq black south smsa reg661-reg668 smsa66 (educ = nearc4
```

# Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs = 3010		
Model	141.146813	15	9.40978752	F( 15, 2994) = 51.01		
Residual	451.494832	2994	.150799877	Prob > F = 0.0000		
Total	592.641645	3009	.196956346	R-squared = 0.2382		
				Adj R-squared = 0.2343		
				Root MSE = .38833		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
educ	.1315038	.0549637	2.39	0.017	.0237335	.2392742
exper	.1082711	.0236586	4.58	0.000	.0618824	.1546598
expersq	-.0023349	.0003335	-7.00	0.000	-.0029888	-.001681
black	-.1467757	.0538999	-2.72	0.007	-.2524603	-.0410912
south	-.1446715	.0272846	-5.30	0.000	-.19817	-.091173
smsa	.1118083	.031662	3.53	0.000	.0497269	.1738898
reg661	-.1078142	.0418137	-2.58	0.010	-.1898007	-.0258278
reg662	-.0070465	.0329073	-0.21	0.830	-.0715696	.0574767
reg663	.0404445	.0317806	1.27	0.203	-.0218694	.1027585
reg664	-.0579172	.0376059	-1.54	0.124	-.1316532	.0158189
reg665	.0384577	.0469387	0.82	0.413	-.0535777	.130493
reg666	.0550887	.0526597	1.05	0.296	-.0481642	.1583416
reg667	.026758	.0488287	0.55	0.584	-.0689832	.1224992
reg668	-.1908912	.0507113	-3.76	0.000	-.2903238	-.0914586
smsa66	.0185311	.0216086	0.86	0.391	-.0238381	.0609003
_cons	3.773965	.934947	4.04	0.000	1.940762	5.607169

Instrumented:	educ
Instruments:	exper expersq black south smsa reg661 reg662 reg663 reg664 reg665 reg666 reg667 reg668 smsa66 nearc4

The estimated return to education has increased to about 13.2%, but notice how wide the 95% confidence interval is: 2.4% to 23.9%. By contrast, the OLS confidence interval is about 6.8% to 8.2%, which is much tighter. Of course, OLS could be inconsistent, in which case a tighter CI is of little value. But the estimated return to education is higher with IV, something that seems a bit counterintuitive.

One possible explanation is that *educ* suffers from classical errors-in-variables. Therefore, while OLS would tend to overestimate the return to schooling because of omitted “ability,” classical measurement error in *educ* leads to an attenuation bias. Measurement error may help explain why the IV estimate is larger, but it is not entirely convincing. It seems unlikely that *educ* satisfies the CEV assumptions. For example, if we think the measurement error is due to

truncation – people are asked about highest grade completed, not actual years of schooling – then *educ* is always less than or equal to *educ\**. And the measurement error could not be independent of *educ\**. If we think the mismeasurement is due to is unobserved quality of schooling, it seems likely that quality of schooling – part of the measurement error – is positively correlated with actual amount of schooling. This, too, violates the CEV assumptions. Another possibility for the much higher IV estimate comes out of the recent treatment effect literature, which is covered in Section 21.4. Of course, we must also remember that the point estimates – particularly the IV estimate – are subject to substantial sampling variation. At this point, we do not even know if OLS and IV are statistically different from each other. See Problem 6.1.

d. When *nearc2* is added to the reduced form of *educ* it has a coefficient (standard error) of .123 (.077), compared with .321 (.089) for *nearc4*. Therefore, *nearc4* has a much stronger ceteris paribus relationship with *educ*; *nearc2* is only marginally statistically significant once *nearc4* has been included. The joint *F* test gives  $F = 7.89$  with  $p\text{-value} = .004$ .

The 2SLS estimate of the return to education becomes about 15.7%, with 95% CI given by 5.4% to 26%. The CI is still very wide.

**5.5.** Under the null hypothesis that  $q$  and  $\mathbf{z}_2$  are uncorrelated,  $\mathbf{z}_1$  and  $\mathbf{z}_2$  are exogenous in (5.55) because each is uncorrelated with  $u_1$ . Unfortunately,  $y_2$  is correlated with  $u_1$ , and so the regression of  $y_1$  on  $\mathbf{z}_1, y_2, \mathbf{z}_2$  does not produce a consistent estimator of  $\mathbf{0}$  on  $\mathbf{z}_2$  even when  $E(\mathbf{z}_2'q) = \mathbf{0}$ . We could find that  $\hat{\psi}_1$  from this regression is statistically different from zero even when  $q$  and  $\mathbf{z}_2$  are uncorrelated – in which case we would incorrectly conclude that  $\mathbf{z}_2$  is not a valid IV candidate. Or, we might fail to reject  $H_0 : \psi_1 = \mathbf{0}$  when  $\mathbf{z}_2$  and  $q$  are correlated – in

which case we incorrectly conclude that the elements in  $\mathbf{z}_2$  are valid as instruments.

The point of this exercise is that one cannot simply add instrumental variable candidates in the structural equation and then test for significance of these variables using OLS estimation. This is the sense in which identification cannot be tested: we cannot test whether all of the IV candidates are uncorrelated with  $q$ . With a single endogenous variable, we must take a stand that at least one element of  $\mathbf{z}_2$  is uncorrelated with  $q$ .

5.6. a. By definition, the reduced form is the linear projection

$$L(q_1|1, \mathbf{x}, q_2) = \pi_0 + \mathbf{x}\pi_1 + \pi_2 q_2,$$

and we want to show that  $\pi_1 = \mathbf{0}$  when  $q_2$  is uncorrelated with  $\mathbf{x}$ . Now, because  $q_2$  is a linear function of  $q$  and  $a_2$ , and  $a_2$  is uncorrelated with  $\mathbf{x}$ ,  $q_2$  is uncorrelated with  $\mathbf{x}$  if and only if  $q$  is uncorrelated with  $\mathbf{x}$ . Assuming then that  $q$  and  $\mathbf{x}$  are uncorrelated,  $q_1$  is also uncorrelated with  $\mathbf{x}$ . A basic fact about linear projections is that, because  $q_1$  and  $q_2$  are each uncorrelated with the vector  $\mathbf{x}$ ,  $\pi_1 = \mathbf{0}$ . This claim follows from Property LP.7:  $\pi_1$  can be obtained by first projecting  $\mathbf{x}$  on 1,  $q_2$  and obtaining the population residuals, say  $\mathbf{r}$ . Then, project  $q_1$  onto  $\mathbf{r}$ . But because  $\mathbf{x}$  and  $q_2$  are orthogonal,  $\mathbf{r} = \mathbf{x} - \mu_{\mathbf{x}}$ . Projecting  $q_1$  on  $(\mathbf{x} - \mu_{\mathbf{x}})$  just gives the zero vector because  $E[(\mathbf{x} - \mu_{\mathbf{x}})'q_1] = \mathbf{0}$ . Therefore,  $\pi_1 = \mathbf{0}$ .

b. If  $q_2$  and  $\mathbf{x}$  are correlated then  $\pi_1 \neq \mathbf{0}$ , and  $\mathbf{x}$  appears in the reduced form for  $q_1$ . It is not realistic to assume that  $q_2$  and  $\mathbf{x}$  are uncorrelated. Under the multiple indicator assumptions, assuming  $\mathbf{x}$  and  $q_2$  are uncorrelated is the same as assuming  $q$  and  $\mathbf{x}$  are uncorrelated. If we believe  $q$  and  $\mathbf{x}$  are uncorrelated then there is no need to collect indicators on  $q$  to consistently estimate  $\beta$ : we could simply put  $q$  into the error term and estimate  $\beta$  from an OLS regression of  $y$  on 1,  $\mathbf{x}$ . (Of course, if  $q$  and  $\mathbf{x}$  are uncorrelated we could, in general, gain efficiency for

estimating  $\beta$  by including  $q$  as an extra regressor.)

5.7. a. If we plug  $q = (1/\delta_1)q_1 - (1/\delta_1)a_1$  into equation (5.45) we get

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_K x_K + \eta_1 q_1 + v - \eta_1 a_1, \quad (5.56)$$

where  $\eta_1 \equiv (1/\delta_1)$ . Now, because the  $z_h$  are redundant in (5.45), they are uncorrelated with the structural error,  $v$  (by definition of redundancy). Further, we have assumed that the  $z_h$  are uncorrelated with  $a_1$ . Since each  $x_j$  is also uncorrelated with  $v - \eta_1 a_1$  we can estimate (5.56) by 2SLS using instruments  $(1, x_1, \dots, x_K, z_1, z_2, \dots, z_M)$  to get consistent of the  $\beta_j$  and  $\eta_1$ .

Given all of the zero correlation assumptions, what we need for identification is that at least one of the  $z_h$  appears in the reduced form for  $q_1$ . More formally, in the linear projection

$$q_1 = \pi_0 + \pi_1 x_1 + \dots + \pi_K x_K + \pi_{K+1} z_1 + \dots + \pi_{K+M} z_M + r_1,$$

at least one of  $\pi_{K+1}, \dots, \pi_{K+M}$  must be different from zero.

b. We need family background variables to be redundant in the  $\log(\text{wage})$  equation once ability (and other factors, such as *educ* and *exper*), have been controlled for. The idea here is that family background may influence ability but should have no partial effect on  $\log(\text{wage})$  once ability has been accounted for. For the rank condition to hold, we need family background variables to be correlated with the indicator,  $q_1$  say *IQ*, once the  $x_j$  have been netted out. This is likely to be true if we think that family background and ability are (partially) correlated.

c. Applying the procedure to the data set in NLS80.RAW gives the following results:

```
. ivreg lwage exper tenure educ married south urban black (iq = meduc feduc sibs
Instrumental variables (2SLS) regression
```

Source	SS	df	MS	Number of obs =	722
-----+-----				F( 8, 713) =	25.81
Model	19.6029198	8	2.45036497	Prob > F =	0.0000

Residual		107.208996	713	.150363248		R-squared	=	0.1546
-----								
Total		126.811916	721	.175883378		Adj R-squared	=	0.1451
						Root MSE	=	.38777

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
iq	.0154368	.0077077	2.00	0.046	.0003044	.0305692
exper	.0162185	.0040076	4.05	0.000	.0083503	.0240867
tenure	.0076754	.0030956	2.48	0.013	.0015979	.0137529
educ	.0161809	.0261982	0.62	0.537	-.035254	.0676158
married	.1901012	.0467592	4.07	0.000	.0982991	.2819033
south	-.047992	.0367425	-1.31	0.192	-.1201284	.0241444
urban	.1869376	.0327986	5.70	0.000	.1225442	.2513311
black	.0400269	.1138678	0.35	0.725	-.1835294	.2635832
_cons	4.471616	.468913	9.54	0.000	3.551	5.392231

Instrumented: iq  
Instruments: exper tenure educ married south urban black meduc feduc sibs

. ivreg lwage exper tenure educ married south urban black (kww = meduc feduc

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs =	722
-----					
Model	19.820304	8	2.477538	F( 8, 713) =	25.70
Residual	106.991612	713	.150058361	Prob > F	= 0.0000
-----					
Total	126.811916	721	.175883378	R-squared	= 0.1563
				Adj R-squared	= 0.1468
				Root MSE	= .38737

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
kww	.0249441	.0150576	1.66	0.098	-.0046184	.0545067
exper	.0068682	.0067471	1.02	0.309	-.0063783	.0201147
tenure	.0051145	.0037739	1.36	0.176	-.0022947	.0125238
educ	.0260808	.0255051	1.02	0.307	-.0239933	.0761549
married	.1605273	.0529759	3.03	0.003	.0565198	.2645347
south	-.091887	.0322147	-2.85	0.004	-.1551341	-.0286399
urban	.1484003	.0411598	3.61	0.000	.0675914	.2292093
black	-.0424452	.0893695	-0.47	0.635	-.2179041	.1330137
_cons	5.217818	.1627592	32.06	0.000	4.898273	5.537362

Instrumented: kww  
Instruments: exper tenure educ married south urban black meduc feduc sibs

Even though there are 935 men in the sample, only 722 are used for the estimation because data are missing on *meduc* and *feduc*.

The return to education is estimated to be small and insignificant whether *IQ* or *KWW* used is used as the indicator. This could be because family background variables do not satisfy the

appropriate redundancy condition, or they might be correlated with  $a_1$ . (In both first-stage regressions, the  $F$  statistic for joint significance of *meduc*, *feduc* and *sibs* have p-values below .002, so it seems the family background variables have some partial correlation with the ability indicators.)

**5.8. a.** Plug in the indicator  $q_1$  for  $q$  and the measurement  $x_K$  for  $x_K^*$ , being sure to keep track of the errors:

$$\begin{aligned} y &= \gamma_0 + \beta_1 x_1 + \dots + \beta_K x_K + \gamma_1 q_1 + v - \beta_K e_K + \gamma_1 a_1, \\ &\equiv \gamma_0 + \beta_1 x_1 + \dots + \beta_K x_K + \gamma_1 q_1 + u \end{aligned}$$

where  $\gamma_1 = 1/\delta_1$  Now, if the variables  $z_1, \dots, z_M$  are redundant in the structural equation (so they are uncorrelated with  $v$ ), and uncorrelated with the measurement error  $e_K$  and the indicator error  $a_1$  we can use these as IVs for  $x_K$  and  $q_1$  in 2SLS. We need  $M \geq 2$  because we have two explanatory variables,  $x_q$  and  $q_1$ , that are possibly correlated with the composite error  $u$ .

**b.** The Stata results are:

```
. ivreg lwage exper tenure married south urban black (educ iq = kww meduc feduc
Instrumental variables (2SLS) regression
```

Source	SS	df	MS	Number of obs = 722		
Model	-.295429993	8	-.036928749	F( 8, 713) = 18.74		
Residual	127.107346	713	.178271172	Prob > F = 0.0000		
Total	126.811916	721	.175883378	R-squared =		
				Adj R-squared =		
				Root MSE = .42222		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
educ	.1646904	.1132659	1.45	0.146	-.0576843	.3870651
iq	-.0102736	.0200124	-0.51	0.608	-.0495638	.0290166
exper	.0313987	.0122537	2.56	0.011	.007341	.0554564
tenure	.0070476	.0033717	2.09	0.037	.0004279	.0136672
married	.2133365	.0535285	3.99	0.000	.1082442	.3184289
south	-.0941667	.0506389	-1.86	0.063	-.1935859	.0052525
urban	.1680721	.0384337	4.37	0.000	.0926152	.2435289
black	-.2345713	.2247568	-1.04	0.297	-.6758356	.2066929



<code>_cons</code>		4.932962	.4870124	10.13	0.000	3.976812	5.889112
-----							
Instrumented:		<code>educ iq</code>					
Instruments:		<code>exper tenure married south urban black kww meduc feduc sibs</code>					
-----							

The estimated return to education is very large, but imprecisely estimated. The 95% confidence interval is very wide, and easily includes zero. Interestingly, the coefficient on *iq* is actually negative, and not statistically different from zero. The large IV estimate of the return to education and the insignificant ability indicator lend some support to the idea that omitted ability is less of a problem than schooling measurement error in the standard  $\log(\text{wage})$  model estimated by OLS. But the evidence is not very convincing given the very wide confidence interval for the *educ* coefficient.

**5.9.** Define  $\theta_4 = \beta_4 - \beta_3$ , so that  $\beta_4 = \beta_3 + \theta_4$ . Plugging this expression into the equation and rearranging gives

$$\begin{aligned}\log(\text{wage}) &= \beta_0 + \beta_1 \text{exper} + \beta_2 \text{exper}^2 + \beta_3 (\text{twoyr} + \text{fouryr}) + \theta_4 \text{fouryr} + u \\ &= \beta_0 + \beta_1 \text{exper} + \beta_2 \text{exper}^2 + \beta_3 \text{totcoll} + \theta_4 \text{fouryr} + u,\end{aligned}$$

where  $\text{totcoll} = \text{twoyr} + \text{fouryr}$ . Now, just estimate the latter equation by 2SLS using *exper*,  $\text{exper}^2$ , *dist2yr* and *dist4yr* as the full set of instruments. We can use the *t* statistic on  $\hat{\theta}_4$  to test  $H_0 : \theta_4 = 0$  against  $H_1 : \theta_4 > 0$ .

**5.10.** a. For  $\hat{\beta}_1$ , the lower right hand element in the general formula (5.24) with  $\mathbf{x} = (1, x)$  and  $\mathbf{z} = (1, z)$  is

$$\sigma^2[\text{Cov}(z, x)^2 / \text{Var}(z)].$$

Alternatively, you can derive this formula directly by writing

$$\sqrt{N}(\hat{\beta}_1 - \beta_1) = \left( N^{-1} \sum_{i=1}^N (z_i - \bar{z})(x_i - \bar{x}) \right)^{-1} \left[ N^{-1/2} \sum_{i=1}^N (z_i - \bar{z})u_i \right]$$

Now,  $\rho_{zx}^2 = [\text{Cov}(z, x)]^2 / (\sigma_z^2 \sigma_x^2)$ , so simple algebra shows that the asymptotic variance is  $\sigma^2 / (\rho_{zx}^2 \sigma_x^2)$ . The asymptotic variance for the OLS estimator is  $\sigma^2 / \sigma_x^2$ . Thus, the difference is the presence of  $\rho_{zx}^2$  in the denominator of the IV asymptotic variance.

b. Naturally, as the error variance  $\sigma^2$  increases so does the asymptotic variance of the IV estimator. More variance in  $x$  in the population is better for estimating  $\beta_1$ : as  $\sigma_x^2$  increases the asymptotic variance decreases. These effects are identical to the findings for OLS. A larger correlation between  $z$  and  $x$  reduces the asymptotic variance of the IV estimator. As  $\rho_{zx} \rightarrow 0$  the asymptotic variance increases without bound. This illustrates why an instrument that is only weakly correlated with  $x$  can lead to very imprecise IV estimators.

**5.11.** Following the hint, let  $y_2^0$  be the linear projection of  $y_2$  on  $\mathbf{z}_2$ , let  $a_2$  be the projection error, and assume that  $\lambda_2$  is known. (The results on generated regressors in Section 6.1.1 show that the argument carries over to the case when  $\lambda_2$  is estimated.) Plugging in  $y_2 = y_2^0 + a_2$  gives

$$y_1 = \mathbf{z}_1 \delta_1 + \alpha_1 y_2^0 + \alpha_1 a_2 + u_1.$$

Effectively, we regress  $y_1$  on  $\mathbf{z}_1, y_2^0$ . The key consistency condition is that each explanatory is orthogonal to the composite error,  $\alpha_1 a_2 + u_1$ . By assumption,  $E(\mathbf{z}_1' u_1) = \mathbf{0}$ . Further,  $E(y_2^0 a_2) = 0$  by construction. The problem is that, in general,  $E(\mathbf{z}_1' a_2) \neq \mathbf{0}$  because  $\mathbf{z}_1$  was not included in the linear projection for  $y_2$ . Therefore, OLS will be inconsistent for all parameters in general. Contrast this conclusion with 2SLS when  $y_2^*$  is the projection on  $\mathbf{z}_1$  and  $\mathbf{z}_2$ :

$$\begin{aligned} y_2 &= y_2^* + r_2 = \mathbf{z} \boldsymbol{\pi}_2 + r_2 \\ E(\mathbf{z}' r_2) &= \mathbf{0} \end{aligned}$$

The second step regression (assuming that  $\boldsymbol{\pi}_2$  is known) is

$$y_1 = \mathbf{z}_1 \delta_1 + \alpha_1 y_2^* + \alpha_1 r_2 + u_1.$$

By construction  $r_2$  is uncorrelated with  $\mathbf{z}$ , and so  $E(\mathbf{z}'_1 r_2) = \mathbf{0}$  and  $E(y_2^* r_2) = 0$ .

The lesson is that one must be very careful if manually carrying out 2SLS by explicitly doing the first- and second- stage regressions: all exogenous variables must be included in the first stage.

**5.12.** This problem is essentially proven by the hint. Given the description of  $\Pi$ , the only way the  $K$  columns of  $\Pi$  can be linearly dependent is if the last column can be written as a linear combination of the first  $K - 1$  columns. This is true if and only if each  $\theta_j$  is zero. Thus, if at least one  $\theta_j$  is different from zero,  $\text{rank}(\Pi) = K$ .

**5.13.** a. In a simple regression model with a single IV, the IV estimate of the slope can be written as

$$\begin{aligned}\hat{\beta}_1 &= \left( \sum_{i=1}^N (z_i - \bar{z})(y_i - \bar{y}) \right) / \left( \sum_{i=1}^N (z_i - \bar{z})(x_i - \bar{x}) \right) \\ &= \left( \sum_{i=1}^N z_i (y_i - \bar{y}) \right) / \left( \sum_{i=1}^N z_i (x_i - \bar{x}) \right).\end{aligned}$$

Now the numerator can be written as

$\sum_{i=1}^N z_i (y_i - \bar{y}) = \sum_{i=1}^N z_i y_i - \left( \sum_{i=1}^N z_i \right) \bar{y} = N_1 \bar{y}_1 - N_1 \bar{y} = N_1 (\bar{y}_1 - \bar{y})$  where  $N_1 = \sum_{i=1}^N z_i$  is the number of observations in the sample with  $z_i = 1$  and  $\bar{y}_1$  is the average of the  $y_i$  over the observations with  $z_i = 1$ . Next, write  $\bar{y}$  as a weighted average:

$$\bar{y} = (N_0/N)\bar{y}_0 + (N_1/N)\bar{y}_1$$

where the notation should be clear. Straightforward algebra shows that

$$\begin{aligned}\bar{y}_1 - \bar{y} &= [(N - N_1)/N]\bar{y} - (N_0/N)\bar{y}_0 \\ &= (N_0/N)(\bar{y}_1 - \bar{y}_0).\end{aligned}$$

Therefore, the numerator of the IV estimate is  $(N_0 N_1 / N)(\bar{y}_1 - \bar{y}_0)$ . The same argument shows

that the denominator is  $(N_0N_1/N)(\bar{x}_1 - \bar{x}_0)$ . Taking the ratio proves the result.

b. If  $x$  is also binary – representing some “treatment” –  $\bar{x}_1$  is the fraction of observations receiving treatment when  $z_i = 1$  and  $\bar{x}_0$  is the fraction receiving treatment when  $z_i = 0$ . Suppose  $x_i = 1$  if person  $i$  participates in a job training program, and let  $z_i = 1$  if person  $i$  is eligible for participation in the program. Then  $\bar{x}_1$  is the fraction of people participating in the program out of those made eligible, and  $\bar{x}_0$  is the fraction of people participating who are not eligible. (When eligibility is necessary for participation,  $\bar{x}_0 = 0$ .) Generally,  $\bar{x}_1 - \bar{x}_0$  is the difference in participation rates when  $z = 1$  and  $z = 0$ . So the difference in the mean response between the  $z = 1$  and  $z = 0$  groups gets divided by the difference in participation rates across the two groups.

**5.14.** a. Taking the linear projection of (5.1) under the assumption that  $(x_1, \dots, x_{K-1}, z_i, \dots, z_M)$  are uncorrelated with  $u$  gives

$$\begin{aligned} L(y|\mathbf{z}) &= \beta_0 + \beta_1x_1 + \dots + \beta_{K-1}x_{K-1} + \beta_KL(x_K|\mathbf{z}) + L(u|\mathbf{z}) \\ &= \beta_0 + \beta_1x_1 + \dots + \beta_{K-1}x_{K-1} + \beta_Kx_K^* \end{aligned}$$

because  $L(u|\mathbf{z}) = 0$ .

b. By the law of iterated projections,

$$L(y|1, x_1, \dots, x_{K-1}, x_K^*) = \beta_0 + \beta_1x_1 + \dots + \beta_{K-1}x_{K-1} + \beta_Kx_K^*.$$

Consistency of OLS for the  $\beta_j$  from the regression  $y$  on  $1, x_1, \dots, x_{K-1}, x_K^*$  follows immediately from our treatment of OLS from Chapter 4: OLS consistently estimates the parameters in a linear projection provided there is not perfect collinearity in  $(1, x_1, \dots, x_{K-1}, x_K^*)$ .

c. I should have said explicitly to assume  $E(\mathbf{z}'\mathbf{z})$  is nonsingular – that is, 2SLS.2a holds. Then,  $x_K^*$  is not a perfect linear combination of  $(x_1, \dots, x_{K-1})$  if and only if at least one element

of  $z_1, \dots, z_M$  has nonzero coefficient in  $L(x_K|1, x_1, \dots, x_{K-1}, z_1, \dots, z_M)$ . In the model with a single endogenous explanatory variable, we know this condition is equivalent to Assumption 2SLS.2b, the standard rank condition.

5.15. In  $L(\mathbf{x}|\mathbf{z}) = \mathbf{z}\Pi$  we can write

$$\Pi = \begin{pmatrix} \Pi_{11} & \mathbf{0} \\ \Pi_{12} & \mathbf{I}_{K_2} \end{pmatrix},$$

where  $\mathbf{I}_{K_2}$  is the  $K_2 \times K_2$  identity matrix,  $\mathbf{0}$  is the  $L_1 \times K_2$  zero matrix,  $\Pi_{11}$  is  $L_1 \times K_1$ , and  $\Pi_{12}$  is  $K_2 \times K_1$ . As in Problem 5.12, the rank condition holds if and only if  $\text{rank}(\Pi) = K$ .

a. If for some  $x_j$ , the vector  $\mathbf{z}_1$  does not appear in  $L(x_j|\mathbf{z})$ , then  $\Pi_{11}$  has a column which is entirely zeros. Then that column of  $\Pi$  can be written as a linear combination of the last  $K_2$  columns of  $\Pi$  – because any  $K_2 \times 1$  vector in  $\Pi_{12}$  can be written as a linear combination of the columns of  $\mathbf{I}_{K_2}$ . This implies  $\text{rank}(\Pi) < K$ . Therefore, a necessary condition for the rank condition is that no columns of  $\Pi_{11}$  be exactly zero, which means that at least one  $z_h$  must appear in the reduced form of each  $x_j$ ,  $j = 1, \dots, K_1$ .

b. Suppose  $K_1 = 2$  and  $L_1 = 2$ , where  $z_1$  appears in the reduced form from both  $x_1$  and  $x_2$ , but  $z_2$  appears in neither reduced form. Then the  $2 \times 2$  matrix  $\Pi_{11}$  has zeros in its second row, which means that the second row of  $\Pi$  is all zeros. In that case, it cannot have rank  $K$ . Intuitively, while we began with two instruments, only one of them turned out to be partially correlated with  $x_1$  and  $x_2$ .

c. Without loss of generality, assume that  $z_j$  appears in the reduced form for  $x_j$ ; we can simply reorder the elements of  $\mathbf{z}_1$  to ensure this is the case. Then  $\Pi_{11}$  is a  $K_1 \times K_1$  diagonal matrix with nonzero diagonal elements. Looking at

$$\mathbf{\Pi} = \begin{pmatrix} \mathbf{\Pi}_{11} & \mathbf{0} \\ \mathbf{\Pi}_{12} & \mathbf{I}_{K_2} \end{pmatrix}$$

we see that if  $\mathbf{\Pi}_{11}$  is diagonal with all nonzero diagonals then  $\mathbf{\Pi}$  is lower triangular with all diagonal elements nonzero. Therefore,  $\text{rank } \mathbf{\Pi} = K$ .

**5.16.** a. The discussion below equation (5.24) implies directly that

$$\text{Avar} \sqrt{N} (\tilde{\beta} - \beta) = \sigma_u^2 / \text{Var}(w^*)$$

because there are no other explanatory variables – exogenous or endogenous – in the equation.

Remember, the expression

$$\sigma_u^2 [\text{E}(\mathbf{x}^* \mathbf{x}^*)]^{-1}$$

has the same form as that for OLS but with  $\mathbf{x}^*$  replacing  $\mathbf{x}$ . So any algebra derived for OLS can be applied to 2SLS.

b. We can write

$$v = u - \mathbf{h}\gamma,$$

so if  $\text{E}(\mathbf{g}'u) = \mathbf{0}$  and  $\text{E}(\mathbf{g}'\mathbf{h}) = \mathbf{0}$  then

$$\text{E}(\mathbf{g}'v) = \text{E}(\mathbf{g}'u) - \text{E}(\mathbf{g}'\mathbf{h})\gamma = \mathbf{0}.$$

c. For the hint here to be entirely correct, I should have stated that  $\text{E}(w) = 0$ . As we will see, when  $w$  has a nonzero mean,  $\check{r}$  differs from  $w^*$  by an additive constant [which, of course, implies  $\text{Var}(\check{r}) = \text{Var}(w^*)$ ].

Again using the discussion following equation (5.24),

$$\text{Avar} \sqrt{N} (\hat{\beta} - \beta) = \sigma_v^2 / \text{Var}(\check{r}),$$

where  $\sigma_v^2 = \text{Var}(v)$ ,  $\check{r}$  is the population residual from the regression  $\check{w}$  on  $1, \mathbf{h}$ , and  $\check{w}$  are the

population fitted values from the linear projection of  $w$  on  $\mathbf{g}$ ,  $\mathbf{h}$ .

Because  $E(\mathbf{g}'\mathbf{h}) = \mathbf{0}$ , we can write

$$\check{w} = \mathbf{g}\boldsymbol{\pi}_1 + \mathbf{h}\boldsymbol{\pi}_2$$

where

$$\begin{aligned}\boldsymbol{\pi}_1 &= [E(\mathbf{g}'\mathbf{g})]^{-1}E(\mathbf{g}'w) \\ \boldsymbol{\pi}_2 &= [E(\mathbf{h}'\mathbf{h})]^{-1}E(\mathbf{h}'w).\end{aligned}$$

Note that

$$w^* = L(w|\mathbf{g}) = \mathbf{g}\boldsymbol{\pi}_1.$$

Next,

$$\begin{aligned}L(\check{w}|1, \mathbf{h}) &= L(\mathbf{g}\boldsymbol{\pi}_1 + \mathbf{h}\boldsymbol{\pi}_2|1, \mathbf{h}) = L(\mathbf{g}|1, \mathbf{h})\boldsymbol{\pi}_1 + \mathbf{h}\boldsymbol{\pi}_2 \\ &= L(\mathbf{g}|1)\boldsymbol{\pi}_1 + \mathbf{h}\boldsymbol{\pi}_2\end{aligned}$$

because  $E(\mathbf{h}) = \mathbf{0}$  and  $E(\mathbf{g}'\mathbf{h}) = \mathbf{0}$  are assumed. Now  $L(\mathbf{g}|1) = E(\mathbf{g})$ , and so

$$L(\check{w}|1, \mathbf{h}) = \eta_1 + \mathbf{h}\boldsymbol{\pi}_2$$

where  $\eta_1 = \boldsymbol{\mu}_g\boldsymbol{\pi}_1$ . Therefore,

$$\begin{aligned}\check{r} &= \check{w} - L(\check{w}|1, \mathbf{h}) = (\mathbf{g}\boldsymbol{\pi}_1 + \mathbf{h}\boldsymbol{\pi}_2) - (\eta_1 + \mathbf{h}\boldsymbol{\pi}_2) = -\eta_1 + \mathbf{g}\boldsymbol{\pi}_1 \\ &= -\eta_1 + w^*\end{aligned}$$

It follows that  $\text{Var}(\check{r}) = \text{Var}(w^*)$  and so we have shown

$$\text{Avar}\sqrt{N}(\hat{\beta} - \beta) = \sigma_v^2/\text{Var}(w^*),$$

d. Because  $E(\mathbf{h}'v) = \mathbf{0}$  by definition, we have

$$\sigma_u^2 = \text{Var}(\mathbf{h}\boldsymbol{\gamma}) + \sigma_v^2 = \boldsymbol{\gamma}'\boldsymbol{\Sigma}_h\boldsymbol{\gamma} + \sigma_v^2 \geq \sigma_v^2,$$

with strict inequality if  $\boldsymbol{\Sigma}_h$  is positive definite and  $\boldsymbol{\gamma} \neq \mathbf{0}$  (and even in some cases where

$\Sigma_h \equiv \text{Var}(\mathbf{h})$  is not positive definite). This means that, asymptotically, we generally get a smaller asymptotic variance for estimate  $\beta$  by including exogenous variables that are uncorrelated with the instruments  $\mathbf{g}$ :

$$\begin{aligned} \text{Avar}\sqrt{N}(\tilde{\beta} - \beta) - \text{Avar}\sqrt{N}(\hat{\beta} - \beta) &= \frac{\sigma_u^2}{\text{Var}(w^*)} - \frac{\sigma_v^2}{\text{Var}(w^*)} \\ &= \frac{\gamma' \Sigma_h \gamma}{\text{Var}(w^*)} \geq 0. \end{aligned}$$



## Solutions to Chapter 6 Problems

6.1. a. Here is abbreviated Stata output for testing the null hypothesis that *educ* is

exogenous:

```
. use card

. qui reg educ nearc4 nearc2 exper expersq black south smsa reg661-reg668
  smsa66

. predict v2hat, resid

. reg lwage educ exper expersq black south smsa reg661-reg668 smsa66 v2hat
```

Source	SS	df	MS	Number of obs = 3010		
Model	178.100803	16	11.1313002	F( 16, 2993) = 80.37		
Residual	414.540842	2993	.138503455	Prob > F = 0.0000		
				R-squared = 0.3005		
				Adj R-squared = 0.2968		
Total	592.641645	3009	.196956346	Root MSE = .37216		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
educ	.1570594	.0482814	3.25	0.001	.0623912	.2517275
exper	.1188149	.0209423	5.67	0.000	.0777521	.1598776
expersq	-.0023565	.0003191	-7.38	0.000	-.0029822	-.0017308
black	-.1232778	.0478882	-2.57	0.010	-.2171749	-.0293806
south	-.1431945	.0261202	-5.48	0.000	-.1944098	-.0919791
smsa	.100753	.0289435	3.48	0.001	.0440018	.1575042
reg661	-.102976	.0398738	-2.58	0.010	-.1811588	-.0247932
reg662	-.0002286	.0310325	-0.01	0.994	-.0610759	.0606186
reg663	.0469556	.0299809	1.57	0.117	-.0118296	.1057408
reg664	-.0554084	.0359807	-1.54	0.124	-.1259578	.0151411
reg665	.0515041	.0436804	1.18	0.238	-.0341426	.1371509
reg666	.0699968	.0489487	1.43	0.153	-.0259797	.1659734
reg667	.0390596	.0456842	0.85	0.393	-.050516	.1286352
reg668	-.1980371	.0482417	-4.11	0.000	-.2926273	-.1034468
smsa66	.0150626	.0205106	0.73	0.463	-.0251538	.0552789
v2hat	-.0828005	.0484086	-1.71	0.087	-.177718	.0121169
_cons	3.339687	.821434	4.07	0.000	1.729054	4.950319

The  $t$  statistic on  $\hat{v}_2$  is  $-1.71$ , which is not significant at the 5% level against a two-sided alternative. The negative correlation between  $u_1$  and *educ* is essentially the same finding that the 2SLS estimated return to education is larger than the OLS estimate. In any case, I would call this marginal evidence that *educ* is endogenous. The quandary is that the OLS and 2SLS

point estimates are quite different.

b. To test the single over indentifying restriction we obtain the 2SLS residuals:

```
. qui reg lwage educ exper expersq black south smsa reg661-reg668 smsa66
      (nearc4 nearc2 exper expersq black south smsa reg661-reg668 smsa66)

. predict uhat1, resid

. qui reg ulhat exper expersq black south smsa reg661-reg668 smsa66 nearc4
      nearc2

. di e(r2)
.00041467

. di 3010*e(r2)
1.2481535

. di chiprob(1,3010*e(r2))
.26390545
```

The test statistic is the sample size times the R-squared from this regression, or about 1.25.

The  $p$ -value, obtained from  $\chi^2_1$  distribution, is about .264, so the instruments pass the over identification test.

**6.2.** We first obtain the reduced form residuals,  $\hat{v}_{21}$  and  $\hat{v}_{22}$ , for *educ* and *IQ*, respectively.

The regression output is suppressed:

```
. qui reg educ exper tenure married south urban black kww meduc feduc sibs

. predict v21hat, resid
(213 missing values generated)

. qui reg iq exper tenure married south urban black kww meduc feduc sibs

. predict v22hat, resid
(213 missing values generated)

. qui reg lwage exper tenure married south urban black educ iq v21hat v22hat

. test v21hat v22hat

( 1)  v21hat = 0
( 2)  v22hat = 0

      F( 2, 711) =      4.20
      Prob > F =      0.0153
```

The  $p$ -value of the joint  $F$  test, which is justified asymptotically, is .0153. Therefore, the

test finds fairly strong evidence for endogeneity of at least one of *educ* and *IQ*, although this conclusion relies on the instruments being truly exogenous. If you look back at Problem 5.8, this IV solution did not seem to work very well. So we still do not know what should be treated as exogenous in this method.

**6.3. a.** We need prices to satisfy two requirements. First, *calories* and *protein* must be partially correlated with prices of food. While this is easy to test separately by estimating the two reduced forms, the rank condition could still be violated. (Problem 5.15c contains a sufficient condition for the rank condition to hold.) In addition, we must also assume prices are exogenous in the productivity equation. Ideally, prices vary because of things like transportation costs that are not systematically related to regional variations in individual productivity. A potential problem is that prices reflect food quality and that features of the food other than calories and protein appear in the disturbance  $u_1$ .

b. Since there are two endogenous explanatory variables we need at least two prices.

c. We would first estimate the two reduced forms for *calories* and *protein* by regressing each on a constant, *exper*,  $exper^2$ , *educ*, and the  $M$  prices,  $p_1, \dots, p_M$ . We obtain the residuals,  $\hat{v}_{21}$  and  $\hat{v}_{22}$ . Then we would run the regression  $\log(produced)$  on  $1, exper, exper^2, educ, \hat{v}_{21}, \hat{v}_{22}$  and do a joint significance test on  $\hat{v}_{21}$  and  $\hat{v}_{22}$ . We could use a standard  $F$  test or use a heteroskedasticity-robust test.

**6.4.a.** Since  $y = \mathbf{x}\boldsymbol{\beta} + q + v$  it follows that

$$E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} + E(q|\mathbf{x}) + E(v|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} + \mathbf{x}\boldsymbol{\delta} = \mathbf{x}(\boldsymbol{\beta} + \boldsymbol{\delta}) \equiv \mathbf{x}\boldsymbol{\gamma}.$$

Since  $E(y|\mathbf{x})$  is linear in  $\mathbf{x}$  there is no functional form misspecification in this conditional expectation. Therefore, no functional form test will detect correlation between  $q$  and  $\mathbf{x}$ , no matter how strong it is:  $\boldsymbol{\delta}$  can be anything.

b. Since  $E(v|\mathbf{x}, q) = 0$ ,  $\text{Var}(v|\mathbf{x}, q) = E(v^2|\mathbf{x}, q) = \sigma_v^2 = E(v^2|\mathbf{x}) = \text{Var}(v|\mathbf{x})$ . Therefore,  $\text{Var}(y|\mathbf{x}) = \text{Var}(q + v|\mathbf{x}) = \text{Var}(q|\mathbf{x}) + \text{Var}(v|\mathbf{x}) + 2E(qv|\mathbf{x})$ , where we use  $\text{Cov}(q, v|\mathbf{x}) = E(qv|\mathbf{x})$  because  $E(v|\mathbf{x}) = 0$ . Now

$$E(qv|\mathbf{x}) = E[E(qv|\mathbf{x}, q)|\mathbf{x}] = E[qE(v|\mathbf{x}, q)|\mathbf{x}] = E[q \cdot 0|\mathbf{x}] = 0.$$

Therefore,  $\text{Var}(y|\mathbf{x}) = \text{Var}(q|\mathbf{x}) + \text{Var}(v|\mathbf{x}) = \sigma_q^2 + \sigma_v^2$ , so that  $y$  is conditionally homoskedastic. But if  $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\gamma}$  and  $\text{Var}(y|\mathbf{x})$  is constant, a test for heteroskedasticity will always have a limiting chi-square distribution. It will have no power for detecting omitted variables.

c. Since  $E(u^2|\mathbf{x}) = \text{Var}(u|\mathbf{x}) + [E(u|\mathbf{x})]^2$  and  $\text{Var}(u|\mathbf{x})$  is constant,  $E(u^2|\mathbf{x})$  is constant if and only if  $E[(u|\mathbf{x})]^2$  is constant. If  $E(u|\mathbf{x}) \neq E(u)$  then  $E(u|\mathbf{x})$  is not constant, so  $[E(u|\mathbf{x})]^2$  generally will be a function of  $\mathbf{x}$ . So  $E(u^2|\mathbf{x})$  depends on  $\mathbf{x}$ , which means that  $u^2$  can be correlated with functions of  $\mathbf{x}$ , say  $\mathbf{h}(\mathbf{x})$ . It follows that regression tests of the form (6.36) can be expected, at least in some cases, to detect “heteroskedasticity”. (If the goal is to determine when heteroskedasticity-robust inference is called for, the regression-based tests do the right thing.)

**6.5.** a. For simplicity, absorb the intercept in  $\mathbf{x}$ , so  $y = \mathbf{x}\boldsymbol{\beta} + u$ ,  $E(u|\mathbf{x}) = 0$ ,  $\text{Var}(u|\mathbf{x}) = \sigma^2$ . In these tests,  $\hat{\sigma}^2$  is implicitly  $SSR/N$  – there is no degrees of freedom adjustment. (In any case, the  $df$  adjustment makes no difference asymptotically.) So  $\hat{u}_i^2 - \hat{\sigma}^2$  has a zero sample average, which means that

$$N^{-1/2} \sum_{i=1}^N (\mathbf{h}_i - \boldsymbol{\mu}_h)' (\hat{u}_i^2 - \hat{\sigma}^2) = N^{-1/2} \sum_{i=1}^N \mathbf{h}_i' (\hat{u}_i^2 - \hat{\sigma}^2).$$

Next,  $N^{-1/2} \sum_{i=1}^N (\mathbf{h}_i - \boldsymbol{\mu}_h)' = O_p(1)$  by the central limit theorem and  $\hat{\sigma}^2 - \sigma^2 = o_p(1)$ . So

$N^{-1/2} \sum_{i=1}^N (\mathbf{h}_i - \boldsymbol{\mu}_h)' (\hat{\sigma}^2 - \sigma^2) = O_p(1) \cdot o_p(1) = o_p(1)$ . Therefore, so far we have

$$N^{-1/2} \sum_{i=1}^N \mathbf{h}_i' (\hat{u}_i^2 - \hat{\sigma}^2) = N^{-1/2} \sum_{i=1}^N (\mathbf{h}_i - \boldsymbol{\mu}_h)' (\hat{u}_i^2 - \hat{\sigma}^2) + o_p(1).$$

We are done with this part if we show

$N^{-1/2} \sum_{i=1}^N (\mathbf{h}_i - \boldsymbol{\mu}_h)' \hat{u}_i^2 = N^{-1/2} \sum_{i=1}^N (\mathbf{h}_i - \boldsymbol{\mu}_h)' \hat{u}_i^2 + o_p(1)$ . Now, as in Problem 4.4, we can

write  $\hat{u}_i^2 = u_i^2 - 2u_i \mathbf{x}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + [\mathbf{x}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]^2$ , so

$$\begin{aligned} N^{-1/2} \sum_{i=1}^N (\mathbf{h}_i - \boldsymbol{\mu}_h)' \hat{u}_i^2 &= N^{-1/2} \sum_{i=1}^N (\mathbf{h}_i - \boldsymbol{\mu}_h)' \hat{u}_i^2 \\ &\quad - 2 \left[ N^{-1/2} \sum_{i=1}^N (\mathbf{h}_i - \boldsymbol{\mu}_h)' \mathbf{x}_i \right] (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\quad + \left[ N^{-1/2} \sum_{i=1}^N (\mathbf{h}_i - \boldsymbol{\mu}_h)' (\mathbf{x}_i \otimes \mathbf{x}_i) \right] \{ \text{vec}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \}, \end{aligned} \tag{6.62}$$

where the expression for the third term follows from

$[\mathbf{x}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]^2 = \mathbf{x}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{x}_i = (\mathbf{x}_i \otimes \mathbf{x}_i) \cdot \text{vec}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})']$ . Dropping the “ $-2$ ” the second term can be written as  $(N^{-1} \sum_{i=1}^N u_i (\mathbf{h}_i - \boldsymbol{\mu}_h)' \mathbf{x}_i) \sqrt{N} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = o_p(1) \cdot O_p(1)$  because  $\sqrt{N} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_p(1)$  and, under  $E(u_i | \mathbf{x}_i) = 0$ ,  $E[u_i (\mathbf{h}_i - \boldsymbol{\mu}_h)' \mathbf{x}_i] = 0$ ; the law-of-large-numbers implies that the sample average is  $o_p(1)$ . The third term can be written as

$$N^{-1/2} \left[ N^{-1} \sum_{i=1}^N (\mathbf{h}_i - \boldsymbol{\mu}_h)' (\mathbf{x}_i \otimes \mathbf{x}_i) \right] \{ \text{vec}[\sqrt{N} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sqrt{N} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \} = N^{-1/2} \cdot O_p(1) \cdot O_p(1),$$

where we again use the fact that sample averages are  $O_p(1)$  by the law of large numbers and  $\text{vec}[\sqrt{N} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sqrt{N} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] = O_p(1)$ . We have shown that the last two terms in (6.62) are  $o_p(1)$ , which proves part a.

b. By part a, the asymptotic variance of  $N^{-1/2} \sum_{i=1}^N \mathbf{h}_i' (\hat{u}_i^2 - \sigma^2)$  is

$$\text{Var}[(\mathbf{h}_i - \boldsymbol{\mu}_h)' (u_i^2 - \sigma^2)] = E[(u_i^2 - \sigma^2)^2 (\mathbf{h}_i - \boldsymbol{\mu}_h)' (\mathbf{h}_i - \boldsymbol{\mu}_h)]. \text{ Now}$$

$(u_i^2 - \sigma^2)^2 = u_i^4 - 2u_i^2\sigma^2 + \sigma^4$ . Under the null,  $E(u_i^2|\mathbf{x}_i) = \text{Var}(u_i|\mathbf{x}_i) = \sigma^2$  [since  $E(u_i|\mathbf{x}_i) = 0$  is assumed] and therefore, when we add (6.37),  $E[(u_i^2 - \sigma^2)^2|\mathbf{x}_i] = \kappa^2 - \sigma^4 \equiv \eta^2$ . A standard iterated expectations argument gives

$$E[(u_i^2 - \sigma^2)^2(\mathbf{h}_i - \boldsymbol{\mu}_h)'(\mathbf{h}_i - \boldsymbol{\mu}_h)] = E\{E[(u_i^2 - \sigma^2)^2(\mathbf{h}_i - \boldsymbol{\mu}_h)'(\mathbf{h}_i - \boldsymbol{\mu}_h)]|\mathbf{x}_i\} = E\{E[(u_i^2 - \sigma^2)^2|$$

which is what we wanted to show. (Whether we carry out the calculation for a random draw  $i$  or for random variables representing the population is a matter of taste.)

c. From part b and Lemma 3.8, the following statistic has an asymptotic  $\chi_Q^2$  distribution:

$$\left[ N^{-1/2} \sum_{i=1}^N (\hat{u}_i^2 - \hat{\sigma}^2) \mathbf{h}_i \right] \{ \eta^2 E[(\mathbf{h}_i - \boldsymbol{\mu}_h)'(\mathbf{h}_i - \boldsymbol{\mu}_h)] \}^{-1} \left[ N^{-1/2} \sum_{i=1}^N \mathbf{h}_i' (\hat{u}_i^2 - \hat{\sigma}^2) \right].$$

Using again the fact that  $\sum_{i=1}^N (\hat{u}_i^2 - \hat{\sigma}^2) = 0$ , we can replace  $\mathbf{h}_i$  with  $\mathbf{h}_i - \bar{\mathbf{h}}$  in the two vectors forming the quadratic form. Then, again by Lemma 3.8, we can replace the matrix in the quadratic form with a consistent estimator, which is

$$\hat{\eta}^2 \left[ N^{-1} \sum_{i=1}^N (\mathbf{h}_i - \bar{\mathbf{h}})'(\mathbf{h}_i - \bar{\mathbf{h}}) \right],$$

where  $\hat{\eta}^2 = N^{-1} \sum_{i=1}^N (\hat{u}_i^2 - \hat{\sigma}^2)^2$ . The computable statistic, after simple algebra, can be written as

$$\hat{\eta}^{-2} \left( \sum_{i=1}^N (\hat{u}_i^2 - \hat{\sigma}^2)(\mathbf{h}_i - \bar{\mathbf{h}}) \right) \left( \sum_{i=1}^N (\mathbf{h}_i - \bar{\mathbf{h}})'(\mathbf{h}_i - \bar{\mathbf{h}}) \right)^{-1} \left( \sum_{i=1}^N (\mathbf{h}_i - \bar{\mathbf{h}})'(\hat{u}_i^2 - \hat{\sigma}^2) \right).$$

Now  $\hat{\eta}^2$  is just the total sum of squares of the  $\hat{u}_i^2$  divided by  $N$ . The numerator of the statistic is simply the explained sum of squares from the regression  $\hat{u}_i^2$  on  $1, \mathbf{h}_i$ ,  $i = 1, \dots, N$ . Therefore, the test statistic is  $N$  times the usual (centered)  $R$ -squared from the regression  $\hat{u}_i^2$  on  $1, \mathbf{h}_i, i = 1, \dots, N$ , or  $NR_c^2$ .

d. Without assumption (6.37) we need to estimate  $E[(u_i^2 - \sigma^2)^2(\mathbf{h}_i - \boldsymbol{\mu}_h)'(\mathbf{h}_i - \boldsymbol{\mu}_h)]$  generally. Hopefully, the approach is by now pretty clear. We replace the population expected value with the sample average and replace any unknown parameters –  $\beta, \sigma^2$ , and  $\boldsymbol{\mu}_h$  in this case – with their consistent estimators (under  $H_0$ ). So a generally consistent estimator of  $\text{Avar}\left(N^{-1/2} \sum_{i=1}^N \mathbf{h}_i'(\hat{u}_i^2 - \hat{\sigma}^2)\right)$  is

$$N^{-1} \sum_{i=1}^N (\hat{u}_i^2 - \hat{\sigma}^2)^2 (\mathbf{h}_i - \bar{\mathbf{h}})' (\mathbf{h}_i - \bar{\mathbf{h}}),$$

and the test statistic robust to heterokurtosis can be written as

$$\left( \sum_{i=1}^N (\hat{u}_i^2 - \hat{\sigma}^2)(\mathbf{h}_i - \bar{\mathbf{h}}) \right) \left( \sum_{i=1}^N (\hat{u}_i^2 - \hat{\sigma}^2)^2 (\mathbf{h}_i - \bar{\mathbf{h}})' (\mathbf{h}_i - \bar{\mathbf{h}}) \right)^{-1} \\ \cdot \left( \sum_{i=1}^N (\mathbf{h}_i - \bar{\mathbf{h}})' (\hat{u}_i^2 - \hat{\sigma}^2) \right),$$

which is easily seen to be the explained sum of squares from the regression of 1 on  $(\hat{u}_i^2 - \hat{\sigma}^2)(\mathbf{h}_i - \bar{\mathbf{h}}), i = 1, \dots, N$  (without an intercept). Since the total sum of squares, without demeaning, of unity is simply  $N$ , the statistic is equivalent to  $N - SSR_0$ , where  $SSR_0$  is the sum of squared residuals.

**6.6.** Here is my Stata session using the data NLS80.RAW:

```
. qui reg lwage exper tenure married south urban black educ
. predict lwageh
(option xb assumed; fitted values)
. gen lwagehsq = lwageh^2
. predict uhat, resid
. gen uhatsq = uhat^2
. reg uhatsq lwageh lwagehsq
```

Source		SS	df	MS	Number of obs =	935
--------	--	----	----	----	-----------------	-----

Model	.288948987	2	.144474493	F( 2, 932) =	2.43
Residual	55.3447136	932	.05938274	Prob > F	= 0.0883
				R-squared	= 0.0052
				Adj R-squared	= 0.0031
Total	55.6336626	934	.059564949	Root MSE	= .24369

uhatsq	Coef.	Std. Err.	t	P> t	[95% Conf. Interval
lwageh	3.027285	1.880375	1.61	0.108	-.6629745 6.717544
lwagehsq	-.2280088	.1390444	-1.64	0.101	-.5008853 .0448677
_cons	-9.901227	6.353656	-1.56	0.119	-22.37036 2.567902

An asymptotically valid test for heteroskedasticity is just the  $F$  statistic for joint significance of  $\hat{y}$  and  $\hat{y}^2$ , and this yields  $p$  – value = .088 (although this version maintains Assumption (6.37) under the null, along with homoskedasticity). Thus, there is only modest evidence of heteroskedasticity. It could be ignored or heteroskedasticity-robust standard errors and test statistics can be used.

#### 6.7. a. The simple regression results are:

```
. use hprice
```

```
. reg lprice ldist if y81
```

Source	SS	df	MS	Number of obs =	142
Model	3.86426989	1	3.86426989	F( 1, 140) =	30.79
Residual	17.5730845	140	.125522032	Prob > F	= 0.0000
				R-squared	= 0.1803
				Adj R-squared	= 0.1744
Total	21.4373543	141	.152037974	Root MSE	= .35429

lprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval
ldist	.3648752	.0657613	5.55	0.000	.2348615 .4948889
_cons	8.047158	.6462419	12.45	0.000	6.769503 9.324813

This regression suggests a strong link between housing price and distance from the incinerator (as distance increases, so does housing price). The elasticity is .365 and the  $t$  statistic is 5.55. However, this is not a good causal regression: the incinerator may have been put near homes with lower values to begin with. If so, we would expect the positive



relationship found in the simple regression even if the new incinerator had no effect on housing prices.

b. The parameter  $\delta_3$  should be positive: after the incinerator is built a house should be worth relatively more the farther it is from the incinerator. Here is the Stata session:

```
. gen y81ldist = y81*ldist
. reg lprice y81 ldist y81ldist
```

Source	SS	df	MS	Number of obs = 321		
Model	24.3172548	3	8.10575159	F( 3, 317) = 69.22		
Residual	37.1217306	317	.117103251	Prob > F = 0.0000		
Total	61.4389853	320	.191996829	R-squared = 0.3958		
				Adj R-squared = 0.3901		
				Root MSE = .3422		

lprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
y81	-.0113101	.8050622	-0.01	0.989	-1.59525	1.57263
ldist	.316689	.0515323	6.15	0.000	.2153005	.4180775
y81ldist	.0481862	.0817929	0.59	0.556	-.1127394	.2091117
_cons	8.058468	.5084358	15.85	0.000	7.058133	9.058803

The coefficient on *ldist* reveals the shortcoming of the regression in part a. This coefficient measures the relationship between *lprice* and *ldist* in 1978, before the incinerator was even being rumored. The effect of the incinerator is given by the coefficient on the interaction, *y81ldist*. While the direction of the effect is as expected, it is not especially large, and it is statistically insignificant, anyway. Therefore, at this point, we cannot reject the null hypothesis that building the incinerator had no effect on housing prices.

c. Adding the variables listed in the problem gives

```
. reg lprice y81 ldist y81ldist lintst lintstsq larea lland age agesq rooms
baths
```

Source	SS	df	MS	Number of obs = 321		
Model	48.7611143	11	4.43282858	F( 11, 309) = 108.04		
Residual	12.677871	309	.041028709	Prob > F = 0.0000		
Total	61.4389853	320	.191996829	R-squared = 0.7937		
				Adj R-squared = 0.7863		
				Root MSE = .20256		

lprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
y81	-.229847	.4877198	-0.47	0.638	-1.189519	.7298249
ldist	.0866424	.0517205	1.68	0.095	-.0151265	.1884113
y81ldist	.0617759	.0495705	1.25	0.214	-.0357625	.1593143
lintst	.9633332	.3262647	2.95	0.003	.3213517	1.605315
lintstsq	-.0591504	.0187723	-3.15	0.002	-.096088	-.0222128
larea	.3548562	.0512328	6.93	0.000	.2540468	.4556655
lland	.109999	.0248165	4.43	0.000	.0611683	.1588297
age	-.0073939	.0014108	-5.24	0.000	-.0101699	-.0046178
agesq	.0000315	8.69e-06	3.63	0.000	.0000144	.0000486
rooms	.0469214	.0171015	2.74	0.006	.0132713	.0805715
baths	.0958867	.027479	3.49	0.001	.041817	.1499564
_cons	2.305525	1.774032	1.30	0.195	-1.185185	5.796236

The incinerator effect is now larger (the elasticity is about .062) and the  $t$  statistic is larger, but the  $p$ -value for the interaction term is still fairly large, .214. Against a one-sided alternative, the  $p$ -value is .107, so it is almost significant at the 10% level. Still, using these two years of data and controlling for the listed factors, the evidence that housing prices were adversely affected by the new incinerator is somewhat weak.

#### 6.8. a. The following is my Stata session:

```
. use fertill1
. gen agesq = age^2
. reg kids educ age agesq black east northcen west farm othrural town smcity
y74-y84
```

Source	SS	df	MS	Number of obs =	1129
Model	399.610888	17	23.5065228	F( 17, 1111) =	9.72
Residual	2685.89841	1111	2.41755033	Prob > F =	0.0000
Total	3085.5093	1128	2.73538059	R-squared =	0.1295
				Adj R-squared =	0.1162
				Root MSE =	1.5548

kids	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
educ	-.1284268	.0183486	-7.00	0.000	-.1644286	-.092425
age	.5321346	.1383863	3.85	0.000	.2606065	.8036626
agesq	-.005804	.0015643	-3.71	0.000	-.0088733	-.0027347
black	1.075658	.1735356	6.20	0.000	.7351631	1.416152
east	.217324	.1327878	1.64	0.102	-.0432192	.4778672
northcen	.363114	.1208969	3.00	0.003	.125902	.6003261
west	.1976032	.1669134	1.18	0.237	-.1298978	.5251041
farm	-.0525575	.14719	-0.36	0.721	-.3413592	.2362443

othrural	-.1628537	.175442	-0.93	0.353	-.5070887	.1813814
town	.0843532	.124531	0.68	0.498	-.1599893	.3286957
smcity	.2118791	.160296	1.32	0.187	-.1026379	.5263961
y74	.2681825	.172716	1.55	0.121	-.0707039	.6070689
y76	-.0973795	.1790456	-0.54	0.587	-.448685	.2539261
y78	-.0686665	.1816837	-0.38	0.706	-.4251483	.2878154
y80	-.0713053	.1827707	-0.39	0.697	-.42992	.2873093
y82	-.5224842	.1724361	-3.03	0.003	-.8608214	-.184147
y84	-.5451661	.1745162	-3.12	0.002	-.8875846	-.2027477
_cons	-7.742457	3.051767	-2.54	0.011	-13.73033	-1.754579

The estimate says that a women with about eight more years of education has about one fewer child (gotten from  $.128(8) = 1.024$ ), other factors fixed. The coefficient is very statistically significant. Also, there has been a notable secular decline in fertility over this period: on average, with other factors held fixed, a women in 1984 had about half a child less (.545) than a similar woman in 1972, the base year. The effect is also statistically significant with  $p$ -value = .002.

#### b. Estimating the reduced form for *educ* gives

```
. reg educ age agesq black east northcen west farm othrural town smcity
y74-y84 meduc feduc
```

Source	SS	df	MS	Number of obs =	1129
Model	2256.26171	18	125.347873	F( 18, 1110) =	24.82
Residual	5606.85432	1110	5.05122011	Prob > F =	0.0000
Total	7863.11603	1128	6.97084755	R-squared =	0.2869
				Adj R-squared =	0.2754
				Root MSE =	2.2475

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
age	-.2243687	.2000013	-1.12	0.262	-.616792	.1680546
agesq	.0025664	.0022605	1.14	0.256	-.001869	.0070018
black	.3667819	.2522869	1.45	0.146	-.1282311	.861795
east	.2488042	.1920135	1.30	0.195	-.1279462	.6255546
northcen	.0913945	.1757744	0.52	0.603	-.2534931	.4362821
west	.1010676	.2422408	0.42	0.677	-.3742339	.5763691
farm	-.3792615	.2143864	-1.77	0.077	-.7999099	.0413869
othrural	-.560814	.2551196	-2.20	0.028	-1.061385	-.060243
town	.0616337	.1807832	0.34	0.733	-.2930816	.416349
smcity	.0806634	.2317387	0.35	0.728	-.3740319	.5353588
y74	.0060993	.249827	0.02	0.981	-.4840872	.4962858
y76	.1239104	.2587922	0.48	0.632	-.3838667	.6316874
y78	.2077861	.2627738	0.79	0.429	-.3078033	.7233755
y80	.3828911	.2642433	1.45	0.148	-.1355816	.9013638
y82	.5820401	.2492372	2.34	0.020	.0930108	1.071069

y84	.4250429	.2529006	1.68	0.093	-.0711741	.92126
meduc	.1723015	.0221964	7.76	0.000	.1287499	.2158531
feduc	.2074188	.0254604	8.15	0.000	.1574629	.2573747
_cons	13.63334	4.396773	3.10	0.002	5.006421	22.26027

```
. test meduc feduc
```

```
( 1) meduc = 0
```

```
( 2) feduc = 0
```

```
F( 2, 1110) = 155.79
Prob > F = 0.0000
```

The joint  $F$  test shows that *educ* is significantly partially correlated with *meduc* and *feduc*; the  $t$  statistics also show this clearly. If we make the test robust to heteroskedasticity of unknown form, the  $F$  statistic drops to 131.37 but the  $p$ -value is still zero to four decimal places.

To test the null that *educ* is exogenous, we need to reduced form residuals and then include them in the OLS regression. I suppress the output here:

```
. predict v2hat, resid
```

```
. reg kids educ age agesq black east northcen west farm othrural town smcity
y74-y84 v2hat
```

Source	SS	df	MS	Number of obs =	1129
Model	400.802376	18	22.2667987	F( 18, 1110) =	9.21
Residual	2684.70692	1110	2.41865489	Prob > F =	0.0000
Total	3085.5093	1128	2.73538059	R-squared =	0.1299
				Adj R-squared =	0.1158
				Root MSE =	1.5552

kids	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
educ	-.1527395	.0392012	-3.90	0.000	-.2296562	-.0758227
age	.5235536	.1389568	3.77	0.000	.2509059	.7962013
agesq	-.005716	.0015697	-3.64	0.000	-.0087959	-.0026362
black	1.072952	.173618	6.18	0.000	.7322958	1.413609
east	.2285554	.1337787	1.71	0.088	-.0339322	.491043
northcen	.3744188	.1219925	3.07	0.002	.1350569	.6137807
west	.2076398	.1675628	1.24	0.216	-.1211357	.5364153
farm	-.0770015	.1512869	-0.51	0.611	-.373842	.2198389
othrural	-.1952451	.1814491	-1.08	0.282	-.5512671	.1607769
town	.08181	.1246122	0.66	0.512	-.162692	.3263119
smcity	.2124996	.160335	1.33	0.185	-.1020943	.5270936
y74	.2721292	.172847	1.57	0.116	-.0670145	.6112729
y76	-.0945483	.1791319	-0.53	0.598	-.4460236	.2569269

y78	-.0572543	.1824512	-0.31	0.754	-.4152424	.3007337
y80	-.053248	.1846139	-0.29	0.773	-.4154795	.3089836
y82	-.4962149	.1764897	-2.81	0.005	-.842506	-.1499238
y84	-.5213604	.1778207	-2.93	0.003	-.8702631	-.1724578
v2hat	.0311374	.0443634	0.70	0.483	-.0559081	.1181829
_cons	-7.241244	3.134883	-2.31	0.021	-13.39221	-1.09028

The  $t$  statistic on *v2hat* is .702, so there is little evidence that *educ* is endogenous in the equation. Still, we can see if 2SLS produces very different estimates:

```
. ivreg kids age agesq black east northcen west farm othrural town smcity
    y74-y84 (educ = meduc feduc)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs =	1129
Model	395.36632	17	23.2568424	F( 17, 1111) =	7.72
Residual	2690.14298	1111	2.42137082	Prob > F =	0.0000
Total	3085.5093	1128	2.73538059	R-squared =	0.1281
				Adj R-squared =	0.1148
				Root MSE =	1.5561

kids	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
educ	-.1527395	.0392232	-3.89	0.000	-.2296993	-.0757796
age	.5235536	.1390348	3.77	0.000	.2507532	.796354
agesq	-.005716	.0015705	-3.64	0.000	-.0087976	-.0026345
black	1.072952	.1737155	6.18	0.000	.732105	1.4138
east	.2285554	.1338537	1.71	0.088	-.0340792	.4911901
northcen	.3744188	.122061	3.07	0.002	.1349228	.6139148
west	.2076398	.1676568	1.24	0.216	-.1213199	.5365995
farm	-.0770015	.1513718	-0.51	0.611	-.3740083	.2200053
othrural	-.1952451	.181551	-1.08	0.282	-.5514666	.1609764
town	.08181	.1246821	0.66	0.512	-.162829	.3264489
smcity	.2124996	.160425	1.32	0.186	-.1022706	.5272698
y74	.2721292	.172944	1.57	0.116	-.0672045	.6114629
y76	-.0945483	.1792324	-0.53	0.598	-.4462205	.2571239
y78	-.0572543	.1825536	-0.31	0.754	-.415443	.3009343
y80	-.053248	.1847175	-0.29	0.773	-.4156825	.3091865
y82	-.4962149	.1765888	-2.81	0.005	-.8427	-.1497297
y84	-.5213604	.1779205	-2.93	0.003	-.8704586	-.1722623
_cons	-7.241244	3.136642	-2.31	0.021	-13.39565	-1.086834

Instrumented: educ

Instruments: age agesq black east northcen west farm othrural town smcity  
y74 y76 y78 y80 y82 y84 meduc feduc

The estimated coefficient on *educ* is larger in magnitude than before, but the test for endogeneity shows that we can reasonably attribute the difference between OLS and 2SLS to sampling error.

c. Since there is little evidence that *educ* is endogenous, we could just use OLS. I did it both ways. First, I just added interactions  $y74 \cdot educ, y76 \cdot educ, \dots, y84 \cdot educ$  to the model in part a and used OLS. Some of the interactions, particularly in the last two years, are marginally significant and negative, showing that the effect of education has become stronger over time. But the joint  $F$  test for the interaction terms yields  $p - value = .180$ , and so we do not reject the model without the interactions. Still, the possibility that the link between fertility and education has become stronger over time is deserves attention, especially using more recent data.

To estimate the full model by 2SLS, I obtained instruments by interacting all year dummies with both *meduc* and *feduc*. The Stata command is then

```
. ivreg kids age agesq black east northcen west farm othrural town smcity y74
      (educ y74educ-y84educ = meduc feduc y74meduc-y84feduc )

. test y74educ y76educ y78educ y80educ y82educ y84educ
```

Qualitatively, the results are similar to the OLS estimates. The  $p - value$  for the joint  $F$  test on the interactions is .205 – again, this has asymptotic justification under Assumption 2SLS.3, the homoskedasticity assumption – so again there is no strong evidence favoring including of the interactions of year dummies and education.

#### 6.9. a. The Stata results are

```
. use injury
. reg ldurat afchnge highearn afhigh male married head-construc if ky
```

Source	SS	df	MS	Number of obs = 5349		
Model	358.441793	14	25.6029852	F( 14, 5334) = 16.37		
Residual	8341.41206	5334	1.56381928	Prob > F = 0.0000		
Total	8699.85385	5348	1.62674904	R-squared = 0.0412		
				Adj R-squared = 0.0387		
				Root MSE = 1.2505		

ldurat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
afchnge	.0106274	.0449167	0.24	0.813	-.0774276	.0986824

highearn	.1757598	.0517462	3.40	0.001	.0743161	.2772035
afhigh	.2308768	.0695248	3.32	0.001	.0945798	.3671738
male	-.0979407	.0445498	-2.20	0.028	-.1852766	-.0106049
married	.1220995	.0391228	3.12	0.002	.0454027	.1987962
head	-.5139003	.1292776	-3.98	0.000	-.7673372	-.2604634
neck	.2699126	.1614899	1.67	0.095	-.0466737	.5864988
upextr	-.178539	.1011794	-1.76	0.078	-.376892	.0198141
trunk	.1264514	.1090163	1.16	0.246	-.0872651	.340168
lowback	-.0085967	.1015267	-0.08	0.933	-.2076305	.1904371
lowextr	-.1202911	.1023262	-1.18	0.240	-.3208922	.0803101
occdis	.2727118	.210769	1.29	0.196	-.1404816	.6859052
manuf	-.1606709	.0409038	-3.93	0.000	-.2408591	-.0804827
construc	.1101967	.0518063	2.13	0.033	.0086352	.2117581
_cons	1.245922	.1061677	11.74	0.000	1.03779	1.454054

The estimated coefficient on the interaction term is actually higher now – .231 – than in equation (6.54), and it has a large  $t$  statistic (3.32 compare with 2.78). Adding the other explanatory variables only slightly increased the standard error on the interaction term.

b. The small  $R$ -squared, on the order of 4.1%, or 3.9% if we used the adjusted  $R$ -squared, means that we do not explain much of the variation in time on workers compensation using the variables included in the regression. This is often the case in the social sciences: it is very difficult to include the multitude of factors that can affect something like *durat*. The low  $R$ -squared means that making predictions of  $\log(\text{durat})$  would be very difficult given the factors we have included in the regression: the variation in the unobservables pretty much swamps the explained variation. However, the low  $R$ -squared does not mean we have a biased or inconsistent estimator of the effect of the policy change. Provided the Kentucky policy change provides a good natural experiment, the OLS estimator is consistent. With over 5,000 observations, we can get a reasonably precise estimate of the effect, although the 95% confidence interval is pretty wide.

c. Using the data for Michigan to estimate the basic regression gives

```
. reg ldurat afchnge highearn afhigh if mi
```

Source	SS	df	MS	Number of obs =	1524
Model	34.3850177	3	11.4616726	F( 3, 1520) =	6.05
				Prob > F =	0.0004

Residual		2879.96981	1520	1.89471698		R-squared	=	0.0118
						Adj R-squared	=	0.0098
Total		2914.35483	1523	1.91356194		Root MSE	=	1.3765
ldurat		Coef.	Std. Err.	t	P> t	[95% Conf. Interval		
afchnge		.0973808	.0847879	1.15	0.251	-.0689329	.2636945	
highearn		.1691388	.1055676	1.60	0.109	-.0379348	.3762124	
afhigh		.1919906	.1541699	1.25	0.213	-.1104176	.4943988	
_cons		1.412737	.0567172	24.91	0.000	1.301485	1.523989	

The coefficient on the interaction term, .192, is remarkably similar to that for Kentucky.

Unfortunately, because of the many fewer observations, the  $t$  statistic is insignificant at the 10% level against a one-sided alternative. Asymptotic theory roughly predicts that the standard error for Michigan will be about  $(5,626/1,524)^{1/2} \approx 1.92$  larger than that for Kentucky (assuming the same error variance and same fraction of observations in the different groups). In fact, the ratio of standard errors is about 2.23. The difference precision in the KY and MI cases shows the importance of a large sample size for this kind of policy analysis.

**6.10.** a. As suggested by the hint, we can write  $\sqrt{N}(\hat{\beta} - \beta) = N^{-1/2} \sum_{i=1}^N \mathbf{A}^{-1} \mathbf{z}_i' u_i$ , where  $\mathbf{A} \equiv E(\mathbf{z}_i' \mathbf{z}_i)$ , plus a term we can ignore by the asymptotic equivalence lemma. Further,  $\sqrt{N}(\bar{\mathbf{x}} - \mu) = N^{-1/2} \sum_{i=1}^N (\mathbf{x}_i - \mu)$ . When we stack these two representations, we see that the asymptotic covariance between  $\sqrt{N}(\hat{\beta} - \beta)$  and  $\sqrt{N}(\bar{\mathbf{x}} - \mu)$  is  $E[\mathbf{A}^{-1} \mathbf{z}_i' u_i (\mathbf{x}_i - \mu)] = \mathbf{A}^{-1} E[u_i \mathbf{z}_i' (\mathbf{x}_i - \mu)]$ . Because  $E(u_i | \mathbf{x}_i) = 0$ , the standard iterated expectations argument shows that  $E[u_i \mathbf{z}_i' (\mathbf{x}_i - \mu)] = 0$  because  $\mathbf{z}_i$  is a function of  $\mathbf{x}_i$ . This completes the proof.

b. While the delta method leads to the same place, it is not needed because of linearity of  $\hat{\beta}$  in the data. We can write  $\hat{\alpha}_1 = \hat{\beta}_1 + \hat{\beta}_3 \bar{x}_2 = \hat{\beta}_1 + \hat{\beta}_3 \mu_2 + \hat{\beta}_3 (\bar{x}_2 - \mu_2) \equiv \tilde{\alpha}_1 + \hat{\beta}_3 (\bar{x}_2 - \mu_2)$ , and so  $\sqrt{N}(\hat{\alpha}_1 - \alpha_1) = \sqrt{N}(\tilde{\alpha}_1 - \alpha_1) + \hat{\beta}_3 [\sqrt{N}(\bar{x}_2 - \mu_2)]$ . Now  $\hat{\beta}_3 [\sqrt{N}(\bar{x}_2 - \mu_2)] = \beta_3 [\sqrt{N}(\bar{x}_2 - \mu_2)] + o_p(1)$  because  $\hat{\beta}_3 - \beta_3 = o_p(1)$  and



$\sqrt{N}(\bar{x}_2 - \mu_2) = O_p(1)$ . So we have

$$\sqrt{N}(\hat{\alpha}_1 - \alpha_1) = \sqrt{N}(\tilde{\alpha}_1 - \alpha_1) + \beta_3[\sqrt{N}(\bar{x}_2 - \mu_2)] + o_p(1).$$

By part a, we know that  $\sqrt{N}(\hat{\beta} - \beta)$  and  $\sqrt{N}(\bar{x}_2 - \mu_2)$  are asymptotically jointly normal and asymptotically independent (uncorrelated). Because  $\sqrt{N}(\tilde{\alpha}_1 - \alpha_1)$  is just a deterministic linear combination of  $\sqrt{N}(\hat{\beta} - \beta)$  it follows that  $\sqrt{N}(\tilde{\alpha}_1 - \alpha_1)$  and  $\sqrt{N}(\bar{x}_2 - \mu_2)$  are asymptotically uncorrelated. Therefore,.

$$\begin{aligned} \text{Avar}[\sqrt{N}(\hat{\alpha}_1 - \alpha_1)] &= \text{Avar}[\sqrt{N}(\tilde{\alpha}_1 - \alpha_1)] + \beta_3^2 \text{Avar}[\sqrt{N}(\bar{x}_2 - \mu_2)] \\ &= \text{Avar}[\sqrt{N}(\tilde{\alpha}_1 - \alpha_1)] + \beta_3^2 \sigma_2^2, \end{aligned}$$

where  $\sigma_2^2 = \text{Var}(x_2)$ . Therefore, by the convention introduced in Section 3.5, we write

$$\text{Avar}(\hat{\alpha}_1) = \text{Avar}(\tilde{\alpha}_1) + \beta_3^2(\sigma_2^2/N),$$

which is what we wanted to show.

c. As stated in the hint, the standard error we get from the regression in Problem 4.8d is really  $\text{se}(\tilde{\alpha}_1)$ , as it does not account for the sampling variation in  $\bar{x}_2$ . So

$$\text{se}(\hat{\alpha}_1) = \{[\text{se}(\tilde{\alpha}_1)]^2 + \hat{\beta}_3^2(\hat{\sigma}_2^2/N)\}^{1/2} = \{[\text{se}(\tilde{\alpha}_1)]^2 + \hat{\beta}_3^2[\text{se}(\bar{x}_2)]^2\}^{1/2}$$

since  $\text{se}(\bar{x}_2) = \sigma^2/\sqrt{N}$ .

d. The standard error reported for the education variable in Problem 4.8d,  $\text{se}(\hat{\alpha}_1)$ , is about .00698, the coefficient on the interaction term ( $\hat{\beta}_3$ ) is about .00455, and the sample standard deviation of *exper* is about 4.375. Plugging these numbers into the formula from part c gives  $\text{se}(\hat{\alpha}_1) = [(.00698)^2 + (.00455)^2(4.375)^2/935]^{1/2} \approx .00701$ . For practical purposes, this is not much bigger than .00698: the effect of accounting for estimation of the population mean of *exper* is very modest.

**6.11.** The following is Stata output for answering the first three parts:

```
. use cps78_85
```

```
. reg lwage y85 educ y85educ exper expersq union female y85fem
```

Source	SS	df	MS	Number of obs = 1084		
Model	135.992074	8	16.9990092	F( 8, 1075) = 99.80		
Residual	183.099094	1075	.170324738	Prob > F = 0.0000		
Total	319.091167	1083	.29463635	R-squared = 0.4262		
				Adj R-squared = 0.4219		
				Root MSE = .4127		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
y85	.1178062	.1237817	0.95	0.341	-.125075	.3606874
educ	.0747209	.0066764	11.19	0.000	.0616206	.0878212
y85educ	.0184605	.0093542	1.97	0.049	.000106	.036815
exper	.0295843	.0035673	8.29	0.000	.0225846	.036584
expersq	-.0003994	.0000775	-5.15	0.000	-.0005516	-.0002473
union	.2021319	.0302945	6.67	0.000	.1426888	.2615749
female	-.3167086	.0366215	-8.65	0.000	-.3885663	-.244851
y85fem	.085052	.051309	1.66	0.098	-.0156251	.185729
_cons	.4589329	.0934485	4.91	0.000	.2755707	.642295

a. The return to another year of education increased by about .0185, or 1.85 percentage points, between 1978 and 1985. The  $t$  statistic on  $y85educ$  is 1.97, which is marginally significant at the 5% level against a two-sided alternative.

b. The coefficient on  $y85fem$  is positive and shows that the estimated gender gap declined by about 8.5 percentage points. It is still very large, with the gender difference in  $lwage$  in 1985 estimated at about  $-.232$ . The  $t$  statistic on  $y85fem$  is only significant at about the 10% level against a two-sided alternative. Still, this is suggestive of some closing of wage differentials between women and men at given levels of education and workforce experience.

c. Only the coefficient on  $y85$  changes if wages are measured in 1978 dollars. In fact, you can check that when 1978 wages are used, the coefficient on  $y85$  becomes about  $-.383 = .118 - \log(1.65) \approx .118 - .501$ .

d. To answer this question, I just took the squared OLS residuals and regressed those on the year dummy,  $y85$ . The coefficient is about .042 with a standard error of about .022, which

gives a  $t$  statistic of about 1.91. So there is some evidence that the variance of the unexplained part of log wages (or even log real wages) has increased over time.

e. As the equation is written in the problem, the coefficient  $\delta_0$  is the growth in nominal wages for a male with no years of education! For a male with 12 years of education, we want  $\theta_0 \equiv \delta_0 + 12\delta_1$ .

Many packages have simple commands that deliver standard errors and tests for linear combinations. But a general way to obtain the standard error for  $\hat{\theta}_0 \equiv \hat{\delta}_0 + 12\hat{\delta}_1$  is to replace  $y85 \cdot educ$  with  $y85 \cdot (educ - 12)$  and reestimate the equation. Simple algebra shows that, in the new equation,  $\theta_0$  is the coefficient on  $educ$ . In Stata we have

```
. gen y85educ_12 = y85*(educ - 12)
```

```
. reg lwage y85 educ y85educ_12 exper expersq union female y85fem
```

Source	SS	df	MS	Number of obs =	1084
Model	135.992074	8	16.9990092	F( 8, 1075) =	99.80
Residual	183.099094	1075	.170324738	Prob > F =	0.0000
Total	319.091167	1083	.29463635	R-squared =	0.4262
				Adj R-squared =	0.4219
				Root MSE =	.4127

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
y85	.3393326	.0340099	9.98	0.000	.2725993	.4060659
educ	.0747209	.0066764	11.19	0.000	.0616206	.0878212
y85educ_12	.0184605	.0093542	1.97	0.049	.000106	.036815
exper	.0295843	.0035673	8.29	0.000	.0225846	.036584
expersq	-.0003994	.0000775	-5.15	0.000	-.0005516	-.0002473
union	.2021319	.0302945	6.67	0.000	.1426888	.2615749
female	-.3167086	.0366215	-8.65	0.000	-.3885663	-.244851
y85fem	.085052	.051309	1.66	0.098	-.0156251	.185729
_cons	.4589329	.0934485	4.91	0.000	.2755707	.642295

So the growth in nominal wages for a man with  $educ = 12$  is about .339, or 33.9%. [We could use the more accurate estimate, .404, obtained from  $\exp(.339) - 1 = .404$ .] The 95% confidence interval goes from about 27.3 to 40.6.

Stata users can verify that the command

. lincom y85 + 12\*y85educ

after estimation of the original equation delivers the same estimate and inference.

**6.12.** Under the assumptions listed,  $E(\mathbf{x}'u) = 0$ ,  $E(\mathbf{z}'u) = 0$ , and the rank conditions hold for OLS and 2SLS, so we can write

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}) = \mathbf{A}_*^{-1} \left( N^{-1/2} \sum_{i=1}^N \mathbf{x}_i^{*'} u_i \right) + o_p(1),$$

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) = \mathbf{A}^{-1} \left( N^{-1/2} \sum_{i=1}^N \mathbf{x}_i' u_i \right) + o_p(1)$$

where  $\mathbf{A} = E(\mathbf{x}_i' \mathbf{x}_i)$ ,  $\mathbf{A}_* = E(\mathbf{x}_i^{*'} \mathbf{x}_i^*)$ , and  $\mathbf{x}_i^* = \mathbf{z}_i \boldsymbol{\Pi}$ . Further, because of the homoskedasticity assumptions,  $E(u_i^2 \mathbf{x}_i' \mathbf{x}_i) = \sigma^2 \mathbf{A}$ ,  $E(u_i^2 \mathbf{x}_i^{*'} \mathbf{x}_i^*) = \sigma^2 \mathbf{A}_*$ , and  $E(u_i^2 \mathbf{x}_i^{*'} \mathbf{x}_i) = \sigma^2 E(\mathbf{x}_i^{*'} \mathbf{x}_i)$ . But we know from Chapter 5 that  $E(\mathbf{x}_i^{*'} \mathbf{x}_i) = \mathbf{A}_*$ . Next, we can stack the above equations to obtain that OLS and 2SLS, when appropriately centered and scaled, are jointly asymptotically normal with variance-covariance matrix

$$\begin{pmatrix} \mathbf{V}_1 & \mathbf{C} \\ \mathbf{C}' & \mathbf{V}_2 \end{pmatrix},$$

where  $\mathbf{V}_1 = \text{Avar}[\sqrt{N}(\hat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta})]$ ,  $\mathbf{V}_2 = \text{Avar}[\sqrt{N}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})]$ , and

$\mathbf{C} = \mathbf{A}_*^{-1} E(u_i^2 \mathbf{x}_i^{*'} \mathbf{x}_i) \mathbf{A}^{-1} = \sigma^2 \mathbf{A}^{-1}$ . Therefore, we can write the asymptotic variance matrix of both estimators as

$$\sigma^2 \begin{pmatrix} \mathbf{A}_*^{-1} & \mathbf{A}^{-1} \\ \mathbf{A}^{-1} & \mathbf{A}^{-1} \end{pmatrix}.$$

Now, the asymptotic variance of any linear combination is easy to obtain. In particular, the asymptotic variance of  $\sqrt{N}(\hat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}) - \sqrt{N}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})$  is simply

$\sigma^2(\mathbf{A}_*^{-1} + \mathbf{A}^{-1} - \mathbf{A}^{-1} - \mathbf{A}^{-1}) = \sigma^2 \mathbf{A}_*^{-1} - \sigma^2 \mathbf{A}^{-1}$ , which is the difference in the asymptotic

variances, as we wanted to show.

**6.13.** This is a simple application of the law of iterated expectations. The statement of the problem should add the requirement  $\rho_1 \neq 0$ . By the LIE,

$$E(u_1|\mathbf{z}) = E[E(u_1|\mathbf{z}, v_2)|\mathbf{z}] = E(\rho_1 v_2|\mathbf{z}) = \rho_1 E(v_2|\mathbf{z})$$

and so if  $E(u_1|\mathbf{z}) = 0$  then  $E(v_2|\mathbf{z}) = 0$ , too.

**6.14.** a. First,  $y_2$  is a function of  $(\mathbf{z}, v_2)$ , and so, from the structural equation,

$$E(y_1|\mathbf{z}, v_2) = \mathbf{z}_1\boldsymbol{\delta}_1 + \mathbf{g}(y_2)\boldsymbol{\alpha}_1 + E(u_1|\mathbf{z}, v_2) = \mathbf{z}_1\boldsymbol{\delta}_1 + \mathbf{g}(y_2)\boldsymbol{\alpha}_1 + E(u_1|v_2),$$

where

$$E(u_1|\mathbf{z}, v_2) = E(u_1|v_2)$$

follows because  $(u_1, v_2)$  is independent of  $\mathbf{z}$ . (Note that, in general, it is not enough to assume that  $u_1$  and  $v_2$  are separately independent of  $\mathbf{z}$ ; joint independence is needed.)

b. If  $E(u_1|v_2) = \rho_1 v_2$  then, under the previous assumptions,

$$E(y_1|\mathbf{z}, v_2) = \mathbf{z}_1\boldsymbol{\delta}_1 + \mathbf{g}(y_2)\boldsymbol{\alpha}_1 + \rho_1 v_2.$$

Therefore, in the first step, we would run OLS of  $y_{i2}$  on  $\mathbf{z}_i, i = 1, \dots, N$ , and obtain the OLS residuals,  $\hat{v}_{i2}$ . In the second step, we would regress  $y_{i1}$  on  $\mathbf{z}_{i1}, \mathbf{g}(y_{i2}), \hat{v}_{i2}, i = 1, \dots, N$ . By the usual two-step estimation results, all coefficients are  $\sqrt{N}$ -consistent and asymptotically normal for the corresponding population parameter. The interesting thing about this method is that, if  $G_1 > 1$  we have more than one endogenous explanatory variable –  $g_1(y_1), \dots, g_{G_1}(y_2)$  – but adding a single regressor,  $\hat{v}_{i2}$ , cleans up the endogeneity. This occurs because all endogenous regressors are a function of  $y_2$ , and we have assumed  $y_2$  is an additive function of  $\mathbf{z}$  and an independent error, which pretty much restricts  $y_2$  to be continuous. (We can easily replace the

linear function  $\mathbf{z}\pi_2$  with known nonlinear functions of  $\mathbf{z}$ .)

As specific examples, the second stage regression might be

$$y_{i1} \text{ or } \mathbf{z}_{i1}, y_{i2}, y_{i2}^2, y_{i2}^3, \hat{v}_{i2}, i = 1, \dots, N$$

or

$$y_{i1} \text{ or } \mathbf{z}_{i1}, 1[a_1 < y_{i2} \leq a_2], \dots, 1[a_{m-1} < y_{i2} \leq a_m], 1[y_{i2} > a_M], \hat{v}_{i2}, i = 1, \dots, N.$$

In the latter case, dummies for whether  $y_{i2}$  falls into one of the intervals

$(-\infty, a_1], (a_1, a_2], \dots, (a_{M-1}, a_M], (a_M, \infty)$  appear in the structural model.

c. If  $\rho_1 = 0$ , no adjustment is needed to the asymptotic variance, so we can use the usual  $t$  statistic on  $\hat{v}_{i2}$  as a test of endogeneity of  $y_2$ , where the null is exogeneity:  $H_0 : \rho_1 = 0$ .

Actually, nothing guarantees that  $\text{Var}(y_1|\mathbf{z}, v_2)$  does not depend on  $v_2$  – and, under weaker assumptions, it could also depend on  $\mathbf{z}$  – so there is a good case for making the test robust to heteroskedasticity.

d. The estimating equation becomes

$$E(y_1|\mathbf{z}, v_2) = \mathbf{z}_1\delta_1 + \mathbf{g}(y_2)\alpha_1 + \rho_1 v_2 + \xi_1(v_2^2 - \tau_2^2)$$

and now, to implement a two-step control function procedure, we obtain  $\hat{\tau}_2^2$ , the usual OLS error variance estimate, along with  $\hat{\pi}_2$ . The residuals are constructed as before,

$\hat{v}_{i2} = y_{i2} - \mathbf{z}_i\hat{\pi}_2$ . The second-step regression is now

$$y_{i1} \text{ on } \mathbf{z}_{i1}, \mathbf{g}(y_{i2}), \hat{v}_{i2}, (\hat{v}_{i2}^2 - \hat{\tau}_2^2), i = 1, \dots, N$$

Now we can use a heteroskedasticity-robust Wald test of joint significance of  $\hat{v}_{i2}$  and

$(\hat{v}_{i2}^2 - \hat{\tau}_2^2)$ . Under the null  $H_0 : \rho_1 = 0, \xi_1 = 0$ , we do not have to adjust the statistic for the first-step estimation.

e. We would use traditional 2SLS, where we need at least one IV for each  $\mathbf{g}_j(y_2)$ . Methods

for coming up with such IVs are discussed in Section 9.5. Briefly, they will be nonlinear functions of  $\mathbf{z}$ , which is why  $E(u_1|\mathbf{z}) = 0$  should be assumed. Generally, we add enough nonlinear functions, say  $\mathbf{h}(\mathbf{z})$ , to the original instrument list  $\mathbf{z}$ . So, do 2SLS of  $y_1$  on  $\mathbf{z}_1, \mathbf{g}_2$  using IVs  $[\mathbf{z}, \mathbf{h}(\mathbf{z})]$ . 2SLS will be more robust than the method described in part b because the reduced form for  $y_2$  is not restricted in any way, and we need not assume  $u_1$  is independent of  $\mathbf{z}$ .

**6.15.** a. Because  $y_2 = \mathbf{z}\pi_2 + v_2$ , we can find  $E(y_1|\mathbf{z}, v_2)$  or  $E(y_1|\mathbf{z}, y_2)$ ; they are the same.

Now

$$\begin{aligned} E(y_1|\mathbf{z}, v_2) &= \mathbf{z}_1\delta_1 + \mathbf{g}(\mathbf{z}_1, y_2)\alpha_1 + \mathbf{g}(\mathbf{z}_1, y_2)E(v_1|\mathbf{z}, v_2) + E(u_1|\mathbf{z}, v_2) \\ &= \mathbf{z}_1\delta_1 + \mathbf{g}(\mathbf{z}_1, y_2)\alpha_1 + \mathbf{g}(\mathbf{z}_1, y_2)v_2\theta_1 + \rho_1 v_2 \end{aligned}$$

b. The first step is to regress  $y_{i2}$  on  $\mathbf{z}_i$  and get the residuals,  $\hat{v}_{i2}$ . Second, run the regression

$$y_{i1} \text{ on } \mathbf{z}_{i1}, \mathbf{g}(\mathbf{z}_{i1}, y_{i2}), \mathbf{g}(\mathbf{z}_{i1}, y_{i2})\hat{v}_{i2}, \hat{v}_{i2}$$

which means that  $\hat{v}_{i2}$  appears by itself and interacted with all elements of  $\mathbf{g}(\mathbf{z}_{i1}, y_{i2})$ .

c. The null is  $H_0 : \theta_1 = \mathbf{0}, \rho_1 = 0$ , which means we can compute a heteroskedasticity-robust Wald test of joint significance of  $\mathbf{g}(\mathbf{z}_{i1}, y_{i2})\hat{v}_{i2}$  and  $\hat{v}_{i2}$ .

d. For the specific model give, the second-step regression is

$$y_{i1} \text{ on } \mathbf{z}_{i1}, y_{i2}, y_{i2}^2, \mathbf{z}_{i1}y_{i2}, y_{i2}\hat{v}_{i2}, \hat{v}_{i2}, i = 1, \dots, N.$$

In other words,  $\hat{v}_{i2}$  appears by itself and interacted with  $y_{i2}$ , as in Garen (1984).

## Solutions to Chapter 7 Problems

7.1. Write (with probability approaching one)

$$\hat{\beta} = \beta + \left( N^{-1} \sum_{i=1}^N \mathbf{X}_i' \mathbf{X}_i \right)^{-1} \left( N^{-1} \sum_{i=1}^N \mathbf{X}_i' \mathbf{u}_i \right).$$

From Assumption SOLS. 2, the weak law of large numbers, and Slutsky's Theorem,

$$\text{plim} \left( N^{-1} \sum_{i=1}^N \mathbf{X}_i' \mathbf{X}_i \right)^{-1} = \mathbf{A}^{-1}.$$

Further, under SOLS.1, the WLLN implies that  $\text{plim} \left( N^{-1} \sum_{i=1}^N \mathbf{X}_i' \mathbf{u}_i \right) = \mathbf{0}$ . Thus,

$$\text{plim}(\hat{\beta}) = \beta + \text{plim} \left( N^{-1} \sum_{i=1}^N \mathbf{X}_i' \mathbf{X}_i \right)^{-1} \cdot \text{plim} \left( N^{-1} \sum_{i=1}^N \mathbf{X}_i' \mathbf{u}_i \right) = \beta + \mathbf{A}^{-1} \cdot \mathbf{0} = \beta.$$

7.2. a. Under SOLS. 1 and SOLS.2, Theorem 7.2 implies that  $\text{Avar}(\hat{\beta}_{OLS}) = \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1} / N$ ,

where  $\mathbf{A} = E(\mathbf{X}_i' \mathbf{X}_i)$  and  $\mathbf{B} = E(\mathbf{X}_i \mathbf{u}_i \mathbf{u}_i' \mathbf{X}_i')$ . But we have assumed that

$E(\mathbf{X}_i \mathbf{u}_i \mathbf{u}_i' \mathbf{X}_i') = E(\mathbf{X}_i' \mathbf{\Omega} \mathbf{X}_i)$ , which proves the assertion. Effectively, this is what we can expect for the asymptotic variance of OLS under the system version of homoskedasticity. [Note that Assumption SGLS. 3 and  $E(\mathbf{X}_i \mathbf{u}_i \mathbf{u}_i' \mathbf{X}_i') = E(\mathbf{X}_i' \mathbf{\Omega} \mathbf{X}_i)$  are not the same, but both are implied by condition (7.53). There are other cases where they reduce to the same assumption, such as in a SUR model when  $\mathbf{\Omega}$  is diagonal.]

b. The estimator in (7.28) is always valid. An estimator that uses the structure of  $\text{Avar}(\hat{\beta}_{SOLS})$  obtained in part a is obtained as follows. Let  $\hat{\mathbf{\Omega}} = N^{-1} \sum_{i=1}^N \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i'$ , where the  $\hat{\mathbf{u}}_i$  are the  $G \times 1$  system OLS residuals. Then

$$\widehat{\text{Avar}}(\hat{\beta}_{SOLS}) = \left( \sum_{i=1}^N \mathbf{X}_i' \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}_i' \hat{\mathbf{\Omega}} \mathbf{X}_i \right) \left( \sum_{i=1}^N \mathbf{X}_i' \mathbf{X}_i \right)^{-1}$$



is a valid estimator provided the homoskedasticity assumption holds.

c. Using the hint and dropping the division by  $N$  on the right hand side, we have

$$[\text{Avar}(\hat{\beta}_{FGLS})]^{-1} - [\text{Avar}(\hat{\beta}_{SOLS})]^{-1} = E(\mathbf{X}_i' \boldsymbol{\Omega}^{-1} \mathbf{X}_i) - E(\mathbf{X}_i' \mathbf{X}_i) [E(\mathbf{X}_i' \boldsymbol{\Omega} \mathbf{X}_i)]^{-1} E(\mathbf{X}_i' \mathbf{X}_i).$$

Define  $\mathbf{Z}_i \equiv \boldsymbol{\Omega}^{-1/2} \mathbf{X}_i$  and  $\mathbf{W}_i \equiv \boldsymbol{\Omega}^{1/2} \mathbf{X}_i$ . Then the difference can be written as

$$E(\mathbf{Z}_i' \mathbf{Z}_i) - E(\mathbf{Z}_i' \mathbf{W}_i) [E(\mathbf{W}_i' \mathbf{W}_i)]^{-1} E(\mathbf{W}_i' \mathbf{Z}_i).$$

Now, define  $\mathbf{R}_i \equiv \mathbf{Z}_i - \mathbf{W}_i \boldsymbol{\Pi}$ , where  $\boldsymbol{\Pi} \equiv [E(\mathbf{W}_i' \mathbf{W}_i)]^{-1} E(\mathbf{W}_i' \mathbf{Z}_i)$ ;  $\mathbf{R}_i$  is the  $G \times K$  matrix of population residuals from the linear projection of  $\mathbf{Z}_i$  on  $\mathbf{W}_i$ . Straightforward multiplication shows that

$$E(\mathbf{Z}_i' \mathbf{Z}_i) - E(\mathbf{Z}_i' \mathbf{W}_i) [E(\mathbf{W}_i' \mathbf{W}_i)]^{-1} E(\mathbf{W}_i' \mathbf{Z}_i) = E(\mathbf{R}_i' \mathbf{R}_i),$$

which is necessarily positive semi-definite. We have shown that if (7.53) holds along with SGLS.1 and the rank conditions for SGLS and SOLS, then FGLS is more efficient than OLS.

d. If  $\boldsymbol{\Omega} = \sigma^2 \mathbf{I}_G$ ,

$$\text{Avar}[\sqrt{N}(\hat{\beta}_{SOLS} - \beta)] = [E(\mathbf{X}_i' \mathbf{X}_i)]^{-1} E(\mathbf{X}_i' \boldsymbol{\Omega} \mathbf{X}_i) [E(\mathbf{X}_i' \mathbf{X}_i)]^{-1} = \sigma^2 [E(\mathbf{X}_i' \mathbf{X}_i)]^{-1} \text{ and}$$

$$\text{Avar}[\sqrt{N}(\hat{\beta}_{SOLS} - \beta)] = [E(\mathbf{X}_i' \boldsymbol{\Omega}^{-1} \mathbf{X}_i)]^{-1} = [E(\mathbf{X}_i' (\sigma^2 \mathbf{I}_G)^{-1} \mathbf{X}_i)]^{-1} = \sigma^2 [E(\mathbf{X}_i' \mathbf{X}_i)]^{-1}.$$

e. This statement is true provided we consider only asymptotic efficiency under the assumption that SGLS.1 holds. In other words, under SGLS.1, the standard rank conditions, and  $E(\mathbf{u}_i \mathbf{u}_i' | \mathbf{X}_i) = \boldsymbol{\Omega}$ , there is nothing to lose asymptotically by using FGLS. Of course, SOLS is more robust in that it only requires SOLS.1 for consistency (and asymptotic normality). Small sample properties are another issue because it is difficult to characterize the exact properties of FGLS under general conditions.

**7.3.** a. Since OLS equation-by-equation is the same as GLS when  $\boldsymbol{\Omega}$  is diagonal, it suffices to show that the GLS estimators for different equations are asymptotically uncorrelated. This

follows if the asymptotic variance matrix is block diagonal (see Section 3.5), where the blocking is by the parameter vector for each equation. To establish block diagonality, we use the result from Theorem 7.4: under SGLS. 1, SGLS.2, and SGLS.3,

$$\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = [\text{E}(\mathbf{X}_i' \boldsymbol{\Omega}^{-1} \mathbf{X}_i)]^{-1}.$$

Now, we can use the special form of  $\mathbf{X}_i$  for SUR (see Example 7.1), the fact that  $\boldsymbol{\Omega}^{-1}$  is diagonal, and SGLS.3. In the SUR model with diagonal  $\boldsymbol{\Omega}$ , SGLS.3 implies that

$$\text{E}(u_{ig}^2 \mathbf{x}_{ig}' \mathbf{x}_{ig}) = \sigma_g^2 \text{E}(\mathbf{x}_{ig}' \mathbf{x}_{ig}) \text{ for all } g = 1, \dots, G, \text{ and}$$

$$\text{E}(u_{ig} u_{ih} \mathbf{x}_{ig}' \mathbf{x}_{ih}) = \text{E}(u_{ig} u_{ih}) \text{E}(\mathbf{x}_{ig}' \mathbf{x}_{ih}) = \mathbf{0}, \text{ all } g \neq h.$$

Therefore, we have

$$\text{E}(\mathbf{X}_i' \boldsymbol{\Omega}^{-1} \mathbf{X}_i) = \begin{pmatrix} \sigma_1^{-2} \text{E}(\mathbf{x}_{i1}' \mathbf{x}_{i1}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_G^{-2} \text{E}(\mathbf{x}_{iG}' \mathbf{x}_{iG}) \end{pmatrix}.$$

When this matrix is inverted, it is also block diagonal. This shows that  $\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$  is block diagonal, and therefore the  $\sqrt{N}(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g)$  are asymptotically uncorrelated.

b. To test any linear hypothesis, we can either construct the Wald Statistic or we can use the weighted sum of squared residuals form of the statistic as in (7.56) or (7.57). For the restricted SSR we must estimate the model with the restriction  $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$  imposed. See Problem 7.6 for one way to impose general linear restrictions.

c. Actually, for the conclusion to hold about asymptotic equivalence, we need to assume SGLS.1 along with SOLS.2 and SGLS.2. When  $\boldsymbol{\Omega}$  is diagonal in a SUR system, system OLS and GLS are the same. Under SGLS.1 and SGLS.2, GLS and FGLS are asymptotically equivalent (regardless of the structure of  $\boldsymbol{\Omega}$ ) whether or not SGLS.3 holds. Now if

$\hat{\beta}_{SOLS} = \hat{\beta}_{GLS}$  and  $\sqrt{N}(\hat{\beta}_{FGLS} - \hat{\beta}_{GLS}) = o_p(1)$ , then  $\sqrt{N}(\hat{\beta}_{SOLS} - \hat{\beta}_{FGLS}) = o_p(1)$ . Thus, when  $\Omega$  is diagonal, OLS and FGLS are asymptotically equivalent under the exogeneity assumption SGLS.1, even if  $\hat{\Omega}$  is estimated in an unrestricted fashion and even if the system homoskedasticity assumption SGLS.3 does not hold.

If only SOLS.1 holds, we cannot conclude  $\sqrt{N}(\hat{\beta}_{FGLS} - \hat{\beta}_{GLS}) = o_p(1)$ , and so  $\sqrt{N}(\hat{\beta}_{SOLS} - \hat{\beta}_{FGLS})$  is not generally  $o_p(1)$ . It is true that FGLS is still consistent under SOLS.1 because its plim is

$$\begin{pmatrix} \sigma_1^{-2} E(\mathbf{x}'_{i1} \mathbf{x}_{i1}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_G^{-2} E(\mathbf{x}'_{iG} \mathbf{x}_{iG}) \end{pmatrix}^{-1} \begin{pmatrix} \sigma_1^{-2} E(\mathbf{x}'_{i1} u_{i1}) \\ \vdots \\ \sigma_G^{-2} E(\mathbf{x}'_{iG} u_{iG}) \end{pmatrix}$$

and  $E(\mathbf{x}'_{ig} u_{ig}) = \mathbf{0}$ ,  $g = 1, \dots, G$ .

7.4. To make the notation align with the text, use  $\check{\beta}$  to denote the SOLS estimator, and let  $\check{\mathbf{u}}_i$  denote the  $G \times 1$  vector of SOLS residuals that are used in obtaining  $\hat{\Omega}$ . Then it suffices to show that

$$N^{-1/2} \sum_{i=1}^N \check{\mathbf{u}}_i \check{\mathbf{u}}_i' = N^{-1/2} \sum_{i=1}^N \mathbf{u}_i \mathbf{u}_i + o_p(1), \quad (7.82)$$

and this follows if, when we sum across  $N$  and divide by  $\sqrt{N}$ , the last three terms in (7.42) are  $o_p(1)$ . Since the third term is the transpose of the second it suffices to consider only the second and fourth terms. Now

$$\begin{aligned} N^{-1/2} \sum_{i=1}^N \text{vec}[\mathbf{u}_i (\check{\beta} - \beta)' \mathbf{X}_i'] &= N^{-1/2} \sum_{i=1}^N (\mathbf{X}_i \otimes \mathbf{u}_i) \cdot (\check{\beta} - \beta) \\ &= \left[ N^{-1} \sum_{i=1}^N (\mathbf{X}_i \otimes \mathbf{u}_i) \right] \sqrt{N} (\check{\beta} - \beta) = o_p(1) \cdot O_p(1) = o_p(1). \end{aligned}$$

Also,

$$\begin{aligned} N^{-1/2} \sum_{i=1}^N \text{vec} \left[ \mathbf{X}_i (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{X}_i' \right] &= \left[ N^{-1} \sum_{i=1}^N (\mathbf{X}_i \otimes \mathbf{X}_i) \right] \text{vec} \left\{ \sqrt{N} (\check{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sqrt{N} (\check{\boldsymbol{\beta}} - \boldsymbol{\beta})' \right\} / \sqrt{N} \\ &= O_p(1) \cdot O_p(1) \cdot N^{-1/2} = o_p(1). \end{aligned}$$

Together, these imply  $N^{-1/2} \sum_{i=1}^N \check{\mathbf{u}}_i \check{\mathbf{u}}_i' = N^{-1/2} \sum_{i=1}^N \mathbf{u}_i \mathbf{u}_i + o_p(1)$  and so

$$N^{-1/2} \sum_{i=1}^N (\check{\mathbf{u}}_i \check{\mathbf{u}}_i' - \boldsymbol{\Omega}) = N^{-1/2} \sum_{i=1}^N (\mathbf{u}_i \mathbf{u}_i - \boldsymbol{\Omega}) + o_p(1).$$

7.5. This is easy with the hint. Note that

$$\left( \hat{\boldsymbol{\Omega}}^{-1} \otimes \left( \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \right) \right)^{-1} = \hat{\boldsymbol{\Omega}} \otimes \left( \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \right)^{-1}.$$

Therefore,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left( \hat{\boldsymbol{\Omega}} \otimes \left( \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \right) (\hat{\boldsymbol{\Omega}}^{-1} \otimes \mathbf{I}_K) \begin{pmatrix} \sum_{i=1}^N \mathbf{x}_i' y_{i1} \\ \vdots \\ \sum_{i=1}^N \mathbf{x}_i' y_{iG} \end{pmatrix} = \left( \mathbf{I}_G \otimes \left( \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \right) \begin{pmatrix} \sum_{i=1}^N \mathbf{x}_i' y_{i1} \\ \vdots \\ \sum_{i=1}^N \mathbf{x}_i' y_{iG} \end{pmatrix} \\ &= \begin{pmatrix} \left( \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \right)^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \left( \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \right)^{-1} & \mathbf{0} & \vdots \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \left( \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^N \mathbf{x}_i' y_{i1} \\ \sum_{i=1}^N \mathbf{x}_i' y_{i2} \\ \vdots \\ \sum_{i=1}^N \mathbf{x}_i' y_{iG} \end{pmatrix} = \begin{pmatrix} \check{\boldsymbol{\beta}}_1 \\ \check{\boldsymbol{\beta}}_2 \\ \vdots \\ \check{\boldsymbol{\beta}}_G \end{pmatrix} \end{aligned}$$

where  $\check{\boldsymbol{\beta}}_g$  is the OLS estimator for equation  $g$ .

7.6. The model for a random draw from the population is  $\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{u}_i$ , which can be written as

$$\mathbf{y}_i = \mathbf{X}_{i1} \boldsymbol{\beta}_1 + \mathbf{X}_{i2} \boldsymbol{\beta}_2 + \mathbf{u}_i,$$

where the partition of  $\mathbf{X}_i$  is defined in the problem. Now, if  $\boldsymbol{\beta}_1 = \mathbf{R}_1^{-1}(\mathbf{r} - \mathbf{R}_2\boldsymbol{\beta}_2)$ , we just plug this into the previous equation:

$$\begin{aligned}\mathbf{y}_i &= \mathbf{X}_{i1}\boldsymbol{\beta}_1 + \mathbf{X}_{i2}\boldsymbol{\beta}_2 + \mathbf{u}_i = \mathbf{X}_{i1}\mathbf{R}_1^{-1}(\mathbf{r} - \mathbf{R}_2\boldsymbol{\beta}_2) + \mathbf{X}_{i2}\boldsymbol{\beta}_2 + \mathbf{u}_i \\ &= \mathbf{X}_{i1}\mathbf{R}_1^{-1}\mathbf{r} + (\mathbf{X}_{i2} - \mathbf{X}_{i1}\mathbf{R}_2)\boldsymbol{\beta}_2 + \mathbf{u}_i.\end{aligned}$$

Bringing  $\mathbf{X}_{i1}\mathbf{R}_1^{-1}\mathbf{r}$  to the left hand side gives

$$\mathbf{y}_i - \mathbf{X}_{i1}\mathbf{R}_1^{-1}\mathbf{r} = (\mathbf{X}_{i2} - \mathbf{X}_{i1}\mathbf{R}_2)\boldsymbol{\beta}_2 + \mathbf{u}_i.$$

If we define  $\tilde{\mathbf{y}}_i = \mathbf{y}_i - \mathbf{X}_{i1}\mathbf{R}_1^{-1}\mathbf{r}$  and  $\tilde{\mathbf{X}}_{i2} \equiv \mathbf{X}_{i2} - \mathbf{X}_{i1}\mathbf{R}_2$ , then we get the desired equation:

$$\tilde{\mathbf{y}}_i = \tilde{\mathbf{X}}_{i2}\boldsymbol{\beta}_2 + \mathbf{u}_i.$$

(Note that  $\tilde{\mathbf{y}}_i$  and  $\tilde{\mathbf{X}}_{i2}$  are functions of the data for observation  $i$  and the known matrices  $\mathbf{R}_1, \mathbf{R}_2$ , and the known vector  $\mathbf{r}$ .)

This general result is very convenient for computing the weighted SSR form of the  $F$  statistic (under SGL.3). Let  $\hat{\boldsymbol{\Omega}}$  denote the estimate of  $\boldsymbol{\Omega}$  based on estimation of the unconstrained system; typically,  $\hat{\boldsymbol{\Omega}} = N^{-1} \sum_{i=1}^N \tilde{\mathbf{u}}_i \tilde{\mathbf{u}}_i'$  where  $\tilde{\mathbf{u}}_i$  are the system OLS residuals. Using this matrix, we estimate  $\mathbf{y}_i = \mathbf{X}_{i1}\boldsymbol{\beta}_1 + \mathbf{X}_{i2}\boldsymbol{\beta}_2 + \mathbf{u}_i$  and then  $\tilde{\mathbf{y}}_i = \tilde{\mathbf{X}}_{i2}\boldsymbol{\beta}_2 + \mathbf{u}_i$  by FGLS using  $\hat{\boldsymbol{\Omega}}$ . Let  $\hat{\mathbf{u}}_i$  denote the FGLS residuals from the unrestricted model and let  $\tilde{\mathbf{u}}_i = \tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_{i2}\tilde{\boldsymbol{\beta}}_2$  denote the restricted FGLS residuals, where  $\tilde{\boldsymbol{\beta}}_2$  is the FGLS estimator from the restricted estimation. Then the  $F$  statistic computed from (7.57) has an approximate  $\mathfrak{F}_{Q, NG-K}$  distribution under  $H_0$  (assuming SGLS.1, SGLS.2, and SGLS.3 hold).

7.7. a. First, the diagonal elements of  $\boldsymbol{\Omega}$  are easily found because

$E(u_{it}^2) = E[E(u_{it}^2 | \mathbf{x}_{it})] = \sigma_t^2$  by iterated expectations. Now, consider  $E(u_{it}u_{is})$ , and take  $s < t$  without loss of generality. Under  $E(u_{it} | \mathbf{x}_{it}, u_{i,t-1}, \dots) = 0$ ,  $E(u_{it} | u_{is}) = 0$  because  $u_{is}$  is a subset of the larger conditioning set. Applying LIE again we have

$$E(u_{it}u_{is}) = E[E(u_{it}u_{is}|u_{is})] = E[E(u_{it}|u_{is})u_{is}] = 0.$$

So

$$\mathbf{\Omega} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_T^2 \end{pmatrix}.$$

b. The GLS estimator is

$$\begin{aligned} \boldsymbol{\beta}^* &\equiv \left( \sum_{i=1}^N \mathbf{X}_i' \mathbf{\Omega}^{-1} \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}_i' \mathbf{\Omega}^{-1} \mathbf{y}_i \right) \\ &= \left( \sum_{i=1}^N \sum_{t=1}^T \sigma_t^{-2} \mathbf{x}_{it}' \mathbf{x}_{it} \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T \sigma_t^{-2} \mathbf{x}_{it}' y_{it} \right), \end{aligned}$$

which is a weighted least squares estimator with every observation for time period  $t$  weighted by  $\sigma_t^{-2}$ , the inverse of the variance.

c. If, say,  $y_{it} = \beta_0 + \beta_1 y_{i,t-1} + u_{it}$ , then  $y_{it}$  is clearly correlated with  $u_{it}$ , which says that  $\mathbf{x}_{i,t+1} = y_{it}$  is correlated with  $u_{it}$ . Thus, SGLS.1 cannot hold. Generally, SGLS.1 fails to hold whenever there is feedback from  $y_{it}$  to  $\mathbf{x}_{is}, s > t$ . Nevertheless, because  $\mathbf{\Omega}^{-1}$  is diagonal,

$\mathbf{X}_i' \mathbf{\Omega}^{-1} \mathbf{u}_i = \sum_{t=1}^T \mathbf{x}_{it}' \sigma_t^{-2} u_{it}$ , and so

$$E(\mathbf{X}_i' \mathbf{\Omega}^{-1} \mathbf{u}_i) - \sum_{t=1}^T \sigma_t^{-2} E(\mathbf{x}_{it}' u_{it}) = \mathbf{0},$$

where we use  $E(\mathbf{x}_{it}' u_{it}) = 0$  under  $E(u_{it} | \mathbf{x}_{it}, u_{i,t-1}, \dots) = 0$ . It follows that the GLS estimator is GLS is consistent in this case without SGLS.1.

d. First, since  $\mathbf{\Omega}^{-1}$  is diagonal,  $\mathbf{X}_i' \mathbf{\Omega}^{-1} = (\sigma_1^{-2} \mathbf{x}_{i1}', \sigma_2^{-2} \mathbf{x}_{i2}', \dots, \sigma_T^{-2} \mathbf{x}_{iT}')'$ , and so

$$E(\mathbf{X}_i' \boldsymbol{\Omega}^{-1} \mathbf{u}_i \mathbf{u}_i' \boldsymbol{\Omega}^{-1} \mathbf{X}_i) = \sum_{t=1}^T \sum_{s=1}^T \sigma_t^{-2} \sigma_s^{-2} E(u_{it} u_{is} \mathbf{x}_{it}' \mathbf{x}_{is}).$$

First consider the terms for  $s \neq t$ . Under  $E(u_{it} | \mathbf{x}_{it}, u_{i,t-1}, \dots) = 0$ ,  $E(u_{it} | \mathbf{x}_{it}, u_{is}, \mathbf{x}_{is}) = 0$  for  $s < t$ , and so by the LIE,  $E(u_{it} u_{is} \mathbf{x}_{it}' \mathbf{x}_{is}) = 0$ , all  $t \neq s$ . Next, for each  $t$ ,

$$\begin{aligned} E(u_{it}^2 \mathbf{x}_{it}' \mathbf{x}_{it}) &= E[E(u_{it}^2 \mathbf{x}_{it}' \mathbf{x}_{it} | \mathbf{x}_{it})] = E[E(u_{it}^2 | \mathbf{x}_{it}) \mathbf{x}_{it}' \mathbf{x}_{it}] \\ &= E(\sigma_t^2 \mathbf{x}_{it}' \mathbf{x}_{it}) = \sigma_t^2 E(\mathbf{x}_{it}' \mathbf{x}_{it}), \quad t = 1, 2, \dots, T. \end{aligned}$$

It follows that

$$E(\mathbf{X}_i' \boldsymbol{\Omega}^{-1} \mathbf{u}_i \mathbf{u}_i' \boldsymbol{\Omega}^{-1} \mathbf{X}_i) = \sum_{t=1}^T \sigma_t^{-2} E(\mathbf{x}_{it}' \mathbf{x}_{it}) = E(\mathbf{X}_i' \boldsymbol{\Omega}^{-1} \mathbf{X}_i).$$

e. First, run pooled OLS across all  $i$  and  $t$  and let  $\check{u}_{it}$  denote the pooled OLS residuals.

Then, for each  $t$ , define

$$\hat{\sigma}_t^2 = N^{-1} \sum_{i=1}^N \check{u}_{it}^2$$

(We might replace  $N$  with  $N - K$  as a degree-of-freedom adjustment.) By standard arguments,

$$\hat{\sigma}_t^2 \xrightarrow{P} \sigma_t^2 \text{ as } N \rightarrow \infty.$$

f. What we need to show is that replacing the  $\sigma_t^2$  with the  $\hat{\sigma}_t^2$  does not affect the  $\sqrt{N}$ -asymptotic distribution of the FGLS estimator. We know this generally under SGLS.1, but we have relaxed that assumption. To show it holds in the current setting we need to show

$$\begin{aligned} N^{-1} \sum_{i=1}^N \sum_{t=1}^T \hat{\sigma}_t^{-2} \mathbf{x}_{it}' \mathbf{x}_{it} &= N^{-1} \sum_{i=1}^N \sum_{t=1}^T \sigma_t^{-2} \mathbf{x}_{it}' \mathbf{x}_{it} + o_p(1) \\ N^{-1/2} \sum_{i=1}^N \sum_{t=1}^T \hat{\sigma}_t^{-2} \mathbf{x}_{it}' u_{it} &= N^{-1/2} \sum_{i=1}^N \sum_{t=1}^T \sigma_t^{-2} \mathbf{x}_{it}' u_{it} + o_p(1). \end{aligned}$$

The first follows from the consistency of each  $\hat{\sigma}_t^2$  using standard arguments we have used

before. The second requirement follows from

$$\begin{aligned} N^{-1/2} \sum_{i=1}^N \sum_{t=1}^T \hat{\sigma}_t^{-2} \mathbf{x}_{it}' u_{it} - N^{-1/2} \sum_{i=1}^N \sum_{t=1}^T \sigma_t^{-2} \mathbf{x}_{it}' u_{it} &= \sum_{t=1}^T \left[ N^{-1/2} \sum_{i=1}^N \mathbf{x}_{it}' u_{it} \right] (\hat{\sigma}_t^{-2} - \sigma_t^{-2}) \\ &= \sum_{t=1}^T O_p(1) \cdot o_p(1) = o_p(1) \end{aligned}$$

because  $N^{-1/2} \sum_{i=1}^N \mathbf{x}_{it}' u_{it}$  satisfies the CLT under Under  $E(u_{it} | \mathbf{x}_{it}, u_{i,t-1}, \dots) = 0$  and second moment assumptions.

So now we know all inference is as if we are applying pooled OLS to

$$(y_{it}/\sigma_t) = (\mathbf{x}_{it}/\sigma_t)\boldsymbol{\beta} + e_{it}, \quad t = 1, 2, \dots, T$$

where this equation satisfies POLS.1, POLS.2, and POLS.3. Thus, we can use the usual statistics – standard errors, confidence intervals,  $t$  and  $F$  statistics – from the regression

$$(y_{it}/\hat{\sigma}_t) = (\mathbf{x}_{it}/\hat{\sigma}_t), \quad t = 1, \dots, T; i = 1, \dots, N.$$

For  $F$  testing, note that the  $\hat{\sigma}_t^2$  should be obtained using the pooled OLS residuals for the unrestricted model.

g. If  $\sigma_t^2 = \sigma^2$  for all  $t = 1, \dots, T$ , inference is very easy because with weighted least squares method reduces to pooled OLS. Thus, we can use the standard errors and test statistics reported by a standard OLS regression pooled across  $i$  and  $t$ .

**7.8.** Here is some Stata output:

```
. use fringe
. gen hrvac = vacdays/annhrs
. gen hrsick = sicklve/annhrs
. gen hrins = insur/annhrs
. gen hrpens = pension/annhrs
. sureg (hrearn hrvac hrsick hrins hrpens = educ exper expersq tenure
```



tenuresq union south nrtheast nrthcen married white male), corr

Seemingly unrelated regression

Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
hrearn	616	12	4.3089	0.2051	158.93	0.0000
hrvac	616	12	.1389899	0.3550	339.01	0.0000
hrsick	616	12	.056924	0.2695	227.23	0.0000
hrins	616	12	.1573797	0.3891	392.27	0.0000
hrpens	616	12	.2500388	0.3413	319.16	0.0000

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
hrearn						
educ	.4588139	.068393	6.71	0.000	.3247662	.5928617
exper	-.0758428	.0567371	-1.34	0.181	-.1870455	.0353598
expersq	.0039945	.0011655	3.43	0.001	.0017102	.0062787
tenure	.1100846	.0829207	1.33	0.184	-.052437	.2726062
tenuresq	-.0050706	.0032422	-1.56	0.118	-.0114252	.0012839
union	.8079933	.4034789	2.00	0.045	.0171892	1.598797
south	-.4566222	.5458508	-0.84	0.403	-1.52647	.6132258
nrtheast	-1.150759	.5993283	-1.92	0.055	-2.32542	.0239032
nrthcen	-.6362663	.5501462	-1.16	0.247	-1.714533	.4420005
married	.6423882	.4133664	1.55	0.120	-.167795	1.452571
white	1.140891	.6054474	1.88	0.060	-.0457639	2.327546
male	1.784702	.3937853	4.53	0.000	1.012897	2.556507
_cons	-2.632127	1.215291	-2.17	0.030	-5.014054	-.2501997
hrvac						
educ	.0201829	.0022061	9.15	0.000	.015859	.0245068
exper	.0066493	.0018301	3.63	0.000	.0030623	.0102363
expersq	-.0001492	.0000376	-3.97	0.000	-.0002229	-.0000755
tenure	.012386	.0026747	4.63	0.000	.0071436	.0176284
tenuresq	-.0002155	.0001046	-2.06	0.039	-.0004205	-.0000106
union	.0637464	.0130148	4.90	0.000	.0382378	.0892549
south	-.0179005	.0176072	-1.02	0.309	-.05241	.016609
nrtheast	-.0169824	.0193322	-0.88	0.380	-.0548728	.0209081
nrthcen	.0002511	.0177458	0.01	0.989	-.03453	.0350321
married	.0227586	.0133337	1.71	0.088	-.0033751	.0488923
white	.0084869	.0195296	0.43	0.664	-.0297905	.0467642
male	.0569525	.0127021	4.48	0.000	.0320568	.0818482
_cons	-.1842348	.039201	-4.70	0.000	-.2610674	-.1074022
hrsick						
educ	.0096054	.0009035	10.63	0.000	.0078346	.0113763
exper	.002145	.0007495	2.86	0.004	.0006759	.0036141
expersq	-.0000383	.0000154	-2.48	0.013	-.0000684	-8.08e-06
tenure	.0050021	.0010954	4.57	0.000	.002855	.0071491
tenuresq	-.0001391	.0000428	-3.25	0.001	-.0002231	-.0000552
union	-.0046655	.0053303	-0.88	0.381	-.0151127	.0057816
south	-.011942	.0072111	-1.66	0.098	-.0260755	.0021916
nrtheast	-.0026651	.0079176	-0.34	0.736	-.0181833	.0128531
nrthcen	-.0222014	.0072679	-3.05	0.002	-.0364462	-.0079567
married	.0038338	.0054609	0.70	0.483	-.0068694	.014537
white	.0038635	.0079984	0.48	0.629	-.0118132	.0195401
male	.0042538	.0052022	0.82	0.414	-.0059423	.01445

	_cons	-.0937606	.016055	-5.84	0.000	-.1252278	-.0622935
-----							
hrins							
	educ	.0080042	.002498	3.20	0.001	.0031082	.0129002
	exper	.0054052	.0020723	2.61	0.009	.0013436	.0094668
	expersq	-.0001266	.0000426	-2.97	0.003	-.00021	-.0000431
	tenure	.0116978	.0030286	3.86	0.000	.0057618	.0176338
	tenuresq	-.0002466	.0001184	-2.08	0.037	-.0004787	-.0000146
	union	.1441536	.0147368	9.78	0.000	.11527	.1730372
	south	.0196786	.0199368	0.99	0.324	-.0193969	.0587541
	nrtheast	-.0052563	.0218901	-0.24	0.810	-.0481601	.0376474
	nrthcen	.0242515	.0200937	1.21	0.227	-.0151315	.0636345
	married	.0365441	.0150979	2.42	0.016	.0069527	.0661355
	white	.0378883	.0221136	1.71	0.087	-.0054535	.0812301
	male	.1120058	.0143827	7.79	0.000	.0838161	.1401955
	_cons	-.1180824	.0443877	-2.66	0.008	-.2050807	-.0310841
-----							
hrpens							
	educ	.0390226	.0039687	9.83	0.000	.031244	.0468012
	exper	.0083791	.0032924	2.55	0.011	.0019262	.0148321
	expersq	-.0001595	.0000676	-2.36	0.018	-.0002921	-.000027
	tenure	.0243758	.0048118	5.07	0.000	.0149449	.0338067
	tenuresq	-.0005597	.0001881	-2.97	0.003	-.0009284	-.0001909
	union	.1621404	.0234133	6.93	0.000	.1162513	.2080296
	south	-.0130816	.0316749	-0.41	0.680	-.0751632	.049
	nrtheast	-.0323117	.0347781	-0.93	0.353	-.1004755	.0358521
	nrthcen	-.0408177	.0319241	-1.28	0.201	-.1033878	.0217525
	married	-.0051755	.023987	-0.22	0.829	-.0521892	.0418381
	white	.0395839	.0351332	1.13	0.260	-.0292758	.1084437
	male	.0952459	.0228508	4.17	0.000	.0504592	.1400325
	_cons	-.4928338	.0705215	-6.99	0.000	-.6310534	-.3546143
-----							

Correlation matrix of residuals:

	hrearn	hrvac	hrsick	hrins	hrpens
hrearn	1.0000				
hrvac	0.2719	1.0000			
hrsick	0.2541	0.5762	1.0000		
hrins	0.2609	0.6701	0.2922	1.0000	
hrpens	0.2786	0.7070	0.4569	0.6345	1.0000

Breusch-Pagan test of independence: chi2(10) = 1393.265, Pr = 0.0000

. test married

```
( 1) [hrearn]married = 0
( 2) [hrvac]married = 0
( 3) [hrsick]married = 0
( 4) [hrins]married = 0
( 5) [hrpens]married = 0
```

```
      chi2( 5) =    14.48
Prob > chi2 =    0.0128
```

. lincom [hrpens]educ - [hrins]educ

```
( 1) - [hrins]educ + [hrpens]educ = 0
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
(1)	.0310184	.0030676	10.11	0.000	.025006	.0370308

The first test shows that there is some evidence that marital status affects at least one of the five forms of compensation. In fact, it has the largest economic effect on hourly earnings: .642, but its  $t$  statistic is only about 1.54. The most statistically significant effect is on *hrins*: .037 with  $t = 2.42$ . It is marginally significant and positive for *hrvac* as well.

The `lincom` command tests whether another year of education has the same effect on *hrpens* and *hrins*. The  $t$  statistic is 10.11 and the  $p$ -value is effectively zero. The estimate in the *hrpens* equation (with standard error) is .039 (.004) while the estimate in the *hrins* equation is .008 (.003). Thus, each is positive and statistically significant, and they are significantly different from one another.

All of the standard errors and statistics reported above assume that SGLS.3 holds, so that there can be no system heteroskedasticity. This is unlikely to hold in this example.

7.9. The Stata session follows, including a test for serial correlation before computing the fully robust standard errors:

```
. use jtrain1
. xtset fcode year
      panel variable:  fcode (strongly balanced)
      time variable:  year, 1987 to 1989
      delta: 1 unit
. reg lscrap d89 grant grant_1 lscrap_1 if year != 1987
```

Source	SS	df	MS	Number of obs =	108
Model	186.376973	4	46.5942432	F( 4, 103) =	153.67
Residual	31.2296502	103	.303200488	Prob > F =	0.0000
Total	217.606623	107	2.03370676	R-squared =	0.8565
				Adj R-squared =	0.8509
				Root MSE =	.55064

lscrap	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
--------	-------	-----------	---	------	---------------------	--

d89	-.1153893	.1199127	-0.96	0.338	-.3532078	.1224292
grant	-.1723924	.1257443	-1.37	0.173	-.4217765	.0769918
grant_1	-.1073226	.1610378	-0.67	0.507	-.426703	.2120579
lscrap_1	.8808216	.0357963	24.61	0.000	.809828	.9518152
_cons	-.0371354	.0883283	-0.42	0.675	-.2123137	.138043

The estimated effect of *grant*, and its lag, are now the expected sign (if we think the job training program should reduce the scrap rate), but neither is strongly statistically significant.

The variable *grant* would be if we use a 10% significance level and a one-sided test. The results are certainly different from when we omit the lag of  $\log(\text{scrap})$ .

Now test for  $AR(1)$  serial correlation:

```
. gen uhat_1 = 1.uhat
(417 missing values generated)
```

```
. reg lscrap grant grant_1 lscrap_1 uhat_1 if d89
```

Source	SS	df	MS	Number of obs =	54
Model	94.4746525	4	23.6186631	F( 4, 49) =	73.47
Residual	15.7530202	49	.321490208	Prob > F =	0.0000
Total	110.227673	53	2.07976741	R-squared =	0.8571
				Adj R-squared =	0.8454
				Root MSE =	.567

lscrap	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
grant	.0165089	.215732	0.08	0.939	-.4170208	.4500385
grant_1	-.0276544	.1746251	-0.16	0.875	-.3785767	.3232679
lscrap_1	.9204706	.0571831	16.10	0.000	.8055569	1.035384
uhat_1	.2790328	.1576739	1.77	0.083	-.0378247	.5958904
_cons	-.232525	.1146314	-2.03	0.048	-.4628854	-.0021646

The estimate of  $\rho$  is about .28, and it is marginally significant with  $t = 1.77$ . (Note we are relying on asymptotics with  $N = 54$ .) One could probably make a case for ignoring the serial correlation. But it is easy enough to obtain the serial-correlation and heteroskedasticity-robust standard errors:

```
. reg lscrap d89 grant grant_1 lscrap_1 if year != 1987, robust cluster(fcode
```

Linear regression	Number of obs =	108
	F( 4, 53) =	77.24
	Prob > F =	0.0000
	R-squared =	0.8565

Root MSE = .55064

(Std. Err. adjusted for 54 clusters in fcode)

lscrap	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
d89	-.1153893	.1145118	-1.01	0.318	-.3450708	.1142922
grant	-.1723924	.1188807	-1.45	0.153	-.4108369	.0660522
grant_1	-.1073226	.1790052	-0.60	0.551	-.4663616	.2517165
lscrap_1	.8808216	.0645344	13.65	0.000	.7513821	1.010261
_cons	-.0371354	.0893147	-0.42	0.679	-.216278	.1420073

The robust standard errors for *grant* and *grant\_1* are actually smaller than the usual ones, but each is still not statistically significant at the 5% level against a one-sided alternative. In addition, they are not jointly significant, as the *p*-value is about .33:

```
. test grant grant_1

( 1)  grant = 0
( 2)  grant_1 = 0

      F( 2,    53) =    1.14
      Prob > F =    0.3266
```

### 7.10. The Stata results are:

```
. use gpa

. reg trmgpa spring cumgpa crsgpa frstsem season sat verbmhath hsperc hssize
  black female
```

Source	SS	df	MS	Number of obs = 732	
Model	218.156689	11	19.8324263	F( 11, 720) =	70.64
Residual	202.140267	720	.280750371	Prob > F =	0.0000
Total	420.296956	731	.574961636	R-squared =	0.5191
				Adj R-squared =	0.5117
				Root MSE =	.52986

trmgpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
spring	-.0121568	.0464813	-0.26	0.794	-.1034118	.0790983
cumgpa	.3146158	.0404916	7.77	0.000	.2351201	.3941115
crsgpa	.9840371	.0960343	10.25	0.000	.7954964	1.172578
frstsem	.7691192	.1204162	6.39	0.000	.5327104	1.005528
season	-.0462625	.0470985	-0.98	0.326	-.1387292	.0462042
sat	.0014097	.0001464	9.63	0.000	.0011223	.0016972
verbmhath	-.112616	.1306157	-0.86	0.389	-.3690491	.1438171
hsperc	-.0066014	.0010195	-6.48	0.000	-.0086029	-.0045998
hssize	-.0000576	.0000994	-0.58	0.562	-.0002527	.0001375
black	-.2312855	.0543347	-4.26	0.000	-.3379589	-.1246122
female	.2855528	.0509641	5.60	0.000	.1854967	.3856089

```

      _cons |   -2.067599   .3381007   -6.12   0.000   -2.731381   -1.403818
-----+-----

```

```

. reg trmgpa spring cumgpa crsgpa frstsem season sat verbmth hsperc hssize
  black female, robust cluster(id)

```

```

Linear regression                               Number of obs =      732
                                                F( 11,    365) =    71.31
                                                Prob > F      =    0.0000
                                                R-squared     =    0.5191
                                                Root MSE     =    .52986

```

(Std. Err. adjusted for 366 clusters in id)

trmgpa	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
spring	-.0121568	.0395519	-0.31	0.759	-.089935	.0656215
cumgpa	.3146158	.0514364	6.12	0.000	.2134669	.4157647
crsgpa	.9840371	.09182	10.72	0.000	.8034745	1.1646
frstsem	.7691192	.1437178	5.35	0.000	.4865003	1.051738
season	-.0462625	.0431631	-1.07	0.285	-.131142	.038617
sat	.0014097	.0001743	8.09	0.000	.001067	.0017525
verbmth	-.112616	.1495196	-0.75	0.452	-.4066441	.1814121
hsperc	-.0066014	.0011954	-5.52	0.000	-.0089522	-.0042506
hssize	-.0000576	.0001066	-0.54	0.589	-.0002673	.0001521
black	-.2312855	.0695278	-3.33	0.001	-.368011	-.0945601
female	.2855528	.0511767	5.58	0.000	.1849146	.386191
_cons	-2.067599	.3327336	-6.21	0.000	-2.721915	-1.413284

Some of the fully robust standard errors are actually smaller than the corresponding nonrobust standard error, although the one on *cumgpa* is quite a bit larger, and drops the *t* statistic from 10.25 to 6.12. No variable that was statistically significant based on the usual *t* statistic becomes statistically insignificant, although the length of some confidence intervals change. The *t* statistics for the key variable, *season*, are similarly and show *season* is not statistically significant.

7.11. a. The following Stata output should be self-explanatory. There is clearly strong positive serial correlation in the errors of the static model ( $\hat{\rho} = .792$ ,  $t_{\hat{\rho}} = 28.84$ ) and the fully robust standard errors are much larger than the nonrobust ones. Not, for example, that the *t* statistic on the log of the conviction probability, *lprbconv* goes from  $-20.69$  to  $-7.75$ .

```

. use cornwell

```

```
. xtset county year
      panel variable:  county (strongly balanced)
      time variable:  year, 81 to 87
              delta:  1 unit
```

```
. reg lcrmte lprbarr lprbconv lprbpris lavgsen lpolpc d82-d87
```

Source	SS	df	MS	Number of obs =	630
Model	117.644669	11	10.6949699	F( 11, 618) =	74.49
Residual	88.735673	618	.143585231	Prob > F =	0.0000
				R-squared =	0.5700
				Adj R-squared =	0.5624
Total	206.380342	629	.328108652	Root MSE =	.37893

lcrmte	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
lprbarr	-.7195033	.0367657	-19.57	0.000	-.7917042	-.6473024
lprbconv	-.5456589	.0263683	-20.69	0.000	-.5974413	-.4938765
lprbpris	.2475521	.0672268	3.68	0.000	.1155314	.3795728
lavgsen	-.0867575	.0579205	-1.50	0.135	-.2005023	.0269872
lpolpc	.3659886	.0300252	12.19	0.000	.3070248	.4249525
d82	.0051371	.057931	0.09	0.929	-.1086284	.1189026
d83	-.043503	.0576243	-0.75	0.451	-.1566662	.0696601
d84	-.1087542	.057923	-1.88	0.061	-.222504	.0049957
d85	-.0780454	.0583244	-1.34	0.181	-.1925835	.0364928
d86	-.0420791	.0578218	-0.73	0.467	-.15563	.0714719
d87	-.0270426	.056899	-0.48	0.635	-.1387815	.0846963
_cons	-2.082293	.2516253	-8.28	0.000	-2.576438	-1.588149

```
. predict uhat, resid
```

```
. gen uhat_1 = 1.uhat
(90 missing values generated)
```

```
. reg uhat uhat_1
```

Source	SS	df	MS	Number of obs =	540
Model	46.6680407	1	46.6680407	F( 1, 538) =	831.46
Residual	30.1968286	538	.056127934	Prob > F =	0.0000
				R-squared =	0.6071
				Adj R-squared =	0.6064
Total	76.8648693	539	.142606437	Root MSE =	.23691

uhat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
uhat_1	.7918085	.02746	28.84	0.000	.7378666	.8457504
_cons	1.74e-10	.0101951	0.00	1.000	-.0200271	.0200271

```
. reg lcrmte lprbarr lprbconv lprbpris lavgsen lpolpc d82-d87, cluster(county)
```

Linear regression	Number of obs =	630
	F( 11, 89) =	37.19
	Prob > F =	0.0000
	R-squared =	0.5700
	Root MSE =	.37893

(Std. Err. adjusted for 90 clusters in county)

lcrmte	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
lprbarr	-.7195033	.1095979	-6.56	0.000	-.9372719	-.5017347
lprbconv	-.5456589	.0704368	-7.75	0.000	-.6856152	-.4057025
lprbpris	.2475521	.1088453	2.27	0.025	.0312787	.4638255
lavgsen	-.0867575	.1130321	-0.77	0.445	-.3113499	.1378348
lpolpc	.3659886	.121078	3.02	0.003	.1254092	.6065681
d82	.0051371	.0367296	0.14	0.889	-.0678439	.0781181
d83	-.043503	.033643	-1.29	0.199	-.1103509	.0233448
d84	-.1087542	.0391758	-2.78	0.007	-.1865956	-.0309127
d85	-.0780454	.0385625	-2.02	0.046	-.1546683	-.0014224
d86	-.0420791	.0428788	-0.98	0.329	-.1272783	.0431201
d87	-.0270426	.0381447	-0.71	0.480	-.1028353	.0487502
_cons	-2.082293	.8647054	-2.41	0.018	-3.800445	-.3641423

. drop uhat uhat\_1

b. We lose the first year, 1981, when we add the lag of  $\log(crmte)$ :

. gen lcrmte\_1 = l.lcrmte  
(90 missing values generated)

. reg lcrmte lcrmte\_1 lprbarr lprbconv lprbpris lavgsen lpolpc d83-d87

Source	SS	df	MS	Number of obs =	540
Model	163.287174	11	14.8442885	F( 11, 528) =	464.68
Residual	16.8670945	528	.031945255	Prob > F =	0.0000
Total	180.154268	539	.334237975	R-squared =	0.9064
				Adj R-squared =	0.9044
				Root MSE =	.17873

lcrmte	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
lcrmte_1	.8263047	.0190806	43.31	0.000	.7888214	.8637879
lprbarr	-.1668349	.0229405	-7.27	0.000	-.2119007	-.1217691
lprbconv	-.1285118	.0165096	-7.78	0.000	-.1609444	-.0960793
lprbpris	-.0107492	.0345003	-0.31	0.755	-.078524	.0570255
lavgsen	-.1152298	.030387	-3.79	0.000	-.174924	-.0555355
lpolpc	.101492	.0164261	6.18	0.000	.0692234	.1337606
d83	-.0649438	.0267299	-2.43	0.015	-.1174537	-.0124338
d84	-.0536882	.0267623	-2.01	0.045	-.1062619	-.0011145
d85	-.0085982	.0268172	-0.32	0.749	-.0612797	.0440833
d86	.0420159	.026896	1.56	0.119	-.0108203	.0948522
d87	.0671272	.0271816	2.47	0.014	.0137298	.1205245
_cons	-.0304828	.1324195	-0.23	0.818	-.2906166	.229651

Not surprisingly, the coefficient on the lagged crime rate is very large and statistically significant. Further, including it makes all other coefficients much smaller in magnitude. The



variable  $\log(\text{prbpris})$  now has a negative sign, although it is insignificant. Adding the lagged crime rate does not change the positive coefficient on the size of the police force: it is smaller but now even more statistically significant.

c. There is little evidence of serial correlation in the model with a lagged dependent variable. The coefficient on  $\hat{u}_{t-1}$  is small and statistically insignificant:

```
. predict uhat, resid
(90 missing values generated)

. gen uhat_1 = l.uhat
(180 missing values generated)

. reg lcrmte lcrmte_1 lprbarr lprbconv lprbpris lavgsen lpolpc d84-d87
    uhat_1
```

Source	SS	df	MS	Number of obs =	450
Model	138.488359	11	12.5898508	F( 11, 438) =	370.77
Residual	14.8729012	438	.033956395	Prob > F =	0.0000
				R-squared =	0.9030
				Adj R-squared =	0.9006
Total	153.36126	449	.341561826	Root MSE =	.18427

	lcrmte	Coef.	Std. Err.	t	P> t	[95% Conf. Interval
	lcrmte_1	.829714	.0248121	33.44	0.000	.7809485 .8784796
	lprbarr	-.1576381	.0278786	-5.65	0.000	-.2124305 -.1028457
	lprbconv	-.1293032	.0191735	-6.74	0.000	-.1669868 -.0916197
	lprbpris	-.0040031	.0395191	-0.10	0.919	-.0816738 .0736675
	lavgsen	-.1241479	.034481	-3.60	0.000	-.1919166 -.0563791
	lpolpc	.1107055	.0187613	5.90	0.000	.0738323 .1475788
	d84	.0103772	.0277393	0.37	0.709	-.0441415 .0648959
	d85	.0557956	.0277577	2.01	0.045	.0012407 .1103505
	d86	.107831	.0277087	3.89	0.000	.0533724 .1622895
	d87	.1333345	.0279635	4.77	0.000	.0783751 .1882938
	uhat_1	-.0592978	.0601101	-0.99	0.324	-.177438 .0588423
	_cons	.0126059	.1524765	0.08	0.934	-.2870706 .3122823

d. None of the  $\log(\text{wage})$  variables is statistically significant, and the magnitudes are pretty small in all cases. The  $p$ -value for the joint test, made fully robust, is .33, which means the  $\log(\text{wage})$  variables are jointly insignificant, too. (Plus, the different signs on the wage variables is hard to explain, except to conclude that each is estimated with substantial sampling

error.)

```
. reg lcrmte lcrmte_1 lprbarr lprbconv lprbpris lavgsen lpolpc d83-d87
    lwcon-lwloc, cluster(county)
```

Linear regression

Number of obs = 540  
F( 20, 89) = 398.63  
Prob > F = 0.0000  
R-squared = 0.9077  
Root MSE = .17895

(Std. Err. adjusted for 90 clusters in county)

lcrmte	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
lcrmte_1	.8087768	.0406432	19.90	0.000	.7280195	.889534
lprbarr	-.1746053	.0495539	-3.52	0.001	-.2730678	-.0761428
lprbconv	-.1337714	.0289031	-4.63	0.000	-.1912012	-.0763415
lprbpris	-.0195318	.0404094	-0.48	0.630	-.0998243	.0607608
lavgsen	-.1108926	.0455404	-2.44	0.017	-.2013804	-.0204049
lpolpc	.1050704	.0575404	1.83	0.071	-.0092612	.219402
d83	-.0729231	.0293628	-2.48	0.015	-.1312664	-.0145799
d84	-.0652494	.0226239	-2.88	0.005	-.1102026	-.0202962
d85	-.0258059	.0413435	-0.62	0.534	-.1079545	.0563428
d86	.0263763	.0393741	0.67	0.505	-.0518591	.1046118
d87	.0465632	.0441727	1.05	0.295	-.041207	.1343334
lwcon	-.0283133	.0272813	-1.04	0.302	-.0825207	.025894
lwtuc	-.0034567	.0208431	-0.17	0.869	-.0448715	.0379582
lwtrd	.0121236	.0496718	0.24	0.808	-.0865733	.1108205
lwfir	.0296003	.0184296	1.61	0.112	-.0070189	.0662195
lwser	.012903	.0269695	0.48	0.634	-.0406847	.0664908
lwmfg	-.0409046	.0508117	-0.81	0.423	-.1418664	.0600573
lwfed	.1070534	.0760639	1.41	0.163	-.044084	.2581908
lwsta	-.0903894	.0548237	-1.65	0.103	-.199323	.0185442
lwloc	.0961124	.1355681	0.71	0.480	-.1732585	.3654833
_cons	-.6438061	.7958054	-0.81	0.421	-2.225055	.9374423

```
. testparm lwcon-lwloc
```

```
( 1) lwcon = 0
( 2) lwtuc = 0
( 3) lwtrd = 0
( 4) lwfir = 0
( 5) lwser = 0
( 6) lwmfg = 0
( 7) lwfed = 0
( 8) lwsta = 0
( 9) lwloc = 0
```

F( 9, 89) = 1.15  
Prob > F = 0.3338

## 7.12. Wealth at the beginning of the year cannot be strictly exogenous in a savings

equation: if saving increases unexpectedly this year – so that the disturbance in year  $t$  is positive – beginning of year wealth is higher next year. This is analogous to Example 7.8, where cumulative grade point average at the start of the semester cannot be strictly exogenous in an equation to explain current-term GPA.

**7.13.** a. The Stata output is below. Married men are estimated to have a scoring average about 1.2 points higher, and assists are .42 higher. The coefficient in the *rebounds* equation is  $-.24$ , but it is not statistically significant. The coefficient in the *assists* equation is significant at the 5% level against a two-sided alternative ( $p$ -value = .048).

```
. use nbasal
. sureg (points rebounds assists = age exper expersq coll guard forward black
Seemingly unrelated regression
```

Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
points	282	8	5.352116	0.1750	59.80	0.0000
rebounds	282	8	2.375338	0.3123	128.07	0.0000
assists	282	8	1.64516	0.3727	167.51	0.0000

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	Interval
points						
age	-1.214936	.27656	-4.39	0.000	-1.756984	-.6728889
exper	2.261943	.3759275	6.02	0.000	1.525138	2.998747
expersq	-.0649631	.0204961	-3.17	0.002	-.1051347	-.0247915
coll	-1.011535	.4026169	-2.51	0.012	-1.800649	-.2224201
guard	1.997013	.9380542	2.13	0.033	.1584603	3.835565
forward	1.348821	.9390513	1.44	0.151	-.4916863	3.189327
black	1.476842	.8171698	1.81	0.071	-.1247815	3.078465
marr	1.236043	.696663	1.77	0.076	-.1293911	2.601478
_cons	34.8283	6.771391	5.14	0.000	21.55662	48.09998
rebounds						
age	-.2818077	.1227409	-2.30	0.022	-.5223754	-.04124
exper	.830967	.1668415	4.98	0.000	.5039637	1.15797
expersq	-.0344878	.0090964	-3.79	0.000	-.0523165	-.0166591
coll	-.3689707	.1786866	-2.06	0.039	-.71919	-.0187514
guard	-2.727081	.4163206	-6.55	0.000	-3.543054	-1.911107
forward	.0896382	.4167631	0.22	0.830	-.7272024	.9064789
black	1.003824	.3626705	2.77	0.006	.2930028	1.714645
marr	-.2406585	.309188	-0.78	0.436	-.8466559	.3653389
_cons	10.87601	3.005231	3.62	0.000	4.985864	16.76615

assists							
age	-.3013925	.0850104	-3.55	0.000	-.4680097	-.1347752	
exper	.6633331	.1155545	5.74	0.000	.4368506	.8898157	
expersq	-.0222961	.0063002	-3.54	0.000	-.0346442	-.009948	
coll	-.1894703	.1237584	-1.53	0.126	-.4320323	.0530916	
guard	2.478626	.2883437	8.60	0.000	1.913482	3.043769	
forward	.4804238	.2886502	1.66	0.096	-.0853202	1.046168	
black	-.1528242	.2511857	-0.61	0.543	-.645139	.3394907	
marr	.4236511	.2141437	1.98	0.048	.0039371	.843365	
_cons	7.501437	2.081423	3.60	0.000	3.421922	11.58095	

b. The Stata test command gives

```
. test marr

( 1)  [points]marr = 0
( 2)  [rebounds]marr = 0
( 3)  [assists]marr = 0

      chi2( 3) =    12.02
Prob > chi2 =    0.0073
```

The rejection is very strong, presumably coming mainly from the points and assists equations. Rather than thinking being married causes a basketball player to be more productive, it could be that the more productive players – at least when it comes to points and assists – are more likely to be married.

**7.14.** Let  $\hat{\beta}$  be the estimator that uses  $\hat{\Omega}$  and let  $\check{\beta}$  be the estimator that uses  $\hat{\Lambda}$ . Because SGLS.1 to SGLS.3 hold,

$$\text{Avar}[\sqrt{N}(\hat{\beta} - \beta)] = [E(\mathbf{X}_i' \hat{\Omega}^{-1} \mathbf{X}_i)]^{-1}.$$

Further, we know from the general result for FGLS,

$$\text{Avar}[\sqrt{N}(\check{\beta} - \beta)] = [E(\mathbf{X}_i' \Lambda^{-1} \mathbf{X}_i)]^{-1} E(\mathbf{X}_i' \Lambda^{-1} \mathbf{u}_i \mathbf{u}_i' \Lambda^{-1} \mathbf{X}_i) [E(\mathbf{X}_i' \Lambda^{-1} \mathbf{X}_i)]^{-1}.$$

Now, because  $E(\mathbf{u}_i \mathbf{u}_i' | \mathbf{X}_i) = \Omega$ , it follows that

$$E(\mathbf{X}_i' \Lambda^{-1} \mathbf{u}_i \mathbf{u}_i' \Lambda^{-1} \mathbf{X}_i) = E(\mathbf{X}_i' \Lambda^{-1} \Omega \Lambda^{-1} \mathbf{X}_i)$$

by a simple iterated expectations argument. So, we have to show that

$$[E(\mathbf{X}_i' \mathbf{\Lambda}^{-1} \mathbf{X}_i)]^{-1} E(\mathbf{X}_i' \mathbf{\Lambda}^{-1} \mathbf{\Omega} \mathbf{\Lambda}^{-1} \mathbf{X}_i) [E(\mathbf{X}_i' \mathbf{\Lambda}^{-1} \mathbf{X}_i)]^{-1} - [E(\mathbf{X}_i' \mathbf{\Omega}^{-1} \mathbf{X}_i)]^{-1}$$

is positive semi-definite. We use the standard trick of showing that

$$E(\mathbf{X}_i' \mathbf{\Omega}^{-1} \mathbf{X}_i) - E(\mathbf{X}_i' \mathbf{\Lambda}^{-1} \mathbf{X}_i) [E(\mathbf{X}_i' \mathbf{\Lambda}^{-1} \mathbf{\Omega} \mathbf{\Lambda}^{-1} \mathbf{X}_i)]^{-1} E(\mathbf{X}_i' \mathbf{\Lambda}^{-1} \mathbf{X}_i)$$

is positive semi-definite. To this end, define  $\mathbf{Z}_i \equiv \mathbf{\Omega}^{-1/2} \mathbf{X}_i$  and  $\mathbf{W}_i \equiv \mathbf{\Omega}^{-1/2} \mathbf{\Lambda}^{-1} \mathbf{X}_i$ . Then

straightforward algebra shows that the difference above can be written as

$E(\mathbf{Z}_i' \mathbf{Z}_i) - E(\mathbf{Z}_i' \mathbf{W}_i) [E(\mathbf{W}_i' \mathbf{W}_i)]^{-1} E(\mathbf{W}_i' \mathbf{Z}_i)$  which is easily seen to be  $E(\mathbf{R}_i' \mathbf{R}_i)$ , where  $\mathbf{R}_i$  is the

$G \times K$  matrix of population residuals from the regression of  $\mathbf{Z}_i$  on  $\mathbf{W}_i$ :  $\mathbf{R}_i = \mathbf{Z}_i - \mathbf{W}_i \mathbf{\Pi}$  where

$\mathbf{\Pi} = [E(\mathbf{W}_i' \mathbf{W}_i)]^{-1} E(\mathbf{W}_i' \mathbf{Z}_i)$ . Matrices of the form  $E(\mathbf{R}_i' \mathbf{R}_i)$  are always positive semi-definite

because for a nonrandom vector  $\mathbf{a}$ ,  $\mathbf{a}' E(\mathbf{R}_i' \mathbf{R}_i) \mathbf{a} = E[(\mathbf{a} \mathbf{R}_i)' (\mathbf{a} \mathbf{R}_i)] \geq 0$ .

**7.15.** Let  $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}')'$  be the FGLS estimator from the full model. Then, because SGLS.1 through SGLS.3 hold, we know

$$\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})] = [E(\mathbf{W}_i' \mathbf{\Omega}^{-1} \mathbf{W}_i)]^{-1}$$

where  $\mathbf{W}_i = (\mathbf{X}_i, \mathbf{Z}_i)$ . Using partitioned matrix multiplication,

$$E(\mathbf{W}_i' \mathbf{\Omega}^{-1} \mathbf{W}_i) = \begin{pmatrix} E(\mathbf{X}_i' \mathbf{\Omega}^{-1} \mathbf{X}_i) & E(\mathbf{X}_i' \mathbf{\Omega}^{-1} \mathbf{Z}_i) \\ E(\mathbf{Z}_i' \mathbf{\Omega}^{-1} \mathbf{X}_i) & E(\mathbf{Z}_i' \mathbf{\Omega}^{-1} \mathbf{Z}_i) \end{pmatrix}.$$

Further, because  $E(\mathbf{X}_i \otimes \mathbf{Z}_i) = \mathbf{0}$ , it follows that  $E(\mathbf{X}_i' \mathbf{\Omega}^{-1} \mathbf{Z}_i) = \mathbf{0}$ . Therefore,  $E(\mathbf{W}_i' \mathbf{\Omega}^{-1} \mathbf{W}_i)$  is block diagonal and is equal to

$$\begin{pmatrix} E(\mathbf{X}_i' \mathbf{\Omega}^{-1} \mathbf{X}_i) & \mathbf{0} \\ \mathbf{0} & E(\mathbf{Z}_i' \mathbf{\Omega}^{-1} \mathbf{Z}_i) \end{pmatrix}$$

Inverting this matrix gives

$$\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})] = \begin{pmatrix} [\text{E}(\mathbf{X}_i' \boldsymbol{\Omega}^{-1} \mathbf{X}_i)]^{-1} & \mathbf{0} \\ \mathbf{0} & [\text{E}(\mathbf{Z}_i' \boldsymbol{\Omega}^{-1} \mathbf{Z}_i)]^{-1} \end{pmatrix}$$

and  $\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$  is the upper left hand block:

$$\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = [\text{E}(\mathbf{X}_i' \boldsymbol{\Omega}^{-1} \mathbf{X}_i)]^{-1}.$$

Now let  $\tilde{\boldsymbol{\beta}}$  be the FGLS estimator from  $\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{v}_i$ . We know this estimator is consistent for  $\boldsymbol{\beta}$  because  $\mathbf{v}_i = \mathbf{Z}_i \boldsymbol{\gamma} + \mathbf{u}_i$ , and so

$$\text{E}(\mathbf{X}_i \otimes \mathbf{v}_i) = \mathbf{0}$$

because  $\text{E}(\mathbf{X}_i \otimes \mathbf{Z}_i) = \mathbf{0}$  and  $\text{E}(\mathbf{u}_i | \mathbf{X}_i, \mathbf{Z}_i) = \mathbf{0}$ . Now, FGLS of  $\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{v}_i$  using a consistent estimator of  $\boldsymbol{\Lambda} = \text{E}(\mathbf{v}_i \mathbf{v}_i')$  generally has asymptotic variance

$$[\text{E}(\mathbf{X}_i' \boldsymbol{\Lambda}^{-1} \mathbf{X}_i)]^{-1} \text{E}(\mathbf{X}_i' \boldsymbol{\Lambda}^{-1} \mathbf{v}_i \mathbf{v}_i' \boldsymbol{\Lambda}^{-1} \mathbf{X}_i) [\text{E}(\mathbf{X}_i' \boldsymbol{\Lambda}^{-1} \mathbf{X}_i)]^{-1}$$

Let  $\mathbf{r}_i = \mathbf{Z}_i \boldsymbol{\gamma}$  so that we can write

$$\mathbf{v}_i \mathbf{v}_i' = (\mathbf{r}_i + \mathbf{u}_i)(\mathbf{r}_i + \mathbf{u}_i)' = \mathbf{r}_i \mathbf{r}_i' + \mathbf{r}_i \mathbf{u}_i' + \mathbf{u}_i \mathbf{r}_i' + \mathbf{u}_i \mathbf{u}_i'.$$

Now  $\text{E}(\mathbf{r}_i \mathbf{u}_i' | \mathbf{X}_i) = \mathbf{0}$  because  $\text{E}(\mathbf{u}_i | \mathbf{X}_i, \mathbf{Z}_i) = \mathbf{0}$  and  $\mathbf{r}_i$  is a function of  $\mathbf{Z}_i$ . Therefore,

$$\text{E}(\mathbf{v}_i \mathbf{v}_i' | \mathbf{X}_i) = \text{E}(\mathbf{r}_i \mathbf{r}_i' | \mathbf{X}_i) + \text{E}(\mathbf{u}_i \mathbf{u}_i' | \mathbf{X}_i) = \text{E}(\mathbf{r}_i \mathbf{r}_i' | \mathbf{X}_i) + \boldsymbol{\Omega}.$$

Using iterated expectations,

$$\begin{aligned} \text{E}(\mathbf{X}_i' \boldsymbol{\Lambda}^{-1} \mathbf{v}_i \mathbf{v}_i' \boldsymbol{\Lambda}^{-1} \mathbf{X}_i) &= \text{E}[\text{E}(\mathbf{X}_i' \boldsymbol{\Lambda}^{-1} \mathbf{v}_i \mathbf{v}_i' \boldsymbol{\Lambda}^{-1} \mathbf{X}_i | \mathbf{X}_i)] = \text{E}[\mathbf{X}_i' \boldsymbol{\Lambda}^{-1} \text{E}(\mathbf{v}_i \mathbf{v}_i' | \mathbf{X}_i) \boldsymbol{\Lambda}^{-1} \mathbf{X}_i] \\ &= \text{E}[\mathbf{X}_i' \boldsymbol{\Lambda}^{-1} \text{E}(\mathbf{r}_i \mathbf{r}_i' | \mathbf{X}_i) \boldsymbol{\Lambda}^{-1} \mathbf{X}_i] + \text{E}(\mathbf{X}_i' \boldsymbol{\Lambda}^{-1} \boldsymbol{\Omega} \boldsymbol{\Lambda}^{-1} \mathbf{X}_i) \\ &= \text{E}(\mathbf{X}_i' \boldsymbol{\Lambda}^{-1} \mathbf{r}_i \mathbf{r}_i' \boldsymbol{\Lambda}^{-1} \mathbf{X}_i) + \text{E}(\mathbf{X}_i' \boldsymbol{\Lambda}^{-1} \boldsymbol{\Omega} \boldsymbol{\Lambda}^{-1} \mathbf{X}_i) \end{aligned}$$

We have shown that

$$\begin{aligned} \text{Avar}[\sqrt{N}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})] &= \mathbf{A}_2^{-1} \{ \text{E}(\mathbf{X}_i' \boldsymbol{\Lambda}^{-1} \mathbf{r}_i \mathbf{r}_i' \boldsymbol{\Lambda}^{-1} \mathbf{X}_i) + \text{E}(\mathbf{X}_i' \boldsymbol{\Lambda}^{-1} \boldsymbol{\Omega} \boldsymbol{\Lambda}^{-1} \mathbf{X}_i) \} \mathbf{A}_2^{-1} \\ &= \mathbf{A}_2^{-1} \text{E}(\mathbf{X}_i' \boldsymbol{\Lambda}^{-1} \mathbf{r}_i \mathbf{r}_i' \boldsymbol{\Lambda}^{-1} \mathbf{X}_i) \mathbf{A}_2^{-1} + \mathbf{A}_2^{-1} \text{E}(\mathbf{X}_i' \boldsymbol{\Lambda}^{-1} \boldsymbol{\Omega} \boldsymbol{\Lambda}^{-1} \mathbf{X}_i) \mathbf{A}_2^{-1} \\ &\equiv \mathbf{C}_2 + \mathbf{A}_2^{-1} \text{E}(\mathbf{X}_i' \boldsymbol{\Lambda}^{-1} \boldsymbol{\Omega} \boldsymbol{\Lambda}^{-1} \mathbf{X}_i) \mathbf{A}_2^{-1} \end{aligned}$$

where  $\mathbf{A}_2 \equiv \mathbf{E}(\mathbf{X}_i' \mathbf{\Lambda}^{-1} \mathbf{X}_i)$  and  $\mathbf{C}_2 \equiv \mathbf{A}_2^{-1} \mathbf{E}(\mathbf{X}_i' \mathbf{\Lambda}^{-1} \mathbf{r}_i \mathbf{r}_i' \mathbf{\Lambda}^{-1} \mathbf{X}_i) \mathbf{A}_2^{-1}$ . Note that  $\mathbf{C}_2$  is positive semi-definite.

Now, Problem 7.14 established that

$$\mathbf{A}_2^{-1} \mathbf{E}(\mathbf{X}_i' \mathbf{\Lambda}^{-1} \mathbf{\Omega} \mathbf{\Lambda}^{-1} \mathbf{X}_i) \mathbf{A}_2^{-1} - \mathbf{A}_1^{-1}$$

is positive semi-definite. Therefore,

$$\text{Avar}[\sqrt{N}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})] - \text{Avar}[\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \mathbf{C}_2 + [\mathbf{A}_2^{-1} \mathbf{E}(\mathbf{X}_i' \mathbf{\Lambda}^{-1} \mathbf{\Omega} \mathbf{\Lambda}^{-1} \mathbf{X}_i) \mathbf{A}_2^{-1} - \mathbf{A}_1^{-1}]$$

and each matrix is positive-semi-definite. We have shown the result.

Interestingly, the proof shows that the asymptotic inefficiency of  $\tilde{\boldsymbol{\beta}}$  has two sources. First, we have omitted variables that are uncorrelated with  $\mathbf{X}_i$ . The second piece is due to using the wrong variance matrix,  $\mathbf{\Lambda}$ . If we could effectively use  $\mathbf{\Omega}$  in obtaining the estimator with  $\mathbf{Z}_i$  omitted – which we can in principle if we observe  $\mathbf{Z}_i$  – then the only source of inefficiency would be due to omitting  $\mathbf{Z}_i$  (as happens in the single-equation case).

## Solutions to Chapter 8 Problems

**8.1.** Letting  $Q(\mathbf{b})$  denote the objective function in equation (8.27), it follows from multivariable calculus that

$$\frac{\partial Q(\mathbf{b})'}{\partial \mathbf{b}} = -2 \left( \sum_{i=1}^N \mathbf{Z}_i' \mathbf{X}_i \right)' \hat{\mathbf{W}} \left( \sum_{i=1}^N \mathbf{Z}_i' (\mathbf{y}_i - \mathbf{X}_i \mathbf{b}) \right).$$

Evaluating the derivative at the solution  $\hat{\boldsymbol{\beta}}$  gives

$$\left( \sum_{i=1}^N \mathbf{Z}_i' \mathbf{X}_i \right)' \hat{\mathbf{W}} \left( \sum_{i=1}^N \mathbf{Z}_i' (\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}) \right) = \mathbf{0}.$$

In terms of full data matrices, we can write, after simple algebra,

$$(\mathbf{X}' \mathbf{Z} \hat{\mathbf{W}} \mathbf{Z}' \mathbf{X}) \hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{Z} \hat{\mathbf{W}} \mathbf{Z}' \mathbf{Y}).$$

Solving for  $\hat{\boldsymbol{\beta}}$  gives (8.28).

**8.2. a.** We can apply general GMM theory to obtain consistency and  $\sqrt{N}$  asymptotic normality of the 3SLS estimator (GMM version). The four assumptions given in the problem are sufficient for SIV.1 to SIV.3, where  $\hat{\mathbf{W}} = \left( N^{-1} \sum_{i=1}^N \mathbf{Z}_i' \hat{\boldsymbol{\Omega}} \mathbf{Z}_i \right)^{-1}$  and  $\mathbf{W} \equiv [E(\mathbf{Z}_i' \boldsymbol{\Omega} \mathbf{Z}_i)]^{-1} = \text{plim}(\hat{\mathbf{W}})$ . (This assumes  $\text{plim} \hat{\boldsymbol{\Omega}} = \boldsymbol{\Omega} \equiv E(\mathbf{u}_i \mathbf{u}_i')$ , something that holds quite generally.) However, without SIV.5, 3SLS is not necessarily an asymptotically efficient GMM estimator.

b. The asymptotic variance of the 3SLS estimator is given in equation (8.29) with the choice of  $\mathbf{W}$  in part a:

$$\text{Avar} \sqrt{N} (\hat{\boldsymbol{\beta}}_{3SLS} - \boldsymbol{\beta}) = (\mathbf{C}' \mathbf{W} \mathbf{C})^{-1} (\mathbf{C}' \mathbf{W} \boldsymbol{\Lambda} \mathbf{W} \mathbf{C}) (\mathbf{C}' \mathbf{W} \mathbf{C})^{-1},$$

where  $\boldsymbol{\Lambda} \equiv E(\mathbf{Z}_i' \mathbf{u}_i \mathbf{u}_i' \mathbf{Z}_i)$ , as in the text. (Note this expression collapses to  $(\mathbf{C}' \mathbf{W} \mathbf{C})^{-1}$  when



$\Lambda = \mathbf{W}^{-1}$ , as happens under SIV.5.)

c. A consistent estimator of  $\text{Avar} \sqrt{N} (\hat{\beta}_{3SLS} - \beta)$  is given in equation (8.31) with

$\hat{\Lambda} \equiv N^{-1} \sum_{i=1}^N \mathbf{Z}_i' \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \mathbf{Z}_i$  and  $\hat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i \hat{\beta}_{3SLS}$  the 3SLS residuals:

$$[(\mathbf{X}'\mathbf{Z}/N)\hat{\mathbf{W}}(\mathbf{Z}'\mathbf{X}/N)]^{-1}(\mathbf{X}'\mathbf{Z}/N)\hat{\mathbf{W}}\left(N^{-1}\sum_{i=1}^N \mathbf{Z}_i' \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \mathbf{Z}_i\right)\hat{\mathbf{W}}(\mathbf{Z}'\mathbf{X}/N)[(\mathbf{X}'\mathbf{Z}/N)\hat{\mathbf{W}}(\mathbf{Z}'\mathbf{X}/N)]^{-1}.$$

The estimator of  $\text{Avar}(\hat{\beta}_{3SLS})$  is simply this expression divided by  $N$ . Even though the formula looks complicated, it can be programmed fairly easily in a matrix-based language. Of course, if we doubt SIV.5 in the first place, we would probably use the more general minimum chi-square estimator, as it is asymptotically more efficient. (If we were going to obtain the robust variance matrix estimate for 3SLS anyway, it is no harder to obtain the minimum chi-square estimate and its asymptotic variance estimate.)

**8.3.** First, we can always write  $\mathbf{x}$  as its linear projection plus an error:  $\mathbf{x} = \mathbf{x}^* + \mathbf{e}$ , where  $\mathbf{x}^* = \mathbf{z}\Pi$  and  $E(\mathbf{z}'\mathbf{e}) = \mathbf{0}$ . Therefore,  $E(\mathbf{z}'\mathbf{x}) = E(\mathbf{z}'\mathbf{x}^*)$ , which verifies the first part of the hint. To verify the second step, let  $\mathbf{h} \equiv \mathbf{h}(\mathbf{z})$ , and write the linear projection as

$$\mathbf{L}(\mathbf{y}|\mathbf{z}, \mathbf{h}) = \mathbf{z}\Pi_1 + \mathbf{h}\Pi_2$$

where  $\Pi_1$  is  $M \times K$  and  $\Pi_2$  is  $Q \times K$ . Then we must show that  $\Pi_2 = \mathbf{0}$ . But, from the two-step projection theorem (see Property LP.7 in Chapter 2),

$$\Pi_2 = [E(\mathbf{s}'\mathbf{s})]^{-1}E(\mathbf{s}'\mathbf{r}), \text{ where } \mathbf{s} \equiv \mathbf{h} - \mathbf{L}(\mathbf{h}|\mathbf{z}) \text{ and } \mathbf{r} \equiv \mathbf{x} - \mathbf{L}(\mathbf{x}|\mathbf{z}).$$

Now, by the assumption that  $E(\mathbf{x}|\mathbf{z}) = \mathbf{L}(\mathbf{x}|\mathbf{z})$ ,  $\mathbf{r}$  is also equal to  $\mathbf{x} - E(\mathbf{x}|\mathbf{z})$ . Therefore,  $E(\mathbf{r}|\mathbf{z}) = \mathbf{0}$ , and so  $\mathbf{r}$  is uncorrelated with all functions of  $\mathbf{z}$ . But  $\mathbf{s}$  is simply a function of  $\mathbf{z}$  since  $\mathbf{h} \equiv \mathbf{h}(\mathbf{z})$ . Therefore,  $E(\mathbf{s}'\mathbf{r}) = \mathbf{0}$ , and this shows that  $\Pi_2 = \mathbf{0}$ .

**8.4.a.** For the system in (8.12), we show that, for each  $g$ ,  $\text{rank } E[(\mathbf{z}, \mathbf{h})'\mathbf{x}_g] = \text{rank } E(\mathbf{z}'\mathbf{x}_g)$

for any function  $\mathbf{h}=\mathbf{h}(\mathbf{z})$ . Now, by Problem 8.3,  $L(\mathbf{x}_g|\mathbf{z}, \mathbf{h}) = L(\mathbf{x}_g|\mathbf{z}) = \mathbf{z}\mathbf{\Pi}_1$  when  $E(\mathbf{x}_g|\mathbf{z})$  is linear in  $\mathbf{z}$  and  $\mathbf{h}$  is any function of  $\mathbf{z}$ . As in Problem 8.3,  $E(\mathbf{z}'\mathbf{x}_g) = E(\mathbf{z}'\mathbf{x}_g^*) = E(\mathbf{z}'\mathbf{z})\mathbf{\Pi}_1$ . Also, if we let  $\mathbf{e}_g=\mathbf{x}_g - \mathbf{x}_g^*$ , then  $E(\mathbf{h}'\mathbf{e}_g) = \mathbf{0}$ , and so  $E[(\mathbf{z}, \mathbf{h})'\mathbf{x}_g] = E[(\mathbf{z}, \mathbf{h})'\mathbf{x}_g^*] = E[(\mathbf{z}, \mathbf{h})'\mathbf{z}]\mathbf{\Pi}_1$ . But  $\text{rank } E[(\mathbf{z}, \mathbf{h})'\mathbf{z}] = \text{rank } E(\mathbf{z}'\mathbf{z})$ , which means that  $\text{rank } E[(\mathbf{z}, \mathbf{h})'\mathbf{z}]\mathbf{\Pi}_1 = \text{rank } E[(\mathbf{z}'\mathbf{z})\mathbf{\Pi}_1]$ . We have shown that  $\text{rank } E[(\mathbf{z}, \mathbf{h})'\mathbf{x}_g] = \text{rank } E[(\mathbf{z}'\mathbf{x}_g)]$ , which means adding  $\mathbf{h}$  to the instrument list does not help satisfy the rank condition.

b. If  $E(\mathbf{x}_g|\mathbf{z})$  is nonlinear in  $\mathbf{z}$ , then  $L(\mathbf{x}_g|\mathbf{z}, \mathbf{h})$  will generally depend on  $\mathbf{h}$ . This can certainly help in satisfying the rank condition. For example, if  $K_g < M$  (the dimension of  $\mathbf{z}$ ) then the order condition fails for equating  $g$  using instruments  $\mathbf{z}$ . But we can add nonlinear functions of  $\mathbf{z}$  to the instrument list that are partially correlated with  $\mathbf{x}_g$  and satisfy the order and rank condition. We use this fact in Section 9.5.

**8.5.** This follows directly from the hint. Straightforward matrix algebra shows that  $(\mathbf{C}'\mathbf{\Lambda}^{-1}\mathbf{C})-(\mathbf{C}'\mathbf{W}\mathbf{C})(\mathbf{C}'\mathbf{W}\mathbf{A}\mathbf{W}\mathbf{C})^{-1}(\mathbf{C}'\mathbf{W}\mathbf{C})$  can be written as

$$\mathbf{C}'\mathbf{\Lambda}^{-1/2}[\mathbf{I}_L - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}']\mathbf{\Lambda}^{-1/2}\mathbf{C},$$

where  $\mathbf{D} \equiv \mathbf{\Lambda}^{1/2}\mathbf{W}\mathbf{C}$ . Since this is a matrix quadratic form in the  $L \times L$  symmetric, idempotent matrix  $\mathbf{I}_L - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$ , it is necessarily itself positive semi-definite.

**8.6. a.** First,  $\mathbf{\Omega}^{-1}\mathbf{u}_i = (\sigma^{11}u_{i1} + \sigma^{12}u_{i2}, \sigma^{12}u_{i1} + \sigma^{22}u_{i2})'$ . Therefore,

$$\begin{aligned} \mathbf{Z}_i'\mathbf{\Omega}^{-1}\mathbf{u}_i &= \begin{pmatrix} \mathbf{z}_{i1}' & 0 \\ 0 & \mathbf{z}_{i2}' \end{pmatrix} (\sigma^{11}u_{i1} + \sigma^{12}u_{i2}, \sigma^{12}u_{i1} + \sigma^{22}u_{i2})' \\ &= \begin{pmatrix} \mathbf{z}_{i1}'(\sigma^{11}u_{i1} + \sigma^{12}u_{i2}) \\ \mathbf{z}_{i2}'(\sigma^{12}u_{i1} + \sigma^{22}u_{i2}) \end{pmatrix}. \end{aligned}$$

The expected value of this vector depends on  $E(\mathbf{z}_{i1}'u_{i1})$ ,  $E(\mathbf{z}_{i1}'u_{i2})$ ,  $E(\mathbf{z}_{i2}'u_{i1})$ , and  $E(\mathbf{z}_{i2}'u_{i2})$ . If

$E(\mathbf{z}'_{i1}u_{i2}) \neq \mathbf{0}$  or  $E(\mathbf{z}'_{i2}u_{i1}) \neq \mathbf{0}$  then  $E(\mathbf{Z}'_i\boldsymbol{\Omega}^{-1}\mathbf{u}_i) \neq \mathbf{0}$  except by fluke. In fact, if  $E(\mathbf{z}'_{i1}u_{i1}) = \mathbf{0}$ ,  $E(\mathbf{z}'_{i2}u_{i2}) = \mathbf{0}$ , and  $\sigma^{12} \neq 0$  then  $E(\mathbf{Z}'_i\boldsymbol{\Omega}^{-1}\mathbf{u}_i) \neq \mathbf{0}$  if  $E(\mathbf{z}'_{i1}u_{i2}) \neq \mathbf{0}$  or  $E(\mathbf{z}'_{i2}u_{i1}) \neq \mathbf{0}$ .

b. When  $\sigma_{12} = 0, \sigma^{12} = 0$ , in which case  $E(\mathbf{z}'_{i1}u_{i1}) = \mathbf{0}$  and  $E(\mathbf{z}'_{i2}u_{i2}) = \mathbf{0}$  imply  $E(\mathbf{Z}'_i\boldsymbol{\Omega}^{-1}\mathbf{u}_i) = \mathbf{0}$ .

c. If the same instruments are valid in each equation – so  $E(\mathbf{z}'_i u_{i1}) = E(\mathbf{z}'_i u_{i2}) = \mathbf{0}$  – then  $E(\mathbf{Z}'_i\boldsymbol{\Omega}^{-1}\mathbf{u}_i) = \mathbf{0}$  without restrictions on  $\boldsymbol{\Omega}$ .

**8.7.** When  $\hat{\boldsymbol{\Omega}}$  is diagonal and  $\mathbf{z}_i$  has the form in (8.15),  $\sum_{i=1}^N \mathbf{z}'_i \hat{\boldsymbol{\Omega}} \mathbf{z}_i = \mathbf{Z}'(\mathbf{I}_N \otimes \hat{\boldsymbol{\Omega}})\mathbf{Z}$  is a block diagonal matrix with  $g^{th}$  block  $\hat{\sigma}_g^2 \left( \sum_{i=1}^N \mathbf{z}'_{ig} \mathbf{z}_{ig} \right) \equiv \hat{\sigma}_g^2 \mathbf{z}'_g \mathbf{z}_g$ , where  $\mathbf{Z}_g$  denotes the  $N \times L_g$  observation matrix of instruments for the  $g^{th}$  equation. Further,  $\mathbf{Z}'\mathbf{X}$  is block diagonal with  $g^{th}$  block  $\mathbf{Z}'_g \mathbf{X}_g$ . Using these facts, it is now straightforward to show that the 3SLS estimator consists of  $[\mathbf{X}'_g \mathbf{Z}_g (\mathbf{Z}'_g \mathbf{Z}_g)^{-1} \mathbf{Z}'_g \mathbf{X}_g]^{-1} \mathbf{X}'_g \mathbf{Z}_g (\mathbf{Z}'_g \mathbf{Z}_g)^{-1} \mathbf{Z}'_g \mathbf{Y}_g$  stacked from  $g = 1, \dots, G$ . This is just the system 2SLS estimator or, equivalently, 2SLS equation-by-equation.

**8.8. a.** With  $\mathbf{Z}_1 = (\mathbf{z}'_{i1}, \mathbf{z}'_{i2}, \dots, \mathbf{z}'_{iT})'$  and  $\mathbf{X}_i = (\mathbf{x}'_{i1}, \mathbf{x}'_{i2}, \dots, \mathbf{x}'_{iT})'$ ,

$$\mathbf{Z}'_i \mathbf{Z}_i = \sum_{t=1}^T \mathbf{z}'_{it} \mathbf{z}_{it}, \quad \mathbf{Z}'_i \mathbf{X}_i = \sum_{t=1}^T \mathbf{z}'_{it} \mathbf{x}_{it}, \quad \text{and} \quad \mathbf{Z}'_i \mathbf{y}_i = \sum_{t=1}^T \mathbf{z}'_{it} y_{it}.$$

Summing over all  $i$  gives

$$\mathbf{Z}'\mathbf{Z} = \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}'_{it} \mathbf{z}_{it}, \quad \mathbf{Z}'\mathbf{X} = \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}'_{it} \mathbf{x}_{it}, \quad \text{and} \quad \mathbf{Z}'\mathbf{Y} = \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}'_{it} y_{it}.$$

b.  $\text{rank } E\left(\sum_{t=1}^T \mathbf{z}'_{it} \mathbf{x}_{it}\right) = K$ .

c. Let  $\hat{\mathbf{u}}_i$  be the  $T \times 1$  vector of pooled 2SLS residuals,  $\hat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}$ . Then we just use (8.31) with  $\hat{\mathbf{W}} = (\mathbf{Z}'\mathbf{Z}/N)^{-1}$  and  $\hat{\boldsymbol{\Lambda}} = N^{-1} \sum_{i=1}^N \mathbf{Z}'_i \hat{\mathbf{u}}_i \hat{\mathbf{u}}'_i \mathbf{Z}_i$ , cancelling  $N$  everywhere:

$$[(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\left(\sum_{i=1}^N \mathbf{Z}'_i \hat{\mathbf{u}}_i \hat{\mathbf{u}}'_i \mathbf{Z}_i\right) \cdot (\mathbf{Z}'\mathbf{Z}^{-1})(\mathbf{Z}'\mathbf{X})[(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X})]^{-1}. \quad (8.67)$$

d. Using reasoning almost identical to Problem 7.7, (8.65) implies that, for  $s < t$ ,

$$\begin{aligned} E(u_{it}u_{is}\mathbf{z}'_{it}\mathbf{z}_{is}) &= E[E(u_{it}u_{is}\mathbf{z}'_{it}\mathbf{z}_{is}|\mathbf{z}_{it}, u_{is}, \mathbf{z}_{is})] \\ &= E[E(u_{it}|\mathbf{z}'_{it}, u_{is}, \mathbf{z}_{is})u_{is}, \mathbf{z}'_{it}, \mathbf{z}_{is}] \\ &= E[0 \cdot u_{is}\mathbf{z}'_{it}, \mathbf{z}_{is}] = \mathbf{0} \end{aligned}$$

because  $E(u_{it}|\mathbf{z}_{it}, u_{is}, \mathbf{z}_{is}) = 0$  for  $s < t$ . A similar argument works for  $t > s$ . So for all  $t \neq s$ ,

$$E(u_{it}u_{is}\mathbf{z}'_{it}, \mathbf{z}_{is}) = 0.$$

Similarly, (8.66) and iterated expectations implies that

$$\begin{aligned} E(u_{it}^2\mathbf{z}'_{it}\mathbf{z}_{it}) &= E[E(u_{it}^2\mathbf{z}'_{it}\mathbf{z}_{it}|\mathbf{z}_{it})] \\ &= E[E(u_{it}^2|\mathbf{z}_{it})\mathbf{z}'_{it}\mathbf{z}_{it}] = \sigma^2 E[(\mathbf{z}'_{it}\mathbf{z}_{it}), t = 1, \dots, T]. \end{aligned}$$

Together, these results imply that

$$\text{Var}(\mathbf{z}'_i \mathbf{u}_i) = \sigma^2 \sum_{t=1}^T E[(\mathbf{z}'_{it}\mathbf{z}_{it})].$$

A consistent estimator of this matrix is  $\hat{\sigma}^2(\mathbf{Z}'\mathbf{Z}/N)$ , where  $\hat{\sigma}^2 = 1/(NT) \sum_{i=1}^N \sum_{t=1}^T \hat{u}_{it}^2$ , by the usual law-of-large-numbers arguments. A degrees of freedom adjustment replaces  $NT$  with  $NT - K$ . Replacing  $\sum_{i=1}^N \mathbf{Z}'_i \hat{\mathbf{u}}_i \hat{\mathbf{u}}'_i \mathbf{Z}_i$  in (8.67) with  $\hat{\sigma}^2(\mathbf{Z}'\mathbf{Z})$  [since  $\hat{\sigma}^2(\mathbf{Z}'\mathbf{Z}/N)$  can play the role of  $\hat{\mathbf{\Lambda}}$  under the maintained assumptions] and cancelling gives the estimated asymptotic variance of  $\hat{\beta}$  as

$$\hat{\sigma}^2[(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X})]^{-1}.$$

This is exactly the variance estimator that would be computed from the pooled 2SLS estimation. This means that the usual 2SLS standard errors and test statistics are

asymptotically valid.

e. If the unconditional variance changes across  $t$ , the simplest approach is to weight the variables in each time period by  $1/\hat{\sigma}_t$ , where  $\hat{\sigma}_t^2$  is a consistent estimator of  $\sigma_t^2 = \text{Var}(u_{it})$ . A consistent estimator of  $\hat{\sigma}_t^2$  is

$$\hat{\sigma}_t^2 = N^{-1} \sum_{i=1}^N \hat{u}_{it}^2.$$

Now, apply pooled 2SLS to the equation

$$(y_{it}/\hat{\sigma}_t) = (\mathbf{x}_{it}/\hat{\sigma}_t)\boldsymbol{\beta} + \text{error}_{it}$$

using instruments  $\mathbf{z}_{it}/\hat{\sigma}_t$ . The usual statistics from this procedure are asymptotically valid: it can be shown that it has the same  $\sqrt{N}$ -asymptotic distribution as if we knew the  $\sigma_t^2$ . This estimator is a generalized instrumental variables (GIV) estimator except it is consistent under the contemporaneous exogeneity assumption only. It turns out to be identical to the GMM estimator that uses weighting matrix  $\left(N^{-1} \sum_{i=1}^N \sum_{t=1}^T \hat{\sigma}_t^2 \mathbf{z}_{it}' \mathbf{z}_{it}\right)^{-1}$  – the optimal weighting matrix under the assumptions in the problem. See Im, Ahn, Schmidt, and Wooldridge (1999, Section 2) for discussion of a more general result.

**8.9** The optimal instruments are given in Theorem 8.5, with  $G = 1$ :

$$\mathbf{z}_i^* = [\omega(\mathbf{z}_i)]^{-1} \text{E}(\mathbf{x}_i|\mathbf{z}_i), \quad \omega(\mathbf{z}_i) = \text{E}(u_i^2|\mathbf{z}_i).$$

If  $\text{E}(u_i^2|\mathbf{z}_i) = \sigma^2$  and  $\text{E}(\mathbf{x}_i|\mathbf{z}_i) = \mathbf{z}_i\boldsymbol{\Pi}$ , then the optimal instruments are  $\sigma^{-2} \mathbf{z}_i\boldsymbol{\Pi}$ . The constant multiple  $\sigma^{-2}$  clearly has no effect on the optimal IV estimator, so the optimal instruments are  $\mathbf{z}_i\boldsymbol{\Pi}$ . These are the optimal IVs underlying 2SLS, except that  $\boldsymbol{\Pi}$  is replaced with its  $\sqrt{N}$ -consistent OLS estimator. The 2SLS estimator has the same asymptotic variance whether  $\boldsymbol{\Pi}$  or  $\hat{\boldsymbol{\Pi}}$  is used, and so 2SLS is asymptotically efficient.

If  $E(u|\mathbf{x}) = 0$  and  $E(u^2|\mathbf{x}) = \sigma^2$ , the optimal instruments are  $\sigma^{-2} E(\mathbf{x}|\mathbf{x}) = \sigma^{-2} \mathbf{x}$ , and this leads to the OLS estimator.

**8.10.a.** Write  $u_{it} = \rho u_{i,t-1} + e_{it}$ , and plug into  $y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + u_{it}$  to get

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \rho u_{i,t-1} + e_{it}, t = 2, \dots, T.$$

Under the assumption

$$E(u_{it}|\mathbf{z}_{it}, u_{it,t-1}, \mathbf{x}_{i,t-1}, \mathbf{z}_{i,t-1}, \mathbf{x}_{i,t-2}, \dots, u_{i1}, \mathbf{x}_{i1}, \mathbf{z}_{i1}) = 0 \quad (8.68)$$

the previous assumption satisfies the dynamic completeness assumption when  $\rho = 0$ . If we assume that  $E(u_{it}^2|\mathbf{z}_{it}, u_{i,t-1})$  is constant under  $H_0$ , then it satisfies the requisite homoskedasticity assumption as well. As shown in Problem 8.8, pooled 2SLS estimation of this equation using instruments  $(\mathbf{z}_{it}, u_{i,t-1})$  results in valid test statistics.

Now we apply the results from Section 6.1.3: when  $\rho = 0$ , replacing  $u_{i,t-1}$  with the initial 2SLS residuals  $\hat{u}_{i,t-1}$  has no effect as  $N$  gets large, provided that (8.68) holds. Thus, we can estimate

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \rho \hat{u}_{i,t-1} + \text{error}_{it}, t = 2, \dots, T,$$

by pooled 2SLS using instruments  $(\mathbf{z}_{it}, \hat{u}_{i,t-1})$ , and obtain the usual  $t$  statistic for  $\hat{\rho}$ .

b. If  $E(u_{it}^2|\mathbf{z}_{it}, u_{i,t-1})$  is not constant, we can use the usual heteroskedasticity-robust  $t$  statistic from pooled 2SLS for  $\hat{\rho}$ . This allows for dynamic forms of heteroskedasticity, such as ARCH and GARCH, as well as static forms of heteroskedasticity.

**8.11. a.** This is a simple application of Theorem 8.5 when  $G = 1$ . Without the  $i$  subscript,  $\mathbf{x}_1 = (\mathbf{z}_1, y_2)$  and so  $E(\mathbf{x}_1|\mathbf{z}) = [\mathbf{z}_1, E(y_2|\mathbf{z})]$ . Further,  $\Omega(\mathbf{z}) = \text{Var}(u_1|\mathbf{z}) = \sigma_1^2$ . It follows that the optimal instruments are  $(1/\sigma_1^2)[\mathbf{z}_1, E(y_2|\mathbf{z})]$ . Dropping the division by  $\sigma_1^2$  clearly does not affect the optimal instruments.

b. If  $y_2$  is binary then  $E(y_2|\mathbf{z}) = P(y_2 = 1|\mathbf{z}) = F(\mathbf{z})$ , and so the optimal IVs are  $[\mathbf{z}_1, F(\mathbf{z})]$ .

**8.12.** a. As long as  $E(\mathbf{Z}'\mathbf{u}) = \mathbf{0}$  holds the estimator is consistent. After all, it is a GMM estimator with a particular weighting matrix that satisfies all of the GMM regularity conditions.

b. Unless the optimal weighting matrix  $\hat{\mathbf{W}}$  consistently estimates  $[\text{Var}(\mathbf{Z}'_i\mathbf{u}_i)]^{-1}$ , the statistic fails to be asymptotically chi-square.

c. Since  $\hat{\mathbf{\Omega}}$  and  $\hat{\mathbf{\Lambda}}$  converge to the same constant matrix  $\mathbf{\Lambda} = \mathbf{\Omega}$ , there is no difference in asymptotic efficiency (at least using the usual  $\sqrt{N}$ -asymptotic distribution).

**8.13.** a. The optimal instrumental variable is

$$\mathbf{z}^* = [E(u_1^2|\mathbf{z})]^{-1} \cdot E[\mathbf{z}_1, y_2, \mathbf{z}_1 y_2 | \mathbf{z}] = (\sigma_1^2)^{-1} [\mathbf{z}_1, E(y_2|\mathbf{z}), \mathbf{z}_1 E(y_2|\mathbf{z})] = (\sigma_1^2)^{-1} [\mathbf{z}_1, \mathbf{z}\pi_2, \mathbf{z}_1(\mathbf{z}\pi_2)].$$

b. The coefficients  $\pi_2$  can be estimated by running an OLS of  $y_2$  on  $\mathbf{z}$ . Since the inverse of the variance is a scalar that does not depend on  $\mathbf{z}$ , it cancels out in the IV estimation. Thus we can operationalize the optimal IV estimator by using  $[\mathbf{z}_1, \mathbf{z}\hat{\pi}_2, \mathbf{z}_1(\mathbf{z}\hat{\pi}_2)]$  as the IVs. The estimator has the same  $\sqrt{N}$ -asymptotic distribution as if we knew  $\pi_2$ .

**8.14.** a. With  $\mathbf{y}_{it2} = \mathbf{z}_{it}\mathbf{\Pi}_2 + \mathbf{v}_{it2}$  and  $E(\mathbf{z}'_{it}u_{it1}) = \mathbf{0}, t = 1, \dots, T$  maintained,  $E(\mathbf{y}'_{it2}u_{it1}) = \mathbf{0}$  is the same as  $E(\mathbf{v}'_{it2}u_{it1}) = \mathbf{0}$ . We can always write the linear projection of  $u_{it1}$  onto  $\mathbf{v}_{it2}$  as

$$\begin{aligned} u_{it1} &= \mathbf{v}_{it2}\mathbf{p}_1 + e_{it1} \\ E(\mathbf{v}'_{it2}e_{it1}) &= \mathbf{0}, t = 1, \dots, T \end{aligned}$$

where we assume that the coefficients  $\mathbf{p}_1$  do not change over time. Thus, we can write the extended equation

$$y_{it1} = \eta_{t1} + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + \mathbf{y}_{it2}\boldsymbol{\alpha}_2 + \mathbf{v}_{it2}\mathbf{p}_1 + e_{it1}, t = 1, \dots, T$$

Now the control function procedure is clear. (1) Estimate the reduced form  $\mathbf{y}_{it2} = \mathbf{z}_{it}\mathbf{\Pi}_2 + \mathbf{v}_{it2}$

by pooled OLS (equation-by-equation if necessary when  $\mathbf{y}_{it2}$  is a vector) and obtain the residuals,  $\hat{\mathbf{v}}_{it2}$ . (2) Run the pooled OLS regression

$$y_{it1} \text{ on } 1, d2_t, \dots, dT_t, \mathbf{z}_{it1}, \mathbf{y}_{it2}, \hat{\mathbf{v}}_{it2}, t = 1, \dots, T; i = 1, \dots, N.$$

and use a fully robust Wald test of  $H_0 : \boldsymbol{\rho}_1 = \mathbf{0}$ . The test has  $G_1$  degrees of freedom in the chi-square distribution, or one can use an  $F$  approximation by dividing the chi-square statistic by  $G_1$ .

b. Extending the discussion in the text around equation (6.32), partition  $\mathbf{z}_{it2} = (\mathbf{g}_{it2}, \mathbf{h}_{it2})$  where  $\mathbf{g}_{it2}$  is  $1 \times G_1$  (the same dimension as  $\mathbf{y}_{it1}$ ) and  $\mathbf{h}_{it2}$  is  $1 \times Q_1$ . Obtain the fitted values  $\hat{\mathbf{y}}_{it2}$  from the first-stage regressions. Then, obtain the residuals,  $\hat{\mathbf{r}}_{it2}$  from the pooled OLS regression

$$\mathbf{h}_{it2} \text{ on } \mathbf{z}_{it1}, \hat{\mathbf{y}}_{it2}, t = 1, \dots, T; i = 1, \dots, N.$$

Let  $\hat{u}_{it1}$  be the P2SLS residuals. Then run the pooled OLS regression

$$\hat{u}_{it1} \text{ on } \hat{\mathbf{r}}_{it2}, t = 1, \dots, T; i = 1, \dots, N,$$

and test the  $\hat{\mathbf{r}}_{it2}$  for joint significance. A fully robust Wald test is most appropriate, and its limiting distribution under the null that all elements of  $\mathbf{z}_{it}$  are exogenous is  $\chi^2_{Q_1}$ .

**8.15.** a. The coefficient shows that a higher fare reduces passenger demand for flights. The estimated elasticity is  $-.565$ , which is fairly large. Even the fully robust 95% confidence interval is pretty narrow, from  $-.696$  to  $-.434$ . Incidentally, the standard error that is robust only to heteroskedasticity and not serial correlation is about  $.0364$ , which is actually slightly smaller than the usual OLS standard error. So it is important to use the fully robust version.

```
. use airfare
. xtset id year
    panel variable:  id (strongly balanced)
    time variable:  year, 1997 to 2000
                delta:  1 unit
```



```
. reg lpassen y98 y99 y00 lfare ldist ldistsq
```

Source	SS	df	MS	Number of obs = 4596		
Model	230.557732	6	38.4262887	F( 6, 4589) = 52.48		
Residual	3360.12968	4589	.732213921	Prob > F = 0.0000		
Total	3590.68741	4595	.781433605	R-squared = 0.0642		
				Adj R-squared = 0.0630		
				Root MSE = .85569		

lpassen	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
y98	.0321212	.0357118	0.90	0.368	-.0378911	.1021335
y99	.081651	.035724	2.29	0.022	.0116148	.1516873
y00	.1380369	.0358761	3.85	0.000	.0677024	.2083713
lfare	-.5647711	.0369644	-15.28	0.000	-.6372392	-.4923031
ldist	-1.54939	.3265076	-4.75	0.000	-2.189502	-.9092778
ldistsq	.1227088	.0247935	4.95	0.000	.0741017	.171316
_cons	13.65144	1.094166	12.48	0.000	11.50635	15.79653

```
. reg lpassen y98 y99 y00 lfare ldist ldistsq, cluster(id)
```

Linear regression

Number of obs = 4596  
F( 6, 1148) = 34.95  
Prob > F = 0.0000  
R-squared = 0.0642  
Root MSE = .85569

(Std. Err. adjusted for 1149 clusters in id)

lpassen	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
y98	.0321212	.0050262	6.39	0.000	.0222597	.0419827
y99	.081651	.0073679	11.08	0.000	.0671949	.0961072
y00	.1380369	.0104857	13.16	0.000	.1174636	.1586101
lfare	-.5647711	.0667107	-8.47	0.000	-.6956597	-.4338826
ldist	-1.54939	.69818	-2.22	0.027	-2.919242	-.179538
ldistsq	.1227088	.0524034	2.34	0.019	.0198916	.2255261
_cons	13.65144	2.316661	5.89	0.000	9.106074	18.1968

b. I use the test that allows the explanatory variables to be non-strictly exogenous. The estimate of  $\rho$  is essentially one. In a pure time series context we would have to worry how this amount of persistence in the errors affects inference. Here, inference is standard because it is with fixed  $T$  and  $N \rightarrow \infty$ . But the “unit root” in  $\{u_{it} : t = 1, \dots, T\}$  is of some concern because it calls into question whether there is a meaningful relationship between passenger demand and airfares. If the error term rarely returns to its mean (which we can take to be zero), in what

sense is do movements in airfare over time cause movements in passenger demand?

```
. predict uhat, resid

. gen uhat_1 = 1.uhat
(1149 missing values generated)

. reg lpassen y99 y00 lfare ldist ldistsq uhat_1, robust
```

```
Linear regression                                Number of obs =      3447
                                                F(   6,   3440) =  7168.14
                                                Prob > F       =   0.0000
                                                R-squared      =   0.9647
                                                Root MSE      =   .1684
```

lpassen	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
y99	.0502195	.0065875	7.62	0.000	.0373036	.0631354
y00	.1105098	.0072252	15.30	0.000	.0963437	.124676
lfare	-.628955	.0095767	-65.68	0.000	-.6477315	-.6101784
ldist	-1.549142	.0726222	-21.33	0.000	-1.691528	-1.406755
ldistsq	.1269054	.0055092	23.04	0.000	.1161037	.1377071
uhat_1	1.005428	.0062555	160.73	0.000	.9931627	1.017693
_cons	13.81801	.2389316	57.83	0.000	13.34955	14.28647

c. The coefficient on  $concen_{it}$  is .360 and the  $t$ -statistic that accounts for heteroskedasticity and serial correlation is 6.15. Therefore, the partial correlation between  $lfare$  and  $concen$  is enough to implement an IV procedure.

```
. reg lfare y98 y99 y00 ldist ldistsq concen, cluster(id)
```

```
Linear regression                                Number of obs =      4596
                                                F(   6,   1148) =  205.63
                                                Prob > F       =   0.0000
                                                R-squared      =   0.4062
                                                Root MSE      =   .33651
```

(Std. Err. adjusted for 1149 clusters in id)

lfare	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
y98	.0211244	.0041474	5.09	0.000	.0129871	.0292617
y99	.0378496	.0051795	7.31	0.000	.0276872	.048012
y00	.09987	.0056469	17.69	0.000	.0887906	.1109493
ldist	-.9016004	.2719464	-3.32	0.001	-1.435168	-.3680328
ldistsq	.1030196	.0201602	5.11	0.000	.0634647	.1425745
concen	.3601203	.058556	6.15	0.000	.2452315	.4750092
_cons	6.209258	.9117551	6.81	0.000	4.420364	7.998151

d. The IV estimates are given below. The estimated elasticity is huge,  $-1.78$ . This seems very large. The fully robust standard error is about twice as large as the usual OLS standard error, and the fully robust 95% confidence interval is  $-2.71$  to  $-.84$ , which is very wide, but it excludes the point estimate from pooled OLS ( $-.5.65$ ).

```
. ivreg lpassen y98 y99 y00 ldist ldistsq (lfare=concen)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs = 4596		
Model	-556.334915	6	-92.7224858	F( 6, 4589) = 20.45		
Residual	4147.02233	4589	.903687586	Prob > F = 0.0000		
				R-squared =		
				Adj R-squared =		
Total	3590.68741	4595	.781433605	Root MSE = .95062		

lpassen	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
lfare	-1.776549	.2358788	-7.53	0.000	-2.238985	-1.314113
y98	.0616171	.0400745	1.54	0.124	-.0169481	.1401824
y99	.1241675	.0405153	3.06	0.002	.044738	.2035971
y00	.2542695	.0456607	5.57	0.000	.1647525	.3437865
ldist	-2.498972	.4058371	-6.16	0.000	-3.294607	-1.703336
ldistsq	.2314932	.0345468	6.70	0.000	.1637648	.2992216
_cons	21.21249	1.891586	11.21	0.000	17.50407	24.9209

Instrumented: lfare

Instruments: y98 y99 y00 ldist ldistsq concen

```
. ivreg lpassen y98 y99 y00 ldist ldistsq (lfare=concen), cluster(id)
```

Instrumental variables (2SLS) regression

Number of obs = 4596  
F( 6, 1148) = 28.02  
Prob > F = 0.0000  
R-squared =  
Root MSE = .95062

(Std. Err. adjusted for 1149 clusters in id)

lpassen	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
lfare	-1.776549	.4753368	-3.74	0.000	-2.709175	-.8439226
y98	.0616171	.0131531	4.68	0.000	.0358103	.0874239
y99	.1241675	.0183335	6.77	0.000	.0881967	.1601384
y00	.2542695	.0458027	5.55	0.000	.164403	.3441359
ldist	-2.498972	.831401	-3.01	0.003	-4.130207	-.8677356
ldistsq	.2314932	.0705247	3.28	0.001	.0931215	.3698649
_cons	21.21249	3.860659	5.49	0.000	13.63775	28.78722

```
-----
Instrumented:  lfare
Instruments:   y98 y99 y00 ldist ldistsq concen
-----
```

e. To compute the asymptotic standard error of  $\sqrt{N}(\hat{\beta}_{1,P2SLS} - \hat{\beta}_{1,POLS})$  using the traditional Hausman approach, we have to maintain enough assumptions so that POLS is relatively efficient under the null. Letting

$$\mathbf{w}_{it} = (1, y98_t, y99_t, y00_t, lfare_{it}, ldist_i, ldist_i^2, concen_{it})$$

we would have to assume, under  $H_0$ ,

$$\begin{aligned} E(\mathbf{w}_{it}' u_{it1}) &= \mathbf{0}, t = 1, \dots, T \\ E(u_{it1}^2 | \mathbf{w}_{it}) &= \sigma^2, t = 1, \dots, T \\ E(u_{it1} u_{ir1} | \mathbf{w}_{it}, \mathbf{w}_{ir}) &= 0, r \neq t. \end{aligned}$$

The first assumption must be maintained under the null for the test to make sense. The second assumption – homoskedasticity – can never be guaranteed, and so it is always a good idea to make tests robust to heteroskedasticity. The current application is a static equation, and so the assumption of no serial correlation is especially strong. In fact, from part b we already have good evidence that there is substantial serial correlation in the errors (although this test maintains contemporaneous exogeneity of  $lfare_{it}$ , along with the distance variables).

f. The Stata commands are given below. The fully robust  $t$  statistic on  $\hat{v}_{it2}$  is 2.92, which is a strong rejection of the null that  $lfare_{it}$  is (contemporaneousl) exogenous – assuming that  $concen_{it}$  is contemporaneously exogenous.

```
. qui reg lfare y98 y99 y00 ldist ldistsq concen
. predict v2hat, resid
. reg lpassen y98 y99 y00 lfare ldist ldistsq v2hat, cluster(id)

Linear regression                                Number of obs =      4596
                                                F(   7,   1148) =     31.50
                                                Prob > F          =    0.0000
                                                R-squared         =    0.0711
```

Root MSE = .85265

(Std. Err. adjusted for 1149 clusters in id

lpassen	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
y98	.0616171	.0112127	5.50	0.000	.0396175	.0836167
y99	.1241675	.0160906	7.72	0.000	.0925973	.1557378
y00	.2542695	.040158	6.33	0.000	.1754782	.3330608
lfare	-1.776549	.4197937	-4.23	0.000	-2.600198	-.9528999
ldist	-2.498972	.767078	-3.26	0.001	-4.004004	-.9939395
ldistsq	.2314932	.0640361	3.62	0.000	.1058524	.3571341
v2hat	1.249653	.4273322	2.92	0.004	.4112137	2.088093
_cons	21.21249	3.46901	6.11	0.000	14.40618	28.0188

**8.16 (Bonus Question).** Consider the GIV estimator with incorrect restrictions imposed on the estimator of  $\mathbf{\Omega}$ . That is, in (8.47) use  $\hat{\mathbf{\Lambda}}$  in place of  $\mathbf{\Omega}$  with  $\hat{\mathbf{\Lambda}} \xrightarrow{p} \mathbf{\Lambda} \neq \mathbf{\Omega}$ .

a. If Assumption GIV.1 holds, that is  $E(\mathbf{Z}_i \otimes \mathbf{u}_i) = \mathbf{0}$ , argue that the GIV estimator is still consistent under an appropriate rank condition (and state the rank condition).

b. Argue that, under the assumptions of part a, the GIV estimator that uses  $\hat{\mathbf{\Lambda}}$  is  $\sqrt{N}$ –asymptotically equivalent to the (infeasible) GIV estimator that uses  $\mathbf{\Lambda}$ .

c. If you insist on using  $\hat{\mathbf{\Lambda}}$  but want to guard against inappropriate inference, what would you do?

**Solution**

a. From equation (8.47), and applying the law of large numbers, the key orthogonality condition for consistency is

$$E(\mathbf{Z}_i' \mathbf{\Lambda}^{-1} \mathbf{u}_i) = \mathbf{0}$$

because  $\hat{\mathbf{\Lambda}} \xrightarrow{p} \mathbf{\Lambda}$ . But if Assumption GIV.1 holds, any linear combination of  $\mathbf{Z}_i$  is uncorrelated with  $\mathbf{u}_i$ , including  $\mathbf{\Lambda}^{-1} \mathbf{Z}_i$ . There are two parts to the rank condition, with the first being the most important:

$$\begin{aligned} \text{rank } E(\mathbf{Z}_i' \mathbf{\Lambda}^{-1} \mathbf{Z}_i) &= L \\ \text{rank } E(\mathbf{Z}_i' \mathbf{\Lambda}^{-1} \mathbf{X}_i) &= K \end{aligned}$$

b. This follows the same line of reasoning that we used for FGLS in Chapter 7 can be used. First, using the same trick with the Kronecker product,

$$\begin{aligned} N^{-1} \sum_{i=1}^N \mathbf{Z}_i' \hat{\mathbf{\Lambda}}^{-1} \mathbf{Z}_i &= N^{-1} \sum_{i=1}^N \mathbf{Z}_i' \mathbf{\Lambda}^{-1} \mathbf{Z}_i + o_p(1) \xrightarrow{p} E(\mathbf{Z}_i' \mathbf{\Lambda}^{-1} \mathbf{Z}_i) \\ N^{-1} \sum_{i=1}^N \mathbf{Z}_i' \hat{\mathbf{\Lambda}}^{-1} \mathbf{X}_i &= N^{-1} \sum_{i=1}^N \mathbf{Z}_i' \mathbf{\Lambda}^{-1} \mathbf{X}_i + o_p(1) \xrightarrow{p} E(\mathbf{Z}_i' \mathbf{\Lambda}^{-1} \mathbf{X}_i) \end{aligned}$$

Second,

$$\begin{aligned} N^{-1/2} \sum_{i=1}^N \mathbf{Z}_i' \hat{\Lambda}^{-1} \mathbf{u}_i - N^{-1/2} \sum_{i=1}^N \mathbf{Z}_i' \Lambda^{-1} \mathbf{u}_i &= N^{-1/2} \sum_{i=1}^N (\mathbf{u}_i \otimes \mathbf{Z}_i)' \text{vec}(\hat{\Lambda}^{-1} - \Lambda^{-1}) \\ &= O_p(1) \cdot o_p(1) = o_p(1). \end{aligned}$$

Combining these asymptotic equivalences shows that replacing  $\Lambda$  with the consistent estimator

$\hat{\Lambda}$  does not affect the  $\sqrt{N}$ -limiting distribution of the GIV estimator.

c. Use a full robust asymptotic variance matrix estimator. Write

$$\widehat{\text{Avar}}[\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{-1}$$

where

$$\begin{aligned} \hat{\mathbf{A}} &= \left( N^{-1} \sum_{i=1}^N \mathbf{X}_i' \hat{\Lambda}^{-1} \mathbf{Z}_i \right) \left( N^{-1} \sum_{i=1}^N \mathbf{Z}_i' \hat{\Lambda}^{-1} \mathbf{Z}_i \right)^{-1} \left( N^{-1} \sum_{i=1}^N \mathbf{Z}_i' \hat{\Lambda}^{-1} \mathbf{X}_i \right) \\ \hat{\mathbf{B}} &= \left( N^{-1} \sum_{i=1}^N \mathbf{X}_i' \hat{\Lambda}^{-1} \mathbf{Z}_i \right) \left( N^{-1} \sum_{i=1}^N \mathbf{Z}_i' \hat{\Lambda}^{-1} \mathbf{Z}_i \right)^{-1} \left( N^{-1} \sum_{i=1}^N \mathbf{Z}_i' \hat{\Lambda}^{-1} \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \hat{\Lambda}^{-1} \mathbf{Z}_i \right) \\ &\quad \cdot \left( N^{-1} \sum_{i=1}^N \mathbf{Z}_i' \hat{\Lambda}^{-1} \mathbf{Z}_i \right)^{-1} \left( N^{-1} \sum_{i=1}^N \mathbf{Z}_i' \hat{\Lambda}^{-1} \mathbf{X}_i \right) \end{aligned}$$

where  $\hat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}$  are the GIV residuals. This asymptotic variance matrix estimator allows  $E(\mathbf{u}_i \mathbf{u}_i') \neq \Lambda$  as well as system heteroskedasticity, that is,  $E(\mathbf{u}_i \mathbf{u}_i' | \mathbf{Z}_i) \neq E(\mathbf{u}_i \mathbf{u}_i')$ . Of course, we get  $\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}})$  as  $\hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{-1} / N$ , whereby all of the divisions by  $N$  disappear.

**8.17 (Bonus Question).** Consider a panel data model with contemporaneously exogenous instruments  $\mathbf{z}_{it}$ :

$$\begin{aligned} y_{it1} &= \mathbf{x}_{it} \boldsymbol{\beta} + u_{it}, \\ E(\mathbf{z}_{it}' u_{it}) &= \mathbf{0}, t = 1, \dots, T, \end{aligned}$$

where  $\mathbf{x}_{it}$  is  $1 \times K$  and  $\mathbf{z}_{it}$  is  $1 \times L$  for all  $t$ ,  $L \geq K$ .

a. If we maintain the assumptions

ASSUMPTION P2SLS.1:  $E(\mathbf{z}_{it}'\mathbf{u}_{it}) = 0$ ,  $t = 1, \dots, T$

ASSUMPTION P2SLS.2: (a)  $\text{rank} \sum_{t=1}^T E(\mathbf{z}_{it}'\mathbf{z}_{it}) = L$ ; (b)  $\text{rank} \sum_{t=1}^T E(\mathbf{z}_{it}'\mathbf{x}_{it}) = K$ ,

argue that the pooled 2SLS (P2SLS) estimator is generally consistent (as always with  $T$  fixed,  $N \rightarrow \infty$ , and random sampling across  $i$ ).

b. Explain how to estimate the asymptotic variance matrix of the P2SLS estimator under the assumptions in part a.

c. Suppose we add the assumption

ASSUMPTION P2SLS.3: (a)  $E(u_{it}^2\mathbf{z}_{it}'\mathbf{z}_{it}) = \sigma^2 E(\mathbf{z}_{it}'\mathbf{z}_{it})$ ,  $t = 1, \dots, T$ ; (b)  $E(u_{it}u_{ir}\mathbf{z}_{it}'\mathbf{z}_{is}) = \mathbf{0}$ ,

$t \neq r$ .

Argue that the usual 2SLS variance matrix estimator that assumes homoskedasticity and ignores the time series component is valid.

d. What would you do if Assumption P2SLS.3(b) holds but not necessarily P2SLS.3(a)?

### Solution

a. Using the general formula for the S2SLS estimator, we can write the P2SLS estimator (with probability approaching one) as

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left[ \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{x}_{it}'\mathbf{z}_{it} \right) \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}_{it}'\mathbf{z}_{it} \right)^{-1} \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}_{it}'\mathbf{x}_{it} \right) \right]^{-1} \\ &\quad \cdot \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{x}_{it}'\mathbf{z}_{it} \right) \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}_{it}'\mathbf{z}_{it} \right)^{-1} \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}_{it}'\mathbf{y}_{it} \right) \\ &= \boldsymbol{\beta} + \left[ \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{x}_{it}'\mathbf{z}_{it} \right) \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}_{it}'\mathbf{z}_{it} \right)^{-1} \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}_{it}'\mathbf{x}_{it} \right) \right]^{-1} \\ &\quad \cdot \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{x}_{it}'\mathbf{z}_{it} \right) \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}_{it}'\mathbf{z}_{it} \right)^{-1} \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}_{it}'\mathbf{u}_{it} \right)\end{aligned}$$



Notice how the law of large numbers implies

$$N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}'_{it} \mathbf{z}_{it} \xrightarrow{p} \sum_{t=1}^T E(\mathbf{z}'_{it} \mathbf{z}_{it})$$

$$N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}'_{it} \mathbf{x}_{it} \xrightarrow{p} \sum_{t=1}^T E(\mathbf{z}'_{it} \mathbf{x}_{it})$$

and the rank condition states that these matrices of ranks  $L$  and  $K$ , respectively. Therefore, the plim can pass through all inverses. We also apply the WLLN and Assumption P2SLS.1 to get

$$N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}'_{it} \mathbf{u}_{it} \xrightarrow{p} \sum_{t=1}^T E(\mathbf{z}'_{it} \mathbf{u}_{it}) = \mathbf{0}.$$

Now we just pass the plim through using Slutsky's Theorem to get  $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ .

b. We have

$$\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}$$

where

$$\mathbf{A} = \left( \sum_{t=1}^T E(\mathbf{x}'_{it} \mathbf{z}_{it}) \right) \left( \sum_{t=1}^T E(\mathbf{z}'_{it} \mathbf{z}_{it}) \right)^{-1} \left( \sum_{t=1}^T E(\mathbf{z}'_{it} \mathbf{x}_{it}) \right)$$

$$\mathbf{B} = \left( \sum_{t=1}^T E(\mathbf{x}'_{it} \mathbf{z}_{it}) \right) \left( \sum_{t=1}^T E(\mathbf{z}'_{it} \mathbf{z}_{it}) \right)^{-1} \left( \sum_{t=1}^T \sum_{r=1}^T E(\mathbf{u}_{it} \mathbf{u}_{ir} \mathbf{z}'_{it} \mathbf{z}_{ir}) \right)$$

$$\cdot \left( \sum_{t=1}^T E(\mathbf{z}'_{it} \mathbf{z}_{it}) \right)^{-1} \left( \sum_{t=1}^T E(\mathbf{z}'_{it} \mathbf{x}_{it}) \right)$$

We can consistently estimate each of these matrices:

$$\begin{aligned}\hat{\mathbf{A}} &= \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{x}'_{it} \mathbf{z}_{it} \right) \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}'_{it} \mathbf{z}_{it} \right)^{-1} \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}'_{it} \mathbf{x}_{it} \right) \\ \hat{\mathbf{B}} &= \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{x}'_{it} \mathbf{z}_{it} \right) \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}'_{it} \mathbf{z}_{it} \right)^{-1} \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \sum_{r=1}^T \hat{u}_{it} \hat{u}_{ir} \mathbf{z}'_{it} \mathbf{z}_{ir} \right) \\ &\quad \cdot \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}'_{it} \mathbf{z}_{it} \right)^{-1} \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}'_{it} \mathbf{x}_{it} \right)\end{aligned}$$

where  $\hat{u}_{it} = y_{it} - \mathbf{x}_{it}' \hat{\boldsymbol{\beta}}$  are the P2SLS residuals.

c. With Assumption P2SLS.3,

$$\sum_{t=1}^T \sum_{r=1}^T E(u_{it} u_{ir} \mathbf{z}'_{it} \mathbf{z}_{ir}) = \sum_{t=1}^T E(u_{it}^2 \mathbf{z}'_{it} \mathbf{z}_{it}) = \sigma^2 \sum_{t=1}^T E(\mathbf{z}'_{it} \mathbf{z}_{it}),$$

where the first equality follows from  $E(u_{it} u_{ir} \mathbf{z}'_{it} \mathbf{z}_{ir}) = \mathbf{0}$ ,  $t \neq r$ , and the second follows from

$E(u_{it}^2 \mathbf{z}'_{it} \mathbf{z}_{it}) = \sigma^2 E(\mathbf{z}'_{it} \mathbf{z}_{it})$ ,  $t = 1, \dots, T$ . Therefore,

$$\mathbf{B} = \sigma^2 \left( \sum_{t=1}^T E(\mathbf{x}'_{it} \mathbf{z}_{it}) \right) \left( \sum_{t=1}^T E(\mathbf{z}'_{it} \mathbf{z}_{it}) \right)^{-1} \left( \sum_{t=1}^T E(\mathbf{z}'_{it} \mathbf{x}_{it}) \right) = \sigma^2 \mathbf{A},$$

and so

$$\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \sigma^2 \mathbf{A}^{-1}$$

When we use  $\hat{\mathbf{A}}$  from part b and a consistent estimator of  $\sigma^2$  (with optional but standard degrees-of-freedom adjustment),

$$\hat{\sigma}^2 = \frac{1}{NT - K} \sum_{i=1}^N \sum_{t=1}^T \hat{u}_{it}^2,$$

then we get

$$\widehat{\text{Avar}}(\hat{\beta}) = \hat{\sigma}^2 \left[ \left( \sum_{i=1}^N \sum_{t=1}^T \mathbf{x}_{it}' \mathbf{z}_{it} \right) \left( \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}_{it}' \mathbf{z}_{it} \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}_{it}' \mathbf{x}_{it} \right) \right]^{-1},$$

which is exactly the standard formula for 2SLS treating the panel data set as one long cross section.

d. We need to make the variance matrix robust to heteroskedasticity only. So

$$\begin{aligned} \hat{\mathbf{B}} = & \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{x}_{it}' \mathbf{z}_{it} \right) \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}_{it}' \mathbf{z}_{it} \right)^{-1} \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \hat{u}_{it}^2 \mathbf{z}_{it}' \mathbf{z}_{it} \right) \\ & \cdot \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}_{it}' \mathbf{z}_{it} \right)^{-1} \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{z}_{it}' \mathbf{x}_{it} \right). \end{aligned}$$

The resulting  $\widehat{\text{Avar}}(\hat{\beta})$  is exactly what would be computed by treating the panel data set as one long cross section with inference robust to heteroskedasticity.

**8.18 (Bonus Question).** Consider the panel data model

$$y_{it} = \mathbf{x}_{it}' \boldsymbol{\beta} + u_{it}, t = 1, \dots, T,$$

where  $\mathbf{x}_{it}$  is a  $1 \times K$  vector and the instruments at time  $t$  are  $\mathbf{z}_{it}$ , a  $1 \times L$  vector for all  $t$ . Suppose the instruments are strictly exogenous in the sense that

$$\text{E}(u_{it} | \mathbf{z}_{i1}, \mathbf{z}_{i2}, \dots, \mathbf{z}_{iT}) = \text{E}(u_{it} | \mathbf{z}_i) = \mathbf{0}, t = 1, \dots, T.$$

Assume that  $\text{E}(\mathbf{u}_i \mathbf{u}_i' | \mathbf{z}_i) = \text{E}(\mathbf{u}_i \mathbf{u}_i') = \boldsymbol{\Omega}$ , where  $\mathbf{z}_i$  is the vector all all exogenous variables in all time periods. Further, assume that  $\boldsymbol{\Omega}$  has the AR(1) form:

$$\boldsymbol{\Omega} = \sigma_e^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 & \rho \\ \rho^{T-1} & \rho^{T-2} & \dots & \rho & 1 \end{pmatrix} \equiv \sigma_e^2 \boldsymbol{\Psi}$$

where  $u_{it} = \rho u_{i,t-1} + e_{it}, t = 1, \dots, T$ .

a. If  $\mathbf{Z}'_i = (\mathbf{z}'_{i1}, \dots, \mathbf{z}'_{iT})$ , find the matrix of transformed instruments,  $\Psi^{-1/2}\mathbf{Z}_i$ .

b. Describe how to implement the GIV estimator as a particular pooled 2SLS estimation

when  $\Omega$  has the AR(1) structure.

c. If you think the AR(1) model might be incorrect, or the system homoskedasticity assumption does not hold, propose a simple method for obtaining valid standard errors and test statistics.

### Solution

From Section 7.8.6, we know that when  $\Psi$  has the AR(1) structure given above,

$$\Psi^{-1/2}\mathbf{Z}_i = \begin{pmatrix} (1 - \rho^2)^{1/2}\mathbf{z}_{i1} \\ \mathbf{z}_{i2} - \rho\mathbf{z}_{i1} \\ \vdots \\ \mathbf{z}_{iT} - \rho\mathbf{z}_{i,T-1} \end{pmatrix}$$

so that, for  $t \geq 2$ , the transformation results in quasi-difference. For  $t = 1$ , the transformation ensures that the transformed errors will have common variance for all  $t = 1, \dots, T$ .

b. We need to estimate  $\rho$ , so we would use pooled 2SLS to get residuals, say  $\check{u}_{it}$ . Then, estimate  $\rho$  from the pooled OLS regression

$$\check{u}_{it} \text{ on } \check{u}_{i,t-1}, t = 2, \dots, T; i = 1, \dots, N.$$

The GIV transformed equation is

$$\begin{aligned} (1 - \rho^2)^{1/2}y_{i1} &= (1 - \rho^2)^{1/2}\mathbf{x}_{i1}\beta + (1 - \rho^2)^{1/2}u_{i1} \\ y_{it} - \rho y_{i,t-1} &= (\mathbf{x}_{it} - \rho\mathbf{x}_{i,t-1})\beta + u_{it} - \rho u_{i,t-1}, t = 2, \dots, T. \end{aligned}$$

The GIV estimator is obtained by replacing  $\rho$  with  $\hat{\rho}$  and estimating

$$\tilde{y}_{it} = \tilde{\mathbf{x}}_{it}\beta + error_{it}, t = 1, \dots, T; i = 1, \dots, N$$

using IVs  $\tilde{\mathbf{z}}_{it}$ , where

$$\begin{aligned}\tilde{\mathbf{z}}_{i1} &= (1 - \hat{\rho}^2)^{1/2} \mathbf{z}_{i1} \\ \tilde{\mathbf{z}}_{it} &= \mathbf{z}_{it} - \hat{\rho} \mathbf{z}_{i,t-1}, t = 2, \dots, T\end{aligned}$$

and where similar definitions hold for  $\tilde{y}_{it}$  and  $\tilde{\mathbf{x}}_{it}$ . As always, the estimation of  $\rho$  has no effect on the  $\sqrt{N}$ -asymptotic distribution under the strict exogeneity assumption on the IVs. The usual P2SLS statistics from the estimation on the transformed variables are asymptotically valid.

c. If we have misspecified  $\text{Var}(\mathbf{u}_i | \mathbf{z}_i)$  then we should make the P2SLS inference from part b fully robust – to heteroskedasticity and serial correlation. In other words, the transformed errors

$$\begin{aligned}e_{i1} &= (1 - \rho^2)^{1/2} u_{i1} \\ e_{it} &= u_{it} - \rho u_{i,t-1}, t = 2, \dots, T\end{aligned}$$

will have serial correlation if the AR(1) model is incorrect, and such errors can always have heteroskedasticity if  $\{u_{it}\}$  does. We know that the GIV estimator that uses an incorrect variance structure is still consistent and  $\sqrt{N}$ -asymptotically normal. We might get a more efficient estimator assuming a simple AR(1) structure than using P2SLS on the original: accounting for the serial correlation at all might be better than ignoring it in estimation. This is the same motivation underlying the generalized estimation equations literature when the explanatory variables are strictly exogenous.

## Solutions to Chapter 9 Problems

9.1. a. No. What causal inference could one draw from this? We may be interested in the tradeoff between wages and benefits, but then either of these can be taken as the dependent variable and estimation of either equation would be by OLS. Of course, if we have omitted some important factors, or have a measurement error problem, OLS could be inconsistent for estimating the tradeoff. But there is no simultaneity problem: wages and benefits are jointly determined, but there is no sense in which an equation for wage and another for benefits satisfy the autonomy requirement.

b. Yes. We can certainly think of an exogenous change in law enforcement expenditures causing a reduction in crime, and we are certainly interested in such counterfactuals. If we could do the appropriate experiment, where expenditures are assigned randomly across cities, then we could estimate the crime equation by OLS. The simultaneous equations model recognizes that cities choose law enforcement expenditures in part based on what they expect the crime rate to be. An SEM is a convenient way to allow expenditures to depend on unobservables (to the econometrician) that affect crime.

c. No. These are both choice variables of the firm, and the parameters in a two-equation system modeling one in terms of the other, and vice versa, have no economic meaning. If we want to know how a change in the price of foreign technology affects foreign technology (FT) purchases, why would we want to hold fixed R&D spending? Clearly FT purchases and R&D spending are simultaneously chosen, but we should use a two-equation SUR setup where neither is an explanatory variable in the other's equation.

d. Yes. We can be interested in the causal effect of alcohol consumption on productivity, and therefore on wage. One's hourly wage is determined by productivity, and other factors;

alcohol consumption is determined by individual choice, where one factor is income.

e. No. These are choice variables by the same household. It makes no sense to think about how exogenous changes in one would affect the other. Further, suppose that we look at the effects of changes in local property tax rates. We would not want to hold fixed family saving and then measure the effect of changing property taxes on housing expenditures. When the property tax changes, a family will generally adjust expenditure in all categories. A SUR system with property tax as an explanatory variable is the appropriate strategy.

f. No. These are both chosen by the firm, presumably to maximize profits. It makes no sense to hold advertising expenditures fixed while looking at how other variables affect price markup.

g. Yes. The outcome variables – quantity demanded and advertising expenditures – are determined by different economic agents. It makes sense to model quantity demanded as a function of advertising expenditures – reflecting that more exposure to the public can affect demand – and at the same time recognize that how much a firm spends on advertising can be determined by how much of the product it can sell.

h. Yes. The rate of HIV infection is determined by many factors, with condom usage being one. We can easily imagine being interested in the effects of making condoms more available on the incidence of HIV. The second equation, which models demand for condoms as a function of HIV incidence, captures the idea that more people might use condoms as the risk of HIV infection increases. Each equation stands on its own.

**9.2.** a. Write the system as

$$\begin{pmatrix} 1 & -\gamma_1 \\ -\gamma_2 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \mathbf{z}_{(1)}\boldsymbol{\delta}_{(1)} + u_1 \\ \mathbf{z}_{(2)}\boldsymbol{\delta}_{(2)} + u_2 \end{pmatrix}.$$

Unique solutions for  $y_1$  and  $y_2$  exist only if the matrix premultiplying  $(y_1, y_2)'$  is nonsingular. But its determinant is  $1 - \gamma_1\gamma_2$ , so a necessary and sufficient condition for the reduced forms to exist is  $\gamma_1\gamma_2 \neq 1$ .

b. The rank condition holds for the first equation if and only if  $\mathbf{z}_{(2)}$  contains an element not in  $\mathbf{z}_{(1)}$  and the coefficient in  $\delta_{(2)}$  on that variable is not zero. Similarly, the rank condition holds for the second equation if and only if  $\mathbf{z}_{(1)}$  contains an element not in  $\mathbf{z}_{(2)}$  and the coefficient in  $\delta_{(1)}$  on that variable is not zero.

9.3. a. We can apply part b of Problem 9.2. First, the only variable excluded from the *support* equation is the variable *mremarr*; since the *support* equation contains one endogenous variable, this equation is identified if and only if  $\delta_{21} \neq 0$ . This ensures that there is an exogenous variable shifting the mother's reaction function that does not also shift the father's reaction function.

The *visits* equation is identified if and only if at least one of *finc* and *fremarr* actually appears in the *support* equation; that is, we need  $\delta_{11} \neq 0$  or  $\delta_{13} \neq 0$ .

b. Each equation can be estimated by 2SLS using instruments  
1, *finc*, *fremarr*, *dist*, *mremarr*.

c. First, obtain the reduced form for *visits* :

$$visits = \pi_{20} + \pi_{21}finc + \pi_{22}fremarr + \pi_{23}dist + \pi_{24}mremarr + v_2.$$

Estimate this equation by OLS, and save the residuals,  $\hat{v}_2$ . Then, run the OLS regression

$$support \text{ on } 1, visits, finc, fremarr, dist, \hat{v}_2$$

and do a (heteroskedasticity-robust)  $t$  test that the coefficient on  $\hat{v}_2$  is zero. If this test rejects we conclude that *visits* is in fact endogenous in the *support* equation.



d. There is one overidentifying restriction in the *visits* equation, assuming that  $\delta_{11}$  and  $\delta_{12}$  are both different from zero. Assuming homoskedasticity of  $u_2$ , the easiest way to test the overidentifying restriction is to first estimate the *visits* equation by 2SLS. as in part b. Let  $\hat{u}_2$  be the 2SLS residuals. Then, run the auxiliary regression

$$\hat{u}_2 \text{ on } 1, \text{finc}, \text{fremarr}, \text{dist}, \text{mremarr};$$

the sample size times the usual  $R$ -squared from this regression is distributed asymptotically as  $\chi^2_1$  under the null hypothesis that all instruments are exogenous.

A heteroskedasticity-robust test is also easy to obtain. Let  $\widehat{\text{support}}$  denote the fitted values from the reduced form regression for *support*. Next, regress *finc* (or *fremarr*) on  $\widehat{\text{support}}, \text{mremarr}, \text{dist}$ , and save the residuals, say  $\hat{r}_1$ . Then, run the simple regression (without intercept) of  $\hat{u}_2$  on  $\hat{r}_1$  and use the heteroskedasticity-robust  $t$  statistic on  $\hat{r}_1$ . (Note that no intercept is needed in this final regression, but including one is harmless.)

**9.4.** a. Because the third equation contains no right hand side endogenous variables, a reduced form exists for the system if and only if the first two equations can be solved for  $y_1$  and  $y_2$  as functions of  $y_3, z_1, z_2, z_3, u_1$ , and  $u_2$ . But this is equivalent to asking when the system

$$\begin{pmatrix} 1 & -\gamma_{12} \\ 1 & -\gamma_{22} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

has a unique solution in  $y_1$  and  $y_2$ . This matrix is nonsingular if and only if  $\gamma_{12} \neq \gamma_{22}$ . This implies that the  $3 \times 3$  matrix  $\Gamma$  in the general SEM notation is nonsingular.

b. The third equation satisfies the rank condition because it includes no right-hand-side endogenous variables. The first equation fails the order condition because there are no excluded exogenous variables in it, but there is one included endogenous variable. This means

it fails the rank condition also. The second equation is just identified according to the order condition because it contains two endogenous variables and also excludes two exogenous variables. To examine the rank condition, write the second equation as  $\mathbf{y}\boldsymbol{\gamma}_2 + \mathbf{z}\boldsymbol{\delta}_2 + u_2 = 0$ , where  $\boldsymbol{\gamma}_2 = (-1, \gamma_{22}, \gamma_{23})'$  and  $\boldsymbol{\delta}_2 = (\delta_{21}, 0, 0)'$ . Write  $\boldsymbol{\beta}_2 = (-1, \gamma_{22}, \gamma_{23}, \delta_{21}, \delta_{22}, \delta_{23})'$  as the vector of parameters for the second equation with only the normalization  $\gamma_{21} = -1$  imposed. Then, the restrictions  $\delta_{22} = 0$  and  $\delta_{23} = 0$  can be written as  $\mathbf{R}_2\boldsymbol{\beta}_2 = \mathbf{0}$ , where

$$\mathbf{R}_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Now letting  $\mathbf{B}$  be the  $6 \times 3$  matrix of all parameters, and imposing all exclusion restrictions in the system,

$$\mathbf{R}_2\mathbf{B} = \begin{pmatrix} \delta_{12} & 0 & \delta_{32} \\ \delta_{13} & 0 & \delta_{33} \end{pmatrix}.$$

The rank condition requires this matrix have rank equal to two. Provided the vector  $(\delta_{32}, \delta_{33})'$  is not a multiple of  $(\delta_{12}, \delta_{13})'$ , or  $\delta_{12}\delta_{33} \neq \delta_{13}\delta_{32}$ , the rank condition is satisfied.

**9.5. a.** Let  $\boldsymbol{\beta}_1$  denote the  $7 \times 1$  vector of parameters in the first equation with only the normalization restriction imposed:

$$\boldsymbol{\beta}_1' = (-1, \gamma_{12}, \gamma_{13}, \delta_{11}, \delta_{12}, \delta_{13}, \delta_{14}).$$

The restrictions  $\delta_{12} = 0$  and  $\delta_{13} + \delta_{14} = 1$  are obtained by choosing

$$\mathbf{R}_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

Because  $\mathbf{R}_1$  has two rows, and  $G - 1 = 2$ , the order condition is satisfied. Now we need to check the rank condition. Letting  $\mathbf{B}$  denote the  $7 \times 3$  matrix of all structural parameters with

only the three normalizations, straightforward matrix multiplication gives

$$\mathbf{R}_1 \mathbf{B} = \begin{pmatrix} \delta_{12} & \delta_{22} & \delta_{32} \\ \delta_{13} + \delta_{14} - 1 & \delta_{23} + \delta_{24} - \gamma_{21} & \delta_{33} + \delta_{34} - \gamma_{31} \end{pmatrix}.$$

By definition of the constraints on the first equation, the first column of  $\mathbf{R}_1 \mathbf{B}$  is zero. Next, we use the constraints in the remainder of the system to get the expression for  $\mathbf{R}_1 \mathbf{B}$  with all information imposed. But  $\gamma_{23} = 0, \delta_{22} = 0, \delta_{23} = 0, \delta_{24} = 0, \gamma_{31} = 0$ , and  $\gamma_{32} = 0$ , and so  $\mathbf{R}_1 \mathbf{B}$  becomes

$$\mathbf{R}_1 \mathbf{B} = \begin{pmatrix} 0 & 0 & \delta_{32} \\ 0 & -\gamma_{21} & \delta_{33} + \delta_{34} - \gamma_{31} \end{pmatrix}.$$

Identification requires  $\gamma_{21} \neq 0$  and  $\delta_{32} \neq 0$ .

b. It is easy to see how to estimate the first equation under the given assumptions. Set  $\delta_{14} = 1 - \delta_{13}$  and plug into the equation. After simple algebra we get

$$y_1 - z_4 = \gamma_{12}y_2 + \gamma_{13}y_3 + \delta_{11}z_1 + \delta_{13}(z_3 - z_4) + u_1.$$

This equation can be estimated by 2SLS using instruments  $(z_1, z_2, z_3, z_4)$ . Note that, if we just count instruments, there are just enough instruments to estimate this equation.

**9.6.** a. If  $\gamma_{13} = 0$  then the two equations constitute a linear SEM. In that case, the first equation is identified if and only if  $\delta_{23} \neq 0$  and the second equation is identified if and only if  $\delta_{12} \neq 0$ .

b. If we plug the second equation into the first we obtain

$$(1 - \gamma_{12}\gamma_{21} - \gamma_{13}\gamma_{21}z_1)y_1 = (\delta_{10}\gamma_{12}\delta_{20}) + (\gamma_{12}\delta_{21} + \gamma_{13}\delta_{20} + \delta_{11})z_1 \\ + \gamma_{13}\delta_{21}z_1^2 + \delta_{12}z_2 + \gamma_{12}\delta_{23}z_3 + \gamma_{13}\delta_{23}z_1z_3 + u_1 + (\gamma_{12} + \gamma_{13}z_1)u_2.$$

This can be solved for  $y_1$  provided  $(1 - \gamma_{12}\gamma_{21} - \gamma_{13}\gamma_{21}z_1) \neq 0$ . Given the solution for  $y_1$ , we

can use the second equation to get  $y_2$ . Note that both are nonlinear in  $z_1$  unless  $\gamma_{13} = 0$ .

c. Since  $E(u_1|\mathbf{z}) = E(u_2|\mathbf{z}) = 0$ , we can use part (b) to get

$$E(y_1|z_1, z_2, z_3) = [(\delta_{10} + \gamma_{12}\delta_{20}) + (\gamma_{12}\delta_{21} + \gamma_{13}\delta_{20} + \delta_{11})z_1 + \gamma_{13}\delta_{21}z_1^2 + \delta_{12}z_2 + \gamma_{12}\delta_{23}z_3 + \gamma_{13}\delta_{23}z_1z_3]/(1 - \gamma_{12}\gamma_{21} - \gamma_{13}\gamma_{21}z_1).$$

Again, this is a nonlinear function of the exogenous variables appearing in the system unless  $\gamma_{13} = 0$ . If  $\gamma_{21} = 0$ ,  $E(y_1|z_1, z_2, z_3)$  becomes linear in  $z_2$  and quadratic in  $z_1$  and  $z_3$ .

d. If  $\gamma_{13} = 0$ , we saw in part a that the first equation is identified. If we include  $\gamma_{13}y_2z_1$  in the model, we need at least one instrument for it. But regardless of the value of  $\gamma_{13}$ , terms  $z_1^2$  and  $z_1z_3$  – as well as many other nonlinear functions of  $\mathbf{z}$  – are partially correlated with  $y_2z_1$ . In other words, the linear projection of  $y_2z_1$  onto  $1, z_1, z_2, z_3, z_1^2$  and  $z_1z_3$  will – except by fluke – depend on at least one of the last two terms. In any case, we can test this using OLS with  $y_2z_1$  as the dependent variable and a heteroskedasticity-robust test of two exclusion restrictions. Identification of the second equation is no problem, as  $z_3$  is always available as an IV for  $y_2$ . To enhance efficiency when  $\gamma_{13} \neq 0$ , we could add  $z_1^2$  and  $z_1z_3$  (say) to the instrument list.

e. We could use IVs  $(1, z_1, z_2, z_3, z_1^2, z_1z_3)$  in estimating the equation

$$y_1 = \delta_{10} + \gamma_{12}y_2 + \gamma_{13}y_2z_1 + \delta_{11}z_1 + \delta_{12}z_2 + u_1$$

by 2SLS, which implies a single overidentifying restriction. We can add other IVs –  $z_2^2, z_3^2, z_1z_2$ , and  $z_2z_3$  seem natural – or even reciprocals, such as  $1/z_1$  (or  $1/(1 + |z_1|)$  if  $z_1$  can equal zero).

f. We can use the instruments in part e for *both* equations. With a large sample size we might expand the list of IVs as discussed in part e.

g. Technically, the parameters in the first equation can be consistently estimated if  $\gamma_{13} \neq 0$  because  $E(y_2|\mathbf{z})$  is a nonlinear function of  $\mathbf{z}$ , and so  $z_1^2, z_1z_2$ , and other nonlinear functions

would generally be partially correlated with  $y_2$  and  $y_2 z_1$ . But, if  $\gamma_{13} = 0$  also,  $E(y_2|\mathbf{z})$  is linear in  $z_1$  and  $z_2$ , and additional nonlinear functions are not partially correlated with  $y_2$ ; thus, there is no instrument for  $y_2$ . Since the equation is not identified when  $\gamma_{13} = 0$  (and  $\delta_{23} = 0$ ),  $H_0 : \gamma_{13} = 0$  cannot be tested.

9.7. a. Because *alcohol* and *educ* are endogenous in the first equation, we need at least two elements in  $(\mathbf{z}_{(2)}, \mathbf{z}_{(3)})$  that are not also in  $\mathbf{z}_{(1)}$ . Ideally, we have at least one such element in  $\mathbf{z}_{(2)}$  and at least one such element in  $\mathbf{z}_{(3)}$ .

b. Let  $\mathbf{z}$  denote all nonredundant exogenous variables in the system. Then use these as instruments in a 2SLS analysis.

c. The matrix of instruments for each  $i$  is

$$\mathbf{z}_i = \begin{pmatrix} \mathbf{z}_i & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{z}_i, \text{educ}_i) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{z}_i \end{pmatrix}.$$

d.  $\mathbf{z}_{(3)} = \mathbf{z}$ . That is, we should not make any exclusion restrictions in the reduced form for *educ*.

9.8. a. I interact *nearc4* with experience and its quadratic, and the race indicator. The Stata output follows.

```
. use card
. gen educsq = educ^2
. gen nearc4exper = nearc4*exper
. gen nearc4expersq = nearc4*expersq
. gen nearc4black = nearc4*black
. reg educsq exper expersq black south smsa reg661-reg668 smsa66 nearc4
  nearc4exper nearc4expersq nearc4black, robust
```

Linear regression

Number of obs = 3010

F( 18, 2991) = 233.34  
 Prob > F = 0.0000  
 R-squared = 0.4505  
 Root MSE = 52.172

educsq	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
exper	-18.01791	1.229128	-14.66	0.000	-20.42793	-15.60789
expersq	.3700966	.058167	6.36	0.000	.2560452	.4841479
black	-21.04009	3.569591	-5.89	0.000	-28.03919	-14.04098
south	-.5738389	3.973465	-0.14	0.885	-8.36484	7.217162
smsa	10.38892	3.036816	3.42	0.001	4.434463	16.34338
reg661	-6.175308	5.574484	-1.11	0.268	-17.10552	4.754903
reg662	-6.092379	4.254714	-1.43	0.152	-14.43484	2.250083
reg663	-6.193772	4.010618	-1.54	0.123	-14.05762	1.670077
reg664	-3.413348	5.069994	-0.67	0.501	-13.35438	6.527681
reg665	-12.31649	5.439968	-2.26	0.024	-22.98295	-1.650031
reg666	-13.27102	5.693005	-2.33	0.020	-24.43362	-2.10842
reg667	-10.83381	5.814901	-1.86	0.063	-22.23542	.567801
reg668	8.427749	6.627727	1.27	0.204	-4.567616	21.42312
smsa66	-.4621454	3.058084	-0.15	0.880	-6.458307	5.534016
nearc4	-12.25914	7.012394	-1.75	0.081	-26.00874	1.490464
nearc4exper	4.192304	1.55785	2.69	0.007	1.137738	7.24687
nearc4expe~q	-.1623635	.0753242	-2.16	0.031	-.310056	-.014671
nearc4black	-4.789202	4.247869	-1.13	0.260	-13.11824	3.53984
_cons	307.212	6.617862	46.42	0.000	294.2359	320.188

. test nearc4exper nearc4expersq nearc4black

( 1) nearc4exper = 0  
 ( 2) nearc4expersq = 0  
 ( 3) nearc4black = 0

F( 3, 2991) = 3.72  
 Prob > F = 0.0110

. ivreg lwage exper expersq black south smsa reg661-reg668 smsa66  
 (educ educsq = nearc4 nearc4exper nearc4expersq nearc4black)

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs =	3010
Model	116.731381	16	7.29571132	F( 16, 2993) =	45.92
Residual	475.910264	2993	.159007773	Prob > F =	0.0000
Total	592.641645	3009	.196956346	R-squared =	0.1970
				Adj R-squared =	0.1927
				Root MSE =	.39876

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
educ	.3161298	.1457578	2.17	0.030	.0303342	.6019254
educsq	-.0066592	.0058401	-1.14	0.254	-.0181103	.0047918
exper	.0840117	.0361077	2.33	0.020	.0132132	.1548101
expersq	-.0007825	.0014221	-0.55	0.582	-.0035709	.0020058
black	-.1360751	.0455727	-2.99	0.003	-.2254322	-.0467181

south	-.141488	.0279775	-5.06	0.000	-.1963451	-.0866308
smsa	.1072011	.0290324	3.69	0.000	.0502755	.1641267
reg661	-.1098848	.0428194	-2.57	0.010	-.1938432	-.0259264
reg662	.0036271	.0325364	0.11	0.911	-.0601688	.0674231
reg663	.0428246	.0315082	1.36	0.174	-.0189554	.1046045
reg664	-.0639842	.0391843	-1.63	0.103	-.1408151	.0128468
reg665	.0480365	.0445934	1.08	0.281	-.0394003	.1354734
reg666	.0672512	.0498043	1.35	0.177	-.0304028	.1649052
reg667	.0347783	.0471451	0.74	0.461	-.0576617	.1272183
reg668	-.1933844	.0512395	-3.77	0.000	-.2938526	-.0929161
smsa66	.0089666	.0222745	0.40	0.687	-.0347083	.0526414
_cons	2.610889	.9706341	2.69	0.007	.7077116	4.514067

---

Instrumented: educ educsq  
Instruments: exper expersq black south smsa reg661 reg662 reg663 reg664  
reg665 reg666 reg667 reg668 smsa66 nearc4 nearc4exper  
nearc4expersq nearc4black

---

The heteroskedasticity-robust Wald test, reported in the form of an  $F$  statistic, shows that the three interaction terms are partially correlated with  $educ^2$ : the  $p$ -value = .011. (Whether the partial correlation is strong enough is a reasonable concern.)

The 2SLS estimate of  $\beta_{educ^2}$  is  $-.0067$  with  $t = -1.14$ . Without stronger evidence, we can safely leave  $educ^2$  out of the wage equation

b. If  $E(u_2|\mathbf{z}) = 0$ , as we would typically assume, than any function of  $\mathbf{z}$  is uncorrelated with  $u_1$ , including interactions of the form  $black \cdot z_j$  for any exogenous variable  $z_j$ . Such interactions are likely to be correlated with  $black \cdot educ$  if  $z_j$  is correlated with  $educ$ .

c. The 2SLS estimates, first using  $black \cdot nearc4$  as the IV for  $black \cdot educ$  and then using  $black \cdot \widehat{educ}$  as the IV, are given by the Stata output. The heteroskedasticity-robust standard errors are computed. The standard error using  $black \cdot \widehat{educ}$  as the IV is much smaller than the standard error using  $black \cdot nearc4$  as the IV. (The point estimate is also substantially higher.)

```
. ivreg lwage exper expersq black south smsa reg661-reg668 smsa66
      (educ blackeduc = nearc4 nearc4black), robust
```

Instrumental variables (2SLS) regression	Number of obs =	3010
	F( 16, 2993) =	52.35
	Prob > F =	0.0000
	R-squared =	0.2435
	Root MSE =	.38702

lwage	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
educ	.1273557	.0561622	2.27	0.023	.0172352	.2374762
blackeduc	.0109036	.0399278	0.27	0.785	-.0673851	.0891923
exper	.1059116	.0249463	4.25	0.000	.0569979	.1548253
expersq	-.0022406	.0004902	-4.57	0.000	-.0032017	-.0012794
black	-.282765	.5012131	-0.56	0.573	-1.265522	.6999922
south	-.1424762	.0298942	-4.77	0.000	-.2010914	-.083861
smsa	.1111555	.0310592	3.58	0.000	.050256	.1720551
reg661	-.1103479	.0418554	-2.64	0.008	-.1924161	-.0282797
reg662	-.0081783	.0339196	-0.24	0.809	-.0746863	.0583298
reg663	.0382413	.0335008	1.14	0.254	-.0274456	.1039283
reg664	-.0600379	.0398032	-1.51	0.132	-.1380824	.0180066
reg665	.0337805	.0519109	0.65	0.515	-.0680042	.1355652
reg666	.0498975	.0559569	0.89	0.373	-.0598204	.1596155
reg667	.0216942	.0528376	0.41	0.681	-.0819075	.1252959
reg668	-.1908353	.0506182	-3.77	0.000	-.2900853	-.0915853
smsa66	.0180009	.0205709	0.88	0.382	-.0223337	.0583356
_cons	3.84499	.9545666	4.03	0.000	1.973317	5.716663

Instrumented: educ blackeduc  
Instruments: exper expersq black south smsa reg661 reg662 reg663 reg664 reg665 reg666 reg667 reg668 smsa66 nearc4 nearc4black

```
. ivreg lwage exper expersq black south smsa reg661-reg668 smsa66
      (educ blackeduc = nearc4 blackeduc), robust
```

Instrumental variables (2SLS) regression	Number of obs =	3010
	F( 16, 2993) =	52.52
	Prob > F	= 0.0000
	R-squared	= 0.2501
	Root MSE	= .38535

lwage	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
educ	.1178141	.0554036	2.13	0.034	.0091811	.226447
blackeduc	.035984	.0105707	3.40	0.001	.0152573	.0567106
exper	.1004843	.0241951	4.15	0.000	.0530436	.147925
expersq	-.0020235	.0003597	-5.63	0.000	-.0027288	-.0013183
black	-.5955669	.1587782	-3.75	0.000	-.9068923	-.2842415
south	-.1374265	.0294259	-4.67	0.000	-.1951236	-.0797294
smsa	.1096541	.0306748	3.57	0.000	.0495083	.1697998
reg661	-.1161759	.0409317	-2.84	0.005	-.196433	-.0359189
reg662	-.0107817	.0335743	-0.32	0.748	-.0766127	.0550494
reg663	.0331736	.0326007	1.02	0.309	-.0307484	.0970955
reg664	-.064916	.0388398	-1.67	0.095	-.1410715	.0112395
reg665	.023022	.0505787	0.46	0.649	-.0761506	.1221946
reg666	.0379568	.0534653	0.71	0.478	-.0668757	.1427892
reg667	.0100466	.0513629	0.20	0.845	-.0906637	.1107568
reg668	-.1907066	.0502527	-3.79	0.000	-.2892399	-.0921733
smsa66	.0167814	.0203639	0.82	0.410	-.0231472	.0567101
_cons	4.00836	.9416251	4.26	0.000	2.162062	5.854658



```
Instrumented:  educ blackeduc
Instruments:  exper expersq black south smsa reg661 reg662 reg663 reg664
              reg665 reg666 reg667 reg668 smsa66 nearc4 blackeduchat
```

---

d. Suppose  $E(educ|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2$  and  $\text{Var}(u_1|\mathbf{z}) = \sigma_1^2$ . Then by Theorem 8.5, the optimal IVs for *educ* and *black* • *educ* are

$$\begin{aligned}\sigma_1^{-2}E(educ|\mathbf{z}) &= \sigma_1^{-2}\mathbf{z}\boldsymbol{\pi}_2 \\ \sigma_1^{-2}E(black \cdot educ|\mathbf{z}) &= \sigma_1^{-2}black \cdot E(educ|\mathbf{z}) = \sigma_1^{-2}black \cdot (\mathbf{z}\boldsymbol{\pi}_2).\end{aligned}$$

We can drop the constant  $\sigma_1^{-2}$ , and so the optimal IVs can be taken to be

$$[\mathbf{z}_1, \mathbf{z}\boldsymbol{\pi}_2, black \cdot (\mathbf{z}\boldsymbol{\pi}_2)].$$

When we operationalize this procedure, we can use

$$[\mathbf{z}_{i1}, \mathbf{z}_i\hat{\boldsymbol{\pi}}_2, black_i \cdot (\mathbf{z}_i\hat{\boldsymbol{\pi}}_2)] = [\mathbf{z}_{i1}, \widehat{educ_i}, black_i \cdot \widehat{educ_i}]$$

as the optimal IVs. Nothing is lost asymptotically by including  $black_i \cdot \widehat{educ_i}$  in the reduced form for  $educ_i$  along with  $\mathbf{z}_i$ . So using 2SLS with IVs  $[\mathbf{z}_i, black_i \cdot \widehat{educ_i}]$  produces the asymptotically efficient IV estimator.

**9.9.** a. The Stata output for 3SLS estimation of (9.28) and (9.29), along with the output for 2SLS on each equation, is given below. For coefficients that are statistically significant, the 3SLS and 2SLS are reasonably close. For coefficients estimated imprecisely, there are some differences between 2SLS and 3SLS, but these are not unexpected. Generally, the 3SLS standard errors are smaller (but recall that none of the standard errors are robust to heteroskedasticity).

```
. reg3 (hours lwage educ age kidslt6 kidsge6 nwifeinc)
      (lwage hours educ exper expersq)
```

Three-stage least-squares regression

Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
hours	428	6	1368.362	-2.1145	34.54	0.0000

```
lwage          428      4      .6892584      0.0895      79.87      0.0000
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
hours						
lwage	1676.933	431.169	3.89	0.000	831.8577	2522.009
educ	-205.0267	51.84729	-3.95	0.000	-306.6455	-103.4078
age	-12.28121	8.261529	-1.49	0.137	-28.47351	3.911094
kidslt6	-200.5673	134.2685	-1.49	0.135	-463.7287	62.59414
kidsge6	-48.63986	35.95137	-1.35	0.176	-119.1032	21.82352
nwifeinc	.3678943	3.451518	0.11	0.915	-6.396957	7.132745
_cons	2504.799	535.8919	4.67	0.000	1454.47	3555.128
lwage						
hours	.000201	.0002109	0.95	0.340	-.0002123	.0006143
educ	.1129699	.0151452	7.46	0.000	.0832858	.1426539
exper	.0208906	.0142782	1.46	0.143	-.0070942	.0488753
expersq	-.0002943	.0002614	-1.13	0.260	-.0008066	.000218
_cons	-.7051103	.3045904	-2.31	0.021	-1.302097	-.1081241

Endogenous variables: hours lwage

Exogenous variables: educ age kidslt6 kidsge6 nwifeinc exper expersq

```
. ivreg hours educ age kidslt6 kidsge6 nwifeinc (lwage = exper expersq)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs =	428
Model	-456272250	6	-76045375	F( 6, 421) =	3.41
Residual	713583270	421	1694972.14	Prob > F =	0.0027
Total	257311020	427	602601.92	R-squared =	
				Adj R-squared =	
				Root MSE =	1301.

hours	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
lwage	1544.819	480.7387	3.21	0.001	599.8713	2489.766
educ	-177.449	58.1426	-3.05	0.002	-291.7349	-63.16302
age	-10.78409	9.577347	-1.13	0.261	-29.60946	8.041289
kidslt6	-210.8339	176.934	-1.19	0.234	-558.6179	136.9501
kidsge6	-47.55708	56.91786	-0.84	0.404	-159.4357	64.3215
nwifeinc	-9.249121	6.481116	-1.43	0.154	-21.9885	3.490256
_cons	2432.198	594.1719	4.09	0.000	1264.285	3600.111

Instrumented: lwage

Instruments: educ age kidslt6 kidsge6 nwifeinc exper expersq

```
. ivreg lwage educ exper expersq (hours = age kidslt6 kidsge6 nwifeinc)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs =	428
Model	24.8336445	4	6.20841113	F( 4, 423) =	18.80
				Prob > F =	0.0000

Residual		198.493796	423	.469252474	R-squared	=	0.1112
-----							
Total		223.327441	427	.523015084	Adj R-squared	=	0.1028
					Root MSE	=	.68502

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
hours	.0001608	.0002154	0.75	0.456	-.0002626	.0005842
educ	.1111175	.0153319	7.25	0.000	.0809814	.1412536
exper	.032646	.018061	1.81	0.071	-.0028545	.0681465
expersq	-.0006765	.0004426	-1.53	0.127	-.0015466	.0001935
_cons	-.69279	.3066002	-2.26	0.024	-1.29544	-.0901403

Instrumented: hours

Instruments: educ exper expersq age kidslt6 kidsge6 nwifeinc

More efficient GMM estimators can be obtained, along with robust standard errors. The weighting matrix is the optimal one that allows for system heteroskedasticity. The (valid) standard errors from the GMM estimation are quite a bit larger than the usual 3SLS standard errors. The coefficient estimates change some, but the magnitudes are similar and all qualitative conclusions hold.

```
. gmm (hours - {b1}*lwage - {b2}*educ - {b3}*age - {b4}*kidslt6
      - {b5}*kidsge6 - {b6}*nwifeinc - {b7})
      (lwage - {b8}*hours - {b9}*educ - {b10}*age - {b11}*exper
      - {b12}*expersq - {b13}),
      instruments(educ age kidslt6 kidsge6 nwifeinc exper
      expersq) winitial(identity)
```

Step 1

```
Iteration 0: GMM criterion Q(b) = 1.339e+11
Iteration 1: GMM criterion Q(b) = 350.92722
Iteration 2: GMM criterion Q(b) = 350.92722
```

Step 2

```
Iteration 0: GMM criterion Q(b) = .08810078
Iteration 1: GMM criterion Q(b) = .00317682
Iteration 2: GMM criterion Q(b) = .00317682
```

GMM estimation

Number of parameters = 13

Number of moments = 16

Initial weight matrix: Identity

Number of obs = 428

GMM weight matrix: Robust

	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval	
/b1	1606.63	618.7605	2.60	0.009	393.882	2819.379

/b2	-179.4037	69.66358	-2.58	0.010	-315.9418	-42.86564
/b3	-10.29206	10.88031	-0.95	0.344	-31.61708	11.03295
/b4	-238.2421	212.7398	-1.12	0.263	-655.2044	178.7202
/b5	-46.23438	59.7293	-0.77	0.439	-163.3017	70.83289
/b6	-10.39225	5.576519	-1.86	0.062	-21.32203	.5375285
/b7	2385.333	630.2282	3.78	0.000	1150.108	3620.558
/b8	.0002704	.0003012	0.90	0.369	-.00032	.0008608
/b9	.1153863	.0158272	7.29	0.000	.0843655	.146407
/b10	.0031491	.00654	0.48	0.630	-.0096691	.0159673
/b11	.0336343	.0246234	1.37	0.172	-.0146266	.0818952
/b12	-.0008599	.0006453	-1.33	0.183	-.0021247	.000405
/b13	-.9910496	.5850041	-1.69	0.090	-2.137637	.1555373

---

Instruments for equation 1: educ age kidslt6 kidsge6 nwifeinc exper expersq  
\_cons

Instruments for equation 2: educ age kidslt6 kidsge6 nwifeinc exper expersq  
\_cons

b. Using  $\gamma$  as coefficients on endogenous variables, the three-equation system can be expressed as

$$\begin{aligned}
hours &= \gamma_{12}lwage + \gamma_{13}educ + \delta_{11} + \delta_{12}age + \delta_{13}kidslt6 + \delta_{14}kidsge6 + \delta_{15}nwifeinc + u_1 \\
lwage &= \gamma_{21}hours + \gamma_{23}educ + \delta_{21} + \delta_{12}age + \delta_{22}exper + \delta_{23}exper^2 + u_2 \\
educ &= \delta_{31} + \delta_{32}age + \delta_{33}kidslt6 + \delta_{34}kidsge6 + \delta_{35}nwifeinc + \delta_{36}exper + \delta_{37}exper^2 \\
&\quad + \delta_{38}motheduc + \delta_{39}fatheduc + \delta_{3,10}huseduc + u_3
\end{aligned}$$

The IVs for the first equation are all appearing in the third equation, plus *educ*. For the second and third equations, the valid IVs are all variables appearing in the third equation (which is also a reduced form for *educ*).

The following Stata output produces the GMM estimates using the optimal weighting matrix, along with the valid standard errors.

```
. gmm (hours - {b1}*lwage - {b2}*educ - {b3}*age - {b4}*kidslt6 -
      {b5}*kidsge6 - {b6}*nwifeinc - {b7})
      (lwage - {b8}*hours - {b9}*educ - {b10}*age - {b11}*exper
      - {b12}*expersq - {b13})
      (educ - {b14}*age - {b15}*kidslt6 - {b16}*kidsge6 - {b17}*nwifeinc
      - {b18}*exper - {b19}*expersq - {b20}*motheduc - {b21}*fatheduc
      - {b22}*huseduc - {b23}),
      instruments(age kidslt6 kidsge6 nwifeinc exper expersq
      motheduc fatheduc huseduc)
      instruments(1: educ) winitial(identity)
```

```
Step 1
Iteration 0:   GMM criterion Q(b) = 1.344e+11
Iteration 1:   GMM criterion Q(b) = 116344.22
```

Iteration 2: GMM criterion Q(b) = 116344.22

Step 2

Iteration 0: GMM criterion Q(b) = .22854932

Iteration 1: GMM criterion Q(b) = .0195166

Iteration 2: GMM criterion Q(b) = .0195166

GMM estimation

Number of parameters = 23

Number of moments = 31

Initial weight matrix: Identity

Number of obs = 428

GMM weight matrix: Robust

	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval	
/b1	1195.924	383.1797	3.12	0.002	444.9052	1946.942
/b2	-140.7006	46.54044	-3.02	0.003	-231.9181	-49.48298
/b3	-9.160326	8.687135	-1.05	0.292	-26.1868	7.866146
/b4	-298.5054	169.207	-1.76	0.078	-630.1449	33.13422
/b5	-72.19858	43.44575	-1.66	0.097	-157.3507	12.95353
/b6	-6.128558	3.9233	-1.56	0.118	-13.81809	1.560969
/b7	2297.594	511.7623	4.49	0.000	1294.559	3300.63
/b8	.0002455	.0002852	0.86	0.389	-.0003134	.0008044
/b9	.1011445	.0233025	4.34	0.000	.0554724	.1468167
/b10	.0007376	.0061579	0.12	0.905	-.0113317	.0128069
/b11	.031704	.0194377	1.63	0.103	-.0063933	.0698013
/b12	-.000695	.0003921	-1.77	0.076	-.0014635	.0000735
/b13	-.6841501	.6119956	-1.12	0.264	-1.883639	.5153392
/b14	-.0061529	.0135359	-0.45	0.649	-.0326827	.0203769
/b15	.5229541	.209776	2.49	0.013	.1118006	.9341075
/b16	-.1128699	.070775	-1.59	0.111	-.2515864	.0258465
/b17	.0273278	.0093602	2.92	0.004	.0089821	.0456735
/b18	.0284576	.0333103	0.85	0.393	-.0368295	.0937447
/b19	-.0001177	.0010755	-0.11	0.913	-.0022256	.0019902
/b20	.1226775	.0302991	4.05	0.000	.0632923	.1820627
/b21	.0976753	.0284494	3.43	0.001	.0419155	.153435
/b22	.3352722	.0357583	9.38	0.000	.2651873	.4053572
/b23	5.858645	.7413707	7.90	0.000	4.405585	7.311704

Instruments for equation 1: educ age kidslt6 kidsge6 nwifeinc exper expersq

motheduc fatheduc huseduc \_cons

Instruments for equation 2: age kidslt6 kidsge6 nwifeinc exper expersq

motheduc fatheduc huseduc \_cons

Instruments for equation 3: age kidslt6 kidsge6 nwifeinc exper expersq

motheduc fatheduc huseduc \_cons

**9.10.** a. No. 2SLS estimation of the first equation uses all nonredundant elements of  $\mathbf{z}_1$  and  $\mathbf{z}_2$  – call these  $\mathbf{z}$  – in the first stage regression for  $y_2$ . Therefore, the exclusion restrictions in the second equation are not imposed.

b. No, except in the extremely rare case where the covariance between the structural errors is estimated to be zero. (If we impose a zero covariance, then the 2SLS estimates and 3SLS estimates will be the same.) Effectively, each equation – including the second – is overidentified.

c. This just follows from the first two parts. Because 2SLS puts no restrictions on the reduced form for  $y_2$ , whereas 3SLS assumes only  $z_2$  appears in the reduced form for  $y_2$ , 2SLS will be more robust for estimating the parameters in the first equation.

**9.11.** a. Because  $z_2$  and  $z_3$  are both omitted from the first equation, we just need  $\delta_{22} \neq 0$  or  $\delta_{23} \neq 0$ . The second equation is identified if and only if  $\delta_{11} \neq 0$ .

b. After substitution and straightforward algebra, it can be seen that  $\pi_{11} = \delta_{11}/(1 - \gamma_{12}\gamma_{21})$ .

c. We can estimate the system by 3SLS; for the second equation, this is identical to 2SLS since it is just identified. Or, we could just use 2SLS on each equation. Given  $\hat{\delta}_{11}$ ,  $\hat{\gamma}_{12}$ , and  $\hat{\gamma}_{21}$ , we would form  $\hat{\pi}_{11} = \hat{\delta}_{11}/(1 - \hat{\gamma}_{12}\hat{\gamma}_{21})$ .

d. Whether we estimate the parameters by 2SLS or 3SLS, we will generally inconsistently estimate  $\delta_{11}$  and  $\gamma_{12}$ . (We are estimating the second equation by 2SLS so we will still consistently estimate  $\gamma_{21}$  provided we have not misspecified this equation.) So our estimate of  $\pi_{11} = \partial E(y_2|z)/\partial z_1$  will be inconsistent in any case.

e. We can just estimate the reduced form  $E(y_2|z_1, z_2, z_3)$  by ordinary least squares.

f. Consistency of OLS for  $\pi_{11}$  does not hinge on the validity of the exclusion restrictions in the structural model, whereas using an SEM does. Of course, if the SEM is correctly specified, we obtain a more efficient estimator of the reduced form parameters by imposing the restrictions in estimating  $\pi_{11}$ .

**9.12.** a. Generally,  $E(y_2^2|z) = \text{Var}(y_2|z) + [E(y_2|z)]^2$ ; when  $\gamma_{13} = 0$  and  $u_1$  and  $u_2$  are

homoskedastic,  $\text{Var}(y_2|\mathbf{z})$  is constant, say  $\tau_2^2$ . (This is easily seen from the reduced form for  $y_2$ , which is linear when  $\gamma_{13} = 0$ .) Therefore,  $E(y_2^2|\mathbf{z}) = \tau_2^2 + (\pi_{20} + \mathbf{z}\pi_2)^2$ .

b. We do not really need to use part a; in fact, it turns out to be a red herring for this problem. Since  $\gamma_{13} = 0$ ,

$$\begin{aligned} E(y_1|\mathbf{z}) &= \delta_{10} + \gamma_{12}E(y_2|\mathbf{z}) + \mathbf{z}_1\delta_1 + E(u_1|\mathbf{z}) \\ &= \delta_{10} + \gamma_{12}E(y_2|\mathbf{z}) + \mathbf{z}_1\delta_1. \end{aligned}$$

c. When  $\gamma_{13} = 0$ , *any* nonlinear function of  $\mathbf{z}$ , including  $(\pi_{20} + \mathbf{z}\pi_2)^2$ , has zero coefficient in  $E(y_1|\mathbf{z}) = \delta_{10} + \gamma_{12}(\pi_{20} + \mathbf{z}\pi_2) + \mathbf{z}_1\delta_1$ . Plus, if  $\gamma_{13} = 0$ , then the parameters  $\pi_{20}$  and  $\pi_2$  are consistently estimated from the first stage regression  $y_{i2}$  on  $1, \mathbf{z}_1, i = 1, \dots, N$ . Therefore, the regression  $y_{i1}$  on  $1, (\hat{\pi}_{20} + \mathbf{z}_i\hat{\pi}_2), (\hat{\pi}_{20} + \mathbf{z}_i\hat{\pi}_2)^2, \mathbf{z}_{i1}, i = 1, \dots, N$  consistently estimates  $\delta_{10}, \gamma_{12}, 0$ , and  $\delta_1$ , respectively. But this is just the regression  $y_{i1}$  on  $1, \hat{y}_{i2}, (\hat{y}_{i2})^2, \mathbf{z}_{i1}, i = 1, \dots, N$ .

d. Because  $E(u_1|\mathbf{z}) = 0$  and  $\text{Var}(u_1|\mathbf{z}) = \sigma_1^2$ , we can immediately apply Theorem 8.5 to conclude that the optimal IVS for estimating the first equation are  $[1, E(y_2|\mathbf{z}), E(y_2^2|\mathbf{z}), \mathbf{z}_1]/\sigma_1^2$ , and we can drop the division by  $\sigma_1^2$ . But, if  $\gamma_{13} = 0$ , then  $E(y_2|\mathbf{z})$  is linear in  $\mathbf{z}$  and, from part a,  $E(y_2^2|\mathbf{z}) = \tau_2^2 + [E(y_2|\mathbf{z})]^2$ . So the optimal IVs are a linear combination of  $\{1, E(y_2|\mathbf{z}), [E(y_2|\mathbf{z})]^2, \mathbf{z}_1\}$ , which means they are a linear combination of  $\{1, \mathbf{z}, [E(y_2|\mathbf{z})]^2\}$ . We never do worse asymptotically by using more IVs, so we can use  $\{1, \mathbf{z}, [E(y_2|\mathbf{z})]^2\}$  as an optimal set. Why would we use this larger set instead of  $\{1, E(y_2|\mathbf{z}), [E(y_2|\mathbf{z})]^2, \mathbf{z}_1\}$ ? For one, the larger set will generally yield overidentifying restrictions. In addition, if  $\gamma_{13} \neq 0$ , we will generally be better off using more instruments:  $\mathbf{z}$  rather than only  $L(y_2|1, \mathbf{z})$ .

e. The estimates below are similar to those reported in Section 9.5.2, where we just added  $educ^2, age^2$ , and  $nwifeinc^2$  to the IV list and using 2SLS with  $lwage = \log(wage)$  and  $lwagesq = [\log(wage)]^2$  as endogenous explanatory variables. In particular, the coefficient on

*lwagesq* is still statistically insignificant. The standard errors reported here are robust to

heteroskedasticity (unlike in the text).

```
. gen lwagesq = lwage^2

. qui reg lwage educ age kidslt6 kidsge6 nwifeinc exper expersq

. predict lwagehat
(option xb assumed; fitted values)

. gen lwagehatsq = lwagehat^2

. ivreg hours (lwage lwagesq = exper expersq lwagehatsq)
      educ age kidslt6 kidsge6 nwifeinc, robust
```

Instrumental variables (2SLS) regression	Number of obs =	428
	F( 7, 420) =	2.88
	Prob > F =	0.0059
	R-squared =	
	Root MSE =	1177.

hours	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
lwage	1846.902	856.1346	2.16	0.032	164.0599	3529.745
lwagesq	-373.16	401.9586	-0.93	0.354	-1163.261	416.9412
educ	-103.2347	71.99564	-1.43	0.152	-244.7513	38.28199
age	-9.425115	8.848798	-1.07	0.287	-26.81856	7.968333
kidslt6	-187.0236	177.1703	-1.06	0.292	-535.2747	161.2274
kidsge6	-55.70163	45.86915	-1.21	0.225	-145.8633	34.46007
nwifeinc	-7.5979	4.491138	-1.69	0.091	-16.42581	1.230009
_cons	1775.847	881.2631	2.02	0.045	43.61095	3508.082

```
Instrumented: lwage lwagesq
Instruments: educ age kidslt6 kidsge6 nwifeinc exper expersq lwagehatsq
```

```
. gen educsq = educ^2

. gen agesq = age^2

. gen nwifeincsq = nwifeinc^2

. ivreg hours (lwage lwagesq = exper expersq educsq agesq nwifeincsq)
      educ age kidslt6 kidsge6 nwifeinc, robust
```

Instrumental variables (2SLS) regression	Number of obs =	428
	F( 7, 420) =	3.53
	Prob > F =	0.0011
	R-squared =	
	Root MSE =	1161

	Robust
--	--------



hours	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
lwage	1873.62	792.3005	2.36	0.018	316.2521	3430.989
lwagesq	-437.2911	313.7658	-1.39	0.164	-1054.038	179.4559
educ	-87.8511	50.01226	-1.76	0.080	-186.1566	10.45442
age	-9.142303	8.512639	-1.07	0.283	-25.87499	7.590381
kidslt6	-185.0554	179.9036	-1.03	0.304	-538.6789	168.5682
kidsge6	-58.18949	44.6767	-1.30	0.193	-146.0073	29.6283
nwifeinc	-7.233422	4.037903	-1.79	0.074	-15.17044	.7035957
_cons	1657.926	671.9361	2.47	0.014	337.1489	2978.702

---

Instrumented: lwage lwagesq  
Instruments: educ age kidslt6 kidsge6 nwifeinc exper expersq educsq agesq  
nwifeincsq

---

9.13. a. The first equation is identified if, and only if,  $\delta_{22} \neq 0$  (the rank condition).

b. Here is the Stata output:

```
. use openness
. reg open lpcinc lland, robust
```

Linear regression

Number of obs =	114
F( 2, 111) =	22.22
Prob > F	= 0.0000
R-squared	= 0.4487
Root MSE	= 17.796

---

open	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
lpcinc	.5464812	1.436115	0.38	0.704	-2.299276	3.392238
lland	-7.567103	1.141798	-6.63	0.000	-9.829652	-5.304554
_cons	117.0845	18.24808	6.42	0.000	80.92473	153.2443

---

With  $t = -6.63$ , we can conclude that it shows that  $\log(\text{lland})$  is very statistically significant in the reduced form for *open*. The negative coefficient implies that smaller countries are more “open.”

c. Here is the Stata output. First 2SLS, the OLS, both with heteroskedasticity-robust standard errors.

```
. ivreg inf (open = lland) lpcinc, robust
```

Instrumental variables (2SLS) regression

Number of obs =	114
F( 2, 111) =	2.53
Prob > F	= 0.0844

R-squared = 0.0309  
Root MSE = 23.836

inf	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
open	-.3374871	.1524489	-2.21	0.029	-.6395748	-.0353994
lpcinc	.3758247	1.378542	0.27	0.786	-2.355848	3.107497
_cons	26.89934	10.9199	2.46	0.015	5.260821	48.53785

Instrumented: open  
Instruments: lpcinc lland

. reg inf open lpcinc, robust

Linear regression

Number of obs = 114  
F( 2, 111) = 3.84  
Prob > F = 0.0243  
R-squared = 0.0453  
Root MSE = 23.658

inf	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
open	-.2150695	.0794571	-2.71	0.008	-.3725191	-.0576199
lpcinc	.0175683	1.278747	0.01	0.989	-2.516354	2.55149
_cons	25.10403	9.99078	2.51	0.013	5.306636	44.90143

The IV estimate is larger in magnitude – by more than 50% – but its standard error is almost twice as large as the OLS standard error. There is some but not overwhelming evidence that *open* is actually endogenous. The variable-addition Hausman test, made robust to heteroskedasticity, has  $t = 1.48$ . With  $N = 114$ , we might not expect very strong evidence.

d. If we add  $\gamma_{13}open^2$  to the equation, we need an IV for it. Since  $\log(land)$  is partially correlated with *open*,  $[\log(land)]^2$  is a natural candidate. A regression of  $open^2$  on  $\log(land)$ ,  $[\log(land)]^2$ , and  $\log(pcinc)$  gives a heteroskedasticity-robust  $t$  statistic on  $[\log(land)]^2$  of about 2. This is borderline, but we will go ahead. The Stata output for 2SLS is

. ivreg inf (open opensq = lland llandsq) lpcinc, robust

Instrumental variables (2SLS) regression

Number of obs = 114  
F( 3, 110) = 1.44  
Prob > F = 0.2350

R-squared =  
Root MSE = 24.

inf	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
open	-1.198637	.7139934	-1.68	0.096	-2.613604	.2163303
opensq	.0075781	.0053779	1.41	0.162	-.0030796	.0182358
lpcinc	.5066092	1.490845	0.34	0.735	-2.447896	3.461114
_cons	43.17124	18.37223	2.35	0.021	6.761785	79.5807

Instrumented: open opensq  
Instruments: lpcinc lland llandsq

The squared term indicates that the impact of *open* on *inf* diminishes; the estimate would be significant at about the 8.1% level against a one-sided alternative.

e. Here is the Stata output for implementing the method described in the problem:

```
. qui reg open lpcinc lland
. predict openhat
(option xb assumed; fitted values)
. gen openhatsq = openhat^2
. reg inf openhat openhatsq lpcinc, robust
```

Linear regression

Number of obs =	114
F( 3, 110) =	2.52
Prob > F =	0.0620
R-squared =	0.0575
Root MSE =	23.612

inf	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
openhat	-.8648092	.5762007	-1.50	0.136	-2.006704	.2770854
openhatsq	.0060502	.0050906	1.19	0.237	-.0040383	.0161387
lpcinc	.0412172	1.293368	0.03	0.975	-2.521935	2.604369
_cons	39.17831	15.99614	2.45	0.016	7.477717	70.8789

Qualitatively, the results are similar to the appropriate IV method from part d, but the coefficient on *openhat* is quite a bit smaller in magnitude using the “forbidden regression.” If  $\gamma_{13} = 0$ ,  $E(\text{open}|\text{lpcinc}, \text{lland})$  is linear, and  $\text{Var}(\text{open}|\text{lpcinc}, \text{lland})$  is constant then, as shown

in Problem 9.12, both methods are consistent. But the forbidden regression implemented in this part is unnecessary, less robust, and we cannot trust the standard errors, anyway.

Incidentally, using *openhatsq* as an IV, rather than a regressor, gives very similar estimates to using *llandsq* as an IV for *opensq*.

**9.14.** a. Agree. In equation (9.13), the reduced form variance matrix,  $\Lambda = E(\mathbf{v}'\mathbf{v})$  is always identified. Now the structural variance matrix can be written as  $\Sigma = \Gamma'\Lambda\Gamma$ , so if  $\Gamma$  and  $\Lambda$  are both identified, so is  $\Sigma$ .

b. Disagree. In many cases a linear version of the model is not identified because there are not enough instruments. In that case, identification of the more general model hinges on the model actually having nonlinearities, a tenuous situation.

c. Disagree.  $E(\mathbf{u}|\mathbf{z}) = \mathbf{0}$  implies that  $z_h$  is uncorrelated with  $u_g$  for all  $h = 1, \dots, L$  and  $g = 1, \dots, G$ . This kind of orthogonality, along with the rank condition, is sufficient for the GMM, traditional, or GIV versions of 3SLS to be consistent. We need not restrict  $\text{Var}(\mathbf{u}|\mathbf{z})$  for consistency, and robust inference is easily obtained.

d. Disagree. Even true SEMs can have other problems that cause endogeneity, name, omitted variables and measurement error. In these cases, some variables may be valid instruments in one equation but not other equations.

e. Disagree. Control function approaches generally require more assumptions in order to be consistent. Take equations (9.70) and (9.71) as an example. The CF approach proposed there basically requires assumption (9.72), which can be very restrictive, especially if  $y_2$  or  $y_3$  exhibit discreteness. By contrast, we can use an IV approach that specifies  $\mathbf{z}$  and nonlinear functions of  $\mathbf{z}$  as instruments in directly estimating (9.70) (by 2SLS or GMM). When they are consistent, we might expect CF approaches to be more efficient asymptotically.

**9.15.** a. The model with  $\alpha_{13} = 0$ ,  $\gamma_{11} = \mathbf{0}$ , and  $\gamma_{12} = \mathbf{0}$  is

$$y_1 = \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_{11} y_2 + \alpha_{12} y_3 + u_3$$

and we maintain that this equation is identified. Necessary is that  $L \geq L_1 + 2$ , as is assumed in the problem. The rank condition is more complicated, and we assume it holds. In other words, we assume we have enough relevant instruments for  $y_2$  and  $y_3$ . If (9.71) also holds, and say  $E(u_2|\mathbf{z}) = E(u_3|\mathbf{z}) = 0$ , then

$$E(y_2 y_3 | \mathbf{z}) = (\mathbf{z} \boldsymbol{\beta}_2)(\mathbf{z} \boldsymbol{\beta}_3) + E(u_2 u_3 | \mathbf{z})$$

$$E(y_2 \mathbf{z}_1 | \mathbf{z}) = (\mathbf{z} \boldsymbol{\beta}_2) \mathbf{z}_1$$

$$E(y_3 \mathbf{z}_1 | \mathbf{z}) = (\mathbf{z} \boldsymbol{\beta}_3) \mathbf{z}_1$$

which means that squares and cross products in  $\mathbf{z}$  are natural instruments for the interaction terms. Note that there are plenty of these interaction terms, and they are likely to provide enough variation because the linear version of the model is identified. For example, if  $z_L$  has zero coefficient in the reduced form for  $y_3$ , then  $z_L \mathbf{z}_1$  appears in  $E(y_2 \mathbf{z}_1 | \mathbf{z})$  but not in  $E(y_3 \mathbf{z}_1 | \mathbf{z})$ . Further, suppose  $E(u_2 u_3 | \mathbf{z})$  is actually constant. then squares of elements in  $\mathbf{z}_2$ , where  $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$ , appear in  $E(y_2 y_3 | \mathbf{z})$  but not the other expectations. If  $E(u_2 u_3 | \mathbf{z})$  is not constant, it would cancel out with those squares only by fluke. Further, even if we only take (9.71) to be linear projections, that would only helps our cause because if these are not conditional expectations then even more nonlinear functions of  $\mathbf{z}$  would be useful as instruments.

b. There are no overidentification restrictions. We have the same number of instruments as explanatory variables. That is one drawback to using fitted values as IVs: there are no overidentification restrictions to test.

c. There are  $L - L_1 - 2$  overidentification restrictions. The  $L_1$  subvector  $\mathbf{z}_1$  acts as its own instruments, and we need two more elements of  $\mathbf{z}$  to instrument for  $y_2$  and  $y_3$ . The rest of the

explanatory variables are taking care of by the interaction terms.

d. We can get many overidentification restrictions by using as IVs the nonredundant elements of  $(\mathbf{z}_i, \mathbf{z}_i \otimes \mathbf{z}_i)$  – that is, the levels, squares, and cross products of  $\mathbf{z}_i$ . For example, if  $L = 5, L_1 = 2$ , and  $\mathbf{z}$  and  $\mathbf{z}_1$  include a constant, then there are  $5 + 4 + 6 = 15$  instruments and  $2 + 2 + 2 + 2 = 8$  explanatory variables.

e. To solve this problem, we also need to assume  $E(u_2 u_3 | \mathbf{z})$  is constant, and to explicitly state that  $\mathbf{z}_1$  includes a constant. By Theorem 8.5, the optimal IVs are then given in part a, along with  $\mathbf{z}_1$ , because  $\text{Var}(u_1 | \mathbf{z}) = \sigma_1^2$ . If  $E(u_2 u_3 | \mathbf{z})$  is constant then  $(\mathbf{z}\boldsymbol{\beta}_2)(\mathbf{z}\boldsymbol{\beta}_3) + E(u_2 u_3 | \mathbf{z}) = (\mathbf{z}\boldsymbol{\beta}_2)(\mathbf{z}\boldsymbol{\beta}_3) + \text{constant}$ . We operationalize the IVs by replacing  $\boldsymbol{\beta}_2$  and  $\boldsymbol{\beta}_3$  with the estimators from the first-stage regressions.

The estimators from parts c and d would also be asymptotically efficient because linear combinations of the instruments in those two parts are the optimal instruments.

Asymptotically, it does not hurt to use redundant instruments; asymptotically, the optimal linear combination will be picked out via the first-stage regression.

## Solutions to Chapter 10 Problems

10.1. a. Because investment is likely to be affected by macroeconomic factors, it is important to allow for these by including separate time intercepts; this is done by using  $T - 1$  time period dummies.

b. Putting the unobserved effect  $c_i$  in the equation is a simple way to account for time-constant features of a county that affect investment and might also be correlated with the tax variable. Something like “average” county economic climate, which affects investment, could easily be correlated with tax rates because tax rates are, at least to a certain extent, selected by state and local officials. If only a cross section were available, we would have to find an instrument for the tax variable that is uncorrelated with  $c_i$  and correlated with the tax rate. This is often a difficult task.

c. Standard investment theories suggest that, *ceteris paribus*, larger marginal tax rates decrease investment.

d. I would start with a fixed effects analysis to allow arbitrary correlation between all time-varying explanatory variables and  $c_i$ . (Actually, doing pooled OLS is a useful initial exercise; these results can be compared with those from an FE analysis). Such an analysis assumes strict exogeneity of  $z_{it}$ ,  $tax_{it}$ , and  $disaster_{it}$  in the sense that these are uncorrelated with the errors  $u_{is}$  for all  $t$  and  $s$ .

I have no strong intuition for the likely serial correlation properties of the  $\{u_{it}\}$ . These might have little serial correlation because we have allowed for  $c_i$ , in which case I would use standard fixed effects. However, it seems more likely that the  $u_{it}$  are positively autocorrelated, in which case I might use first differencing instead. In either case, I would compute the fully robust standard errors along with the usual ones. In either case we can test for serial correlation

in  $\{u_{it}\}$ .

e. If  $tax_{it}$  and  $disaster_{it}$  do not have lagged effects on investment, then the only possible violation of the strict exogeneity assumption is if future values of these variables are correlated with  $u_{it}$ . It seems reasonable not to worry whether future natural disasters are determined by past investment. On the other hand, state officials might look at the levels of past investment in determining future tax policy, especially if there is a target level of tax revenue the officials are trying to achieve. This could be similar to setting property tax rates: sometimes property tax rates are set depending on recent housing values because a larger base means a smaller rate can achieve the same amount of revenue. Given that we allow  $tax_{it}$  to be correlated with  $c_i$ , feedback might not be much of a problem. But it cannot be ruled out ahead of time.

**10.2.** a.  $\theta_2$ ,  $\delta_2$ , and  $\gamma$  can be consistently estimated (assuming all elements of  $\mathbf{z}_{it}$  are time-varying). The first period intercept,  $\theta_1$ , and the coefficient on  $female_1$ ,  $\delta_1$ , cannot be estimated.

b. Everything else equal,  $\theta_2$  measures the growth in wage for men over the period. This is because, if we set  $female_i = 0$  and  $\mathbf{z}_{i1} = \mathbf{z}_{i2}$ , the change in log wage is, on average,  $\theta_2$  (set  $d2_1 = 0$  and  $d2_2 = 1$ ). We can think of this as being the growth in wage rates (for males) due to aggregate factors in the economy. The parameter  $\delta_2$  measures the *difference* in wage growth rates between women and men, all else equal. If  $\delta_2 = 0$  then, for men and women with the same characteristics, average wage growth is the same.

c. Write

$$\begin{aligned}\log(wage_{i1}) &= \theta_1 + \mathbf{z}_{i1}\gamma + \delta_1 female_i + c_i + u_{i1} \\ \log(wage_{i2}) &= \theta_1 + \theta_2 + \mathbf{z}_{i2}\gamma + \delta_1 female_i + \delta_2 female_i + c_i + u_{i2},\end{aligned}$$

where I have used the fact that  $d2_1 = 0$  and  $d2_2 = 1$ . Subtracting the first equation from the



second gives

$$\Delta \log(wage_i) = \theta_2 + \Delta \mathbf{z}_i \boldsymbol{\gamma} + \delta_2 female_i + \Delta u_i.$$

This equation shows explicitly that the growth in wages depends on  $\Delta \mathbf{z}_i$  and gender. If

$\mathbf{z}_{i1} = \mathbf{z}_{i2}$  then  $\Delta \mathbf{z}_i = \mathbf{0}$ , and the growth in wage for men is  $\theta_2$  and that for women is  $\theta_2 + \delta_2$ , just as above. This shows that we can allow for  $c_i$  and still test for a gender differential in the growth of wages. But we cannot say anything about the wage differential between men and women for a given year.

**10.3.** a. Let  $\bar{\mathbf{x}}_i = (\mathbf{x}_{i1} + \mathbf{x}_{i2})/2$ ,  $\bar{y}_i = (y_{i1} + y_{i2})/2$ ,  $\ddot{\mathbf{x}}_{i1} = \mathbf{x}_{i1} - \bar{\mathbf{x}}_i$ ,  $\ddot{\mathbf{x}}_{i2} = \mathbf{x}_{i2} - \bar{\mathbf{x}}_i$ , and similarly for  $\ddot{y}_{i1}$  and  $\ddot{y}_{i2}$ . For  $T = 2$  the fixed effects estimator can be written as

$$\hat{\boldsymbol{\beta}}_{FE} = \left[ \sum_{i=1}^N (\ddot{\mathbf{x}}_{i1}' \ddot{\mathbf{x}}_{i1} + \ddot{\mathbf{x}}_{i2}' \ddot{\mathbf{x}}_{i2}) \right]^{-1} \left[ \sum_{i=1}^N (\ddot{\mathbf{x}}_{i1}' \ddot{y}_{i1} + \ddot{\mathbf{x}}_{i2}' \ddot{y}_{i2}) \right].$$

Now, by simple algebra,

$$\begin{aligned} \ddot{\mathbf{x}}_{i1} &= (\mathbf{x}_{i1} - \mathbf{x}_{i2})/2 = -\Delta \mathbf{x}_i/2 \\ \ddot{\mathbf{x}}_{i2} &= (\mathbf{x}_{i2} - \mathbf{x}_{i1})/2 = \Delta \mathbf{x}_i/2 \\ \ddot{y}_{i1} &= (y_{i1} - y_{i2})/2 = -\Delta y_i/2 \\ \ddot{y}_{i2} &= (y_{i2} - y_{i1})/2 = \Delta y_i/2 \end{aligned}$$

Therefore,

$$\begin{aligned} \ddot{\mathbf{x}}_{i1}' \ddot{\mathbf{x}}_{i1} + \ddot{\mathbf{x}}_{i2}' \ddot{\mathbf{x}}_{i2} &= \Delta \mathbf{x}_i' \Delta \mathbf{x}_i/4 + \Delta \mathbf{x}_i' \Delta \mathbf{x}_i/4 = \Delta \mathbf{x}_i' \Delta \mathbf{x}_i/2 \\ \ddot{\mathbf{x}}_{i1}' \ddot{y}_{i1} + \ddot{\mathbf{x}}_{i2}' \ddot{y}_{i2} &= \Delta \mathbf{x}_i' \Delta y_i/4 + \Delta \mathbf{x}_i' \Delta y_i/4 = \Delta \mathbf{x}_i' \Delta y_i/2 \end{aligned}$$

and so

$$\begin{aligned}\hat{\beta}_{FE} &= \left( \sum_{i=1}^N \Delta \mathbf{x}_i' \Delta \mathbf{x}_i / 2 \right)^{-1} \left( \sum_{i=1}^N \Delta \mathbf{x}_i' \Delta y_i / 2 \right) \\ &= \left( \sum_{i=1}^N \Delta \mathbf{x}_i' \Delta \mathbf{x}_i \right)^{-1} \left( \sum_{i=1}^N \Delta \mathbf{x}_i' \Delta y_i \right) = \hat{\beta}_{FD}.\end{aligned}$$

b. Let  $\hat{u}_{i1} = \ddot{y}_{i1} - \ddot{\mathbf{x}}_{i1}' \hat{\beta}_{FE}$  and  $\hat{u}_{i2} = \ddot{y}_{i2} - \ddot{\mathbf{x}}_{i2}' \hat{\beta}_{FE}$  be the fixed effects residuals for the two time periods for cross section observation  $i$ . Since  $\hat{\beta}_{FE} = \hat{\beta}_{FD}$ , and using the representations above, we have

$$\begin{aligned}\hat{u}_{i1} &= -\Delta y_i / 2 - (-\Delta \mathbf{x}_i / 2)' \hat{\beta}_{FD} = -(\Delta y_i - \Delta \mathbf{x}_i' \hat{\beta}_{FD}) / 2 \equiv -\hat{e}_i / 2 \\ \hat{u}_{i2} &= \Delta y_i / 2 - (\Delta \mathbf{x}_i / 2)' \hat{\beta}_{FD} = (\Delta y_i - \Delta \mathbf{x}_i' \hat{\beta}_{FD}) / 2 \equiv \hat{e}_i / 2,\end{aligned}$$

where  $\hat{e}_i \equiv \Delta y_i - \Delta \mathbf{x}_i' \hat{\beta}_{FD}$  are the first difference residuals,  $i = 1, 2, \dots, N$ . Therefore,

$$\sum_{i=1}^N (\hat{u}_{i1}^2 + \hat{u}_{i2}^2) = (1/2) \sum_{i=1}^N \hat{e}_i^2.$$

This shows that the sum of squared residuals from the fixed effects regression is exactly one half the sum of squared residuals from the first difference regression. Since we know the variance estimate for fixed effects is the  $SSR$  divided by  $N - K$  (when  $T = 2$ ), and the variance estimate for first difference is the  $SSR$  divided by  $N - K$ , the error variance from fixed effects is always half the size as the error variance for first difference estimation, that is,  $\hat{\sigma}_u^2 = \hat{\sigma}_e^2 / 2$  (contrary to what the problem asks you to show). What I wanted you to show is that the variance matrix estimates of  $\hat{\beta}_{FE}$  and  $\hat{\beta}_{FD}$  are identical. This is easy since the variance matrix estimate for fixed effects is

$$\hat{\sigma}_u^2 \left[ \sum_{i=1}^N (\ddot{\mathbf{x}}_{i1}' \ddot{\mathbf{x}}_{i1} + \ddot{\mathbf{x}}_{i2}' \ddot{\mathbf{x}}_{i2}) \right]^{-1} = (\hat{\sigma}_e^2 / 2) \left( \sum_{i=1}^N \Delta \mathbf{x}_i' \Delta \mathbf{x}_i / 2 \right)^{-1} = \hat{\sigma}_e^2 \left( \sum_{i=1}^N \Delta \mathbf{x}_i' \Delta \mathbf{x}_i \right)^{-1},$$

which is the variance matrix estimator for FD estimator. Thus, the standard errors, and the fact all other test statistics ( $F$  statistics) will be numerically identical using the two approaches.

**10.4.** a. Including the aggregate time effect,  $d2_t$ , can be very important. Without it, we must assume that any change in average  $y$  over the two time periods is due to the program, and not to external factors. For example, if  $y_{it}$  is the unemployment rate for city  $i$  at time  $t$ , and  $prog_{it}$  denotes a job creation program, we want to be sure that we account for the fact that the general economy may have worsened or improved over the period. If  $d2_t$  is omitted, and  $\theta_2 < 0$  (an improving economy, since unemployment has fallen), we might attribute a decrease in unemployment to the job creation program, when in fact it had nothing to do with it. For general  $T$ , each time period should have its own intercept (otherwise the analysis is not entirely convincing).

b. The presence of  $c_i$  allows program participation to be correlated with unobserved individual heterogeneity, something crucial in contexts where the experimental group is not randomly assigned. Two examples are when individuals “self-select” into the program and when program administrators target specific groups that may benefit more or less from the program.

c. If we first difference the equation, use the fact that  $prog_{i1} = 0$  for all  $i$ ,  $d2_1 = 0$ , and  $d2_2 = 1$ , we get

$$y_{i2} - y_{i1} = \theta_2 + \delta_1 prog_{i2} + u_{i2} - u_{i1},$$

or

$$\Delta y_i = \theta_2 + \delta_1 prog_{i2} + \Delta u_i.$$

Now, the  $FE$  (and  $FD$ ) estimates of  $\theta_2$  and  $\delta_1$  are just the OLS estimators from this equation

(on cross section data). From basic two-variable regression with a dummy independent variable,  $\hat{\theta}_2$  is the average value of  $\Delta y$  over the group with  $prog_{12} = 0$  – that is, the control group. Also,  $\hat{\theta}_2$  and  $\hat{\delta}_1$  is the average value of  $\Delta y$  over the group with  $prog_{i1} = 1$  – that is, the treatment group. Thus, as asserted, we have

$$\hat{\theta}_2 = \overline{\Delta y}_{control}, \hat{\delta}_1 = \overline{\Delta y}_{treat} - \overline{\Delta y}_{control}.$$

If we did not include the  $d2_t$ ,  $\hat{\delta}_1 = \overline{\Delta y}_{treat}$ , the average change of the treated group. This demonstrates the claim in part b that without the aggregate time effect any change in the average value of  $y$  for the treated group is attributed to the program. Differencing and averaging over the treated group allows program participation to depend on time-constant unobservables affecting the level of  $y$ , but that does not account for external factors that affect  $y$  for everyone.

d. In general, for  $T$  time periods we have

$$y_{it} = \theta_1 + \theta_2 d2_t + \theta_3 d3_t + \dots + \theta_T dT_t + \delta_1 prog_{it} + c_i + u_{it};$$

that is, we have separate year intercepts, an unobserved effect  $c_i$ , and the program indicator.

e. First, the model from part d is more flexible because it allows any sequence of program participation. Equation (10.89), when extended to  $T > 2$ , applies only when treatment is ongoing. In addition, (10.89) is restrictive in terms of aggregate time effects: it assumes that any aggregate time effects correspond to the start of the program only. It is better to use the unobserved effects model from part d, and estimate it using either *FE* or *FD*.

**10.5. a.** Write  $\mathbf{v}_i \mathbf{v}_i' = c_i^2 \mathbf{j}_T \mathbf{j}_T' + \mathbf{u}_i \mathbf{u}_i' + \mathbf{j}_T (c_i \mathbf{u}_i') + (c_i \mathbf{u}_i') \mathbf{j}_T'$ . Under RE.1,  $E(\mathbf{u}_i | \mathbf{x}_i, c_i) = \mathbf{0}$ , which implies that  $E(c_i \mathbf{u}_i' | \mathbf{x}_i) = \mathbf{0}$  by iterated expectations. Under RE.3a,  $E(\mathbf{u}_i \mathbf{u}_i' | \mathbf{x}_i, c_i) = \sigma_u^2 \mathbf{I}_T$ , which implies that  $E(\mathbf{u}_i \mathbf{u}_i' | \mathbf{x}_i) = \sigma_u^2 \mathbf{I}_T$  (again, by iterated expectations). Therefore,

$$E(\mathbf{v}_i \mathbf{v}_i' | \mathbf{x}_i) = E(c_i^2 | \mathbf{x}_i) \mathbf{j}_T \mathbf{j}_T' + E(\mathbf{u}_i \mathbf{u}_i' | \mathbf{x}_i) = h(\mathbf{x}_i) \mathbf{j}_T \mathbf{j}_T' + \sigma_u^2 \mathbf{I}_T,$$

where  $h(\mathbf{x}_i) \equiv \text{Var}(c_i | \mathbf{x}_i) = E(c_i^2 | \mathbf{x}_i)$  (by RE.1b). This shows that the conditional variance matrix of  $\mathbf{v}_i$  given  $\mathbf{x}_i$  has the same covariance for all  $t \neq s$ ,  $h(\mathbf{x}_i)$ , and the same variance for all  $t$ ,  $h(\mathbf{x}_i) + \sigma_u^2$ . Therefore, while the variances and covariances depend on  $\mathbf{x}_i$ , they do not depend on time.

b. The RE estimator is still consistent and  $\sqrt{N}$ -asymptotically normal without Assumption RE.3b, but the usual random effects variance estimator of  $\text{Avar}(\hat{\boldsymbol{\beta}}_{RE})$  is no longer valid because  $E(\mathbf{v}_i \mathbf{v}_i' | \mathbf{x}_i)$  does not have the form (10.30) (because it depends on  $\mathbf{x}_i$ ). The robust variance matrix estimator given in (7.52) should be used in obtaining standard errors and Wald statistics.

**10.6.** a. By stacking the formulas for the *FD* and *FE* estimators, and using standard asymptotic arguments, we have, under *FE*.1 and the rank conditions,

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{G}^{-1} \left( N^{-1} \sum_{i=1}^N \mathbf{s}_i \right) + o_p(1),$$

where  $\mathbf{G}$  is the  $2K \times 2K$  block diagonal matrix with blocks  $\mathbf{A}_1$  and  $\mathbf{A}_2$ , respectively, and  $\mathbf{s}_i$  is the  $2K \times 1$  vector

$$\mathbf{s}_i \equiv \begin{pmatrix} \Delta \mathbf{X}_i' \Delta \mathbf{u}_i \\ \ddot{\mathbf{X}}_i' \ddot{\mathbf{u}}_i \end{pmatrix}.$$

b. Let  $\Delta \hat{\mathbf{u}}_i$  denote the  $(T-1) \times 1$  vector of *FD* residuals, and let  $\hat{\ddot{\mathbf{u}}}_i$  denote the  $T \times 1$  vector of *FE* residuals. Plugging these into the formula for  $\mathbf{s}_i$  gives  $\hat{\mathbf{s}}_i$ . Let  $\hat{\mathbf{D}} = N^{-1} \sum_{i=1}^N \hat{\mathbf{s}}_i \hat{\mathbf{s}}_i'$ , and define  $\hat{\mathbf{G}}$  by replacing  $\mathbf{A}_1$  and  $\mathbf{A}_2$  with their obvious consistent estimators. Then

$\widehat{\text{Avar}}[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})] = \hat{\mathbf{G}}^{-1} \hat{\mathbf{D}} \hat{\mathbf{G}}^{-1}$  is a consistent estimator of  $\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})]$ .

c. Let  $\mathbf{R}$  be the  $K \times 2K$  partitioned matrix  $\mathbf{R} = [\mathbf{I}_K \mid -\mathbf{I}_K]$ . Then the null hypothesis imposed by the Hausman test is  $H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{0}$ . We can form a Wald-type statistic,

$$H = (\mathbf{R}\hat{\boldsymbol{\theta}})' [\mathbf{R}\hat{\mathbf{G}}^{-1} \hat{\mathbf{D}} \hat{\mathbf{G}}^{-1} \mathbf{R}']^{-1} (\mathbf{R}\hat{\boldsymbol{\theta}}).$$

Under FE.1 and the rank conditions for *FD* and *FE*,  $H$  has a limiting  $\chi_K^2$  distribution. The statistic requires no particular second moment assumptions of the kind in FE.3. Note that

$$\mathbf{R}\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}}_{FD} - \hat{\boldsymbol{\beta}}_{FE}.$$

**10.7. a.** The random effects estimates are given below. The coefficient on *season* is  $-.044$ , which means that being in season is estimated to reduce an athlete's term GPA by  $.044$  points. The nonrobust  $t$  statistic is only  $-1.12$ .

```
. use gpa
. xtset id term
      panel variable:  id (strongly balanced)
      time variable:  term, 8808 to 8901, but with gaps
                  delta:  1 unit

. xtreg trmgpa spring crsgpa frstsem season sat verbmath hsperc hssize black

Random-effects GLS regression              Number of obs   =          732
Group variable: id                        Number of groups  =          366

R-sq:  within  = 0.2067                    Obs per group: min =
       between = 0.5390                      avg   =          2.
       overall  = 0.4785                      max   =

Random effects u_i ~Gaussian              Wald chi2(10)     =       512.77
corr(u_i, X)      = 0 (assumed)           Prob > chi2       =       0.0000
```

trmgpa	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
spring	-.0606536	.0371605	-1.63	0.103	-.1334868	.0121797
crsgpa	1.082365	.0930877	11.63	0.000	.8999166	1.264814
frstsem	.0029948	.0599542	0.05	0.960	-.1145132	.1205028
season	-.0440992	.0392381	-1.12	0.261	-.1210044	.032806
sat	.0017052	.0001771	9.63	0.000	.0013582	.0020523
verbmath	-.15752	.16351	-0.96	0.335	-.4779937	.1629538
hsperc	-.0084622	.0012426	-6.81	0.000	-.0108977	-.0060268
hssize	-.0000775	.0001248	-0.62	0.534	-.000322	.000167
black	-.2348189	.0681573	-3.45	0.001	-.3684048	-.1012331
female	.358153	.0612948	5.84	0.000	.2380173	.4782886
_cons	-1.73492	.3566599	-4.86	0.000	-2.43396	-1.035879

```

sigma_u | .37185442
sigma_e | .40882825
rho     | .4527451   (fraction of variance due to u_i)
-----

```

b. Below are the fixed effects estimates with nonrobust standard errors. The time-constant variables have been dropped. The coefficient on *season* is now larger in magnitude,  $-.057$ , and it is more statistically significant with  $t = -1.37$ .

```

. xtreg trmgpa spring crsgpa frstsem season, fe

Fixed-effects (within) regression              Number of obs   =       732
Group variable: id                          Number of groups  =       366

R-sq:  within = 0.2069                      Obs per group:  min =
        between = 0.0333                      avg   =       2.
        overall = 0.0613                      max   =

corr(u_i, Xb)  = -0.0893                    F(4,362)        =      23.61
                                                Prob > F        =      0.0000
-----

```

	trmgpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	spring	-.0657817	.0391404	-1.68	0.094	-.1427528	.0111895
	crsgpa	1.140688	.1186538	9.61	0.000	.9073505	1.374025
	frstsem	.0128523	.0688364	0.19	0.852	-.1225172	.1482218
	season	-.0566454	.0414748	-1.37	0.173	-.1382072	.0249165
	_cons	-.7708055	.3305004	-2.33	0.020	-1.420747	-.1208636

```

-----
sigma_u | .67913296
sigma_e | .40882825
rho     | .73400603   (fraction of variance due to u_i)
-----
F test that all u_i=0:      F(365, 362) =      5.40          Prob > F = 0.0000

```

c. The following Stata output gives the nonrobust and fully robust regression-based Hausman test. Whether we test the three time averages, *crsgpabar*, *frstsembar*, and *seasonbar*, or just *seasonbar*, the  $p$ -value is large (.068 for the joint nonrobust test, .337 for the single nonrobust test). And the findings do not depend on using a robust test: the  $p$ -values are a little smaller but not close to being significant.

For comparison, the traditional way of computing the Hausman statistic – directly forming the quadratic form in the FE and RE estimates is included at the end, computed two different ways. The first uses the difference in the estimated variance matrices, and the value of the

statistic, 1.81, is very close to the nonrobust, regression-based statistic, 1.83. But the degrees of freedom reported by Stata are incorrect: it should be three, not four. Thus, the  $p$ -value reported by Stata using the hausman command is too large. This will be the case whenever aggregate time variables – most commonly, time period dummies – are included among the coefficients to test.

If we impose that the RE estimate of the variance  $\sigma_u^2$  is used to estimate both the FE and RE asymptotic variances, Stata then recognizes that the variance matrix has rank three rather than rank four. (The same is true if we use the FE estimate of  $\sigma_u^2$  in both places.) The  $p$ -value in this case agrees very closely with that for the nonrobust, regression-based test (and both statistics are 1.83 rounded to two decimal places.)

```
. egen crsgpabar = mean(crsgpa), by(id)
. egen frstsembar = mean(frstsem), by(id)
. egen seasonbar = mean(season), by(id)
. xtreg trmgpa spring crsgpa frstsem season sat verbmh hspcr hssize black
    female crsgpabar frstsembar seasonbar, re

Random-effects GLS regression                Number of obs      =       732
Group variable: id                          Number of groups   =       366

R-sq:   within  = 0.2069                    Obs per group: min =
        between = 0.5408                                avg  =       2.
        overall  = 0.4802                                max  =

Random effects u_i ~Gaussian                Wald chi2(13)       =     513.77
corr(u_i, X)      = 0 (assumed)             Prob > chi2         =     0.0000
```

trmgpa	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
spring	-.0657817	.0391404	-1.68	0.093	-.1424954	.0109321
crsgpa	1.140688	.1186538	9.61	0.000	.9081308	1.373245
frstsem	.0128523	.0688364	0.19	0.852	-.1220646	.1477692
season	-.0566454	.0414748	-1.37	0.172	-.1379345	.0246438
sat	.0016681	.0001804	9.24	0.000	.0013145	.0020218
verbmh	-.1316461	.1654748	-0.80	0.426	-.4559708	.1926785
hspcr	-.0084655	.0012554	-6.74	0.000	-.0109259	-.006005
hssize	-.0000783	.000125	-0.63	0.531	-.0003232	.0001666
black	-.2447934	.0686106	-3.57	0.000	-.3792676	-.1103192
female	.3357016	.0711808	4.72	0.000	.1961898	.4752134



crsgpabar	-.1861551	.2011254	-0.93	0.355	-.5803537	.2080434
frstsembar	-.078244	.1461014	-0.54	0.592	-.3645975	.2081095
seasonbar	.1243006	.1293555	0.96	0.337	-.1292315	.3778326
_cons	-1.423761	.5183296	-2.75	0.006	-2.439668	-.4078539
-----						
sigma_u	.37185442					
sigma_e	.40882825					
rho	.4527451	(fraction of variance due to u_i)				
-----						

. test crsgpabar frstsembar seasonbar

```
( 1) crsgpabar = 0
( 2) frstsembar = 0
( 3) seasonbar = 0
```

```
      chi2( 3) =      1.83
Prob > chi2 =      0.6084
```

. test seasonbar

```
( 1) seasonbar = 0
```

```
      chi2( 1) =      0.92
Prob > chi2 =      0.3366
```

. xtreg trmgpa spring crsgpa frstsem season sat verbmth hspcr hssize black  
female crsgpabar frstsembar seasonbar , re cluster(id)

Random-effects GLS regression	Number of obs	=	732
Group variable: id	Number of groups	=	366

R-sq: within = 0.2069	Obs per group: min =	
between = 0.5408	avg =	2.
overall = 0.4802	max =	

Random effects u_i ~Gaussian	Wald chi2(13)	=	629.75
corr(u_i, X) = 0 (assumed)	Prob > chi2	=	0.0000

(Std. Err. adjusted for 366 clusters in id)

trmgpa	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval	
spring	-.0657817	.0394865	-1.67	0.096	-.1431737	.0116104
crsgpa	1.140688	.1317893	8.66	0.000	.8823856	1.39899
frstsem	.0128523	.0684334	0.19	0.851	-.1212746	.1469793
season	-.0566454	.0411639	-1.38	0.169	-.1373251	.0240344
sat	.0016681	.0001848	9.03	0.000	.0013059	.0020304
verbmath	-.1316461	.166478	-0.79	0.429	-.4579371	.1946448
hsperc	-.0084655	.0013131	-6.45	0.000	-.0110391	-.0058918
hssize	-.0000783	.0001172	-0.67	0.504	-.000308	.0001514
black	-.2447934	.075569	-3.24	0.001	-.392906	-.0966808
female	.3357016	.067753	4.95	0.000	.2029081	.4684951
crsgpabar	-.1861551	.1956503	-0.95	0.341	-.5696227	.1973125
frstsembar	-.078244	.1465886	-0.53	0.594	-.3655525	.2090644
seasonbar	.1243006	.1342238	0.93	0.354	-.1387732	.3873743
cons	-1.423761	.4571037	-3.11	0.002	-2.319668	-.5278545

```

-----+-----
sigma_u | .37185442
sigma_e | .40882825
rho     | .4527451   (fraction of variance due to u_i)
-----+-----

```

```
. test crsgpabar frstsembar seasonbar
```

```

( 1) crsgpabar = 0
( 2) frstsembar = 0
( 3) seasonbar = 0

```

```

      chi2( 3) =      1.95
Prob > chi2 =      0.5829

```

```
. test seasonbar
```

```
( 1) seasonbar = 0
```

```

      chi2( 1) =      0.86
Prob > chi2 =      0.3544

```

```

. * The traditional Hausman test that incorrectly includes the coefficients
. * on "spring" (the time dummy) among those being tested.

```

```
. qui xtreg trmgpa spring crsgpa frstsem season, fe
```

```
. estimates store fe
```

```
. qui xtreg trmgpa spring crsgpa frstsem season sat verbmh hspc hssize
      black female, re
```

```
. estimates store re
```

```
. hausman fe re
```

```

-----+-----
              Coefficients
              (b)      (B)      (b-B)      sqrt(diag(V_b-V_B))
              fe      re      Difference      S.E.
-----+-----
spring | -.0657817 | -.0606536 | -.0051281 | .012291
crsgpa |  1.140688 |  1.082365 |  .0583227 | .0735758
frstsem |  .0128523 |  .0029948 |  .0098575 | .0338223
season | -.0566454 | -.0440992 | -.0125462 | .0134363
-----+-----

```

```

      b = consistent under Ho and Ha; obtained from xtreg
      B = inconsistent under Ha, efficient under Ho; obtained from xtreg

```

```
Test:  Ho:  difference in coefficients not systematic
```

```

      chi2(4) = (b-B)'[(V_b-V_B)^(-1)](b-B)
              =      1.81
Prob>chi2 =      0.7702

```

```
. qui xtreg trmgpa spring crsgpa frstsem season, fe
```

```
. estimates store fe
```

```
. qui xtreg trmgpa spring crsgpa frstsem season sat verbmh hspc hssize black
```

```
. estimates store re

. hausman fe re, sigmamore
```

Note: the rank of the differenced variance matrix (3) does not equal the number of coefficients being tested (4); be sure this is what you expect, or there may be problems computing the test. Examine the output of your estimators for anything unexpected and possibly consider scaling your variables so that the coefficients are on a similar scale.

	---- Coefficients ----			
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	fe	re	Difference	S.E.
spring	-.0657817	-.0606536	-.0051281	.0121895
crsgpa	1.140688	1.082365	.0583227	.0734205
frstsem	.0128523	.0029948	.0098575	.0337085
season	-.0566454	-.0440992	-.0125462	.013332

b = consistent under Ho and Ha; obtained from xtreg  
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

chi2(3) = (b-B)'[(V\_b-V\_B)^(-1)](b-B)  
= 1.83  
Prob>chi2 = 0.6077  
(V\_b-V\_B is not positive definite)

**10.8. a.** The Stata output is below. The coefficients on the lagged “clear-up” percentages are very close in magnitude. For example, if the first lag is 10 percentage points higher, the crime rate is estimated to fall by about 18.5 percent, a very large effect. The estimate of  $\rho$  in the AR(1) serial correlation test is .574 and  $t = 5.82$ , so there is very strong evidence of serial correlation.

```
. use norway

. xtset district year, delta(6)
    panel variable:  district (strongly balanced)
    time variable:   year, 72 to 78
               delta: 6 units

. reg lcrime d78 clrprc_1 clrprc_2
```

Source	SS	df	MS	Number of obs =	106
Model	18.7948264	3	6.26494214	F( 3, 102) =	30.27
Residual	21.1114968	102	.206975459	Prob > F =	0.0000
				R-squared =	0.4710
				Adj R-squared =	0.4554

Total		39.9063233	105	.380060222	Root MSE	=	.45495
lcrime		Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
d78		-.0547246	.0944947	-0.58	0.564	-.2421544	.1327051
clrprc_1		-.0184955	.0053035	-3.49	0.001	-.0290149	-.007976
clrprc_2		-.0173881	.0054376	-3.20	0.002	-.0281735	-.0066026
_cons		4.18122	.1878879	22.25	0.000	3.808545	4.553894

```
. predict vhat, resid
. gen vhat_1 = 1.vhat
(53 missing values generated)
. reg vhat vhat_1
```

Source	SS	df	MS	Number of obs = 53		
				F( 1, 51) = 33.85		
Model	3.8092697	1	3.8092697	Prob > F = 0.0000		
Residual	5.73894345	51	.112528303	R-squared = 0.3990		
				Adj R-squared = 0.3872		
Total	9.54821315	52	.183619484	Root MSE = .33545		
vhat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
vhat_1	.5739582	.0986485	5.82	0.000	.3759132	.7720033
_cons	-3.01e-09	.0460779	-0.00	1.000	-.0925053	.0925053

b. The fixed effects estimates are given below. The coefficient on *clrprc\_1* falls dramatically in magnitude, and becomes statistically insignificant. The coefficient on *clrprc\_2* falls somewhat but is still practically large and statistically significant.

To obtain the heteroskedasticity-robust standard error for FE, we must use the FD estimation (which is the same as FE because  $T = 2$ ) in order to make the calculation simple. Stock and Watson (2008, *Econometrica*) show that just applying the usual heteroskedasticity-robust standard error using pooled regression on the time-demeaned data does not produce valid standard errors. The reason is simple: as shown in the text, the time demeaning induces serial correlation in the errors. Of course, one can always use the fully robust standard errors, which allow for any kind of serial correlation in the original errors and

any kind of heteroskedasticity. In this example, obtaining the heteroskedasticity-robust standard errors has little effect on inference.

```
. xtreg lcrime d78 clrprc_1 clrprc_2, fe
```

```
Fixed-effects (within) regression               Number of obs   =       106
Group variable: district                       Number of groups =        53

R-sq:  within = 0.4209                        Obs per group:  min =
        between = 0.4798                                avg =       2.
        overall = 0.4234                                max =

                                                F(3,50)         =      12.12
corr(u_i, Xb)  = 0.3645                        Prob > F         =      0.0000
```

lcrime	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
d78	.0856556	.0637825	1.34	0.185	-.0424553	.2137665
clrprc_1	-.0040475	.0047199	-0.86	0.395	-.0135276	.0054326
clrprc_2	-.0131966	.0051946	-2.54	0.014	-.0236302	-.0027629
_cons	3.350995	.2324736	14.41	0.000	2.884058	3.817932
sigma_u	.47140473					
sigma_e	.2436645					
rho	.78915666	(fraction of variance due to u_i)				

```
F test that all u_i=0:      F(52, 50) =      5.88      Prob > F = 0.0000
```

```
. reg clcrime cclrprc_1 cclrprc_2, robust
```

```
Linear regression                               Number of obs =       53
                                                F( 2,    50) =       4.78
                                                Prob > F      =     0.0126
                                                R-squared     =     0.1933
                                                Root MSE     =     .34459
```

clcrime	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
cclrprc_1	-.0040475	.0042659	-0.95	0.347	-.0126158	.0045207
cclrprc_2	-.0131966	.0047286	-2.79	0.007	-.0226942	-.003699
_cons	.0856556	.0554876	1.54	0.129	-.0257945	.1971057

c. I use the FD regression to easily allow for heteroskedasticity. The two-sided  $p$ -value is .183. Because we do not reject  $H_0 : \beta_1 = \beta_2$  at even the 15% level, we might justify estimating a model with  $\beta_1 = \beta_2$  (and the pooled OLS results suggest it, too). The variable *avgclr* is the average of *clrprc\_1* and *clrprc\_2*, and so we can use it as the only explanatory

variable. Imposing the restriction gives a large estimated effect – a 10 percentage point increase in the average clear-up rate decreases crime by about 16.7% – and the heteroskedasticity-robust  $t$  statistic is  $-2.89$ .

```
. qui reg clcrime cclrprc_1 cclrprc_2, robust
. lincom cclrprc_1 - cclrprc_2
( 1) cclrprc_1 - cclrprc_2 = 0
```

clcrime	Coef.	Std. Err.	t	P> t	[95% Conf. Interval
(1)	.009149	.0067729	1.35	0.183	-.0044548 .0227529

```
. reg clcrime cavgclr, robust
```

Linear regression

Number of obs = 53  
F( 1, 51) = 8.38  
Prob > F = 0.0056  
R-squared = 0.1747  
Root MSE = .34511

clcrime	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval
cavgclr	-.0166511	.0057529	-2.89	0.006	-.0282006 -.0051016
_cons	.0993289	.0554764	1.79	0.079	-.0120446 .2107024

```
. reg clcrime cavgclr
```

Source	SS	df	MS	Number of obs =
Model	1.28607105	1	1.28607105	53
Residual	6.07411496	51	.119100293	F( 1, 51) = 10.80
Total	7.36018601	52	.141542039	Prob > F = 0.0018
				R-squared = 0.1747
				Adj R-squared = 0.1586
				Root MSE = .34511

clcrime	Coef.	Std. Err.	t	P> t	[95% Conf. Interval
cavgclr	-.0166511	.0050672	-3.29	0.002	-.0268239 -.0064783
_cons	.0993289	.0625916	1.59	0.119	-.0263289 .2249867

**10.9. a.** The RE and FE estimates, with fully robust standard errors for each, are given below. The variable-addition Hausman test is obtained by adding time averages of all variables

except the year dummies; all other explanatory variables change across  $i$  and  $t$ . For comparison, the traditional nonrobust Hausman statistic is computed. This version uses the RE estimate of  $\sigma_u^2$  in estimating both the RE and FE asymptotic variances and properly computes the degrees of freedom (which is five for this application).

While there are differences in the RE and FE estimates, the signs are the same and the magnitudes are similar. The fully robust Hausman test gives a strong statistical rejection of the RE assumption that county heterogeneity is uncorrelated with the criminal justice variables. Therefore, for magnitudes, we should prefer the FE estimates. (Remember, though, that both RE and FE maintain strict exogeneity conditional on the heterogeneity.) The nonrobust Hausman test gives a substantially larger statistic, 78.79 compared with 60.53, but the conclusion is the same.

```
. xtreg lcrmte lprbarr lprbconv lprbpris lavgsen lpolpc d82-d87, re
      cluster(county)

Random-effects GLS regression              Number of obs   =       630
Group variable: county                    Number of groups  =        90

R-sq:  within  = 0.4287                    Obs per group: min =
      between = 0.4533                               avg  =       7.
      overall  = 0.4454                               max  =

Random effects u_i ~Gaussian              Wald chi2(11)      =    156.83
corr(u_i, X)      = 0 (assumed)           Prob > chi2        =     0.0000
```

(Std. Err. adjusted for 90 clusters in county)

	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval	
lcrmte						
lprbarr	-.4252097	.0629147	-6.76	0.000	-.5485202	-.3018993
lprbconv	-.3271464	.0499587	-6.55	0.000	-.4250636	-.2292292
lprbpris	-.1793507	.0457547	-3.92	0.000	-.2690283	-.0896731
lavgsen	-.0083696	.0322377	-0.26	0.795	-.0715543	.0548152
lpolpc	.4294148	.0878659	4.89	0.000	.2572007	.6016288
d82	.0137442	.0164857	0.83	0.404	-.0185671	.0460556
d83	-.075388	.0194832	-3.87	0.000	-.1135743	-.0372017
d84	-.1130975	.0217025	-5.21	0.000	-.1556335	-.0705614
d85	-.1057261	.0254587	-4.15	0.000	-.1556242	-.0558279
d86	-.0795307	.0239141	-3.33	0.001	-.1264014	-.0326599
d87	-.0424581	.0246408	-1.72	0.085	-.0907531	.005837
_cons	-1.672632	.5678872	-2.95	0.003	-2.785671	-.5595939

```

-----+-----
      sigma_u | .30032934
      sigma_e | .13871215
         rho  | .82418424   (fraction of variance due to u_i)
-----+-----

. xtreg lcrmte lprbarr lprbconv lprbpris lavgsen lpolpc d82-d87, fe
      cluster(county)

Fixed-effects (within) regression              Number of obs   =       630
Group variable: county                        Number of groups  =        90

R-sq:  within = 0.4342                        Obs per group: min =
       between = 0.4066                                avg   =       7.
       overall = 0.4042                                max   =

                                                F(11,89)          =      11.49
corr(u_i, Xb)  = 0.2068                        Prob > F           =      0.0000

                               (Std. Err. adjusted for 90 clusters in county)
-----+-----
      lcrmte |               Coef.   Robust      t    P>|t|    [95% Conf. Interval
-----+-----
      lprbarr |   -.3597944   .0594678   -6.05   0.000   -.4779557   -.2416332
      lprbconv |   -.2858733   .051522   -5.55   0.000   -.3882464   -.1835001
      lprbpris |   -.1827812   .0452811   -4.04   0.000   -.2727538   -.0928085
      lavgsen |   -.0044879   .0333499   -0.13   0.893   -.0707535   .0617777
      lpolpc   |   .4241142   .0849052    5.00   0.000   .2554095   .592819
      d82      |   .0125802   .0160066    0.79   0.434   -.0192246   .044385
      d83      |  -.0792813   .0195639   -4.05   0.000   -.1181544   -.0404081
      d84      |  -.1177281   .0217118   -5.42   0.000   -.160869   -.0745872
      d85      |  -.1119561   .0256583   -4.36   0.000   -.1629386   -.0609736
      d86      |  -.0818268   .0236276   -3.46   0.001   -.1287745   -.0348792
      d87      |  -.0404704   .0241765   -1.67   0.098   -.0885087   .0075678
      _cons    |  -1.604135   .5102062   -3.14   0.002   -2.617904   -.5903664
-----+-----
      sigma_u | .43487416
      sigma_e | .13871215
         rho  | .90765322   (fraction of variance due to u_i)
-----+-----

. egen lprbatb = mean(lprbarr), by(county)
. egen lprbctb = mean(lprbconv), by(county)
. egen lprbptb = mean(lprbpris), by(county)
. egen lavgtb = mean(lavgsen), by(county)
. egen lpoltb = mean(lpolpc), by(county)

. qui xtreg lcrmte lprbarr lprbconv lprbpris lavgsen lpolpc d82-d87 lprbatb
      lprbctb lprbptb lavgtb lpoltb, re cluster(county)

. test  lprbatb lprbctb lprbptb lavgtb lpoltb

( 1)  lprbatb = 0
( 2)  lprbctb = 0

```



```

( 3)  lprbptb = 0
( 4)  lavgtb = 0
( 5)  lpoltb = 0

      chi2( 5) =    60.53
Prob > chi2 =    0.0000

```

```

. qui xtreg lcrmte lprbarr lprbconv lprbpris lavgsen lpolpc d82-d87, fe
. estimates store fe

. qui xtreg lcrmte lprbarr lprbconv lprbpris lavgsen lpolpc d82-d87, re
. estimates store re

. hausman fe re, sigmamore

```

Note: the rank of the differenced variance matrix (3) does not equal the number of coefficients being tested (4); be sure this is what you expect, or there may be problems computing the test. Examine the output of your estimators for anything unexpected and possibly consider scaling your variables so that the coefficients are on a similar scale.

	---- Coefficients ----		(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
	(b) fe	(B) re		
lprbarr	-.3597944	-.4252097	.0654153	.0133827
lprbconv	-.2858733	-.3271464	.0412731	.0084853
lprbpris	-.1827812	-.1793507	-.0034305	.0065028
lavgsen	-.0044879	-.0083696	.0038816	.0037031
lpolpc	.4241142	.4294148	-.0053005	.0103217
d82	.0125802	.0137442	-.001164	.0010763
d83	-.0792813	-.075388	-.0038933	.0008668
d84	-.1177281	-.1130975	-.0046306	.0013163
d85	-.1119561	-.1057261	-.00623	.0014304
d86	-.0818268	-.0795307	-.0022962	.0007719
d87	-.0404704	-.0424581	.0019876	.001219

b = consistent under Ho and Ha; obtained from xtreg  
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

```

      chi2(5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
              =      78.79
Prob>chi2 =      0.0000
(V_b-V_B is not positive definite)

```

b. Below is the Stata output for fixed effects with the nine wage variables; the inference is fully robust. The log wage variables are jointly significant at the 1.2% significance level.

Unfortunately, one of the most significant wage variables, *lwtuc*, has a positive, statistically

significant coefficient. It is difficult to give that a causal interpretation. The coefficient on the manufacturing wage implies that a ceteris paribus 10% increase in manufacturing wage reduces the crime rate by about three percent.

```
. xtreg lcrmte lprbarr lprbconv lprbpris lavgsen lpolpc d82-d87 lwcon lwtuc
      lwtrd lwfir lwser lwmfg lwfed lwsta lwloc, fe cluster(county)

Fixed-effects (within) regression              Number of obs   =       630
Group variable: county                       Number of groups  =        90

R-sq:   within  = 0.4575                     Obs per group:   min =
        between = 0.2518                               avg   =       7.
        overall  = 0.2687                               max   =

                                                F(20,89)         =      13.39
corr(u_i, Xb)  = 0.0804                      Prob > F          =      0.0000
```

(Std. Err. adjusted for 90 clusters in county)

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
lcrmte						
lprbarr	-.3563515	.0615202	-5.79	0.000	-.4785909	-.2341121
lprbconv	-.2859539	.0507647	-5.63	0.000	-.3868224	-.1850855
lprbpris	-.1751355	.0457628	-3.83	0.000	-.2660652	-.0842058
lavgsen	-.0028739	.0333994	-0.09	0.932	-.0692379	.06349
lpolpc	.4229	.0822172	5.14	0.000	.2595362	.5862639
d82	.0188915	.0221161	0.85	0.395	-.0250527	.0628357
d83	-.055286	.0306016	-1.81	0.074	-.1160906	.0055187
d84	-.0615162	.0406029	-1.52	0.133	-.1421934	.0191609
d85	-.0397115	.0603405	-0.66	0.512	-.1596068	.0801837
d86	-.0001133	.0720231	-0.00	0.999	-.1432217	.1429952
d87	.0537042	.0847749	0.63	0.528	-.1147418	.2221501
lwcon	-.0345448	.0253345	-1.36	0.176	-.0848839	.0157943
lwtuc	.0459747	.0161107	2.85	0.005	.013963	.0779864
lwtrd	-.0201766	.0313131	-0.64	0.521	-.0823951	.0420418
lwfir	-.0035445	.0130093	-0.27	0.786	-.0293937	.0223047
lwser	.0101264	.0202248	0.50	0.618	-.0300599	.0503128
lwmfg	-.3005691	.1063746	-2.83	0.006	-.5119331	-.089205
lwfed	-.3331226	.2245785	-1.48	0.142	-.7793554	.1131101
lwsta	.0215209	.1051755	0.20	0.838	-.1874606	.2305023
lwloc	.1810215	.1629903	1.11	0.270	-.1428368	.5048797
_cons	.8931726	1.479937	0.60	0.548	-2.04743	3.833775
sigma_u	.47756823					
sigma_e	.13700505					
rho	.92395784	(fraction of variance due to u_i)				

```
. testparm lwcon-lwloc
```

- ( 1) lwcon = 0
- ( 2) lwtuc = 0
- ( 3) lwtrd = 0

```
( 4)  lwfir = 0
( 5)  lwser = 0
( 6)  lwmfg = 0
( 7)  lwfed = 0
( 8)  lwsta = 0
( 9)  lwloc = 0
```

```
F( 9, 89) = 2.54
Prob > F = 0.0121
```

c. First, we need to compute the changes in log wages. Then, we just use pooled OLS.

Rather than difference the year dummies we just include dummies for 1983 through 1987.

Both the usual and full robust standard errors are computed and compared with those from FE.

The nonrobust FD and FE standard errors are similar, and often very different from the comparable robust standard errors. In fact, in many cases the robust standard errors are double or more than the nonrobust ones, although some nonrobust ones are actually smaller. The wage variables generally have much smaller coefficients when FD is used, but they are still jointly significant using a robust test.

```
. gen clwcon = lwcon - lwcon[_n-1] if year > 81
(90 missing values generated)

. gen clwtuc = lwtuc - lwtuc[_n-1] if year > 81
(90 missing values generated)

. gen clwtrd = lwtrd - lwtrd[_n-1] if year > 81
(90 missing values generated)

. gen clwfir = lwfir - lwfir[_n-1] if year > 81
(90 missing values generated)

. gen clwser = lwser - lwser[_n-1] if year > 81
(90 missing values generated)

. gen clwmfg = lwmfg - lwmfg[_n-1] if year > 81
(90 missing values generated)

. gen clwfed = lwfed - lwfed[_n-1] if year > 81
(90 missing values generated)

. gen clwsta = lwsta - lwsta[_n-1] if year > 81
(90 missing values generated)

. gen clwloc = lwloc - lwloc[_n-1] if year > 81
(90 missing values generated)
```

```
. reg clcrmrte clprbarr clprbcon clprbpri clavgscn clpolpc clwcon-clwloc
d83-d87
```

Source	SS	df	MS	Number of obs =	540
Model	9.86742162	19	.51933798	F( 19, 520) =	21.90
Residual	12.3293822	520	.02371035	Prob > F =	0.0000
				R-squared =	0.4445
				Adj R-squared =	0.4242
Total	22.1968038	539	.041181454	Root MSE =	.15398

clcrmrte	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
clprbarr	-.3230993	.0300195	-10.76	0.000	-.3820737	-.2641248
clprbcon	-.2402885	.0182474	-13.17	0.000	-.2761362	-.2044407
clprbpri	-.1693859	.02617	-6.47	0.000	-.2207978	-.117974
clavgscn	-.0156167	.0224126	-0.70	0.486	-.0596469	.0284136
clpolpc	.3977221	.026987	14.74	0.000	.3447051	.450739
clwcon	-.0442368	.0304142	-1.45	0.146	-.1039865	.015513
clwtuc	.0253997	.0142093	1.79	0.074	-.002515	.0533144
clwtrd	-.0290309	.0307907	-0.94	0.346	-.0895203	.0314586
clwfir	.009122	.0212318	0.43	0.668	-.0325886	.0508326
clwser	.0219549	.0144342	1.52	0.129	-.0064016	.0503113
clwmfg	-.1402482	.1019317	-1.38	0.169	-.3404967	.0600003
clwfed	.0174221	.1716065	0.10	0.919	-.319705	.3545493
clwsta	-.0517891	.0957109	-0.54	0.589	-.2398166	.1362385
clwloc	-.0305153	.1021028	-0.30	0.765	-.2311	.1700694
d83	-.1108653	.0268105	-4.14	0.000	-.1635355	-.0581951
d84	-.0374103	.024533	-1.52	0.128	-.0856063	.0107856
d85	-.0005856	.024078	-0.02	0.981	-.0478877	.0467164
d86	.0314757	.0245099	1.28	0.200	-.0166749	.0796262
d87	.0388632	.0247819	1.57	0.117	-.0098218	.0875482
_cons	.0198522	.0206974	0.96	0.338	-.0208086	.060513

```
. reg clcrmrte clprbarr clprbcon clprbpri clavgscn clpolpc clwcon-clwloc
d83-d87, cluster(county)
```

Linear regression

```
Number of obs = 540
F( 19, 89) = 13.66
Prob > F = 0.0000
R-squared = 0.4445
Root MSE = .15398
```

(Std. Err. adjusted for 90 clusters in county)

clcrmrte	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
clprbarr	-.3230993	.0584771	-5.53	0.000	-.4392919	-.2069066
clprbcon	-.2402885	.0403223	-5.96	0.000	-.320408	-.1601689
clprbpri	-.1693859	.0459288	-3.69	0.000	-.2606455	-.0781263
clavgscn	-.0156167	.0267541	-0.58	0.561	-.0687765	.0375432
clpolpc	.3977221	.1038642	3.83	0.000	.1913461	.604098
clwcon	-.0442368	.0165835	-2.67	0.009	-.0771879	-.0112856
clwtuc	.0253997	.0123845	2.05	0.043	.000792	.0500075
clwtrd	-.0290309	.0180398	-1.61	0.111	-.0648755	.0068138
clwfir	.009122	.006921	1.32	0.191	-.0046299	.0228739
clwser	.0219549	.0180754	1.21	0.228	-.0139606	.0578703

clwmfg	-.1402482	.1190279	-1.18	0.242	-.3767541	.0962578
clwfed	.0174221	.1326	0.13	0.896	-.2460511	.2808954
clwsta	-.0517891	.0674058	-0.77	0.444	-.185723	.0821449
clwloc	-.0305153	.1269012	-0.24	0.811	-.2826652	.2216346
d83	-.1108653	.0270368	-4.10	0.000	-.1645868	-.0571437
d84	-.0374103	.0237018	-1.58	0.118	-.0845052	.0096845
d85	-.0005856	.0256369	-0.02	0.982	-.0515257	.0503544
d86	.0314757	.0214193	1.47	0.145	-.011084	.0740353
d87	.0388632	.0263357	1.48	0.144	-.0134653	.0911917
_cons	.0198522	.0180545	1.10	0.274	-.0160217	.0557261

. test clwcon clwtuc clwtrd clwfir clwser clwmfg clwfed clwsta clwloc

```
( 1) clwcon = 0
( 2) clwtuc = 0
( 3) clwtrd = 0
( 4) clwfir = 0
( 5) clwser = 0
( 6) clwmfg = 0
( 7) clwfed = 0
( 8) clwsta = 0
( 9) clwloc = 0
```

F( 9, 89) = 2.38  
Prob > F = 0.0184

. xtreg lcrmte lprbarr lprbconv lprbpris lavgsen lpolpc d82-d87 lwcon lwtuc  
lwtrd lwfir lwser lwmfg lwfed lwsta lwloc, fe

Fixed-effects (within) regression	Number of obs	=	630
Group variable: county	Number of groups	=	90

R-sq: within	=	0.4575	Obs per group: min	=	
between	=	0.2518	avg	=	7.
overall	=	0.2687	max	=	

	F(20,520)	=	21.92
corr(u_i, Xb) = 0.0804	Prob > F	=	0.0000

lcrmte	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
lprbarr	-.3563515	.0321591	-11.08	0.000	-.4195292	-.2931738
lprbconv	-.2859539	.0210513	-13.58	0.000	-.3273099	-.2445979
lprbpris	-.1751355	.0323403	-5.42	0.000	-.2386693	-.1116017
lavgsen	-.0028739	.0262108	-0.11	0.913	-.054366	.0486181
lpolpc	.4229	.0263942	16.02	0.000	.3710476	.4747524
d82	.0188915	.0251244	0.75	0.452	-.0304662	.0682492
d83	-.055286	.0330287	-1.67	0.095	-.1201721	.0096001
d84	-.0615162	.0410805	-1.50	0.135	-.1422204	.0191879
d85	-.0397115	.0561635	-0.71	0.480	-.1500468	.0706237
d86	-.0001133	.0680124	-0.00	0.999	-.1337262	.1334996
d87	.0537042	.0798953	0.67	0.502	-.1032532	.2106615
lwcon	-.0345448	.0391616	-0.88	0.378	-.1114792	.0423896
lwtuc	.0459747	.019034	2.42	0.016	.0085817	.0833677
lwtrd	-.0201766	.0406073	-0.50	0.619	-.0999511	.0595979
lwfir	-.0035445	.028333	-0.13	0.900	-.0592058	.0521168
lwser	.0101264	.0191915	0.53	0.598	-.027576	.0478289

lwmfg	-.3005691	.1094068	-2.75	0.006	-.5155028	-.0856354
lwfed	-.3331226	.176448	-1.89	0.060	-.6797612	.013516
lwsta	.0215209	.1130648	0.19	0.849	-.2005991	.2436409
lwloc	.1810215	.1180643	1.53	0.126	-.0509203	.4129632
_cons	.8931726	1.424067	0.63	0.531	-1.90446	3.690805

---

sigma_u	.47756823	
sigma_e	.13700505	
rho	.92395784	(fraction of variance due to u_i)

---

F test that all u\_i=0:      F(89, 520) =      39.12      Prob > F = 0.0000

d. There is strong evidence of negative serial correlation in the FD equation, suggesting that if the idiosyncratic errors follow an AR(1) process, the coefficient is less than unity.

```
. qui reg clcrmrte clprbarr clprbcon clprbpri clavgscn clpolpc clwcon-clwloc
d83-d87

. predict ehat, resid
(90 missing values generated)

. gen ehat_1 = 1.ehat
(180 missing values generated)

. reg ehat ehat_1
```

Source	SS	df	MS	Number of obs =	450
Model	.490534556	1	.490534556	F( 1, 448) =	21.29
Residual	10.3219221	448	.023040005	Prob > F =	0.0000
				R-squared =	0.0454
				Adj R-squared =	0.0432
Total	10.8124566	449	.024081195	Root MSE =	.15179

---

ehat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval
ehat_1	-.222258	.0481686	-4.61	0.000	-.3169225    -.1275936
_cons	5.97e-10	.0071554	0.00	1.000	-.0140624    .0140624

---

**10.10. a.** To allow for different intercepts in the original model we can include a year dummy for 1993 in the FD equation. (The three years of data are 1987, 1990, and 1993.) There is no evidence of serial correlation in the FD errors,  $e_{it} = u_{it} - u_{i,t-1}$ , as the coefficient on  $\hat{e}_{i,t-1}$  is puny and so is its  $t$  statistic. It appears that a random walk for  $\{u_{it} : t = 1, 2, 3\}$  is a reasonably characterization, although concluding this with  $T = 3$  is tenuous.

```
. use murder
. xtset id year
```

```

panel variable:  id (strongly balanced)
time variable:  year, 87 to 93, but with gaps
delta: 1 unit

```

```
. reg cmrdрте d93 cexec cunem
```

Source	SS	df	MS	Number of obs =	102
Model	46.7620386	3	15.5873462	F( 3, 98) =	0.84
Residual	1812.28688	98	18.4927233	Prob > F =	0.4736
Total	1859.04892	101	18.406425	R-squared =	0.0252
				Adj R-squared =	-0.0047
				Root MSE =	4.3003

cmrdрте	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
d93	-1.296717	1.016118	-1.28	0.205	-3.313171	.7197368
cexec	-.1150682	.1473871	-0.78	0.437	-.407553	.1774167
cunem	.1630854	.3079049	0.53	0.598	-.447942	.7741127
_cons	1.51099	.6608967	2.29	0.024	.1994622	2.822518

```
. predict ehat, resid
(51 missing values generated)
```

```
. gen ehat_1 = 1.ehat
(102 missing values generated)
```

```
. reg ehat ehat_1
```

Source	SS	df	MS	Number of obs =	51
Model	.075953071	1	.075953071	F( 1, 49) =	0.06
Residual	58.3045094	49	1.18988795	Prob > F =	0.8016
Total	58.3804625	50	1.16760925	R-squared =	0.0013
				Adj R-squared =	-0.0191
				Root MSE =	1.0908

ehat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
ehat_1	.0065807	.0260465	0.25	0.802	-.0457618	.0589231
_cons	-9.10e-10	.1527453	-0.00	1.000	-.3069532	.3069532

b. To make all of the FE and FD estimates comparable, the year dummies are differenced along with the other variables in the FD estimation, and no constant is included. (The *R*-squared for the FD equation is computed using the usual total sum of squares, but the FE and FD *R*-squareds are not directly comparable.) The FE and FD coefficient estimates are similar but, especially for the execution variable, the FD standard error is much smaller. Because these are fully robust it is sensible to compare them. Because we found no serial correlation in the

FD errors, it makes sense that the FD estimator is more efficient than FE (whose idiosyncratic errors appear to follow a random walk).

```
. gen cd90 = d90 - d90[_n-1] if year > 87
(51 missing values generated)
```

```
. gen cd93 = d93 - d93[_n-1] if year > 87
(51 missing values generated)
```

```
. reg cmrdte cd90 cd93 cexec cunem, nocons tsscons cluster(id)
```

```
Linear regression                                Number of obs =      102
                                                F(   4,    50) =      2.95
                                                Prob > F       =    0.0291
                                                R-squared      =    0.0252
                                                Root MSE      =    4.3003
```

(Std. Err. adjusted for 51 clusters in id)

cmrdte	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
cd90	1.51099	1.056408	1.43	0.159	-.6108675	3.632848
cd93	1.725264	.8603626	2.01	0.050	-.0028256	3.453353
cexec	-.1150682	.0386021	-2.98	0.004	-.1926027	-.0375337
cunem	.1630854	.2998749	0.54	0.589	-.439231	.7654018

```
. xtreg mrdte d90 d93 exec unem, fe cluster(id)
```

```
Fixed-effects (within) regression                Number of obs      =      153
Group variable: id                             Number of groups   =       51

R-sq:  within = 0.0734                          Obs per group: min =
        between = 0.0037                          avg =               3.
        overall = 0.0108                          max =

                                                F(4,50)           =      1.80
corr(u_i, Xb) = 0.0010                          Prob > F           =    0.1443
```

(Std. Err. adjusted for 51 clusters in id)

mrdte	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
d90	1.556215	1.119004	1.39	0.170	-.6913706	3.8038
d93	1.733242	.8685105	2.00	0.051	-.0112126	3.477697
exec	-.1383231	.0805733	-1.72	0.092	-.3001593	.0235132
unem	.2213158	.374899	0.59	0.558	-.5316909	.9743225
_cons	5.822104	2.814864	2.07	0.044	.1682823	11.47593
sigma_u	8.7527226					
sigma_e	3.5214244					
rho	.86068589	(fraction of variance due to u_i)				



c. The explanatory variable  $exec_{it}$  might fail strict exogeneity if states increase future executions in response to current positive shocks to the murder rate. Given the relatively short stretch of time, feedback from murder rates to future executions may not be much of a concern, as the judicial process in capital cases tends to move slowly. (Of course, if it were sped up because of an increase in murder rates, that could violate strict exogeneity.) With a longer time series we could add  $exec_{i,t+1}$  (and even values from further in the future) and estimate the equation by FE or FD, testing  $exec_{i,t+1}$  for statistical significance.

**10.11.** a. The key coefficient is  $\beta_1$ . Because AFDC participation gives women access to better nutrition and prenatal care, we hope that AFDC participation causes the percent of low-weight births to fall. This only makes sense with a *ceteris paribus* thought experiment, holding fixed economic and other variables, such as demographic variables and quality of other kinds of health care. A reasonable expectation is  $\beta_2 < 0$ : more physicians means relatively fewer low-weight births. The variable *bedspc* is another proxy for health-care availability, and we expect  $\beta_3 < 0$ . Higher per capita income should lead to lower *lowbrth*, too ( $\beta_4 < 0$ ). The effect of population on a per capita variable is ambiguous, especially because it is total population and not population density.

b. The Stata output follows. Both the usual and fully robust standard errors are computed. The standard errors robust to serial correlation (and heteroskedasticity) are, as expected, somewhat larger. (If you test for AR(1) serial correlation in the composite error,  $v_{it}$  it is very strong. In fact, the estimated  $\rho$  is slightly above one). Only the per capita income variable is statistically significant. The estimate implies that a 10 percent rise in per capita income is associated with a .25 percentage point fall in the percent of low-weight births.

```
. reg lowbrth d90 afdcprc lphypc lbedspc lpcinc lpopul
```

Source	SS	df	MS	Number of obs =	100
Model	33.7710894	6	5.6285149	F( 6, 93) =	5.19
Residual	100.834005	93	1.08423661	Prob > F =	0.0001
				R-squared =	0.2509
				Adj R-squared =	0.2026
Total	134.605095	99	1.35964742	Root MSE =	1.0413

lowbrth	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
d90	.5797136	.2761244	2.10	0.038	.0313853	1.128042
afdcprc	.0955932	.0921802	1.04	0.302	-.0874584	.2786448
lphypc	.3080648	.71546	0.43	0.668	-1.112697	1.728827
lbedspc	.2790041	.5130275	0.54	0.588	-.7397668	1.297775
lpcinc	-2.494685	.9783021	-2.55	0.012	-4.4374	-.5519711
lpopul	.739284	.7023191	1.05	0.295	-.6553826	2.133951
_cons	26.57786	7.158022	3.71	0.000	12.36344	40.79227

```
. reg lowbrth d90 afdcprc lphypc lbedspc lpcinc lpopul, cluster(state)
```

Linear regression	Number of obs =	100
	F( 6, 49) =	4.73
	Prob > F =	0.0007
	R-squared =	0.2509
	Root MSE =	1.0413

(Std. Err. adjusted for 50 clusters in state

lowbrth	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
d90	.5797136	.2214303	2.62	0.012	.1347327	1.024694
afdcprc	.0955932	.1199883	0.80	0.429	-.1455324	.3367188
lphypc	.3080648	.9063342	0.34	0.735	-1.513282	2.129411
lbedspc	.2790041	.7853754	0.36	0.724	-1.299267	1.857275
lpcinc	-2.494685	1.203901	-2.07	0.044	-4.914014	-.0753567
lpopul	.739284	.9041915	0.82	0.418	-1.077757	2.556325
_cons	26.57786	9.29106	2.86	0.006	7.906773	45.24894

c. The FD (equivalently, FE) estimates are given below. The heteroskedasticity-robust standard error for the AFDC variable is actually smaller. In any case, removing the state unobserved effect changes the sign on the AFDC participation variable, and it is marginally statistically significant. Oddly, physicians-per-capita now has a positive, significant effect on percent of low-weight births. The hospital beds-per-capita variable has the expected negative effect.

```
. reg clowbrth cafdcprc clphypc clbedspc clpcinc clpopul
```

Source	SS	df	MS	Number of obs =	50
Model	.861531934	5	.172306387	F( 5, 44) =	2.53
Residual	3.00026764	44	.068187901	Prob > F =	0.0428
				R-squared =	0.2231
				Adj R-squared =	0.1348
Total	3.86179958	49	.078812236	Root MSE =	.26113

clowbrth	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
cafdcprc	-.1760763	.0903733	-1.95	0.058	-.3582116	.006059
clphypc	5.894509	2.816689	2.09	0.042	.2178452	11.57117
clbedspc	-1.576195	.8852111	-1.78	0.082	-3.360221	.2078308
clpcinc	-.8455268	1.356773	-0.62	0.536	-3.579924	1.88887
clpopul	3.441116	2.872175	1.20	0.237	-2.347372	9.229604
_cons	.1060158	.3090664	0.34	0.733	-.5168667	.7288983

```
. reg clowbrth cafdcprc clphypc clbedspc clpcinc clpopul, robust
```

Linear regression

Number of obs = 50  
F( 5, 44) = 1.97  
Prob > F = 0.1024  
R-squared = 0.2231  
Root MSE = .26113

clowbrth	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
cafdcprc	-.1760763	.0767568	-2.29	0.027	-.3307695	-.021383
clphypc	5.894509	3.098646	1.90	0.064	-.3504018	12.13942
clbedspc	-1.576195	1.236188	-1.28	0.209	-4.067567	.9151775
clpcinc	-.8455268	1.484034	-0.57	0.572	-3.8364	2.145346
clpopul	3.441116	2.687705	1.28	0.207	-1.975596	8.857829
_cons	.1060158	.3675668	0.29	0.774	-.6347664	.8467981

d. Adding a quadratic in *afdcprc* yields a diminishing impact of AFDC participation. The turning point in the quadratic is at about *afdcprc* = 6.4, and only four states have AFDC participation rates above 6.4 percent. So, the largest effect is at low AFDC participation rates and the effect is negative until *afdcprc* = 6.4. It is not clear this makes sense: if AFDC participation increases then more women in living in poverty get better prenatal care for their children. But the quadratic is not statistically significant at the usual levels and we could safely drop it.

```
. reg clowbrth cafdcprc cafdcpsq clphypc clbedspc clpcinc clpopul, robust
```

Linear regression

```
Number of obs =      50
F(   6,   43) =    2.07
Prob > F      =   0.0762
R-squared     =   0.2499
Root MSE     =   .25956
```

clowbrth	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
cafdcprc	-.5035049	.2612029	-1.93	0.061	-1.030271	.023261
cafdcpsq	.0396094	.0317531	1.25	0.219	-.0244267	.1036456
clphypc	6.620885	3.448026	1.92	0.061	-.332723	13.57449
clbedspc	-1.407963	1.344117	-1.05	0.301	-4.118634	1.302707
clpcinc	-.9987865	1.541609	-0.65	0.521	-4.107738	2.110165
clpopul	4.429026	2.925156	1.51	0.137	-1.470113	10.32817
_cons	.1245915	.386679	0.32	0.749	-.655221	.9044041

```
. di abs(_b[cafdcprc]/(2*_b[cafdcpsq]))
6.3558685
```

```
. sum afdcprc if d90
```

Variable	Obs	Mean	Std. Dev.	Min	Max
afdcprc	50	4.162976	1.317277	1.688183	7.358795

```
. count if afdcprc >= 6.4 & d90
4
```

**10.12. a.** Even if  $c_i$  is uncorrelated with  $\mathbf{x}_{it}$  for all  $t$ , the usual OLS standard errors do not account for the serial correlation in  $v_{it} = c_i + u_{it}$ . You can see that the fully robust standard errors are substantially larger than the usual ones, in some cases more than double.

```
. use wagepan
```

```
. reg lwage educ black hisp exper expersq married union d81-d87, cluster(nr)
```

Linear regression

```
Number of obs =    4360
F(  14,   544) =   47.10
Prob > F      =   0.0000
R-squared     =   0.1893
Root MSE     =   .48033
```

(Std. Err. adjusted for 545 clusters in nr)

lwage	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
-------	-------	---------------------	---	------	---------------------	--

educ	.0913498	.0110822	8.24	0.000	.0695807	.1131189
black	-.1392342	.0505238	-2.76	0.006	-.2384798	-.0399887
hisp	.0160195	.0390781	0.41	0.682	-.060743	.092782
exper	.0672345	.0195958	3.43	0.001	.0287417	.1057273
expersq	-.0024117	.0010252	-2.35	0.019	-.0044255	-.0003979
married	.1082529	.026034	4.16	0.000	.0571135	.1593924
union	.1824613	.0274435	6.65	0.000	.1285531	.2363695
d81	.05832	.028228	2.07	0.039	.0028707	.1137693
d82	.0627744	.0369735	1.70	0.090	-.0098538	.1354027
d83	.0620117	.046248	1.34	0.181	-.0288348	.1528583
d84	.0904672	.057988	1.56	0.119	-.0234407	.204375
d85	.1092463	.0668474	1.63	0.103	-.0220644	.240557
d86	.1419596	.0762348	1.86	0.063	-.007791	.2917102
d87	.1738334	.0852056	2.04	0.042	.0064611	.3412057
_cons	.0920558	.1609365	0.57	0.568	-.2240773	.4081888

. reg lwage educ black hisp exper expersq married union d81-d87

Source	SS	df	MS	Number of obs =	4360
Model	234.048277	14	16.7177341	F( 14, 4345) =	72.46
Residual	1002.48136	4345	.230720682	Prob > F =	0.0000
Total	1236.52964	4359	.283672779	R-squared =	0.1893
				Adj R-squared =	0.1867
				Root MSE =	.48033

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
educ	.0913498	.0052374	17.44	0.000	.0810819	.1016177
black	-.1392342	.0235796	-5.90	0.000	-.1854622	-.0930062
hisp	.0160195	.0207971	0.77	0.441	-.0247535	.0567925
exper	.0672345	.0136948	4.91	0.000	.0403856	.0940834
expersq	-.0024117	.00082	-2.94	0.003	-.0040192	-.0008042
married	.1082529	.0156894	6.90	0.000	.0774937	.1390122
union	.1824613	.0171568	10.63	0.000	.1488253	.2160973
d81	.05832	.0303536	1.92	0.055	-.0011886	.1178286
d82	.0627744	.0332141	1.89	0.059	-.0023421	.1278909
d83	.0620117	.0366601	1.69	0.091	-.0098608	.1338843
d84	.0904672	.0400907	2.26	0.024	.011869	.1690654
d85	.1092463	.0433525	2.52	0.012	.0242533	.1942393
d86	.1419596	.046423	3.06	0.002	.0509469	.2329723
d87	.1738334	.049433	3.52	0.000	.0769194	.2707474
_cons	.0920558	.0782701	1.18	0.240	-.0613935	.2455051

b. The random effects estimates on the time-constant variables are similar to the pooled OLS estimates. The coefficients on the quadratic in experience for RE show an initially stronger effect of experience, but with the slope diminishing more rapidly. There are important differences in the variables that change across individual and time; they are notably lower for random effects. The random effects marriage premium is about 6.4%, while the pooled OLS

estimate is about 10.8%. For union status, the random effects estimate is 10.6% compared with a pooled OLS estimate of 18.2%.

Note that the RE standard errors for the coefficients on the time-constant explanatory variables are similar to the fully robust POLS standard errors. However, the RE standard errors for *married* and *union* are substantially smaller than the robust POLS standard errors, suggestive of the relative efficiency of RE. To be fair, we should compute the fully robust standard errors for RE. As shown below, these are somewhat larger than the usual RE standard errors, but for the *married* and *union* still not nearly as large as the robust standard errors for POLS. An important conclusion is that, even though RE might not be the asymptotically efficient FGLS estimator, it appears to be more efficient than POLS, at least for the time-varying explanatory variables.

```
. xtreg lwage educ black hisp exper expersq married union d81-d87, re
```

```
Random-effects GLS regression              Number of obs      =       4360
Group variable: nr                        Number of groups   =        545

R-sq:  within  = 0.1799                   Obs per group: min =
       between = 0.1860                                     avg  =        8.
       overall  = 0.1830                                     max  =

Random effects u_i ~Gaussian              Wald chi2(14)       =       957.77
corr(u_i, X)      = 0 (assumed)           Prob > chi2         =        0.0000
```

	lwage	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
educ		.0918763	.0106597	8.62	0.000	.0709836	.1127689
black		-.1393767	.0477228	-2.92	0.003	-.2329117	-.0458417
hisp		.0217317	.0426063	0.51	0.610	-.0617751	.1052385
exper		.1057545	.0153668	6.88	0.000	.0756361	.1358729
expersq		-.0047239	.0006895	-6.85	0.000	-.0060753	-.0033726
married		.063986	.0167742	3.81	0.000	.0311091	.0968629
union		.1061344	.0178539	5.94	0.000	.0711415	.1411273
d81		.040462	.0246946	1.64	0.101	-.0079385	.0888626
d82		.0309212	.0323416	0.96	0.339	-.0324672	.0943096
d83		.0202806	.041582	0.49	0.626	-.0612186	.1017798
d84		.0431187	.0513163	0.84	0.401	-.0574595	.1436969
d85		.0578155	.0612323	0.94	0.345	-.0621977	.1778286
d86		.0919476	.0712293	1.29	0.197	-.0476592	.2315544
d87		.1349289	.0813135	1.66	0.097	-.0244427	.2943005
_cons		.0235864	.1506683	0.16	0.876	-.271718	.3188907

```

-----
sigma_u | .32460315
sigma_e | .35099001
rho     | .46100216   (fraction of variance due to u_i)
-----

. xtreg lwage educ black hisp exper expersq married union d81-d87, re
  cluster(nr)

Random-effects GLS regression              Number of obs   =       4360
Group variable: nr                        Number of groups  =       545

R-sq:  within = 0.1799                    Obs per group: min =
       between = 0.1860                                     avg =      8.
       overall = 0.1830                                     max =

Random effects u_i ~Gaussian                Wald chi2(14)      =      610.97
corr(u_i, X)      = 0 (assumed)             Prob > chi2       =      0.0000

                                         (Std. Err. adjusted for 545 clusters in nr)
-----
      lwage |               Coef.   Robust      z    P>|z|    [95% Conf. Interval
-----+-----
      educ |      .0918763   .0111455    8.24   0.000    .0700315    .1137211
     black |     -.1393767   .0509251   -2.74   0.006   -.2391882   -.0395653
      hisp |      .0217317   .0399157    0.54   0.586   -.0565015    .099965
      exper |      .1057545   .016379    6.46   0.000    .0736522    .1378568
    expersq |     -.0047239   .0007917   -5.97   0.000   -.0062756   -.0031723
    married |      .063986    .0189722    3.37   0.001    .0268013    .1011708
      union |      .1061344   .020844    5.09   0.000    .065281    .1469879
       d81 |      .040462    .0275684    1.47   0.142   -.0135711    .0944951
       d82 |      .0309212   .0350705    0.88   0.378   -.0378158    .0996581
       d83 |      .0202806   .043861    0.46   0.644   -.0656853    .1062466
       d84 |      .0431187   .0555848    0.78   0.438   -.0658254    .1520628
       d85 |      .0578155   .0645584    0.90   0.370   -.0687167    .1843476
       d86 |      .0919476   .0747028    1.23   0.218   -.0544671    .2383623
       d87 |      .1349289   .0848618    1.59   0.112   -.0313971    .3012549
      _cons |      .0235864   .1599577    0.15   0.883   -.289925    .3370977
-----
sigma_u | .32460315
sigma_e | .35099001
rho     | .46100216   (fraction of variance due to u_i)
-----

```

c. The variable  $exper_{it}$  is redundant because everyone in the sample works every year, so  $exper_{i,t+1} = exper_{it} + 1$ ,  $t = 1, \dots, 7$ , for all  $i$ . The effects of the initial levels of experience,  $exper_{i1}$ , cannot be distinguished from  $c_i$  because we are allowing  $exper_{i1}$  to be correlated with  $c_i$ . Then, because each experience variable follows the same linear time trend, the effects cannot be separated from the aggregate time effects (year dummies).

TheFE estimates follow. The marriage and union premiums fall even more, although both are still statistically significant and economically relevant. The fully robust standard errors are somewhat larger than the usual FE standard errors.

```
. xtreg lwage expersq married union d81-d87, fe
```

```
Fixed-effects (within) regression      Number of obs      =      4360
Group variable: nr                    Number of groups   =      545

R-sq:  within  = 0.1806                Obs per group: min =
      between  = 0.0286                                avg  =      8.
      overall  = 0.0888                                max  =

                                         F(10,3805)         =      83.85
corr(u_i, Xb)  = -0.1222                Prob > F           =      0.0000
```

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
expersq	-.0051855	.0007044	-7.36	0.000	-.0065666	-.0038044
married	.0466804	.0183104	2.55	0.011	.0107811	.0825796
union	.0800019	.0193103	4.14	0.000	.0421423	.1178614
d81	.1511912	.0219489	6.89	0.000	.1081584	.194224
d82	.2529709	.0244185	10.36	0.000	.2050963	.3008454
d83	.3544437	.0292419	12.12	0.000	.2971125	.4117749
d84	.4901148	.0362266	13.53	0.000	.4190894	.5611402
d85	.6174823	.0452435	13.65	0.000	.5287784	.7061861
d86	.7654966	.0561277	13.64	0.000	.6554532	.8755399
d87	.9250249	.0687731	13.45	0.000	.7901893	1.059861
_cons	1.426019	.0183415	77.75	0.000	1.390058	1.461979
sigma_u	.39176195					
sigma_e	.35099001					
rho	.55472817	(fraction of variance due to u_i)				

```
F test that all u_i=0:      F(544, 3805) =      9.16      Prob > F = 0.0000
```

```
. xtreg lwage expersq married union d81-d87, fe cluster(nr)
```

```
Fixed-effects (within) regression      Number of obs      =      4360
Group variable: nr                    Number of groups   =      545

R-sq:  within  = 0.1806                Obs per group: min =
      between  = 0.0286                                avg  =      8.
      overall  = 0.0888                                max  =

                                         F(10,544)         =      46.59
corr(u_i, Xb)  = -0.1222                Prob > F           =      0.0000
```

(Std. Err. adjusted for 545 clusters in nr)

lwage	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
-------	-------	------------------	---	------	---------------------	--



expersq	-.0051855	.0008102	-6.40	0.000	-.0067771	-.0035939
married	.0466804	.0210038	2.22	0.027	.0054218	.0879389
union	.0800019	.0227431	3.52	0.000	.0353268	.1246769
d81	.1511912	.0255648	5.91	0.000	.1009733	.2014091
d82	.2529709	.0286624	8.83	0.000	.1966684	.3092733
d83	.3544437	.0348608	10.17	0.000	.2859655	.422922
d84	.4901148	.0454581	10.78	0.000	.4008199	.5794097
d85	.6174823	.0568088	10.87	0.000	.5058908	.7290737
d86	.7654966	.071244	10.74	0.000	.6255495	.9054436
d87	.9250249	.0840563	11.00	0.000	.7599103	1.09014
_cons	1.426019	.0209824	67.96	0.000	1.384802	1.467235
<hr/>						
sigma_u	.39176195					
sigma_e	.35099001					
rho	.55472817	(fraction of variance due to u_i)				

```
. gen d81educ = d81*educ
. gen d82educ = d82*educ
. gen d83educ = d83*educ
. gen d84educ = d84*educ
. gen d85educ = d85*educ
. gen d86educ = d86*educ
. gen d87educ = d87*educ

. xtreg lwage expersq married union d81-d87 d81educ-d87educ, fe cluster(nr)

Fixed-effects (within) regression               Number of obs   =        4360
Group variable: nr                           Number of groups =         545

R-sq:    within  = 0.1814                      Obs per group: min =
          between = 0.0211                          avg  =           8.
          overall  = 0.0784                          max  =

                                         F(17,544)       =       28.33
corr(u_i, Xb)  = -0.1732                      Prob > F        =       0.0000
```

union	.0789759	.022762	3.47	0.001	.0342638	.123688
d81	.0984201	.1463954	0.67	0.502	-.1891495	.3859897
d82	.2472016	.1490668	1.66	0.098	-.0456155	.5400186
d83	.408813	.1716953	2.38	0.018	.071546	.74608
d84	.6399247	.1873708	3.42	0.001	.2718659	1.007984
d85	.7729397	.2090195	3.70	0.000	.3623554	1.183524
d86	.9699322	.2463734	3.94	0.000	.4859724	1.453892
d87	1.188777	.2580167	4.61	0.000	.6819456	1.695608
d81educ	.0049906	.0122858	0.41	0.685	-.0191429	.0291241
d82educ	.001651	.012194	0.14	0.892	-.0223021	.025604
d83educ	-.0026621	.0136788	-0.19	0.846	-.0295319	.0242076
d84educ	-.0098257	.0146869	-0.67	0.504	-.0386757	.0190243
d85educ	-.0092145	.0151166	-0.61	0.542	-.0389085	.0204796
d86educ	-.0121382	.0168594	-0.72	0.472	-.0452558	.0209794
d87educ	-.0157892	.0163557	-0.97	0.335	-.0479172	.0163389
_cons	1.436283	.0227125	63.24	0.000	1.391668	1.480897
<hr/>						
sigma_u	.39876325					
sigma_e	.3511451					
rho	.56324361	(fraction of variance due to u_i)				
<hr/>						

```
. testparm d81educ-d87educ
```

```
( 1) d81educ = 0
( 2) d82educ = 0
( 3) d83educ = 0
( 4) d84educ = 0
( 5) d85educ = 0
( 6) d86educ = 0
( 7) d87educ = 0
```

```
F( 7, 544) = 0.43
Prob > F = 0.8851
```

e. First, I created the lead variable, and then included it in the FE estimation with fully robust inference. As you can see, *unionp1* is statistically significant with  $p$ -value = .029, and its coefficient is not small. It seems  $union_{it}$  fails the strict exogeneity assumption, and we possibly should use an IV approach as described in Chapter 11. (However, coming up with instruments is not trivial.)

```
. gen unionp1 = union[_n+1] if year < 1987
(545 missing values generated)
```

```
. xtreg lwage expersq married union unionp1 d81-d86, fe cluster(nr)
```

```
Fixed-effects (within) regression           Number of obs   =       3815
Group variable: nr                         Number of groups  =        545

R-sq:   within  = 0.1474                   Obs per group:   min =
        between = 0.0305                   avg  =           7.
```

```

overall = 0.0744
corr(u_i, Xb) = -0.1262
F(10,544) = 31.29
Prob > F = 0.0000

```

(Std. Err. adjusted for 545 clusters in nr)

lwage	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
expersq	-.0054448	.0009786	-5.56	0.000	-.0073671	-.0035226
married	.0448778	.0235662	1.90	0.057	-.0014141	.0911697
union	.0763554	.0236392	3.23	0.001	.0299202	.1227906
unionpl	.0497356	.0227497	2.19	0.029	.0050477	.0944236
d81	.1528275	.0257236	5.94	0.000	.1022977	.2033573
d82	.2576486	.0304975	8.45	0.000	.1977413	.3175558
d83	.3618296	.0384587	9.41	0.000	.2862839	.4373754
d84	.5023642	.0517471	9.71	0.000	.4007155	.6040129
d85	.6342402	.065288	9.71	0.000	.5059928	.7624876
d86	.7841312	.0826431	9.49	0.000	.6217924	.9464699
_cons	1.417924	.0225168	62.97	0.000	1.373693	1.462154
sigma_u	.39716048					
sigma_e	.35740734					
rho	.5525375	(fraction of variance due to u_i)				

f. The Stata output shows that the union premium for Hispanic men is well below that of non-black men: about 13 percentage points lower. The difference is statistically significant, too. The estimated union “premium” is actually about –3.5% for Hispanics, although it is not statistically different from zero. The estimated wage premium for black men is about 7.1 percentage points higher than the base group, but the difference is not statistically significant.

```

. gen black_union = black*union
. gen hisp_union = hisp*union
. xtreg lwage expersq married union black_union hisp_union d81-d87, fe

Fixed-effects (within) regression      Number of obs      =      4360
Group variable: nr                    Number of groups    =       545

R-sq:  within  = 0.1830                Obs per group: min =
      between  = 0.0267                                avg   =      8.
      overall  = 0.0871                                max   =

corr(u_i, Xb) = -0.1360
F(12,3803) = 70.99
Prob > F = 0.0000

```

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
-------	-------	-----------	---	------	---------------------	--

```

      expersq |    -0.005308    .0007048    -7.53    0.000    -0.0066898    -0.0039262
      married |    .0461639    .0182922     2.52    0.012     .0103004     .0820275
        union |    .0957205    .0244326     3.92    0.000     .0478183     .1436227
black_union |    .0714378    .0532042     1.34    0.179    -0.0328737     .1757492
hisp_union |   -0.1302478    .0485409    -2.68    0.007    -0.2254166    -0.0350791
      d81 |    .1507003    .0219236     6.87    0.000     .1077172     .1936833
      d82 |    .2545937    .0243936    10.44    0.000     .206768     .3024194
      d83 |    .3576139     .029227    12.24    0.000     .3003119     .414916
      d84 |    .4947141    .0362132    13.66    0.000     .423715     .5657132
      d85 |    .6236823    .0452345    13.79    0.000     .5349961     .7123686
      d86 |    .7750896    .0561524    13.80    0.000     .664998     .8851813
      d87 |    .9344805    .0687783    13.59    0.000     .7996347     1.069326
      _cons |    1.42681    .0183207    77.88    0.000     1.390891     1.462729
-----+-----
      sigma_u |    .393505
      sigma_e |    .35056318
        rho |    .55752062    (fraction of variance due to u_i)
-----+-----
F test that all u_i=0:      F(544, 3803) =      9.17      Prob > F = 0.0000

. xtreg lwage expersq married union black_union hisp_union d81-d87, fe
  cluster(nr)

Fixed-effects (within) regression      Number of obs      =      4360
Group variable: nr                    Number of groups    =      545

R-sq:  within  = 0.1830                Obs per group: min =
      between = 0.0267                      avg  =      8.
      overall  = 0.0871                      max  =

corr(u_i, Xb) = -0.1360                F(12,544)          =      40.16
                                      Prob > F            =      0.0000

                                      (Std. Err. adjusted for 545 clusters in nr)
-----+-----
      lwage |      Coef.    Robust      t    P>|t|    [95% Conf. Interval
-----+-----
      expersq |   -0.005308    .0008095    -6.56    0.000    -0.0068982    -0.0037178
      married |   .0461639    .0209641     2.20    0.028     .0049834     .0873444
        union |   .0957205    .0304494     3.14    0.002     .0359077     .1555333
black_union |   .0714378    .0600866     1.19    0.235    -0.0465925     .189468
hisp_union |  -0.1302478    .0493283    -2.64    0.009    -0.2271451    -0.0333505
      d81 |   .1507003     .025519     5.91    0.000     .1005725     .2008281
      d82 |   .2545937    .0286062     8.90    0.000     .1984014     .3107859
      d83 |   .3576139    .0348463    10.26    0.000     .2891641     .4260638
      d84 |   .4947141    .0453929    10.90    0.000     .4055473     .5838809
      d85 |   .6236823     .056721    11.00    0.000     .5122633     .7351014
      d86 |   .7750896     .071142    10.89    0.000     .635343     .9148363
      d87 |   .9344805    .0838788    11.14    0.000     .7697145     1.099247
      _cons |    1.42681    .0209431    68.13    0.000     1.385671     1.467949
-----+-----
      sigma_u |    .393505
      sigma_e |    .35056318
        rho |    .55752062    (fraction of variance due to u_i)
-----+-----

. lincom union + hisp_union

```

( 1) union + hisp\_union = 0

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
(1)	-.0345273	.0388508	-0.89	0.375	-.1108432	.0417886

g. We reject the null hypothesis that  $\{union_{it} : t = 1, \dots, T\}$  is strictly exogenous even

when the union premium is allowed to differ by race and ethnicity.

```
. xtreg lwage expersq married union black_union hisp_union d81-d86 unionpl, fe
cluster(nr)
```

```
Fixed-effects (within) regression      Number of obs      =      3815
Group variable: nr                    Number of groups    =      545
```

```
R-sq:  within = 0.1497                  Obs per group: min =
       between = 0.0293                  avg =      7.
       overall = 0.0735                  max =
```

```
corr(u_i, Xb) = -0.1386                F(12,544)          =      27.21
                                           Prob > F            =      0.0000
```

(Std. Err. adjusted for 545 clusters in nr)

lwage	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
expersq	-.0055477	.0009833	-5.64	0.000	-.0074793	-.0036162
married	.044659	.0235905	1.89	0.059	-.0016807	.0909986
union	.0886305	.031931	2.78	0.006	.0259075	.1513536
black_union	.0849246	.0627531	1.35	0.177	-.0383434	.2081926
hisp_union	-.1179177	.0525974	-2.24	0.025	-.2212365	-.0145988
d81	.1522945	.0256633	5.93	0.000	.1018831	.2027058
d82	.2589966	.0304368	8.51	0.000	.1992085	.3187846
d83	.3643699	.0385023	9.46	0.000	.2887385	.4400013
d84	.506142	.0518005	9.77	0.000	.4043885	.6078955
d85	.6393639	.0654104	9.77	0.000	.5108761	.7678517
d86	.7921705	.0828197	9.57	0.000	.6294849	.954856
unionpl	.0502016	.0227235	2.21	0.028	.0055649	.0948382
_cons	1.418021	.022508	63.00	0.000	1.373808	1.462234
sigma_u	.39854263					
sigma_e	.35703614					
rho	.5547681	(fraction of variance due to u_i)				

**10.13.** a. Showing that this procedure is consistent with fixed  $T$  as  $N \rightarrow \infty$  requires some algebra. First, in the sum of squared residuals, we can “concentrate out” the  $a_i$  by finding  $\hat{a}_i(\mathbf{b})$  as a function of  $(\mathbf{x}_i, \mathbf{y}_i)$  and  $\mathbf{b}$ , substituting back into the sum of squared residuals, and then

minimizing with respect to  $\mathbf{b}$  only. Straightforward algebra gives the first order conditions for each  $i$  as

$$\sum_{t=1}^T (y_{it} - \hat{a}_i - \mathbf{x}_{it}\mathbf{b})/h_{it} = 0$$

which implies

$$\hat{a}_i(\mathbf{b}) = w_i \left( \sum_{t=1}^T y_{it}/h_{it} \right) - w_i \left( \sum_{t=1}^T (\mathbf{x}_{it}/h_{it}) \right) \mathbf{b} \equiv \bar{y}_i^w - \bar{\mathbf{x}}_i^w \mathbf{b},$$

where  $w_i \equiv \left( \sum_{t=1}^T (1/h_{it}) \right)^{-1} > 0$  and  $\bar{y}_i^w \equiv w_i \left( \sum_{t=1}^T y_{it}/h_{it} \right)$ , and a similar definition holds for  $\bar{\mathbf{x}}_i^w$ . Note that  $\bar{y}_i^w$  and  $\bar{\mathbf{x}}_i^w$  are weighted averages with weights  $w_i/h_{it}$ ,  $t = 1, 2, \dots, T$ . If  $h_{it}$  equals the same constant for all  $t$ ,  $y_i^{-w}$  and  $\mathbf{x}_i^{-w}$  are simply weighted averages. If  $h_{it}$  equals the same constant for all  $t$ ,  $\bar{y}_i^w$  and  $\bar{\mathbf{x}}_i^w$  are the usual time averages.

Now we can plug each  $\hat{a}_i(\mathbf{b})$  into the SSR to get the problem solved by  $\hat{\boldsymbol{\beta}}_{FEWLS}$ :

$$\min_{\mathbf{b} \in \mathbb{R}^K} \sum_{i=1}^N \sum_{t=1}^T [(y_{it} - \bar{y}_i^w) - (\mathbf{x}_{it} - \bar{\mathbf{x}}_i^w)\mathbf{b}]^2/h_{it}.$$

This is just a pooled weighted least squares regression of  $(y_{it} - \bar{y}_i^w)$  on  $(\mathbf{x}_{it} - \bar{\mathbf{x}}_i^w)$  with weights  $1/h_{it}$ . Equivalently, define  $\tilde{y}_{it} \equiv (y_{it} - \bar{y}_i^w)/\sqrt{h_{it}}$ ,  $\tilde{\mathbf{x}}_{it} \equiv (\mathbf{x}_{it} - \bar{\mathbf{x}}_i^w)/\sqrt{h_{it}}$ , all

$t = 1, \dots, T, i = 1, \dots, N$ . Then  $\hat{\boldsymbol{\beta}}$  can be expressed in usual pooled OLS form:

$$\hat{\boldsymbol{\beta}}_{FEWLS} = \left( \sum_{i=1}^N \sum_{t=1}^T \tilde{\mathbf{x}}_{it}' \tilde{\mathbf{x}}_{it} \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T \tilde{\mathbf{x}}_{it}' \tilde{y}_{it} \right) \quad (10.90)$$

Note carefully how the initial  $y_{it}$  are weighted by  $1/h_{it}$  to obtain  $y_i^{-w}$ , but where the usual  $1/\sqrt{h_{it}}$  weighting shows up in the sum of squared residuals on the time-demeaned data (where the demeaning is a weighted average). Given (10.90), we can easily study the asymptotic

$(N \rightarrow \infty)$  properties of  $\hat{\beta}$ . First, we can write  $\bar{y}_i^w = \mathbf{x}_i^{-w}\beta + c_i + \bar{u}_i^w$  where  $\bar{u}_i^w \equiv w_i \left( \sum_{t=1}^T u_{it}/h_{it} \right)$ .

Subtracting this equation from  $y_{it} = \mathbf{x}_{it}\beta + c_i + u_{it}$  for all  $t$  gives  $\tilde{y}_{it} = \tilde{\mathbf{x}}_{it}\beta + \tilde{u}_{it}$ , where

$\tilde{u}_{it} \equiv (u_{it} - \bar{u}_i^w)/\sqrt{h_{it}}$ . When we plug this in for  $\tilde{y}_{it}$  in (10.90) and divide by  $N$  in the appropriate places we get

$$\hat{\beta}_{FEWLS} = \beta + \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \tilde{\mathbf{x}}_{it}' \tilde{\mathbf{x}}_{it} \right) \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \tilde{\mathbf{x}}_{it}' u_{it} / \sqrt{h_{it}} \right). \quad (10.91)$$

From this last equation we can immediately read off the consistency of  $\hat{\beta}_{FEWLS}$  regardless of whether  $\text{Var}(u_{it}|\mathbf{x}_i, \mathbf{h}_i, c_i) = \sigma_u^2 h_{it}$ . Why? We assumed that  $E(u_{it}|\mathbf{x}_i, \mathbf{h}_i, c_i) = 0$ , which means  $u_{it}$  is uncorrelated with any function of  $(\mathbf{x}_i, \mathbf{h}_i)$ , including  $\tilde{\mathbf{x}}_{it}$ . Therefore,  $E(\tilde{\mathbf{x}}_{it}' u_{it}) = \mathbf{0}$ ,  $t = 1, \dots, T$  under  $E(u_{it}|\mathbf{x}_i, \mathbf{h}_i, c_i) = 0$  with any restrictions on the conditional second moments of  $\{u_{it} : t = 1, \dots, T\}$ . As long as we assume that  $\sum_{t=1}^T E(\tilde{\mathbf{x}}_{it}' \tilde{\mathbf{x}}_{it})$  has rank  $K$ , we can apply the consistency result for pooled OLS to conclude  $\text{plim}(\hat{\beta}_{FEWLS}) = \beta$ . (We can even show that  $E(\hat{\beta}_{FEWLS}|\mathbf{X}, \mathbf{H}) = \beta$ , that is, the FEWLS estimator is conditionally unbiased.)

b. It is clear from (10.91) that  $\hat{\beta}_{FEWLS}$  is  $\sqrt{N}$ -asymptotically normal under mild assumptions because we can write

$$\sqrt{N}(\hat{\beta}_{FEWLS} - \beta) = \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T \tilde{\mathbf{x}}_{it}' \tilde{\mathbf{x}}_{it} \right) \left( N^{-1/2} \sum_{i=1}^N \sum_{t=1}^T \tilde{\mathbf{x}}_{it}' u_{it} / \sqrt{h_{it}} \right)$$

The asymptotic variance is generally

$$\text{Avar} \sqrt{N}(\hat{\beta}_{FEWLS} - \beta) = \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1},$$

where

$$\mathbf{A} \equiv \sum_{t=1}^T \mathbf{E}(\tilde{\mathbf{x}}_{it}' \tilde{\mathbf{x}}_{it})$$

$$\mathbf{B} \equiv \mathbf{E} \left[ \left( \sum_{t=1}^T \tilde{\mathbf{x}}_{it}' u_{it} / \sqrt{h_{it}} \right) \left( \sum_{t=1}^T \tilde{\mathbf{x}}_{it}' u_{it} / \sqrt{h_{it}} \right)' \right].$$

If we assume that  $\text{Cov}(u_{it}, u_{is} | \mathbf{x}_i, \mathbf{h}_i, c_i) = \mathbf{E}(u_{it} u_{is} | \mathbf{x}_i, \mathbf{h}_i, c_i) = 0$ ,  $t \neq s$  then by a standard iterated expectations argument,

$$\mathbf{E} \left[ \left( \tilde{\mathbf{x}}_{it}' u_{it} / \sqrt{h_{it}} \right) \left( \tilde{\mathbf{x}}_{is}' u_{is} / \sqrt{h_{is}} \right)' \right] = \mathbf{E} \left[ \left( u_{it} u_{is} \tilde{\mathbf{x}}_{it}' \tilde{\mathbf{x}}_{is} / \sqrt{h_{it} h_{is}} \right) \right] = \mathbf{0}, \quad t \neq s.$$

Further, given the variance assumption  $\text{Var}(u_{it} | \mathbf{x}_i, \mathbf{h}_i, c_i) = \mathbf{E}(u_{it}^2 | \mathbf{x}_i, \mathbf{h}_i, c_i) = \sigma_u^2 h_{it}$ , iterated expectations implies

$$\mathbf{E} \left[ \left( \tilde{\mathbf{x}}_{it}' u_{it} / \sqrt{h_{it}} \right) \left( \tilde{\mathbf{x}}_{it}' u_{it} / \sqrt{h_{it}} \right)' \right] = \mathbf{E}(u_{it}^2 \tilde{\mathbf{x}}_{it}' \tilde{\mathbf{x}}_{it} / h_{it}) = \sigma_u^2 \mathbf{E}(\tilde{\mathbf{x}}_{it}' \tilde{\mathbf{x}}_{it} / h_{it}).$$

It follows then that

$$\mathbf{B} = \sigma_u^2 \mathbf{A}$$

and so

$$\text{Avar} \sqrt{N} (\hat{\boldsymbol{\beta}}_{FEWLS} - \boldsymbol{\beta}) = \sigma_u^2 \mathbf{A}^{-1}.$$

c. The same subtleties that arise in estimating  $\sigma_u^2$  for the usual fixed effects estimator crop up here as well. Assume the zero conditional covariance assumption and correct variance specification in part b. Then the residuals from the pooled OLS regression

$$\tilde{y}_{it} \text{ on } \tilde{\mathbf{x}}_{it}, \quad t = 1, \dots, T, \quad i = 1, \dots, N, \quad (10.92)$$

say  $\hat{u}_{it}$ , are estimating  $\ddot{u}_{it} = (u_{it} - \bar{u}_i^w) / \sqrt{h_{it}}$  in the sense that we obtain  $\hat{u}_{it}$  from  $\ddot{u}_{it}$  by replacing  $\boldsymbol{\beta}$  with  $\hat{\boldsymbol{\beta}}_{FEWLS}$ ). Now

$$\mathbf{E}(\ddot{u}_{it}^2) = \mathbf{E}[(\ddot{u}_{it}^2 / h_{it})] - 2\mathbf{E}[(u_{it} \bar{u}_i^w) / h_{it}] = \mathbf{E}[(\bar{u}_i^w)^2 / h_{it}] = \sigma_u^2 - 2\sigma_u^2 \mathbf{E}[(w_i / h_{it})] + \sigma_u^2 \mathbf{E}[(w_i / h_{it})],$$



where the law of iterated expectations is applied several times, and  $E[(\bar{u}_i^w)^2|\mathbf{x}_i, \mathbf{h}_i] = \sigma_u^2 w_i$  has been used. Therefore,  $E(\ddot{u}_{it}^2) = \sigma_u^2[1 - E(w_i/h_{it})]$ ,  $t = 1, \dots, T$ , and so

$$\sum_{t=1}^T E(\ddot{u}_{it}^2) = \sigma_u^2 \{T - E[w_i \cdot \sum_{t=1}^T (1/h_{it})]\} = \sigma_u^2(T-1).$$

This contains the usual result for the within transformation as a special case. A consistent estimator of  $\sigma_u^2$  is  $SSR/[N(T-1) - K]$ , where SSR is the usual sum of squared residuals from (10.92), and the subtraction of  $K$  is optional as a degrees-of-freedom adjustment. The estimator of  $\text{Avar}(\hat{\boldsymbol{\beta}}_{FEWLS})$  is then

$$\hat{\sigma}_u^2 \left( \sum_{i=1}^N \sum_{t=1}^T \tilde{\mathbf{x}}_{it}' \tilde{\mathbf{x}}_{it} \right)^{-1}.$$

d. If we want to allow serial correlation in the  $\{u_{it}\}$ , or allow  $\text{Var}(u_{it}|\mathbf{x}_i, \mathbf{h}_i, c_i) \neq \sigma_u^2 h_{it}$ , then we can just apply the robust formula for the pooled OLS regression (10.92). See equation (7.77) in the text.

**10.14.** a. Because  $E(h_i|\mathbf{z}_i) = 0$ ,  $\mathbf{z}_i\boldsymbol{\gamma}$  and  $h_i$  are uncorrelated, so  $\text{Var}(c_i) = \text{Var}(\mathbf{z}_i\boldsymbol{\gamma}) + \sigma_h^2 = \boldsymbol{\gamma}'\text{Var}(\mathbf{z}_i)\boldsymbol{\gamma} + \sigma_h^2 \geq \sigma_h^2$ . Assuming that  $\text{Var}(\mathbf{z}_i)$  is positive definite – which we must to satisfy the RE rank condition – strict inequality holds whenever  $\boldsymbol{\gamma} \neq \mathbf{0}$ . Of course  $\boldsymbol{\gamma}'\text{Var}(\mathbf{z}_i)\boldsymbol{\gamma} > 0$  is possible even if  $\text{Var}(\mathbf{z}_i)$  is not positive definite.

b. If we estimate the model by fixed effects, the associated estimate of the variance of the unobserved effect is  $\sigma_c^2$ . If we estimate the model by random effects (with, of course,  $\mathbf{z}_i$  included), the variance component is  $\sigma_h^2$ . This makes intuitive sense: with random effects we are able to explicitly control for time-constant variances, and so the  $\mathbf{z}_i$  are effectively taken out of  $c_i$  with  $h_i$  as the remainder.

c. Using equation (10.81), we obtain  $\lambda_h$  and  $\lambda_c$  as follows.

$$\lambda_h = 1 - \{1/[1 + T(\sigma_h^2/\sigma_u^2)]\}^{1/2}$$

$$\lambda_c = 1 - \{1/[1 + T(\sigma_c^2/\sigma_u^2)]\}^{1/2}$$

so

$$\lambda_c - \lambda_h = \{1/[1 + T(\sigma_h^2/\sigma_u^2)]\}^{1/2} - \{1/[1 + T(\sigma_c^2/\sigma_u^2)]\}^{1/2}.$$

Therefore,  $\lambda_c - \lambda_h \geq 0$  iff  $1/[1 + T(\sigma_h^2/\sigma_u^2)] \geq 1/[1 + T(\sigma_c^2/\sigma_u^2)]$  (because  $1 + T(\sigma_h^2/\sigma_u^2)$  and  $1 + T(\sigma_c^2/\sigma_u^2)$  are positive). We conclude that

$$\lambda_c - \lambda_h \geq 0 \text{ iff } T(\sigma_c^2/\sigma_u^2) \geq T(\sigma_h^2/\sigma_u^2)$$

which holds because we already showed that  $\sigma_c^2 \geq \sigma_h^2$  (often with strict inequality).

d. If we use FE to estimate the heterogeneity variance then we are estimating  $\sigma_c^2$ , which means  $\lambda_c$  is effectively the quasi-time demeaning parameter used in subsequent RE estimation. If we use POLS, then we estimate  $\sigma_h^2$ , which then delivers the appropriate quasi-time demeaning parameter  $\lambda_h$ . Thus, we should use POLS, not FE, as the initial estimator for RE estimation when the model includes time-constant variables.

e. Because Problem 7.15 contains a more general result, a separate proof is not provided here. One need not have an RE structure and, as mentioned in the test, we do not need homoskedasticity of  $\text{Var}(c_i|\mathbf{z}_i)$ , either.

**10.15.** a. Because  $\mathbf{v}_i$  is independent of  $\mathbf{x}_i$ ,  $\bar{\mathbf{v}}_i$  is also independent of  $\mathbf{x}_i$ . Therefore,

$$E(v_{it}|\mathbf{x}_i, \bar{\mathbf{v}}_i) = E(v_{it}|\bar{\mathbf{v}}_i), t = 1, \dots, T.$$

Because we assume linearity, we know  $E(v_{it}|\bar{\mathbf{v}}_i)$  is the linear projection (with intercept zero because  $E(v_{it}) = 0$  for all  $t$ ):

$$E(v_{it}|\bar{v}_i) = \frac{\text{Cov}(\bar{v}_i, v_{it})}{\text{Var}(\bar{v}_i)} \cdot \bar{v}_i.$$

Now

$$\begin{aligned} \text{Cov}(\bar{v}_i, v_{it}) &= T^{-1} \text{Cov}\left(\sum_{r=1}^T v_{ir}, v_{it}\right) = T^{-1} \left[ \text{Var}(v_{it}) + \sum_{r \neq t}^T \text{Cov}(v_{ir}, v_{it}) \right] \\ &= T^{-1}[(\sigma_c^2 + \sigma_u^2) + (T-1)\sigma_c^2] = \sigma_c^2 + \sigma_u^2/T \end{aligned}$$

Also, because  $\bar{v}_i = c_i + \bar{u}_i$ , and  $\{u_{it} : t = 1, \dots, T\}$  is serially uncorrelated with constant variance, and each  $u_{it}$  is uncorrelated with  $c_i$ ,

$$\text{Var}(\bar{v}_i) = \text{Var}(c_i) + \text{Var}(\bar{u}_i) = \sigma_c^2 + \sigma_u^2/T = \text{Cov}(\bar{v}_i, v_{it})$$

We have shown that the slope in the population regression of  $v_{it}$  on  $\bar{v}_i$  is unity, and so

$$E(v_{it}|\bar{v}_i) = \bar{v}_i.$$

b. Because  $\bar{y}_i = \bar{\mathbf{x}}_i\boldsymbol{\beta} + \bar{v}_i$ ,  $E(v_{it}|\mathbf{x}_i, \bar{v}_i) = E(v_{it}|\mathbf{x}_i, \bar{y}_i)$ , and so

$$\begin{aligned} E(y_{it}|\mathbf{x}_i, \bar{y}_i) &= E(\mathbf{x}_{it}\boldsymbol{\beta} + v_{it}|\mathbf{x}_i, \bar{y}_i) = \mathbf{x}_{it}\boldsymbol{\beta} + E(v_{it}|\mathbf{x}_i, \bar{y}_i) = \mathbf{x}_{it}\boldsymbol{\beta} + \bar{v}_i \\ &= \mathbf{x}_{it}\boldsymbol{\beta} + (\bar{y}_i - \bar{\mathbf{x}}_i\boldsymbol{\beta}). \end{aligned}$$

c. We can rewrite equation in part b as follows.

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}\boldsymbol{\beta} + \bar{y}_i - \bar{\mathbf{x}}_i\boldsymbol{\beta} + r_{it} \\ E(r_{it}|\mathbf{x}_i, \bar{y}_i) &= 0 \end{aligned}$$

To impose the coefficient of unity on  $\bar{y}_i$  and the common vector on  $\mathbf{x}_{it}$  and  $\bar{\mathbf{x}}_i$ , write

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\beta} + r_{it}, t = 1, \dots, T,$$

which we can estimate consistently by POLS because  $E(r_{it}|\mathbf{x}_i) = 0$  for all  $t$ .

d. The RE estimator is not based on  $E(y_{it}|\mathbf{x}_i, \bar{y}_i) = E(y_{it}|\mathbf{x}_{it}, \bar{y}_i, \bar{\mathbf{x}}_i)$ , but on

$E(y_{it}|\mathbf{x}_i) = E(y_{it}|\mathbf{x}_{it})$ . We now see that the FE estimator can be given an interpretation as an

estimator that “controls for”  $\bar{y}_i$ , along with  $\bar{\mathbf{x}}_i$  (even though it does not need to under the RE assumptions).

e. Under Assumptions RE1 to RE3 we can derive

$$L(v_{it}|\mathbf{x}_i, \bar{v}_i) = L(v_{it}|\bar{v}_i), t = 1, \dots, T$$

without further assumptions. First, we know that the LP has the form

$$L(v_{it}|\mathbf{x}_i, \bar{v}_i) = \mathbf{x}_i \boldsymbol{\psi}_t + \rho_t \bar{v}_i \equiv \mathbf{w}_i \boldsymbol{\delta}_t$$

where

$$\boldsymbol{\delta}_t = [E(\mathbf{w}_i' \mathbf{w}_i)]^{-1} E(\mathbf{w}_i' v_{it})$$

But  $E(\mathbf{w}_i' \mathbf{w}_i)$  is block diagonal because  $E(\mathbf{x}_i' \bar{v}_i) = \mathbf{0}$ . Further,  $E(\mathbf{x}_i' v_{it}) = \mathbf{0}$ , and so

$$\boldsymbol{\delta}_t = \begin{pmatrix} \mathbf{0} \\ \rho_t \end{pmatrix}$$

It is easy to see that

$$\rho_t = \frac{E(\bar{v}_i v_{it})}{E(\bar{v}_i^2)} = \frac{\text{Cov}(\bar{v}_i, v_{it})}{\text{Var}(\bar{v}_i)} = 1,$$

**10.16.** a. By independence between  $(v_{i1}, v_{i2}, \dots, v_{iT}, v_{i,T+1})$  and  $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT}, \mathbf{x}_{i,T+1})$ , it follows immediately that

$$E(v_{i,T+1}|\mathbf{x}_i, \mathbf{x}_{i,T+1}, v_{i1}, \dots, v_{iT}) = E(v_{i,T+1}|v_{i1}, \dots, v_{iT}).$$

Because we are assuming that all conditional expectations involving  $\{v_{it}\}$  are linear, we know

$E(v_{i,T+1}|v_{i1}, \dots, v_{iT})$  is a linear function of  $(v_{i1}, \dots, v_{iT})$ . The tricky part is to show that

$E(v_{i,T+1}|v_{i1}, \dots, v_{iT}) = E(v_{i,T+1}|\bar{v}_i)$ . Intuitively it makes sense that the elements in  $(v_{i1}, \dots, v_{iT})$

should get equal weight under the RE variance-covariance structure. One way to verify that

each element gets the same weight, and to determine that common weight, is to note that the

vector in the LP, say  $\mathbf{p}_T$ , satisfies

$$\begin{pmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \cdots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma_c^2 \\ \sigma_c^2 & \cdots & \sigma_c^2 & \sigma_c^2 + \sigma_u^2 \end{pmatrix} \mathbf{p}_T = \begin{pmatrix} \sigma_c^2 \\ \sigma_c^2 \\ \vdots \\ \sigma_c^2 \end{pmatrix}.$$

If we hypothesis that  $\mathbf{p}_T = \eta_T \mathbf{j}_T'$ , where  $\mathbf{j}_T$  is the  $T \times 1$  vector of ones, then

$$(T\sigma_c^2 + \sigma_u^2)\eta_T = \sigma_c^2$$

so

$$\eta_T = \frac{\sigma_c^2}{(T\sigma_c^2 + \sigma_u^2)}$$

This must be the unique solution because the RE variance matrix is assumed to be nonsingular ( $\sigma_u^2 > 0$ ). So we have shown

$$\begin{aligned} E(v_{i,T+1} | \mathbf{x}_i, \mathbf{x}_{i,T+1}, v_{i1}, \dots, v_{iT}) &= \left[ \frac{\sigma_c^2}{(T\sigma_c^2 + \sigma_u^2)} \right] (v_{i1} + v_{i2} + \dots + v_{iT}) \\ &= \left[ \frac{T\sigma_c^2}{(T\sigma_c^2 + \sigma_u^2)} \right] \bar{v}_i \\ &= \left[ \frac{\sigma_c^2}{(\sigma_c^2 + \sigma_u^2/T)} \right] \bar{v}_i \end{aligned}$$

which is what the problem asked to show.

b. This is straightforward given part a:

$$\begin{aligned} E(y_{i,T+1} | \mathbf{x}_i, \mathbf{x}_{i,T+1}, v_{i1}, \dots, v_{iT}) &= \mathbf{x}_{i,T+1} \boldsymbol{\beta} + E(v_{i,T+1} | \mathbf{x}_i, \mathbf{x}_{i,T+1}, v_{i1}, \dots, v_{iT}) \\ &= \mathbf{x}_{i,T+1} \boldsymbol{\beta} + E(v_{i,T+1} | \bar{v}_i) = \mathbf{x}_{i,T+1} \boldsymbol{\beta} + \left[ \frac{\sigma_c^2}{(\sigma_c^2 + \sigma_u^2/T)} \right] \bar{v}_i \end{aligned}$$

c. If we condition only on the history of covariates, and not the past composite errors, we get

$$E(y_{i,T+1}|\mathbf{x}_i, \mathbf{x}_{i,T+1}) = \mathbf{x}_{i,T+1}\boldsymbol{\beta} + E(v_{i,T+1}|\mathbf{x}_i, \mathbf{x}_{i,T+1}) = \mathbf{x}_{i,T+1}\boldsymbol{\beta} + 0 = \mathbf{x}_{i,T+1}\boldsymbol{\beta}$$

because we are assuming RE.1 for all  $T + 1$  time periods.

d. Forecast errors are given by:

$$\begin{aligned} y_{i,T+1} - E(y_{i,T+1}|\mathbf{x}_i, \mathbf{x}_{i,T+1}, v_{i1}, \dots, v_{iT}) &= (\mathbf{x}_{i,T+1}\boldsymbol{\beta} + v_{i,T+1}) - \left\{ \mathbf{x}_{i,T+1}\boldsymbol{\beta} - \left[ \frac{\sigma_c^2}{(\sigma_c^2 + \sigma_u^2/T)} \right] \bar{v}_i \right\} \\ &= v_{i,T+1} - \left[ \frac{\sigma_c^2}{(\sigma_c^2 + \sigma_u^2/T)} \right] \bar{v}_i \end{aligned}$$

Let  $\theta = \sigma_c^2/(\sigma_c^2 + \sigma_u^2/T)$ . Then the variance of the forecast error is

$$\begin{aligned} \text{Var}(v_{i,T+1} - \theta \bar{v}_i) &= \text{Var}(v_{i,T+1}) + \theta^2 \text{Var}(\bar{v}_i) - 2\theta \text{Cov}(v_{i,T+1}, \bar{v}_i) \\ &= \sigma_c^2 + \sigma_u^2 - 2\theta(\sigma_c^2) + \theta^2(\sigma_c^2 + \sigma_u^2/T) \\ &= \sigma_c^2 + \sigma_u^2 + \frac{\sigma_c^2 \sigma_c^2}{(\sigma_c^2 + \sigma_u^2/T)} - 2 \frac{\sigma_c^2 \sigma_c^2}{(\sigma_c^2 + \sigma_u^2/T)} \\ &= \sigma_c^2 + \sigma_u^2 - \sigma_c^2 \theta. \end{aligned}$$

If we use  $E(y_{i,T+1}|\mathbf{x}_i, \mathbf{x}_{i,T+1})$  to forecast  $y_{i,T+1}$  the forecast error is simply

$v_{i,T+1} = y_{i,T+1} - E(y_{i,T+1}|\mathbf{x}_i, \mathbf{x}_{i,T+1})$ , which has variance  $\sigma_c^2 + \sigma_u^2$ . Because  $\sigma_c^2 \geq 0$  and  $\theta \geq 0$ ,

$$\text{Var}(v_{i,T+1} - \theta \bar{v}_i) \leq \text{Var}(v_{i,T+1})$$

with strict inequality when  $\sigma_c^2 > 0$ . Of course, all we have really shown is that

$\text{Var}[E(v_{i,T+1}|\bar{v}_i)] \leq \text{Var}(v_{i,T+1})$ , which we already know from general properties of conditional expectations. Using more information in a conditional mean results in a smaller prediction error variance.

e. We can use  $N$  cross section observations and the first  $T$  time periods to estimate the parameters by random effects. Let  $\hat{\boldsymbol{\beta}}_{RE}$  be the RE estimator, and let

$$\hat{\theta} = \frac{\hat{\sigma}_c^2}{(\hat{\sigma}_c^2 + \hat{\sigma}_u^2/T)}.$$

Then

$$\hat{y}_{i,T+1} = \mathbf{x}_{i,T+1}\hat{\boldsymbol{\beta}}_{RE} + \hat{\theta}\hat{\bar{v}}_i$$

where

$$\hat{\bar{v}}_i = \frac{1}{T} \sum_{t=1}^T \hat{v}_{it} \text{ and } \hat{v}_{it} = y_{it} - \mathbf{x}_{it}\hat{\boldsymbol{\beta}}_{RE}.$$

**10.17.** a. By the usual averaging across time, the quasi-time-demeaned equation can be written, for each time period, as

$$\begin{aligned} y_{it} - \lambda \bar{y}_i &= (1 - \lambda)\alpha + (\mathbf{d}_t - \lambda \bar{\mathbf{d}})\boldsymbol{\eta} + (\mathbf{w}_{it} - \lambda \bar{\mathbf{w}}_i)\boldsymbol{\delta} + (v_{it} - \lambda \bar{v}_i) \\ &= [(1 - \lambda)\alpha + (1 - \lambda)\bar{\mathbf{d}}\boldsymbol{\eta}] + (\mathbf{d}_t - \bar{\mathbf{d}})\boldsymbol{\eta} + (\mathbf{w}_{it} - \lambda \bar{\mathbf{w}}_i)\boldsymbol{\delta} + (v_{it} - \lambda \bar{v}_i) \end{aligned}$$

which is what we wanted to show by letting  $\mu = (1 - \lambda)\alpha + (1 - \lambda)\bar{\mathbf{d}}\boldsymbol{\eta}$ .

b. The first part – the  $\sqrt{N}$ -asymptotic representation of the RE estimator – is just the usual linear representation of a pooled OLS estimator laid out in Chapter 7. It also follows from the discussion of random effects in Section 10.7.2. For the second part, write

$v_{it} - \lambda \bar{v}_i = (c_i + u_{it}) - (\lambda c_i + \lambda \bar{u}_i) = (1 - \lambda)c_i + u_{it} - \lambda \bar{u}_i$ , and plug in:

$$\begin{aligned} \sum_{t=1}^T (\mathbf{d}_t - \bar{\mathbf{d}})(v_{it} - \lambda \bar{v}_i) &= \sum_{t=1}^T (\mathbf{d}_t - \bar{\mathbf{d}})(1 - \lambda)c_i + \sum_{t=1}^T (\mathbf{d}_t - \bar{\mathbf{d}})u_{it} - \sum_{t=1}^T (\mathbf{d}_t - \bar{\mathbf{d}})\lambda \bar{u}_i \\ &= (1 - \lambda)c_i \sum_{t=1}^T (\mathbf{d}_t - \bar{\mathbf{d}}) + \sum_{t=1}^T (\mathbf{d}_t - \bar{\mathbf{d}})u_{it} - (\lambda \bar{u}_i) \sum_{t=1}^T (\mathbf{d}_t - \bar{\mathbf{d}}) \\ &= \sum_{t=1}^T (\mathbf{d}_t - \bar{\mathbf{d}})u_{it} \end{aligned}$$

because  $\sum_{t=1}^T (\mathbf{d}_t - \bar{\mathbf{d}}) = \mathbf{0}$ .

c. Actually, there is nothing to do here. This is just the usual first-order representation of the fixed effects estimator, which follows from the general pooled OLS results.

d. From part b we can write

$$\mathbf{A}_1 \sqrt{N} (\hat{\boldsymbol{\beta}}_{RE} - \boldsymbol{\beta}) = \left( N^{-1/2} \sum_{i=1}^N \sum_{t=1}^T \mathbf{g}'_{it} (v_{it} - \lambda \bar{v}_i) \right) + o_p(1) \equiv \left( N^{-1/2} \sum_{i=1}^N \sum_{t=1}^T \mathbf{r}_{it} \right) + o_p(1),$$

where  $\mathbf{r}_{it} = [(\mathbf{d}_t - \bar{\mathbf{d}})u_{it}, (\mathbf{w}_{it} - \lambda \bar{\mathbf{w}}_i)(v_{it} - \lambda \bar{v}_i)]'$ . From part c we can write

$$\mathbf{A}_2 \sqrt{N} (\hat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta}) = \left( N^{-1/2} \sum_{i=1}^N \sum_{t=1}^T \mathbf{h}'_{it} u_{it} \right) + o_p(1) \equiv \left( N^{-1/2} \sum_{i=1}^N \sum_{t=1}^T \mathbf{s}_{it} \right) + o_p(1),$$

where  $\mathbf{s}_{it} = [(\mathbf{d}_t - \bar{\mathbf{d}})u_{it}, (\mathbf{w}_{it} - \bar{\mathbf{w}}_i)u_{it}]'$ . But the first  $R$  elements of  $\mathbf{r}_{it}$  and  $\mathbf{s}_{it}$  are  $(\mathbf{d}_t - \bar{\mathbf{d}})'u_{it}$ ,

which implies that

$$\mathbf{A}_1 \sqrt{N} (\hat{\boldsymbol{\beta}}_{RE} - \boldsymbol{\beta}) - \mathbf{A}_2 \sqrt{N} (\hat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta}) = N^{-1/2} \sum_{i=1}^N \sum_{t=1}^T \begin{pmatrix} \mathbf{0} \\ (\mathbf{w}_{it} - \lambda \bar{\mathbf{w}}_i)'e_{it} - (\mathbf{w}_{it} - \bar{\mathbf{w}}_i)'u_{it} \end{pmatrix} + o_p(1),$$

where “ $\mathbf{0}$ ” is  $R \times 1$  and  $e_{it} = v_{it} - \lambda \bar{v}_i$ . The second part of the vector is  $M \times 1$  and generally satisfies the central limit theorem. Under standard rank assumptions

$\text{Var}[(\mathbf{w}_{it} - \lambda \bar{\mathbf{w}}_i)'e_{it} - (\mathbf{w}_{it} - \bar{\mathbf{w}}_i)'u_{it}]$  has rank  $M$

e. If there were no  $\mathbf{w}_{it}$ , part d would imply that the limiting distribution of the difference between RE and FE is degenerate. In other words, we cannot compute the Hausman test comparing the FE and RE estimators if the only time-varying covariates are aggregates. (In fact, the FE and RE estimates are numerically identical in this case.) More generally, the variance-covariance matrix of the difference has rank  $M$ , not  $M + R$  (whether or not we assume RE.3 under  $H_0$ ). A properly computed Hausman test will have only  $M$  degrees-of-freedom, not  $M + R$ . The regression-based test from equation (10.88) forces one to get the degrees-of-freedom correct, as there is obviously no value in adding  $\bar{\mathbf{d}}$ , a vector of constants, to the regression.



**10.18. a.** The Stata results are given below. All estimates are numerically identical.

```
. reg lwage d81-d87
```

Source	SS	df	MS	Number of obs =	4360
Model	92.9668229	7	13.2809747	F( 7, 4352) =	50.54
Residual	1143.56282	4352	.262767192	Prob > F =	0.0000
Total	1236.52964	4359	.283672779	R-squared =	0.0752
				Adj R-squared =	0.0737
				Root MSE =	.51261

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
d81	.1193902	.0310529	3.84	0.000	.0585107	.1802697
d82	.1781901	.0310529	5.74	0.000	.1173106	.2390696
d83	.2257865	.0310529	7.27	0.000	.1649069	.286666
d84	.2968181	.0310529	9.56	0.000	.2359386	.3576976
d85	.3459333	.0310529	11.14	0.000	.2850538	.4068128
d86	.4062418	.0310529	13.08	0.000	.3453623	.4671213
d87	.4730023	.0310529	15.23	0.000	.4121228	.5338818
_cons	1.393477	.0219577	63.46	0.000	1.350429	1.436525

```
. xtreg lwage d81-d87, re
```

Random-effects GLS regression	Number of obs =	4360
Group variable: nr	Number of groups =	545
R-sq: within = 0.0000	Obs per group: min =	
between = 0.0000	avg =	8.
overall = 0.0752	max =	

Random effects u_i ~Gaussian	Wald chi2(7) =	738.94
corr(u_i, X) = 0 (assumed)	Prob > chi2 =	0.0000

lwage	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
d81	.1193902	.021487	5.56	0.000	.0772765	.1615039
d82	.1781901	.021487	8.29	0.000	.1360764	.2203038
d83	.2257865	.021487	10.51	0.000	.1836728	.2679001
d84	.2968181	.021487	13.81	0.000	.2547044	.3389318
d85	.3459333	.021487	16.10	0.000	.3038196	.388047
d86	.4062418	.021487	18.91	0.000	.3641281	.4483555
d87	.4730023	.021487	22.01	0.000	.4308886	.515116
_cons	1.393477	.0219577	63.46	0.000	1.350441	1.436513
sigma_u	.37007665					
sigma_e	.35469771					
rho	.52120938	(fraction of variance due to u_i)				

```
. xtreg lwage d81-d87, fe
```

Fixed-effects (within) regression	Number of obs =	4360
Group variable: nr	Number of groups =	545

R-sq: within = 0.1625  
 between = .  
 overall = 0.0752

Obs per group: min =  
 avg = 8.  
 max =

corr(u\_i, Xb) = 0.0000  
 F(7,3808) = 105.56  
 Prob > F = 0.0000

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
d81	.1193902	.021487	5.56	0.000	.0772631	.1615173
d82	.1781901	.021487	8.29	0.000	.136063	.2203172
d83	.2257865	.021487	10.51	0.000	.1836594	.2679135
d84	.2968181	.021487	13.81	0.000	.254691	.3389452
d85	.3459333	.021487	16.10	0.000	.3038063	.3880604
d86	.4062418	.021487	18.91	0.000	.3641147	.4483688
d87	.4730023	.021487	22.01	0.000	.4308753	.5151294
_cons	1.393477	.0151936	91.71	0.000	1.363689	1.423265
sigma_u	.39074676					
sigma_e	.35469771					
rho	.54824631	(fraction of variance due to u_i)				

F test that all u\_i=0: F(544, 3808) = 9.71 Prob > F = 0.0000

. reg d.(lwage d81-d87), nocons

Source	SS	df	MS	Number of obs =	3815
Model	19.3631642	7	2.76616631	F( 7, 3808) =	14.06
Residual	749.249837	3808	.196756785	Prob > F =	0.0000
				R-squared =	0.0252
				Adj R-squared =	0.0234
Total	768.613001	3815	.201471298	Root MSE =	.44357

D.lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
d81						
D1.	.1193902	.0190006	6.28	0.000	.0821379	.1566425
d82						
D1.	.1781901	.0268709	6.63	0.000	.1255074	.2308728
d83						
D1.	.2257865	.03291	6.86	0.000	.1612636	.2903093
d84						
D1.	.2968181	.0380011	7.81	0.000	.2223136	.3713226
d85						
D1.	.3459333	.0424866	8.14	0.000	.2626347	.4292319
d86						
D1.	.4062418	.0465417	8.73	0.000	.3149927	.4974908
d87						
D1.	.4730023	.0502708	9.41	0.000	.3744421	.5715626

b. The Stata output follows. The POLS and RE estimates are identical on the year dummies and the three time-constant variables. This is a general result: if the model includes only aggregate time effects and individual-specific covariates that have no time variation, POLS = RE (and, in particular, there is no efficiency gain in using RE).

When FE is used, of course the time-constant variables drop out. The estimates on the year dummies are the same as POLS and RE. (Recall that the “constant” reported by FE is the average of the estimated heterogeneity terms. When POLS and RE include time-constant variables the FE “constant” does not equal the intercept from POLS/RE.)

```
. reg lwage d81-d87 educ black hisp
```

Source	SS	df	MS	Number of obs =	4360
Model	179.091659	10	17.9091659	F( 10, 4349) =	73.66
Residual	1057.43798	4349	.243145087	Prob > F =	0.0000
				R-squared =	0.1448
				Adj R-squared =	0.1429
Total	1236.52964	4359	.283672779	Root MSE =	.4931

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
d81	.1193902	.029871	4.00	0.000	.0608279	.1779526
d82	.1781901	.029871	5.97	0.000	.1196277	.2367524
d83	.2257865	.029871	7.56	0.000	.1672241	.2843488
d84	.2968181	.029871	9.94	0.000	.2382557	.3553804
d85	.3459333	.029871	11.58	0.000	.287371	.4044957
d86	.4062418	.029871	13.60	0.000	.3476794	.4648041
d87	.4730023	.029871	15.83	0.000	.41444	.5315647
educ	.0770943	.0043766	17.62	0.000	.0685139	.0856747
black	-.1225637	.0237021	-5.17	0.000	-.1690319	-.0760955
hisp	.024623	.0213056	1.16	0.248	-.0171468	.0663928
_cons	.4966384	.0566686	8.76	0.000	.3855391	.6077377

```
. xtreg lwage d81-d87 educ black hisp, re
```

Random-effects GLS regression	Number of obs =	4360
Group variable: nr	Number of groups =	545
R-sq: within = 0.1625	Obs per group: min =	
between = 0.1296	avg =	8.
overall = 0.1448	max =	
Random effects u_i ~Gaussian	Wald chi2(10) =	819.51
corr(u_i, X) = 0 (assumed)	Prob > chi2 =	0.0000

lwage	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
d81	.1193902	.021487	5.56	0.000	.0772765	.1615039
d82	.1781901	.021487	8.29	0.000	.1360764	.2203038
d83	.2257865	.021487	10.51	0.000	.1836728	.2679001
d84	.2968181	.021487	13.81	0.000	.2547044	.3389318
d85	.3459333	.021487	16.10	0.000	.3038196	.388047
d86	.4062418	.021487	18.91	0.000	.3641281	.4483555
d87	.4730023	.021487	22.01	0.000	.4308886	.515116
educ	.0770943	.009177	8.40	0.000	.0591076	.0950809
black	-.1225637	.0496994	-2.47	0.014	-.2199728	-.0251546
hisp	.024623	.0446744	0.55	0.582	-.0629371	.1121831
_cons	.4966384	.1122718	4.42	0.000	.2765897	.7166871
sigma_u	.34337144					
sigma_e	.35469771					
rho	.48377912	(fraction of variance due to u_i)				

```
. xtreg lwage d81-d87 educ black hisp, fe
note: educ omitted because of collinearity
note: black omitted because of collinearity
note: hisp omitted because of collinearity
```

```
Fixed-effects (within) regression      Number of obs   =      4360
Group variable: nr                     Number of groups =      545

R-sq:  within  = 0.1625                Obs per group:  min =
        between = .                                avg   =      8.
        overall = 0.0752                                max   =

                                         F(7,3808)       =    105.56
corr(u_i, Xb) = 0.0000                 Prob > F        =    0.0000
```

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
d81	.1193902	.021487	5.56	0.000	.0772631	.1615173
d82	.1781901	.021487	8.29	0.000	.136063	.2203172
d83	.2257865	.021487	10.51	0.000	.1836594	.2679135
d84	.2968181	.021487	13.81	0.000	.254691	.3389452
d85	.3459333	.021487	16.10	0.000	.3038063	.3880604
d86	.4062418	.021487	18.91	0.000	.3641147	.4483688
d87	.4730023	.021487	22.01	0.000	.4308753	.5151294
educ	(omitted)					
black	(omitted)					
hisp	(omitted)					
_cons	1.393477	.0151936	91.71	0.000	1.363689	1.423265
sigma_u	.39074676					
sigma_e	.35469771					
rho	.54824631	(fraction of variance due to u_i)				

```
F test that all u_i=0:      F(544, 3808) =      8.45      Prob > F = 0.0000
```

c. The reported standard errors for POLS and RE are not the same. The POLS standard errors assume, in addition to homoskedasticity, no serial correlation in the composite error – in other words, that there is no unobserved heterogeneity. At least the RE standard errors allow for the standard RE structure, which means constant variance and correlations that are the same across all pairs  $(t,s)$ . This may be too restrictive, but it is less restrictive than the usual OLS standard errors.

d. The fully robust POLS standard errors – that allow any kind of serial correlation and heteroskedasticity – are reported below. We prefer these to the usual RE standard errors because, as noted in part c, the usual RE standard errors impose a special kind of serial correlation. Notice that the fully robust POLS standard errors are not uniformly larger than the usual RE standard errors.

```
. reg lwage d81-d87 educ black hisp, cluster(nr)
```

Linear regression

Number of obs = 4360

F( 10, 544) = 49.41

Prob > F = 0.0000

R-squared = 0.1448

Root MSE = .4931

(Std. Err. adjusted for 545 clusters in nr

-----

lwage	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
d81	.1193902	.0244086	4.89	0.000	.0714435	.1673369
d82	.1781901	.0241987	7.36	0.000	.1306558	.2257243
d83	.2257865	.0243796	9.26	0.000	.1778968	.2736761
d84	.2968181	.0271485	10.93	0.000	.2434894	.3501468
d85	.3459333	.0263181	13.14	0.000	.2942358	.3976309
d86	.4062418	.0273064	14.88	0.000	.3526029	.4598807
d87	.4730023	.025996	18.20	0.000	.4219374	.5240672
educ	.0770943	.0090198	8.55	0.000	.0593763	.0948122
black	-.1225637	.0532662	-2.30	0.022	-.2271964	-.017931
hisp	.024623	.0411235	0.60	0.550	-.0561573	.1054033
_cons	.4966384	.1097474	4.53	0.000	.2810579	.7122189

e. The fully robust standard errors for RE are given below. They are numerically identical

to the fully robust POLS standard errors. Because we really have only one estimator – remember, POLS = RE in this setup – there is one asymptotic variance. While there could be different ways to estimate that asymptotic variance, in this case the estimators are the same, and that is appealing because it means inference does not rely on the particular pre-programmed command.

```
. xtreg lwage d81-d87 educ black hisp, re cluster(nr)
```

```
Random-effects GLS regression                Number of obs    =      4360
Group variable: nr                          Number of groups   =       545

R-sq:   within  = 0.1625                     Obs per group:   min =
        between  = 0.1296                               avg  =      8.
        overall  = 0.1448                               max  =

Random effects u_i ~Gaussian                Wald chi2(10)     =     494.13
corr(u_i, X)      = 0 (assumed)              Prob > chi2       =      0.0000
```

(Std. Err. adjusted for 545 clusters in nr)

lwage	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval	
d81	.1193902	.0244086	4.89	0.000	.0715502	.1672302
d82	.1781901	.0241987	7.36	0.000	.1307616	.2256186
d83	.2257865	.0243796	9.26	0.000	.1780033	.2735696
d84	.2968181	.0271485	10.93	0.000	.2436081	.3500281
d85	.3459333	.0263181	13.14	0.000	.2943508	.3975159
d86	.4062418	.0273064	14.88	0.000	.3527222	.4597613
d87	.4730023	.025996	18.20	0.000	.422051	.5239536
educ	.0770943	.0090198	8.55	0.000	.0594157	.0947728
black	-.1225637	.0532662	-2.30	0.021	-.2269636	-.0181638
hisp	.024623	.0411235	0.60	0.549	-.0559775	.1052236
_cons	.4966384	.1097474	4.53	0.000	.2815375	.7117392
sigma_u	.34337144					
sigma_e	.35469771					
rho	.48377912	(fraction of variance due to u_i)				

## Solutions to Chapter 11 Problems

11.1. a. It is important to remember that, any time we put a variable in a regression model (whether we are using cross section or panel data), we are controlling for the effects of that variable on the dependent variable. The whole point of regression analysis is that it allows the explanatory variables to be correlated while estimating *ceteris paribus* effect of each explanatory variable. Thus, the inclusion of  $y_{i,t-1}$  in the equation allows  $prog_{it}$  to be correlated with  $y_{i,t-1}$ , and also recognizes that, due to inertia,  $y_{it}$  is often strongly related to  $y_{i,t-1}$ .

An assumption that implies pooled OLS is consistent is

$$E(u_{it} | \mathbf{z}_i, \mathbf{x}_{it}, y_{i,t-1}, prog_{it}) = 0, \text{ all } t,$$

which is implied by but is weaker than dynamic completeness. Without additional assumptions, the pooled OLS standard errors and test statistics need to be adjusted for heteroskedasticity and serial correlation (although the latter will not be present under dynamic completeness).

When  $y_{i,t-1}$  is added to a regression model in an astructural way, we can think of the goal as being to estimate

$$E(y_{it} | \mathbf{z}_i, \mathbf{x}_{it}, y_{i,t-1}, prog_{it}),$$

which means that we are controlling for differences in the lagged response when gauging the effect of the program. Of course, we might not have the conditional mean correctly specified; we may be simply estimating a linear projection.

b. As we discussed in Section 7.8.2, this statement is incorrect. Provided our interest is in  $E(y_{it} | \mathbf{z}_i, \mathbf{x}_{it}, y_{i,t-1}, prog_{it})$ , we are not especially concerned about serial correlation in the implied errors,

$$u_{it} \equiv y_{it} - E(y_{it} | \mathbf{z}_i, \mathbf{x}_{it}, y_{i,t-1}, prog_{it}).$$

Nor does serial correlation cause inconsistency in the OLS estimators.

c. Such a model is the standard unobserved effects model:

$$y_{it} = \theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \delta_1 prog_{it} + c_i + u_{it}, \quad t = 1, 2, \dots, T,$$

where the  $\theta_t$  are the time effects (that can be treated as parameters). We would probably assume that  $(\mathbf{x}_{it}, prog_{it})$  is strictly exogenous; the weakest form of strict exogeneity is that  $(\mathbf{x}_{it}, prog_{it})$  is uncorrelated with  $u_{is}$  for all  $t$  and  $s$ . Then we could estimate the equation by fixed effects or first differencing. If the  $u_{it}$  are serially uncorrelated, FE is preferred. We could also do a GLS analysis after the fixed effects or first-differencing transformations, but we should have a large  $N$ .

d. A model that incorporates features from parts a and c is

$$y_{it} = \theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \delta_1 prog_{it} + \rho_1 y_{i,t-1} + c_i + u_{it}, \quad t = 1, \dots, T.$$

Now, program participation can depend on unobserved city heterogeneity as well as on lagged  $y_{it}$  (we assume that  $y_{i0}$  is observed). Fixed effects and first-differencing are both inconsistent and  $N \rightarrow \infty$  with fixed  $T$ .

Assuming that  $E(u_{it} | \mathbf{x}_i, prog_i, y_{i,t-1}, y_{i,t-2}, \dots, y_{i0}) = 0$ , a consistent procedure is obtained by first differencing, to get

$$y_{it} = \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \delta_1 \Delta prog_{it} + \rho_1 \Delta y_{i,t-1} + \Delta u_{it}, \quad t = 2, \dots, T.$$

At time  $t$  and  $\Delta \mathbf{x}_{it}, \Delta prog_{it}$  can be used as their own instruments, along with  $y_{i,t-j}$  for  $j \geq 2$ .

Either pooled 2SLS or a GMM procedure can be used. Past and future values of  $\mathbf{x}_{it}$  can also be used as instruments because  $\{\mathbf{x}_{it}\}$  is strictly exogenous.

**11.2.** a. OLS estimation on the first-differenced equation is inconsistent (for all parameters)



if  $\text{Cov}(\Delta w_i, \Delta u_i) \neq 0$ . Because  $w_{it}$  is correlated with  $u_{it}$ , for all  $t$  we cannot assume that  $\Delta w_i$  and  $\Delta u_i$  are uncorrelated.

b. Because  $u_{it}$  is uncorrelated with  $\mathbf{z}_{i1}, \mathbf{z}_{i2}$ , for  $t = 1, 2$ ,  $\Delta u_i = u_{i2} - u_{i1}$  is uncorrelated with  $\mathbf{z}_{i1}$  and  $\mathbf{z}_{i2}$ , and so  $(\mathbf{z}_{i1}, \mathbf{z}_{i2})$  are exogenous in the equation

$$\Delta y_i = \Delta \mathbf{z}_i \boldsymbol{\gamma} + \delta \Delta w_i + \Delta u_i$$

The linear projection of  $\Delta w_i$  on  $(\mathbf{z}_{i1}, \mathbf{z}_{i2})$  can be written as

$$\Delta w_i = \mathbf{z}_{i1} \boldsymbol{\pi}_1 + \mathbf{z}_{i2} \boldsymbol{\pi}_2 + r_i, \quad E(\mathbf{z}_{it}' r_i) = 0, \quad t = 1, 2.$$

The question is whether the rank condition holds. Rewrite this linear projection in terms of  $\Delta \mathbf{z}_i$  and, say,  $\mathbf{z}_{i1}$  as

$$\Delta w_i = \mathbf{z}_{i1} (\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2) + (\mathbf{z}_{i2} - \mathbf{z}_{i1}) \boldsymbol{\pi}_2 + r_i = \mathbf{z}_{i1} \boldsymbol{\lambda}_1 + \Delta \mathbf{z}_i \boldsymbol{\pi}_2 + r_i,$$

where  $\boldsymbol{\lambda}_1 \equiv \boldsymbol{\pi}_1 - \boldsymbol{\pi}_2$ . If  $\boldsymbol{\lambda}_1 = 0$ , that is  $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$ , then the reduced form of  $\Delta w_i$  depends only on  $\Delta \mathbf{z}_i$ . Because  $\Delta \mathbf{z}_i$  appears in the equation for  $\Delta y_i$ , there are no instruments for  $\Delta w_i$ . Thus, the change in  $w_{it}$  must depend on the level of  $\mathbf{z}_{it}$ , and not just on the change in  $\mathbf{z}_{it}$ .

c. With  $T \geq 2$  time periods we can write the differenced equation as

$$\Delta y_{it} = \Delta \mathbf{z}_{it} \boldsymbol{\delta} + \gamma \Delta w_{it} + \Delta u_{it}, \quad t = 2, \dots, T.$$

Now, under the assumption that  $w_{is}$  is uncorrelated with  $u_{it}$  for  $s < t$ , we have natural instruments for  $\Delta w_{it}$ . At time  $t$ ,  $\Delta u_{it}$  depends on  $u_{it}$  and  $u_{1,t-1}$ . Thus, valid instruments at time  $t$  in the FD equation are  $w_{i,t-2}, \dots, w_{i1}$ . We need  $T \geq 3$  for an IV procedure to work. With  $T = 3$  we have the cross sectional equation

$$\Delta y_{i3} = \Delta \mathbf{z}_{i3} \boldsymbol{\delta} + \gamma \Delta w_{i3} + \Delta u_{i3}$$

and we can instrument for  $\Delta w_{i3}$  with  $w_{i1}$  (and possibly  $\mathbf{z}_{ir}$  from earlier time periods).

With  $T \geq 4$ , we can implement an IV estimator by using the simple pooled IV estimator described Section 11.4. Or, we can use the more efficient GMM procedure. Write the  $T - 1$  time periods as

$$\Delta \mathbf{y}_i = \Delta \mathbf{Z}_i \boldsymbol{\beta} + \gamma \Delta w_i + \Delta u_i,$$

where each data vector or matrix has  $T - 1$  rows. The matrix that includes all possible instruments of observation  $i$  (with  $T - 1$  rows) is

$$\begin{pmatrix} \mathbf{z}_i & \mathbf{0} & 0 & \mathbf{0} & 0 & 0 & 0 & \cdots & \mathbf{0} & 0 & 0 \\ \mathbf{0} & \mathbf{z}_i & w_{i1} & \mathbf{0} & \cdots & 0 & 0 & \cdots & \mathbf{0} & 0 & 0 \\ \mathbf{0} & \mathbf{0} & 0 & \mathbf{z}_i & w_{i2} & w_{i1} & 0 & \cdots & \mathbf{0} & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & 0 & \mathbf{0} & 0 & 0 & 0 & 0 & \mathbf{z}_i & w_{i,T-2} \cdots & w_{i1} \end{pmatrix}.$$

Putting in the levels of all  $\mathbf{z}_{it}$  for instruments in each time period is perhaps using too many overidentifying restrictions. The dimension could be reduced substantially by using only  $(\mathbf{z}_{it}, \mathbf{z}_{i,t-1})$  at period  $t$  rather than  $\mathbf{z}_i$ . Further, periods for  $t \geq 3$  one would use only  $w_{i,t-2}$  and  $w_{i,t-3}$  as the IVs.

d. Generally, the IV estimator applied to the time-demeaned equation is inconsistent. This is because  $w_{i,t-j}$  is generally correlated with  $\ddot{u}_{it}$ , as the latter depends on the idiosyncratic errors in all time periods.

**11.3.** Writing  $y_{it} = \beta x_{it} + c_i + u_{it} - \beta r_{it}$ , the fixed effects estimator  $\hat{\beta}_{FE}$  can be written as

$$\beta + \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i) \right)^2 \left( N^{-1} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i) (u_{it} - \bar{u}_i - \beta(r_{it} - \bar{r}_i)) \right).$$

Now  $x_{it} - \bar{x}_i = (x_{it}^* - \bar{x}_i^*) + (r_{it} - \bar{r}_i)$ . Then, because  $E(r_{it} | \mathbf{x}_i^*, c_i) = 0$  for all  $t$ ,  $(x_{it}^* - \bar{x}_i^*)$  and  $(r_{it} - \bar{r}_i)$  are uncorrelated, and so

$$\text{Var}(x_{it} - \bar{x}_i) = \text{Var}(x_{it}^* - \bar{x}_i^*) + \text{Var}(r_{it} - \bar{r}_i), \text{ all } t.$$

Similarly, under (11.42),  $(x_{it} - \bar{x}_i)$  and  $(u_{it} - \bar{u}_i)$  are uncorrelated for all  $t$ . Now

$E[(x_{it} - \bar{x}_i)(r_{it} - \bar{r}_i)] = E[\{(x_{it}^* - \bar{x}_i^*) + (r_{it} - \bar{r}_i)\}(r_{it} - \bar{r}_i)] = \text{Var}(r_{it} - \bar{r}_i)$ . By the law of large numbers and the assumption of constant variances across  $t$ ,

$$N^{-1} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i) \xrightarrow{p} \sum_{t=1}^T \text{Var}(x_{it} - \bar{x}_i) = T[\text{Var}(x_{it}^* - \bar{x}_i^*) + \text{Var}(r_{it} - \bar{r}_i)]$$

and

$$N^{-1} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)[u_{it} - \bar{u}_i - \beta(r_{it} - \bar{r}_i)] \xrightarrow{p} -T\beta \cdot \text{Var}(r_{it} - \bar{r}_i).$$

Therefore,

$$\text{plim} \hat{\beta}_{FE} = \beta - \beta \left( \frac{\text{Var}(r_{it} - \bar{r}_i)}{[\text{Var}(x_{it}^* - \bar{x}_i^*) + \text{Var}(r_{it} - \bar{r}_i)]} \right) = \beta \left( 1 - \frac{\text{Var}(r_{it} - \bar{r}_i)}{[\text{Var}(x_{it}^* - \bar{x}_i^*) + \text{Var}(r_{it} - \bar{r}_i)]} \right).$$

**11.4. a.** For each  $i$  we can average across  $t$  and rearrange to get

$$c_i = \bar{y}_i - \bar{\mathbf{x}}_i \boldsymbol{\beta} - \bar{u}_i.$$

Because  $E(\bar{u}_i) = 0$ ,  $\mu_c \equiv E(c_i) = E(\bar{y}_i - \bar{\mathbf{x}}_i \boldsymbol{\beta})$ . By the law of large numbers,

$$N^{-1} \sum_{i=1}^N c_i = N^{-1} \sum_{i=1}^N (\bar{y}_i - \bar{\mathbf{x}}_i \boldsymbol{\beta}) \xrightarrow{p} \mu_c.$$

Now replace  $\boldsymbol{\beta}$  with  $\hat{\boldsymbol{\beta}}_{FE}$  and call the estimator  $\hat{\mu}_c$ :

$$\begin{aligned} \hat{\mu}_c &= N^{-1} \sum_{i=1}^N (\bar{y}_i - \bar{\mathbf{x}}_i \hat{\boldsymbol{\beta}}_{FE}) = N^{-1} \sum_{i=1}^N (\bar{y}_i - \bar{\mathbf{x}}_i \boldsymbol{\beta}) - \left( N^{-1} \sum_{i=1}^N \bar{\mathbf{x}}_i \right) (\hat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta}) \\ &= N^{-1} \sum_{i=1}^N c_i + O_p(1) \cdot o_p(1) = N^{-1} \sum_{i=1}^N c_i + o_p(1) \xrightarrow{p} \mu_c, \end{aligned}$$

where we use  $N^{-1} \sum_{i=1}^N \bar{\mathbf{x}}_i = O_p(1)$  (by the law of large numbers – see Lemma 3.2) and

$$\hat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta} = o_p(1).$$

b. There is more than one way to estimate  $\mu_g$ , but a simple approach is to first difference, giving

$$\Delta y_{it} = g_i + \mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}, t = 2, \dots, T.$$

Then we can estimate  $\boldsymbol{\beta}$  by fixed effects on the first differences (using  $t = 2, \dots, T$ ), and then apply the estimator from part a to the first differenced data. This means we just replace  $\bar{y}_i$  with  $\overline{\Delta y}_i$  and  $\bar{\mathbf{x}}_i$  with  $\overline{\Delta \mathbf{x}}_i$  everywhere (and the time averages are based on  $T - 1$ , not  $T$ , time periods).

**11.5.** a.  $E(\mathbf{v}_i|\mathbf{z}_i, \mathbf{x}_i) = \mathbf{Z}_i[E(\mathbf{a}_i|\mathbf{z}_i, \mathbf{x}_i) - \boldsymbol{\alpha}] + E(\mathbf{u}_i|\mathbf{z}_i, \mathbf{x}_i) = \mathbf{Z}_i(\boldsymbol{\alpha} - \boldsymbol{\alpha}) + \mathbf{0} = \mathbf{0}$ . Next,

$$\begin{aligned} \text{Var}(\mathbf{v}_i|\mathbf{z}_i, \mathbf{x}_i) &= \mathbf{Z}_i \text{Var}(\mathbf{a}_i|\mathbf{z}_i, \mathbf{x}_i) \mathbf{Z}_i' + \text{Var}(\mathbf{u}_i|\mathbf{z}_i, \mathbf{x}_i) + \text{Cov}(\mathbf{a}_i, \mathbf{u}_i|\mathbf{z}_i, \mathbf{x}_i) + \text{Cov}(\mathbf{u}_i, \mathbf{a}_i|\mathbf{z}_i, \mathbf{x}_i) \\ &= \mathbf{Z}_i \text{Var}(\mathbf{a}_i|\mathbf{z}_i, \mathbf{x}_i) \mathbf{Z}_i' + \text{Var}(\mathbf{u}_i|\mathbf{z}_i, \mathbf{x}_i) \end{aligned}$$

because  $\mathbf{a}_i$  and  $\mathbf{u}_i$  are uncorrelated, conditional on  $(\mathbf{z}_i, \mathbf{x}_i)$ , by Assumption FE.1' and the usual iterated expectations argument,

$$\text{Var}(\mathbf{v}_i|\mathbf{z}_i, \mathbf{x}_i) = \mathbf{Z}_i \boldsymbol{\Lambda} \mathbf{Z}_i' + \sigma_u^2 \mathbf{I}_T.$$

Therefore, under the assumptions given, which shows that the conditional variance depends on  $\mathbf{z}_i$ . Unlike in the standard random effects model, there is conditional heteroskedasticity.

b. If we use the usual RE analysis, we are applying FGLS to the equation

$$\mathbf{y}_i = \mathbf{Z}_i \boldsymbol{\alpha} + \mathbf{X}_i \boldsymbol{\beta} + \mathbf{v}_i, \text{ where } \mathbf{v}_i = \mathbf{Z}_i(\mathbf{a}_i - \boldsymbol{\alpha}) + \mathbf{u}_i. \text{ From part a, we know that } E(\mathbf{v}_i|\mathbf{x}_i, \mathbf{z}_i) = \mathbf{0},$$

and so the usual RE estimator is consistent (as  $N \rightarrow \infty$  for fixed  $T$ ) and  $\sqrt{N}$ -asymptotically normal, provided the rank condition, Assumption RE.2, holds. (Remember, a feasible GLS analysis with any  $\hat{\boldsymbol{\Omega}}$  will be consistent provided  $\hat{\boldsymbol{\Omega}}$  converges in probability to a nonsingular matrix as  $N \rightarrow \infty$ . It need not be the case that  $\text{Var}(\mathbf{v}_i|\mathbf{x}_i, \mathbf{z}_i) = \text{plim}(\hat{\boldsymbol{\Omega}})$ , or even that

$$\text{Var}(\mathbf{v}_i) = \text{plim}(\hat{\mathbf{\Omega}}).$$

From part a, we know that  $\text{Var}(\mathbf{v}_i|\mathbf{x}_i, \mathbf{z}_i)$  depends on  $\mathbf{z}_i$  unless we restrict almost all elements of  $\mathbf{\Lambda}$  to be zero (all but those corresponding to the constant in  $\mathbf{z}_{it}$ ). Therefore, the usual random effects inference – that is, based on the usual RE variance matrix estimator – will be invalid.

c. We can easily make the RE analysis fully robust to an arbitrary  $\text{Var}(\mathbf{v}_i|\mathbf{x}_i, \mathbf{z}_i)$ , as in equation (7.52). Naturally, we expand the set of explanatory variables to  $(\mathbf{z}_{it}, \mathbf{x}_{it})$ , and we estimate  $\mathbf{\alpha}$  along with  $\mathbf{\beta}$ .

**11.6.** No. Assumption (11.42) maintains strict exogeneity of  $\{w_{it}^*\}$  in (11.41), and strict exogeneity clearly fails when  $w_{it}^* = y_{i,t-1}^*$ .

**11.7.** When  $\lambda_t = \lambda/T$  for all  $t$ , we can rearrange (11.6) to get

$$y_{it} = \mathbf{x}_{it}\mathbf{\beta} + \psi + \bar{\mathbf{x}}_i\boldsymbol{\lambda} + r_{it}, t = 1, 2, \dots, T.$$

Let  $\hat{\mathbf{\beta}}$  (along with  $\hat{\lambda}$ ) denote the pooled OLS estimator from this equation. By standard results on partitioned regression [for example, Davidson and MacKinnon (1993, Section 1.4)],  $\hat{\mathbf{\beta}}$  can be obtained by the following two-step procedure:

(i) Regress  $\mathbf{x}_{it}$  on  $\bar{\mathbf{x}}_i$  across all  $t$  and  $i$ , and save the  $1 \times K$  vectors of residuals, say  $\hat{\mathbf{g}}_{it}$ ,  $t = 1, \dots, T$ ;  $i = 1, \dots, N$ .

(ii) Regress  $y_{it}$  on  $\hat{\mathbf{g}}_{it}$  across all  $t$  and  $i$ . The OLS vector on  $\hat{\mathbf{g}}_{it}$  is  $\hat{\mathbf{\beta}}$ .

We want to show that  $\hat{\mathbf{\beta}}$  is the FE estimator. Given that the FE estimator can be obtained by pooled OLS of  $y_{it}$  on  $(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$ , it suffices to show that  $\hat{\mathbf{g}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$  for all  $t$  and  $i$ . But,

$$\hat{\mathbf{g}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i \left( \sum_{i=1}^N \sum_{t=1}^T \bar{\mathbf{x}}_i' \bar{\mathbf{x}}_i \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T \bar{\mathbf{x}}_i' \mathbf{x}_{it} \right)$$

and

$$\sum_{i=1}^N \sum_{t=1}^T \bar{\mathbf{x}}_i' \bar{\mathbf{x}}_{it} = \sum_{i=1}^N \bar{\mathbf{x}}_i' \sum_{t=1}^T \mathbf{x}_{it} = \sum_{i=1}^N T \bar{\mathbf{x}}_i' \bar{\mathbf{x}}_{it} = \sum_{i=1}^N \sum_{t=1}^T \bar{\mathbf{x}}_i' \bar{\mathbf{x}}_i.$$

It follows that  $\hat{\mathbf{g}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i \mathbf{I}_K = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$ . This completes the proof.

**11.8.** a. This is just a special case of Problem 8.8, where we now apply the results to the FD equation and account for the loss of the first time period. The rank condition is

$$\text{rank}\left(\sum_{t=2}^T \mathbf{E}(\mathbf{z}_{it}' \Delta \mathbf{x}_{it})\right) = K.$$

b. Again, Problem 8.8 provides the answer. Letting  $e_{it} = \Delta u_{it}, t \geq 2$ , two sufficient conditions are  $\text{Var}(e_{it} | \mathbf{z}_{it}) = \sigma_e^2, t = 2, \dots, T$  and  $\mathbf{E}(e_{it} | \mathbf{z}_{it}, e_{i,t-1}, \dots, \mathbf{z}_{i2}, e_{i2}) = 0, t = 2, \dots, T$ .

c. As in the case of pooled OLS after first differencing, this is only useful (and can only be implemented) when  $T \geq 3$ . First, estimate equation (11.100) by pooled 2SLS and obtain the residuals,  $e_{it}, t = 2, \dots, T, i = 1, \dots, N$ . Then, estimate the augmented equation,

$$\Delta y_{it} = \Delta \mathbf{x}_{it} \beta + \rho \hat{e}_{i,t-1} + \text{error}_{it}, t = 3, \dots, T$$

by pooled 2SLS, using IVs  $(\mathbf{z}_{it}, \hat{e}_{i,t-1})$ . If we strengthen the condition from part b to

$$\mathbf{E}(e_{it} | \mathbf{z}_{it}, \Delta \mathbf{x}_{i,t-1}, e_{i,t-1}, \dots, \mathbf{z}_{i2}, \Delta \mathbf{x}_{i2}, e_{i2}) = 0$$

then, under  $H_0$ , the usual  $t$  statistic on  $\hat{e}_{i,t-1}$  is distributed as asymptotically standard normal, provided we add a dynamic homoskedasticity assumption. See Problem 8.10 for verification in a general IV setting.

**11.9.** a. We can apply Problem 8.8.b because we are applying pooled 2SLS – this time to the time-demeaned equation. Therefore, the rank condition is

$$\text{rank}\left(\sum_{t=1}^T \mathbf{E}(\ddot{\mathbf{z}}_{it}' \ddot{\mathbf{x}}_{it})\right) = K.$$

The rank condition clearly fails if  $\mathbf{x}_{it}$  contains any time-constant explanatory variables (across all  $i$ , as usual). The condition  $\text{rank}\left(\sum_{t=1}^T E(\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{z}}_{it})\right) = L$  also should be assumed, and this rules out time-constant instruments (and perfectly collinear instruments). If the rank condition holds, we can always redefine  $\mathbf{z}_{it}$  so that  $\sum_{t=2}^T E(\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{z}}_{it})$  has full rank.

b. We can apply the results on GMM estimation in Chapter 8. In particular, in equation (8.25), take  $\mathbf{C} = E(\ddot{\mathbf{Z}}_i'\ddot{\mathbf{X}}_i)$ ,  $\mathbf{W} = [E(\ddot{\mathbf{Z}}_i'\ddot{\mathbf{Z}}_i)]^{-1}$ ,  $\mathbf{\Lambda} = E(\mathbf{Z}_i'\ddot{\mathbf{u}}_i\ddot{\mathbf{u}}_i'\ddot{\mathbf{Z}}_i)$ .

A key point is that  $\ddot{\mathbf{Z}}_i'\ddot{\mathbf{u}}_i = (\mathbf{Q}_T'\mathbf{Z}_i)'(\mathbf{Q}_T\mathbf{u}_i) = \mathbf{Z}_i'\mathbf{Q}_T\mathbf{u}_i = \ddot{\mathbf{Z}}_i'\mathbf{u}_i$ , where  $\mathbf{Q}_T$  is the  $T \times T$  time-demeaning matrix defined in Chapter 10. Under Assumption FEIV.3,  $E(\mathbf{u}_i\mathbf{u}_i'|\ddot{\mathbf{Z}}_i) = \sigma_u^2\mathbf{I}_T$  (by the usual iterated expectations argument), and so  $\mathbf{\Lambda} = \sigma_u^2 E(\ddot{\mathbf{Z}}_i'\ddot{\mathbf{Z}}_i)$ . If we plug these choices of  $\mathbf{C}$ ,  $\mathbf{W}$ , and  $\mathbf{\Lambda}$  into (8.29) and simplify, we obtain

$$\text{Avar}\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sigma_u^2 \{E(\ddot{\mathbf{X}}_i'\ddot{\mathbf{Z}}_i)[E(\ddot{\mathbf{Z}}_i'\ddot{\mathbf{Z}}_i)]^{-1}E(\ddot{\mathbf{Z}}_i'\ddot{\mathbf{X}}_i)\}^{-1}.$$

c. The argument is very similar to the case of the fixed effects estimator. First, we already showed in Chapter 10 that  $\sum_{t=1}^T E(\ddot{u}_{it}^2) = (T-1)\sigma_u^2$ . If  $\hat{\ddot{u}}_{it} = \ddot{y}_{it} - \ddot{\mathbf{x}}_{it}\hat{\boldsymbol{\beta}}$  are the pooled 2SLS residuals applied to the time-demeaned data, then  $[N(T-1)]^{-1} \sum_{i=1}^N \sum_{t=1}^T \hat{\ddot{u}}_{it}^2$  is a consistent estimator of  $\sigma_u^2$ . Typically,  $N(T-1)$  would be replaced by  $N(T-1) - K$  as a degrees of freedom adjustment.

d. From Problem 5.1 – which is purely algebraic, and so applies directly to pooled 2SLS, even with lots of dummy variables – the 2SLS estimates, including  $\hat{\boldsymbol{\beta}}$ , can be obtained as follows. First, run the regression  $\mathbf{x}_{it}$  on  $d1_i, \dots, dN_i, \mathbf{z}_{it}$  across all  $t$  and  $i$ , and obtain the residuals, say  $\hat{\mathbf{r}}_{it}$ . Second, obtain  $\hat{c}_1, \dots, \hat{c}_N, \hat{\boldsymbol{\beta}}$  from pooled regression  $y_{it}$  on  $d1_i, \dots, dN_i, \mathbf{x}_{it}, \hat{\mathbf{r}}_{it}$ . Now, by algebra of partial regression,  $\hat{\boldsymbol{\beta}}$  and the coefficient on  $\hat{\mathbf{r}}_{it}$ , say  $\hat{\boldsymbol{\delta}}$ , from this last regression can be obtained by first partialling out the dummy variables,  $d1_i, \dots, dN_i$ . As we

know from Chapter 10, this partialling out is equivalent to time demeaning all variables.

Therefore,  $\hat{\beta}$  and  $\hat{\delta}$  can be obtained from the pooled regression  $\ddot{y}_{it}$  on  $\ddot{\mathbf{x}}_{it}, \hat{\mathbf{r}}_{it}$ , where we use the fact that the time average of  $\hat{\mathbf{r}}_{it}$  for each  $i$  is identically zero.

Now consider the 2SLS estimator of  $\beta$  from

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\beta + \ddot{u}_{it} \quad (11.102)$$

using IVs  $\ddot{\mathbf{z}}_{it}$ . Again appealing to Problem 5.1, the pooled 2SLS estimator can be obtained from regressing  $\ddot{\mathbf{x}}_{it}$  on  $\ddot{\mathbf{z}}_{it}$  and saving the residuals, say  $\hat{\mathbf{s}}_{it}$ , and then running the OLS regression  $\ddot{y}_{it}$  on  $\ddot{\mathbf{x}}_{it}, \hat{\mathbf{s}}_{it}$ . By partial regression and the fact that regressing on  $d1_i, \dots, dN_i$  results in time demeaning,  $\hat{\mathbf{s}}_{it} = \hat{\mathbf{r}}_{it}$  for all  $i$  and  $t$ . This proves that the 2SLS estimates of  $\beta$  from (11.102) and

$$y_{it} = c_1 d1_i + c_2 d2_i + \dots + c_N dN_i + \mathbf{x}_{it}\beta + u_{it} \quad (11.103)$$

are identical.

e. By writing down the first order condition for the 2SLS estimates from (11.103) (with  $dn_i$  as their own instruments, and  $\hat{\mathbf{x}}_{it}$  as the IVs for  $\mathbf{x}_{it}$ ), it is easy to show that  $\hat{c}_i = \bar{y}_i - \bar{\mathbf{x}}_i\hat{\beta}$ , where  $\hat{\beta}$  is the FE2SLS estimator. Therefore, the 2SLS residuals from (11.103) are computed as

$$y_{it} - \hat{c}_i - \mathbf{x}_{it}\hat{\beta} = y_{it} - (\bar{y}_i - \bar{\mathbf{x}}_i\hat{\beta}) - \mathbf{x}_{it}\hat{\beta} = (y_{it} - \bar{y}_i) - (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\hat{\beta} = \ddot{y}_{it} - \ddot{\mathbf{x}}_{it}\hat{\beta},$$

which are exactly the 2SLS residuals from (11.102). Because the  $N$  dummy variables are explicitly included in (11.103), the degrees of freedom in estimating  $\sigma_u^2$  from part c are properly calculated.

The general, messy estimator in equation (8.31) should be used, where  $\mathbf{X}$  and  $\mathbf{Z}$  are replaced with  $\ddot{\mathbf{X}}$  and  $\ddot{\mathbf{Z}}$ , respectively,  $\hat{\mathbf{W}} = (\ddot{\mathbf{Z}}'\ddot{\mathbf{Z}}/N)^{-1}$ ,  $\hat{\mathbf{u}}_i = \ddot{\mathbf{y}}_i - \ddot{\mathbf{X}}_i\hat{\beta}$ , and

$$\hat{\Lambda} = N^{-1} \sum_{i=1}^N \ddot{\mathbf{Z}}_i' \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \ddot{\mathbf{Z}}_i.$$



**11.10.** Let  $\tilde{\mathbf{a}}_i$ ,  $i = 1, \dots, N$ , and  $\tilde{\boldsymbol{\beta}}$  be the OLS estimates from the pooled OLS regression (11.101). By partial regression,  $\tilde{\boldsymbol{\beta}}$  can be obtained by first regressing  $y_{it}$  on  $d1_i \mathbf{z}_{it}$ ,  $d2_i \mathbf{z}_{it}$ , ...,  $dN_i \mathbf{z}_{it}$  and obtaining the residuals,  $\check{y}_{it}$ , and likewise for  $\check{\mathbf{x}}_{it}$ . Then, we regress  $\check{y}_{it}$  on  $\check{\mathbf{x}}_{it}$ ,  $t = 1, \dots, T$ ;  $i = 1, \dots, N$ . But regressing on  $d1_i \mathbf{z}_{it}$ ,  $d2_i \mathbf{z}_{it}$ , ...,  $dN_i \mathbf{z}_{it}$  across all  $t$  and  $i$  is the same as regressing on  $\mathbf{z}_{it}$ ,  $t = 1, \dots, T$ , for each cross section observation,  $i$ . Therefore, we can write

$$\begin{aligned}\check{y}_{it} &= y_{it} - \mathbf{z}_{it}[(\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \mathbf{Z}_i' \mathbf{y}_i] \\ \check{\mathbf{y}}_i &= \mathbf{M}_i \mathbf{y}_i\end{aligned}$$

where  $\mathbf{M}_i = \mathbf{I}_T - \mathbf{Z}_i[(\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \mathbf{Z}_i']$ . A similar expression holds for  $\check{\mathbf{x}}_{it}$ . We have shown that regression (11.101) is identical to the pooled OLS regression  $\check{y}_{it}$  on  $\check{\mathbf{x}}_{it}$ ,  $t = 1, \dots, T$ ,  $i = 1, \dots, N$ . The residuals from the two regressions are exactly the same by the two-step projection result. The regression in (11.101) results in  $NT - NJ - K = N(T - J) - K$  degrees of freedom, which is exactly what we need in (11.76).

**11.11.** Differencing twice and using the resulting cross section is easily done in most statistical packages. Alternatively, Equivalently, use FE on the FD equation (which is the same as FD on the FD equation).I can use fixed effects on the first differences

The Stata output follows. The estimates from the random growth model are pretty bad: the estimates on the grant variables are of the “wrong” sign, and they are very imprecise.

The joint  $F$  test for the 53 different firm intercepts (when we treat the heterogeneity as estimable parameters) is significant at the 5% level( $p$ -value = .033), which does suggest a random growth model is appropriate. (But remember, this statistic is only valid under restrictive assumptions.) It is hard to know what to make of the poor estimates, but it does cast doubt on the standard unobserved effects model without a random growth term.

```
. xtreg clscrap d89 cgrant cgrant_1, fe
```

```
Fixed-effects (within) regression      Number of obs   =      108
Group variable: fcode                  Number of groups =       54

R-sq:  within  = 0.0577                Obs per group: min =
        between = 0.0476                    avg   =      2.
        overall = 0.0050                    max   =

corr(u_i, Xb) = -0.4011                F(3,51)         =      1.04
                                         Prob > F        =      0.3826
```

clscrap	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
d89	-.2377384	.1407362	-1.69	0.097	-.5202783	.0448014
cgrant	.1564748	.2632934	0.59	0.555	-.3721088	.6850584
cgrant_1	.6099015	.6343411	0.96	0.341	-.6635913	1.883394
_cons	-.2240491	.114748	-1.95	0.056	-.4544153	.0063171
sigma_u	.50956703					
sigma_e	.49757778					
rho	.51190251	(fraction of variance due to u_i)				
F test that all u_i=0:		F(53, 51) =	1.67		Prob > F = 0.0334	

**11.12. a.** Using only the changes from 1990 to 1993 and estimating the first-differenced equation by OLS gives:

```
. reg cmrdрте cexec cunem if d93
```

Source	SS	df	MS	Number of obs =	51	
				F( 2, 48) =	2.96	
Model	6.8879023	2	3.44395115	Prob > F =	0.0614	
Residual	55.8724857	48	1.16401012	R-squared =	0.1097	
				Adj R-squared =	0.0727	
Total	62.760388	50	1.25520776	Root MSE =	1.0789	
cmrdrtc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
cexec	-.1038396	.0434139	-2.39	0.021	-.1911292	-.01655
cunem	-.0665914	.1586859	-0.42	0.677	-.3856509	.252468
_cons	.4132665	.2093848	1.97	0.054	-.0077298	.8342628

The coefficient on *cexec* means that one more execution reduces the murder rate by about .10, and the effect is statistically significant.

b. If executions in the future respond to changes in the past murder rate, then *exec* may not be strictly exogenous. If executions more than three years ago have a partial effect on the

murder rate, this would also violate strict exogeneity because, effectively, we do not have enough lags. In principle, we could handle the latter problem by collecting more data and including more lags.

If we assume that only  $exec_{it}$  appears in the equation at time  $t$ , so that current and past executions are uncorrelated with  $u_{it}$ , then we can difference away  $c_i$  and apply IV:

$$\Delta mrdrte_{it} = \delta_0 + \beta_1 \Delta exec_{it} + \beta_2 \Delta unem_{it} + \Delta u_{it}.$$

A valid IV for  $\Delta exec_{it}$  is  $\Delta exec_{i,t-1}$  because, by assumption,  $exec_{i,t-1}$  and  $exec_{i,t-2}$  are both uncorrelated with  $u_{it}$  and  $u_{i,t-1}$ . This results in a cross section IV estimation.

c. To test the rank condition, we regress  $\Delta exec_{it}$  on 1,  $\Delta exec_{i,t-1}$ ,  $\Delta unem_{it}$  for 1993, and do a  $t$  test on  $\Delta exec_{i,t-1}$ :

```
. reg cexec cexec_1 cunem if d93
```

Source	SS	df	MS	Number of obs = 51		
Model	281.429488	2	140.714744	F( 2, 48) = 20.09		
Residual	336.217571	48	7.00453273	Prob > F = 0.0000		
Total	617.647059	50	12.3529412	R-squared = 0.4556		
				Adj R-squared = 0.4330		
				Root MSE = 2.6466		

cexec	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
cexec_1	-1.08241	.1707822	-6.34	0.000	-1.42579	-.7390289
cunem	.0400493	.3892505	0.10	0.918	-.7425912	.8226898
_cons	.3139609	.5116532	0.61	0.542	-.7147868	1.342709

Interestingly, there is a one-for-one negative relationship between the change in lagged executions and the change in current executions. Certainly the rank condition passes.

The IV estimates are below:

```
. reg cmrdrte cexec cunem (cexec_1 cunem) if d93
```

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs = 51		
Model	6.87925253	2	3.43962627	F( 2, 48) = 1.31		
				Prob > F = 0.2796		

Residual		55.8811355	48	1.16419032		R-squared	=	0.1096
Total		62.760388	50	1.25520776		Adj R-squared	=	0.0725
						Root MSE	=	1.079

cmrdrte		Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
cexec		-.1000972	.0643241	-1.56	0.126	-.2294293	.029235
cunem		-.0667262	.1587074	-0.42	0.676	-.3858289	.2523764
_cons		.410966	.2114237	1.94	0.058	-.0141298	.8360617

The point estimate on  $\Delta exec$  is essentially the same as the OLS estimate, but, of course, the IV standard error is larger. We can justify the POLS estimator on the FD equation (as the null of exogeneity of  $\Delta exec$  would not be rejected).

d. The following Stata command gives the results without Texas:

```
. reg cmrdrte cexec cunem if (d93==1 & state!= "TX")
```

Source		SS	df	MS		Number of obs =	50
Model		.755191109	2	.377595555		F( 2, 47) =	0.32
Residual		55.7000012	47	1.18510641		Prob > F	= 0.7287
Total		56.4551923	49	1.15214678		R-squared	= 0.0134
						Adj R-squared	= -0.0286
						Root MSE	= 1.0886

cmrdrte		Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
cexec		-.067471	.104913	-0.64	0.523	-.2785288	.1435868
cunem		-.0700316	.1603712	-0.44	0.664	-.3926569	.2525936
_cons		.4125226	.2112827	1.95	0.057	-.0125233	.8375686

Instrumental variables (2SLS) regression

Source		SS	df	MS		Number of obs =	50
Model		-1.65785462	2	-.828927308		F( 2, 47) =	0.11
Residual		58.1130469	47	1.23644781		Prob > F	= 0.8939
Total		56.4551923	49	1.15214678		R-squared	=
						Adj R-squared	=
						Root MSE	= 1.112

cmrdrte		Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
cexec		.082233	.804114	0.10	0.919	-1.535436	1.699902
cunem		-.0826635	.1770735	-0.47	0.643	-.4388895	.2735624
_cons		.3939505	.2373797	1.66	0.104	-.0835958	.8714968

The OLS estimate is smaller in magnitude and not statistically significant, while the IV

estimate actually changes sign (but, statistically, is not different from zero). Clearly, including Texas in the estimation has a big impact. It is easy to see why this is the case by listing the change in the murder rates and executions for Texas along with the averages for all states:

```
. list cmrdрте cexec if (d93==1 & state == "TX")
```

```

+-----+
|      cmrdрте      cexec |
+-----+
132. |    -2.200001         23 |
+-----+

```

```
. sum cmrdрте cexec if d93==1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
cmrdрте	51	.2862745	1.120361	-2.200001	3.099998
cexec	51	.6470588	3.514675	-3	23

Texas has the largest drop in the murder rate from 1990 to 1993, and also the largest increase in the number of executions. This does not necessarily mean Texas is an outlier, but it clearly is an influential observation. And it is clear why including Texas makes for a fairly strong deterrent effect.

**11.13. a.** The following Stata output estimates the reduced form for  $\Delta \log(pris)$  and tests joint significance of *final1* and *final2*, and also tests equality of the coefficients on *final1* and *final2*. The latter is actually not very interesting. Technically, because we do not reject, we could reduce our instrument to *final1* + *final2*, but we could always look ex post for restrictions on the parameters in a reduced form.

```
. use prison
```

```
. xtset state year
      panel variable:  state (strongly balanced)
      time variable:  year, 80 to 93
              delta:  1 unit
```

```
. reg gpris final1 final2 gpolpc gincpc cunem cblack cmetro cag0_14 cag15_17
      cag18_24 cag25_34 y81-y93
```

Source	SS	df	MS	Number of obs =	714
				F( 24, 689) =	5.15

Model	.481041472	24	.020043395	Prob > F	=	0.0000
Residual	2.68006631	689	.003889791	R-squared	=	0.1522
<hr/>						
Total	3.16110778	713	.004433531	Adj R-squared	=	0.1226
				Root MSE	=	.06237

gpris	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
final1	-.077488	.0259556	-2.99	0.003	-.1284496	-.0265265
final2	-.0529558	.0184078	-2.88	0.004	-.0890979	-.0168136
gpolpc	-.0286921	.0440058	-0.65	0.515	-.1150937	.0577094
gincpc	.2095521	.1313169	1.60	0.111	-.0482772	.4673815
cunem	.1616595	.3111688	0.52	0.604	-.4492935	.7726124
cblack	-.0044763	.0262118	-0.17	0.864	-.055941	.0469883
cmetro	-1.418389	.7860435	-1.80	0.072	-2.961717	.1249393
cag0_14	2.617307	1.582611	1.65	0.099	-.4900126	5.724627
cag15_17	-1.608738	3.755564	-0.43	0.669	-8.982461	5.764986
cag18_24	.9533678	1.731188	0.55	0.582	-2.445669	4.352405
cag25_34	-1.031684	1.763248	-0.59	0.559	-4.493667	2.4303
y81	.0124113	.013763	0.90	0.367	-.0146111	.0394337
y82	.0773503	.0156924	4.93	0.000	.0465396	.108161
y83	.0767785	.0153929	4.99	0.000	.0465559	.1070011
y84	.0289763	.0176504	1.64	0.101	-.0056787	.0636314
y85	.0279051	.0164176	1.70	0.090	-.0043295	.0601397
y86	.0541489	.0179305	3.02	0.003	.018944	.0893539
y87	.0312716	.0171317	1.83	0.068	-.002365	.0649082
y88	.019245	.0170725	1.13	0.260	-.0142754	.0527654
y89	.0184651	.0172867	1.07	0.286	-.0154759	.052406
y90	.0635926	.0165775	3.84	0.000	.0310442	.0961411
y91	.0263719	.0168913	1.56	0.119	-.0067927	.0595366
y92	.0190481	.0179372	1.06	0.289	-.0161701	.0542663
y93	.0134109	.0189757	0.71	0.480	-.0238461	.050668
_cons	.0272013	.0170478	1.60	0.111	-.0062705	.0606731

```
. test final1 final2
```

```
( 1) final1 = 0
```

```
( 2) final2 = 0
```

```

F( 2, 689) = 8.56
Prob > F = 0.0002
```

```
. test final1 = final2
```

```
( 1) final1 - final2 = 0
```

```

F( 1, 689) = 0.60
Prob > F = 0.4401
```

Jointly, *final1* and *final2* are pretty significant. Next, test for serial correlation in

$a_{it} \equiv \Delta v_{it}$ :

```
. predict ahat, resid
```

```
. gen ahat_1 = 1.ahat
```

(51 missing values generated)

```
. reg ahat ahat_1
```

Source	SS	df	MS	Number of obs =	663
Model	.051681199	1	.051681199	F( 1, 661) =	14.33
Residual	2.38322468	661	.003605484	Prob > F =	0.0002
				R-squared =	0.0212
				Adj R-squared =	0.0197
Total	2.43490588	662	.003678106	Root MSE =	.06005

ahat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval
ahat_1	.1426247	.0376713	3.79	0.000	.0686549 .2165945
_cons	4.24e-11	.002332	0.00	1.000	-.004579 .004579

There is strong evidence of positive serial correlation, although the estimated size of the AR(1) coefficient, .143, is not especially large. Still, a fully robust variance matrix should be used for the joint significance test of *final1* and *final2*. These two IVs are much more significant when the robust variance matrix is used:

```
. qui reg gpris final1 final2 gpolpc gincpc cunem cblack cmetro cag0_14
      cag15_17 cag18_24 cag25_34 y81-y93, cluster(state)

. test final1 final2

( 1) final1 = 0
( 2) final2 = 0

      F( 2, 50) = 18.82
      Prob > F = 0.000
```

b. First, we do pooled 2SLS to obtain the 2SLS residuals,  $\hat{e}_{it}$ . Then we add the lagged residual to the equation, and use it as its own IV:

```
. ivreg gcriv gpolpc gincpc cunem cblack cmetro cag0_14 cag15_17 cag18_24
      cag25_34 y81-y93 (gpris = final1 final2)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs =	714
Model	-.696961613	23	-.030302679	F( 23, 690) =	6.08
Residual	6.28846843	690	.009113722	Prob > F =	0.0000
				R-squared =	
				Adj R-squared =	
Total	5.59150682	713	.007842226	Root MSE =	.09547

gcriv	Coef.	Std. Err.	t	P> t	[95% Conf. Interval
-------	-------	-----------	---	------	---------------------

gpris	-1.031956	.3699628	-2.79	0.005	-1.758344	-.3055684
gpolpc	.035315	.0674989	0.52	0.601	-.0972128	.1678428
gincpc	.9101992	.2143266	4.25	0.000	.4893885	1.33101
cunem	.5236958	.4785632	1.09	0.274	-.415919	1.46331
cblack	-.0158476	.0401044	-0.40	0.693	-.0945889	.0628937
cmetro	-.591517	1.298252	-0.46	0.649	-3.140516	1.957482
cag0_14	3.379384	2.634893	1.28	0.200	-1.793985	8.552753
cag15_17	3.549945	5.766302	0.62	0.538	-7.771659	14.87155
cag18_24	3.358348	2.680839	1.25	0.211	-1.905233	8.621929
cag25_34	2.319993	2.706345	0.86	0.392	-2.993667	7.633652
y81	-.0560732	.0217346	-2.58	0.010	-.0987471	-.0133992
y82	.0284616	.0384773	0.74	0.460	-.047085	.1040082
y83	.024703	.0373965	0.66	0.509	-.0487216	.0981276
y84	.0128703	.0293337	0.44	0.661	-.0447236	.0704643
y85	.0354026	.0275023	1.29	0.198	-.0185956	.0894008
y86	.0921857	.0343884	2.68	0.008	.0246672	.1597042
y87	.004771	.0290145	0.16	0.869	-.0521964	.0617383
y88	.0532706	.0273221	1.95	0.052	-.0003738	.106915
y89	.0430862	.0275204	1.57	0.118	-.0109476	.0971201
y90	.1442652	.0354625	4.07	0.000	.0746379	.2138925
y91	.0618481	.0276502	2.24	0.026	.0075595	.1161366
y92	.0266574	.0285333	0.93	0.350	-.0293651	.0826799
y93	.0222739	.0296099	0.75	0.452	-.0358624	.0804103
_cons	.0148377	.0275197	0.54	0.590	-.0391948	.0688702

```

Instrumented:  gpris
Instruments:   gpolpc gincpc cunem cblack cmetro cag0_14 cag15_17 cag18_24
               cag25_34 y81 y82 y83 y84 y85 y86 y87 y88 y89 y90 y91 y92 y93
               final1 final2

```

```
. predict ehat, resid
```

```
. gen ehat_1 = 1.ehat
(51 missing values generated)
```

```
. ivreg gcriv gpolpc gincpc cunem cblack cmetro cag0_14 cag15_17 cag18_24
      cag25_34 y81-y93 ehat_1 (gpris = final1 final2)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs =	663
Model	-.815873465	23	-.035472759	F( 23, 639) =	5.14
Residual	5.90425699	639	.009239839	Prob > F =	0.0000
Total	5.08838353	662	.00768638	R-squared =	
				Adj R-squared =	
				Root MSE =	.09612

gcriv	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
gpris	-1.084446	.4071905	-2.66	0.008	-1.884039	-.2848525
gpolpc	.0179121	.0719595	0.25	0.804	-.1233935	.1592176
gincpc	.7492611	.2421405	3.09	0.002	.2737738	1.224748
cunem	.1979701	.515973	0.38	0.701	-.8152375	1.211178
cblack	-.0102865	.0424589	-0.24	0.809	-.0936622	.0730893
cmetro	-.5272326	1.357715	-0.39	0.698	-3.193354	2.138889
cag0_14	3.284496	3.045539	1.08	0.281	-2.695979	9.26497



cag15_17	.066451	6.105497	0.01	0.991	-11.92281	12.05571
cag18_24	3.094998	2.830038	1.09	0.275	-2.462301	8.652297
cag25_34	2.716353	2.799581	0.97	0.332	-2.781137	8.213843
y81	-.0782703	.0350721	-2.23	0.026	-.1471409	-.0093998
y82	.0090276	.0225246	0.40	0.689	-.0352036	.0532588
y83	(dropped)					
y84	-.0113602	.0314408	-0.36	0.718	-.0731	.0503796
y85	.015744	.0309473	0.51	0.611	-.0450267	.0765148
y86	.0752485	.027649	2.72	0.007	.0209547	.1295424
y87	-.0205808	.0282106	-0.73	0.466	-.0759774	.0348159
y88	.0265964	.0315542	0.84	0.400	-.0353661	.0885589
y89	.0182293	.0327158	0.56	0.578	-.0460142	.0824727
y90	.1275351	.0235386	5.42	0.000	.0813126	.1737575
y91	.0435859	.0315328	1.38	0.167	-.0183346	.1055064
y92	.0121958	.0354112	0.34	0.731	-.0573406	.0817321
y93	.0016107	.0365807	0.04	0.965	-.0702221	.0734435
ehat_1	.0763754	.0456451	1.67	0.095	-.0132571	.166008
_cons	.0441747	.0477902	0.92	0.356	-.0496701	.1380195

---

Instrumented: gpris  
Instruments: gpolpc gincpc cunem cblack cmetro cag0\_14 cag15\_17 cag18\_24  
cag25\_34 y81 y82 y83 y84 y85 y86 y87 y88 y89 y90 y91 y92 y93  
ehat\_1 final1 final2

---

There is only marginal evidence of positive serial correlation, and it is practically small, anyway ( $\hat{\rho} = .076$ ).

c. Adding a state effect to the change (FD) equation changes very little. In this example, there seems to be little need for a random growth model. The estimated prison effect becomes a little smaller in magnitude,  $-.959$ . Here is the Stata output:

```
. xtivreg gcriv gpolpc gincpc cunem cblack cmetro cag0_14 cag15_17 cag18_24
      cag25_34 y81-y93 (gpris = final1 final2), fe
```

Fixed-effects (within) IV regression	Number of obs	=	714
Group variable: state	Number of groups	=	51
R-sq: within =	Obs per group: min =		14
between = 0.0001	avg =		14.
overall = 0.1298	max =		14
	Wald chi2(23)	=	179.24
corr(u_i, Xb) = -0.2529	Prob > chi2	=	0.0000

gcriv	Coef.	Std. Err.	z	P> z	[95% Conf. Interval
gpris	-.9592287	.3950366	-2.43	0.015	-1.733486 - .1849713
gpolpc	.04445	.0664696	0.67	0.504	-.0858281 .1747281
gincpc	1.027161	.2157944	4.76	0.000	.6042122 1.450111
cunem	.6560942	.4698359	1.40	0.163	-.2647672 1.576956
cblack	.0706601	.1496426	0.47	0.637	-.2226339 .3639542

cmetro	3.229287	4.683812	0.69	0.491	-5.950815	12.40939
cag0_14	1.14119	2.749679	0.42	0.678	-4.248082	6.530462
cag15_17	1.402606	6.330461	0.22	0.825	-11.00487	13.81008
cag18_24	1.169114	2.866042	0.41	0.683	-4.448225	6.786453
cag25_34	-2.089449	3.383237	-0.62	0.537	-8.720471	4.541574
y81	-.0590819	.0230252	-2.57	0.010	-.1042104	-.0139534
y82	.0033116	.0388056	0.09	0.932	-.0727459	.0793691
y83	.0080099	.0378644	0.21	0.832	-.066203	.0822228
y84	-.0019285	.0293861	-0.07	0.948	-.0595243	.0556672
y85	.0220412	.0276807	0.80	0.426	-.032212	.0762945
y86	.075621	.0338898	2.23	0.026	.0091981	.1420438
y87	-.0124835	.0294198	-0.42	0.671	-.0701453	.0451783
y88	.0329977	.0286125	1.15	0.249	-.0230817	.0890771
y89	.018718	.0292666	0.64	0.522	-.0386434	.0760794
y90	.1157811	.0354143	3.27	0.001	.0463703	.1851919
y91	.0378784	.0290414	1.30	0.192	-.0190417	.0947984
y92	-.0006633	.0305014	-0.02	0.983	-.0604449	.0591184
y93	-.0007561	.0317733	-0.02	0.981	-.0630306	.0615184
_cons	.0014574	.0296182	0.05	0.961	-.0565932	.0595079

---

sigma_u	.03039696	
sigma_e	.0924926	
rho	.09747751	(fraction of variance due to u_i)

---

F test that all u\_i=0:      F(50,640) =      0.69      Prob > F      = 0.9459

---

Instrumented:      gpris  
Instruments:      gpolpc gincpc cunem cblack cmetro cag0\_14 cag15\_17 cag18\_24  
                      cag25\_34 y81 y82 y83 y84 y85 y86 y87 y88 y89  
                      y90 y91 y92 y93 final1 final2

---

d. When we use the property crime rate, the estimated elasticity with respect to prison size is substantially smaller, but still negative and marginally significant:

```
. ivreg gcrip gpolpc gincpc cunem cblack cmetro cag0_14 cag15_17 cag18_24
      cag25_34 y81-y93 (gpris = final1 final2)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS		Number of obs =	714
Model	1.07170564	23	.046595897		F( 23, 690) =	22.83
Residual	1.5490539	690	.002245006		Prob > F =	0.0000
Total	2.62075954	713	.00367568		R-squared =	0.4089
					Adj R-squared =	0.3892
					Root MSE =	.04738

gcrip	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
gpris	-.3285567	.1836195	-1.79	0.074	-.6890768	.0319633
gpplpc	.014567	.033501	0.43	0.664	-.051209	.0803431
gincpc	.0560822	.1063744	0.53	0.598	-.1527741	.2649385
cunem	.8583588	.2375199	3.61	0.000	.3920102	1.324707
cblack	-.0507462	.0199046	-2.55	0.011	-.089827	-.0116654
cmetro	.0404892	.6443472	0.06	0.950	-1.224627	1.305606

cag0_14	1.890526	1.307747	1.45	0.149	-.6771151	4.458167
cag15_17	5.699448	2.861925	1.99	0.047	.0803221	11.31857
cag18_24	1.712283	1.330551	1.29	0.199	-.9001312	4.324698
cag25_34	2.027833	1.34321	1.51	0.132	-.6094366	4.665102
y81	-.0771684	.0107873	-7.15	0.000	-.0983483	-.0559886
y82	-.0980884	.019097	-5.14	0.000	-.1355836	-.0605932
y83	-.1093989	.0185606	-5.89	0.000	-.1458409	-.0729569
y84	-.0810119	.0145589	-5.56	0.000	-.1095968	-.0524269
y85	-.031369	.0136499	-2.30	0.022	-.0581693	-.0045687
y86	-.0169451	.0170676	-0.99	0.321	-.0504558	.0165656
y87	-.0310865	.0144005	-2.16	0.031	-.0593605	-.0028125
y88	-.0437643	.0135605	-3.23	0.001	-.0703891	-.0171396
y89	-.0359254	.0136589	-2.63	0.009	-.0627434	-.0091074
y90	-.0298029	.0176007	-1.69	0.091	-.0643603	.0047544
y91	-.0505269	.0137233	-3.68	0.000	-.0774713	-.0235824
y92	-.1024579	.0141616	-7.23	0.000	-.1302629	-.0746529
y93	-.0867254	.014696	-5.90	0.000	-.1155796	-.0578712
_cons	.0857682	.0136586	6.28	0.000	.0589509	.1125856

---

Instrumented: gpris  
Instruments: gpolpc gincpc cunem cblack cmetro cag0\_14 cag15\_17 cag18\_24  
cag25\_34 y81 y82 y83 y84 y85 y86 y87 y88 y89 y90 y91 y92 y93  
final1 final2

---

The test for serial correlation yields a coefficient on  $\hat{e}_{i,t-1}$  of  $-.024$  ( $t = -.52$ ), and so we conclude that serial correlation is not an issue.

**11.14. a.** The fixed effects estimate of the first-difference equations are given below. We have included year dummies without differencing them, since we are not interested in the time effects in the original model:

```
. use ezunem

. xtset city year
      panel variable:  city (strongly balanced)
      time variable:  year, 1980 to 1988
      delta:  1 unit

. gen cezt = d.ezt
(22 missing values generated)

. xtreg guclms cez cezt d82-d88, fe
```

Fixed-effects (within) regression	Number of obs	=	176
Group variable: city	Number of groups	=	22
R-sq: within = 0.6406	Obs per group: min =		
between = 0.0094	avg =		8.
overall = 0.6205	max =		
	F(9,145)	=	28.71
corr(u_i, Xb) = -0.0546	Prob > F	=	0.0000

guclms	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
cez	.1937324	.3448663	0.56	0.575	-.4878818	.8753467
cezt	-.0783638	.0679161	-1.15	0.250	-.2125972	.0558697
d82	.7787595	.0675022	11.54	0.000	.6453442	.9121748
d83	-.0331192	.0675022	-0.49	0.624	-.1665345	.1002961
d84	-.0127177	.0713773	-0.18	0.859	-.153792	.1283566
d85	.3616479	.0762138	4.75	0.000	.2110144	.5122814
d86	.3277739	.0742264	4.42	0.000	.1810684	.4744793
d87	.089568	.0742264	1.21	0.230	-.0571375	.2362735
d88	.0185673	.0742264	0.25	0.803	-.1281381	.1652728
_cons	-.3216319	.0477312	-6.74	0.000	-.4159708	-.2272931
sigma_u	.05880562					
sigma_e	.22387933					
rho	.06454083	(fraction of variance due to u_i)				

F test that all u\_i=0: F(21, 145) = 0.49 Prob > F = 0.9712

. test cez cezt

( 1) cez = 0  
( 2) cezt = 0

F( 2, 145) = 3.22  
Prob > F = 0.0428

The coefficient  $\hat{\delta}_2 = -.078$  gives the difference in annual growth rate due to *EZ* designation. It is not significant at the usual 5% level. Note that this formulation does not give the coefficient  $\hat{\delta}_1$  a simple interpretation because zone designation happened either at  $t = 5$  (if in 1984) or  $t = 6$  (if in 1985). A better formulation centers the linear trend at the time of designation before constructing the interactions:

```
. egen nyrsez =sum(ez), by(city)

. gen ezt0 = 0 if ~ez
(46 missing values generated)

. replace ezt0 = ez*(t-5) if nyrsez == 5
(30 real changes made)

. replace ezt0 = ez*(t-6) if nyrsez == 4
(16 real changes made)

. gen cezt0 = ezt0 - ezt0[_n-1] if year > 1980
(22 missing values generated)

. xtreg guclms cez cezt0 d82-d88, fe
```

Fixed-effects (within) regression                      Number of obs                      =                      176

Group variable: city	Number of groups	=	22
R-sq: within = 0.6406	Obs per group: min	=	
between = 0.0025	avg	=	8.
overall = 0.6185	max	=	
	F(9,145)	=	28.72
corr(u_i, Xb) = -0.0630	Prob > F	=	0.0000

guclms	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
cez	-.2341545	.0924022	-2.53	0.012	-.4167837	-.0515252
cezt0	-.082805	.0715745	-1.16	0.249	-.224269	.058659
d82	.7787595	.0675005	11.54	0.000	.6453475	.9121716
d83	-.0331192	.0675005	-0.49	0.624	-.1665312	.1002928
d84	-.0028809	.0720513	-0.04	0.968	-.1452874	.1395256
d85	.355169	.0740177	4.80	0.000	.208876	.5014621
d86	.3297926	.0749318	4.40	0.000	.181693	.4778922
d87	.0915867	.0749318	1.22	0.224	-.0565129	.2396864
d88	.0205861	.0749318	0.27	0.784	-.1275135	.1686857
_cons	-.3216319	.0477301	-6.74	0.000	-.4159685	-.2272953
sigma_u	.06091433					
sigma_e	.22387389					
rho	.06893091	(fraction of variance due to u_i)				
F test that all u_i=0:		F(21, 145) =	0.50		Prob > F = 0.9681	

Now the coefficient on *cez* is the estimated effect of the *EZ* in the first year of designation, and that gets added to  $-.083 \cdot (\text{years since initial designation})$ . This is easier to read.

b. Setting  $\delta_1 = 0$  gives a within  $R$ -squared of about .640, compared with that of the original model in Example 11.4 of about .637. The difference is minor, and we would probably go with the simpler, basic model in Example 11.4. With more years of data, the trend effect in part a might become significant.

c. Because the general model contains  $c_i + g_it$ , we cannot distinguish the effects of a time-constant variable,  $w_i$ , or its interaction with a linear time trend – at least if we stay in a fixed effects framework. If we assume  $c_i$  and  $g_i$  are uncorrelated with  $ez_{it}$  we could include  $w_i$  and  $w_it$ .

d. Yes. Provided  $\{e_{it} : t = 1, \dots, T\}$  has the kind of variation that it does in this data set,  $w_i e_{it}$  is linearly independent from other covariates included in the model. Therefore, we can

estimate  $\eta$ . If we add  $h_i e z_{it}$  to the model, where  $h_i$  is additional unobserved heterogeneity, then  $\eta$  would not be identified (again, allowing  $h_i$  to be correlated with  $e z_{it}$ ).

**11.15. a.** We would have to assume that  $grant_{it}$  is uncorrelated with the idiosyncratic errors,  $u_{is}$ , for all  $t$  and  $s$ . One way to think of this assumption is that while grant designation may depend on firm heterogeneity  $c_i$ , it is not related to idiosyncratic fluctuations in any time period. Further, one must assume the grants have an effect on scrap rates only through their effects on job training – the standard assumption for an instrument.

**b.** The following simple regression shows that  $\Delta hrsemp_{it}$  and  $\Delta grant_{it}$  are highly positively correlated, as expected:

```
. reg chrsemp cgrant if d88
```

Source	SS	df	MS	Number of obs = 125		
Model	18117.5987	1	18117.5987	F( 1, 123) = 79.37		
Residual	28077.3319	123	228.270991	Prob > F = 0.0000		
Total	46194.9306	124	372.539763	R-squared = 0.3922		
				Adj R-squared = 0.3873		
				Root MSE = 15.109		

chrsemp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
cgrant	27.87793	3.129216	8.91	0.000	21.68384	34.07202
_cons	.5093234	1.558337	0.33	0.744	-2.57531	3.593956

Unfortunately, this is on a bigger sample than we can use to estimate the scrap rate equation, because the scrap rate is missing for so many firms. Restricted to that sample, we get:

```
. reg chrsemp cgrant if d88 & clscrap ~= .
```

Source	SS	df	MS	Number of obs = 45		
Model	6316.65458	1	6316.65458	F( 1, 43) = 22.23		
Residual	12217.3517	43	284.124457	Prob > F = 0.0000		
Total	18534.0062	44	421.227414	R-squared = 0.3408		
				Adj R-squared = 0.3255		
				Root MSE = 16.856		

chrsemp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
cgrant	24.43691	5.182712	4.72	0.000	13.98498	34.88885
_cons	1.580598	3.185483	0.50	0.622	-4.84354	8.004737

-----

So there is still a pretty strong relationship, but we will be using IV on a small sample ( $N = 45$ ).

c. The IV estimate is:

```
. ivreg clscrap (chrsemp = cgrant) if d88
```

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs = 45		
Model	.274951237	1	.274951237	F( 1, 43)	=	3.20
Residual	17.0148885	43	.395695081	Prob > F	=	0.0808
				R-squared	=	0.0159
				Adj R-squared	=	-0.0070
Total	17.2898397	44	.392950903	Root MSE	=	.62904

clscrap	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
chrsemp	-.0141532	.0079147	-1.79	0.081	-.0301148	.0018084
_cons	-.0326684	.1269512	-0.26	0.798	-.2886898	.223353

Instrumented: chrsemp  
Instruments: cgrant

-----

The estimate says that 10 more hours training per employee would lower the average scrap rate by about 14.2 percent, which is a large economic effect. It is marginally statistically significant (assuming we can trust the asymptotic distribution theory for IV with 45 observations).

d. The OLS estimates is only about  $-.0076$  – about half of the IV estimate – with  $t = -1.68$ .

```
. reg clscrap chrsemp if d88
```

Source	SS	df	MS	Number of obs = 45		
Model	1.07071245	1	1.07071245	F( 1, 43)	=	2.84
Residual	16.2191273	43	.377189007	Prob > F	=	0.0993
				R-squared	=	0.0619
				Adj R-squared	=	0.0401
Total	17.2898397	44	.392950903	Root MSE	=	.61416

clscrap	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
chrsemp	-.0076007	.0045112	-1.68	0.099	-.0166984	.0014971

_cons		-.1035161	.103736	-1.00	0.324	-.3127197	.1056875
-------	--	-----------	---------	-------	-------	-----------	----------

---

e. Any effect pretty much disappears using two years of differences (even though you can verify the rank condition easily holds):

```
. ivreg clscrap d89 (chrsemp = cgrant)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs =	91
Model	.538688387	2	.269344194	F( 2, 88) =	0.90
Residual	33.2077492	88	.377360787	Prob > F =	0.4087
				R-squared =	0.0160
				Adj R-squared =	-0.0064
Total	33.7464376	90	.374960418	Root MSE =	.6143

clscrap	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
chrsemp	-.0028567	.0030577	-0.93	0.353	-.0089332	.0032198
d89	-.1387379	.1296916	-1.07	0.288	-.3964728	.1189969
_cons	-.1548094	.0973592	-1.59	0.115	-.3482902	.0386715

Instrumented: chrsemp  
Instruments: d89 cgrant

---

**11.16.** a. Just use fixed effects, or first differencing. Of course  $w_i$  gets eliminated by either transformation.

b. Take the expectation of the structural equation conditional on  $(w_i, \mathbf{x}_i, r_i)$  :

$$\begin{aligned} E(y_{it}|w_i, \mathbf{x}_i, r_i) &= \gamma w_i + \mathbf{x}_{it}\boldsymbol{\beta} + E(c_i|w_i, \mathbf{x}_i, r_i) + E(u_{it}|w_i, \mathbf{x}_i, r_i) \\ &= \gamma w_i + \mathbf{x}_{it}\boldsymbol{\beta} + \delta_0 + \delta_i r_i + \bar{\mathbf{x}}_i \boldsymbol{\delta}_2. \end{aligned}$$

c. Provided a standard rank condition holds for the explanatory variables,  $\gamma$  is identified because it appears in a conditional expectation containing observable variables:  $E(y_{it}|w_i, \mathbf{x}_i, r_i)$ .

The pooled OLS estimation

$$y_{it} \text{ on } 1, w_i, \mathbf{x}_{it}, r_i, \bar{\mathbf{x}}_i, t = 1, \dots, T; i = 1, \dots, N$$

consistently estimates all parameters.

d. Following the hint, we can write



$$y_{it} = \delta_0 + \gamma w_i + \mathbf{x}_{it}\boldsymbol{\beta} + \delta_i r_i + \bar{\mathbf{x}}_i \delta_2 + a_i + u_{it}, t = 1, \dots, T, \quad (11.104)$$

where  $a_i = c_i - E(c_i|w_i, \mathbf{x}_i, r_i)$ . Under the assumptions given, the composite error,  $v_{it} \equiv a_i + u_{it}$ , is easily shown to have variance-covariance matrix that has the random effect form. In particular,  $\text{Var}(v_{it}|w_i, \mathbf{x}_i, r_i) = \sigma_a^2 + \sigma_u^2$  and  $\text{Cov}(v_{it}, v_{is}|w_i, \mathbf{x}_i, r_i) = \sigma_a^2$ . [The arguments for obtaining these expressions should be familiar. For example, since  $a_i$  is a function of  $c_i, \mathbf{x}_i$ , and  $r_i$ , we can replace  $c_i$  with  $a_i$  in all of the assumptions concerning the first and second moments of  $\{u_{it} : t = 1, \dots, T\}$ . Therefore,

$$E(a_i u_{it}|w_i, \mathbf{x}_i, a_i, r_i) = a_i E(u_{it}|w_i, \mathbf{x}_i, a_i, r_i) = 0$$

and so, by iterated expectations,

$$\text{Cov}(a_i, u_{it}|w_i, \mathbf{x}_i, r_i) = E(a_i u_{it}|w_i, \mathbf{x}_i, r_i) = 0.]$$

We conclude that  $\text{Var}(\mathbf{v}_i|w_i, \mathbf{x}_i, r_i) = \text{Var}(\mathbf{v}_i)$  has the random effects form, and so we should just apply the usual random effects estimator to (11.104). This is asymptotically more efficient than the pooled OLS estimator.

**11.17.** To obtain (11.81), we used (11.80) and the representation

$\sqrt{N}(\hat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta}) = \mathbf{A}^{-1} \left( N^{-1/2} \sum_{i=1}^N \ddot{\mathbf{X}}_i' \mathbf{u}_i \right) + o_p(1)$ . Simple algebra and standard properties of  $O_p(1)$  and  $o_p(1)$  give

$$\begin{aligned} \sqrt{N}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) &= N^{-1/2} \sum_{i=1}^N [(\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \mathbf{Z}_i' (\mathbf{y}_i - \mathbf{x}_i \boldsymbol{\beta}) - \boldsymbol{\alpha}] - \left( N^{-1} \sum_{i=1}^N [(\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \mathbf{Z}_i' \mathbf{X}_i] \right) \sqrt{N}(\hat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta}) \\ &= N^{-1/2} \sum_{i=1}^N (\mathbf{s}_i - \boldsymbol{\alpha}) - \mathbf{C} \mathbf{A}^{-1} N^{-1/2} \sum_{i=1}^N \ddot{\mathbf{X}}_i' \mathbf{u}_i + o_p(1) \end{aligned}$$

where  $\mathbf{C} \equiv E[(\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \mathbf{Z}_i' \mathbf{X}_i]$  and  $\mathbf{s}_i \equiv (\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \mathbf{Z}_i' (\mathbf{y}_i - \mathbf{x}_i \boldsymbol{\beta})$ . By definition,  $E(\mathbf{s}_i) = \boldsymbol{\alpha}$ . By combining terms in the sum we have

$$\sqrt{N}(\hat{\alpha} - \alpha) = N^{-1/2} \sum_{i=1}^N [(s_i - \alpha) - \mathbf{C}\mathbf{A}^{-1}\ddot{\mathbf{X}}_i'\mathbf{u}_i] + o_p(1),$$

which implies by the central limit theorem and the asymptotic equivalence lemma that

$\sqrt{N}(\hat{\alpha} - \alpha)$  is asymptotically normal with zero mean and variance  $E(\mathbf{r}_i\mathbf{r}_i')$ , where

$\mathbf{r}_i \equiv (s_i - \alpha) - \mathbf{C}\mathbf{A}^{-1}\ddot{\mathbf{X}}_i'\mathbf{u}_i$ . If we replace  $\alpha$ ,  $\mathbf{C}$ ,  $\mathbf{A}$ , and  $\beta$  with their consistent estimators, we get exactly (11.81) because the  $\hat{\mathbf{u}}_i$  are the  $T \times 1$  FE residuals.

**11.18.** a. Using equation (8.47) we have

$$\begin{aligned} \hat{\beta}_{REIV} = & \left[ \left( \sum_{i=1}^N \mathbf{X}_i' \hat{\Omega}^{-1} \mathbf{Z}_i \right) \left( \sum_{i=1}^N \mathbf{Z}_i' \hat{\Omega}^{-1} \mathbf{Z}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{Z}_i' \hat{\Omega}^{-1} \mathbf{X}_i \right) \right]^{-1} \\ & \cdot \left( \sum_{i=1}^N \mathbf{X}_i' \hat{\Omega}^{-1} \mathbf{Z}_i \right) \left( \sum_{i=1}^N \mathbf{Z}_i' \hat{\Omega}^{-1} \mathbf{Z}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{Z}_i' \hat{\Omega}^{-1} \mathbf{y}_i \right) \end{aligned}$$

where  $\hat{\Omega}$  has the RE form (and is probably estimated from the pooled 2SLS residuals).

By arguments very similar to that for FGLS, we can show

$$\sqrt{N}(\hat{\beta}_{REIV} - \beta) = \mathbf{A}^{-1} \mathbf{C}' \mathbf{D}^{-1} \left( N^{-1/2} \sum_{i=1}^N \mathbf{Z}_i' \mathbf{\Lambda}^{-1} \mathbf{u}_i \right) + o_p(1)$$

where

$$\begin{aligned} \mathbf{\Lambda} &= \text{plim}(\hat{\Omega}) \\ \mathbf{C} &= E(\mathbf{Z}_i' \mathbf{\Lambda}^{-1} \mathbf{X}_i) \\ \mathbf{D} &= E(\mathbf{Z}_i' \mathbf{\Lambda}^{-1} \mathbf{Z}_i) \\ \mathbf{A} &= \mathbf{C}' \mathbf{D}^{-1} \mathbf{C} \end{aligned}$$

Note that this formulation recognizes that  $\hat{\Omega}$  is not generally consistent for  $E(\mathbf{v}_i\mathbf{v}_i')$ . It follows that

$$\sqrt{N}(\hat{\beta}_{REIV} - \beta) \xrightarrow{d} \text{Normal}(\mathbf{0}, \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1})$$

where

$$\mathbf{B} = \mathbf{C}'\mathbf{D}^{-1}\mathbf{E}(\mathbf{Z}_i'\mathbf{\Lambda}^{-1}\mathbf{u}_i\mathbf{u}_i'\mathbf{\Lambda}^{-1}\mathbf{Z}_i)\mathbf{D}^{-1}\mathbf{C}$$

b. Consistent estimators of  $\mathbf{A}$  and  $\mathbf{B}$  are

$$\begin{aligned}\hat{\mathbf{A}} &= \hat{\mathbf{C}}'\hat{\mathbf{D}}^{-1}\hat{\mathbf{C}} \\ \hat{\mathbf{B}} &= \hat{\mathbf{C}}'\hat{\mathbf{D}}^{-1}\left(N^{-1}\sum_{i=1}^N\mathbf{Z}_i'\hat{\mathbf{\Omega}}^{-1}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\hat{\mathbf{\Omega}}^{-1}\mathbf{Z}_i\right)\hat{\mathbf{D}}^{-1}\hat{\mathbf{C}}\end{aligned}$$

where

$$\begin{aligned}\hat{\mathbf{C}} &= N^{-1}\sum_{i=1}^N\mathbf{Z}_i'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}_i \\ \hat{\mathbf{D}} &= N^{-1}\sum_{i=1}^N\mathbf{Z}_i'\hat{\mathbf{\Omega}}^{-1}\mathbf{Z}_i \\ \hat{\mathbf{u}}_i &= \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_{REIV}\end{aligned}$$

**11.19.** a. Below is the Stata output. The *concen* variable is positive and statistically significant using both RE and FE estimation of the reduced form, and using fully robust (that is, to any serial correlation and heteroskedasticity) standard errors. The coefficient is somewhat larger for RE compared with FE, and its standard error is somewhat smaller for RE. We conclude that *concen* is suitably partially correlated with *lfare* in order to apply REIV and FEIV.

```
. xtreg lfare concen ldist ldistsq y98 y99 y00, re cluster(id)
```

Random-effects GLS regression	Number of obs	=	4596
Group variable: id	Number of groups	=	1149
R-sq: within = 0.1348	Obs per group: min =		
between = 0.4176	avg =		4.
overall = 0.4030	max =		

Random effects u_i ~Gaussian	Wald chi2(7)	=	386792.48
corr(u_i, X) = 0 (assumed)	Prob > chi2	=	0.0000

(Std. Err. adjusted for 1149 clusters in id)

---

	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
lfare					

concen	.2089935	.0422459	4.95	0.000	.126193	.2917939
ldist	-.8520921	.2720902	-3.13	0.002	-1.385379	-.3188051
ldistsq	.0974604	.0201417	4.84	0.000	.0579833	.1369375
y98	.0224743	.0041461	5.42	0.000	.014348	.0306005
y99	.0366898	.0051318	7.15	0.000	.0266317	.046748
y00	.098212	.0055241	17.78	0.000	.0873849	.109039
_cons	6.222005	.9144067	6.80	0.000	4.429801	8.014209
sigma_u	.31933841					
sigma_e	.10651186					
rho	.89988885	(fraction of variance due to u_i)				

```
. xtreg lfare concen y98 y99 y00, fe cluster(id)
```

```
Fixed-effects (within) regression      Number of obs      =      4596
Group variable: id                    Number of groups   =      1149
```

```
R-sq:  within = 0.1352                Obs per group: min =
        between = 0.0576                avg =      4.
        overall = 0.0083                max =
```

```
corr(u_i, Xb) = -0.2033                F(4,1148)           =      120.06
                                         Prob > F            =      0.0000
```

(Std. Err. adjusted for 1149 clusters in id)

lfare	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
concen	.168859	.0494587	3.41	0.001	.0718194	.2658985
y98	.0228328	.004163	5.48	0.000	.0146649	.0310007
y99	.0363819	.0051275	7.10	0.000	.0263215	.0464422
y00	.0977717	.0055054	17.76	0.000	.0869698	.1085735
_cons	4.953331	.0296765	166.91	0.000	4.895104	5.011557
sigma_u	.43389176					
sigma_e	.10651186					
rho	.94316439	(fraction of variance due to u_i)				

b. The REIV estimates without *ldist* and *ldistsq* from Stata are given below. For comparison, the REIV estimates in Example 11.1 are also reported. Dropping the distance variables changes the estimated elasticity to  $-.654$ , which is notably larger in magnitude than  $-.508$ . This is a good example of how relevant time-constant variables – when they are available – should be controlled for in an RE analysis.

```
. xtivreg lpassen y98 y99 y00 (lfare=concen), re
```

```
G2SLS random-effects IV regression      Number of obs      =      4596
Group variable: id                    Number of groups   =      1149
```

```
Obs per group: min =
                avg =      4.
                max =
```

```
Wald chi2(4)      =    219.33
Prob > chi2       =    0.0000
```

lpassen	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
lfare	-.6540984	.4019123	-1.63	0.104	-1.441832	.1336351
y98	.0342955	.011701	2.93	0.003	.011362	.057229
y99	.0847852	.0154938	5.47	0.000	.0544178	.1151525
y00	.146605	.0390819	3.75	0.000	.070006	.2232041
_cons	9.28363	2.032528	4.57	0.000	5.299949	13.26731
sigma_u	.91384976					
sigma_e	.16964171					
rho	.9666879	(fraction of variance due to u_i)				

```
Instrumented:    lfare
Instruments:    y98 y99 y00 concen
```

```
. xtivreg lpassen ldist ldistsq y98 y99 y00 (lfare=concen), re
```

G2SLS random-effects IV regression	Number of obs	=	4596
Group variable: id	Number of groups	=	1149

```
R-sq:  within = 0.4075          Obs per group: min =
        between = 0.0542          avg = 4.
        overall = 0.0641          max =
```

corr(u_i, X)	= 0 (assumed)	Wald chi2(6)	= 231.10
		Prob > chi2	= 0.0000

lpassen	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
lfare	-.5078762	.229698	-2.21	0.027	-.958076	-.0576763
ldist	-1.504806	.6933147	-2.17	0.030	-2.863678	-.1459338
ldistsq	.1176013	.0546255	2.15	0.031	.0105373	.2246652
y98	.0307363	.0086054	3.57	0.000	.0138699	.0476027
y99	.0796548	.01038	7.67	0.000	.0593104	.0999992
y00	.1325795	.0229831	5.77	0.000	.0875335	.1776255
_cons	13.29643	2.626949	5.06	0.000	8.147709	18.44516
sigma_u	.94920686					
sigma_e	.16964171					
rho	.96904799	(fraction of variance due to u_i)				

```
Instrumented:    lfare
Instruments:     ldist ldistsq y98 y99 y00 concen
```

c. Now we have three endogenous variables:  $lfare$ ,  $(ldist - \mu_1) \cdot lfare$ , and

$(ldist^2 - \mu_2) \cdot lfare$ . We can use

$$concen, (ldist - \mu_1) \cdot concen, \text{ and } (ldist^2 - \mu_2) \cdot concen$$

as instruments. In other words, we add the interactions  $(ldist - \mu_1) \cdot concen$  and  $(ldist^2 - \mu_2) \cdot concen$  as extra IVs to account for the endogenous interactions in the structural model.

In practice, we replace  $\mu_1$  and  $\mu_2$  with the sample averages.

d. The Stata output below provides the estimates. Something interesting happens here. The REIV and FEIV estimates of the coefficient on *lfare* now much closer to each other, and much larger in magnitude than the estimates in Table 11.1. In particular, the estimated elasticity at the mean of *ldist* and *ldistsq* is about  $-1$  for REIV and FEIV. Interestingly, the REIV and FEIV estimates with the interactions are close to the RE and FE estimates without the interactions.

```
. egen mu_ldist = mean(ldist)
. gen dmlldist = ldist-mu_ldist
. egen mu_ldistsq = mean(ldistsq)
. gen dmlldistsq = ldistsq-mu_ldistsq
. gen ldist_lfare = dmlldist*lfare
. gen ldistsq_lfare = dmlldistsq*lfare
. gen ldist_concen = dmlldist*concen
. gen ldistsq_concen = dmlldistsq*concen
. xtivreg lpassen ldist ldistsq y98 y99 y00 (lfare ldist_lfare ldistsq_lfare
    = concen ldist_concen ldistsq_concen), re
```

G2SLS random-effects IV regression	Number of obs	=	4596
Group variable: id	Number of groups	=	1149
R-sq: within = 0.1319	Obs per group: min =		
between = 0.0006	avg =		4.
overall = 0.0016	max =		
	Wald chi2(8)	=	180.72
corr(u_i, X) = 0 (assumed)	Prob > chi2	=	0.0000

lpassen	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
lfare	-1.048873	.3250545	-3.23	0.001	-1.685969	-.4117783
ldist_lfare	29.63707	7.957828	3.72	0.000	14.04001	45.23413
ldistsq_lfare	-2.330287	.638173	-3.65	0.000	-3.581083	-1.079491
ldist	-157.8477	42.76284	-3.69	0.000	-241.6613	-74.03409
ldistsq	12.45005	3.437782	3.62	0.000	5.712121	19.18798
y98	.0319578	.0105546	3.03	0.002	.0112713	.0526444
y99	.080002	.0127579	6.27	0.000	.0549969	.1050071
y00	.1570325	.026578	5.91	0.000	.1049406	.2091244
_cons	504.8691	131.4462	3.84	0.000	247.2392	762.499
sigma_u	1.3686882					
sigma_e	.19436268					
rho	.98023276	(fraction of variance due to u_i)				
Instrumented:	lfare ldist_lfare ldistsq_lfare					
Instruments:	ldist ldistsq y98 y99 y00 concen ldist_concen ldistsq_concen					

```
. xtivreg lpassen y98 y99 y00 (lfare ldist_lfare ldistsq_lfare
= concen ldist_concen ldistsq_concen), fe
```

```
Fixed-effects (within) IV regression      Number of obs      =      4596
Group variable: id                       Number of groups    =      1149
```

```
R-sq:  within =      .
       between = 0.0016
       overall = 0.0016
Obs per group: min =
               avg =      4.
               max =
```

```
corr(u_i, Xb) = -0.9913
Wald chi2(6)      =      4.40e+06
Prob > chi2       =      0.0000
```

lpassen	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
lfare	-1.011863	.3214187	-3.15	0.002	-1.641832	-.3818937
ldist_lfare	24.11579	6.951145	3.47	0.001	10.4918	37.73979
ldistsq_lf~e	-1.905021	.5593222	-3.41	0.001	-3.001273	-.8087699
y98	.0322146	.0102786	3.13	0.002	.0120689	.0523603
y99	.080772	.0123315	6.55	0.000	.0566026	.1049414
y00	.155485	.0260008	5.98	0.000	.1045244	.2064456
_cons	11.33584	1.694321	6.69	0.000	8.015032	14.65665
sigma_u	6.6845875					
sigma_e	.19436268					
rho	.99915529	(fraction of variance due to u_i)				

```
F test that all u_i=0:      F(1148,3441) =      72.37      Prob > F      = 0.0000
```

```
Instrumented: lfare ldist_lfare ldistsq_lfare
Instruments: ldist ldistsq y98 y99 y00 concen ldist_concen ldistsq_concen
```

e. We can use the command `xtivreg2`, a user-written program for Stata. The 95%

confidence interval for  $\alpha_1$  is  $[-2.408, .385]$ , which includes zero. The fully robust joint test of the two interaction terms gives  $p$ -value = .101, so we might be justified in dropping them. The robust standard error

```
. xtivreg2 lpassen y98 y99 y00 (lfare ldist_lfare ldistsq_lfare
    = concen ldist_concen ldistsq_concen), fe cluster(id)
```

#### FIXED EFFECTS ESTIMATION

```
-----
Number of groups =      1149                      Obs per group: min =
                                                    avg =      4.
                                                    max =
```

#### IV (2SLS) estimation

Estimates efficient for homoskedasticity only

Statistics robust to heteroskedasticity and clustering on id

```
Number of clusters (id) = 1149                      Number of obs =      4596
                                                    F(   6, 1148) =      14.90
                                                    Prob > F       =      0.0000
Total (centered) SS      = 128.0991685              Centered R2     = -0.0148
Total (uncentered) SS   = 128.0991685              Uncentered R2   = -0.0148
Residual SS             = 129.9901441              Root MSE       =      .1942
```

lpassen	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval	
lfare	-1.011863	.7124916	-1.42	0.156	-2.408321	.384595
ldist_lfare	24.11579	11.26762	2.14	0.032	2.03166	46.19993
ldistsq_lfare	-1.905021	.8941964	-2.13	0.033	-3.657614	-.1524285
y98	.0322146	.0167737	1.92	0.055	-.0006613	.0650905
y99	.080772	.0261059	3.09	0.002	.0296053	.1319387
y00	.155485	.0625692	2.49	0.013	.0328515	.2781184

```
-----
Instrumented:      lfare ldist_lfare ldistsq_lfare
Included instruments: y98 y99 y00
Excluded instruments: concen ldist_concen ldistsq_concen
-----
```

```
. test ldist_lfare ldistsq_lfare
```

```
( 1)  ldist_lfare = 0
( 2)  ldistsq_lfare = 0
```

```
      chi2(  2) =      4.59
Prob > chi2 =      0.1008
```

f. In general, the estimated elasticities can be obtained from



$$\frac{\widehat{lpassen}}{lfare} = \hat{\alpha}_1 + \hat{\gamma}_1(ldist - \hat{\mu}_1) + \hat{\gamma}_2(ldist^2 - \hat{\mu}_2)$$

for any value of *ldist*. Calculations are given below. For *dist* = 500 the estimated elasticity is about .047 with a very small *t* statistic. For *dist* = 1,500, the estimated elasticity is -1.77 with fully robust *t* = -1.55. So the magnitude of the elasticity increases substantially as the route distance increase, but the estimates contain substantial noise.

```
. sum ldist ldistsq if y00
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ldist	1149	6.696482	.6595331	4.553877	7.909857
ldistsq	1149	45.27747	8.729749	20.73779	62.56583

```
. di log(500) - 6.696482
-.4818739
```

```
. di (log(500))^2 - 45.27747
-6.6561162
```

```
. lincom lfare - .4818739*ldist_lfare - 6.6561162*ldistsq_lfare
( 1)  lfare - .4818739*ldist_lfare - 6.656116*ldistsq_lfare = 0
```

lpassen	Coef.	Std. Err.	z	P> z	[95% Conf. Interval
(1)	.0474081	.7405447	0.06	0.949	-1.404033 1.498849

```
. di log(1500) - 6.696482
.61673839
```

```
. di (log(1500))^2 - 45.27747
8.2057224
```

```
. lincom lfare +.61673839*ldist_lfare +8.2057224*ldistsq_lfare
( 1)  lfare + .6167384*ldist_lfare + 8.205722*ldistsq_lfare = 0
```

lpassen	Coef.	Std. Err.	z	P> z	[95% Conf. Interval
(1)	-1.770802	1.142114	-1.55	0.121	-4.009305 .4677006

## Solutions to Chapter 12 Problems

**12.1.** a. Take the conditional expectation of equation (12.4) with respect to  $\mathbf{x}$ , and use

$$E(u|\mathbf{x}) = 0:$$

$$\begin{aligned} E\{[y - m(\mathbf{x}, \boldsymbol{\theta})]^2|\mathbf{x}\} &= E(u^2|\mathbf{x}) + 2[m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})]E(u|\mathbf{x}) + E\{[m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})]^2|\mathbf{x}\} \\ &= E(u^2|\mathbf{x}) + 0 + [m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})]^2 \\ &= E(u^2|\mathbf{x}) + [m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})]^2. \end{aligned}$$

The first term does not depend on  $\boldsymbol{\theta}$  and the second term is clearly minimized at  $\boldsymbol{\theta} = \boldsymbol{\theta}_o$  for any  $\mathbf{x}$ . Therefore, the parameters of a correctly specified conditional mean function minimize the squared error conditional on any value of  $\mathbf{x}$ .

b. Part a shows that

$$E\{[y - m(\mathbf{x}, \boldsymbol{\theta}_o)]^2|\mathbf{x}\} \leq E\{[y - m(\mathbf{x}, \boldsymbol{\theta})]^2|\mathbf{x}\}, \text{ all } \boldsymbol{\theta} \in \boldsymbol{\Theta}, \text{ all } \mathbf{x} \in \mathcal{X}.$$

If we take the expected value of both sides – with respect to the distribution of  $\mathbf{x}$ , of course – and apply iterated expectations, we conclude

$$E\{[y - m(\mathbf{x}, \boldsymbol{\theta}_o)]^2\} \leq E\{[y - m(\mathbf{x}, \boldsymbol{\theta})]^2\}, \text{ all } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

In other words, if we know  $\boldsymbol{\theta}_o$  solves the population minimization problem conditional on any  $\mathbf{x}$ , then it also solves the unconditional population problem. Of course, conditional on a particular value of  $\mathbf{x}$ ,  $\boldsymbol{\theta}_o$  would usually not be the unique solution. (For example, in the linear case  $m(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}\boldsymbol{\theta}$ , any  $\boldsymbol{\theta}$  such as that  $\mathbf{x}(\boldsymbol{\theta}_o - \boldsymbol{\theta}) = 0$  sets  $m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})$  to zero.)

Uniqueness of  $\boldsymbol{\theta}_o$  as a population minimizer is realistic only after we integrate out  $\mathbf{x}$  to obtain  $E\{[y - m(\mathbf{x}, \boldsymbol{\theta})]^2\}$ .

**12.2.** a. Since  $u = y - E(y|\mathbf{x})$ ,

$$\text{Var}(y|\mathbf{x}) = \text{Var}(u|\mathbf{x}) = E(u^2|\mathbf{x})$$

because  $E(u|\mathbf{x}) = 0$ . So  $E(u^2|\mathbf{x}) = \exp(\alpha_o + \mathbf{x}\boldsymbol{\gamma}_o)$ .

b. If we knew the  $u_i = y_i - m(\mathbf{x}_i, \boldsymbol{\theta}_o)$ , then we could do a nonlinear regression of  $u_i^2$  on  $\exp(\alpha + \mathbf{x}\boldsymbol{\gamma})$  and just use the asymptotic theory for nonlinear regression. The NLS estimators of  $\alpha$  and  $\boldsymbol{\gamma}$  would then solve

$$\min_{\alpha, \boldsymbol{\gamma}} \sum_{i=1}^N [u_i^2 - \exp(\alpha + \mathbf{x}_i \boldsymbol{\gamma})]^2.$$

The problem is that  $\boldsymbol{\theta}_o$  is unknown. When we replace  $\boldsymbol{\theta}_o$  with its NLS estimator,  $\hat{\boldsymbol{\theta}}$  – that is we replace  $u_i^2$  with  $\hat{u}_i^2$ , the squared NLS residuals – we are solving the problem

$$\min_{\alpha, \boldsymbol{\gamma}} \sum_{i=1}^N \{[y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}})]^2 - \exp(\alpha + \mathbf{x}_i \boldsymbol{\gamma})\}^2.$$

This objective function has the form of a two-step M-estimator in Section 12.4. Since  $\hat{\boldsymbol{\theta}}$  is generally consistent for  $\boldsymbol{\theta}_o$ , the two-step M-estimator is generally consistent for  $\alpha_o$  and  $\boldsymbol{\gamma}_o$  (under weak regularity and identification conditions). In fact,  $\sqrt{N}$ -consistency of  $\hat{\alpha}$  and  $\hat{\boldsymbol{\gamma}}$  holds very generally.

c. We now estimate  $\boldsymbol{\theta}_o$  by solving

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^N [y_i - m(\mathbf{x}_i, \boldsymbol{\theta})]^2 / \exp(\hat{\alpha} + \mathbf{x}_i \hat{\boldsymbol{\gamma}}),$$

where  $\hat{\alpha}$  and  $\hat{\boldsymbol{\gamma}}$  are from part b. The general theory of WNLS under WNLS.1 to WNLS.3 can be applied.

d. Using the definition of  $v$ , write  $u^2 = \exp(\alpha_o + \mathbf{x}\boldsymbol{\gamma}_o)v^2$ . Taking logs gives  $\log(u^2) = \alpha_o + \mathbf{x}\boldsymbol{\gamma}_o + \log(v^2)$ . Now, if  $v$  is independent of  $\mathbf{x}$ , so is  $\log(v^2)$ . Therefore,  $E[\log(u^2)|\mathbf{x}] = \alpha_o + \mathbf{x}\boldsymbol{\gamma}_o + E[\log(v^2)|\mathbf{x}] = \alpha_o + \mathbf{x}\boldsymbol{\gamma}_o + \kappa_o$ , where  $\kappa_o \equiv E[\log(v^2)]$ . So, if we

could observe the  $u_i$ , and OLS regression of  $\log(u_i^2)$  on  $1, \mathbf{x}_i$  would be consistent for  $(\alpha_o + \kappa_o, \boldsymbol{\gamma}_o)$ ; in fact, it would be unbiased. By two-step estimation theory, consistency still holds if  $u_i$  is replaced with  $\hat{u}_i$ , by essentially the same argument in part b. So, if  $m(\mathbf{x}, \boldsymbol{\theta})$  is linear in  $\boldsymbol{\theta}$ , we can carry out a weighted NLS procedure without ever doing nonlinear estimation.

e. If we have misspecified the variance function – or, for example, we use the approach in part d but  $v$  is not independent of  $\mathbf{x}$  – then we should use a fully robust variance-covariance matrix in equation (12.60) with  $\hat{h}_i = \exp(\hat{\alpha} + \mathbf{x}_i \hat{\boldsymbol{\gamma}})$ .

**12.3.** a. The approximate elasticity is

$$\partial \log[\hat{E}(y|\mathbf{z})]/\partial \log(z_1) = \partial[\hat{\theta}_1 + \hat{\theta}_2 \log(z_1) + \hat{\theta}_3 z_2]/\partial \log(z_1) = \hat{\theta}_2.$$

b. This is approximated by  $100 \cdot \partial \log[\hat{E}(y|\mathbf{z})]/\partial z_2 = 100 \cdot \hat{\theta}_3$ .

c. Since  $\partial \hat{E}(y|\mathbf{z})/\partial z_2 = \exp[\hat{\theta}_1 + \hat{\theta}_2 \log(z_1) + \hat{\theta}_3 z_2 + \hat{\theta}_4 z_2^2] \cdot (\hat{\theta}_3 + 2\hat{\theta}_4 z_2)$ , the estimated turning point is  $\hat{z}_2^* = \hat{\theta}_3/(-2\hat{\theta}_4)$ . This is a consistent estimator of  $z_2^* = \theta_3/(-2\theta_4)$ .

d. Since  $\nabla_{\boldsymbol{\theta}} m(\mathbf{x}, \boldsymbol{\theta}) = \exp(\mathbf{x}_1 \boldsymbol{\theta}_1 + \mathbf{x}_2 \boldsymbol{\theta}_2) \mathbf{x}$ , the gradient of the mean function evaluated under the null is

$$\nabla_{\boldsymbol{\theta}} \tilde{m}_i = \exp(\mathbf{x}_{i1} \tilde{\boldsymbol{\theta}}_1) \mathbf{x}_i \equiv \tilde{m}_i \mathbf{x}_i,$$

where  $\tilde{\boldsymbol{\theta}}_1$  is the restricted NLS estimator. From regression (12.72), we can compute the usual

LM statistic as  $NR_u^2$  from the regression  $\tilde{u}_i$  on  $\tilde{m}_i \mathbf{x}_{i1}, \tilde{m}_i \mathbf{x}_{i2}$ ,  $i = 1, \dots, N$ , where  $\tilde{u}_i = y_i - \tilde{m}_i$ .

For the robust test, we first regress  $\tilde{m}_i \mathbf{x}_{i2}$  on  $\tilde{m}_i \mathbf{x}_{i1}$  and obtain the  $1 \times K_2$  residuals,  $\tilde{\mathbf{r}}_i$ . Then we compute the statistic as in regression (12.75).

**12.4.** a. Write the objective function as  $(1/2) \sum_{i=1}^N [y_i - m(\mathbf{x}_i, \boldsymbol{\theta})]^2 / h(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})$ . The objective function, for any value of  $\boldsymbol{\gamma}$ , is

$$q(\mathbf{w}_i, \boldsymbol{\theta}; \boldsymbol{\gamma}) = (1/2)[y_i - m(\mathbf{x}_i, \boldsymbol{\theta})]^2/h(\mathbf{x}_i, \boldsymbol{\gamma}).$$

Taking the gradient with respect to  $\boldsymbol{\theta}$  gives

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} q(w_i, \boldsymbol{\theta}; \boldsymbol{\gamma}) &= -\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta})[y_i - m(\mathbf{x}_i, \boldsymbol{\theta})]/h(\mathbf{x}_i, \boldsymbol{\gamma}) \\ &= -\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta})u_i(\boldsymbol{\theta})/h(\mathbf{x}_i, \boldsymbol{\gamma}).\end{aligned}$$

Taking the transpose gives us the score with respect to  $\boldsymbol{\theta}$  for any  $\boldsymbol{\theta}$  and any  $\boldsymbol{\gamma}$ .

b. This follows because, under WNLS.1,  $u_i \equiv u_i(\boldsymbol{\theta}_o)$  has a zero mean given  $\mathbf{x}_i$ :

$$\mathbb{E}[\mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma})|\mathbf{x}_i] = -\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta}_o)' \mathbb{E}(u_i|\mathbf{x}_i)/h(\mathbf{x}_i, \boldsymbol{\gamma}) = \mathbf{0};$$

the value of  $\boldsymbol{\gamma}$  plays no role.

c. First, the Jacobian of  $\mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma})$  with respect to  $\boldsymbol{\gamma}$  is

$\nabla_{\boldsymbol{\gamma}} \mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}) = \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta}_o)' u_i \nabla_{\boldsymbol{\gamma}} h(\mathbf{x}_i, \boldsymbol{\gamma})/[h(\mathbf{x}_i, \boldsymbol{\gamma})]^2$ . Everything but  $u_i$  is a function only of  $\mathbf{x}_i$ , so

$$\mathbb{E}[\nabla_{\boldsymbol{\gamma}} \mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma})|\mathbf{x}_i] = \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta}_o)' \mathbb{E}(u_i|\mathbf{x}_i) \nabla_{\boldsymbol{\gamma}} h(\mathbf{x}_i, \boldsymbol{\gamma})/[h(\mathbf{x}_i, \boldsymbol{\gamma})]^2 = \mathbf{0}.$$

It follows by the LIE that the unconditional expectation is zero, too. In other words, we have shown that the key condition (12.37) holds (and we did not rely on Assumption WNLS.3).

d. We would just use equation (12.60), which can be written as

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}) = \left( \sum_{i=1}^N \nabla_{\boldsymbol{\theta}} \check{m}_i' \nabla_{\boldsymbol{\theta}} \check{m}_i \right)^{-1} \left( \sum_{i=1}^N \check{u}_i^2 \nabla_{\boldsymbol{\theta}} \check{m}_i' \nabla_{\boldsymbol{\theta}} \check{m}_i \right) \left( \sum_{i=1}^N \nabla_{\boldsymbol{\theta}} \check{m}_i' \nabla_{\boldsymbol{\theta}} \check{m}_i \right)^{-1},$$

where  $\check{u}_i \equiv \hat{u}_i/\hat{h}_i^{1/2}$  and  $\nabla_{\boldsymbol{\theta}} \check{m}_i \equiv \nabla_{\boldsymbol{\theta}} \hat{m}_i/\hat{h}_i^{1/2}$  are the standardized residuals and gradient, respectively.

e. Under Assumption WNLS.3 (along with WNLS.1),

$$\text{Var}(y_i|\mathbf{x}_i) = \mathbb{E}(u_i^2|\mathbf{x}_i) = \sigma_o^2 h(\mathbf{x}_i, \boldsymbol{\gamma}_o),$$

and  $\hat{\boldsymbol{\gamma}}$  is  $\sqrt{N}$ -consistent for  $\boldsymbol{\gamma}_o$ . This ensures that the asymptotic variance of  $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$  does

not depend on that of  $\sqrt{N}(\hat{\gamma} - \gamma_o)$ . Further,

$$\begin{aligned}\mathbf{A}_o &= E[\nabla_{\theta} m(\mathbf{x}_i, \theta_o)' \nabla_{\theta} m(\mathbf{x}_i, \theta_o) / h(\mathbf{x}_i, \gamma_o)] \\ \mathbf{B}_o &= E[\mathbf{s}_i(\theta_o; \gamma_o) \mathbf{s}_i(\theta_o; \gamma_o)'] = E\{u_i^2 \nabla_{\theta} m(\mathbf{x}_i, \theta_o)' \nabla_{\theta} m(\mathbf{x}_i, \theta_o) / [h(\mathbf{x}_i, \gamma_o)]^2\}.\end{aligned}$$

By iterated expectations and WNLS.3,

$$\begin{aligned}E\{u_i^2 \nabla_{\theta} m(\mathbf{x}_i, \theta_o)' \nabla_{\theta} m(\mathbf{x}_i, \theta_o) / [h(\mathbf{x}_i, \gamma_o)]^2\} &= E\{u_i^2 \nabla_{\theta} m(\mathbf{x}_i, \theta_o)' \nabla_{\theta} m(\mathbf{x}_i, \theta_o) / [h(\mathbf{x}_i, \gamma_o)]^2\} \\ &= E(E\{u_i^2 \nabla_{\theta} m(\mathbf{x}_i, \theta_o)' \nabla_{\theta} m(\mathbf{x}_i, \theta_o) / [h(\mathbf{x}_i, \gamma_o)]^2\} | \mathbf{x}_i) \\ &= E\{E(u_i^2 | \mathbf{x}_i) \nabla_{\theta} m(\mathbf{x}_i, \theta_o)' \nabla_{\theta} m(\mathbf{x}_i, \theta_o) / [h(\mathbf{x}_i, \gamma_o)]^2\} \\ &= E\{\sigma_o^2 h(\mathbf{x}_i, \gamma_o) \nabla_{\theta} m(\mathbf{x}_i, \theta_o)' \nabla_{\theta} m(\mathbf{x}_i, \theta_o) / [h(\mathbf{x}_i, \gamma_o)]^2\} \\ &= \sigma_o^2 E[\nabla_{\theta} m(\mathbf{x}_i, \theta_o)' \nabla_{\theta} m(\mathbf{x}_i, \theta_o) / h(\mathbf{x}_i, \gamma_o)] \\ &= \sigma_o^2 \mathbf{A}_o.\end{aligned}$$

Therefore,

$$\text{Avar}[\sqrt{N}(\hat{\theta} - \theta_o)] = \sigma_o^2 \mathbf{A}_o^{-1}$$

an a consistent estimator is

$$\hat{\sigma}^2 \left[ N^{-1} \sum_{i=1}^N \nabla_{\theta} m(\mathbf{x}_i, \hat{\theta})' \nabla_{\theta} m(\mathbf{x}_i, \hat{\theta}) / h(\mathbf{x}_i, \hat{\gamma}) \right]^{-1}$$

Dividing this expression by  $N$  to get  $\widehat{\text{Avar}}(\hat{\theta})$  delivers (12.59).

**12.5. a.** We need the gradient of  $m(\mathbf{x}_i, \theta)$  evaluated under the null hypothesis. By the chain rule,

$$\begin{aligned}\nabla_{\beta} m(\mathbf{x}, \theta) &= g[\mathbf{x}\beta + \delta_1(\mathbf{x}\beta)^2 + \delta_2(\mathbf{x}\beta)^3] \cdot [\mathbf{x} + 2\delta_1(\mathbf{x}\beta)^2 \mathbf{x} + 3\delta_2(\mathbf{x}\beta)^2 \mathbf{x}], \\ \nabla_{\delta} m(\mathbf{x}, \theta) &= g[\mathbf{x}\beta + \delta_1(\mathbf{x}\beta)^2 + \delta_2(\mathbf{x}\beta)^3] \cdot [(\mathbf{x}\beta)^2, (\mathbf{x}\beta)^3]\end{aligned}$$

The gradients with  $\delta_1 = \delta_2 = 0$  are

$$\begin{aligned}\nabla_{\beta} m(\mathbf{x}, \beta, 0) &= g(\mathbf{x}\beta) \cdot \mathbf{x} \\ \nabla_{\delta} m(\mathbf{x}, \beta, 0) &= g(\mathbf{x}\beta) \cdot [(\mathbf{x}\beta)^2, (\mathbf{x}\beta)^3].\end{aligned}$$

Let  $\tilde{\beta}$  denote the NLS estimator with  $\delta_1 = \delta_2 = 0$  imposed. Then  $\nabla_{\beta} m(\mathbf{x}_i, \tilde{\theta}) = g(\mathbf{x}_i \tilde{\beta}) \mathbf{x}_i$  and  $\nabla_{\delta} m(\mathbf{x}_i, \tilde{\theta}) = g(\mathbf{x}_i \tilde{\beta})[(\mathbf{x}_i \tilde{\beta})^2, (\mathbf{x}_i \tilde{\beta})^3]$ . Therefore, the usual LM statistic can be obtained as  $NR_u^2$  from the regression  $\tilde{u}_i$  on  $\tilde{g}_i \mathbf{x}_i, \tilde{g}_i \cdot (\mathbf{x}_i \tilde{\beta})^2, \tilde{g}_i \cdot (\mathbf{x}_i \tilde{\beta})^3$ , where  $\tilde{g}_i \equiv g(\mathbf{x}_i \tilde{\beta})$ . If  $G(\cdot)$  is the identity function,  $g(\cdot) \equiv 1$ , and the auxiliary regression is

$$\tilde{u}_i \text{ on } \mathbf{x}_i, (\mathbf{x}_i \tilde{\beta})^2, (\mathbf{x}_i \tilde{\beta})^3,$$

which is a versino of RESET.

b. The VAT version of the test is obtained as follows. As with the LM test, first estimate the model under the null and obtain the NLS estimator,  $\tilde{\beta}$ , as before. Then estimate the auxiliary model with  $(\mathbf{x}_i \tilde{\beta})^2$  and  $(\mathbf{x}_i \tilde{\beta})^3$  as explanatory variables. In other words, act as if the mean function is

$$G[\mathbf{x}_i \beta + \delta_1 (\mathbf{x}_i \tilde{\beta})^2 + \delta_2 (\mathbf{x}_i \tilde{\beta})^3]$$

and estimate  $\delta_1$  and  $\delta_2$  along with  $\beta$ . A joint Wald test, made robust to heteroskedasticity if necessary, of  $H_0 : \delta_1 = 0, \delta_2 = 0$  is asymptotically equivalent (has the same asymptotic size and asymptotic power against local alternatives) to the LM test. Given the way modern software works, this often affords some computational simplification (albeit modest). When  $G(\cdot)$  is the identify function, this variable addition approach gives the RESET test in its traditional form.

One danger in using the VAT is that it is tempting to use the second-step estimates of  $\delta_1$ ,  $\delta_2$ , and even  $\beta$  as generally valid estimators. But they are not. If the null is false,  $\tilde{\beta}$  is inconsistent for  $\beta$  (because  $\tilde{\beta}$  is imposed with  $\delta_1 = 0, \delta_2 = 0$ ) and so the added variables are not correct under the alternative. The VAT should be used only for testing purposes (just like the LM statistic).

**12.6. a.** The pooled NLS estimator of  $\boldsymbol{\theta}_o$  solves

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^N \sum_{t=1}^T [y_{it} - m(\mathbf{x}_{it}, \boldsymbol{\theta})]^2/2,$$

and so, to put this into the standard M-estimation framework, we can take the objective function for a random draw  $i$  to be  $q_i(\boldsymbol{\theta}) \equiv q(\mathbf{w}_i, \boldsymbol{\theta}) \equiv \sum_{t=1}^T [y_{it} - m(\mathbf{x}_{it}, \boldsymbol{\theta})]^2/2$ . The score for random draw  $i$  is  $\mathbf{s}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} q_i(\boldsymbol{\theta}) = -\sum_{t=1}^T \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta})' u_{it}(\boldsymbol{\theta})$ . Without further assumptions, a consistent estimator of  $\mathbf{B}_o$  is

$$\hat{\mathbf{B}} = N^{-1} \sum_{i=1}^N \mathbf{s}_i(\hat{\boldsymbol{\theta}}) \mathbf{s}_i(\hat{\boldsymbol{\theta}})'$$

where  $\hat{\boldsymbol{\theta}}$  is the pooled NLS estimator. The Hessian for observation  $i$ , which can be computed as the Jacobian of the score, can be written as

$$\mathbf{H}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathbf{s}_i(\boldsymbol{\theta}) = -\sum_{t=1}^T \nabla_{\boldsymbol{\theta}}^2 m(\mathbf{x}_{it}, \boldsymbol{\theta}) u_{it}(\boldsymbol{\theta}) + \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta})' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}).$$

When we plug in  $\boldsymbol{\theta}_o$  and use the fact that  $E(u_{it}|\mathbf{x}_{it}) = 0$ , all  $t = 1, \dots, T$ , then

$$\begin{aligned} \mathbf{A}_o \equiv E[\mathbf{H}_i(\boldsymbol{\theta}_o)] &= -\sum_{t=1}^T E[\nabla_{\boldsymbol{\theta}}^2 m(\mathbf{x}_{it}, \boldsymbol{\theta}_o) u_{it}] + \sum_{t=1}^T E[\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)] \\ &= \sum_{t=1}^T E[\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)] \end{aligned}$$

because  $E[\nabla_{\boldsymbol{\theta}}^2 m(\mathbf{x}_{it}, \boldsymbol{\theta}_o) u_{it}] = \mathbf{0}$ ,  $t = 1, \dots, T$  by iterated expectations. By the usual law of large numbers argument,

$$\hat{\mathbf{A}} \equiv N^{-1} \sum_{i=1}^N \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \hat{\boldsymbol{\theta}})' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \hat{\boldsymbol{\theta}}) \equiv N^{-1} \sum_{i=1}^N \hat{\mathbf{A}}_i$$

is a consistent estimator of  $\mathbf{A}_o$ . Then, we just use the usual sandwich formula in equation



(12.49).

b. As in the hint we show that  $\mathbf{B}_o = \sigma_o^2 \mathbf{A}_o$ . First, write  $\mathbf{s}_i(\boldsymbol{\theta}) = \sum_{t=1}^T \mathbf{s}_{it}(\boldsymbol{\theta})$ , where  $\mathbf{s}_{it}(\boldsymbol{\theta}) \equiv -\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}) u_{it}(\boldsymbol{\theta})$ . Under dynamic completeness of the mean, these scores are serially uncorrelated across  $t$  (when evaluated at  $\boldsymbol{\theta}_o$ , of course). The argument is very similar to the linear regression case from Chapter 7.

Let  $r < t$  for concreteness. Then

$$E[\mathbf{s}_{it}(\boldsymbol{\theta}_o) \mathbf{s}_{ir}(\boldsymbol{\theta}_o)' | \mathbf{x}_{it}, \mathbf{x}_{ir}, u_{ir}] = E(u_{it} | \mathbf{x}_{it}, \mathbf{x}_{ir}, u_{ir}) u_{ir} \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{ir}, \boldsymbol{\theta}_o) = 0$$

because  $E(u_{it} | \mathbf{x}_{it}, u_{i,t-1}, \mathbf{x}_{i,t-1}, \dots) = 0$  and  $r < t$ . Now apply the LIE to conclude

$E[\mathbf{s}_{it}(\boldsymbol{\theta}_o) \mathbf{s}_{ir}(\boldsymbol{\theta}_o)'] = \mathbf{0}$ . So we have shown that  $\mathbf{B}_o = \sum_{t=1}^T E[\mathbf{s}_{it}(\boldsymbol{\theta}_o) \mathbf{s}_{it}(\boldsymbol{\theta}_o)']$ . But for each  $t$ , apply iterated expectations:

$$\begin{aligned} E[\mathbf{s}_{it}(\boldsymbol{\theta}_o) \mathbf{s}_{it}(\boldsymbol{\theta}_o)'] &= E(u_{it}^2 \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)) \\ &= E[E(u_{it}^2 | \mathbf{x}_{it}) \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)] \\ &= \sigma_o^2 E[\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)] \end{aligned}$$

where the last equality follows because  $E(u_{it}^2 | \mathbf{x}_{it}) = \sigma_o^2$ . It follows that

$$\mathbf{B}_o = \sum_{t=1}^T E[\mathbf{s}_{it}(\boldsymbol{\theta}_o) \mathbf{s}_{it}(\boldsymbol{\theta}_o)'] = \sigma_o^2 \sum_{t=1}^T E[\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)] = \sigma_o^2 \mathbf{A}_o.$$

Next, the usual two-step estimation argument – see Lemma 12.1 – shows that

$$(NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T \hat{u}_{it}^2 \xrightarrow{P} T^{-1} \sum_{t=1}^T E(u_{it}^2) = \sigma_o^2 \text{ as } N \rightarrow \infty.$$

The degrees of freedom correction – putting  $NT - P$  in place of  $NT$  – does not affect consistency. The variance matrix obtained by ignoring the time dimension and assuming homoskedasticity is simply

$$\hat{\sigma}^2 \left( \sum_{i=1}^N \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \hat{\boldsymbol{\theta}})' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \hat{\boldsymbol{\theta}}) \right)^{-1},$$

and we just showed that  $N$  times this matrix is a consistent estimator of  $\text{Avar} \sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$ .

c. As we just saw in part b,  $\mathbf{B}_o = \sigma_o^2 \mathbf{A}_o$ , which means by slightly extending the argument before (12.69) we can use an extension of the LM statistic there. Namely,

$$LM = \left( \sum_{i=1}^N \tilde{\mathbf{s}}_i \right)' (\tilde{\sigma}^2 \tilde{\mathbf{M}})^{-1} \left( \sum_{i=1}^N \tilde{\mathbf{s}}_i \right).$$

It is convenient to choose

$$\tilde{\mathbf{M}} = \sum_{i=1}^N \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \tilde{\boldsymbol{\theta}})' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \tilde{\boldsymbol{\theta}}),$$

where  $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\delta}}')$ . So the LM statistic can be written as

$$\tilde{\sigma}^{-2} \left( \sum_{i=1}^N \sum_{t=1}^T \tilde{u}_{it} \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \tilde{\boldsymbol{\theta}}) \right) \left( \sum_{i=1}^N \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \tilde{\boldsymbol{\theta}})' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \tilde{\boldsymbol{\theta}}) \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \tilde{\boldsymbol{\theta}})' \tilde{u}_{it} \right)$$

where we take

$$\tilde{\sigma}^2 = (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T \tilde{u}_{it}^2.$$

(It is common not to use the degrees of freedom adjustment when estimating  $\sigma_o^2$  under the null.). Finally, the LM statistic can be written as

$$\begin{aligned} LM &= \frac{\left( \sum_{i=1}^N \sum_{t=1}^T \tilde{u}_{it} \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \tilde{\boldsymbol{\theta}}) \right) \left( \sum_{i=1}^N \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \tilde{\boldsymbol{\theta}})' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \tilde{\boldsymbol{\theta}}) \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \tilde{\boldsymbol{\theta}})' \tilde{u}_{it} \right)}{(NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T \tilde{u}_{it}^2} \\ &= NT \cdot \frac{\left( \sum_{i=1}^N \sum_{t=1}^T \tilde{u}_{it} \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \tilde{\boldsymbol{\theta}}) \right) \left( \sum_{i=1}^N \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \tilde{\boldsymbol{\theta}})' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \tilde{\boldsymbol{\theta}}) \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \tilde{\boldsymbol{\theta}})' \tilde{u}_{it} \right)}{\sum_{i=1}^N \sum_{t=1}^T \tilde{u}_{it}^2} \\ &= NTR_u^2 \end{aligned}$$

because the numerator is the explained sum of squares from the pooled OLS regression

$$\tilde{u}_{it} \text{ on } \nabla_{\theta} m(\mathbf{x}_{it}, \tilde{\theta}), t = 1, \dots, T; i = 1, \dots, N$$

and the numerator is the (uncentered) total sum of squares.

**12.7.** a. For each  $i$  and  $g$ , define  $u_{ig} \equiv y_{ig} - m(\mathbf{x}_{ig}, \theta_{og})$ , so that  $E(u_{ig}|\mathbf{x}_i) = 0$ ,  $g = 1, \dots, G$ . Further, let  $\mathbf{u}_i$  be the  $G \times 1$  vector containing the  $u_{ig}$ . Then  $E(\mathbf{u}_i \mathbf{u}_i' | \mathbf{x}_i) = E(\mathbf{u}_i \mathbf{u}_i') = \mathbf{\Omega}_o$ . Let  $\check{\mathbf{u}}_i$  be the vector of nonlinear least squares residuals for each observation  $i$ . That is, compute the NLS estimates for each equation  $g$  and collect the residuals. Then, by standard arguments (apply Lemma 12.1), a consistent estimator of  $\mathbf{\Omega}_o$  is

$$\hat{\mathbf{\Omega}} \equiv N^{-1} \sum_{i=1}^N \check{\mathbf{u}}_i \check{\mathbf{u}}_i'$$

because each NLS estimator,  $\hat{\theta}_g$  is consistent for  $\theta_{og}$  as  $N \rightarrow \infty$ .

b. This part involves several steps, and I will sketch how each one goes. First, let  $\gamma$  be the vector of distinct elements of  $\mathbf{\Omega}$  – the nuisance parameters in the context of two-step M-estimation. Then, the score for observation  $i$  is

$$\begin{aligned} \mathbf{s}(\mathbf{w}_i, \theta; \gamma) &= -\nabla_{\theta} \mathbf{m}(\mathbf{x}_i, \theta)' \mathbf{\Omega}^{-1} \mathbf{u}_i(\theta) \\ &= -[\mathbf{u}_i(\theta) \otimes \nabla_{\theta} \mathbf{m}(\mathbf{x}_i, \theta)]' \text{vec}(\mathbf{\Omega}^{-1}) \end{aligned}$$

where  $\mathbf{m}(\mathbf{x}_i, \theta)$  is the  $G \times 1$  vector of conditional mean functions. With this expression we can verify condition (12.37), even though the actual derivatives are complicated. It is clear that  $\nabla_{\gamma} \mathbf{s}(\mathbf{w}_i, \theta; \gamma)$  is a linear combination of  $\mathbf{u}_i(\theta)$ , where the linear combination is a function of  $\mathbf{x}_i$  (and the parameter values). Therefore, because  $E(\mathbf{u}_i | \mathbf{x}_i) = \mathbf{0}$ ,  $E[\nabla_{\gamma} \mathbf{s}(\mathbf{w}_i, \theta_o; \gamma) | \mathbf{x}_i] = \mathbf{0}$  for any  $\gamma$ , that is, any  $\mathbf{\Omega}$ . Its unconditional expectation is zero, too, which verifies (12.37). This shows that we do not have to adjust for the first-stage estimation of  $\mathbf{\Omega}_o$ . (Note: This problem assumes

that  $\text{Var}(\mathbf{u}_i|\mathbf{x}_i) = \mathbf{\Omega}_o$ , but it is clear that (12.37) holds without any assumption about  $\text{Var}(\mathbf{u}_i|\mathbf{x}_i)$ . We just need the estimator we use,  $\hat{\mathbf{\Omega}}$ , to converge to its limit at the usual  $\sqrt{N}$  rate.)

Next we obtain  $\mathbf{B}_o \equiv E[\mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o) \mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o)']$ :

$$\begin{aligned} E[\mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o) \mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o)'] &= E[\nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' \mathbf{\Omega}_o^{-1} \mathbf{u}_i \mathbf{u}_i' \mathbf{\Omega}_o^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)] \\ &= E\{E[\nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' \mathbf{\Omega}_o^{-1} \mathbf{u}_i \mathbf{u}_i' \mathbf{\Omega}_o^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o) | \mathbf{x}_i]\} \\ &= E[\nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' \mathbf{\Omega}_o^{-1} E(\mathbf{u}_i \mathbf{u}_i' | \mathbf{x}_i) \mathbf{\Omega}_o^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)] \\ &= E[\nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' \mathbf{\Omega}_o^{-1} \mathbf{\Omega}_o \mathbf{\Omega}_o^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)] \\ &= E[\nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' \mathbf{\Omega}_o^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)]. \end{aligned}$$

Next, we have to derive  $\mathbf{A}_o \equiv E[\mathbf{H}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o)]$ , and show that  $\mathbf{B}_o = \mathbf{A}_o$ . The Hessian itself is complicated, but its expected value is not. The Jacobian of  $\mathbf{s}_i(\boldsymbol{\theta}; \boldsymbol{\gamma})$  with respect to  $\boldsymbol{\theta}$  can be written

$$\mathbf{H}_i(\boldsymbol{\theta}; \boldsymbol{\gamma}) = \nabla_{\boldsymbol{\theta}} \mathbf{m}(\mathbf{x}_i, \boldsymbol{\theta})' \mathbf{\Omega}^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}(\mathbf{x}_i, \boldsymbol{\theta}) + [\mathbf{I}_P \otimes \mathbf{u}_i(\boldsymbol{\theta})'] \mathbf{F}(\mathbf{x}_i, \boldsymbol{\theta}; \boldsymbol{\gamma}),$$

where  $\mathbf{F}(\mathbf{x}_i, \boldsymbol{\theta}; \boldsymbol{\gamma})$  is a  $GP \times P$  matrix, where  $P$  is the total number of parameters, that involves Jacobians of the rows  $\mathbf{\Omega}^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta})$  with respect to  $\boldsymbol{\theta}$ . The key is that  $\mathbf{F}(\mathbf{x}_i, \boldsymbol{\theta}; \boldsymbol{\gamma})$  depends on  $\mathbf{x}_i$ , not on  $\mathbf{y}_i$ . So,

$$\begin{aligned} E[\mathbf{H}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o) | \mathbf{x}_i] &= \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' \mathbf{\Omega}_o^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o) + [\mathbf{I}_P \otimes E(\mathbf{u}_i | \mathbf{x}_i)'] \mathbf{F}(\mathbf{x}_i, \boldsymbol{\theta}_o; \boldsymbol{\gamma}_o) \\ &= \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' \mathbf{\Omega}_o^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o). \end{aligned}$$

Now iterated expectations gives  $\mathbf{A}_o = E[\nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' \mathbf{\Omega}_o^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)]$ . We have verified (12.37) and also that  $\mathbf{A}_o = \mathbf{B}_o$ . Therefore, from Theorem 12.3,

$$\text{Avar} \sqrt{N} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = \mathbf{A}_o^{-1} = \{E[\nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' \mathbf{\Omega}_o^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)]\}^{-1}.$$

c. As usual, we replace expectations with sample averages and unknown parameters, and divide the result by  $N$  to get  $\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}})$ :

$$\begin{aligned}\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}) &= \left( N^{-1} \sum_{i=1}^N \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\hat{\boldsymbol{\theta}})' \hat{\boldsymbol{\Omega}}^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\hat{\boldsymbol{\theta}}) \right)^{-1} / N \\ &= \left( \sum_{i=1}^N \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\hat{\boldsymbol{\theta}})' \hat{\boldsymbol{\Omega}}^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\hat{\boldsymbol{\theta}}) \right)^{-1}.\end{aligned}$$

The estimate  $\hat{\boldsymbol{\Omega}}$  can be based on the multivariate NLS residuals or can be updated after the nonlinear SUR estimates have been obtained.

d. First, note that  $\nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)$  is a block-diagonal matrix that has  $G$  rows, with blocks  $\nabla_{\boldsymbol{\theta}_g} m_{ig}(\boldsymbol{\theta}_{og})$ , a  $1 \times P_g$  vector. (I assume that there are no cross-equation restrictions imposed in the nonlinear SUR estimation.) If  $\boldsymbol{\Omega}_o$  is diagonal, so is its inverse. Standard matrix multiplication shows that

$$\nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' \boldsymbol{\Omega}_o^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o) = \begin{pmatrix} \sigma_{o1}^{-2} \nabla_{\boldsymbol{\theta}_1} m_{i1}' \nabla_{\boldsymbol{\theta}_1} m_{i1}^o & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sigma_{o2}^{-2} \nabla_{\boldsymbol{\theta}_2} m_{i2}' \nabla_{\boldsymbol{\theta}_2} m_{i2}^o & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \sigma_{oG}^{-2} \nabla_{\boldsymbol{\theta}_G} m_{iG}' \nabla_{\boldsymbol{\theta}_G} m_{iG}^o \end{pmatrix}.$$

Taking expectations and inverting the result shows that

$\text{Avar} \sqrt{N} (\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_{og}) = \sigma_{og}^2 [E(\nabla_{\boldsymbol{\theta}_g} m_{ig}' \nabla_{\boldsymbol{\theta}_g} m_{ig}^o)]^{-1}$ ,  $g = 1, \dots, G$ . (Note also that the nonlinear SUR estimators are asymptotically uncorrelated across equations.) These asymptotic variances are easily seen to be the same as those for nonlinear least squares on each equation.

e. I cannot see a nonlinear analog of Theorem 7.7. The first hint given in Problem 7.5 does not extend readily to nonlinear models, even when the same regressors appear in each equation. The key is that  $\mathbf{X}_i$  is replaced with  $\nabla_{\boldsymbol{\theta}} \mathbf{m}(\mathbf{x}_i, \boldsymbol{\theta}_o)$ . While this  $G \times P$  matrix has a block-diagonal form, as described in part d, the blocks are not the same even when the same regressors appear in each equation. In the linear case,  $\nabla_{\boldsymbol{\theta}_g} m_g(\mathbf{x}_i, \boldsymbol{\theta}_{og}) = \mathbf{x}_i$  for all  $g$ . But, unless

$\boldsymbol{\theta}_{og}$  is the same in all equations – a very restrictive assumption –  $\nabla_{\boldsymbol{\theta}_g} m_g(\mathbf{x}_i, \boldsymbol{\theta}_{og})$  varies across  $g$ . For example, if  $m_g(\mathbf{x}_i, \boldsymbol{\theta}_{og}) = \exp(\mathbf{x}_i \boldsymbol{\theta}_{og})$  then  $\nabla_{\boldsymbol{\theta}_g} m_g(\mathbf{x}_i, \boldsymbol{\theta}_{og}) = \exp(\mathbf{x}_i \boldsymbol{\theta}_{og}) \mathbf{x}_i$ , and the gradients differ across  $g$ .

**12.8.** As stated in the hint, we can use (12.37) and a modified version of (12.76),

$$N^{-1/2} \sum_{i=1}^N \mathbf{s}_i(\tilde{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}) = N^{-1/2} \sum_{i=1}^N \mathbf{s}_i(\boldsymbol{\theta}_o; \hat{\boldsymbol{\gamma}}) + \mathbf{A}_o \sqrt{N} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) + o_p(1),$$

to show  $\sqrt{N}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) = \mathbf{A}_o^{-1} N^{-1/2} \sum_{i=1}^N \mathbf{s}_i(\tilde{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}) + o_p(1)$ ; this is just standard algebra. Under (12.37),

$$N^{-1/2} \sum_{i=1}^N \mathbf{s}_i(\tilde{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}) = N^{-1/2} \sum_{i=1}^N \mathbf{s}_i(\tilde{\boldsymbol{\theta}}; \boldsymbol{\gamma}_o) = o_p(1),$$

by a similar mean value expansion used for the unconstrained two-step M-estimator:

$$N^{-1/2} \sum_{i=1}^N \mathbf{s}_i(\tilde{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}) = N^{-1/2} \sum_{i=1}^N \mathbf{s}_i(\tilde{\boldsymbol{\theta}}; \boldsymbol{\gamma}_o) + E[\nabla_{\boldsymbol{\gamma}} \mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o)] \sqrt{N} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_o) + o_p(1),$$

and use  $E[\nabla_{\boldsymbol{\gamma}} \mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o)] = \mathbf{0}$ . Now, the second-order Taylor expansion gives

$$\begin{aligned} \sum_{i=1}^N q(\mathbf{w}_i, \tilde{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}) - \sum_{i=1}^N q(\mathbf{w}_i, \hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}) &= \sum_{i=1}^N \mathbf{s}_i(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}) + (1/2)(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})' \left( \sum_{i=1}^N \ddot{\mathbf{H}}_i \right) (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) \\ &= (1/2)(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})' \left( \sum_{i=1}^N \ddot{\mathbf{H}}_i \right) (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}). \end{aligned}$$

Therefore,

$$\begin{aligned} 2 \left( \sum_{i=1}^N q(\mathbf{w}_i, \tilde{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}) - \sum_{i=1}^N q(\mathbf{w}_i, \hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}) \right) &= [\sqrt{N}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})]' \mathbf{A}_o [\sqrt{N}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})] + o_p(1) \\ &= \left( N^{-1/2} \sum_{i=1}^N \tilde{\mathbf{s}}_i \right) \mathbf{A}_o^{-1} \left( N^{-1/2} \sum_{i=1}^N \tilde{\mathbf{s}}_i \right) + o_p(1), \end{aligned}$$

where  $\tilde{\mathbf{s}}_i = \mathbf{s}_i(\tilde{\boldsymbol{\theta}}; \boldsymbol{\gamma}_o)$ . Again, this shows the asymptotic equivalence of the QLR and LM statistics. To complete the problem, we should verify that the *LM* statistic is not affected by  $\hat{\boldsymbol{\gamma}}$  either, but that follows from  $N^{-1/2} \sum_{i=1}^N \mathbf{s}_i(\tilde{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}) = N^{-1/2} \sum_{i=1}^N \mathbf{s}_i(\tilde{\boldsymbol{\theta}}; \boldsymbol{\gamma}_o) + o_p(1)$ .

**12.9.** a. We cannot say anything in general about  $\text{Med}(y|\mathbf{x})$  because

$$\text{Med}(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\beta}_o) + \text{Med}(u|\mathbf{x})$$

and  $\text{Med}(u|\mathbf{x})$  could be a general function of  $\mathbf{x}$ .

b. If  $u$  and  $\mathbf{x}$  are independent, then  $E(u|\mathbf{x})$  and  $\text{Med}(u|\mathbf{x})$  are both constants, say  $\alpha$  and  $\delta$ . Then  $E(y|\mathbf{x}) - \text{Med}(y|\mathbf{x}) = [m(\mathbf{x}, \boldsymbol{\beta}_o) + \alpha] - [m(\mathbf{x}, \boldsymbol{\beta}_o) + \delta] = \alpha - \delta$ , which does not depend on  $\mathbf{x}$ .

c. When  $u$  and  $\mathbf{x}$  are independent, the partial effects of  $x_j$  on the conditional mean and conditional median are the same, and there is no ambiguity about what is “the effect of  $x_j$  on  $y$ ,” at least when only the mean and median are under consideration. In this case, we could interpret large differences between LAD and NLS as perhaps indicating an outlier problem. But it could be just that  $u$  and  $\mathbf{x}$  are not independent, and so the function  $m(\mathbf{x}, \boldsymbol{\beta}_o)$  cannot be both the mean and the median (or differ from each of these by a constant).

**12.10.** The conditional mean function is  $m(\mathbf{x}_i, n_i, \boldsymbol{\beta}) = n_i p(\mathbf{x}_i, \boldsymbol{\beta})$ . So we would, as usual, minimize the sum of squared residuals,  $\sum_{i=1}^N [y_i - n_i p(\mathbf{x}_i, \boldsymbol{\beta})]^2$  with respect to  $\boldsymbol{\beta}$ . This gives the NLS estimator, say  $\check{\boldsymbol{\beta}}$ . Define the weights as  $\hat{h}_i \equiv n_i p(\mathbf{x}_i, \check{\boldsymbol{\beta}})[1 - p(\mathbf{x}_i, \check{\boldsymbol{\beta}})]$ . Then the weighted NLS estimator minimizes  $\sum_{i=1}^N [y_i - n_i p(\mathbf{x}_i, \boldsymbol{\beta})]^2 / \hat{h}_i$ .

**12.11.** a. The key to the derivation is to verify condition (12.37), which is similar to Problem 12.7. In fact, this contains Problem 12.7 as a special case. In particular, write the score (with respect to  $\boldsymbol{\theta}$ ) for observation  $i$  as

$$\begin{aligned}\mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}; \boldsymbol{\gamma}) &= -\nabla_{\boldsymbol{\theta}} \mathbf{m}(\mathbf{x}_i, \boldsymbol{\theta})' [\mathbf{W}(\mathbf{x}_i, \boldsymbol{\gamma})]^{-1} \mathbf{u}_i(\boldsymbol{\theta}) \\ &= -[\mathbf{u}_i(\boldsymbol{\theta}) \otimes \nabla_{\boldsymbol{\theta}} \mathbf{m}(\mathbf{x}_i, \boldsymbol{\theta})]' \text{vec}\{[\mathbf{W}_i(\boldsymbol{\gamma})]^{-1}\}.\end{aligned}$$

The Jacobina of  $\mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}; \boldsymbol{\gamma})$  with respect to  $\boldsymbol{\gamma}$  is generally complicated, but it is clear that  $\nabla_{\boldsymbol{\gamma}} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}; \boldsymbol{\gamma})$  is a linear combination of  $\mathbf{u}_i(\boldsymbol{\theta})$ , where the linear combination is a function of  $\mathbf{x}_i$  (and the parameter values). Therefore, because  $E(\mathbf{u}_i|\mathbf{x}_i) = \mathbf{0}$ ,  $E[\nabla_{\boldsymbol{\gamma}} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o; \boldsymbol{\gamma})|\mathbf{x}_i] = \mathbf{0}$  for any  $\boldsymbol{\gamma}$ , which verifies (12.37). Notice that we do not need to assume  $\text{Var}(\mathbf{u}_i|\mathbf{x}_i) = \mathbf{W}(\mathbf{x}_i, \boldsymbol{\gamma}_o)$  for some  $\boldsymbol{\gamma}_o$ .

Without assuming (12.96) there are no simplifications for

$$\begin{aligned}\mathbf{B}_o &\equiv E[\mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}^*) \mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}^*)'] \\ &= E\{\nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' [\mathbf{W}_i(\boldsymbol{\gamma}^*)]^{-1} \mathbf{u}_i \mathbf{u}_i' [\mathbf{W}_i(\boldsymbol{\gamma}^*)]^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)\}\end{aligned}$$

where  $\boldsymbol{\gamma}^* = \text{plim}(\hat{\boldsymbol{\gamma}})$ . A consistent estimator of  $\mathbf{B}_o$  is

$$\hat{\mathbf{B}} = N^{-1} \sum_{i=1}^N \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\hat{\boldsymbol{\theta}})' [\mathbf{W}_i(\hat{\boldsymbol{\gamma}})]^{-1} \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' [\mathbf{W}_i(\hat{\boldsymbol{\gamma}})]^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\hat{\boldsymbol{\theta}})$$

where  $\hat{\boldsymbol{\theta}}$  is the WMNLS estimator.

We also need to consistently estimate  $\mathbf{A}_o \equiv E[\mathbf{H}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o)]$ . Again, the argument is similar to that in Problem 12.7, and uses that the mean function is correctly specified. We can write the Hessian as

$$\mathbf{H}_i(\boldsymbol{\theta}; \boldsymbol{\gamma}) = \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta})' [\mathbf{W}_i(\boldsymbol{\gamma})]^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}) + [\mathbf{I}_P \otimes \mathbf{u}_i(\boldsymbol{\theta})'] \mathbf{F}(\mathbf{x}_i, \boldsymbol{\theta}; \boldsymbol{\gamma}),$$

where  $\mathbf{F}(\mathbf{x}_i, \boldsymbol{\theta}; \boldsymbol{\gamma})$  is a  $GP \times P$  matrix, where  $P$  is the total number of parameters, that involves Jacobians of the rows  $[\mathbf{W}_i(\boldsymbol{\gamma})]^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta})$  with respect to  $\boldsymbol{\theta}$ . Therefore,

$$\begin{aligned}E[\mathbf{H}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}^*)|\mathbf{x}_i] &= \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' [\mathbf{W}_i(\boldsymbol{\gamma}^*)]^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o) + [\mathbf{I}_P \otimes E(\mathbf{u}_i|\mathbf{x}_i)'] \mathbf{F}(\mathbf{x}_i, \boldsymbol{\theta}_o; \boldsymbol{\gamma}^*) \\ &= \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' [\mathbf{W}_i(\boldsymbol{\gamma}^*)]^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o),\end{aligned}$$



and so  $\mathbf{A}_o = E[\nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' [\mathbf{W}_i(\boldsymbol{\gamma}^*)]^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)]$ . A consistent estimator of  $\mathbf{A}_o$  is

$$\hat{\mathbf{A}} = N^{-1} \sum_{i=1}^N \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\hat{\boldsymbol{\theta}})' [\mathbf{W}_i(\hat{\boldsymbol{\gamma}})]^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\hat{\boldsymbol{\theta}}).$$

When we form

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{-1} / N,$$

simple algebra shows this expression is the same as (12.98).

b. If we assume (12.96) then

$$\begin{aligned} \mathbf{B}_o &= E(E\{\nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' [\mathbf{W}_i(\boldsymbol{\gamma}_o)]^{-1} \mathbf{u}_i \mathbf{u}_i' [\mathbf{W}_i(\boldsymbol{\gamma}_o)]^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)\} | \mathbf{x}_i) \\ &= E\{\nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' [\mathbf{W}_i(\boldsymbol{\gamma}_o)]^{-1} E(\mathbf{u}_i \mathbf{u}_i' | \mathbf{x}_i) [\mathbf{W}_i(\boldsymbol{\gamma}_o)]^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)\} | \mathbf{x}_i) \\ &= E\{\nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)' [\mathbf{W}_i(\boldsymbol{\gamma}_o)]^{-1} \mathbf{W}_i(\boldsymbol{\gamma}_o) [\mathbf{W}_i(\boldsymbol{\gamma}_o)]^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{m}_i(\boldsymbol{\theta}_o)\} \\ &= \mathbf{A}_o \end{aligned}$$

and so

$$\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)] = \mathbf{A}_o^{-1}.$$

c. We can apply Problem 12.8 once we have properly chosen the objective function to ensure  $\mathbf{B}_o = \mathbf{A}_o$  when (12.96) holds. That objective function, with nuisance parameters  $\boldsymbol{\gamma}$ , is

$$q(\mathbf{w}_i, \boldsymbol{\theta}; \boldsymbol{\gamma}) = (1/2)[\mathbf{y}_i - \mathbf{m}(\mathbf{x}_i, \boldsymbol{\theta})]' [\mathbf{W}_i(\boldsymbol{\gamma})]^{-1} [\mathbf{y}_i - \mathbf{m}(\mathbf{x}_i, \boldsymbol{\theta})]$$

The division by two ensures

$$E[\mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o) \mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o)'] = E[\mathbf{H}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o)]$$

Now, letting  $\tilde{\boldsymbol{\theta}}$  and  $\hat{\boldsymbol{\theta}}$  be the restricted and unrestricted estimators, respectively – where both use  $\hat{\boldsymbol{\gamma}}$  as the nuisance parameter estimator – the QLR statistic is

$$\begin{aligned}
QLR &= 2 \left( \sum_{i=1}^N q(\mathbf{w}_i, \tilde{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}) - \sum_{i=1}^N q(\mathbf{w}_i, \hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}) \right) \\
&= \sum_{i=1}^N \tilde{\mathbf{u}}_i' \hat{\mathbf{W}}_i^{-1} \tilde{\mathbf{u}}_i - \sum_{i=1}^N \hat{\mathbf{u}}_i' \hat{\mathbf{W}}_i^{-1} \hat{\mathbf{u}}_i
\end{aligned}$$

where  $\hat{\mathbf{W}}_i \equiv \mathbf{W}_i(\hat{\boldsymbol{\gamma}})$ ,  $\tilde{\mathbf{u}}_i \equiv \mathbf{y}_i - \mathbf{m}(\mathbf{x}_i, \tilde{\boldsymbol{\theta}})$ , and  $\hat{\mathbf{u}}_i \equiv \mathbf{y}_i - \mathbf{m}(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$ . Under  $H_0$  and standard regularity conditions,  $QLR \xrightarrow{d} \chi_Q^2$ , where  $Q$  is the number of restrictions.

And  $F$ -type statistic is obtained as

$$F = \frac{\left( \sum_{i=1}^N \tilde{\mathbf{u}}_i' \hat{\mathbf{W}}_i^{-1} \tilde{\mathbf{u}}_i - \sum_{i=1}^N \hat{\mathbf{u}}_i' \hat{\mathbf{W}}_i^{-1} \hat{\mathbf{u}}_i \right) / Q}{\left( \sum_{i=1}^N \hat{\mathbf{u}}_i' \hat{\mathbf{W}}_i^{-1} \hat{\mathbf{u}}_i \right) / (NG - P)}$$

which can be treated as an approximate  $\mathcal{F}_{Q, NG-P}$  random variable. Note that under (12.96)

$$\begin{aligned}
E[\mathbf{u}_i' \mathbf{W}_i(\boldsymbol{\gamma}_o)^{-1} \mathbf{u}_i] &= E\{E[\mathbf{u}_i' \mathbf{W}_i(\boldsymbol{\gamma}_o)^{-1} \mathbf{u}_i | \mathbf{x}_i]\} \\
&= E\{\text{tr } E[\mathbf{W}_i(\boldsymbol{\gamma}_o)^{-1} \mathbf{u}_i \mathbf{u}_i' | \mathbf{x}_i]\} = G
\end{aligned}$$

because  $E(\mathbf{u}_i \mathbf{u}_i' | \mathbf{x}_i) = \mathbf{W}_i(\boldsymbol{\gamma}_o)$ . Therefore,

$$(NG)^{-1} \sum_{i=1}^N \hat{\mathbf{u}}_i' \hat{\mathbf{W}}_i^{-1} \hat{\mathbf{u}}_i \xrightarrow{p} 1$$

and using  $NG - P$  is a degrees-of-freedom adjustment.

**12.12.** a. We can appeal to equation (12.41) and the discussion that follows about the scores for the two problems being uncorrelated. We have

$$\mathbf{s}_i(\boldsymbol{\theta}; \boldsymbol{\delta}) = -\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \mathbf{v}(\mathbf{w}_i, \boldsymbol{\delta}), \boldsymbol{\theta})' [y_i - m(\mathbf{x}_i, \mathbf{v}(\mathbf{w}_i, \boldsymbol{\delta}), \boldsymbol{\theta})]$$

and we know, because  $E(y_i | \mathbf{x}_i, \mathbf{w}_i) = m(\mathbf{x}_i, \mathbf{v}(\mathbf{w}_i, \boldsymbol{\delta}_o), \boldsymbol{\theta}_o)$ ,

$$E[\mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\delta}_o) | \mathbf{x}_i, \mathbf{w}_i] = \mathbf{0}.$$

As usual, this means any function of  $(\mathbf{x}_i, \mathbf{w}_i)$  is uncorrelated with  $\mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\delta}_o)$ , including  $\mathbf{r}(\mathbf{w}_i, \boldsymbol{\delta}_o)$ .

It follows that

$$\mathbf{D}_o = \mathbf{B}_o + \mathbf{F}_o \mathbf{E}[\mathbf{r}(\mathbf{w}_i, \boldsymbol{\delta}_o) \mathbf{r}(\mathbf{w}_i, \boldsymbol{\delta}_o)'] \mathbf{F}_o'$$

where

$$\begin{aligned} \mathbf{B}_o &= \mathbf{E}[\mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\delta}_o) \mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\delta}_o)'] \\ \mathbf{F}_o &= \mathbf{E}[\nabla_{\boldsymbol{\delta}} \mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\delta}_o)]. \end{aligned}$$

the matrix  $\mathbf{F}_o \mathbf{E}[\mathbf{r}(\mathbf{w}_i, \boldsymbol{\delta}_o) \mathbf{r}(\mathbf{w}_i, \boldsymbol{\delta}_o)'] \mathbf{F}_o'$  is at least p.s.d., and so  $\mathbf{D}_o - \mathbf{B}_o$  is p.s.d. The asymptotic variance of the two-step estimator of  $\boldsymbol{\theta}_o$  (standardized by  $\sqrt{N}$ ) is

$$\mathbf{A}_o^{-1} \mathbf{D}_o \mathbf{A}_o^{-1}$$

and that of the estimator where  $\boldsymbol{\delta}_o$  is known is

$$\mathbf{A}_o^{-1} \mathbf{B}_o \mathbf{A}_o^{-1}.$$

b. To estimate  $\mathbf{A}_o$  under correct specification of the mean it is convenient to use

$$\mathbf{A}_o = \mathbf{E}[\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \mathbf{v}(\mathbf{w}_i, \boldsymbol{\delta}_o), \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \mathbf{v}(\mathbf{w}_i, \boldsymbol{\delta}_o), \boldsymbol{\theta}_o)]$$

and so

$$\hat{\mathbf{A}} = N^{-1} \sum_{i=1}^N \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \mathbf{v}(\mathbf{w}_i, \hat{\boldsymbol{\delta}}), \hat{\boldsymbol{\theta}})' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \mathbf{v}(\mathbf{w}_i, \hat{\boldsymbol{\delta}}), \hat{\boldsymbol{\theta}})]$$

Further,

$$\hat{\mathbf{B}} = N^{-1} \sum_{i=1}^N \mathbf{s}_i(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\delta}}) \mathbf{s}_i(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\delta}})'].$$

It remains to consistently estimate  $\mathbf{F}_o$ . But by the product and chain rules,

$$\begin{aligned} \nabla_{\boldsymbol{\delta}} \mathbf{s}_i(\boldsymbol{\theta}; \boldsymbol{\delta}) &= -\nabla_{\boldsymbol{\delta}} \{ \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \mathbf{v}(\mathbf{w}_i, \boldsymbol{\delta}), \boldsymbol{\theta})' \} [y_i - m(\mathbf{x}_i, \mathbf{v}(\mathbf{w}_i, \boldsymbol{\delta}), \boldsymbol{\theta})] \\ &\quad + \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \mathbf{v}(\mathbf{w}_i, \boldsymbol{\delta}), \boldsymbol{\theta})' \nabla_{\mathbf{v}} m(\mathbf{x}_i, \mathbf{v}(\mathbf{w}_i, \boldsymbol{\delta}), \boldsymbol{\theta}) \nabla_{\boldsymbol{\delta}} \mathbf{v}(\mathbf{w}_i, \boldsymbol{\delta}). \end{aligned}$$

When we plug in  $(\boldsymbol{\theta}_o, \boldsymbol{\delta}_o)$  the first term has zero mean because the conditional mean is

correctly specified – much like the argument for the Hessian. Therefore,

$$\begin{aligned}\mathbf{F}_o &= E[\nabla_{\delta} \mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\delta}_o)] \\ &= E[\nabla_{\theta} m(\mathbf{x}_i, \mathbf{v}(\mathbf{w}_i, \boldsymbol{\delta}_o), \boldsymbol{\theta}_o)' \nabla_{\mathbf{v}} m(\mathbf{x}_i, \mathbf{v}(\mathbf{w}_i, \boldsymbol{\delta}_o), \boldsymbol{\theta}_o) \nabla_{\delta} \mathbf{v}(\mathbf{w}_i, \boldsymbol{\delta}_o)]\end{aligned}$$

and

$$\hat{\mathbf{F}} = N^{-1} \sum_{i=1}^N \nabla_{\theta} m(\mathbf{x}_i, \mathbf{v}(\mathbf{w}_i, \hat{\boldsymbol{\delta}}), \hat{\boldsymbol{\theta}})' \nabla_{\mathbf{v}} m(\mathbf{x}_i, \mathbf{v}(\mathbf{w}_i, \hat{\boldsymbol{\delta}}), \hat{\boldsymbol{\theta}}) \nabla_{\delta} \mathbf{v}(\mathbf{w}_i, \hat{\boldsymbol{\delta}})$$

is consistent for  $\mathbf{F}_o$ . Finally, let

$$\hat{\mathbf{C}} = N^{-1} \sum_{i=1}^N \mathbf{r}_i(\hat{\boldsymbol{\delta}}) \mathbf{r}_i(\hat{\boldsymbol{\delta}})'$$

Then

$$\widehat{\text{Avar}}[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)] = \hat{\mathbf{A}}^{-1}(\hat{\mathbf{B}} + \hat{\mathbf{F}}\hat{\mathbf{C}}\hat{\mathbf{F}}')\hat{\mathbf{A}}^{-1}$$

which, numerically, will always be larger (in the matrix sense) than  $\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}$ .

**12.13.** a. Strict exogeneity is not needed because the population objective function is

$$(1/2) \sum_{t=1}^T E\{[y_{it} - m(\mathbf{x}_{it}, \boldsymbol{\theta})]^2 / h(\mathbf{x}_{it}, \boldsymbol{\gamma}^*)\},$$

and  $\boldsymbol{\theta}_o$  minimizes this function provided

$$E(y_{it} | \mathbf{x}_{it}) = m(\mathbf{x}_{it}, \boldsymbol{\theta}_o), \quad t = 1, \dots, T.$$

We do not need

$$E(y_{it} | \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT}) = m(\mathbf{x}_{it}, \boldsymbol{\theta}_o).$$

The proof of the claim that  $\boldsymbol{\theta}_o$  is a minimizer of the population objective function could use the score – assuming that  $m(\mathbf{x}_t, \cdot)$  is continuously differentiable and  $\boldsymbol{\theta}_o \in \text{int}(\boldsymbol{\Theta})$  – and follow Problem 12.4. But we can show directly that, for each  $t = 1, \dots, T$ ,

$$E\{[y_{it} - m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)]^2/h(\mathbf{x}_{it}, \boldsymbol{\gamma}^*)\} \leq E\{[y_{it} - m(\mathbf{x}_{it}, \boldsymbol{\theta})]^2/h(\mathbf{x}_{it}, \boldsymbol{\gamma}^*)\}, \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

and then inequality clearly holds when we sum over  $t$ . Identification requires that strict inequality holds for  $\boldsymbol{\theta} \neq \boldsymbol{\theta}_o$  when we sum across  $t$ .

To establish the above inequality, we follow Problem 12.1. Applied to a given  $t$ , we have

$$E\{[y_{it} - m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)]^2|\mathbf{x}_{it}\} \leq E\{[y_{it} - m(\mathbf{x}_{it}, \boldsymbol{\theta})]^2|\mathbf{x}_{it}\}$$

for any  $\mathbf{x}_{it}$ . Because  $h(\mathbf{x}_{it}, \boldsymbol{\gamma}^*) > 0$  the inequality continues to hold if we divide each side by  $h(\mathbf{x}_{it}, \boldsymbol{\gamma}^*)$ . Further, because  $h(\mathbf{x}_{it}, \boldsymbol{\gamma}^*)$  is a function of  $\mathbf{x}_{it}$ , we can bring it inside both conditional expectations:

$$E\left\{\frac{[y_{it} - m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)]^2}{h(\mathbf{x}_{it}, \boldsymbol{\gamma}^*)} \middle| \mathbf{x}_{it}\right\} \leq E\left\{\frac{[y_{it} - m(\mathbf{x}_{it}, \boldsymbol{\theta})]^2}{h(\mathbf{x}_{it}, \boldsymbol{\gamma}^*)} \middle| \mathbf{x}_{it}\right\}$$

and then take the expected value with respect to  $\mathbf{x}_{it}$  on both sides.

b. For each  $t$  we can use the same argument for WNLS on a single cross section to show

$$E[\nabla_{\boldsymbol{\gamma}} \mathbf{s}_{it}(\boldsymbol{\theta}_o; \boldsymbol{\gamma})] = \mathbf{0}$$

for any  $\boldsymbol{\gamma}$ , where

$$\mathbf{s}_{it}(\boldsymbol{\theta}; \boldsymbol{\gamma}) = -\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta})' u_{it}(\boldsymbol{\theta})/h(\mathbf{x}_{it}, \boldsymbol{\gamma})$$

Because

$$\mathbf{s}_i(\boldsymbol{\theta}; \boldsymbol{\gamma}) = \sum_{t=1}^T \mathbf{s}_{it}(\boldsymbol{\theta}; \boldsymbol{\gamma})$$

it follows that condition (12.37) holds, so we can ignore the estimation of  $\boldsymbol{\gamma}^*$  in obtaining

$\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)]$ . But then we just need to estimate

$$\mathbf{B}_o = E[\mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}^*) \mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}^*)']$$

$$\mathbf{A}_o = \sum_{t=1}^T E[\mathbf{H}_{it}(\boldsymbol{\theta}_o; \boldsymbol{\gamma}^*)] = \sum_{t=1}^T E\{u_{it}^2 \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o) / [h(\mathbf{x}_{it}, \boldsymbol{\gamma})]^2\}$$

where  $u_{it} = y_{it} - m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)$ . Consistent estimators are

$$\begin{aligned} \hat{\mathbf{B}} &= N^{-1} \sum_{i=1}^N \mathbf{s}_i(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}) \mathbf{s}_i(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}})' = N^{-1} \sum_{i=1}^N \left[ \sum_{t=1}^T \mathbf{s}_{it}(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}) \right] \left[ \sum_{t=1}^T \mathbf{s}_{it}(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}) \right]' \\ &= N^{-1} \sum_{i=1}^N \left[ \sum_{t=1}^T \hat{u}_{it} \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \hat{\boldsymbol{\theta}})' / h(\mathbf{x}_{it}, \hat{\boldsymbol{\gamma}}) \right] \left[ \sum_{t=1}^T \hat{u}_{it} \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \hat{\boldsymbol{\theta}}) / h(\mathbf{x}_{it}, \hat{\boldsymbol{\gamma}}) \right] \end{aligned}$$

and

$$\hat{\mathbf{A}} = N^{-1} \sum_{i=1}^N \sum_{t=1}^T \hat{u}_{it}^2 \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \hat{\boldsymbol{\theta}})' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \hat{\boldsymbol{\theta}}) / [h(\mathbf{x}_{it}, \hat{\boldsymbol{\gamma}})]^2.$$

Notice how  $\hat{\mathbf{B}}$  includes terms involving  $\hat{u}_{it}\hat{u}_{ir}$  for  $t \neq r$ , thereby allowing for serial correlation.

Further, terms involving  $\hat{u}_{it}^2/[h(\mathbf{x}_{it}, \hat{\boldsymbol{\gamma}})]^2$  mean we are not assuming the variance function is correctly specified.

c. For any  $\boldsymbol{\gamma}$  we can write

$$\mathbf{s}_{it}(\boldsymbol{\theta}_o; \boldsymbol{\gamma}) \mathbf{s}_{ir}(\boldsymbol{\theta}_o; \boldsymbol{\gamma})' = u_{it} u_{ir} \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{ir}, \boldsymbol{\theta}_o) / [h(\mathbf{x}_{it}, \boldsymbol{\gamma}) h(\mathbf{x}_{ir}, \boldsymbol{\gamma})].$$

Take  $r < t$ . Then, by dynamic completeness – that is,  $E(y_{it} | \mathbf{x}_{it}, y_{i,t-1}, \mathbf{x}_{i,t-1}, \dots, y_{i1}, \mathbf{x}_{i1}) = 0 -$

$E(u_{it} | u_{ir}, \mathbf{x}_{it}, \mathbf{x}_{ir}) = 0$ , and so

$$E[\mathbf{s}_{it}(\boldsymbol{\theta}_o; \boldsymbol{\gamma}) | u_{ir}, \mathbf{x}_{it}, \mathbf{x}_{ir}] = \mathbf{0}.$$

Therefore,

$$E[\mathbf{s}_{it}(\boldsymbol{\theta}_o; \boldsymbol{\gamma}) \mathbf{s}_{ir}(\boldsymbol{\theta}_o; \boldsymbol{\gamma})' | u_{ir}, \mathbf{x}_{it}, \mathbf{x}_{ir}] = \mathbf{0}$$

and so  $E[\mathbf{s}_{it}(\boldsymbol{\theta}_o; \boldsymbol{\gamma}) \mathbf{s}_{ir}(\boldsymbol{\theta}_o; \boldsymbol{\gamma})'] = \mathbf{0}$ . It follows that we need not estimate the terms

$E[\mathbf{s}_{it}(\boldsymbol{\theta}_o; \boldsymbol{\gamma})\mathbf{s}_{it}'(\boldsymbol{\theta}_o; \boldsymbol{\gamma})]$ , and so a consistent estimator of  $\mathbf{B}_o$  is

$$N^{-1} \sum_{i=1}^N \sum_{t=1}^T \hat{u}_{it}^2 \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \hat{\boldsymbol{\theta}})' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \hat{\boldsymbol{\theta}}) / [h(\mathbf{x}_{it}, \hat{\boldsymbol{\gamma}})]^2,$$

and we need not change  $\hat{\mathbf{A}}$  because the conditional mean for each  $t$  is assumed to be correctly specified.

Remember that, because our analysis is for fixed  $T$  and  $N \rightarrow \infty$ , and we are using the usual  $\sqrt{N}$ -limiting distribution, there is nothing wrong with using the fully robust form even under dynamic completeness. There is a sense that imposing zero correlation in the scores when they are uncorrelated leads to better finite-sample inference, but that is difficult to establish in any generality.

d. Again, we can keep  $\hat{\mathbf{A}}$  the same. For  $\hat{\mathbf{B}}$  we can use either of the estimators in parts b or c. But if we want to use both dynamic completeness and a correctly specified conditional variance, we can simplify  $\hat{\mathbf{B}}$  even further.

$$\begin{aligned} \mathbf{B}_o &= \sum_{t=1}^T E\{u_{it}^2 \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o) / [h(\mathbf{x}_{it}, \boldsymbol{\gamma}_o)]^2\} \\ &= \sum_{t=1}^T E(E\{u_{it}^2 \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o) / [h(\mathbf{x}_{it}, \boldsymbol{\gamma}_o)]^2 | \mathbf{x}_{it}\}) \\ &= \sum_{t=1}^T E\{[\sigma_o^2 h(\mathbf{x}_{it}, \boldsymbol{\gamma}_o)] \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o) / [h(\mathbf{x}_{it}, \boldsymbol{\gamma}_o)]^2\} \\ &= \sigma_o^2 \sum_{t=1}^T E\{\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_{it}, \boldsymbol{\theta}_o) / [h(\mathbf{x}_{it}, \boldsymbol{\gamma}_o)]\} = \sigma_o^2 \mathbf{A}_o. \end{aligned}$$

So

$$\widehat{\text{Avar}}[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)] = \hat{\sigma}^2 \hat{\mathbf{A}}^{-1}$$

where

$$\hat{\sigma}^2 = (NT - P)^{-1} \sum_{i=1}^N \sum_{t=1}^T \hat{u}_{it}^2 / h(\mathbf{x}_{it}, \hat{\gamma})$$

is easily shown to be consistent for  $\sigma_o^2$ : by iterated expectations,

$$E[u_{it}^2 / h(\mathbf{x}_{it}, \gamma_o)] = \sigma_o^2, t = 1, \dots, T.$$

**12.14.** Write the score evaluated at  $\theta_o$  as

$$\begin{aligned} \mathbf{s}_i(\theta_o) &= -\mathbf{x}_i' \{ \tau 1[y_i - \mathbf{x}_i \theta_o \geq 0] - (1 - \tau) 1[y_i - \mathbf{x}_i \theta_o < 0] \} \\ &= -\mathbf{x}_i' \{ \tau 1[u_i \geq 0] - (1 - \tau) 1[u_i < 0] \} \end{aligned}$$

where  $u_i \equiv y_i - \mathbf{x}_i \theta_o$ . Therefore,

$$\begin{aligned} \mathbf{s}_i(\theta_o) \mathbf{s}_i(\theta_o)' &= \{ \tau 1[u_i \geq 0] - (1 - \tau) 1[u_i < 0] \}^2 \mathbf{x}_i' \mathbf{x}_i \\ &= (\tau^2 1[u_i \geq 0] + (1 - \tau)^2 1[u_i < 0]) \mathbf{x}_i' \mathbf{x}_i \end{aligned}$$

where this expression uses the hint that  $1[u_i \geq 0] \cdot 1[u_i < 0] = 0$  and the square of an indicator function is just itself.

Now take the expectation conditional on  $\mathbf{x}_i$ :

$$\begin{aligned} E[\mathbf{s}_i(\theta_o) \mathbf{s}_i(\theta_o)' | \mathbf{x}_i] &= \{ \tau^2 E(1[u_i \geq 0] | \mathbf{x}_i) + (1 - \tau)^2 E(1[u_i < 0] | \mathbf{x}_i) \} \mathbf{x}_i' \mathbf{x}_i \\ &= [\tau^2 (1 - \tau) + (1 - \tau)^2 \tau] \mathbf{x}_i' \mathbf{x}_i \\ &= \tau(1 - \tau) \mathbf{x}_i' \mathbf{x}_i, \end{aligned}$$

where we use the fact that  $E(1[u_i < 0] | \mathbf{x}_i) = P(y_i < \mathbf{x}_i \theta_o | \mathbf{x}_i) = \tau$  – see the discussion below equation (12.110). Now apply iterated expectations to get (12.115).

**12.15.** a.  $\hat{\theta}$  (approximately) solves the first order condition

$$\sum_{i=1}^N \sum_{t=1}^T -\mathbf{x}_{it}' \{ \tau 1[y_{it} - \mathbf{x}_{it} \hat{\theta} \geq 0] - (1 - \tau) 1[y_{it} - \mathbf{x}_{it} \hat{\theta} < 0] \} = \mathbf{0},$$

so the score function for time period  $t$  is



$$\mathbf{s}_{it}(\boldsymbol{\theta}) = -\mathbf{x}_{it}'\{\tau 1[y_{it} - \mathbf{x}_{it}\boldsymbol{\theta} \geq 0] - (1 - \tau)1[y_{it} - \mathbf{x}_{it}\boldsymbol{\theta} < 0]\}$$

and the score for random draw  $i$  is

$$\mathbf{s}_i(\boldsymbol{\theta}) = \sum_{t=1}^T \mathbf{s}_{it}(\boldsymbol{\theta}).$$

b. We have to show that the scores  $\{\mathbf{s}_{it}(\boldsymbol{\theta}_o) : t = 1, \dots, T\}$  in part a are serially uncorrelated. Now

$$\mathbf{s}_{it}(\boldsymbol{\theta}_o) = -\mathbf{x}_{it}'\{\tau 1[u_{it} \geq 0] - (1 - \tau)1[u_{it} < 0]\}$$

and, under dynamic completeness of the quantile,

$$\begin{aligned} E\{\tau 1[u_{it} \geq 0] - (1 - \tau)1[u_{it} < 0] | \mathbf{x}_{it}, y_{i,t-1}, \dots, y_{i1}, \mathbf{x}_{i1}\} \\ = E\{\tau 1[u_{it} \geq 0] - (1 - \tau)1[u_{it} < 0] | \mathbf{x}_{it}\} \\ = \tau(1 - \tau) - (1 - \tau)\tau = 0. \end{aligned}$$

Therefore,

$$E[\mathbf{s}_{it}(\boldsymbol{\theta}_o) | \mathbf{x}_{it}, y_{i,t-1}, \dots, y_{i1}, \mathbf{x}_{i1}] = -\mathbf{x}_{it}' E[\tau 1[u_{it} \geq 0] - (1 - \tau)1[u_{it} < 0] | \mathbf{x}_{it}, y_{i,t-1}, \dots, y_{i1}, \mathbf{x}_{i1}] = \mathbf{0},$$

and it follows that if  $r < t$ ,  $\mathbf{s}_{ir}(\boldsymbol{\theta}_o)$  is uncorrelated with  $\mathbf{s}_{it}(\boldsymbol{\theta}_o)$ . Therefore,

$$\mathbf{B}_o = \sum_{t=1}^T E[\mathbf{s}_{it}(\boldsymbol{\theta}_o)\mathbf{s}_{it}(\boldsymbol{\theta}_o)'] = \tau(1 - \tau) \sum_{t=1}^T E(\mathbf{x}_{it}'\mathbf{x}_{it})$$

and a consistent estimator is

$$\hat{\mathbf{B}} = \tau(1 - \tau)N^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathbf{x}_{it}'\mathbf{x}_{it}$$

c. This follows in a way similar to the to the cross section case. Now

$$\mathbf{a}(\boldsymbol{\theta}) = \sum_{i=1}^T E[\mathbf{s}_{it}(\boldsymbol{\theta})]$$

and we need its Jacobian. We use  $E[\mathbf{s}_{it}(\boldsymbol{\theta})] = E\{E[\mathbf{s}_{it}(\boldsymbol{\theta})|\mathbf{x}_{it}]\}$  for each  $t$ , and then, just as in Section 12.10.2,

$$\nabla_{\boldsymbol{\theta}} E[\mathbf{s}_{it}(\boldsymbol{\theta})|\mathbf{x}_{it}] = f_{u_t}(\mathbf{x}_{it}(\boldsymbol{\theta} - \boldsymbol{\theta}_o)|\mathbf{x}_{it})\mathbf{x}_{it}'\mathbf{x}_{it}$$

Then

$$\mathbf{A}_o = \sum_{t=1}^T E\{\nabla_{\boldsymbol{\theta}} E[\mathbf{s}_{it}(\boldsymbol{\theta}_o)|\mathbf{x}_{it}]\} = \sum_{t=1}^T E[f_{u_t}(0|\mathbf{x}_{it})\mathbf{x}_{it}'\mathbf{x}_{it}].$$

**12.16.** a. Because  $\text{Med}(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2$ , we would use LAD.

b. From

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1$$

we have

$$\begin{aligned} \text{Med}(y_1|y_2, \mathbf{z}) &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \text{Med}(u_1|y_2, \mathbf{z}) \\ &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 v_2 \\ &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 (y_2 - \mathbf{z}\boldsymbol{\pi}_2) \end{aligned}$$

We can use a control function approach but based on LAD. So, in the first stage, estimate  $\boldsymbol{\pi}_2$  by LAD and compute, for each  $i$ ,

$$\hat{v}_{i2} = y_{i2} - \mathbf{z}_i\hat{\boldsymbol{\pi}}_2.$$

Then use LAD in the second stage. Using dummy arguments of optimization,

$$\min_{\mathbf{d}_1, \alpha_1, r_1} \sum_{i=1}^N |y_{i1} - \mathbf{z}_{i1}\mathbf{d}_1 - \alpha_1 y_{i2} - r_1 \hat{v}_{i2}|$$

to get  $\hat{\boldsymbol{\delta}}_1$ ,  $\hat{\alpha}_1$ , and  $\hat{\rho}_1$ . These estimators are generally consistent by the two-step estimation result discussed in Section 12.4.1.

c. It is natural to use the LAD  $t$  statistic for  $\hat{\rho}_1$  from the control function procedure in part

b. We know from Chapter 6 that if we were using OLS in both stages then we could ignore the first-stage estimation of  $\pi_2$  under  $H_0 : \rho_1 = 0$ . That seems very likely the case here, too, but it does not follow from the results presented in the text (which assume smooth objective functions with nonsingular expected Hessians).

d. As mentioned in part c, an analytical calculation requires an extended set of tools, such as those in Newey and McFadden (1994). A computationally intensive solution is to bootstrap the two-step estimation method (being sure to recompute  $\hat{\pi}_2$  with every bootstrap sample in order to account for its sampling distribution).

**12.17.** a. We use a mean value expansion, similar to the delta method from Chapter 3 but now allowing for the randomness of  $\mathbf{w}_i$ . By a mean value expansion, we can write

$$N^{-1/2} \sum_{i=1}^N \mathbf{g}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = N^{-1/2} \sum_{i=1}^N \mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o) + \left( N^{-1} \sum_{i=1}^N \ddot{\mathbf{G}}_i \right) \sqrt{N} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o),$$

where  $\ddot{\mathbf{G}}_i$  is the  $M \times P$  Jacobian of  $\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta})$  evaluated at mean values between  $\boldsymbol{\theta}_o$  and  $\hat{\boldsymbol{\theta}}$ . Now, because  $\sqrt{N} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \stackrel{a}{\sim} \text{Normal}(\mathbf{0}, \mathbf{A}_o^{-1} \mathbf{B}_o \mathbf{A}_o^{-1})$ , it follows that  $\sqrt{N} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = O_p(1)$ . Further, by Lemma 12.1,  $N^{-1} \sum_{i=1}^N \ddot{\mathbf{G}}_i \xrightarrow{p} E[\nabla_{\boldsymbol{\theta}} \mathbf{g}(\mathbf{w}, \boldsymbol{\theta}_o)] = \mathbf{G}_o$  (the mean values all converge in probability to  $\boldsymbol{\theta}_o$ ). Therefore,

$$\left( N^{-1} \sum_{i=1}^N \ddot{\mathbf{G}}_i \right) \sqrt{N} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = \mathbf{G}_o \sqrt{N} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) + o_p(1),$$

and so

$$N^{-1/2} \sum_{i=1}^N \mathbf{g}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = N^{-1/2} \sum_{i=1}^N \mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o) + \mathbf{G}_o \sqrt{N} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) + o_p(1).$$

Because  $\sqrt{N} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = -N^{-1/2} \sum_{i=1}^N \mathbf{A}_o^{-1} \mathbf{s}_i(\boldsymbol{\theta}_o) = o_p(1)$ , we can write

$$\sqrt{N}\hat{\boldsymbol{\delta}} = N^{-1/2} \sum_{i=1}^N \mathbf{g}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = N^{-1/2} \sum_{i=1}^N [\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o) - \mathbf{G}_o \mathbf{A}_o^{-1} \mathbf{s}_i(\boldsymbol{\theta}_o)] + o_p(1)$$

or, subtracting  $\sqrt{N}\boldsymbol{\delta}_o$  from both sides,

$$\sqrt{N}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_o) = N^{-1/2} \sum_{i=1}^N [\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o) - \boldsymbol{\delta}_o - \mathbf{G}_o \mathbf{A}_o^{-1} \mathbf{s}_i(\boldsymbol{\theta}_o)] + o_p(1).$$

Now

$$\begin{aligned} E[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o) - \boldsymbol{\delta}_o - \mathbf{G}_o \mathbf{A}_o^{-1} \mathbf{s}_i(\boldsymbol{\theta}_o)] &= E[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o)] - \boldsymbol{\delta}_o - \mathbf{G}_o \mathbf{A}_o^{-1} E[\mathbf{s}_i(\boldsymbol{\theta}_o)] \\ &= \boldsymbol{\delta}_o - \boldsymbol{\delta}_o = \mathbf{0}. \end{aligned}$$

Therefore, by the CLT for i.i.d. sequences,

$$\sqrt{N}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_o) \stackrel{a}{\sim} \text{Normal}(\mathbf{0}, \mathbf{D}_o)$$

where

$$\mathbf{D}_o = \text{Var}(\mathbf{g}_i - \boldsymbol{\delta}_o - \mathbf{G}_o \mathbf{A}_o^{-1} \mathbf{s}_i),$$

where hopefully the shorthand is clear. This differs from the usual delta method result because the randomness in  $\mathbf{g}_i = \mathbf{g}_i(\boldsymbol{\theta}_o)$  must be accounted for.

b. We assume we have  $\hat{\mathbf{A}}$  consistent for  $\mathbf{A}_o$ . By the usual arguments,

$\hat{\mathbf{G}} = N^{-1} \sum_{i=1}^N \nabla_{\boldsymbol{\theta}} \mathbf{g}(\mathbf{w}_i, \hat{\boldsymbol{\theta}})$  is consistent for  $\mathbf{G}_o$ . Then

$$\hat{\mathbf{D}} = N^{-1} \sum_{i=1}^N (\hat{\mathbf{g}}_i - \hat{\boldsymbol{\delta}} - \hat{\mathbf{G}} \hat{\mathbf{A}}^{-1} \hat{\mathbf{s}}_i)(\hat{\mathbf{g}}_i - \hat{\boldsymbol{\delta}} - \hat{\mathbf{G}} \hat{\mathbf{A}}^{-1} \hat{\mathbf{s}}_i)'$$

is consistent for  $\mathbf{D}_o$ , where the “ $\wedge$ ” denotes evaluation at  $\hat{\boldsymbol{\theta}}$ .

c. Using the shorthand notation, if  $E(\mathbf{s}_i | \mathbf{x}_i) = \mathbf{0}$  then  $\mathbf{g}_i$  is uncorrelated with  $\mathbf{s}_i$  because the premise of the problem is that  $\mathbf{g}_i$  is a function of  $\mathbf{x}_i$ . Therefore,  $(\mathbf{g}_i - \boldsymbol{\delta}_o)$  is uncorrelated with

$\mathbf{G}_o \mathbf{A}_o^{-1} \mathbf{s}_i$ , which means

$$\begin{aligned}
\mathbf{D}_o &= \text{Var}(\mathbf{g}_i - \boldsymbol{\delta}_o - \mathbf{G}_o \mathbf{A}_o^{-1} \mathbf{s}_i) \\
&= \text{Var}(\mathbf{g}_i - \boldsymbol{\delta}_o) + \text{Var}(\mathbf{G}_o \mathbf{A}_o^{-1} \mathbf{s}_i) \\
&= \text{Var}(\mathbf{g}_i) + \mathbf{G}_o \mathbf{A}_o^{-1} \mathbf{B}_o \mathbf{A}_o^{-1} \mathbf{G}_o' \\
&= \text{Var}(\mathbf{g}_i) + \mathbf{G}_o [\text{Avar } \sqrt{N} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)] \mathbf{G}_o',
\end{aligned}$$

which is what we wanted to show.

## Solutions to Chapter 13 Problems

**13.1.** No. We know that  $\theta_o$  solves

$$\max_{\theta \in \Theta} E[\log f(\mathbf{y}_i | \mathbf{x}_i; \theta)],$$

where the expectation is over the joint distribution of  $(\mathbf{x}_i, \mathbf{y}_i)$ . Therefore, because  $\exp(\cdot)$  is an increasing function,  $\theta_o$  also maximizes  $\exp\{E[\log f(\mathbf{y}_i | \mathbf{x}_i; \theta)]\}$  over  $\Theta$ . The problem is that the expectation and the exponential function cannot be interchanged:

$E[f(\mathbf{y}_i | \mathbf{x}_i; \theta)] \neq \exp\{E[\log f(\mathbf{y}_i | \mathbf{x}_i; \theta)]\}$ . In fact, Jensen's inequality tells us that

$$E[f(\mathbf{y}_i | \mathbf{x}_i; \theta)] > \exp\{E[\log f(\mathbf{y}_i | \mathbf{x}_i; \theta)]\}$$

**13.2.** a. Because

$$f(y | \mathbf{x}_i) = (2\pi\sigma_o^2)^{-1/2} \exp[-(y - m(\mathbf{x}_i, \boldsymbol{\beta}_o))^2 / (2\sigma_o^2)],$$

it follows that for observation  $i$  the log likelihood is

$$\ell_i(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} [y_i - m(\mathbf{x}_i, \boldsymbol{\beta})]^2.$$

Only the last of these terms depends on  $\boldsymbol{\beta}$ . Further, for any  $\sigma^2 > 0$ , maximizing  $\sum_{i=1}^N \ell_i(\boldsymbol{\beta}, \sigma^2)$  with respect to  $\boldsymbol{\beta}$  is the same as minimizing

$$\sum_{i=1}^N [y_i - m(\mathbf{x}_i, \boldsymbol{\beta})]^2,$$

which means the MLE  $\hat{\boldsymbol{\beta}}$  is the NLS estimator.

b. First,

$$\nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta}, \sigma^2) = \nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta}) [y_i - m(\mathbf{x}_i, \boldsymbol{\beta})] / \sigma^2;$$

note that  $\nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta})$  is  $1 \times P$ . Next,

$$\frac{\partial \ell_i(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} [y_i - m(\mathbf{x}_i, \boldsymbol{\beta})]^2.$$

For notational simplicity, define the residual function  $u_i(\boldsymbol{\beta}) \equiv y_i - m(\mathbf{x}_i, \boldsymbol{\beta})$ . Then the score is

$$\mathbf{s}_i(\boldsymbol{\theta}) = \begin{pmatrix} \nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta})' u_i(\boldsymbol{\beta}) / \sigma^2 \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} [u_i(\boldsymbol{\beta})]^2 \end{pmatrix},$$

where  $\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}) \equiv \nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta})$ .

Define the errors as  $u_i \equiv u_i(\boldsymbol{\beta}_o)$ , so that  $E(u_i | \mathbf{x}_i) = 0$  and  $E(u_i^2 | \mathbf{x}_i) = \text{Var}(y_i | \mathbf{x}_i) = \sigma_o^2$ .

Then, since  $\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}_o)$  is a function of  $\mathbf{x}_i$ , it is easily seen that  $E[\mathbf{s}_i(\boldsymbol{\theta}_o) | \mathbf{x}_i] = \mathbf{0}$ . Note that we only use the fact that  $E(y_i | \mathbf{x}_i) = m(\mathbf{x}_i, \boldsymbol{\beta}_o)$  and  $\text{Var}(y_i | \mathbf{x}_i) = \sigma_o^2$  in showing this. In other words, only the first two conditional moments of  $y_i$  need to be correctly specified; nothing else about the normal distribution is used.

The equation used to obtain  $\hat{\sigma}^2$  is

$$\sum_{i=1}^N \left( -\frac{1}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} [y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})]^2 \right) = 0,$$

where  $\hat{\boldsymbol{\beta}}$  is the nonlinear least squares estimator. Solving gives

$$\hat{\sigma}^2 = N^{-1} \sum_{i=1}^N \hat{u}_i^2,$$

where  $\hat{u}_i \equiv y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ . Thus, the MLE of  $\sigma^2$  is the sum of squared residuals divided by  $N$ . In practice,  $N$  is often replaced with  $N - P$  as a degrees-of-freedom adjustment, but this makes no difference as  $N \rightarrow \infty$ .

c. The derivations are a bit tedious but fairly straightforward:

$$\mathbf{H}_i(\theta) = \begin{pmatrix} -\nabla_{\beta} m_i(\beta)' \nabla_{\beta} m_i(\beta) / \sigma^2 + \nabla_{\beta}^2 m_i(\beta) u_i(\beta) / \sigma^2 & -\nabla_{\beta} m_i(\beta)' u_i(\beta) / \sigma^4 \\ -\nabla_{\beta} m_i(\beta) u_i(\beta) / \sigma^4 & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6} [u_i(\beta)]^2 \end{pmatrix},$$

where  $\nabla_{\beta}^2 m_i(\beta)$  is the  $P \times P$  Hessian of  $m_i(\beta)$ .

d. From part c and  $E(u_i | \mathbf{x}_i) = 0$ , the off-diagonal blocks are zero. Further,

$$E[\nabla_{\beta} m_i(\beta_o)' \nabla_{\beta} m_i(\beta_o) / \sigma_o^2 - \nabla_{\beta}^2 m_i(\beta_o) u_i / \sigma_o^2 | \mathbf{x}_i] = \nabla_{\beta} m_i(\beta_o)' \nabla_{\beta} m_i(\beta_o) / \sigma_o^2$$

Because ,  $E(u_i^2 | \mathbf{x}_i) = \sigma_o^2$ ,

$$E\left(\frac{1}{\sigma_o^6} u_i^2 - \frac{1}{2\sigma_o^4} \middle| \mathbf{x}_i\right) = \frac{1}{\sigma_o^2} - \frac{1}{2\sigma_o^4} = \frac{1}{2\sigma_o^4}.$$

Therefore,

$$-E[\mathbf{H}_i(\theta_o) | \mathbf{x}_i] = \begin{pmatrix} \nabla_{\beta} m_i(\beta_o)' \nabla_{\beta} m_i(\beta_o) / \sigma_o^2 & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\sigma_o^4} \end{pmatrix} \quad (13.99)$$

where we again use  $E(u_i | \mathbf{x}_i) = 0$  and  $E(u_i^2 | \mathbf{x}_i) = \sigma_o^2$ .

e. To show that  $-E[\mathbf{H}_i(\theta_o) | \mathbf{x}_i]$  equals  $E[\mathbf{s}_i(\theta_o) \mathbf{s}_i(\theta_o)' | \mathbf{x}_i]$ , we need to know that, with  $u_i$  defined as above,  $E(u_i^3 | \mathbf{x}_i) = 0$ , which can be used, along with the zero mean and constant conditional variance, to show

$$E[\mathbf{s}_i(\theta_o) \mathbf{s}_i(\theta_o)' | \mathbf{x}_i] = \begin{pmatrix} \nabla_{\beta} m_i(\beta_o)' \nabla_{\beta} m_i(\beta_o) / \sigma_o^2 & \mathbf{0} \\ \mathbf{0} & E\left[\left(-\frac{1}{2\sigma_o^2} + \frac{1}{2\sigma_o^4} u_i^2\right)^2\right] \end{pmatrix}.$$

Further,  $E(u_i^4 | \mathbf{x}_i) = 3\sigma_o^4$ , and so

$$E\left[\left(-\frac{1}{2\sigma_o^2} + \frac{1}{2\sigma_o^4} u_i^2\right)^2\right] = \frac{1}{4\sigma_o^4} + \frac{3\sigma_o^4}{4\sigma_o^8} - \frac{2\sigma_o^2}{4\sigma_o^6} = \frac{1}{2\sigma_o^4}.$$

Thus, we have shown  $-E[\mathbf{H}_i(\theta_o) | \mathbf{x}_i] = E[\mathbf{s}_i(\theta_o) \mathbf{s}_i(\theta_o)' | \mathbf{x}_i]$ .



f. From general MLE, we know that  $\text{Avar}\sqrt{N}(\hat{\beta} - \beta_o)$  is the  $P \times P$  upper left hand block of  $\{E[\mathbf{A}_i(\theta_o)]\}^{-1}$ , where  $\mathbf{A}_i(\theta_o)$  is the matrix in (13.99). Because this matrix is block diagonal, it is easily seen that

$$\text{Avar}\sqrt{N}(\hat{\beta} - \beta_o) = \sigma_o^2 \{E[\nabla_{\beta} m_i(\beta_o)' \nabla_{\beta} m_i(\beta_o)]\}^{-1},$$

and this is consistently estimated by

$$\hat{\sigma}^2 \left( N^{-1} \sum_{i=1}^N \nabla_{\beta} \hat{m}_i' \nabla_{\beta} \hat{m}_i \right)^{-1}, \quad (13.100)$$

which means that  $\widehat{\text{Avar}}(\hat{\beta})$  is (13.100) divided by  $N$ , or

$$\widehat{\text{Avar}}(\hat{\beta}) = \hat{\sigma}^2 \left( \sum_{i=1}^N \nabla_{\beta} \hat{m}_i' \nabla_{\beta} \hat{m}_i \right)^{-1}. \quad (13.101)$$

If the model is linear,  $\nabla_{\beta} \hat{m}_i = \mathbf{x}_i$ , and we obtain exactly the asymptotic variance estimator for the OLS estimator under homoskedasticity.

**13.3. a.** The conditional log-likelihood for observation  $i$  is

$$\ell_i(\theta) = y_i \log[G(\mathbf{x}_i, \theta)] + (1 - y_i) \log[1 - G(\mathbf{x}_i, \theta)].$$

b. The derivation for the probit case in Example 13.1 extends immediately:

$$\begin{aligned} \mathbf{s}_i(\theta) &= y_i \nabla_{\theta} G(\mathbf{x}_i, \theta)' / G(\mathbf{x}_i, \theta) - (1 - y_i) \nabla_{\theta} G(\mathbf{x}_i, \theta)' / [1 - G(\mathbf{x}_i, \theta)] \\ &= \nabla_{\theta} G(\mathbf{x}_i, \theta)' [y_i - G(\mathbf{x}_i, \theta)] / \{G(\mathbf{x}_i, \theta)[1 - G(\mathbf{x}_i, \theta)]\}. \end{aligned}$$

If we plug in  $\theta_o$  for  $\theta$  and take the expectation conditional on  $\mathbf{x}_i$  we get  $E[\mathbf{s}_i(\theta_o)|\mathbf{x}_i] = \mathbf{0}$

because  $E[y_i - G(\mathbf{x}_i, \theta_o)|\mathbf{x}_i] = 0$ , and the functions multiplying  $y_i - G(\mathbf{x}_i, \theta_o)$  depend only on  $\mathbf{x}_i$ .

c. We need to evaluate the score and the expected Hessian with respect to the full set of parameters, but then evaluate these at the restricted estimates. Now,

$$\nabla_{\theta} G(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{0}) = \phi(\mathbf{x}\boldsymbol{\beta})[\mathbf{x}, (\mathbf{x}\boldsymbol{\beta})^2, (\mathbf{x}\boldsymbol{\beta})^3],$$

a  $1 \times (K + 2)$  vector. Let  $\tilde{\boldsymbol{\beta}}$  denote the probit estimates of  $\boldsymbol{\beta}$ , obtained under the null. The score for observation  $i$ , evaluated under the null estimates, is the  $(K + 2) \times 1$  vector

$$\begin{aligned} \mathbf{s}_i(\tilde{\boldsymbol{\theta}}) &= \nabla_{\theta} G(\mathbf{x}_i, \tilde{\boldsymbol{\beta}}, \mathbf{0})' [y_i - \Phi(\mathbf{x}_i \tilde{\boldsymbol{\beta}})] / \{\Phi(\mathbf{x}_i \tilde{\boldsymbol{\beta}})[1 - \Phi(\mathbf{x}_i \tilde{\boldsymbol{\beta}})]\} \\ &= \phi(\mathbf{x}_i \tilde{\boldsymbol{\beta}}) \tilde{\mathbf{z}}_i' [y_i - \Phi(\mathbf{x}_i \tilde{\boldsymbol{\beta}})] / \{\Phi(\mathbf{x}_i \tilde{\boldsymbol{\beta}})[1 - \Phi(\mathbf{x}_i \tilde{\boldsymbol{\beta}})]\}, \end{aligned}$$

where  $\tilde{\mathbf{z}}_i \equiv [\mathbf{x}_i, (\mathbf{x}_i \tilde{\boldsymbol{\beta}})^2, (\mathbf{x}_i \tilde{\boldsymbol{\beta}})^3]$ . The negative of the expected Hessian, evaluated under the null, is the  $(K + 2) \times (K + 2)$  matrix

$$\mathbf{A}(\mathbf{x}_i, \tilde{\boldsymbol{\theta}}) = [\phi(\mathbf{x}_i \tilde{\boldsymbol{\beta}})]^2 \tilde{\mathbf{z}}_i' \tilde{\mathbf{z}}_i / \{\Phi(\mathbf{x}_i \tilde{\boldsymbol{\beta}})[1 - \Phi(\mathbf{x}_i \tilde{\boldsymbol{\beta}})]\}.$$

These can be plugged into the second expression in equation (13.36) to obtain a nonnegative, well-behaved LM statistic. Simple algebra shows that the statistic can be computed as  $N$  times the explained sum of squares from the regression

$$\frac{\tilde{u}_i}{\sqrt{\tilde{\Phi}_i(1 - \tilde{\Phi}_i)}} \text{ on } \frac{\tilde{\phi}_i \cdot \mathbf{x}_i}{\sqrt{\tilde{\Phi}_i(1 - \tilde{\Phi}_i)}}, \frac{\tilde{\phi}_i \cdot (\mathbf{x}_i \tilde{\boldsymbol{\beta}})^2}{\sqrt{\tilde{\Phi}_i(1 - \tilde{\Phi}_i)}}, \frac{\tilde{\phi}_i \cdot (\mathbf{x}_i \tilde{\boldsymbol{\beta}})^3}{\sqrt{\tilde{\Phi}_i(1 - \tilde{\Phi}_i)}}, i = 1, \dots, N$$

where “ $\sim$ ” denotes evaluation at  $(\tilde{\boldsymbol{\beta}}, \mathbf{0})$  and  $\tilde{u}_i = y_i - \tilde{\Phi}_i$ . Under  $H_0$ ,  $LM$  is distributed asymptotically as  $\chi^2_2$ .

d. The variable addition version of the test is to estimate, in a second step, a probit model with response probability of the form

$$\Phi[\mathbf{x}_i \boldsymbol{\beta} + \delta_1 (\mathbf{x}_i \tilde{\boldsymbol{\beta}})^2 + \delta_2 (\mathbf{x}_i \tilde{\boldsymbol{\beta}})^3]$$

and the compute a Wald test of  $H_0 : \delta_1 = \delta_2 = 0$ . As we discussed in Problem 12.5 in a related context, this is to be used only as a test. The estimates of  $\delta_1$  and  $\delta_2$  obtained by inserting  $\tilde{\boldsymbol{\beta}}$  into the square and quadratic are generally inconsistent if at least one of  $\delta_1$  and  $\delta_2$  is different from zero.

**13.4.** If the density of  $\mathbf{y}$  given  $\mathbf{x}$  is correctly specified then  $E[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o) | \mathbf{x}] = \mathbf{0}$ . But then

$$E[a(\mathbf{x}, \boldsymbol{\theta}_o) \mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o) | \mathbf{x}] = a(\mathbf{x}, \boldsymbol{\theta}_o) E[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o) | \mathbf{x}] = \mathbf{0}$$

which, of course implies an unconditional expectation of zero. The only restriction on  $a(\mathbf{x}, \boldsymbol{\theta}_o)$  would be to ensure the expected value is well defined (but this is usually just assumed, not verified).

**13.5.** a. Because  $\mathbf{s}_i^g(\boldsymbol{\phi}_o) = [\mathbf{G}(\boldsymbol{\theta}_o)]^{-1} \mathbf{s}_i(\boldsymbol{\theta}_o)$ ,

$$\begin{aligned} E[\mathbf{s}_i^g(\boldsymbol{\phi}_o) \mathbf{s}_i^g(\boldsymbol{\phi}_o)' | \mathbf{x}_i] &= E\{[\mathbf{G}(\boldsymbol{\theta}_o)]^{-1} \mathbf{s}_i(\boldsymbol{\theta}_o) \mathbf{s}_i(\boldsymbol{\theta}_o)' [\mathbf{G}(\boldsymbol{\theta}_o)]^{-1} | \mathbf{x}_i\} \\ &= [\mathbf{G}(\boldsymbol{\theta}_o)]^{-1} E[\mathbf{s}_i(\boldsymbol{\theta}_o) \mathbf{s}_i(\boldsymbol{\theta}_o)' | \mathbf{x}_i] [\mathbf{G}(\boldsymbol{\theta}_o)]^{-1} \\ &= [\mathbf{G}(\boldsymbol{\theta}_o)]^{-1} \mathbf{A}_i(\boldsymbol{\theta}_o) [\mathbf{G}(\boldsymbol{\theta}_o)]^{-1}. \end{aligned}$$

where the last equality follows from the conditional information matrix equality.

b. In part a, we just replace  $\boldsymbol{\theta}_o$  with  $\tilde{\boldsymbol{\theta}}$  and  $\boldsymbol{\phi}_o$  with  $\tilde{\boldsymbol{\phi}}$  :

$$\tilde{\mathbf{A}}_i^g = [\mathbf{G}(\tilde{\boldsymbol{\theta}})]^{-1} \mathbf{A}_i(\tilde{\boldsymbol{\theta}}) [\mathbf{G}(\tilde{\boldsymbol{\theta}})]^{-1} \equiv \tilde{\mathbf{G}}'^{-1} \tilde{\mathbf{A}}_i \tilde{\mathbf{G}}^{-1}.$$

c. The expected Hessian form of the statistic is given in the second part of equation (13.36),

but where it depends on  $\tilde{\mathbf{s}}_i^g$  and  $\tilde{\mathbf{A}}_i^g$  :

$$\begin{aligned} LM_g &= \left( \sum_{i=1}^N \tilde{\mathbf{s}}_i^g \right)' \left( \sum_{i=1}^N \tilde{\mathbf{A}}_i^g \right)^{-1} \left( \sum_{i=1}^N \tilde{\mathbf{s}}_i^g \right) \\ &= \left( \sum_{i=1}^N \tilde{\mathbf{G}}'^{-1} \tilde{\mathbf{s}}_i \right)' \left( \sum_{i=1}^N \tilde{\mathbf{G}}'^{-1} \tilde{\mathbf{A}}_i \tilde{\mathbf{G}}^{-1} \right)^{-1} \left( \sum_{i=1}^N \tilde{\mathbf{G}}'^{-1} \tilde{\mathbf{s}}_i \right) \\ &= \left( \sum_{i=1}^N \tilde{\mathbf{s}}_i \right)' \tilde{\mathbf{G}}^{-1} \tilde{\mathbf{G}} \left( \sum_{i=1}^N \tilde{\mathbf{A}}_i \right)^{-1} \tilde{\mathbf{G}}' \tilde{\mathbf{G}}^{-1} \left( \sum_{i=1}^N \tilde{\mathbf{s}}_i \right) \\ &= \left( \sum_{i=1}^N \tilde{\mathbf{s}}_i \right)' \left( \sum_{i=1}^N \tilde{\mathbf{A}}_i \right)^{-1} \left( \sum_{i=1}^N \tilde{\mathbf{s}}_i \right) = LM. \end{aligned}$$

**13.6.** a. No, for two reasons. First, just specifying a distribution of  $y_{it}$  given  $\mathbf{x}_{it}$  says

nothing, in general, about the distribution of  $y_{it}$  given  $\mathbf{x}_i \equiv (\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})$ . We could assume these two are the same, which is the strict exogeneity assumption. But, even under strict exogeneity, we would have to specify something about joint distributions (perhaps via conditional distributions) involving different time periods. We could assume independence (conditional on  $\mathbf{x}_i$ ) or make a dynamic completeness assumption. Either way, without substantially more assumptions, we cannot derive the distribution of  $\mathbf{y}_i$  given  $\mathbf{x}_i$ .

b. This is given in a more general case in equation (18.69) in Chapter 18. It can be derived easily from Example 13.2, which gives  $\ell_i(\theta)$  for the cross section case:

$$\ell_i(\theta) = \sum_{t=1}^T [y_{it}\mathbf{x}_{it}'\theta - \exp(\mathbf{x}_{it}'\theta)] \equiv \sum_{t=1}^T \ell_{it}(\theta).$$

Taking the gradient and transposing gives

$$\mathbf{s}_i(\theta) = \sum_{t=1}^T \mathbf{x}_{it}' [y_{it} - \exp(\mathbf{x}_{it}'\theta)] \equiv \sum_{t=1}^T \mathbf{s}_{it}(\theta).$$

c. First, we need the Hessian for each  $i$ , which is easily obtained as  $\nabla_{\theta}\mathbf{s}_i(\theta)$  :

$$\mathbf{H}_i(\theta) = - \sum_{t=1}^T \exp(\mathbf{x}_{it}'\theta) \mathbf{x}_{it}' \mathbf{x}_{it},$$

which, in this example, does not depend on the  $y_{it}$  (see Problem 13.12 for the notion of a canonical link function). In particular,  $\mathbf{A}_{it}(\theta_o) = -E[\mathbf{H}_{it}(\theta_o)|\mathbf{x}_{it}] = -\mathbf{H}_{it}(\theta_o)$ . Therefore,

$$\hat{\mathbf{A}} = N^{-1} \sum_{i=1}^N \sum_{t=1}^T \exp(\mathbf{x}_{it}'\hat{\theta}) \mathbf{x}_{it}' \mathbf{x}_{it},$$

where  $\hat{\theta}$  is the partial MLE. Further,

$$\hat{\mathbf{B}} = N^{-1} \sum_{i=1}^N \mathbf{s}_i(\hat{\boldsymbol{\theta}}) \mathbf{s}_i(\hat{\boldsymbol{\theta}})',$$

and then  $\text{Avar}(\hat{\boldsymbol{\theta}})$  is estimated as

$$\left( \sum_{i=1}^N \sum_{t=1}^T \exp(\mathbf{x}_{it}\hat{\boldsymbol{\theta}}) \mathbf{x}_{it}' \mathbf{x}_{it} \right)^{-1} \left( \sum_{i=1}^N \mathbf{s}_i(\hat{\boldsymbol{\theta}}) \mathbf{s}_i(\hat{\boldsymbol{\theta}})' \right) \left( \sum_{i=1}^N \sum_{t=1}^T \exp(\mathbf{x}_{it}\hat{\boldsymbol{\theta}}) \mathbf{x}_{it}' \mathbf{x}_{it} \right)^{-1}$$

d. If  $E(y_{it}|\mathbf{x}_{it}, y_{i,t-1}, \mathbf{x}_{i,t-1}, \dots, y_{i1}, \mathbf{x}_{i1}) = E(y_{it}|\mathbf{x}_{it})$  then

$$E[\mathbf{s}_{it}(\boldsymbol{\theta}_o)|\mathbf{x}_{it}, y_{i,t-1}, \mathbf{x}_{i,t-1}, \dots] = \mathbf{x}_{it}' [E(y_{it}|\mathbf{x}_{it}, y_{i,t-1}, \mathbf{x}_{i,t-1}, \dots) - \exp(\mathbf{x}_{it}\boldsymbol{\theta}_o)] = \mathbf{0}.$$

As usual, this finding implies that  $\mathbf{s}_{it}(\boldsymbol{\theta}_o)$  and  $\mathbf{s}_{ir}(\boldsymbol{\theta}_o)$  are uncorrelated,  $t \neq r$ . Therefore,

$$\mathbf{B}_o = \sum_{t=1}^T E[\mathbf{s}_{it}(\boldsymbol{\theta}_o) \mathbf{s}_{it}(\boldsymbol{\theta}_o)'] = \sum_{t=1}^T E(u_{it}^2 \mathbf{x}_{it}' \mathbf{x}_{it}),$$

where  $u_{it} \equiv y_{it} - E(y_{it}|\mathbf{x}_{it}) = y_{it} - \exp(\mathbf{x}_{it}\boldsymbol{\theta}_o)$ . Now, by the Poisson assumption,

$E(u_{it}^2|\mathbf{x}_{it}) = \text{Var}(y_{it}|\mathbf{x}_{it}) = \exp(\mathbf{x}_{it}\boldsymbol{\theta}_o)$ . By iterated expectations,

$$\mathbf{B}_o = \sum_{t=1}^T E[\exp(\mathbf{x}_{it}\boldsymbol{\theta}_o) \mathbf{x}_{it}' \mathbf{x}_{it}] = \mathbf{A}_o.$$

(We have really just verified the conditional information matrix equality for each  $t$  in the special case of Poisson regression with an exponential mean function.) herefore, we can estimate  $\text{Avar}(\hat{\boldsymbol{\theta}})$  as

$$\left( \sum_{i=1}^N \sum_{t=1}^T \exp(\mathbf{x}_{it}\hat{\boldsymbol{\theta}}) \mathbf{x}_{it}' \mathbf{x}_{it} \right)^{-1},$$

which is exactly what we get by using pooled Poisson estimation and ignoring the time dimension.

**13.7. a.** The joint density is simply  $g(y_1|y_2, \mathbf{x}; \boldsymbol{\theta}_o) \cdot h(y_2|\mathbf{x}; \boldsymbol{\theta}_o)$ . The log-likelihood for

observation  $i$  is

$$\ell_i(\theta) \equiv \log[g(y_{i1}|y_{i2}, \mathbf{x}_i; \theta_o)] + \log[h(y_{i2}|\mathbf{x}_i; \theta_o)],$$

and we would use this in a standard MLE analysis (conditional on  $\mathbf{x}_i$ ).

b. First, we know that, for all  $(y_{i2}, \mathbf{x}_i)$ ,  $\theta_o$  minimizes  $E[\ell_{i1}(\theta)|y_{i2}, \mathbf{x}_i]$ . Because  $r_{i2}$  is a function of  $(y_{i2}, \mathbf{x}_i)$ ,

$$E[r_{i2}\ell_{i1}(\theta)|y_{i2}, \mathbf{x}_i] = r_{i2}E[\ell_{i1}(\theta)|y_{i2}, \mathbf{x}_i];$$

because  $r_{i2} \geq 0$ ,  $\theta_o$  maximizes  $E[r_{i2}\ell_{i1}(\theta)|y_{i2}, \mathbf{x}_i]$  for all  $(y_{i2}, \mathbf{x}_i)$ , and therefore  $\theta_o$  maximizes  $E[r_{i2}\ell_{i1}(\theta)]$  by iterated expectations. Similarly,  $\theta_o$  maximizes  $E[\ell_{i1}(\theta)]$ , and so it follows that  $\theta_o$  maximizes  $E[r_{i2}\ell_{i1}(\theta) + \ell_{i2}(\theta)]$ . For identification, we have to assume or verify uniqueness.

c. The score is

$$\mathbf{s}_i(\theta) = r_{i2}\mathbf{s}_{i1}(\theta) + \mathbf{s}_{i2}(\theta),$$

where  $\mathbf{s}_{i1}(\theta) = \nabla_{\theta}\ell_{i1}(\theta)'$  and  $\mathbf{s}_{i2}(\theta) \equiv \nabla_{\theta}\ell_{i2}(\theta)'$ . Therefore,

$$\begin{aligned} E[\mathbf{s}_i(\theta_o)\mathbf{s}_i(\theta_o)'] &= E[r_{i2}\mathbf{s}_i(\theta_o)\mathbf{s}_i(\theta_o)'] + E[\mathbf{s}_{i2}(\theta_o)\mathbf{s}_{i2}(\theta_o)'] \\ &\quad + E[r_{i2}\mathbf{s}_{i1}(\theta_o)\mathbf{s}_{i2}(\theta_o)'] + E[r_{i2}\mathbf{s}_{i2}(\theta_o)\mathbf{s}_{i1}(\theta_o)']. \end{aligned}$$

Now by the usual conditional MLE theory,  $E[\mathbf{s}_i(\theta_o)|y_{i2}, \mathbf{x}_i] = 0$  and, since  $r_{i2}$  and  $\mathbf{s}_{i2}(\theta)$  are functions of  $(y_{i2}, \mathbf{x}_i)$ , it follows that  $E[r_{i2}\mathbf{s}_{i1}(\theta_o)\mathbf{s}_{i2}(\theta_o)']|y_{i2}, \mathbf{x}_i] = 0$ , and so its transpose also has zero conditional expectation. As usual, this implies zero unconditional expectation. We have shown

$$E[\mathbf{s}_i(\theta_o)\mathbf{s}_i(\theta_o)'] = E[r_{i2}\mathbf{s}_{i1}(\theta_o)\mathbf{s}_{i1}(\theta_o)'] + E[\mathbf{s}_{i2}(\theta_o)\mathbf{s}_{i2}(\theta_o)'].$$

Now, by the unconditional information matrix equality for the density  $h(y_2|\mathbf{x}; \theta)$ ,

$$E[\mathbf{s}_{i2}(\theta_o)\mathbf{s}_{i2}(\theta_o)'] = -E[\mathbf{H}_{i2}(\theta_o)],$$

where  $\mathbf{H}_{i2}(\boldsymbol{\theta}_o) = \nabla_{\boldsymbol{\theta}} \mathbf{s}_{i2}(\boldsymbol{\theta})$ . Further, by the conditional IM equality for the density  $g(y_1|y_2, \mathbf{x}; \boldsymbol{\theta})$ ,

$$\mathbb{E}[\mathbf{s}_{i1}(\boldsymbol{\theta}_o) \mathbf{s}_{i1}(\boldsymbol{\theta}_o)' | y_{i2}, \mathbf{x}_i] = -\mathbb{E}[\mathbf{H}_{i1}(\boldsymbol{\theta}_o) | y_{i2}, \mathbf{x}_i], \quad (13.102)$$

where  $\mathbf{H}_{i1}(\boldsymbol{\theta}_o) = \nabla_{\boldsymbol{\theta}} \mathbf{s}_{i1}(\boldsymbol{\theta})$ . Since  $r_{i2}$  is a function of  $(y_{i2}, \mathbf{x}_i)$ , we can put  $r_{i2}$  inside both expectations in (13.102). Then, by iterated expectations,

$$\mathbb{E}[r_{i2} \mathbf{s}_i(\boldsymbol{\theta}_o) \mathbf{s}_i(\boldsymbol{\theta}_o)'] = -\mathbb{E}[r_{i2} \mathbf{H}_{i1}(\boldsymbol{\theta}_o)].$$

Combining all the pieces, we have shown that

$$\begin{aligned} \mathbb{E}[\mathbf{s}_i(\boldsymbol{\theta}_o) \mathbf{s}_i(\boldsymbol{\theta}_o)'] &= -\mathbb{E}[r_{i2} \mathbf{H}_{i1}(\boldsymbol{\theta}_o)] - \mathbb{E}[\mathbf{H}_{i2}(\boldsymbol{\theta}_o)] \\ &= -\{\mathbb{E}[r_{i2} \nabla_{\boldsymbol{\theta}} \mathbf{s}_{i1}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \mathbf{s}_{i2}(\boldsymbol{\theta})]\} \\ &= -\mathbb{E}[\nabla_{\boldsymbol{\theta}}^2 \ell_i(\boldsymbol{\theta})] \equiv -\mathbb{E}[\mathbf{H}_i(\boldsymbol{\theta})]. \end{aligned}$$

So we have verified that an unconditional IM equality holds, which means we can estimate the asymptotic variance of  $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$  by estimating  $\{-\mathbb{E}[\mathbf{H}_i(\boldsymbol{\theta})]\}^{-1}$ .

d. From part c, one consistent estimator of  $\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)]$  is

$$N^{-1} \sum_{i=1}^N (r_{i2} \hat{\mathbf{H}}_{i1} + \hat{\mathbf{H}}_{i2}),$$

where the notation should be obvious. In some cases it may be simpler to use an expected Hessian form for each piece, which we can obtain by looking for consistent estimators of  $-\mathbb{E}[r_{i2} \mathbf{H}_{i1}(\boldsymbol{\theta}_o)]$  and  $-\mathbb{E}[\mathbf{H}_{i2}(\boldsymbol{\theta}_o)]$ . By definition,  $\mathbf{A}_{i2}(\boldsymbol{\theta}_o) \equiv -\mathbb{E}[\mathbf{H}_{i2}(\boldsymbol{\theta}_o) | \mathbf{x}_i]$ , and so  $\mathbb{E}[\mathbf{A}_{i2}(\boldsymbol{\theta}_o)] \equiv -\mathbb{E}[\mathbf{H}_{i2}(\boldsymbol{\theta}_o)]$ . By the usual law of large numbers argument,

$$N^{-1} \sum_{i=1}^N \hat{\mathbf{A}}_{i2} \xrightarrow{p} -\mathbb{E}[\mathbf{H}_{i2}(\boldsymbol{\theta}_o)].$$

Similarly, since  $\mathbf{A}_{i1}(\boldsymbol{\theta}_o) \equiv -\mathbb{E}[\mathbf{H}_{i1}(\boldsymbol{\theta}_o) | y_{i2}, \mathbf{x}_i]$ , and  $r_{i2}$  is a function of  $(y_{i2}, \mathbf{x}_i)$ , it follows that

$E[r_{i2}\mathbf{A}_{i1}(\boldsymbol{\theta}_o)] = -E[r_{i2}\mathbf{H}_{i1}(\boldsymbol{\theta}_o)]$ . Under general regularity conditions,  $N^{-1} \sum_{i=1}^N r_{i2}\hat{\mathbf{A}}_{i1}$  consistently estimates  $-E[r_{i2}\mathbf{H}_{i1}(\boldsymbol{\theta}_o)]$ . This completes what we needed to show.

Interestingly, even though we do not have a true conditional maximum likelihood problem, we can still use the conditional expectations of the Hessians – but conditioned on different sets of variables,  $(y_{i2}, \mathbf{x}_i)$  in one case, and  $\mathbf{x}_i$  in the other – to consistently estimate the asymptotic variance of the partial MLE.

e. **(Bonus Question)** Show that if we were able to use the entire random sample, the resulting conditional MLE would be more efficient than the partial MLE based on the selected sample.

### Solution

We use a standard fact about positive definite matrices: if  $\mathbf{A}$  and  $\mathbf{B}$  are  $P \times P$  positive definite matrices, then  $\mathbf{A} - \mathbf{B}$  is p.s.d. if and only if  $\mathbf{B}^{-1} - \mathbf{A}^{-1}$  is p.s.d. Now, as we showed in part d, the asymptotic variance of the partial MLE is  $\{E[r_{i2}\mathbf{A}_{i1}(\boldsymbol{\theta}_o) + \mathbf{A}_{i2}(\boldsymbol{\theta}_o)]\}^{-1}$ . If we could use the entire random sample for both terms, the asymptotic variance would be  $\{E[\mathbf{A}_{i1}(\boldsymbol{\theta}_o) + \mathbf{A}_{i2}(\boldsymbol{\theta}_o)]\}^{-1}$ . But

$$E[\mathbf{A}_{i1}(\boldsymbol{\theta}_o) + \mathbf{A}_{i2}(\boldsymbol{\theta}_o)] - E[r_{i2}\mathbf{A}_{i1}(\boldsymbol{\theta}_o) + \mathbf{A}_{i2}(\boldsymbol{\theta}_o)] = E[(1 - r_{i2})\mathbf{A}_{i1}(\boldsymbol{\theta}_o)],$$

which is p.s.d. because  $\mathbf{A}_{i1}(\boldsymbol{\theta}_o)$  is p.s.d. and  $1 - r_{i2} \geq 0$ . Intuitively, the larger is  $P(r_{i2} = 1)$  the smaller is the efficiency difference.

**13.8. a.** This is similar to Problem 12.12 for nonlinear regression; here we are specifying a full conditional distribution. We can use the results in Section 12.4.2:

$$\begin{aligned} \text{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)] &= \mathbf{A}_o^{-1}(\mathbf{B}_o + \mathbf{F}_o\mathbf{C}_o\mathbf{F}_o')\mathbf{A}_o^{-1} \\ &= \mathbf{A}_o^{-1} + \mathbf{A}_o^{-1}\mathbf{F}_o\mathbf{C}_o\mathbf{F}_o'\mathbf{A}_o^{-1} \end{aligned}$$



where  $\mathbf{C}_o = E[\mathbf{r}_i(\boldsymbol{\gamma}_o)\mathbf{r}_i(\boldsymbol{\gamma}_o)']$ . We also use the information matrix equality,  $\mathbf{A}_o = \mathbf{B}_o$ , where

$$\mathbf{A}_o = E[\mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o)\mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o)] = -E[\mathbf{H}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o)]$$

and

$$\begin{aligned}\mathbf{s}_i(\boldsymbol{\theta}; \boldsymbol{\gamma}) &= \nabla_{\boldsymbol{\theta}} \log[f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{g}(\mathbf{w}_i, \boldsymbol{\gamma}); \boldsymbol{\theta})]' \\ \mathbf{H}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o) &= \nabla_{\boldsymbol{\theta}} \mathbf{s}_i(\boldsymbol{\theta}; \boldsymbol{\gamma}).\end{aligned}$$

To use the formula we need to characterize  $\mathbf{F}_o$ . First,

$$\nabla_{\boldsymbol{\gamma}} \mathbf{s}_i(\boldsymbol{\theta}; \boldsymbol{\gamma}) = \nabla_{\boldsymbol{\gamma}} \{ \nabla_{\boldsymbol{\theta}} \log[f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{g}(\mathbf{w}_i, \boldsymbol{\gamma}); \boldsymbol{\theta})] \},$$

which generally requires using the chain rule to compute. Write

$$\mathbf{k}(\mathbf{y}, \mathbf{x}, \mathbf{g}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log[f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{g}(\mathbf{w}_i, \boldsymbol{\gamma}); \boldsymbol{\theta})]'.$$

Then

$$\nabla_{\boldsymbol{\gamma}} \mathbf{s}_i(\boldsymbol{\theta}; \boldsymbol{\gamma}) = \nabla_{\mathbf{g}} \mathbf{k}(\mathbf{y}_i, \mathbf{x}_i, \mathbf{g}(\mathbf{w}_i, \boldsymbol{\gamma}), \boldsymbol{\theta}) \nabla_{\boldsymbol{\gamma}} \mathbf{g}(\mathbf{w}_i, \boldsymbol{\gamma})$$

and

$$\mathbf{F}_o = E[\nabla_{\mathbf{g}} \mathbf{k}(\mathbf{y}_i, \mathbf{x}_i, \mathbf{g}(\mathbf{w}_i, \boldsymbol{\gamma}_o), \boldsymbol{\theta}_o) \nabla_{\boldsymbol{\gamma}} \mathbf{g}(\mathbf{w}_i, \boldsymbol{\gamma}_o)]$$

b. Generally,

$$\widehat{\text{Avar}}[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)] = \hat{\mathbf{A}}_o^{-1} + \hat{\mathbf{A}}_o^{-1} \hat{\mathbf{F}}_o \hat{\mathbf{C}}_o \hat{\mathbf{F}}_o' \hat{\mathbf{A}}_o^{-1}$$

where  $\hat{\mathbf{A}}_o$  is one of the various choices for estimating the information matrix, evaluated at  $\hat{\boldsymbol{\theta}}$  and  $\hat{\boldsymbol{\gamma}}$ ,

$$\hat{\mathbf{C}} = N^{-1} \sum_{i=1}^N \mathbf{r}(\mathbf{w}_i, \hat{\boldsymbol{\gamma}}) \mathbf{r}(\mathbf{w}_i, \hat{\boldsymbol{\gamma}})'$$

and

$$\hat{\mathbf{F}} = N^{-1} \sum_{i=1}^N \nabla_{\mathbf{g}} \mathbf{k}(\mathbf{y}_i, \mathbf{x}_i, \mathbf{g}(\mathbf{w}_i, \hat{\boldsymbol{\gamma}}), \hat{\boldsymbol{\theta}}) \nabla_{\boldsymbol{\gamma}} \mathbf{g}(\mathbf{w}_i, \hat{\boldsymbol{\gamma}}).$$

c. It applies directly where the scalar  $\rho_o$  plays the role of  $\gamma_o$ . The score for this problem (with respect to  $\boldsymbol{\theta}$ ) is

$$\mathbf{s}_i(\boldsymbol{\theta}; \boldsymbol{\gamma}) = \frac{\phi(\mathbf{x}_i \boldsymbol{\beta} + \rho g_i(\boldsymbol{\gamma}))}{\{\Phi(\mathbf{x}_i \boldsymbol{\beta} + \rho g_i(\boldsymbol{\gamma})) [1 - \Phi(\mathbf{x}_i \boldsymbol{\beta} + \rho g_i(\boldsymbol{\gamma}))]\}} \begin{pmatrix} \mathbf{x}_i' \\ g_i(\boldsymbol{\gamma}) \end{pmatrix} [y_i - \Phi(\mathbf{x}_i \boldsymbol{\beta} + \rho g_i(\boldsymbol{\gamma}))],$$

where  $g_i(\boldsymbol{\gamma}) = h_i - \mathbf{z}_i \boldsymbol{\gamma}$ . The full Jacobian of  $\mathbf{s}_i(\boldsymbol{\theta}; \boldsymbol{\gamma})$  with respect to  $\boldsymbol{\gamma}$  is complicated, but it is easy to see it has the form

$$\begin{aligned} \nabla_{\boldsymbol{\gamma}} \mathbf{s}_i(\boldsymbol{\theta}; \boldsymbol{\gamma}) &= \mathbf{L}(\mathbf{x}_i, \mathbf{z}_i, h_i; \boldsymbol{\theta}, \rho) [y_i - \Phi(\mathbf{x}_i \boldsymbol{\beta} + \rho g_i(\boldsymbol{\gamma}))] \\ &\quad + \rho \cdot \frac{\phi(\mathbf{x}_i \boldsymbol{\beta} + \rho g_i(\boldsymbol{\gamma}))}{\{\Phi(\mathbf{x}_i \boldsymbol{\beta} + \rho g_i(\boldsymbol{\gamma})) [1 - \Phi(\mathbf{x}_i \boldsymbol{\beta} + \rho g_i(\boldsymbol{\gamma}))]\}} \begin{pmatrix} \mathbf{x}_i' \\ g_i(\boldsymbol{\gamma}) \end{pmatrix} \mathbf{z}_i \phi(\mathbf{x}_i \boldsymbol{\beta} + \rho g_i(\boldsymbol{\gamma})) \end{aligned}$$

When evaluated at the true values  $\boldsymbol{\theta}_o$  and  $\gamma_o$ , the first term has zero expectation conditional on  $(\mathbf{x}_i, \mathbf{z}_i, h_i)$  because

$$E(y_i | \mathbf{x}_i, \mathbf{z}_i, h_i) = \Phi(\mathbf{x}_i \boldsymbol{\beta}_o + \rho_o g_i(\gamma_o))$$

So  $\mathbf{F}_o$  can be estimated by plugging in the estimators and averaging the second term across  $i$ .

d. When  $\rho_o = 0$ , the second term in  $\nabla_{\boldsymbol{\gamma}} \mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o)$  is zero, and so

$$E[\nabla_{\boldsymbol{\gamma}} \mathbf{s}_i(\boldsymbol{\theta}_o; \boldsymbol{\gamma}_o) | \mathbf{x}_i, \mathbf{z}_i, h_i] = \mathbf{0},$$

which means condition (12.37) holds and  $\mathbf{F}_o = \mathbf{0}$ . This implies, from part a,

$$\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)] = \mathbf{A}_o^{-1}.$$

e. Because  $\rho_o$  is an element of  $\boldsymbol{\theta}_o$ , for testing  $H_0 : \rho_o = 0$  we can ignore the fact that  $\boldsymbol{\gamma}_o$  has been estimated in the first stage. In other words, when we run probit of  $y_i$  on  $\mathbf{x}_i, \hat{g}_i$ ,

$i = 1, \dots, N$ , where  $\hat{g}_i = h_i - \mathbf{z}_i \hat{\gamma}$ , we can use a standard probit  $t$  statistic on  $g_i$ .

**13.9.** a. Under the Markov assumption, the joint density of  $(y_{i0}, \dots, y_{iT})$  is given by

$$f_T(y_T|y_{T-1}) \cdot f_{T-1}(y_{T-1}|y_{T-2}) \cdots f_1(y_1|y_0) \cdot f_0(y_0),$$

so we would need to model  $f_0(y_0)$  to obtain a model of the joint density.

b. The log likelihood

$$\ell_i(\boldsymbol{\theta}) = \sum_{t=1}^T \log[f_t(y_{it}|y_{i,t-1}; \boldsymbol{\theta})]$$

is the conditional log-likelihood for the density of  $(y_{i1}, \dots, y_{iT})$  given  $y_{i0}$ , and so the usual theory of conditional maximum likelihood applies. In practice, this is MLE pooled across  $i$  and  $t$ .

c. Because we have the density of  $(y_{i1}, \dots, y_{iT})$  given  $y_{i0}$ , we can use any of the three asymptotic variance estimators implied by the information matrix equality. However, we can also use the simplifications due to dynamic completeness of each conditional density. Let  $\mathbf{s}_{it}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log[f(y_{it}|y_{i,t-1}; \boldsymbol{\theta})]$ ,  $\mathbf{H}_{it}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathbf{s}_{it}(\boldsymbol{\theta})$  and  $\mathbf{A}_{it}(\boldsymbol{\theta}_o) = -E[\mathbf{H}_{it}(\boldsymbol{\theta})|y_{i,t-1}]$ ,  $t = 1, \dots, T$ . Then  $\text{Avar} \sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$  is consistently estimated using the inverse of any of the three matrices in equation (13.50). If we have a canned package that computes a particular MLE, we can just use any of the usual asymptotic variance estimates obtained from the pooled MLE.

**13.10.** a. Because of conditional independence, the joint density [conditional on  $(\mathbf{x}, c)$ ] is the product of the marginal densities [conditional on  $(\mathbf{x}, c)$ ]:

$$f(y_1, y_2, \dots, y_G | \mathbf{x}, c) = \prod_{g=1}^G f_g(y_g | \mathbf{x}, c).$$

b. Let  $g(y_i, \dots, y_G | \mathbf{x})$  be the joint density of  $y_i$  given  $\mathbf{x}_i = \mathbf{x}$ . Then

$$g(y_i, \dots, y_G | \mathbf{x}) = \int_{\mathbb{R}} f(y_i, y_2, \dots, y_G | \mathbf{x}, c) h(c | \mathbf{x}) dc.$$

c. The density  $g(y_i, \dots, y_G | \mathbf{x})$  is now

$$\begin{aligned} g(y_i, \dots, y_G | \mathbf{x}; \boldsymbol{\gamma}_o, \boldsymbol{\delta}_o) &= \int_{\mathbb{R}} f(y_i, y_2, \dots, y_G | \mathbf{x}, c; \boldsymbol{\gamma}_o) h(c | \mathbf{x}; \boldsymbol{\delta}_o) dc \\ &= \int_{\mathbb{R}} \prod_{g=1}^G f(y_g | \mathbf{x}, c; \boldsymbol{\gamma}_o^g) h(c | \mathbf{x}; \boldsymbol{\delta}_o) dc, \end{aligned}$$

and so the log likelihood for observation  $i$  is

$$\begin{aligned} &\log[g(y_{i1}, \dots, y_{iG} | \mathbf{x}_i; \boldsymbol{\gamma}_o, \boldsymbol{\delta}_o)] \\ &= \log \left[ \int_{\mathbb{R}} \prod_{g=1}^G f(y_{ig} | \mathbf{x}_i, c; \boldsymbol{\gamma}_o^g) h(c | \mathbf{x}_i; \boldsymbol{\delta}_o) dc \right]. \end{aligned}$$

d. This setup has some features in common with a linear SUR model, although here the correlation across equations is assumed to come through a single common component,  $c$ . Because of computational issues with general nonlinear models – especially if  $G$  is large and some of the models are for qualitative response – one probably needs to restrict the cross equation correlation somehow.

**13.11.** a. For each  $t \geq 1$ , the density of  $y_{it}$  given  $y_{i,t-1} = y_{i,t-1}, y_{i,t-2} = y_{i,t-2}, \dots, y_{i0} = y_{i0}$  and  $c_i = c$  is

$$f_t(y_t | y_{t-1}, c) = (2\pi\sigma_e^2)^{-1/2} \exp[-(y_t - \rho y_{t-1} - c)^2 / (2\sigma_e^2)].$$

Therefore, the density of  $(y_{i1}, \dots, y_{iT})$  given  $y_{i0} = y_{i0}$  and  $c_i = c$  is obtained by the product of these densities:

$$\prod_{t=1}^T (2\pi\sigma_e^2)^{-1/2} \exp[-(y_t - \rho y_{t-1} - c)^2 / (2\sigma_e^2)].$$

If we plug in the data for observation  $i$  and take the log we get

$$\begin{aligned}
& \sum_{t=1}^T \left\{ -(1/2) \log(\sigma_e^2) - (y_{it} - \rho y_{i,t-1} - c_i)^2 / (2\sigma_e^2) \right\} \\
&= -(T/2) \log(\sigma_e^2) - \sum_{t=1}^T (y_{it} - \rho y_{i,t-1} - c_i)^2 / (2\sigma_e^2),
\end{aligned}$$

where we have dropped the term that does not depend on the parameters.

It is not a good idea to “estimate” the  $c_i$  along with the  $\rho$  and  $\sigma_e^2$ , as the incidental parameters problem causes inconsistency – severe in some cases – in the estimator of  $\rho$ .

b. If we write  $c_i = \alpha_0 + \alpha_1 y_{i0} + a_i$ , under the maintained assumption, then the density of  $(y_{i1}, \dots, y_{iT})$  given  $(y_{i0} = y_0, a_i = a)$  is

$$\prod_{t=1}^T (2\pi\sigma_e^2)^{-1/2} \exp[-(y_{it} - \rho y_{i,t-1} - \alpha_0 - \alpha_1 y_0 - a)^2 / (2\sigma_e^2)].$$

Now, to get the density condition on  $y_{i0} = y_0$  only, we integrate this density over the density of  $a_i$  given  $y_{i0} = y_0$ . But  $a_i$  and  $y_{i0}$  are independent, and  $a_i \sim \text{Normal}(0, \sigma_a^2)$ . So the density of  $(y_{i1}, \dots, y_{iT})$  given  $y_{i0} = y_0$  is

$$\int_{-\infty}^{\infty} \left( \prod_{t=1}^T (2\pi\sigma_e^2)^{-1/2} \exp[-(y_{it} - \rho y_{i,t-1} - \alpha_0 - \alpha_1 y_0 - a)^2 / (2\sigma_e^2)] \right) \sigma_a^{-1} \phi(a/\sigma_a) da.$$

If we now plug in the data  $(y_{i0}, y_{i1}, \dots, y_{iT})$  for each  $i$  and take the log we get a conditional log-likelihood (conditional on  $y_{i0}$ ) for each  $i$ . We can estimate the parameters by maximizing the sum of the log-likelihoods across  $i$ .

c. As before, we can replace  $c_i$  with  $\alpha_0 + \alpha_1 y_{i0} + a_i$ . Then, the density of  $y_{it}$  given  $(y_{i,t-1}, \dots, y_{i1}, y_{i0}, a_i)$  is

$$\text{Normal}[\rho y_{i,t-1} + \alpha_0 + \alpha_1 y_{i0} + a_i + \delta(\alpha_0 + \alpha_1 y_{i0} + a_i) y_{i,t-1}, \sigma_e^2],$$

$t = 1, \dots, T$ . Using the same argument as in part b, we just integrate out  $a_i$  to get the density

of  $(y_{i1}, \dots, y_{iT})$  given  $y_{i0} = y_0$ :

$$\int_{-\infty}^{\infty} \left( \prod_{t=1}^T (2\pi\sigma_e^2)^{-1/2} \exp[-(y_t - \rho y_{t-1} - \alpha_0 - \alpha_1 y_{i0} - \mathbf{a} - \delta(\alpha_0 + \alpha_1 y_{i0} + \mathbf{a})y_{t-1})^2 / (2\sigma_e^2)] \right) \sigma_a^{-1} \phi(\mathbf{a}/\sigma_a) d\mathbf{a}.$$

Numerically, this could be a difficult MLE problem to solve. Assuming we can get the MLEs, we would estimate  $\rho + \delta E(c_i)$  as  $\hat{\rho} + \hat{\delta}(\hat{\alpha}_0 + \hat{\alpha}_1 \bar{y}_0)$ , where  $\bar{y}_0$  is the cross-sectional average of the initial observation.

d. The log likelihood for observation  $i$ , now conditional on  $(y_{i0}, \mathbf{z}_i)$ , is the log of

$$\int_{-\infty}^{\infty} \left( \prod_{t=1}^T (2\pi\sigma_e^2)^{-1/2} \exp[-(y_{it} - \rho y_{i,t-1} - \mathbf{z}_{it}\beta - \alpha_0 - \alpha_1 y_{i0} - \bar{\mathbf{z}}_i - \mathbf{a})^2 / (2\sigma_e^2)] \right) \sigma_a^{-1} \phi(\mathbf{a}/\sigma_a) d\mathbf{a}.$$

The assumption that we can put in the time average,  $\bar{\mathbf{z}}_i$ , to account for correlation between  $c_i$  and  $(y_{i0}, \mathbf{z}_i)$ , may be too strong. It may be better to put in the full vector  $\mathbf{z}_i$ , although this leads to many more parameters to estimate.

**13.12.** a. The first order conditions can be written as

$$\sum_{i=1}^N x_{ij} [y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}})] = 0, j = 1, \dots, K.$$

If, say, the first element of  $\mathbf{x}_i$  is unity,  $x_{i1} \equiv 1$ , then the first entry of the FOC is

$$\sum_{i=1}^N [y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}})] = 0$$

or

$$\sum_{i=1}^N \hat{u}_i = 0.$$

b. For the Bernoulli QLL with mean function  $\Lambda(\mathbf{x}\boldsymbol{\theta}) = \exp(\mathbf{x}\boldsymbol{\theta})/[1 + \exp(\mathbf{x}\boldsymbol{\theta})]$  it is easily

seen that the FOC is

$$\sum_{i=1}^N \mathbf{x}_i' [y_i - \Lambda(\mathbf{x}_i \hat{\boldsymbol{\theta}})] = 0, j = 1, \dots, K.$$

and so the canonical mean function is the logistic function. Therefore,  $g(\mu) = \Lambda^{-1}(\mu)$ , and to find  $\Lambda^{-1}(\mu)$  we need to solve for  $z$  as a function of  $\mu$  in

$\mu = \exp(z)/[1 + \exp(z)] = 1/[\exp(-z) + 1]$ . So

$$\exp(-z) = \frac{1}{\mu} - 1 = \frac{(1 - \mu)}{\mu}$$

or

$$\exp(z) = \frac{\mu}{(1 - \mu)}.$$

Now just take the log to get  $z = \log[\mu/(1 - \mu)]$ .

c. Generally, the FOC for the Poisson QMLE has the form

$$\sum_{i=1}^N \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \hat{\boldsymbol{\theta}})' [y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}})] / m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) = \mathbf{0}$$

and, with  $m(\mathbf{x}, \boldsymbol{\theta}) = \exp(\mathbf{x}\boldsymbol{\theta})$ , we get

$$\sum_{i=1}^N \exp(\mathbf{x}_i \hat{\boldsymbol{\theta}}) \mathbf{x}_i' [y_i - \exp(\mathbf{x}_i \hat{\boldsymbol{\theta}})] / \exp(\mathbf{x}_i \hat{\boldsymbol{\theta}}) = \mathbf{0},$$

or

$$\sum_{i=1}^N \mathbf{x}_i' [y_i - \exp(\mathbf{x}_i \hat{\boldsymbol{\theta}})] = \mathbf{0}.$$

So  $m(z) = \exp(z)$  is the canonical mean function and its inverse is  $g(\mu) = \log(\mu)$ .

d. This is a true statement. The score for observation  $i$  has the form

$$\mathbf{s}_i(\boldsymbol{\theta}) = \mathbf{x}_i' [y_i - m(\mathbf{x}_i, \boldsymbol{\theta})],$$

and therefore the Hessian,  $-\mathbf{x}_i' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta})$  – which has the form  $-r(\mathbf{x}_i, \boldsymbol{\theta}) \mathbf{x}_i' \mathbf{x}_i$  for some function  $r(\cdot) > 0$  – does not depend on  $y_i$ . If  $\boldsymbol{\theta}^*$  is the plim of  $\hat{\boldsymbol{\theta}}$  whether or not the mean is correctly specified, then a consistent estimator of  $-E[\mathbf{H}(\mathbf{x}_i, \boldsymbol{\theta}^*)]$  is

$$N^{-1} \sum_{i=1}^N r(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \mathbf{x}_i' \mathbf{x}_i = \mathbf{0}.$$

By contrast, with any other mean (link) function, the Hessian depends on  $y_i$ , and the estimators based on  $E[\mathbf{H}(\mathbf{x}_i, \boldsymbol{\theta}_o) | \mathbf{x}_i]$  under the assumption the mean is correctly specified are generally inconsistent.

**13.13.** In fact, there is nothing special about the QMLE setup for this problem: the conclusion holds for M-estimation. It is instructive to see the general argument.

To prove the result for general M-estimation (whether a minimization or maximization problem), use a mean value expansion and multiply through by  $N^{-1/2}$ :

$$N^{-1/2} \sum_{i=1}^N q(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = N^{-1/2} \sum_{i=1}^N q(\mathbf{w}_i, \boldsymbol{\theta}^*) + \left( N^{-1} \sum_{i=1}^N \mathbf{s}_i(\tilde{\boldsymbol{\theta}})' \right) \sqrt{N} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$$

where  $\mathbf{s}_i(\boldsymbol{\theta})$  is the  $P \times 1$  score and  $\tilde{\boldsymbol{\theta}}$  is on the line segment between  $\hat{\boldsymbol{\theta}}$  and  $\boldsymbol{\theta}^*$ . By Lemma 12.1,

$$N^{-1} \sum_{i=1}^N \mathbf{s}_i(\tilde{\boldsymbol{\theta}}) \xrightarrow{p} E[\mathbf{s}_i(\boldsymbol{\theta}^*)]$$

because  $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^*$ . Under the regularity conditions in Theorem 12.3,  $E[\mathbf{s}_i(\boldsymbol{\theta}^*)] = \mathbf{0}$ , and so

$N^{-1} \sum_{i=1}^N \mathbf{s}_i(\tilde{\boldsymbol{\theta}}) \xrightarrow{p} \mathbf{0}$ . We also know  $\sqrt{N} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = O_p(1)$ , and so



$$N^{-1/2} \sum_{i=1}^N q(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = N^{-1/2} \sum_{i=1}^N q(\mathbf{w}_i, \boldsymbol{\theta}^*) + o_p(1) \cdot O_p(1) = N^{-1/2} \sum_{i=1}^N q(\mathbf{w}_i, \boldsymbol{\theta}^*) + o_p(1).$$

**13.14 (Bonus Question).** Let  $\{f(y_t|\mathbf{x}_t; \theta) : t = 1, \dots, T\}$  be a sequence of correctly specified densities for  $y_{it}$  given  $\mathbf{x}_{it}$ . That is, assume that there is  $\theta_o \in \text{int}(\Theta)$  such that  $f(y_t|\mathbf{x}_t; \theta_o)$  is the density of  $y_{it}$  given  $\mathbf{x}_{it} = \mathbf{x}_t$ . Also assume that  $\{\mathbf{x}_{it} : t = 1, 2, \dots, T\}$  is strictly exogenous for each  $t$ :  $D(y_{it}|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}) = D(y_{it}|\mathbf{x}_{it})$ .

a. It is true that, under the standard regularity conditions for partial MLE, that

$$E[\mathbf{s}_{it}(\theta_o)|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = \mathbf{0}, \text{ where } \mathbf{s}_{it}(\theta_o) = \nabla_{\theta} \log f_t(y_{it}|\mathbf{x}_{it}; \theta)?$$

b. Under the assumptions given, is  $\{\mathbf{s}_{it}(\theta_o) : t = 1, \dots, T\}$  necessarily serially uncorrelated?

c. Let  $c_i$  be “unobserved heterogeneity” for cross section unit  $i$ , and assume that, for each  $t$ ,

$$D(y_{it}|\mathbf{z}_{i1}, \dots, \mathbf{z}_{iT}, c_i) = D(y_{it}|\mathbf{z}_{it}, c_i)$$

In other words,  $\{\mathbf{z}_{it} : t = 1, \dots, T\}$  is strictly exogenous conditional on  $c_i$ . Further, assume that

$$D(c_i|\mathbf{z}_{i1}, \dots, \mathbf{z}_{iT}) = D(c_i|\bar{\mathbf{z}}_i),$$

where  $\bar{\mathbf{z}}_i = T^{-1}(\mathbf{z}_{i1} + \dots + \mathbf{z}_{iT})$  is the vector of time averages. Assuming that well-behaved, correctly-specified conditional densities are available, how do we choose  $\mathbf{x}_{it}$  to make part a applicable?

### Solution

a. This is true because, by the general theory for partial MLE, we know that

$E[\mathbf{s}_{it}(\theta_o)|\mathbf{x}_{it}] = \mathbf{0}$ ,  $t = 1, \dots, T$ . But if  $D(y_{it}|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}) = D(y_{it}|\mathbf{x}_{it})$  then, for any function  $m_t(y_{it}, \mathbf{x}_{it})$ ,  $E[m_t(y_{it}, \mathbf{x}_{it})|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = E[m_t(y_{it}, \mathbf{x}_{it})|\mathbf{x}_{it}]$ , including the score function.

b. No. Strict exogeneity and complete dynamic specification of the conditional density are entirely different. Saying that  $D(y_{it}|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})$  does not depend on  $\mathbf{x}_{is}$ ,  $s \neq t$ , says nothing about whether  $y_{ir}$ ,  $r < t$ , appears in  $D(y_{it}|\mathbf{x}_{it}, y_{i,t-1}, \mathbf{x}_{i,t-1}, \dots, y_{i1}, \mathbf{x}_{i1})$ . Of course it is possible (if unlikely) for the score to be serially uncorrelated without complete dynamic specification, but

that is still a separate issue from strict exogeneity.

c. We take  $\mathbf{x}_{it} = (\mathbf{z}_{it}, \bar{\mathbf{z}}_i)$ ,  $t = 1, \dots, T$ . If  $g_t(y_t|\mathbf{z}_t, c; \boldsymbol{\gamma})$  is correctly specified for the density of  $y_{it}$  given  $(\mathbf{z}_{it} = \mathbf{z}_t, c_i = c)$ , and  $h(c|\bar{\mathbf{z}}; \boldsymbol{\delta})$  is correctly specified for the density of  $c_i$  given  $\bar{\mathbf{z}}_i = \bar{\mathbf{z}}$ , then the density of  $y_{it}$  given  $\mathbf{z}_i$  is obtained as

$$f_t(y_t|\mathbf{z}_i; \boldsymbol{\theta}_o) = \int_{\mathcal{C}} g_t(y_t|\mathbf{z}_{it}, c; \boldsymbol{\gamma}_o) h(c|\bar{\mathbf{z}}_i; \boldsymbol{\delta}_o) v(dc)$$

and this clearly depends only on  $(\mathbf{z}_{it}, \bar{\mathbf{z}}_i)$ . In other words, under the assumptions given,

$$D(y_{it}|\mathbf{z}_{i1}, \dots, \mathbf{z}_{iT}) = D(y_{it}|\mathbf{z}_{it}, \bar{\mathbf{z}}_i), t = 1, \dots, T$$

which implies

$$D(y_{it}|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}) = D(y_{it}|\mathbf{x}_{it}), t = 1, \dots, T.$$

Incidentally, we have not eliminated the serial dependence in  $\{y_{it}\}$  after only conditioning on  $(\mathbf{z}_{it}, \bar{\mathbf{z}}_i)$ : the part of  $c_i$  not explained by  $\bar{\mathbf{z}}_i$  affects  $y_{it}$  in each time period.

**13.15 (Bonus Question).** Consider the problem of estimating quantiles in a parametric context. In particular, write

$$y = \alpha_o + \mathbf{x}\boldsymbol{\beta}_o + u$$

$$D(u|\mathbf{x}) = \text{Normal}(0, \sigma_o^2 \exp(2\mathbf{x}\boldsymbol{\gamma}_o))$$

This means that  $\alpha_o + \mathbf{x}\boldsymbol{\beta}_o = E(y|\mathbf{x}) = \text{Med}(y|\mathbf{x})$ .

a. For  $0 < \tau < 1$  let  $\eta_\tau$  be the  $\tau^{\text{th}}$  quantile in the standard normal distribution. (So, for example,  $\eta_{.95} = 1.645$ .) Find  $\text{Quant}_\tau(y|\mathbf{x})$  in terms of  $\eta_\tau$  and all of the parameters. When is  $\text{Quant}_\tau(y|\mathbf{x})$  a linear function of  $\mathbf{x}$ ?

b. Given a random sample of size  $N$ , how would you estimate  $\text{Quant}_\tau(y|\mathbf{x})$  for a given  $\tau$ ?

c. Suppose we do not assume normality but use the weaker assumption that  $u/[\sigma_o \exp(\mathbf{x}\boldsymbol{\gamma}_o)]$

is independent of  $\mathbf{x}$ . Can we consistently estimate  $\text{Quant}_\tau(y|\mathbf{x})$  in this case?

**Solution**

a. First note that

$$\text{Quant}_\tau(y|\mathbf{x}) = \alpha_o + \mathbf{x}\boldsymbol{\beta}_o + \text{Quant}_\tau(u|\mathbf{x}).$$

Let  $r = u/[\sigma_o \exp(\mathbf{x}\boldsymbol{\gamma}_o)]$ , so that  $r$  is independent of  $\mathbf{x}$  with a  $\text{Normal}(0, 1)$  distribution. Because  $u$  has a strictly increasing cdf conditional on its quantile  $q_\tau(\mathbf{x})$  is the unique value such that

$$P[u \leq q_\tau(\mathbf{x})|\mathbf{x}] = \tau,$$

or

$$P\left[\frac{u}{\sigma_o \exp(\mathbf{x}\boldsymbol{\gamma}_o)} \leq \frac{q_\tau(\mathbf{x})}{\sigma_o \exp(\mathbf{x}\boldsymbol{\gamma}_o)} \middle| \mathbf{x}\right] = \tau$$

or

$$P\left[r \leq \frac{q_\tau(\mathbf{x})}{\sigma_o \exp(\mathbf{x}\boldsymbol{\gamma}_o)} \middle| \mathbf{x}\right] = \tau.$$

Because  $r$  is independent of  $\mathbf{x}$ , its  $\tau^{\text{th}}$  quantile conditional on  $\mathbf{x}$  is  $\eta_\tau$ . Therefore, we must have

$$\frac{q_\tau(\mathbf{x})}{\sigma_o \exp(\mathbf{x}\boldsymbol{\gamma}_o)} = \eta_\tau$$

or

$$q_\tau(\mathbf{x}) = \eta_\tau \sigma_o \exp(\mathbf{x}\boldsymbol{\gamma}_o).$$

So we have derived

$$\text{Quant}_\tau(y|\mathbf{x}) = \alpha_o + \mathbf{x}\boldsymbol{\beta}_o + \eta_\tau \sigma_o \exp(\mathbf{x}\boldsymbol{\gamma}_o).$$

$\text{Quant}_\tau(y|\mathbf{x})$  is linear in  $\mathbf{x}$  for  $\tau = .5$  because then  $\eta_\tau = 0$ . It is also linear in  $\mathbf{x}$  for any  $\tau$  if

$\boldsymbol{\gamma}_o = \mathbf{0}$ , in which case it can be written as

$$\text{Quant}_\tau(y|\mathbf{x}) = \alpha_o + \mathbf{x}\boldsymbol{\beta}_o + \eta_\tau \sigma_o.$$

Of course if  $\boldsymbol{\gamma}_o = \mathbf{0}$  then  $u$  and  $\mathbf{x}$  are independent, and the quantile functions for different  $\tau$  are parallel lines with different intercepts,  $\alpha_o + \eta_\tau \sigma_o$ .

b. Because we have specified

$$D(y|\mathbf{x}) = \text{Normal}(\alpha_o + \mathbf{x}\boldsymbol{\beta}_o, \sigma_o^2 \exp(2\mathbf{x}\boldsymbol{\gamma}_o))$$

we can use maximum likelihood to estimate all parameters, given a random sample of size  $N$ .

Then

$$\widehat{\text{Quant}}_\tau(y|\mathbf{x}) = \hat{\alpha} + \mathbf{x}\hat{\boldsymbol{\beta}} + \eta_\tau \hat{\sigma} \exp(\mathbf{x}\hat{\boldsymbol{\gamma}}).$$

c. The quantile function still has the form

$$\text{Quant}_\tau(y|\mathbf{x}) = \alpha_o + \mathbf{x}\boldsymbol{\beta}_o + \eta_\tau \sigma_o \exp(\mathbf{x}\boldsymbol{\gamma}_o)$$

but we must treat  $\eta_\tau$  as an unknown parameter because we do not know the distribution of  $r$ .

Note that the distribution of  $r$  may be asymmetric. The key restriction is that  $D(r|\mathbf{x})$  does not depend on  $\mathbf{x}$ .

We know  $\eta_\tau$  is the  $\tau^{th}$  quantile of the random variable  $r$ . If we observed  $\{r_i : i = 1, \dots, N\}$  we could estimate  $\eta_\tau$  as the  $\tau^{th}$  sample quantile of the  $r_i$ . Instead, we can use the standardized residuals

$$\hat{r}_i = \frac{\hat{u}_i}{\hat{\sigma} \exp(\mathbf{x}_i \hat{\boldsymbol{\gamma}})} = \frac{(y_i - \hat{\alpha} - \mathbf{x}_i \hat{\boldsymbol{\beta}})}{\hat{\sigma} \exp(\mathbf{x}_i \hat{\boldsymbol{\gamma}})}$$

and compute  $\hat{\eta}_\tau$  as the  $\tau^{th}$  sample quantile of  $\{\hat{r}_i : i = 1, \dots, N\}$ . Because  $\hat{\eta}_\tau$  solves the problem

$$\min_{h_\tau} \sum_{i=1}^N c_\tau(h_\tau - \hat{r}_i),$$

where  $c_\tau(\cdot)$  is the “check” function defined in Section 12.10, we can conclude  $\hat{\eta}_\tau$  is generally consistent using the consistency result for two-step M-estimators in Section 12.4. Of course, we have to have consistent estimators of the other parameters. From the results of Gouriéroux, Monfort, and Trognon (1984a), the normal QMLE is generally consistent for  $\alpha_o$ ,  $\beta_o$ ,  $\sigma_o$ , and  $\gamma_o$  even if normality does not hold. (As usual, we would need to use a sandwich covariance matrix estimator for inference on these parameters.) Obtaining a valid standard error for  $\hat{\eta}_\tau$ , and then getting the joint variance-covariance matrix of all parameter estimators, is challenging. Probably the nonparametric bootstrap is valid.

## Solutions to Chapter 14 Problems

**14.1.** a. The simplest way to estimate (14.35) is by 2SLS, using instruments  $(\mathbf{x}_1, \mathbf{x}_2)$ .

Nonlinear functions of these can be added to the instrument list, and they would generally improve efficiency if  $\gamma_2 \neq 1$ . If  $E(u_2^2|\mathbf{x}) = \sigma_2^2$ , 2SLS using the given list of instruments is the efficient, single equation GMM estimator. If there is heteroskedasticity an optimal weighting matrix that allows heteroskedasticity of unknown form should be used. Finally, one could try to use the optimal instruments derived in section 14.4.3. Even under homoskedasticity, these are difficult, if not impossible, to find analytically if  $\gamma_2 \neq 1$ .

With  $y_2 \geq 0$ , equation (14.35) is suspect as a structural equation because it is a linear model, and generally there are outcomes where  $\mathbf{x}_2\delta_2 + \gamma_3 y_1 + u_2 < 0$ .

b. No. If  $\gamma_1 = 0$  the parameter  $\gamma_2$  does not appear in the model. Of course, if we knew  $\gamma_1 = 0$ , we would consistently estimate  $\delta_1$  by OLS.

c. We can see this by obtaining  $E(y_1|\mathbf{x})$ :

$$\begin{aligned} E(y_1|\mathbf{x}) &= \mathbf{x}_1\delta_1 + \gamma_1 E(y_2^{\gamma_2}|\mathbf{x}) + E(u_1|\mathbf{x}) \\ &= \mathbf{x}_1\delta_1 + \gamma_1 E(y_2^{\gamma_2}|\mathbf{x}). \end{aligned}$$

Now, when  $\gamma_2 \neq 1$ ,  $E(y_2^{\gamma_2}|\mathbf{x}) \neq [E(y_2|\mathbf{x})]^{\gamma_2}$ , so we cannot write

$$E(y_1|\mathbf{x}) = \mathbf{x}_1\delta_1 + \gamma_1(\mathbf{x}_2\delta_2)^{\gamma_2};$$

in fact, we cannot find  $E(y_1|\mathbf{x})$  without more assumptions. While the regression  $y_2$  on  $\mathbf{x}_2$  consistently estimates  $\delta_2$ , the two-step NLS estimator of  $y_{i1}$  on  $\mathbf{x}_{i1}, (\mathbf{x}_i\hat{\delta}_2)^{\gamma_2}$  will not be consistent for  $\delta_1$  and  $\gamma_2$ . (This is an example of a “forbidden regression,” which we discussed in Chapter 9.) When  $\gamma_2 = 1$  and we impose this in estimation, we obtain the usual 2SLS estimator.

**14.2.** a. When  $\rho_1 = 1$ , we obtain the level-level model,  $hours = -\gamma_1 + \mathbf{z}_1\delta_1 + \gamma_1 wage + u_1$ .

Using the hint, let  $\rho_1 \rightarrow 0$  to get  $hours = \mathbf{z}_1\delta_1 + \gamma_1 \log(wage) + u_1$ .

b. We cannot use a standard  $t$  test after estimating the full model (say, by nonlinear 2SLS), because  $\rho_1$  cannot be estimated under  $H_0$ . The score test and QLR test also fail because of lack of identification under  $H_0$ . What we can do is *fix* a value for  $\rho_1$  – essentially our best guess – and then use a  $t$  test on  $(wage^{\rho_1} - 1)/\rho_1$  after linear 2SLS estimation (or GMM more generally). This need not be a very good test for detecting  $\gamma_1 \neq 0$  if our guess for  $\rho_1$  is not close to the actual value. There is a growing literature on testing hypotheses when parameters are not identified under the null.

c. If  $\text{Var}(u_1|\mathbf{z}) = \sigma_1^2$ , use nonlinear 2SLS, where we would use  $\mathbf{z}$  and functions of  $\mathbf{z}$  as IVs.

If we are not willing to assume homoskedasticity, GMM is generally more efficient.

d. The residual function is  $r(\boldsymbol{\theta}) = [hours - \mathbf{z}_1\delta_1 - \gamma_1(wage^{\rho_1} - 1)/\rho_1]$ , where

$\boldsymbol{\theta} = (\delta_1', \gamma_1, \rho_1)'$ . Using the hint the gradient is

$$\nabla_{\boldsymbol{\theta}} r(\boldsymbol{\theta}) = \{-\mathbf{z}_1, -(wage^{\rho_1} - 1)/\rho_1, \gamma_1[(wage^{\rho_1} - 1) - \rho_1 wage^{\rho_1} \log(wage)]/\rho_1^2\}.$$

The score is just the transpose.

e. Estimate  $\delta_1$  and  $\gamma_1$  by 2SLS, or use the GMM estimator that accounts for

heteroskedasticity, under the restriction  $\rho_1 = 1$ . Suppose the instruments are  $\mathbf{z}_i$ , a  $1 \times L$  vector.

This is just linear estimation because the model is linear under  $H_0$ . Then, taking  $\mathbf{Z}_i = \mathbf{z}_i$ , and

$$\begin{aligned} r_i(\tilde{\boldsymbol{\theta}}) &= [hours_i - \mathbf{z}_{i1}\tilde{\delta}_1 - \tilde{\gamma}_1(wage_i - 1)] \\ \nabla_{\boldsymbol{\theta}} r_i(\tilde{\boldsymbol{\theta}}) &= (-\mathbf{z}_{i1}, -(wage_i - 1), \tilde{\gamma}_1[(wage_i - 1) - wage_i \log(wage_i)]), \end{aligned}$$

use the score statistic in equation (14.32).

**14.3.** Let  $\mathbf{Z}_i^*$  be the  $G \times G$  matrix of optimal instruments in (14.57), where we suppress its



dependence on  $\mathbf{x}_i$ . Let  $\mathbf{Z}_i$  be the  $G \times L$  matrix that is a function of  $\mathbf{x}_i$  and let  $\mathbf{\Xi}_o$  the probability limit of the weighting matrix. Then the asymptotic variance of the GMM estimator has the form (14.10) with  $\mathbf{G}_o = E[\mathbf{Z}_i' \mathbf{R}_o(\mathbf{x}_i)]$ . So, in (14.48) take  $\mathbf{A} \equiv \mathbf{G}_o' \mathbf{\Xi}_o \mathbf{G}_o$  and  $\mathbf{s}(\mathbf{w}_i) \equiv \mathbf{G}_o' \mathbf{\Xi}_o \mathbf{Z}_i' \mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o)$ . The optimal score function is  $\mathbf{s}^*(\mathbf{w}_i) \equiv \mathbf{R}_o(\mathbf{x}_i)' \boldsymbol{\Omega}_o(\mathbf{x}_i)^{-1} \mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o)$ . Now we can verify (14.51) with  $\rho = 1$ :

$$\begin{aligned} E[\mathbf{s}(\mathbf{w}_1) \mathbf{s}^*(\mathbf{w}_1)'] &= \mathbf{G}_o' \mathbf{\Xi}_o E[\mathbf{Z}_i' \mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o) \mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o)' \boldsymbol{\Omega}_o(\mathbf{x}_i)^{-1} \mathbf{R}_o(\mathbf{x}_i)] \\ &= \mathbf{G}_o' \mathbf{\Xi}_o E[\mathbf{Z}_i' E\{\mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o) \mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o)' | \mathbf{x}_i\} \boldsymbol{\Omega}_o(\mathbf{x}_i)^{-1} \mathbf{R}_o(\mathbf{x}_i)] \\ &= \mathbf{G}_o' \mathbf{\Xi}_o E[\mathbf{Z}_i' \boldsymbol{\Omega}_o(\mathbf{x}_i) \boldsymbol{\Omega}_o(\mathbf{x}_i)^{-1} \mathbf{R}_o(\mathbf{x}_i)] = \mathbf{G}_o' \mathbf{\Xi}_o \mathbf{G}_o = \mathbf{A}. \end{aligned}$$

**14.4. a.** The residual function for the conditional mean model  $E(y_i | \mathbf{x}) = m(\mathbf{x}_i, \boldsymbol{\beta}_o)$  is  $r_i(\boldsymbol{\beta}) \equiv y_i - m(\mathbf{x}_i, \boldsymbol{\beta})$ . Then  $\Omega_o(\mathbf{x}_i)$  in (14.55) is just a scalar,  $\Omega_o(\mathbf{x}_i) = \text{Var}(y_i | \mathbf{x}_i) \equiv \omega_o(\mathbf{x}_i)$ . Under WNLS.3,  $\omega_o(\mathbf{x}_i) = \sigma_o^2 h(\mathbf{x}_i, \boldsymbol{\gamma}_o)$  for a known function  $h(\cdot)$ . Further,  $\mathbf{R}_o(\mathbf{x}_i) \equiv E[\nabla_{\boldsymbol{\beta}} \mathbf{r}_i(\boldsymbol{\beta}_o) | \mathbf{x}_i] = -\nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta}_o)$ , and so the optimal instruments are  $\nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta}_o) / \omega_o(\mathbf{x}_i)$ . The asymptotic variance of the efficient IV estimator is obtained from (14.60):

$$\{E[\nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta}_o)' [\omega_o(\mathbf{x}_i)]^{-1} \nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta}_o)]\}^{-1} = \sigma_o^2 \{E[\nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta}_o)' \nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta}_o) / h(\mathbf{x}_i, \boldsymbol{\gamma}_o)]\}^{-1},$$

which is the asymptotic variance of the WNLS estimator under WNLS.1, WNLS.2, and WNLS.3.

b. If  $\text{Var}(y_i | \mathbf{x}_i) = \sigma_o^2$  then NLS achieves the efficiency bound, as it seen by setting  $h(\mathbf{x}, \boldsymbol{\gamma}_o) \equiv 1$  in part a.

c. Now let  $r_{i1}(\boldsymbol{\theta}) \equiv u_i(\boldsymbol{\beta}) \equiv y_i - m(\mathbf{x}_i, \boldsymbol{\beta})$  and  $r_{i2}(\boldsymbol{\beta}, \sigma^2) = [y_i - m(\mathbf{x}_i, \boldsymbol{\beta})]^2 - \sigma^2$ . Let  $\mathbf{r}_i(\boldsymbol{\theta})$  denote the  $2 \times 1$  vector obtained by stacking the two residual functions. Then the moment conditions can be written as

$$E[\mathbf{r}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = \mathbf{0},$$

where  $\boldsymbol{\theta}_o = (\boldsymbol{\beta}'_o, \sigma_o^2)'$ . To obtain the efficient IVs, we first need  $E[\nabla_{\boldsymbol{\theta}} \mathbf{r}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i]$ . But

$$\nabla_{\boldsymbol{\theta}} \mathbf{r}_i(\boldsymbol{\theta}) = \begin{pmatrix} -\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}) & 0 \\ -2\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}) u_i(\boldsymbol{\beta}) & -1 \end{pmatrix}.$$

Evaluating at  $\boldsymbol{\theta}_o$  and using  $E[u_i(\boldsymbol{\beta}_o)|\mathbf{x}_i] = 0$  gives

$$\mathbf{R}_o(\mathbf{x}_i) \equiv \nabla_{\boldsymbol{\theta}} \mathbf{r}_i(\boldsymbol{\theta}) = \begin{pmatrix} -\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}) & 0 \\ 0 & -1 \end{pmatrix}.$$

We also need

$$\boldsymbol{\Omega}_o(\mathbf{x}_i) = E[r_i(\boldsymbol{\theta}_o) \mathbf{r}_i(\boldsymbol{\theta}_o)' | \mathbf{x}_i] = \begin{pmatrix} \sigma_o^2 & E(u_i^3 | \mathbf{x}_i) \\ E(u_i^3 | \mathbf{x}_i) & E(u_i^4 | \mathbf{x}_i) - \sigma_o^4 \end{pmatrix}$$

where  $u_i \equiv y_i - m(\mathbf{x}_i, \boldsymbol{\beta}_o)$ . The optimal IVs are  $[\boldsymbol{\Omega}_o(\mathbf{x}_i)]^{-1} \mathbf{R}_o(\mathbf{x}_i)$ . If  $E(u_i^3 | \mathbf{x}_i) = 0$ , as occurs under conditional symmetry of  $u_i$ , then the asymptotic variance matrix of the optimal IV estimator is block diagonal, and for  $\hat{\boldsymbol{\beta}}$  it is the same as NLS. In other words, adding the moment condition for the homoskedasticity assumption does not improve efficiency over NLS under symmetry, even if  $E(u_i^4 | \mathbf{x}_i)$  is not constant. But there is something subtle here. the NLS estimator is efficient in the class of estimators that only uses information on the first two conditional moments. If we use the information  $E(u_i^3 | \mathbf{x}_i) = 0$  then, in general, we could do better. But, of course, such an estimator would be less robust than NLS.

If, in addition,  $E(u_i^4 | \mathbf{x}_i)$  is constant, then the usual estimator of  $\sigma_o^2$  based on the sum of squared NLS residuals is efficient (among estimators that only use the first two conditional moments, but it happens that  $E(u_i^3 | \mathbf{x}_i) = 0$  and  $E(u_i^4 | \mathbf{x}_i)$  is constant).

**14.5.** We can write the unrestricted linear projection as

$$y_{it} = \pi_{t0} + \mathbf{x}_i \boldsymbol{\pi}_t + v_{it}, \quad t = 1, 2, 3$$

where  $\boldsymbol{\pi}_t$  is  $1 + 3K \times 1$ , and then  $\boldsymbol{\pi}$  is the  $3 + 9K \times 1$  vector obtained by stacking the  $\boldsymbol{\pi}_t$ . Let

$\boldsymbol{\theta} = (\psi, \boldsymbol{\lambda}'_1, \boldsymbol{\lambda}'_2, \boldsymbol{\lambda}'_3, \boldsymbol{\beta}')'$ . With the restrictions imposed on the  $\boldsymbol{\pi}_t$  we have

$$\begin{aligned} \pi_{t0} &= \psi, t = 1, 2, 3, \boldsymbol{\pi}_1 = [(\boldsymbol{\lambda}_1 + \boldsymbol{\beta})', \boldsymbol{\lambda}'_2, \boldsymbol{\lambda}'_3]' \\ \boldsymbol{\pi}_2 &= [\boldsymbol{\lambda}'_1, (\boldsymbol{\lambda}_2 + \boldsymbol{\beta})', \boldsymbol{\lambda}'_3]', \boldsymbol{\pi}_3 = [\boldsymbol{\lambda}'_1, \boldsymbol{\lambda}'_2, (\boldsymbol{\lambda}_3 + \boldsymbol{\beta})']' \end{aligned}$$

Therefore, we can write  $\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\theta}$  for the  $(3 + 9K) \times (1 + 4K)$  matrix  $\mathbf{H}$  defined by

$$\mathbf{H} = \begin{pmatrix} 1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_K & \mathbf{0} & \mathbf{0} & \mathbf{I}_K \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_K & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_K & \mathbf{0} \\ 1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_K & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_K & \mathbf{0} & \mathbf{I}_K \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_K & \mathbf{0} \\ 1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_K & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_K & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_K & \mathbf{I}_K \end{pmatrix}.$$

**14.6.** By this hint, it suffices to show that

$$[\text{Avar} \sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)]^{-1} - [\text{Avar} \sqrt{N}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)]^{-1}$$

is p.s.d. This difference is  $\mathbf{H}'_o \boldsymbol{\Xi}_o^{-1} \mathbf{H}_o - \mathbf{H}'_o \boldsymbol{\Lambda}_o^{-1} \mathbf{H}_o = \mathbf{H}'_o (\boldsymbol{\Xi}_o^{-1} - \boldsymbol{\Lambda}_o^{-1}) \mathbf{H}_o$ . This is positive

semi-definite if  $\boldsymbol{\Xi}_o^{-1} - \boldsymbol{\Lambda}_o^{-1}$  is p.s.d., which again holds by the hint because  $\boldsymbol{\Lambda}_o - \boldsymbol{\Xi}_o$  is assumed to be p.s.d.

**14.7.** With  $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\theta}$ , the minimization problem becomes

$$\min_{\theta \in \mathbb{R}^P} (\hat{\pi} - \mathbf{H}\theta)' \hat{\Xi}^{-1} (\hat{\pi} - \mathbf{H}\theta),$$

where it is assumed that no restrictions are placed on  $\theta$ . The first order condition is easily seen to be

$$-2\mathbf{H}'\hat{\Xi}^{-1}(\hat{\pi} - \mathbf{H}\theta) = \mathbf{0} \text{ or } (\mathbf{H}'\hat{\Xi}^{-1}\mathbf{H})\hat{\theta} = \mathbf{H}'\hat{\Xi}^{-1}\hat{\pi}.$$

Therefore, assuming  $\mathbf{H}'\hat{\Xi}^{-1}\mathbf{H}$  is nonsingular – which occurs w.p.a.1. when  $\mathbf{H}'\Xi_o^{-1}\mathbf{H}$  – is nonsingular – we have  $\hat{\theta} = (\mathbf{H}'\hat{\Xi}^{-1}\mathbf{H})^{-1}\mathbf{H}'\hat{\Xi}^{-1}\hat{\pi}$ .

**14.8.** From the efficiency discussion about maximum likelihood in Section 14.4.2, it is no less asymptotically efficient to use the density of  $(y_{i0}, y_{i1}, \dots, y_{iT})$  than to use the conditional distribution  $(y_{i1}, \dots, y_{iT})$  given  $y_{i0}$ . The cost of the asymptotic efficiency is that if we misspecify  $f_0(y_0; \theta)$ , then the unconditional MLE will generally be inconsistent for  $\theta_o$ . The MLE that conditions on  $y_{i0}$  is consistent provided we have the densities  $f_t(y_t|y_{t-1}; \theta)$  correctly specified,  $t \geq 1$ . As  $f_t(y_t|y_{t-1}; \theta)$  is the density of interest, we are usually willing to put more effort into testing our specification of it.

**14.9.** We have to verify equations (14.49) and (14.50) for the random effects and fixed effects estimators with . The choices of  $\mathbf{s}_{i1}$ ,  $\mathbf{s}_{i2}$  (with added  $i$  subscripts for clarity),  $\mathbf{A}_1$ , and  $\mathbf{A}_2$  are given in the hint. Now, from Chapter 10, we know that  $E(\mathbf{r}_i \mathbf{r}_i' | \mathbf{x}_i) = \sigma_u^2 \mathbf{I}_T$  under RE.1, RE.2, and RE.3, where  $\mathbf{r}_i = \mathbf{v}_i - \lambda \mathbf{j}_T \bar{v}_i$ . Therefore,

$$E(\mathbf{s}_{i1} \mathbf{s}_{i1}') = E(\check{\mathbf{X}}_i' \mathbf{r}_i \mathbf{r}_i' \check{\mathbf{X}}_i) = \sigma_u^2 E(\check{\mathbf{X}}_i' \check{\mathbf{X}}_i) \equiv \sigma_u^2 \mathbf{A}_1$$

by the usual iterated expectations argument. This means that, in (14.49),  $\rho \equiv \sigma_u^2$ . Now, we just need to verify (14.50) for this choice of  $\rho$ . But  $\mathbf{s}_{i2} \mathbf{s}_{i1}' = \check{\mathbf{X}}_i' \mathbf{u}_i \mathbf{r}_i' \check{\mathbf{X}}_i$  and, as described in the hint,

$$\check{\mathbf{X}}_i' \mathbf{r}_i = \check{\mathbf{X}}_i' (\mathbf{v}_i - \lambda \mathbf{j}_T \bar{v}_i) = \check{\mathbf{X}}_i' \mathbf{v}_i = \check{\mathbf{X}}_i' (c_j \mathbf{j}_T + \mathbf{u}_i) = \check{\mathbf{X}}_i' \mathbf{u}_i.$$

Therefore,  $\mathbf{s}_{i2}\mathbf{s}'_{i1} = \ddot{\mathbf{X}}'_i \mathbf{r}_i \mathbf{r}'_i \check{\mathbf{X}}_i$  and so

$$\mathbb{E}(\mathbf{s}_{i2}\mathbf{s}'_{i1}|\mathbf{x}_i) = \ddot{\mathbf{X}}'_i \mathbb{E}(\mathbf{r}_i \mathbf{r}'_i|\mathbf{x}_i) \check{\mathbf{X}}_i = \sigma_u^2 \ddot{\mathbf{X}}'_i \check{\mathbf{X}}_i$$

It follows that

$$\mathbb{E}(\mathbf{s}_{i2}\mathbf{s}'_{i1}) = \ddot{\mathbf{X}}'_i \mathbb{E}(\mathbf{r}_i \mathbf{r}'_i|\mathbf{x}_i) \check{\mathbf{X}}_i = \sigma_u^2 \mathbb{E}(\ddot{\mathbf{X}}'_i \check{\mathbf{X}}_i)$$

Finally,  $\ddot{\mathbf{X}}'_i \check{\mathbf{X}}_i = \ddot{\mathbf{X}}'_i (\mathbf{X}_i - \lambda \mathbf{j}_T \bar{\mathbf{x}}_i) = \ddot{\mathbf{X}}'_i \mathbf{X}_i = \ddot{\mathbf{X}}'_i \check{\mathbf{X}}_i$ , and so  $\mathbb{E}(\mathbf{s}_{i2}\mathbf{s}'_{i1}) = \sigma_u^2 \mathbb{E}(\ddot{\mathbf{X}}'_i \check{\mathbf{X}}_i)$ , and this verifies (14.50) with  $\rho = \sigma_u^2$ .

**14.10.** a. For each  $t$  we have

$$\begin{aligned} \mathbb{E}(v_{it}|\mathbf{x}_i) &= \mathbb{E}(\eta_t c_i + u_{it}|\mathbf{x}_i) = \mathbb{E}(\eta_t c_i|\mathbf{x}_i) + \mathbb{E}(u_{it}|\mathbf{x}_i) \\ &= \eta_t \mathbb{E}(c_i|\mathbf{x}_i) + \mathbb{E}(u_{it}|\mathbf{x}_i) \\ &= \eta_t \cdot 0 + 0 = 0 \end{aligned}$$

because  $\mathbb{E}(c_i|\mathbf{x}_i) = 0$  and  $\mathbb{E}(u_{it}|\mathbf{x}_i) = 0$  (because  $\mathbb{E}(u_{it}|\mathbf{x}_i, c_i) = 0$ ).

b. Under the assumptions – which are the same as Assumptions RE.1 and RE.3 – we know that

$$\begin{aligned} \text{Var}(u_{it}) &= \sigma_u^2, t = 1, \dots, T \\ \text{Cov}(c_i, u_{it}) &= 0, t = 1, \dots, T \\ \text{Cov}(u_{it}, u_{is}) &= 0, t \neq s. \end{aligned}$$

Therefore,

$$\begin{aligned} \text{Var}(v_{it}) &= \text{Var}(\eta_t c_i + u_{it}) = \eta_t^2 \sigma_c^2 + 2\eta_t \text{Cov}(c_i, u_{it}) + \sigma_u^2 \\ &= \eta_t^2 \sigma_c^2 + \sigma_u^2 \end{aligned}$$

and, for  $t \neq s$ ,

$$\begin{aligned} \text{Cov}(v_{it}, v_{is}) &= \text{Cov}(\eta_t c_i + u_{it}, \eta_s c_i + u_{is}) \\ &= \text{Cov}(\eta_t c_i, \eta_s c_i) + \text{Cov}(\eta_t c_i, u_{is}) + \text{Cov}(\eta_s c_i, u_{it}) + \text{Cov}(u_{it}, u_{is}) \\ &= \eta_t \eta_s \text{Cov}(c_i, c_i) = \eta_t \eta_s \sigma_c^2 \end{aligned}$$

c. The usual RE estimator treats the  $\eta_t$  as constant (which can then be normalized to be unity). In other words, it uses a misspecified model for  $\mathbf{\Omega} = \text{Var}(\mathbf{v}_i)$ . As we discussed in Chapter 10, the RE estimator is still consistent and  $\sqrt{N}$ -asymptotically normal, and we can conduct inference using a robust variance matrix estimator.

A more efficient estimator is, of course, FGLS with the correct form of the variance-covariance matrix. Write the  $T$  time periods for draw  $i$  as

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{v}_i$$

$$\text{E}(\mathbf{v}_i | \mathbf{x}_i) = \mathbf{0}$$

where the  $t^{\text{th}}$  row of  $\mathbf{X}_i$  is  $\mathbf{x}_{it}$ . The  $T \times T$  variance-covariance matrix (which is also the conditional on  $\mathbf{x}_i$ ) is

$$\text{Var}(\mathbf{v}_i) = \begin{pmatrix} \sigma_c^2 + \sigma_u^2 & \eta_2 \sigma_c^2 & \cdots & \eta_T \sigma_c^2 \\ \eta_2 \sigma_c^2 & \eta_2^2 \sigma_c^2 + \sigma_u^2 & \cdots & \eta_2 \eta_T \sigma_c^2 \\ \vdots & \vdots & \ddots & \vdots \\ \eta_T \sigma_c^2 & \eta_T \eta_2 \sigma_c^2 & \cdots & \eta_T^2 \sigma_c^2 + \sigma_u^2 \end{pmatrix},$$

where we impose the normalization  $\eta_1 = 1$ . The GLS estimator is

$$\hat{\boldsymbol{\beta}}_{GLS} = \left( \sum_{i=1}^N \mathbf{X}_i' \mathbf{\Omega}^{-1} \mathbf{X}_i \right)^{-1} \left( \sum_{i=1}^N \mathbf{X}_i' \mathbf{\Omega}^{-1} \mathbf{y}_i \right).$$

Of course, we would have to estimate  $\sigma_c^2$ ,  $\sigma_u^2$ , and  $\eta_2, \dots, \eta_T$ . One way to approach estimation of the variance-covariance parameters is to write

$$v_{it} v_{is} = \eta_t \eta_s \sigma_c^2 + d_{ts} \sigma_u^2 + r_{its}$$

$$\text{E}(r_{its}) = 0$$

for all  $t, s = 1, 2, \dots, T$ , where  $d_{ts}$  is a dummy variable equal to one if  $t = s$ , and zero otherwise.

Then we can estimate the parameters by pooled NLS after replacing  $v_{it} v_{is}$  with  $\check{v}_{it} \check{v}_{is}$ , where the

$\check{v}_{it}$  are perhaps the RE residuals (or they could be the POLS residuals). Note that  $\eta_1 \equiv 1$  is imposed. Then we can form  $\hat{\Omega}$  and then use FGLS.

**14.11.** First estimate initial parameters  $\pi_o$  from a set of linear reduced-form equations:

$$\mathbf{y}_i = \mathbf{X}_i \pi_o + \mathbf{u}_i$$

where  $\pi_o$  is  $K \times 1$  and unrestricted. Then estimate  $\pi_o$  by, say, system OLS. Or, if we assumed  $\text{Var}(\mathbf{u}_i | \mathbf{x}_i) = \Omega_o$  then FGLS would be no less asymptotically efficient.

Given  $\hat{\pi}$  and  $\hat{\Xi}$  consistent for

$$\Xi_o = \text{Avar}[\sqrt{N}(\hat{\pi} - \pi_o)],$$

the CMD estimator of  $\theta_o$ ,  $\hat{\theta}$ , solves

$$\min_{\theta \in \Theta} [\hat{\pi} - \mathbf{g}(\theta)]' \hat{\Xi}^{-1} [\hat{\pi} - \mathbf{g}(\theta)],$$

which is algebraically equivalent to a weighted multivariate nonlinear least squares problem where  $\hat{\pi}$  plays the role of the  $K \times 1$  vector of “dependent variables.” As discussed in the case where  $\mathbf{g}(\theta)$  is linear, the asymptotic analysis of the CMD estimator is different from the standard WMNLS problem: here  $K$  is fixed.

After estimation,  $\text{Avar}(\hat{\theta})$  is estimated as

$$(\hat{\mathbf{G}}' \hat{\Xi}^{-1} \hat{\mathbf{G}})^{-1} / N$$

where  $\hat{\mathbf{G}} \equiv \nabla_{\theta} \mathbf{g}(\hat{\theta})$ .

## Solutions to Chapter 15 Problems

**15.1.** a. Because the regressors are all orthogonal by construction – that is,  $dk_i \cdot dm_i = 0$  for  $k \neq m$ , and all  $i$  – the coefficient on  $dm_i$  is obtained from the regression  $y_i$  on  $dm_i, i = 1, \dots, N$ . But this is easily seen to be the fraction of ones in the sample falling into category  $m$  (because it is the average of  $y_i$  over the observations from category  $m$ ). Therefore, the fitted value for any  $i$  is the cell frequency for the appropriate category. These frequencies are all necessarily in  $[0,1]$ .

b. The fitted values for each category will be the same. If we drop  $dm_i$  but add an overall intercept, the overall intercept is the cell frequency for the first category, and the coefficient on  $dm_i$  becomes the difference in cell frequencies between category  $m$  and category one (the base category),  $m = 2, \dots, M$ .

**15.2.** a. First, because utility is increasing in both  $c$  and  $q$ , the budget constraint is binding at the optimum:  $c_i + p_i q_i = m_i$ . Plugging  $c = m_i - p_i q$  into the utility function reduces the problem to

$$\max_{q \geq 0} (m_i - p_i q) + a_i \log(1 + q).$$

Define utility as a function of  $q$ , as

$$s_i(q) \equiv (m_i - p_i q) + a_i \log(1 + q).$$

Then, for all  $q \geq 0$ ,

$$\frac{ds_i}{dq}(q) = -p_i + \frac{a_i}{1 + q}.$$

The optimal solution is  $q_i = 0$  if the marginal utility of charitable giving at  $q = 0$  is nonpositive, that is, if



$$\frac{ds_i}{dq}(0) = -p_i + a_i \leq 0 \text{ or } a_i \leq p_i.$$

(This can also be obtained by solving the Kuhn-Tucker conditions.) Thus, for this utility function,  $a_i$  can be interpreted as the reservation price above which no charitable contribution will be made; in other words, we have the corner solution  $q_i = 0$  whenever the price of charitable giving is too high relative to the marginal utility of charitable giving. On the other hand, if  $a_i > p_i$  then an interior solution exists ( $q_i > 0$ ) and necessarily solves the first order condition

$$\frac{ds_i}{dq}(q_i) = -p_i + \frac{a_i}{1 + q_i} \equiv 0$$

or

$$1 + q_i = a_i/p_i.$$

b. By definition of  $y_i$ ,  $y_i = 1$  if and only if  $a_i/p_i > 1$  or  $\log(a_i/p_i) > 0$ . If

$a_i = \exp(\mathbf{z}_i\boldsymbol{\gamma} + v_i)$ , the condition for  $y_i = 1$  is equivalent to  $\mathbf{z}_i\boldsymbol{\gamma} + v_i - \log p_i > 0$ . Therefore,

$$\begin{aligned} P(y_i = 1 | \mathbf{z}_i, m_i, p_i) &= P(y_i = 1 | \mathbf{z}_i, p_i) \\ &= P(\mathbf{z}_i\boldsymbol{\gamma} + v_i - \log p_i > 0 | \mathbf{z}_i, p_i) = P[v_i/\sigma > (-\mathbf{z}_i\boldsymbol{\gamma} + \log p_i)/\sigma] \\ &= 1 - G[(-\mathbf{z}_i\boldsymbol{\gamma} + \log p_i)/\sigma] = G[(\mathbf{z}_i\boldsymbol{\gamma} - \log p_i)\sigma], \end{aligned}$$

where the last equality follows by symmetry of the distribution of  $v_i/\sigma$ .

**15.3.** a. If  $P(y_i = 1 | \mathbf{z}_1, z_2) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2 + \gamma_2 z_2^2)$  then

$$\frac{\partial P(y = 1 | \mathbf{z}_1, z_2)}{\partial z_2} = (\gamma_1 + 2\gamma_2 z_2) \cdot \phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2 + \gamma_2 z_2^2);$$

for given  $\mathbf{z}$ , the partial effect is estimated as

$$(\hat{\gamma}_1 + 2\hat{\gamma}_2 z_2) \cdot \phi(\mathbf{z}_1\hat{\boldsymbol{\delta}}_1 + \hat{\gamma}_1 z_2 + \hat{\gamma}_2 z_2^2),$$

where, of course, the estimates are the probit estimates.

b. In the model

$$P(y_i = 1|z_i, z_2, d_1) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2 + \gamma_2 d_1 + \gamma_3 z_2 d_1),$$

the partial effect of  $z_2$  is

$$\frac{\partial P(y = 1|z_1, z_2, d_1)}{\partial z_2} = (\gamma_1 + \gamma_3 d_1) \cdot \phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2 + \gamma_2 d_1 + \gamma_3 z_2 d_1).$$

The effect of  $d_1$  is measured as the difference in the probabilities at  $d_1 = 1$  and  $d_1 = 0$  :

$$P(y = 1|z, d_1 = 1) - P(y = 1|z, d_1 = 0) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_2 + (\gamma_1 + \gamma_3)z_2) - \Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2).$$

Again, to estimate these effects at given  $\mathbf{z}$  and – in the first case,  $d_1$  – we just replace the parameters with their probit estimates, and use average or other interesting values of  $\mathbf{z}$ .

c. If the estimated partial effect is for particular values of  $(\mathbf{z}_1, z_2, d_1)$ , for example,

$$(\hat{\gamma}_1 + \hat{\gamma}_3 d_1^o) \cdot \phi(\mathbf{z}_1^o \boldsymbol{\delta}_1 + \hat{\gamma}_1 z_2^o + \hat{\gamma}_2 d_1^o + \hat{\gamma}_3 z_2^o d_1^o),$$

then we can apply the delta method from Chapter 3 (and referred to in Part III). Thus, we would require the full variance matrix of the probit estimates as well as the gradient of the expression of interest, such as  $(\gamma_1 + 2\gamma_3 z_2) \cdot \phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2 + \gamma_2 d_1 + \gamma_3 z_2 d_1)$ , with respect to all probit parameters. Alternatively, the bootstrap would be simpler but require a bit more computation.

If we are interested in the average partial effect (APE) of  $d_1$  going from zero to one then we estimate it as

$$N^{-1} \sum_{i=1}^N [\Phi(\mathbf{z}_{i1}\hat{\boldsymbol{\delta}}_1 + (\hat{\gamma}_1 + \hat{\gamma}_3)z_{i2} + \hat{\gamma}_2) - \Phi(\mathbf{z}_{i1}\hat{\boldsymbol{\delta}}_1 + \hat{\gamma}_1 z_{i2})],$$

that is, we estimate the effect for each unit  $i$  and then average these across all  $i$ . If we want a standard error for this, we would use the extension of the delta method worked out in Problem

12.17 – to account for the averaging as well as estimation of the parameters. The bootstrap can be used, too.

d. **(Bonus Question)** For a fixed value of  $z_2$ , say  $z_2^o$ , how would you estimate the average partial effect of  $d_1$  on the response probability?

### Solution

Now we average out only with respect to  $\mathbf{z}_{i1}$ :

$$N^{-1} \sum_{i=1}^N [\Phi(\mathbf{z}_{i1} \hat{\boldsymbol{\delta}}_1 + (\hat{\gamma}_1 + \hat{\gamma}_3)z_2^o + \hat{\gamma}_2) - \Phi(\mathbf{z}_{i1} \hat{\boldsymbol{\delta}}_1 + \hat{\gamma}_1 z_2^o)].$$

We can then vary  $z_2^o$  to see how the effect of changing  $d_1$  from zero to one varies with  $z_2^o$ .

Again, we can use Problem 12.17 to obtain an asymptotic standard error.

**15.4.** This is the kind of (nonsense) statement that arises out of failure to distinguish between the underlying latent variable model and the model for  $P(y = 1|\mathbf{x})$ . To compare the LPM and probit on equal footing, we must recognize that the LPM assumes  $P(y = 1|\mathbf{x}) = \mathbf{x}\boldsymbol{\gamma}$  while the probit model assumes that  $P(y = 1|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta})$ . So the substantive difference is purely in the functional forms for the response probabilities. And the probit functional form has some attractive properties compared with the linear model:  $\Phi(\mathbf{x}\boldsymbol{\beta})$  is always between zero and one, and the marginal effect of any  $x_j$  is diminishing after some point. The LPM and probit models are both approximations to the true response probability, and the LPM has some deficiencies for describing the partial effects over a broad range of the covariates.

If one insists on focusing on normality of the latent error in the probit case then one must compare that assumption with the the corresponding assumption for the LPM. If we specify a latent variable as  $y^* = \mathbf{x}\boldsymbol{\gamma} + e$  then the LPM is obtained when  $e$  has a uniform distribution over  $[-\alpha, \alpha]$  for some constant  $0 < \alpha < \infty$ . For most purposes, this is much less

plausible than the normality underlying probit.

**15.5.** a. If  $P(y = 1|\mathbf{z}, q) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2 q)$  then

$$\frac{\partial P(y = 1|\mathbf{z}, q)}{\partial z_2} = \gamma_1 q \cdot \phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2 q),$$

assuming that  $z_2$  is not functionally related to  $\mathbf{z}_1$ .

b. Write  $y^* = \mathbf{z}_1\boldsymbol{\delta}_1 + r$ , where  $r = \gamma_1 z_2 q + e$ , and  $e$  is independent of  $(\mathbf{z}, q)$  with a standard normal distribution. Because  $q$  is assumed independent of  $\mathbf{z}$ ,  $q|\mathbf{z} \sim \text{Normal}(0, \gamma_1^2 z_2^2 + 1)$ ; this follows because  $E(r|\mathbf{z}) = \gamma_1 z_2 E(q|\mathbf{z}) + E(e|\mathbf{z}) = 0$ . Also,

$$\text{Var}(r|\mathbf{z}) = \gamma_1^2 z_2^2 \text{Var}(q|\mathbf{z}) + \text{Var}(e|\mathbf{z}) + 2\gamma_1 z_2 \text{Cov}(q, e|\mathbf{z}) = \gamma_1^2 z_2^2 + 1$$

because  $\text{Cov}(q, e|\mathbf{z}) = 0$  by independence between  $e$  and  $(\mathbf{z}, q)$ . Thus,  $r/\sqrt{\gamma_1^2 z_2^2 + 1}$  has a standard normal distribution independent of  $\mathbf{z}$ . It follows that

$$P(y = 1|\mathbf{z}) = \Phi\left(\mathbf{z}_1\boldsymbol{\delta}_1 / \sqrt{\gamma_1^2 z_2^2 + 1}\right). \quad (15.97)$$

c. Because  $P(y = 1|\mathbf{z})$  depends only on  $\gamma_1^2$ , this is what we can estimate along with  $\boldsymbol{\delta}_1$ . (For example,  $\gamma_1 = -2$  and  $\gamma_1 = 2$  give exactly the same model for  $P(y = 1|\mathbf{z})$ .) This is why we define  $\rho_1 = \gamma_1^2$ . Testing  $H_0 : \rho_1 = 0$  is most easily done using the score or LM test because, under  $H_0$ , we have a standard probit model.

Let  $\hat{\boldsymbol{\delta}}_1$  denote the probit estimates under the null that  $\rho_1 = 0$ . Define  $\phi_i = \phi(\mathbf{z}_{i1}\hat{\boldsymbol{\delta}}_1)$ ,  $\hat{\Phi}_i = \Phi(\mathbf{z}_{i1}\hat{\boldsymbol{\delta}}_1)$ ,  $\hat{u}_i = y_i - \hat{\Phi}_i$ , and  $\tilde{u}_i = \hat{u}_i / \sqrt{\hat{\Phi}_i(1 - \hat{\Phi}_i)}$  (the standardized residuals). The gradient of the mean function in (15.97) with respect to  $\boldsymbol{\delta}_1$ , evaluated under the null estimates, is simply  $\hat{\phi}_i \mathbf{z}_{i1}$ . The only other quantity needed is the gradient with respect to  $\rho_1$  evaluated at the null estimates. But the partial derivative of (15.97) with respect to  $\rho_1$  is, for each  $i$ ,

$$-(\mathbf{z}_{i1}\boldsymbol{\delta}_1)(z_{i2}^2/2)(\rho_1 z_{i2}^2 + 1)^{-3/2} \phi\left(\mathbf{z}_{i1}\boldsymbol{\delta}_1/\sqrt{\gamma_1^2 z_{i2}^2 + 1}\right).$$

When we evaluate this at  $\rho_1 = 0$  and  $\hat{\boldsymbol{\delta}}_1$  we get  $-(\mathbf{z}_{i1}\hat{\boldsymbol{\delta}}_1)(z_{i2}^2/2)\hat{\phi}_i$ . Then, the score statistic can be obtained as  $NR_u^2$  from the regression

$$\tilde{u}_i \text{ on } \frac{\hat{\phi}_i \mathbf{z}_{i1}}{\sqrt{\hat{\Phi}_i(1 - \hat{\Phi}_i)}}, \frac{(\mathbf{z}_{i1}\hat{\boldsymbol{\delta}}_1)z_{i2}^2\hat{\phi}_i}{\sqrt{\hat{\Phi}_i(1 - \hat{\Phi}_i)}};$$

under  $H_0$ ,  $NR_u^2 \stackrel{a}{\sim} \chi_1^2$ .

d. The model can be estimated by MLE using the formulation with  $\rho_1$  in place of  $\gamma_1^2$ . It is not a standard probit estimation but a kind of “heteroskedastic probit.”

**15.6.** a. What we would like to know is that, if we exogenously change the number of cigarettes that someone smokes per day, what effect would this have on the probability of missing work over a three-month period? In other words, we want to infer causality, not just find a correlation between missing work and cigarette smoking.

b. Since people choose whether and how much to smoke, we certainly cannot treat the data as coming from the experiment we have in mind in part a. (That is, we cannot randomly assign people a daily cigarette consumption.) It is possible that smokers are less healthy to begin with, or have other attributes that cause them to miss work more often. Or, it could go the other way: cigarette consumption may be related to personality traits that make people harder workers. In any case, *cigs* might be correlated with the unobservables in the equation.

c. If we start with the model

$$P(y = 1|\mathbf{z}, cigs, q_1) = \Phi(\mathbf{z}_{i1}\boldsymbol{\delta}_1 + \gamma_1 cigs + q_1), \quad (15.98)$$

but ignore  $q_1$  when it is correlated with *cigs*, we will not consistently estimate anything of interest, whether the model is linear or nonlinear. Thus, we would not be estimating a causal

effect. If  $q_1$  is independent of *cigs*, the probit ignoring  $q_1$  does estimate the average partial effect of another cigarette.

d. No. There are many people in the working population who do not smoke. Thus, the distribution (conditional or unconditional) of *cigs* piles up at zero. Also, since *cigs* takes on integer values, it cannot be normally distributed. But it is really the pile up at zero that is the most serious issue.

e. Use the Rivers-Vuong test. Obtain the residuals,  $\hat{r}_2$ , from the regression *cigs* on **z**. Then, estimate the probit of *y* on **z**<sub>1</sub>, *cigs*,  $\hat{r}_2$  and use a standard *t* test on  $\hat{r}_2$ . This does not rely on normality of  $r_2$  (or *cigs*). It does, of course, rely on the probit model being correct for *y* under  $H_0$ .

f. Assuming people will not immediately move out of their state of residence when the state implements no smoking laws in the workplace, and that state of residence is roughly independent of general health in the population, a dummy indicator for whether the person works in a state with a new law can be treated as exogenous and excluded from (15.98). (These situations are often called “natural experiments.”) Further, *cigs* is likely to be correlated with the state law indicator because since people will not be able to smoke as much as they otherwise would. Thus, it seems to be a reasonable instrument for *cigs*.

**15.7.** a. The LPM estimates, with the usual and heteroskedasticity-robust standard errors, are given below. Interesting, the robust standard errors on the non-demographic variables are often notably smaller than the usual standard errors. The statistical significance of the OLS coefficients is the same using either set of standard errors.

When *pcnv* goes from .25 to .75, the estimated probability of arrest falls by about .077, or 7.7 percentage points.

```
. use grogger
```

```
. gen arr86 = 0
```

```
. replace arr86 = 1 if narr86 > 0  
(755 real changes made)
```

```
. reg arr86 pcnv avgsen tottime ptime86 inc86 black hispan born60
```

Source	SS	df	MS	Number of obs =	2725
Model	44.9720916	8	5.62151145	F( 8, 2716) =	30.48
Residual	500.844422	2716	.184405163	Prob > F =	0.0000
				R-squared =	0.0824
				Adj R-squared =	0.0797
Total	545.816514	2724	.20037317	Root MSE =	.42942

arr86	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
pcnv	-.1543802	.0209336	-7.37	0.000	-.1954275	-.1133329
avgsen	.0035024	.0063417	0.55	0.581	-.0089326	.0159374
tottime	-.0020613	.0048884	-0.42	0.673	-.0116466	.007524
ptime86	-.0215953	.0044679	-4.83	0.000	-.0303561	-.0128344
inc86	-.0012248	.000127	-9.65	0.000	-.0014738	-.0009759
black	.1617183	.0235044	6.88	0.000	.1156299	.2078066
hispan	.0892586	.0205592	4.34	0.000	.0489454	.1295718
born60	.0028698	.0171986	0.17	0.867	-.0308539	.0365936
_cons	.3609831	.0160927	22.43	0.000	.329428	.3925382

```
. reg arr86 pcnv avgsen tottime ptime86 inc86 black hispan born60, robust
```

Linear regression

```
Number of obs = 2725  
F( 8, 2716) = 37.59  
Prob > F = 0.0000  
R-squared = 0.0824  
Root MSE = .42942
```

arr86	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
pcnv	-.1543802	.018964	-8.14	0.000	-.1915656	-.1171948
avgsen	.0035024	.0058876	0.59	0.552	-.0080423	.0150471
tottime	-.0020613	.0042256	-0.49	0.626	-.010347	.0062244
ptime86	-.0215953	.0027532	-7.84	0.000	-.0269938	-.0161967
inc86	-.0012248	.0001141	-10.73	0.000	-.0014487	-.001001
black	.1617183	.0255279	6.33	0.000	.1116622	.2117743
hispan	.0892586	.0210689	4.24	0.000	.0479459	.1305714
born60	.0028698	.0171596	0.17	0.867	-.0307774	.036517
_cons	.3609831	.0167081	21.61	0.000	.3282214	.3937449

```
. di .5*_b[pcnv]  
-.0771901
```

b. The robust statistic and its  $p$ -value are gotten by using the “test” command after

appending “robust” to the regression command. The  $p$ -values are virtually identical.

```
. qui reg arr86 pcnv avgsen tottime ptime86 inc86 black hispan born60

. test avgsen tottime

( 1)  avgsen = 0
( 2)  tottime = 0

      F( 2, 2716) =    0.18
      Prob > F =    0.8360

. qui reg arr86 pcnv avgsen tottime ptime86 inc86 black hispan born60, robust

. test avgsen tottime

( 1)  avgsen = 0
( 2)  tottime = 0

      F( 2, 2716) =    0.18
      Prob > F =    0.8320
```

c. The probit model estimates follow.

```
. probit arr86 pcnv avgsen tottime ptime86 inc86 black hispan born60

Probit regression                                Number of obs   =      2725
                                                LR chi2(8)      =      249.09
                                                Prob > chi2     =      0.0000
Log likelihood = -1483.6406                    Pseudo R2      =      0.0774
```

arr86	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
pcnv	-.5529248	.0720778	-7.67	0.000	-.6941947	-.4116549
avgsen	.0127395	.0212318	0.60	0.548	-.028874	.0543531
tottime	-.0076486	.0168844	-0.45	0.651	-.0407414	.0254442
ptime86	-.0812017	.017963	-4.52	0.000	-.1164085	-.0459949
inc86	-.0046346	.0004777	-9.70	0.000	-.0055709	-.0036983
black	.4666076	.0719687	6.48	0.000	.3255516	.6076635
hispan	.2911005	.0654027	4.45	0.000	.1629135	.4192875
born60	.0112074	.0556843	0.20	0.840	-.0979318	.1203466
_cons	-.3138331	.0512999	-6.12	0.000	-.4143791	-.213287

Now, we must compute the difference in the normal cdf at the two different values of *pcnv*,

*black* = 1, *hispan* = 0, *born60* = 1, and at the average values of remaining variables.

```
. sum avgsen tottime ptime86 inc86
```

Variable	Obs	Mean	Std. Dev.	Min	Max
avgsen	2725	.6322936	3.508031	0	59.2
tottime	2725	.8387523	4.607019	0	63.4



```

ptime86 |      2725      .387156      1.950051      0      12
inc86   |      2725      54.96705      66.62721      0      541

. di normal(_b[_cons] + _b[pcnv]*.75 + _b[avgsen]*.6322936
+ _b[totttime]*.8387523 + _b[ptime86]* .387156 + _b[inc86]*54.96705
+ _b[black] + _b[born60])
- normal(_b[_cons] + _b[pcnv]*.25 + _b[avgsen]*.6322936
+ _b[totttime]*.8387523 + _b[ptime86]* .387156 + _b[inc86]* 54.96705
+ _b[black] + _b[born60])

-.10166064

```

This last command shows that the probability falls by about .102, which is somewhat larger than the effect obtained from the LPM.

d. To obtain the percent correctly predicted for each outcome, we first generate the predicted values of *arr86* as described on page 465:

```

. predict PHIhat
(option pr assumed; Pr(arr86))

. gen arr86t = PHIhat >= .5

. tab arr86t arr86

```

arr86t	arr86		Total
	0	1	
0	1,903	677	2,580
1	67	78	145
Total	1,970	755	2,725

```

. di 1903/1970
.96598985

. di 78/755
.10331126

. di (1903 + 78)/2725
.72697248

```

For men who were not arrested, the probit predicts correctly about 96.6% of the time. Unfortunately, for the men who were arrested, the probit is correct only about 10.3% of the time. The overall percent correctly predicted is pretty high – 72.7% – but we cannot very well predict the outcome we would most like to predict.

e. Adding the quadratic terms gives

```
. probit arr86 pcnv avgsen tottime ptime86 inc86 black hispan born60
      pcnvsq pt86sq inc86sq

Probit regression                                Number of obs   =       2725
                                                LR chi2(11)         =       336.77
                                                Prob > chi2         =       0.0000
Log likelihood = -1439.8005                    Pseudo R2          =       0.1047
```

arr86	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
pcnv	.2167615	.2604937	0.83	0.405	-.2937968	.7273198
avgsen	.0139969	.0244972	0.57	0.568	-.0340166	.0620105
tottime	-.0178158	.0199703	-0.89	0.372	-.056957	.0213253
ptime86	.7449712	.1438485	5.18	0.000	.4630333	1.026909
inc86	-.0058786	.0009851	-5.97	0.000	-.0078094	-.0039478
black	.4368131	.0733798	5.95	0.000	.2929913	.580635
hispan	.2663945	.067082	3.97	0.000	.1349163	.3978727
born60	-.0145223	.0566913	-0.26	0.798	-.1256351	.0965905
pcnvsq	-.8570512	.2714575	-3.16	0.002	-1.389098	-.3250042
pt86sq	-.1035031	.0224234	-4.62	0.000	-.1474522	-.059554
inc86sq	8.75e-06	4.28e-06	2.04	0.041	3.63e-07	.0000171
_cons	-.337362	.0562665	-6.00	0.000	-.4476423	-.2270817

Note: 51 failures and 0 successes completely determined.

```
. test pcnvsq pt86sq inc86sq

( 1) pcnvsq = 0
( 2) pt86sq = 0
( 3) inc86sq = 0

      chi2( 3) =    38.54
      Prob > chi2 =    0.0000
```

The quadratics are individually and jointly significant. The quadratic in *pcnv* means that, at low levels of *pcnv*, there is actually a positive relationship between probability of arrest and *pcnv*, which does not make much sense. The turning point is easily found as

$.217/(2(.857)) \approx .127$ , and there are many cases – 1,265 – where *pcnv* is less than .127.

```
. sum pcnv

Variable |      Obs      Mean      Std. Dev.      Min      Max
-----+-----
pcnv     |    2725     .3577872     .395192         0         1

. count if pcnv < .127
1265
```

**15.8. a.** The following Stata session answers this part. The difference in estimated probabilities of smoking at 16 and 12 years of education is about  $-.080$ . In other words, for

non-white women at the average family income, women with 16 years of education are, on average, about eight percentage points less likely to smoke.

```
. use bwght
```

```
. gen smokes = cigs > 0
```

```
. tab smokes
```

smokes	Freq.	Percent	Cum.
0	1,176	84.73	84.73
1	212	15.27	100.00
Total	1,388	100.00	

```
. probit smokes motheduc white lfaminc
```

```
Probit regression                                Number of obs   =       1387
                                                LR chi2(3)      =       92.67
                                                Prob > chi2     =       0.0000
Log likelihood = -546.76991                    Pseudo R2      =       0.0781
```

smokes	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
motheduc	-.1450599	.0207899	-6.98	0.000	-.1858074	-.1043124
white	.1896765	.1098805	1.73	0.084	-.0256853	.4050383
lfaminc	-.1669109	.0498894	-3.35	0.001	-.2646923	-.0691296
_cons	1.126276	.2504611	4.50	0.000	.6353817	1.617171

```
. sum faminc
```

Variable	Obs	Mean	Std. Dev.	Min	Max
faminc	1388	29.02666	18.73928	.5	65

```
. di normal(_b[_cons] + _b[motheduc]*16 + _b[lfaminc]*log(29.02666))
    - normal(_b[_cons] + _b[motheduc]*12 + _b[lfaminc]*log(29.02666))
-.08020112
```

b. The variance *faminc* is probably not exogenous because, at a minimum, income is likely correlated with quality of health care. It might also be correlated with unobserved cultural factors that are correlated with smoking.

c. The reduced form equation for *lfaminc* is estimated below. As expected, *fatheduc* has a positive partial effect on *lfaminc*, and the relationship is statistically significant. We need the

residuals from this regression for part d. We lose 196 observations due to missing data on *fatheduc*, and one observation has already been lost due to a missing value for *motheduc*.

```
. reg lfaminc motheduc white fatheduc
```

Source	SS	df	MS	Number of obs =	1191
Model	140.936735	3	46.9789115	F( 3, 1187) =	119.23
Residual	467.690904	1187	.394010871	Prob > F =	0.0000
				R-squared =	0.2316
				Adj R-squared =	0.2296
Total	608.627639	1190	.511451797	Root MSE =	.6277

lfaminc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
motheduc	.0709044	.0098338	7.21	0.000	.0516109	.090198
white	.3452115	.050418	6.85	0.000	.2462931	.4441298
fatheduc	.0616625	.008708	7.08	0.000	.0445777	.0787473
_cons	1.241413	.1103648	11.25	0.000	1.024881	1.457945

```
. predict v2hat, resid
(197 missing values generated)
```

d. To test the null of exogeneity, we estimate the probit that includes  $\hat{v}_2$ :

```
. probit smokes motheduc white lfaminc v2hat
```

```
Iteration 0: log likelihood = -471.77574
Iteration 1: log likelihood = -432.90303
Iteration 2: log likelihood = -432.0639
Iteration 3: log likelihood = -432.06242
Iteration 4: log likelihood = -432.06242
```

Probit regression	Number of obs =	1191
	LR chi2(4) =	79.43
	Prob > chi2 =	0.0000
Log likelihood = -432.06242	Pseudo R2 =	0.0842

smokes	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
motheduc	-.0826247	.0465204	-1.78	0.076	-.173803	.0085536
white	.4611075	.1965245	2.35	0.019	.0759265	.8462886
lfaminc	-.7622559	.3652949	-2.09	0.037	-1.478221	-.046291
v2hat	.6107298	.3708071	1.65	0.100	-.1160387	1.337498
_cons	1.98796	.5996374	3.32	0.001	.8126927	3.163228

There is not strong evidence of endogeneity, but the sign of the coefficient on *v2hat* is what we expect: unobservables that lead to higher income are positively correlated with unobserved

factors affecting birth weight. There is a further problem in that using this test presumes *fatheduc* can be omitted from the birth weight equation. Remember, the test can be interpreted as a test for endogeneity of *lfaminc* only when we maintain that *fatheduc* is exogenous.

Because of the potential endogeneity of this is perhaps not a very good example, but it shows you how to mechanically carry out the tests.

Incidentally, the probit coefficients on *lfaminc* are very different depending on whether we treat it as exogenous or not. This is true even if we use the same samples, as the Stata output below shows. The APE is probably quite different, too. It is hard to know what to do in such cases (which are all too common).

```
. probit smokes motheduc white lfaminc if v2hat != .
```

```
Probit regression                               Number of obs   =       1191
                                                LR chi2(3)      =       76.72
                                                Prob > chi2     =       0.0000
Log likelihood = -433.41656                    Pseudo R2      =       0.0813
```

smokes	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
motheduc	-.1497589	.0225634	-6.64	0.000	-.1939823	-.1055355
white	.2323285	.137875	1.69	0.092	-.0379015	.5025584
lfaminc	-.1719479	.0687396	-2.50	0.012	-.306675	-.0372207
_cons	1.133026	.2990124	3.79	0.000	.5469727	1.71908

15.9. a. Let  $P(y = 1|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$ , where  $x_1 = 1$ . Then for each  $i$ ,

$$\ell_i(\boldsymbol{\beta}) = y_i \log(\mathbf{x}_i \boldsymbol{\beta}) + (1 - y_i) \log(1 - \mathbf{x}_i \boldsymbol{\beta}),$$

which is only well-defined for  $0 < \mathbf{x}_i \boldsymbol{\beta} < 1$ .

b. For any possible estimate  $\hat{\boldsymbol{\beta}}$ , the log-likelihood function is well-defined only if  $0 < \mathbf{x}_i \hat{\boldsymbol{\beta}} < 1$  for all  $i = 1, \dots, N$ . Therefore, during the iterations to obtain the MLE, this condition must be checked. It may be impossible to find an estimate that satisfies these inequalities for every observation, especially if  $N$  is large.

c. This follows from the KLIC, and the discussion of Vuong's model selection statistic in Section 13.11.2: the true density of  $y$  given  $\mathbf{x}$  – evaluated at the true values, of course – maximizes the KLIC. Because the MLEs are consistent for the unknown parameters, asymptotically the true density will produce the highest average log-likelihood function. So, just as we can use an  $R$ -squared to choose among different functional forms for  $E(y|\mathbf{x})$ , we can use values of the log-likelihood to choose among different models for  $P(y = 1|\mathbf{x})$  when  $y$  is binary.

**15.10.** a. There are several possibilities. One is to define  $\hat{p}_i = \hat{P}(y = 1|\mathbf{x}_i)$  – the estimated response probabilities – and obtain the square of the correlation between  $y_i$  and  $\hat{p}_i$ . For the LPM, this is just the usual  $R$ -squared. For the general index model,  $G(\mathbf{x}_i\hat{\boldsymbol{\beta}})$  is the estimate of  $E(y|\mathbf{x}_i)$ , and so it makes sense to compute an analogous goodness-of-fit measure. This measure is always between zero and one.

An alternative is to use the sum of squared residuals form. While this produces the same  $R$ -squared measure for the linear model, it does not for nonlinear models.

b. The Stata output below gives the square of the correlation between  $y_i$  and the fitted probabilities for the LPM and probit. The LPM  $R$ -squared is about .106 and that for probit is higher, about .115. So probit is preferred based on this goodness-of-fit measure, although the improvement is not overwhelming. (It is about an 8.5% increase in the  $R$ -squared.)

```
. reg arr86 pcnv avgsen tottime ptime86 inc86 black hispan born60
    pcnvsq pt86sq inc86sq
```

Source	SS	df	MS	Number of obs =	2725
Model	57.8976285	11	5.26342077	F( 11, 2713) =	29.27
Residual	487.918885	2713	.179844779	Prob > F =	0.0000
Total	545.816514	2724	.20037317	R-squared =	0.1061
				Adj R-squared =	0.1025
				Root MSE =	.42408

arr86	Coef.	Std. Err.	t	P> t	[95% Conf. Interval
-------	-------	-----------	---	------	---------------------

pcnv	.075977	.0803402	0.95	0.344	-.0815573	.2335112
avgsen	.0012998	.0062692	0.21	0.836	-.0109932	.0135927
totttime	-.0022213	.0048287	-0.46	0.646	-.0116897	.007247
ptime86	.1321786	.0230021	5.75	0.000	.0870752	.177282
inc86	-.0018505	.0002737	-6.76	0.000	-.0023872	-.0013139
black	.1447942	.0233225	6.21	0.000	.0990627	.1905258
hispan	.0803938	.0204959	3.92	0.000	.0402047	.1205829
born60	-.0062993	.0170252	-0.37	0.711	-.039683	.0270843
pcnvsq	-.2456865	.0812584	-3.02	0.003	-.4050211	-.0863519
pt86sq	-.0139981	.0020109	-6.96	0.000	-.0179411	-.0100551
inc86sq	3.31e-06	1.09e-06	3.03	0.002	1.17e-06	5.45e-06
_cons	.363352	.0175536	20.70	0.000	.3289323	.3977718

```
. probit arr86 pcnv avgsen tottime ptime86 inc86 black hispan born60
pcnvsq pt86sq inc86sq
```

Probit regression	Number of obs	=	2725
	LR chi2(11)	=	336.77
	Prob > chi2	=	0.0000
Log likelihood = -1439.8005	Pseudo R2	=	0.1047

arr86	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
pcnv	.2167615	.2604937	0.83	0.405	-.2937968	.7273198
avgsen	.0139969	.0244972	0.57	0.568	-.0340166	.0620105
totttime	-.0178158	.0199703	-0.89	0.372	-.056957	.0213253
ptime86	.7449712	.1438486	5.18	0.000	.4630332	1.026909
inc86	-.0058786	.0009851	-5.97	0.000	-.0078094	-.0039478
black	.4368131	.0733798	5.95	0.000	.2929913	.580635
hispan	.2663945	.067082	3.97	0.000	.1349163	.3978727
born60	-.0145223	.0566913	-0.26	0.798	-.1256351	.0965905
pcnvsq	-.8570512	.2714575	-3.16	0.002	-1.389098	-.3250042
pt86sq	-.1035031	.0224234	-4.62	0.000	-.1474522	-.059554
inc86sq	8.75e-06	4.28e-06	2.04	0.041	3.63e-07	.0000171
_cons	-.337362	.0562665	-6.00	0.000	-.4476423	-.2270817

Note: 51 failures and 0 successes completely determined.

```
. predict PHIhat
(option pr assumed; Pr(arr86))
```

```
. corr PHIhat arr86
(obs=2725)
```

	PHIhat	arr86
PHIhat	1.0000	
arr86	0.3396	1.0000

```
. di .3396^2
.11532816
```

**15.11.** We really need to make two assumptions. The first is a conditional independence

assumption: given  $\mathbf{x}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})$ ,  $(y_{i1}, \dots, y_{iT})$  are independent. This allows us to write

$$f(y_{i1}, \dots, y_{iT}|\mathbf{x}_i) = f_1(y_{i1}|\mathbf{x}_i) \cdots f_T(y_{iT}|\mathbf{x}_i),$$

that is, the joint density (conditional on  $\mathbf{x}_i$ ) is the product of the marginal densities (each conditional on  $\mathbf{x}_i$ ). The second assumption is a strict exogeneity assumption:

$D(y_{iT}|\mathbf{x}_i) = D(y_{iT}|\mathbf{x}_{it})$ ,  $t = 1, \dots, T$ . When we add the standard assumption for pooled probit – that  $D(y_{iT}|\mathbf{x}_{it})$  follows a probit model – then

$$f(y_1, \dots, y_T|\mathbf{x}_i) = \prod_{t=1}^T [G(\mathbf{x}_{it}\boldsymbol{\beta})]^{y_t} [1 - G(\mathbf{x}_{it}\boldsymbol{\beta})]^{1-y_t},$$

and so pooled probit is conditional MLE.

**15.12.** We can extend the  $T = 2$  case used to obtain equation (15.81):

$$\begin{aligned} P(y_{i1} = 1|\mathbf{x}_i, c_i, n_i = 1) &= P(y_{i1} = 1, n_i = 1|\mathbf{x}_i, c_i) / P(n_i = 1|\mathbf{x}_i, c_i) \\ &= P(y_{i1} = 1, y_{i2} = 0, y_{i3} = 0|\mathbf{x}_i, c_i) / \{P(y_{i1} = 1, y_{i2} = 0, y_{i3} = 0|\mathbf{x}_i, c_i) \\ &\quad + P(y_{i1} = 0, y_{i2} = 1, y_{i3} = 0|\mathbf{x}_i, c_i) + P(y_{i1} = 0, y_{i2} = 0, y_{i3} = 1|\mathbf{x}_i, c_i)\} \end{aligned}$$

Now, we just use the conditional independence assumption (across  $t$ ) and the logistic functional form:

$$\begin{aligned} P(y_{i1} = 1, y_{i2} = 0, y_{i3} = 0|\mathbf{x}_i, c_i) &= \Lambda(\mathbf{x}_{i1}\boldsymbol{\beta} + c_i) [1 - \Lambda(\mathbf{x}_{i2}\boldsymbol{\beta} + c_i)] \cdot [1 - \Lambda(\mathbf{x}_{i3}\boldsymbol{\beta} + c_i)] \\ P(y_{i1} = 0, y_{i2} = 1, y_{i3} = 0|\mathbf{x}_i, c_i) &= [1 - \Lambda(\mathbf{x}_{i1}\boldsymbol{\beta} + c_i)] \Lambda(\mathbf{x}_{i2}\boldsymbol{\beta} + c_i) \cdot [1 - \Lambda(\mathbf{x}_{i3}\boldsymbol{\beta} + c_i)] \end{aligned}$$

and

$$P(y_{i1} = 0, y_{i2} = 0, y_{i3} = 1|\mathbf{x}_i, c_i) = [1 - \Lambda(\mathbf{x}_{i1}\boldsymbol{\beta} + c_i)] \cdot [1 - \Lambda(\mathbf{x}_{i2}\boldsymbol{\beta} + c_i)] \Lambda(\mathbf{x}_{i3}\boldsymbol{\beta} + c_i).$$

Now, the term

$$1 / \{[1 + \exp(\mathbf{x}_{i1}\boldsymbol{\beta} + c_i)] \cdot [1 + \exp(\mathbf{x}_{i2}\boldsymbol{\beta} + c_i)] \cdot [1 + \exp(\mathbf{x}_{i3}\boldsymbol{\beta} + c_i)]\}$$

appears multiplicatively in both the numerator and denominator, and so it disappears.



Therefore,

$$\begin{aligned} P(y_{i1} = 1 | \mathbf{x}_i, c_i, n_i = 1) &= \exp(\mathbf{x}_{i1}\boldsymbol{\beta} + c_i) / [\exp(\mathbf{x}_{i1}\boldsymbol{\beta} + c_i) + \exp(\mathbf{x}_{i2}\boldsymbol{\beta} + c_i) + \exp(\mathbf{x}_{i3}\boldsymbol{\beta} + c_i)] \\ &= \exp(\mathbf{x}_{i1}\boldsymbol{\beta}) / [\exp(\mathbf{x}_{i1}\boldsymbol{\beta}) + \exp(\mathbf{x}_{i2}\boldsymbol{\beta}) + \exp(\mathbf{x}_{i3}\boldsymbol{\beta})]. \end{aligned}$$

Also,

$$P(y_{i2} = 1 | \mathbf{x}_i, c_i, n_i = 1) = \exp(\mathbf{x}_{i2}\boldsymbol{\beta}) / [\exp(\mathbf{x}_{i1}\boldsymbol{\beta}) + \exp(\mathbf{x}_{i2}\boldsymbol{\beta}) + \exp(\mathbf{x}_{i3}\boldsymbol{\beta})]$$

and

$$P(y_{i3} = 1 | \mathbf{x}_i, c_i, n_i = 1) = \exp(\mathbf{x}_{i3}\boldsymbol{\beta}) / [\exp(\mathbf{x}_{i1}\boldsymbol{\beta}) + \exp(\mathbf{x}_{i2}\boldsymbol{\beta}) + \exp(\mathbf{x}_{i3}\boldsymbol{\beta})].$$

Incidentally, a consistent estimator of  $\boldsymbol{\beta}$  is obtained using only the  $n_i = 1$  observations and applying conditional logit, as described in Chapter 16. This approach would be inefficient because it does not use the  $n_i = 2$  observations.

A similar argument can be used for the three possible configurations with  $n_i = 2$ , which leads to the log-likelihood conditional on  $(\mathbf{x}_i, n_i)$ , where  $c_i$  has dropped out. For example,

$$P(y_{i1} = 1, y_{i2} = 1 | \mathbf{x}_i, c_i, n_i = 2) = \frac{\exp[(\mathbf{x}_{i1} + \mathbf{x}_{i2})\boldsymbol{\beta}]}{\exp[(\mathbf{x}_{i1} + \mathbf{x}_{i2})\boldsymbol{\beta}] + [\exp(\mathbf{x}_{i1} + \mathbf{x}_{i3})\boldsymbol{\beta}] + \exp[(\mathbf{x}_{i2} + \mathbf{x}_{i3})\boldsymbol{\beta}]}$$

**15.13.** a. If there are no covariates, there is no point in using any method other than a straight comparison of means – in particular, the difference-in-differences approach described in Section 6.5.2. The estimated probabilities for the treatment and control groups, both before and after the policy change, will be identical to the sample proportions regardless of the model we use.

b. Let  $d2$  be a binary indicator for the second time period, and let  $dB$  be an indicator for the treatment group. Then a probit model to evaluate the treatment effect is

$$P(y = 1 | \mathbf{x}) = \Phi(\delta_0 + \delta_1 d2 + \delta_2 dB + \delta_3 d2 \cdot dB + \mathbf{x}\boldsymbol{\gamma}),$$

where  $\mathbf{x}$  is a vector of covariates. We would estimate all parameters from a probit of  $y$  on  $1, d2, dB, d2 \cdot dB$ , and  $\mathbf{x}$  using all observations. Once we have the estimates, we need to compute the “difference-in-differences” estimate, which requires either plugging in a value for  $\mathbf{x}$ , say  $\bar{\mathbf{x}}$ , or averaging the differences across  $\mathbf{x}_i$ . In the former case, we have

$$\begin{aligned}\hat{\tau}_{PAE} \equiv & [\Phi(\hat{\delta}_0 + \hat{\delta}_1 + \hat{\delta}_2 + \hat{\delta}_3 + \bar{\mathbf{x}}\hat{\gamma}) - \Phi(\hat{\delta}_0 + \hat{\delta}_2 + \bar{\mathbf{x}}\hat{\gamma})] \\ & - [\Phi(\hat{\delta}_0 + \hat{\delta}_1 + \bar{\mathbf{x}}\hat{\gamma}) - \Phi(\hat{\delta}_0 + \bar{\mathbf{x}}\hat{\gamma})],\end{aligned}$$

and in the latter we have

$$\begin{aligned}\hat{\tau}_{APE} \equiv & N^{-1} \sum_{i=1}^N \{[\Phi(\hat{\delta}_0 + \hat{\delta}_1 + \hat{\delta}_2 + \hat{\delta}_3 + \mathbf{x}_i\hat{\gamma}) - \Phi(\hat{\delta}_0 + \hat{\delta}_2 + \mathbf{x}_i\hat{\gamma})] \\ & - [\Phi(\hat{\delta}_0 + \hat{\delta}_1 + \mathbf{x}_i\hat{\gamma}) - \Phi(\hat{\delta}_0 + \mathbf{x}_i\hat{\gamma})]\}.\end{aligned}$$

Probably  $\hat{\tau}_{APE}$  is preferred as it averages each of the estimated “treatment effects” – see Chapter 21 – across all units.

c. We would have to use the delta method to obtain a valid standard error for either  $\hat{\tau}_{PAE}$  or  $\hat{\tau}_{APE}$ , with the latter using the extension in Problem 12.17.

**15.14.** a. First plug in for  $y_2$  from (15.40):

$$\begin{aligned}y_1 &= 1[\mathbf{z}_1\boldsymbol{\delta}_1 + y_2\mathbf{z}_2\boldsymbol{\alpha}_1 + u_1 > 0] = 1[\mathbf{z}_1\boldsymbol{\delta}_1 + (\mathbf{z}\boldsymbol{\delta}_2 + v_2)\mathbf{z}_2\boldsymbol{\alpha}_1 + u_1 > 0] \\ &= 1[\mathbf{z}_1\boldsymbol{\delta}_1 + (\mathbf{z}\boldsymbol{\delta}_2)\mathbf{z}_2\boldsymbol{\alpha}_1 + u_1 + v_2\mathbf{z}_2\boldsymbol{\alpha}_1 > 0]\end{aligned}$$

Given the assumptions,  $u_1 + v_2\mathbf{z}_2\boldsymbol{\alpha}_1$  has a mean zero normal distribution conditional on  $\mathbf{z}$ . Its variance is

$$\text{Var}(u_1 + v_2\mathbf{z}_2\boldsymbol{\alpha}_1|\mathbf{z}) = 1 + 2\eta_1\mathbf{z}_2\boldsymbol{\alpha}_1 + \tau_2^2(\mathbf{z}_2\boldsymbol{\alpha}_1)^2$$

where  $\eta_1 = \text{Cov}(v_2, u_1)$  and  $\tau_2^2 = \text{Var}(v_2)$ . So we can write

$$\begin{aligned}
P(y_1 = 1|\mathbf{z}) &= 1 - \Phi \left[ \frac{-[\mathbf{z}_1 \boldsymbol{\delta}_1 + (\mathbf{z} \boldsymbol{\delta}_2) \mathbf{z}_2 \boldsymbol{\alpha}_1]}{\sqrt{1 + 2\eta_1 \mathbf{z}_2 \boldsymbol{\alpha}_1 + \tau_2^2 (\mathbf{z}_2 \boldsymbol{\alpha}_1)^2}} \right] \\
&= \Phi \left[ \frac{\mathbf{z}_1 \boldsymbol{\delta}_1 + (\mathbf{z} \boldsymbol{\delta}_2) \mathbf{z}_2 \boldsymbol{\alpha}_1}{\sqrt{1 + 2\eta_1 \mathbf{z}_2 \boldsymbol{\alpha}_1 + \tau_2^2 (\mathbf{z}_2 \boldsymbol{\alpha}_1)^2}} \right]
\end{aligned}$$

which is a heteroskedastic-probit model (but not with exponential heteroskedasticity in the latent error).

b. This two-step procedure is inconsistent because the response probability  $P(y_1 = 1|\mathbf{z})$  does not have the usual probit form

$$\Phi[\mathbf{z}_1 \boldsymbol{\delta}_1 + (\mathbf{z} \boldsymbol{\delta}_2) \mathbf{z}_2 \boldsymbol{\alpha}_1].$$

Under the assumptions given, the first-stage estimation of  $\boldsymbol{\delta}_2$  is not the problem: OLS is consistent. It is the misspecified functional form in the second stage that causes the problem.

c. A control function method works nicely here. Scaled coefficients are easily estimated and then  $\boldsymbol{\delta}_1$  and  $\boldsymbol{\alpha}_1$  can be recovered using the same approach in Section 15.7.2. In addition, average partial effects are easily estimated after control function estimation.

Under (15.40), independence, and bivariate normality, we can write  $u_1$  as in equation (15.42) and then substitute:

$$\begin{aligned}
y_1 &= 1[\mathbf{z}_1 \boldsymbol{\delta}_1 + y_2 \mathbf{z}_2 \boldsymbol{\alpha}_1 + \theta_1 v_2 + e_1 > 0] \\
e_1 | \mathbf{z}, y_2, v_2 &\sim \text{Normal}(0, 1 - \rho_1^2)
\end{aligned}$$

Following the same argument in Section 15.7.2 we have

$$P(y_1 = 1|\mathbf{z}, y_2, v_2) = \Phi(\mathbf{z}_1 \boldsymbol{\delta}_{\rho 1} + y_2 \mathbf{z}_2 \boldsymbol{\alpha}_{\rho 1} + \theta_{\rho 1} v_2)$$

where  $\boldsymbol{\delta}_{\rho 1} = \boldsymbol{\delta}_1 / (1 - \rho_1^2)^{1/2}$ ,  $\boldsymbol{\alpha}_{\rho 1} = \boldsymbol{\alpha}_1 / (1 - \rho_1^2)^{1/2}$ , and  $\theta_{\rho 1} = \theta_1 / (1 - \rho_1^2)^{1/2}$ . Therefore, the following two-step CF method – which extends Procedure 15.1 – consistently estimates the

scaled parameters: (i) Regress  $y_2$  on  $\mathbf{z}$  and obtain the OLS residuals,  $\hat{v}_2$ . (ii) Run a probit of  $y_1$  on  $\mathbf{z}_1$ ,  $y_2\mathbf{z}_2$ , and  $\hat{v}_2$ .

Letting  $\boldsymbol{\beta}_1 = (\boldsymbol{\delta}'_1, \boldsymbol{\alpha}'_1)'$  and  $\boldsymbol{\beta}_{\rho 1} = \boldsymbol{\beta}_1/(1 - \rho_1^2)$ , we use exactly the same unscaling of the parameters as before. Namely,

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_{\rho 1}/(1 + \theta_{\rho 1}^2 \tau_2^2)^{1/2}$$

where  $\tau_2^2 = \text{Var}(v_2)$ . The estimator in equation (15.45) can still be used.

The approach to estimating the APEs follows directly from the estimator of the average structural function in equation (15.47). Allowing for the interactions,

$$\widehat{\text{ASF}}(\mathbf{z}_1, y_2) = N^{-1} \sum_{i=1}^n \Phi(\mathbf{z}_1 \hat{\boldsymbol{\delta}}_{\rho 1} + y_2 \mathbf{z}_2 \hat{\boldsymbol{\alpha}}_{\rho 1} + \hat{\theta}_{\rho 1} \hat{v}_{i2}).$$

As usual, we can take derivatives or changes with respect to the elements of  $(\mathbf{z}_1, y_2)$  to obtain estimated APEs.

**15.15.** a. This example falls into the situation described below equation (12.41). Namely, the scores from the two optimization problems are uncorrelated. This follows because the first problem – OLS regression of  $y_{i2}$  on  $\mathbf{z}_i$  – depends only on the random draws  $(\mathbf{z}_i, y_{i2})$ . In the second stage, we are estimating a model for  $f(y_1|y_2, \mathbf{z})$ . Letting  $\mathbf{s}_i(\boldsymbol{\gamma}_1; \boldsymbol{\delta}_2)$  denote the score for the second-step MLE – with respect to  $\boldsymbol{\gamma}_1$  –  $\mathbf{s}_i(\boldsymbol{\gamma}_1; \boldsymbol{\delta}_2)$  is uncorrelated with any function of  $(\mathbf{z}_i, y_{i2})$  because  $E[\mathbf{s}_i(\boldsymbol{\gamma}_1; \boldsymbol{\delta}_2)|\mathbf{z}_i, y_{i2}] = \mathbf{0}$ . (I do not use a separate notation for the true values of the parameters.)

So that we can apply the results from Section 12.4.2 directly, we set the problem up as a minimization problem. Then, from the usual score formula for the probit model, we have

$$\mathbf{s}_i(\boldsymbol{\gamma}_1; \boldsymbol{\delta}_2) = -\frac{\mathbf{w}_{i1}(\boldsymbol{\delta}_2)' \phi(\mathbf{w}_{i1}(\boldsymbol{\delta}_2) \boldsymbol{\gamma}_1) [y_{i1} - \Phi(\mathbf{w}_{i1}(\boldsymbol{\delta}_2) \boldsymbol{\gamma}_1)]}{\Phi(\mathbf{w}_{i1}(\boldsymbol{\delta}_2) \boldsymbol{\gamma}_1) [1 - \Phi(\mathbf{w}_{i1}(\boldsymbol{\delta}_2) \boldsymbol{\gamma}_1)]}$$

where  $\mathbf{w}_{i1}(\boldsymbol{\delta}_2) = [\mathbf{x}_{i1}, v_{i2}(\boldsymbol{\delta}_2)]$  and  $v_{i2}(\boldsymbol{\delta}_2) = y_{i2} - \mathbf{z}_i' \boldsymbol{\delta}_2$ . The expected Hessian (that is, the expected Jacobian of  $\mathbf{s}_i(\boldsymbol{\gamma}_1; \boldsymbol{\delta}_2)$  with respect to  $\boldsymbol{\gamma}_1$ ) has the usual form for binary response with a correctly specified response probability:

$$\mathbf{A}_1 = E \left\{ \frac{\mathbf{w}_{i1}(\boldsymbol{\delta}_2)' \mathbf{w}_{i1}(\boldsymbol{\delta}_2) [\phi(\mathbf{w}_{i1}(\boldsymbol{\delta}_2) \boldsymbol{\gamma}_1)]^2}{\Phi(\mathbf{w}_{i1}(\boldsymbol{\delta}_2) \boldsymbol{\gamma}_1) [1 - \Phi(\mathbf{w}_{i1}(\boldsymbol{\delta}_2) \boldsymbol{\gamma}_1)]} \right\}$$

Next, we need  $\mathbf{F} = E[\nabla_{\boldsymbol{\delta}_2} \mathbf{s}_i(\boldsymbol{\gamma}_1; \boldsymbol{\delta}_2)]$ . Like the Hessian in the usual binary response model, the Jacobian  $\nabla_{\boldsymbol{\delta}_2} \mathbf{s}_i(\boldsymbol{\gamma}_1; \boldsymbol{\delta}_2)$  is complicated. But its expectation is not. Using the fact that  $E[y_{i1} - \Phi(\mathbf{w}_{i1}(\boldsymbol{\delta}_2) \boldsymbol{\gamma}_1) | \mathbf{z}_i, y_{i2}] = 0$  it is easy to show

$$\begin{aligned} \mathbf{F} &= E \left\{ \frac{\mathbf{w}_{i1}(\boldsymbol{\delta}_2)' \phi(\mathbf{w}_{i1}(\boldsymbol{\delta}_2) \boldsymbol{\gamma}_1)}{\Phi(\mathbf{w}_{i1}(\boldsymbol{\delta}_2) \boldsymbol{\gamma}_1) [1 - \Phi(\mathbf{w}_{i1}(\boldsymbol{\delta}_2) \boldsymbol{\gamma}_1)]} \cdot \nabla_{\boldsymbol{\delta}_2} \Phi(\mathbf{w}_{i1}(\boldsymbol{\delta}_2) \boldsymbol{\gamma}_1) \right\} \\ &= -\theta_{\rho 1} E \left\{ \frac{\mathbf{w}_{i1}(\boldsymbol{\delta}_2)' \mathbf{z}_i \phi(\mathbf{w}_{i1}(\boldsymbol{\delta}_2) \boldsymbol{\gamma}_1)}{\Phi(\mathbf{w}_{i1}(\boldsymbol{\delta}_2) \boldsymbol{\gamma}_1) [1 - \Phi(\mathbf{w}_{i1}(\boldsymbol{\delta}_2) \boldsymbol{\gamma}_1)]} \right\} \end{aligned}$$

Finally, we need a first-order representation for the OLS estimator,  $\hat{\boldsymbol{\delta}}_2$ :

$$\sqrt{N}(\hat{\boldsymbol{\delta}}_2 - \boldsymbol{\delta}_2) = \mathbf{A}_2^{-1} N^{-1/2} \sum_{i=1}^N \mathbf{z}_i' v_{i2} + o_p(1),$$

where  $\mathbf{A}_2 \equiv E(\mathbf{z}_i' \mathbf{z}_i)$ . It follows that the matrix in the middle of the sandwich is

$$\begin{aligned} \mathbf{D} &= \text{Var}\{\mathbf{s}_i(\boldsymbol{\gamma}_1; \boldsymbol{\delta}_2) + \mathbf{F} \mathbf{A}_2^{-1} \mathbf{z}_i' v_{i2}\} \\ &= \text{Var}[\mathbf{s}_i(\boldsymbol{\gamma}_1; \boldsymbol{\delta}_2)] + \mathbf{F} \mathbf{A}_2^{-1} \text{Var}(\mathbf{z}_i' v_{i2}) \mathbf{A}_2^{-1} \mathbf{F}' \\ &= \mathbf{A}_1 + \tau_2^2 \mathbf{F} \mathbf{A}_2^{-1} \mathbf{F}' \end{aligned}$$

because  $\mathbf{s}_i(\boldsymbol{\gamma}_1; \boldsymbol{\delta}_2)$  and  $\mathbf{z}_i' v_{i2}$  are uncorrelated,  $\text{Var}[\mathbf{s}_i(\boldsymbol{\gamma}_1; \boldsymbol{\delta}_2)] = \mathbf{A}_1$  by the information matrix equality, and  $E(v_{i2}^2 | \mathbf{z}_i) = \tau_2^2$  under homoskedasticity for  $v_{i2}$ . (The results that follow do not rely in any crucial way on  $E(v_{i2}^2 | \mathbf{z}_i) = \tau_2^2$ ; we could just drop that and use the more general

formula.) Therefore,

$$\begin{aligned}\text{Avar}[\sqrt{N}(\hat{\gamma}_1 - \gamma_1)] &= \mathbf{A}_1^{-1}(\mathbf{A}_1 + \tau_2^2 \mathbf{F} \mathbf{A}_2^{-1} \mathbf{F}') \mathbf{A}_1^{-1} \\ &= \mathbf{A}_1^{-1} + \tau_2^2 \mathbf{A}_1^{-1} \mathbf{F} \mathbf{A}_2^{-1} \mathbf{F}' \mathbf{A}_1^{-1}.\end{aligned}$$

It is easy to construct consistent estimators of each part using sample averages and plugging in the consistent estimators.

b. If we ignore estimation of  $\delta_2$  we act as if  $\text{Avar}[\sqrt{N}(\hat{\gamma}_1 - \gamma_1)]$  is just  $\mathbf{A}_1^{-1}$ , the inverse of the information matrix from the second stage problem. But the correct matrix differs from  $\mathbf{A}_1^{-1}$  by  $\tau_2^2 \mathbf{A}_1^{-1} \mathbf{F} \mathbf{A}_2^{-1} \mathbf{F}' \mathbf{A}_1^{-1}$ , which is positive semi-definite (and usual positive definite if  $\theta_{\rho 1} \neq 0$ ).

c. In Problem 12.17 we can take  $\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}) = \Phi(\mathbf{w}_{i1}(\delta_2)\gamma_1)\gamma_1$ , but we have to be careful in choosing the “score” with respect to  $\boldsymbol{\theta}$ . The same argument as in Problem 12.17 gives us

$$N^{-1/2} \sum_{i=1}^N \mathbf{g}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = N^{-1/2} \sum_{i=1}^N \mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}) + \mathbf{G} \sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p(1)$$

where  $\mathbf{G} \equiv \text{E}[\nabla_{\boldsymbol{\theta}} \mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta})]$ . For this application,

$$\nabla_{\boldsymbol{\theta}} \mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}) = [\Phi(\mathbf{w}_{i1}(\delta_2)\gamma_1) \mathbf{I}_{K_1+1} + \phi(\mathbf{w}_{i1}(\delta_2)\gamma_1)\gamma_1 \mathbf{w}_{i1}(\delta_2) \phi(\mathbf{w}_{i1}(\delta_2)\gamma_1)\gamma_1' \nabla_{\delta_2} \mathbf{w}_{i1}(\delta_2)']$$

where  $K_1$  is the dimension of  $\mathbf{x}_1$  and  $\nabla_{\delta_2} \mathbf{w}_{i1}(\delta_2)' = [\mathbf{0} | -\mathbf{z}_i']$ . To get a representation for  $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$  we stack the first-order representations obtained in part a:

$$\begin{aligned}\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) &= N^{-1/2} \sum_{i=1}^N \begin{pmatrix} -\mathbf{A}_1^{-1}[\mathbf{s}_i(\gamma_1; \delta_2) + \mathbf{F} \mathbf{A}_2^{-1} \mathbf{z}_i' v_{i2}] \\ \mathbf{A}_2^{-1} \mathbf{z}_i' v_{i2} \end{pmatrix} + o_p(1) \\ &\equiv N^{-1/2} \sum_{i=1}^N \mathbf{e}_i(\boldsymbol{\theta}) + o_p(1).\end{aligned}$$

Then, from Problem 12.17,

$$\mathbf{C} = \text{Avar}[\sqrt{N}(\hat{\boldsymbol{\eta}}_1 - \boldsymbol{\eta}_1)] = \text{Var}[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}) - \boldsymbol{\eta}_1 - \mathbf{G} \mathbf{e}_i(\boldsymbol{\theta})]$$

d. A consistent estimator of the asymptotic variance in part c is

$$\hat{\mathbf{C}} = N^{-1} \sum_{i=1}^N (\hat{\mathbf{g}}_i - \hat{\boldsymbol{\eta}}_1 - \hat{\mathbf{G}}\hat{\mathbf{e}}_i)(\hat{\mathbf{g}}_i - \hat{\boldsymbol{\eta}}_1 - \hat{\mathbf{G}}\hat{\mathbf{e}}_i)'$$

where  $\hat{\mathbf{g}}_i = \Phi(\hat{\mathbf{w}}_{i1}\hat{\boldsymbol{\gamma}}_1)\hat{\boldsymbol{\gamma}}_1$ ,

$$\hat{\mathbf{G}} = N^{-1} \sum_{i=1}^N [\Phi(\hat{\mathbf{w}}_{i1}\hat{\boldsymbol{\gamma}}_1)\hat{\boldsymbol{\gamma}}_1 \mathbf{I}_{K_1+1} + \phi(\hat{\mathbf{w}}_{i1}\hat{\boldsymbol{\gamma}}_1)\hat{\boldsymbol{\gamma}}_1 \hat{\mathbf{w}}_{i1} |\phi(\hat{\mathbf{w}}_{i1}\hat{\boldsymbol{\gamma}}_1)\hat{\boldsymbol{\gamma}}_1' \nabla_{\delta_2} \mathbf{w}_{i1}(\hat{\boldsymbol{\delta}}_2)']$$

and

$$\hat{\mathbf{e}}_i = \begin{pmatrix} -\hat{\mathbf{A}}_1^{-1} [\hat{\mathbf{s}}_i + \hat{\mathbf{F}}\hat{\mathbf{A}}_2^{-1} \mathbf{z}_i' \hat{\mathbf{v}}_{i2}] \\ \hat{\mathbf{A}}_2^{-1} \mathbf{z}_i' \hat{\mathbf{v}}_{i2} \end{pmatrix}.$$

The score  $\hat{\mathbf{s}}_i$  and Hessian  $\hat{\mathbf{A}}_1$  are estimated as usual for a probit model (but with minus signs)

and

$$\hat{\mathbf{F}} = N^{-1} \sum_{i=1}^N \frac{-\theta'_{\rho 1} \hat{\mathbf{w}}_{i1} \mathbf{z}_i \phi(\hat{\mathbf{w}}_{i1}\hat{\boldsymbol{\gamma}}_1)}{\Phi(\hat{\mathbf{w}}_{i1}\hat{\boldsymbol{\gamma}}_1)[1 - \Phi(\hat{\mathbf{w}}_{i1}\hat{\boldsymbol{\gamma}}_1)]}.$$

**15.16.** a. The response probability is

$$p(\mathbf{x}) = 1 - [1 + \exp(\mathbf{x}\boldsymbol{\beta})]^{-\alpha}$$

and, using the chain rule,

$$\frac{\partial p(\mathbf{x})}{\partial x_K} = \alpha \beta_j \exp(\mathbf{x}\boldsymbol{\beta}) [1 + \exp(\mathbf{x}\boldsymbol{\beta})]^{-\alpha-1} = \frac{\alpha \beta_j \exp(\mathbf{x}\boldsymbol{\beta})}{[1 + \exp(\mathbf{x}\boldsymbol{\beta})]^{\alpha+1}}$$

Of course, we get the logit partial effect as a special case when  $\alpha = 1$ .

b. The log likelihood has the usual form for a binary response. Let

$G(\mathbf{x}, \boldsymbol{\theta}) = 1 - [1 + \exp(\mathbf{x}\boldsymbol{\beta})]^{-\alpha}$ , so  $1 - G(\mathbf{x}, \boldsymbol{\theta}) = [1 + \exp(\mathbf{x}\boldsymbol{\beta})]^{-\alpha}$ . Without making the

distinction between generic and “true” values,

$$\ell_i(\beta, \alpha) = -(1 - y_i)\alpha \log[1 + \exp(\mathbf{x}_i\beta)] + y_i \log\{1 - [1 + \exp(\mathbf{x}_i\beta)]^{-\alpha}\}.$$

c. The Stata output is given below. Given the estimated value of  $\alpha$ ,  $\hat{\alpha} = 413,553$ , the model does not seem well determined. (Remember, the logit model imposes  $\alpha = 1$ .) The logit estimates are included for comparison. The  $\hat{\beta}_j$  are all the same sign and of roughly the same statistical significance across the two models. The  $t$  statistic for  $H_0 : \log(\alpha) = 0$  is very small, about .02.

```
. scobit inlf nwifeinc educ exper expersq age kidslt6 kidsge6
```

```
Skewed logistic regression              Number of obs      =          753
                                         Zero outcomes      =          325
Log likelihood = -399.5222              Nonzero outcomes   =          428
```

inlf	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
nwifeinc	-.0148532	.0056874	-2.61	0.009	-.0260003	-.0037061
educ	.1512102	.0277346	5.45	0.000	.0968514	.2055689
exper	.139092	.020757	6.70	0.000	.0984091	.1797749
expersq	-.002257	.0006377	-3.54	0.000	-.0035069	-.0010072
age	-.0587203	.0089444	-6.57	0.000	-.076251	-.0411897
kidslt6	-.9977924	.1426425	-7.00	0.000	-1.277367	-.7182183
kidsge6	.0257666	.045345	0.57	0.570	-.0631079	.1146411
_cons	-13.09326	666.1339	-0.02	0.984	-1318.692	1292.505
/lnalpha	12.93254	666.1327	0.02	0.985	-1292.663	1318.529
alpha	413553.1	2.75e+08			0	

```
Likelihood-ratio test of alpha=1:    chi2(1) =      4.49    Prob > chi2 = 0.0342
```

Note: likelihood-ratio tests are recommended for inference with scobit models

```
. logit inlf nwifeinc educ exper expersq age kidslt6 kidsge6
```

```
Logistic regression              Number of obs      =          753
                                   LR chi2(7)             =          226.22
                                   Prob > chi2            =          0.0000
Log likelihood = -401.76515       Pseudo R2          =          0.2197
```

inlf	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
nwifeinc	-.0213452	.0084214	-2.53	0.011	-.0378509	-.0048394
educ	.2211704	.0434396	5.09	0.000	.1360303	.3063105
exper	.2058695	.0320569	6.42	0.000	.1430391	.2686999
expersq	-.0031541	.0010161	-3.10	0.002	-.0051456	-.0011626
age	-.0880244	.014573	-6.04	0.000	-.116587	-.0594618
kidslt6	-1.443354	.2035849	-7.09	0.000	-1.842373	-1.044335



kidsge6		.0601122	.0747897	0.80	0.422	-.086473	.2066974
_cons		.4254524	.8603697	0.49	0.621	-1.260841	2.111746

---

d. The likelihood ratio statistic for  $H_0 : \alpha = 1$ , reported in the Stata output, is 4.49 with  $p$ -value = .034. So this statistic rejects the logit model, although it is not an overwhelming rejection. Its  $p$ -value is certainly much smaller than the Wald test ( $t$  test) for  $H_0 : \log(\alpha) = 0$ .

e. Given the bizarre value for  $\hat{\alpha}$  and the modest gain in fit, the skewed logit model does not seem worth the effort. Plus, the Stata output below shows that the correlations of the fitted probabilities and *inlf* are very similar across the two models (.5196 for skewed logit, .5179 for logit). The average partial effects are similar, too. For *nwifeinc*, the APE for skewed logit is about -.0041 for skewed logit and about -.0038 for logit. For *kidslt6*, the APEs are -.274 (skewed logit) and -.258 (logit). It is likely these differences can be attributed to sampling error.

```
. qui scobit inlf nwifeinc educ exper expersq age kidslt6 kidsge6

. predict phat_skewlog
(option pr assumed; Pr(inlf))

. predict xbh_sklog, xb

. gen scale_sklog = e(alpha)*exp(xbh_sklog)/((1 + exp(xbh_sklog))^(1 + e(alpha)

. sum scale_sklog
```

Variable	Obs	Mean	Std. Dev.	Min	Max
scale_sklog	753	.2741413	.0891063	.0098302	.3678786

```
. predict phat_skewlog
(option pr assumed; Pr(inlf))

. qui logit inlf nwifeinc educ exper expersq age kidslt6 kidsge6

. predict phat_log
(option pr assumed; Pr(inlf))

. predict xbh_log, xb

. gen scale_log = exp(xbh_log)/((1 + exp(xbh_log))^2 )

. sum scale_log
```

Variable	Obs	Mean	Std. Dev.	Min	Max
scale_log	753	.1785796	.0617942	.0085973	.25

```
. corr phat_skewlog inlf
(obs=753)
```

	phat_s~g	inlf
phat_skewlog	1.0000	
inlf	0.5196	1.0000

```
. corr phat_log inlf
(obs=753)
```

	phat_log	inlf
phat_log	1.0000	
inlf	0.5179	1.0000

```
. di .2741413* (-.0148532)
-.00407188
```

```
. di .1785796*(-.0213452)
-.00381182
```

```
. di .2741413* (-.9977924)
-.27353611
```

```
. di .1785796*(-1.443354)
-.25775358
```

**15.17. a.** We obtain the joint density by the product rule, since we have independence conditional on  $(\mathbf{x}, c)$ :

$$f(y_1, \dots, y_G | \mathbf{x}, c; \boldsymbol{\gamma}_o) = f_1(y_1 | \mathbf{x}, c; \boldsymbol{\gamma}_o^1) f_2(y_2 | \mathbf{x}, c; \boldsymbol{\gamma}_o^2) \cdots f_G(y_G | \mathbf{x}, c; \boldsymbol{\gamma}_o^G).$$

b. The density of  $(y_1, \dots, y_G)$  given  $\mathbf{x}$  is obtained by integrating out with respect to the distribution of  $c$  given  $\mathbf{x}$ :

$$g(y_1, \dots, y_G | \mathbf{x}; \boldsymbol{\gamma}_o) = \int_{-\infty}^{\infty} \left( \prod_{g=1}^G f_g(y_g | \mathbf{x}, c; \boldsymbol{\gamma}_o^g) \right) h(c | \mathbf{x}; \boldsymbol{\delta}_o) dc,$$

where  $c$  is a dummy argument of integration. Because  $c$  appears in each  $D(y_g | \mathbf{x}, c)$ ,  $y_1, \dots, y_G$  are dependent without conditioning on  $c$ .

c. The log-likelihood for each  $i$  is

$$\log \left[ \int_{-\infty}^{\infty} \left( \prod_{g=1}^G f_g(y_{ig} | \mathbf{x}_i, \mathbf{c}; \boldsymbol{\gamma}^g) \right) h(\mathbf{c} | \mathbf{x}_i; \boldsymbol{\delta}) d\mathbf{c} \right].$$

As expected, this depends only on the observed data,  $(\mathbf{x}_i, y_{i1}, \dots, y_{iG})$ , and the unknown parameters.

**15.18.** a. The probability is the same as if we assume (15.73), that is,

$$P(y_{it} = 1 | \mathbf{x}_i, a_i) = \Phi(\psi + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i), \quad t = 1, 2, \dots, T.$$

The fact that  $a_i$  given  $\mathbf{x}_i$  is heteroskedastic has no bearing on the distribution conditional on  $(\mathbf{x}_i, a_i)$ . Only when we “integrate out”  $a_i$  does  $D(a_i | \mathbf{x}_i)$  matter.

b. Let  $g_t(y | \mathbf{x}_i, a_i; \boldsymbol{\theta}) = [\Phi(\psi + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i)]^{y_t} [1 - \Phi(\psi + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i)]^{1-y_t}$ . Then, by the product and integration rules,

$$f(y_1, \dots, y_T | \mathbf{x}; \boldsymbol{\theta}) = \left[ \int_{-\infty}^{\infty} \left( \prod_{t=1}^T g_t(y_t | \mathbf{x}_i, a_i; \boldsymbol{\theta}) \right) h(a_i | \mathbf{x}_i; \boldsymbol{\delta}) da_i \right],$$

where  $h(\cdot | \mathbf{x}_i, \boldsymbol{\delta})$  is the  $\text{Normal}[0, \sigma_a^2 \exp(\bar{\mathbf{x}}_i\boldsymbol{\lambda})]$  density. We get the log-likelihood by plugging in the  $y_{it}$  and taking the natural log. For each  $i$ , the log likelihood depends on  $(\mathbf{x}_i, \mathbf{y}_i)$  and the parameters  $\boldsymbol{\theta}$  and  $\boldsymbol{\delta}$ ;  $a_i$  does not appear.

c. To estimate the APEs we can estimate the average structural function, which in this case is

$$\begin{aligned} \text{ASF}(\mathbf{x}^o) &= E_{c_i}[\Phi(\psi + \mathbf{x}^o\boldsymbol{\beta} + c_i)] \\ &= E_{(\mathbf{x}_i, a_i)}[\Phi(\psi + \mathbf{x}^o\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i)] \\ &= E_{\mathbf{x}_i}\{E[\Phi(\psi + \mathbf{x}^o\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i) | \mathbf{x}_i]\} \end{aligned}$$

To compute  $E[\Phi(\psi + \mathbf{x}^o\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i) | \bar{\mathbf{x}}_i]$  we use a similar trick as before. It is the same as computing

$$E(1[\psi + \mathbf{x}^o\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i + u_{it} > 0] | \mathbf{x}_i)$$

where

$$a_i + u_{it} | \mathbf{x}_i \sim \text{Normal}(0, 1 + \sigma_a^2 \exp(\bar{\mathbf{x}}_i \boldsymbol{\lambda})).$$

because  $u_{it}$  is independent of  $(a_i, \mathbf{x}_i)$  with a standard normal distribution. Now

$$\begin{aligned} E(1[\psi + \mathbf{x}^o \boldsymbol{\beta} + \bar{\mathbf{x}}_i \boldsymbol{\xi} + a_i + u_{it} > 0] | \mathbf{x}_i) &= P[a_i + u_{it} > -(\psi + \mathbf{x}^o \boldsymbol{\beta} + \bar{\mathbf{x}}_i \boldsymbol{\xi}) | \mathbf{x}_i] \\ &= P\left[ \frac{a_i + u_{it}}{[1 + \sigma_a^2 \exp(\bar{\mathbf{x}}_i \boldsymbol{\lambda})]^{1/2}} > \frac{-(\psi + \mathbf{x}^o \boldsymbol{\beta} + \bar{\mathbf{x}}_i \boldsymbol{\xi})}{[1 + \sigma_a^2 \exp(\bar{\mathbf{x}}_i \boldsymbol{\lambda})]^{1/2}} \middle| \mathbf{x}_i \right] \\ &= \Phi\left[ \frac{(\psi + \mathbf{x}^o \boldsymbol{\beta} + \bar{\mathbf{x}}_i \boldsymbol{\xi})}{[1 + \sigma_a^2 \exp(\bar{\mathbf{x}}_i \boldsymbol{\lambda})]^{1/2}} \right]. \end{aligned}$$

(Notice this only depends on  $\bar{\mathbf{x}}_i$ , not on  $\mathbf{x}_i$ . We could relax that assumption.)

The ASF is therefore,

$$\text{ASF}(\mathbf{x}^o) = E_{\bar{\mathbf{x}}_i} \left\{ \Phi \left[ \frac{(\psi + \mathbf{x}^o \boldsymbol{\beta} + \bar{\mathbf{x}}_i \boldsymbol{\xi})}{[1 + \sigma_a^2 \exp(\bar{\mathbf{x}}_i \boldsymbol{\lambda})]^{1/2}} \right] \right\}$$

and a consistent estimator is obtained by using a sample average and plugging in the maximum likelihood estimators:

$$\widehat{\text{ASF}}(\mathbf{x}^o) = N^{-1} \sum_{i=1}^N \Phi \left[ \frac{(\hat{\psi} + \mathbf{x}^o \hat{\boldsymbol{\beta}} + \bar{\mathbf{x}}_i \hat{\boldsymbol{\xi}})}{[1 + \hat{\sigma}_a^2 \exp(\bar{\mathbf{x}}_i \hat{\boldsymbol{\lambda}})]^{1/2}} \right].$$

Now take derivatives and changes with respect to  $\mathbf{x}^o$  (a placeholder).

**15.19.** a. The Stata output is below. We need to assume first-order dynamics for the usual standard errors and test statistics to be valid.

. tab year

81 to 87	Freq.	Percent	Cum.
81	1,738	14.29	14.29
82	1,738	14.29	28.57
83	1,738	14.29	42.86
84	1,738	14.29	57.14
85	1,738	14.29	71.43
86	1,738	14.29	85.71
87	1,738	14.29	100.00

```

Total |      12,166      100.00

. tab black if year == 87

=1 if black |      Freq.      Percent      Cum.
-----+-----
          0 |      1,065      61.28      61.28
          1 |       673      38.72     100.00
-----+-----
Total    |      1,738     100.00

. xtset id year
      panel variable:  id (strongly balanced)
      time variable:  year, 81 to 87
              delta:  1 unit

. gen employ_1 = l.employ
(1738 missing values generated)

. probit employ employ_1 if black

Probit regression                                Number of obs   =       4038
                                                LR chi2(1)      =     1091.27
                                                Prob > chi2     =       0.0000
Log likelihood = -2248.0349                    Pseudo R2      =       0.1953

-----+-----
employ |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval
-----+-----
employ_1 |  1.389433   .0437182    31.78   0.000    1.303747    1.475119
   _cons | -.5396127   .0281709   -19.15   0.000   -.5948268   -.4843987
-----+-----

```

b. After estimating the previous model, the Stata calculations are below. The difference in employment probabilities this year, based on employment status last year, is about .508.

```

. di normal(_b[_cons])
.29473206

. di normal(_b[_cons] + _b[employ_1])
.80228758

. di normal(_b[_cons] + _b[employ_1]) - normal(_b[_cons])
.50755552

```

c. With year dummies, the story is very similar. The estimated state dependence for 1987 is about .472.

```

. probit employ employ_1 y83-y87 if black

Probit regression                                Number of obs   =       4038
                                                LR chi2(6)      =     1156.98
                                                Prob > chi2     =       0.0000
Log likelihood = -2215.1795                    Pseudo R2      =       0.2071

```

employ	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
employ_1	1.321349	.0453568	29.13	0.000	1.232452	1.410247
y83	.3427664	.0749844	4.57	0.000	.1957997	.4897331
y84	.4586078	.0755742	6.07	0.000	.3104851	.6067305
y85	.5200576	.0767271	6.78	0.000	.3696753	.6704399
y86	.3936516	.0774704	5.08	0.000	.2418125	.5454907
y87	.5292136	.0773031	6.85	0.000	.3777023	.680725
_cons	-.8850412	.0556042	-15.92	0.000	-.9940233	-.776059

```
. di normal(_b[_cons] + _b[y87] + _b[employ_1]) - normal(_b[_cons] + _b[y87])
.4718734
```

d. Below gives one way in Stata to estimate the dynamic unobserved effects model.

Compared with not allowing for heterogeneity as in part c, the coefficient on *employ*<sub>-1</sub> has fallen: from about 1.321 to about .899. In addition, the coefficient on the initial condition is .566 and it is very statistically significant. But we cannot know how much the amount of state dependence has changed without computing an average partial effect.

```
. gen employ81 = employ if y81
(10428 missing values generated)

. replace employ81 = employ[_n-1] if y82
(1738 real changes made)

. replace employ81 = employ[_n-2] if y83
(1738 real changes made)

. replace employ81 = employ[_n-3] if y84
(1738 real changes made)

. replace employ81 = employ[_n-4] if y85
(1738 real changes made)

. replace employ81 = employ[_n-5] if y86
(1738 real changes made)

. replace employ81 = employ[_n-6] if y87
(1738 real changes made)

. xtprobit employ employ_1 employ81 y83-y87 if black, re

Random-effects probit regression              Number of obs      =      4038
Group variable: id                          Number of groups    =       673

Random effects u_i ~Gaussian                 Obs per group: min =         6
                                              avg =         6.
                                              max =
```

Log likelihood = -2176.3738      Wald chi2(7) = 677.59  
 Prob > chi2 = 0.0000

employ	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
employ_1	.8987806	.0677058	13.27	0.000	.7660797	1.031482
employ81	.5662897	.0884941	6.40	0.000	.3928444	.739735
y83	.4339911	.0804064	5.40	0.000	.2763974	.5915847
y84	.6563094	.0841199	7.80	0.000	.4914374	.8211814
y85	.7919805	.0887167	8.93	0.000	.618099	.965862
y86	.6896344	.090158	7.65	0.000	.5129279	.8663409
y87	.8382018	.091054	9.21	0.000	.6597393	1.016664
_cons	-1.005103	.0660945	-15.21	0.000	-1.134646	-.8755602
/lnsig2u	-1.178731	.1995372			-1.569817	-.7876454
sigma_u	.5546791	.0553396			.4561615	.6744736
rho	.2352804	.0359014			.1722425	.3126745

Likelihood-ratio test of rho=0: chibar2(01) = 47.90 Prob >= chibar2 = 0.000

e. There is still plenty of evidence of state dependence because of the very statistically significant coefficient on  $employ_{-1}$  ( $t = 13.27$ ). The coefficient still seems quite large, but we still need to compute the APE.

The positive coefficient on  $employ_{81}$  shows that that  $c_i$  and  $employ_{i,81}$  are positively correlated. The estimate of  $\sigma_a^2$  is  $(.5546791)^2$ , or  $\hat{\sigma}_a^2 \approx .308$ .

f. The average state dependence, where we average out the distribution of  $c_i$ , is estimated as

$$N^{-1} \sum_{i=1}^N \left\{ \Phi \left[ \frac{(\hat{\psi} + \hat{\delta}_{87} + \hat{\rho} + \hat{\xi}y_{i0})}{(1 + \hat{\sigma}_a^2)^{1/2}} \right] - \Phi \left[ \frac{(\hat{\psi} + \hat{\delta}_{87} + \hat{\xi}y_{i0})}{(1 + \hat{\sigma}_a^2)^{1/2}} \right] \right\}$$

where  $\hat{\rho}$  is the coefficient on  $y_{-1} = employ_{-1}$ ,  $y_{i0} = employ_{i,1981}$ , and, in this case, the averaging is done across the black men in the sample. The Stata calculations below (done after the calculations in part d) show the estimated state dependence is about .283, which is much lower than the estimate of .472 from part c (where we ignored heterogeneity). Bootstrapping is a convenient way to obtain a standard error, as was done in Example 15.6.

```
. gen stdep = normal((_b[_cons] + _b[employ_1] + _b[employ81]*employ81
```

```

+ _b[y87])/sqrt(1 + e(sigma_u)^2))
- normal((_b[_cons] + _b[employ81]*employ81 + _b[y87])
/sqrt(1 + e(sigma_u)^2)) if black & y87
(11493 missing values generated)

```

```

. sum stdep

```

Variable	Obs	Mean	Std. Dev.	Min	Max
stdep	673	.283111	.0257298	.2353074	.2969392



**15.20. (Bonus Question)** Estimate the CRE probit model report in Table 15.3 using the generalized estimation equation (GEE) approach described in Section 12.9.2, using an exchangeable correlation structure.

a. How do the point estimates compare with the pooled probit estimates in Column (3) of Table 15.3?

b. Does it appear that the GEE approach improves on the efficiency of pooled probit? Explain.

**Solution:**

a. The Stata output for pooled probit and GEE is given below. The pooled probit estimates replicate the numbers in Table 15.3.

```
. probit lfp kids lhinc kidsbar lhincbar educ black age agesq per2-per5,
    cluster(id)
```

```
Iteration 0:  log pseudolikelihood = -17709.021
Iteration 1:  log pseudolikelihood = -16521.245
Iteration 2:  log pseudolikelihood = -16516.437
Iteration 3:  log pseudolikelihood = -16516.436
```

```
Probit regression                                Number of obs   =      28315
                                                Wald chi2(12)   =      538.09
                                                Prob > chi2     =      0.0000
Log pseudolikelihood = -16516.436                Pseudo R2      =      0.0673
```

(Std. Err. adjusted for 5663 clusters in id)

lfp	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval	
kids	-.1173749	.0269743	-4.35	0.000	-.1702435	-.0645064
lhinc	-.0288098	.014344	-2.01	0.045	-.0569234	-.0006961
kidsbar	-.0856913	.0311857	-2.75	0.006	-.146814	-.0245685
lhincbar	-.2501781	.0352907	-7.09	0.000	-.3193466	-.1810097
educ	.0841338	.0067302	12.50	0.000	.0709428	.0973248
black	.2030668	.0663945	3.06	0.002	.0729359	.3331976
age	.1516424	.0124831	12.15	0.000	.127176	.1761089
agesq	-.0020672	.0001553	-13.31	0.000	-.0023717	-.0017628
per2	-.0135701	.0103752	-1.31	0.191	-.0339051	.0067648
per3	-.0331991	.0127197	-2.61	0.009	-.0581293	-.008269
per4	-.0390317	.0136244	-2.86	0.004	-.0657351	-.0123284
per5	-.0552425	.0146067	-3.78	0.000	-.0838711	-.0266139
_cons	-.7260562	.2836985	-2.56	0.010	-1.282095	-.1700173

```
. xtgee lfp kids lhinc kidsbar lhincbar educ black age agesq per2-per5,
      fam(binomial) link(probit) corr(exch) robust
```

```
GEE population-averaged model
Group variable:          id      Number of obs      =      28315
Link:                   probit   Number of groups   =      5663
Family:                 binomial Obs per group: min =
Correlation:            exchangeable          avg =      5.
                                      max =
Scale parameter:        1      Wald chi2(12)      =      536.66
                                      Prob > chi2      =      0.0000
```

(Std. Err. adjusted for clustering on id)

lfp	Coef.	Semirobust Std. Err.	z	P> z	[95% Conf. Interval	
kids	-.1125361	.0281366	-4.00	0.000	-.1676828	-.0573894
lhinc	-.0276543	.014799	-1.87	0.062	-.0566598	.0013511
kidsbar	-.0892543	.0323884	-2.76	0.006	-.1527344	-.0257742
lhincbar	-.252001	.0360377	-6.99	0.000	-.3226337	-.1813684
educ	.0841304	.0066834	12.59	0.000	.0710312	.0972296
black	.205611	.0668779	3.07	0.002	.0745328	.3366893
age	.152809	.0125434	12.18	0.000	.1282245	.1773936
agesq	-.0020781	.0001565	-13.28	0.000	-.0023847	-.0017714
per2	-.0134259	.0103607	-1.30	0.195	-.0337324	.0068807
per3	-.0329993	.0126967	-2.60	0.009	-.0578845	-.0081141
per4	-.0384026	.0136212	-2.82	0.005	-.0650997	-.0117056
per5	-.05451	.0146135	-3.73	0.000	-.083152	-.025868
_cons	-.7532503	.285216	-2.64	0.008	-1.312263	-.1942373

b. Surprisingly, and disappointingly, the GEE approach does not improve the precision of the estimators. In fact, the robust standard errors for GEE are actually slightly above those for pooled probit. This finding is particular puzzling because there is substantial serial correlation in the standardized residuals, written generally after pooled probit estimation as

$$\hat{r}_{it} \equiv \frac{[y_{it} - \Phi(\mathbf{x}_{it}\hat{\boldsymbol{\beta}} + \hat{\psi} + \bar{\mathbf{w}}_i\hat{\boldsymbol{\xi}})]}{\{\Phi(\mathbf{x}_{it}\hat{\boldsymbol{\beta}} + \hat{\psi} + \bar{\mathbf{w}}_i\hat{\boldsymbol{\xi}})[1 - \Phi(\mathbf{x}_{it}\hat{\boldsymbol{\beta}} + \hat{\psi} + \bar{\mathbf{w}}_i\hat{\boldsymbol{\xi}})]\}^{1/2}},$$

where  $\bar{\mathbf{w}}_i$  is the time average of variables that change across  $i$  and  $t$  ( $kids_{it}$  and  $lhinc_{it}$  in this application). The first-order correlation in the  $\{\hat{r}_{it} : t = 2, \dots, T; i = 1, \dots, N\}$  is about .83.

```
. qui probit lfp kids lhinc kidsbar lhincbar educ black age agesq per2-per5
. predict phat
(option pr assumed; Pr(lfp))
```

```
. gen rh = (lfp - phat)/sqrt(phat*(1 - phat))

. gen rh_1 = 1.rh
(5663 missing values generated)

. corr rh rh_1
(obs=22652)
```

	rh	rh_1
rh	1.0000	
rh_1	0.8315	1.0000

c. This is not an answer to a particular question, but serves as an errata for the estimates on Column (4) of Table 15.3. Those estimates were obtained using a version of Stata earlier than 9.0. Using Stata 11.0, a higher value of the log likelihood is found, and the point estimates are different. Note that the estimated value of  $\rho$ , which is the pairwise correlation between any of the two composite errors  $a_i + e_{it}$ , is very large: .95. The estimated scale factor for the coefficients, about .233, is substantially below that in Table 15.3, but the coefficients reported below are substantially higher. I have deleted the details of the numerical iterations.

```
. xtprobit lfp kids lhinc kidsbar lhincbar educ black age agesq per2-per5, re

Random-effects probit regression      Number of obs      =      28315
Group variable: id                   Number of groups    =      5663

Random effects u_i ~Gaussian          Obs per group: min =      5
                                      avg =      5.
                                      max =

Wald chi2(12)                        =      623.40
Log likelihood = -8609.9002           Prob > chi2         =      0.0000
```

lfp	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
kids	-.3970102	.0701298	-5.66	0.000	-.534462	-.2595584
lhinc	-.1003399	.0469979	-2.13	0.033	-.1924541	-.0082258
kidsbar	-.4085664	.0898875	-4.55	0.000	-.5847428	-.2323901
lhincbar	-.8941069	.1199703	-7.45	0.000	-1.129244	-.6589695
educ	.3189079	.024327	13.11	0.000	.2712279	.366588
black	.6388784	.1903525	3.36	0.001	.2657945	1.011962
age	.7282057	.0445623	16.34	0.000	.6408651	.8155462
agesq	-.0098358	.0005747	-17.11	0.000	-.0109623	-.0087094
per2	-.0451653	.0499429	-0.90	0.366	-.1430516	.052721
per3	-.1247056	.0501522	-2.49	0.013	-.2230022	-.026409
per4	-.1356834	.0500679	-2.71	0.007	-.2338147	-.0375522
per5	-.200357	.049539	-4.04	0.000	-.2974515	-.1032624
_cons	-5.359375	1.000514	-5.36	0.000	-7.320346	-3.398404

/lnsig2u	2.947234	.0435842	2.861811	3.032657
sigma_u	4.364995	.0951224	4.182484	4.55547
rho	.9501326	.002065	.945926	.9540279

Likelihood-ratio test of rho=0: chibar2(01) = 1.6e+04 Prob >= chibar2 = 0.000

. \* Scale factor for coefficients:

. di 1/sqrt(1 + e(sigma\_u)^2)

.22331011

## Solutions to Chapter 16 Problems

16.1. a. The Stata output below contains the estimates for 1981 and, for completeness, 1987.

Certainly some magnitudes are fairly different. For example, education has a much larger effect in the latter time period. Also, the effect of experience on the log-odds ratios are quite different.

```
. mlogit status educ exper expersq black if y81, base(0)
```

```
Multinomial logistic regression      Number of obs   =      1737
                                     LR chi2(8)        =      720.39
                                     Prob > chi2       =      0.0000
Log likelihood = -1502.9396          Pseudo R2       =      0.1933
```

	status	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
0		(base outcome)					
1							
	educ	-.47558	.0466559	-10.19	0.000	-.5670238	-.3841361
	exper	3.016025	.4513224	6.68	0.000	2.131449	3.900601
	expersq	-.5953032	.2690175	-2.21	0.027	-1.122568	-.0680386
	black	.8649358	.1302512	6.64	0.000	.6096481	1.120224
	_cons	4.138761	.5276112	7.84	0.000	3.104662	5.17286
2							
	educ	-.1019564	.0495931	-2.06	0.040	-.1991571	-.0047558
	exper	4.101794	.4359451	9.41	0.000	3.247357	4.956231
	expersq	-.7069626	.2628842	-2.69	0.007	-1.222206	-.1917191
	black	.0208189	.1436123	0.14	0.885	-.2606561	.3022938
	_cons	-.0313456	.5828582	-0.05	0.957	-1.173727	1.111035

```
. mlogit status educ exper expersq black if y87, base(0)
```

```
Multinomial logistic regression      Number of obs   =      1717
                                     LR chi2(8)        =      583.72
                                     Prob > chi2       =      0.0000
Log likelihood = -907.85723          Pseudo R2       =      0.2433
```

status		Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
0		(base outcome)					
1							
	educ	-.6736313	.0698999	-9.64	0.000	-.8106325	-.53663
	exper	-.1062149	.173282	-0.61	0.540	-.4458414	.2334116
	expersq	-.0125152	.0252291	-0.50	0.620	-.0619633	.036933
	black	.8130166	.3027231	2.69	0.007	.2196902	1.406343

	_cons	10.27787	1.133336	9.07	0.000	8.056578	12.49917
2							
	educ	-.3146573	.0651096	-4.83	0.000	-.4422699	-.1870448
	exper	.8487367	.1569856	5.41	0.000	.5410507	1.156423
	expersq	-.0773003	.0229217	-3.37	0.001	-.1222261	-.0323746
	black	.3113612	.2815339	1.11	0.269	-.240435	.8631574
	_cons	5.543798	1.086409	5.10	0.000	3.414475	7.673121

b. Just adding year dummies is probably not sufficient, given the findings in part a, but the results are below. Because the model is static and we have panel data, we should use inference robust to arbitrary serial dependence. In this application, the robust standard errors are typically larger but the difference is not huge.

```
. mlogit status educ exper expersq black y82-y87, base(0)
```

Multinomial logistic regression	Number of obs	=	12108
	LR chi2(20)	=	6409.72
	Prob > chi2	=	0.0000
Log likelihood = -8842.6383	Pseudo R2	=	0.2660

	status	Coef.	Std. Err.	z	P> z	[95% Conf. Interval
0		(base outcome)				
1						
	educ	-.5473739	.0189537	-28.88	0.000	-.5845225 - .5102253
	exper	.769957	.0633149	12.16	0.000	.6458621 .8940519
	expersq	-.1153749	.0107134	-10.77	0.000	-.1363729 -.094377
	black	.8773806	.0656223	13.37	0.000	.7487633 1.005998
	y82	.9871298	.0928663	10.63	0.000	.8051152 1.169144
	y83	1.383591	.1035337	13.36	0.000	1.180669 1.586514
	y84	1.587213	.115548	13.74	0.000	1.360743 1.813683
	y85	2.052594	.1307157	15.70	0.000	1.796396 2.308792
	y86	2.652847	.1513588	17.53	0.000	2.356189 2.949505
	y87	2.727265	.1701085	16.03	0.000	2.393858 3.060671
	_cons	5.151552	.2282352	22.57	0.000	4.704219 5.598885
2						
	educ	-.2555556	.0182414	-14.01	0.000	-.291308 - .2198032
	exper	1.823821	.058522	31.16	0.000	1.70912 1.938522
	expersq	-.195654	.0095781	-20.43	0.000	-.2144267 -.1768813
	black	.33846	.0649312	5.21	0.000	.2111972 .4657227
	y82	.5624964	.0936881	6.00	0.000	.3788712 .7461217
	y83	1.225732	.0998516	12.28	0.000	1.030027 1.421438
	y84	1.42652	.1095939	13.02	0.000	1.21172 1.64132
	y85	1.662994	.1243071	13.38	0.000	1.419357 1.906632
	y86	2.029585	.1447257	14.02	0.000	1.745928 2.313242
	y87	1.995639	.1622294	12.30	0.000	1.677675 2.313603
	_cons	1.858323	.225749	8.23	0.000	1.415863 2.300783

```
. mlogit status educ exper expersq black y82-y87, base(0) cluster(id)
```

```

Multinomial logistic regression      Number of obs   =      12108
                                     Wald chi2(20)      =      2742.09
                                     Prob > chi2        =      0.0000
Log pseudolikelihood = -8842.6383    Pseudo R2       =      0.2660

```

(Std. Err. adjusted for 1738 clusters in id)

status	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval	
0	(base outcome)					
1						
educ	-.5473739	.0200999	-27.23	0.000	-.586769	-.5079789
exper	.769957	.0776371	9.92	0.000	.617791	.922123
expersq	-.1153749	.0106075	-10.88	0.000	-.1361653	-.0945846
black	.8773806	.0855443	10.26	0.000	.7097169	1.045044
y82	.9871298	.0760747	12.98	0.000	.8380261	1.136234
y83	1.383591	.0888752	15.57	0.000	1.209399	1.557784
y84	1.587213	.1050477	15.11	0.000	1.381323	1.793103
y85	2.052594	.1275644	16.09	0.000	1.802572	2.302615
y86	2.652847	.1526831	17.37	0.000	2.353593	2.9521
y87	2.727265	.1666166	16.37	0.000	2.400702	3.053827
_cons	5.151552	.2523957	20.41	0.000	4.656866	5.646238
2						
educ	-.2555556	.0177679	-14.38	0.000	-.29038	-.2207312
exper	1.823821	.0731396	24.94	0.000	1.68047	1.967172
expersq	-.195654	.010131	-19.31	0.000	-.2155104	-.1757976
black	.33846	.0783575	4.32	0.000	.1848821	.4920378
y82	.5624964	.0796845	7.06	0.000	.4063177	.7186751
y83	1.225732	.0897086	13.66	0.000	1.049907	1.401558
y84	1.42652	.1027116	13.89	0.000	1.225209	1.627831
y85	1.662994	.124454	13.36	0.000	1.419069	1.90692
y86	2.029585	.1526669	13.29	0.000	1.730363	2.328807
y87	1.995639	.1636634	12.19	0.000	1.674865	2.316414
_cons	1.858323	.2257666	8.23	0.000	1.415829	2.300817

c. The time dummies have very large  $t$  statistics, and the robust joint test gives a  $\chi^2_{12}$  value of 624.28, which implies a zero  $p$ -value to many decimal places.

d. After obtaining the estimates from part c, the following commands produce the change in the estimated employment probabilities. It is about .021 for 1981, and about .058 for 1987.

```

. di exp([2]_cons + [2]educ*16 + [2]exper*5 + [2]expersq*25 + [2]black)
    /(1 + exp([1]_cons + [1]educ*16 + [1]exper*5 + [1]expersq*25 + [1]black)
      + exp([2]_cons + [2]educ*16 + [2]exper*5 + [2]expersq*25 + [2]black))
.89820453

```

```

. di exp([2]_cons + [2]educ*12 + [2]exper*5 + [2]expersq*25 + [2]black)
  /(1 + exp([1]_cons + [1]educ*12 + [1]exper*5 + [1]expersq*25 + [1]black)
    + exp([2]_cons + [2]educ*12 + [2]exper*5 + [2]expersq*25 + [2]black))
.91903414

. di .91903414 - .89820453
.02082961

. di exp([2]_cons + [2]educ*12 + [2]exper*5 + [2]expersq*25 + [2]black + [2]y87
  /(1 + exp([1]_cons + [1]educ*12 + [1]exper*5 + [1]expersq*25 + [1]black
    + [1]y87)
    + exp([2]_cons + [2]educ*12 + [2]exper*5 + [2]expersq*25 + [2]black
    + [2]y87))
.89646574

. di exp([2]_cons + [2]educ*16 + [2]exper*5 + [2]expersq*25 + [2]black + [2]y87
  /(1 + exp([1]_cons + [1]educ*16 + [1]exper*5 + [1]expersq*25 + [1]black
    + [1]y87)
    + exp([2]_cons + [2]educ*16 + [2]exper*5 + [2]expersq*25 + [2]black
    + [2]y87))
.95454392

. di .95454392 - .89646574
.05807818

```

**16.2. a.** The following Stata output contains the linear regression results. Because *pctstck* is discrete (taking on the values 0, 50, and 100), it seems likely that heteroskedasticity is present in a linear model. In fact, the robust standard errors are not very different from the usual ones.

```

. use pension
. tab pctstck

```

0=mstbnds,5 0=mixed,100 =mststcks		Freq.	Percent	Cum.
0		78	34.51	34.51
50		85	37.61	72.12
100		63	27.88	100.00
Total		226	100.00	

```

. reg pctstck choice age educ female black married finc25-finc101 wealth89
  prftshr, robust

```

Linear regression	Number of obs =	194
	F( 14, 179) =	2.15
	Prob > F =	0.0113
	R-squared =	0.0998
	Root MSE =	39.134

---

	Robust
--	--------



pctstck	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
choice	12.04773	5.994437	2.01	0.046	.2188713	23.87658
age	-1.625967	.8327895	-1.95	0.052	-3.269315	.0173813
educ	.7538685	1.172328	0.64	0.521	-1.559493	3.06723
female	1.302856	7.148595	0.18	0.856	-12.80351	15.40922
black	3.967391	8.974971	0.44	0.659	-13.74297	21.67775
married	3.303436	8.369616	0.39	0.694	-13.21237	19.81924
finc25	-18.18567	16.00485	-1.14	0.257	-49.76813	13.39679
finc35	-3.925374	15.86275	-0.25	0.805	-35.22742	27.37668
finc50	-8.128784	15.3762	-0.53	0.598	-38.47072	22.21315
finc75	-17.57921	16.6797	-1.05	0.293	-50.49335	15.33493
finc100	-6.74559	16.7482	-0.40	0.688	-39.7949	26.30372
finc101	-28.34407	16.57814	-1.71	0.089	-61.05781	4.369672
wealth89	-.0026918	.0114136	-0.24	0.814	-.0252142	.0198307
prftshr	15.80791	8.107663	1.95	0.053	-.1909844	31.80681
_cons	134.1161	58.87288	2.28	0.024	17.9419	250.2902

b. With relatively few husband-wife pairs – 23 in this application – we do not expect big differences in standard errors, and we do not see them. On the key variable, *choice*, the cluster-robust standard error is only slightly larger. (Incidentally, this part really should not come until Chapter 20.)

```
. reg pctstck choice age educ female black married finc25-finc101 wealth89
    prftshr, cluster(id)
```

Linear regression	Number of obs =	194
	F( 14, 170) =	2.12
	Prob > F =	0.0128
	R-squared =	0.0998
	Root MSE =	39.134

(Std. Err. adjusted for 171 clusters in id)

pctstck	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
choice	12.04773	6.184085	1.95	0.053	-.1597617	24.25521
age	-1.625967	.8192942	-1.98	0.049	-3.243267	-.0086663
educ	.7538685	1.1803	0.64	0.524	-1.576064	3.083801
female	1.302856	7.000538	0.19	0.853	-12.51632	15.12203
black	3.967391	8.711611	0.46	0.649	-13.22948	21.16426
married	3.303436	8.624168	0.38	0.702	-13.72082	20.32769
finc25	-18.18567	16.82939	-1.08	0.281	-51.40716	15.03583
finc35	-3.925374	16.17574	-0.24	0.809	-35.85656	28.00581
finc50	-8.128784	15.91447	-0.51	0.610	-39.54421	23.28665
finc75	-17.57921	17.2789	-1.02	0.310	-51.68804	16.52963
finc100	-6.74559	17.24617	-0.39	0.696	-40.78983	27.29865
finc101	-28.34407	17.10783	-1.66	0.099	-62.1152	5.42707
wealth89	-.0026918	.0119309	-0.23	0.822	-.0262435	.02086
prftshr	15.80791	8.356266	1.89	0.060	-.6874979	32.30332
_cons	134.1161	58.1316	2.31	0.022	19.36333	248.8688

```
-----
. di _b[_cons] + _b[age]*60 + _b[educ]*12 + _b[female] + _b[finc50] + _b[wealth89]
38.374791
```

```
. di _b[_cons] + _b[age]*60 + _b[educ]*12 + _b[female] + _b[finc50] + _b[wealth89]
50.422517
```

For later use, the predicted *pctstck* for the person described in the problem, with *choice* = 0 is about 38.37. With choice, it is roughly 50.42.

c. The ordered probit estimates follow, including commands that provide the predictions

for *pctstck* with and without choice:

```
. oprobit pctstck choice age educ female black married finc25-finc101 wealth89
prftshr
```

```
Iteration 0: log likelihood = -212.37031
Iteration 1: log likelihood = -202.0094
Iteration 2: log likelihood = -201.9865
Iteration 3: log likelihood = -201.9865
```

Ordered probit regression	Number of obs	=	194
	LR chi2(14)	=	20.77
	Prob > chi2	=	0.1077
Log likelihood = -201.9865	Pseudo R2	=	0.0489

pctstck	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
choice	.371171	.1841121	2.02	0.044	.010318	.7320241
age	-.0500516	.0226063	-2.21	0.027	-.0943591	-.005744
educ	.0261382	.0352561	0.74	0.458	-.0429626	.0952389
female	.0455642	.206004	0.22	0.825	-.3581963	.4493246
black	.0933923	.2820403	0.33	0.741	-.4593965	.6461811
married	.0935981	.2332114	0.40	0.688	-.3634878	.550684
finc25	-.5784299	.423162	-1.37	0.172	-1.407812	.2509524
finc35	-.1346721	.4305242	-0.31	0.754	-.9784841	.7091399
finc50	-.2620401	.4265936	-0.61	0.539	-1.098148	.5740681
finc75	-.5662312	.4780035	-1.18	0.236	-1.503101	.3706385
finc100	-.2278963	.4685942	-0.49	0.627	-1.146324	.6905316
finc101	-.8641109	.5291111	-1.63	0.102	-1.90115	.1729279
wealth89	-.0000956	.0003737	-0.26	0.798	-.0008279	.0006368
prftshr	.4817182	.2161233	2.23	0.026	.0581243	.905312
/cut1	-3.087373	1.623765			-6.269894	.0951479
/cut2	-2.053553	1.618611			-5.225972	1.118865

```
. di b[age]*60 + _b[educ]*12 + _b[female] + _b[finc50] + _b[wealth89]*150
-2.9202491
```

```
. di normal(_b[/cut2] + 2.9202491) - normal(_b[/cut1] + 2.9202491)
.37330935
```

```

. di 1 - normal(_b[/cut2] + 2.9202491)
.19305438

. di 50*.37330935 + 100*.19305438
37.970906

. di _b[age]*60 + _b[educ]*12 + _b[female] + _b[finc50] + _b[wealth89]*150
    + _b[choice]
-2.5490781

. di normal(_b[/cut2] + 2.5490781) - normal(_b[/cut1] + 2.5490781)
.39469838

. di 1 - normal(_b[/cut2] + 2.5490781)
.31011489

. di 50*.39469838 + 100*.31011489
50.746408

. di 50.75 - 37.97
12.78

```

Using ordered probit, the effect of having choice for this person is about 12.8 percentage points more invested in the stock market, which is pretty similar to the 12.1 points obtained with the linear model.

d. We can compute an  $R$ -squared for the ordered probit model by using the squared correlation between the predicted  $pctstck_i$  and the actual. The following Stata session does this, after using the `oprobit` command. The squared correlation for ordered probit is about .097, which is actually slightly below the linear model  $R$ -squared, .098. The correlation between the fitted values for the linear and OP models is very high: .998.

```

. qui oprobit pctstck choice age educ female black married finc25-finc101 wealth89

. predict p1hat p2hat p3hat
(option pr assumed; predicted probabilities)
(32 missing values generated)

. sum p1hat p2hat p3hat

```

Variable	Obs	Mean	Std. Dev.	Min	Max
p1hat	194	.331408	.1327901	.0685269	.8053644
p2hat	194	.3701685	.0321855	.1655734	.3947809
p3hat	194	.2984236	.1245914	.0290621	.6747374

```

. gen pctstck_op = 50*p2hat + 100*p3hat

```

(32 missing values generated)

```
. corr pctstck pctstck_op
(obs=194)
```

	pctstck	pctstck_op
pctstck	1.0000	
pctstck_op	0.3119	1.0000

```
. di .312^2
.097344
```

```
. qui reg pctstck choice age educ female black married finc25-finc101 wealth89
```

```
. predict pctstck_lin
(option xb assumed; fitted values)
(32 missing values generated)
```

```
. corr pctstck_lin pctstck_op
(obs=194)
```

	pctstck_lin	pctstck_op
pctstck_lin	1.0000	
pctstck_op	0.9980	1.0000

**16.3. a.** We can derive the response probabilities from the latent variable formulation in (16.21) and the rule in (16.22).

$$\begin{aligned}\exp(-\mathbf{x}_1\boldsymbol{\delta})y^* &= \exp(-\mathbf{x}_1\boldsymbol{\delta})\mathbf{x}\boldsymbol{\beta} + \exp(-\mathbf{x}_1\boldsymbol{\delta})e \\ &= \exp(-\mathbf{x}_1\boldsymbol{\delta})\mathbf{x}\boldsymbol{\beta} + a\end{aligned}$$

where

$$a|\mathbf{x} \sim \text{Normal}(0, 1).$$

Now  $\alpha_j < y^* \leq \alpha_{j+1}$  if and only if  $\exp(-\mathbf{x}_1\boldsymbol{\delta})\alpha_j < \exp(-\mathbf{x}_1\boldsymbol{\delta})y^* \leq \exp(-\mathbf{x}_1\boldsymbol{\delta})\alpha_{j+1}$ , and so

$$\begin{aligned}P(y = j|\mathbf{x}) &= P[\exp(-\mathbf{x}_1\boldsymbol{\delta})\alpha_j < \exp(-\mathbf{x}_1\boldsymbol{\delta})\mathbf{x}\boldsymbol{\beta} + a \leq \exp(-\mathbf{x}_1\boldsymbol{\delta})\alpha_{j+1}|\mathbf{x}] \\ &= P[\exp(-\mathbf{x}_1\boldsymbol{\delta})\alpha_j - \exp(-\mathbf{x}_1\boldsymbol{\delta})\mathbf{x}\boldsymbol{\beta} < a \leq \exp(-\mathbf{x}_1\boldsymbol{\delta})\alpha_{j+1} - \exp(-\mathbf{x}_1\boldsymbol{\delta})\mathbf{x}\boldsymbol{\beta}|\mathbf{x}] \\ &= \Phi[\exp(-\mathbf{x}_1\boldsymbol{\delta})(\alpha_{j+1} - \mathbf{x}\boldsymbol{\beta})] - \Phi[\exp(-\mathbf{x}_1\boldsymbol{\delta})(\alpha_j - \mathbf{x}\boldsymbol{\beta})].\end{aligned}$$

A similar argument holds at  $j = 0$  and  $j = J$ . Therefore, as described in the text, the response probabilities for the heteroskedastic ordered probit are of the same form as usual ordered probit

but with  $\alpha_j - \mathbf{x}\boldsymbol{\beta}$  everywhere replaced with  $\exp(-\mathbf{x}_1\boldsymbol{\delta})(\alpha_j - \mathbf{x}\boldsymbol{\beta})$ .

b. We can obtain a useful VAT by applying the score statistic – just as in the binary probit case. The score of the log likelihood with respect to  $(\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ , evaluated at  $\boldsymbol{\delta} = \mathbf{0}$ , is easily seen to just be the usual score for ordered probit. For  $0 < j < J$ , the score of the response probability with respect to  $\boldsymbol{\delta}$  evaluated at  $\boldsymbol{\delta} = \mathbf{0}$  is

$$-1[y_i = j] \frac{\mathbf{x}'_{i1}[(\alpha_{j+1} - \mathbf{x}_i\boldsymbol{\beta})\phi(\alpha_{j+1} - \mathbf{x}_i\boldsymbol{\beta}) - (\alpha_j - \mathbf{x}_i\boldsymbol{\beta})\phi(\alpha_j - \mathbf{x}_i\boldsymbol{\beta})]}{\Phi(\alpha_{j+1} - \mathbf{x}_i\boldsymbol{\beta}) - \Phi(\alpha_j - \mathbf{x}_i\boldsymbol{\beta})}.$$

For  $j = 0$  and  $j = J$  we have

$$\begin{aligned} -1[y_i = 0] & \frac{\mathbf{x}'_{i1}[(\alpha_1 - \mathbf{x}_i\boldsymbol{\beta})\phi(\alpha_1 - \mathbf{x}_i\boldsymbol{\beta})]}{\Phi(\alpha_1 - \mathbf{x}_i\boldsymbol{\beta})} \\ 1[y_i = J] & \frac{\mathbf{x}'_{i1}[(\alpha_J - \mathbf{x}_i\boldsymbol{\beta})\phi(\alpha_J - \mathbf{x}_i\boldsymbol{\beta})]}{1 - \Phi(\alpha_J - \mathbf{x}_i\boldsymbol{\beta})} \end{aligned}$$

It is easily seen that these are identical to the scores that would be obtained by adding

$$-\mathbf{x}_{i1}(\alpha_j - \mathbf{x}_i\boldsymbol{\beta})$$

as a set of explanatory variables to the usual OP model and testing their joint significance. In practice, we would replace the  $\alpha_j$  and  $\boldsymbol{\beta}$  with the MLEs from the original OP problem.

d. The ASF can be written as

$$\begin{aligned} \text{ASF}(\mathbf{x}) &= E_{e_i}(1[\alpha_1 - \mathbf{x}\boldsymbol{\beta} < e_i \leq \alpha_2 - \mathbf{x}\boldsymbol{\beta}]) \\ &= P(\alpha_1 - \mathbf{x}\boldsymbol{\beta} < e_i \leq \alpha_2 - \mathbf{x}\boldsymbol{\beta}) \\ &= F_e(\alpha_2 - \mathbf{x}\boldsymbol{\beta}) - F_e(\alpha_1 - \mathbf{x}\boldsymbol{\beta}) \end{aligned}$$

where  $F_e(\cdot)$  is the cdf of  $e_i$ . We do not know  $F_e$  because it depends on the distribution of  $\mathbf{x}_1$ : we have specified  $D(e_i|\mathbf{x}_i) = D(e_i|\mathbf{x}_{i1})$ , not  $D(e_i)$ .

e. From iterated expectations we can write

$$\text{ASF}(\mathbf{x}) = E_{\mathbf{x}_{i1}}\{E(1[\alpha_1 - \mathbf{x}\boldsymbol{\beta} < e_i \leq \alpha_2 - \mathbf{x}\boldsymbol{\beta}]|\mathbf{x}_{i1})\}$$

and the conditional expectation is a conditional probability:

$$\begin{aligned} E(1[\alpha_1 - \mathbf{x}\boldsymbol{\beta} < e_i \leq \alpha_2 - \mathbf{x}\boldsymbol{\beta}] | \mathbf{x}_{i1}) &= P(\alpha_1 - \mathbf{x}\boldsymbol{\beta} < e_i \leq \alpha_2 - \mathbf{x}\boldsymbol{\beta} | \mathbf{x}_{i1}) \\ &= P[\exp(-\mathbf{x}_{i1}\boldsymbol{\delta})(\alpha_1 - \mathbf{x}\boldsymbol{\beta}) < a_i \leq \exp(-\mathbf{x}_{i1}\boldsymbol{\delta})(\alpha_2 - \mathbf{x}\boldsymbol{\beta}) | \mathbf{x}_{i1}] \\ &= \Phi[\exp(-\mathbf{x}_{i1}\boldsymbol{\delta})(\alpha_2 - \mathbf{x}\boldsymbol{\beta})] - \Phi[\exp(-\mathbf{x}_{i1}\boldsymbol{\delta})(\alpha_1 - \mathbf{x}\boldsymbol{\beta})]. \end{aligned}$$

Therefore,

$$\text{ASF}(\mathbf{x}) = E_{\mathbf{x}_{i1}} \{ \Phi[\exp(-\mathbf{x}_{i1}\boldsymbol{\delta})(\alpha_2 - \mathbf{x}\boldsymbol{\beta})] - \Phi[\exp(-\mathbf{x}_{i1}\boldsymbol{\delta})(\alpha_1 - \mathbf{x}\boldsymbol{\beta})] \}.$$

By the law of large numbers, a consistent estimator is

$$N^{-1} \sum_{i=1}^N \{ \Phi[\exp(-\mathbf{x}_{i1}\boldsymbol{\delta})(\alpha_2 - \mathbf{x}\boldsymbol{\beta})] - \Phi[\exp(-\mathbf{x}_{i1}\boldsymbol{\delta})(\alpha_1 - \mathbf{x}\boldsymbol{\beta})] \}$$

and, by Lemma 12.1, consistency is preserved if we insert the (consistent) MLES:

$$\widehat{\text{ASF}}(\mathbf{x}) = N^{-1} \sum_{i=1}^N \{ \Phi[\exp(-\mathbf{x}_{i1}\hat{\boldsymbol{\delta}})(\hat{\alpha}_2 - \mathbf{x}\hat{\boldsymbol{\beta}})] - \Phi[\exp(-\mathbf{x}_{i1}\hat{\boldsymbol{\delta}})(\hat{\alpha}_1 - \mathbf{x}\hat{\boldsymbol{\beta}})] \}.$$

The APEs are estimated by taking derivatives or changes with respect to elements of  $\mathbf{x}$  in

$\widehat{\text{ASF}}(\mathbf{x})$ .

**16.4. a.** The results of the ordered probit estimation using *invest* as the response variable are given below. Every statistic is identical to when *pctstck* is used as the response variable. This is as it should be, as only the order of the outcomes matter – not the magnitudes.

```
. gen invest = 0 if pctstck == 0
(148 missing values generated)
```

```
. replace invest = 1 if pctstck == 50
(85 real changes made)
```

```
. replace invest = 2 if pctstck == 100
(63 real changes made)
```

```
. oprobit invest choice age educ female black married finc25-finc101 wealth89
```

Ordered probit regression	Number of obs	=	194
	LR chi2(14)	=	20.77
	Prob > chi2	=	0.1077

Log likelihood = -201.9865

Pseudo R2

=

0.0489

invest	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
choice	.371171	.1841121	2.02	0.044	.010318	.7320241
age	-.0500516	.0226063	-2.21	0.027	-.0943591	-.005744
educ	.0261382	.0352561	0.74	0.458	-.0429626	.0952389
female	.0455642	.206004	0.22	0.825	-.3581963	.4493246
black	.0933923	.2820403	0.33	0.741	-.4593965	.6461811
married	.0935981	.2332114	0.40	0.688	-.3634878	.550684
finc25	-.5784299	.423162	-1.37	0.172	-1.407812	.2509524
finc35	-.1346721	.4305242	-0.31	0.754	-.9784841	.7091399
finc50	-.2620401	.4265936	-0.61	0.539	-1.098148	.5740681
finc75	-.5662312	.4780035	-1.18	0.236	-1.503101	.3706385
finc100	-.2278963	.4685942	-0.49	0.627	-1.146324	.6905316
finc101	-.8641109	.5291111	-1.63	0.102	-1.90115	.1729279
wealth89	-.0000956	.0003737	-0.26	0.798	-.0008279	.0006368
prftshr	.4817182	.2161233	2.23	0.026	.0581243	.905312
/cut1	-3.087373	1.623765			-6.269894	.0951479
/cut2	-2.053553	1.618611			-5.225972	1.118865

b. One quantity that would change is the estimated expected value, something pretty

obvious because of the rescaling. In particular,

$$\hat{E}(\text{invest}|\mathbf{x}) = \hat{P}(\text{invest} = 1|\mathbf{x}) + 2 \cdot \hat{P}(\text{invest} = 2|\mathbf{x})$$

whereas

$$\begin{aligned}\hat{E}(\text{pctstck}|\mathbf{x}) &= 50 \cdot \hat{P}(\text{pctstck} = 50|\mathbf{x}) + 100 \cdot \hat{P}(\text{pctstck} = 100|\mathbf{x}) \\ &= 50 \cdot \hat{P}(\text{invest} = 1|\mathbf{x}) + 100 \cdot \hat{P}(\text{invest} = 2|\mathbf{x}) \\ &= 50 \cdot \hat{E}(\text{invest}|\mathbf{x}).\end{aligned}$$

Because  $\text{pctstck} = 50 \cdot \text{invest}$ ,  $E(\text{pctstck}|\mathbf{x}) = 50 \cdot E(\text{invest}|\mathbf{x})$ .

**16.5.** a. We have

$$D(y_2|\mathbf{z}) = \text{Normal}(\mathbf{z}\delta_2, \exp(\mathbf{z}\xi_2))$$

which means we should use maximum likelihood to estimate  $\delta_2$  and  $\xi_2$ . In fact,  $\hat{\delta}_2$  is

asymptotically equivalent to a weighted least squares estimator using weights  $\exp(-\mathbf{z}_i\xi_2)$ .

b. By assumption,  $(u_1, e_2)$  is independent of  $\mathbf{z}$  and so  $D(u_1|e_2, \mathbf{z}) = D(u_1|e_2)$ . Because

$(u_1, e_2)$  is bivariate normal with zero mean, we can always write

$$u_1 = \theta_1 e_2 + e_1$$

where

$$e_1|e_2 \sim \text{Normal}(0, \tau_1^2)$$

where  $\tau_1^2 = \sigma_1^2 - \theta_1^2$ , where  $\sigma_1^2 = \text{Var}(u_1)$ . This is necessarily the distribution conditional on  $\mathbf{z}$ , too.

We can write

$$y_2 = \mathbf{z}\boldsymbol{\delta}_2 + \exp(\mathbf{z}\boldsymbol{\xi}_2/2)e_2,$$

which shows that  $y_2$  is a function of  $(\mathbf{z}, e_2)$ . Therefore,  $e_1$  is independent of  $y_2$ , too.

c. We can use the latent variable formulation in equation (16.30) and insert  $u_1 = \theta_1 e_2 + e_1$ :

$$\begin{aligned} y_1^* &= \mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 y_2 + \theta_1 e_2 + e_1 \\ e_1|\mathbf{z}, y_2 &\sim \text{Normal}(0, \tau_1^2) \end{aligned}$$

To obtain an error with a unit variance, we divide by  $\tau_1$ :

$$(y_1^*/\tau_1) = \mathbf{z}_1(\boldsymbol{\delta}_1/\tau_1) + (\gamma_1/\tau_1)y_2 + (\theta_1/\tau_1)e_2 + (e_1/\tau_1)$$

and then the cut parameters also get divided by  $\tau_1$ . For example,  $\alpha_j < y_1^* \leq \alpha_{j+1}$  if and only if

$$(\alpha_j/\tau_1) < (y_1^*/\tau_1) \leq (\alpha_{j+1}/\tau_1)$$

Therefore, if we run ordered probit of

$$y_1 \text{ on } \mathbf{z}_1, y_2, e_2$$

we consistently estimate all parameters multiplied by  $1/\tau_1$ . Of course we do not observe  $e_2$ , but

we can replace it with estimates because  $e_2 = \exp(-\mathbf{z}\boldsymbol{\xi}_2/2)v_2$ .

The two-step approach is to estimate  $\boldsymbol{\delta}_2$  and  $\boldsymbol{\xi}_2$  by the MLE from part a. Then create



$$\begin{aligned}\hat{v}_{i2} &= y_{i2} - \mathbf{z}_i \hat{\boldsymbol{\delta}}_2 \\ \hat{e}_{i2} &= \exp(-\mathbf{z}_i \hat{\boldsymbol{\xi}}_2 / 2) \hat{v}_{i2}\end{aligned}$$

In the second step, estimate the scaled coefficients by OP of

$$y_{i1} \text{ on } \mathbf{z}_{i1}, y_{i2}, \hat{e}_{i2}.$$

Let  $\hat{\alpha}_{\tau j}, j = 1, 2, \dots, J, \hat{\boldsymbol{\delta}}_{\tau 1}, \hat{\gamma}_{\tau 1}$ , and  $\hat{\theta}_{\tau 1}$  be the scaled coefficients.

Incidentally, a simple test of the null that  $y_2$  is exogenous is the usual MLE  $t$  statistic for  $\hat{\theta}_{\tau 1}$ .

d. We can obtain the ASF by averaging out  $e_2$  in response probabilities of the form

$$\Phi(\alpha_{\tau j+1} - \mathbf{z}_1 \boldsymbol{\delta}_{\tau 1} - \gamma_{\tau 1} y_2 - \theta_{\tau 1} e_2) - \Phi(\alpha_{\tau j} - \mathbf{z}_1 \boldsymbol{\delta}_{\tau 1} - \gamma_{\tau 1} y_2 - \theta_{\tau 1} e_2)$$

(for  $0 < j < J$ ). A consistent estimator of the ASF is

$$\widehat{\text{ASF}}(\mathbf{z}_1, y_2) = N^{-1} \sum_{i=1}^N [\Phi(\hat{\alpha}_{\tau j+1} - \mathbf{z}_1 \hat{\boldsymbol{\delta}}_{\tau 1} - \hat{\gamma}_{\tau 1} y_2 - \hat{\theta}_{\tau 1} \hat{e}_{i2}) - \Phi(\hat{\alpha}_{\tau j} - \mathbf{z}_1 \hat{\boldsymbol{\delta}}_{\tau 1} - \hat{\gamma}_{\tau 1} y_2 - \hat{\theta}_{\tau 1} \hat{e}_{i2})].$$

and, as usual, we can compute derivatives or changes with respect to the elements of  $(\mathbf{z}_1, y_2)$ .

e. Now the normal MLE is just applied to

$$\log(y_2) | \mathbf{z} \sim \text{Normal}(\mathbf{z} \boldsymbol{\delta}_2, \exp(\mathbf{z} \boldsymbol{\xi}_2))$$

and  $\hat{v}_{i2} = \log(y_{i2}) - \mathbf{z}_i \hat{\boldsymbol{\delta}}_2$ .

f. Without a distributional assumption for  $D(u_2 | e_2)$ , allowing for endogeneity is tricky. We would still assume that  $(u_1, e_2)$  is independent of  $\mathbf{z}$ . We could just *assume* we can write

$u_2 = \theta_1 e_2 + e_1$  where

$$D(e_1 | e_2, \mathbf{z}) \sim \text{Normal}(0, \tau_1^2).$$

Then the two-step method from part c, with ASF estimated as in part d, applies but where  $\hat{\boldsymbol{\delta}}_2$

and  $\hat{\xi}_2$  are obtained from a suitable estimation procedure. It could be a several step procedure or, more conveniently, a single step based on the normal quasi-MLE. That is, we act *as if*

$$D(y_2|\mathbf{z}) = \text{Normal}(\exp(\mathbf{z}\boldsymbol{\delta}_2), \exp(\mathbf{z}\boldsymbol{\xi}_2))$$

even though it cannot be literally true. As the results of Gourieroux, Monfort, and Trognon (1984a) show, this estimator is generally consistent and  $\sqrt{N}$ -asymptotically normal. Then

$$\begin{aligned}\hat{v}_{i2} &= y_{i2} - \exp(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) \\ \hat{e}_{i2} &= \exp(-\mathbf{z}_i\hat{\boldsymbol{\xi}}_2/2)\hat{v}_{i2}\end{aligned}$$

and the steps in part c can be followed.

A way to make the method more flexible is to add polynomials in  $\hat{e}_{i2}$  to the second-stage OP. For example, if we just add a square, the ASF would be estimated as

$$\begin{aligned}\widehat{\text{ASF}}(\mathbf{z}_1, y_2) &= N^{-1} \sum_{i=1}^N [\Phi(\hat{\alpha}_{\tau,j+1} - \mathbf{z}_1\hat{\boldsymbol{\delta}}_{\tau 1} - \hat{\gamma}_{\tau 1}y_2 - \hat{\theta}_{\tau 1}\hat{e}_{i2} - \hat{\eta}_{\tau 1}\hat{e}_{i2}^2) \\ &\quad - \Phi(\hat{\alpha}_{\tau j} - \mathbf{z}_1\hat{\boldsymbol{\delta}}_{\tau 1} - \hat{\gamma}_{\tau 1}y_2 - \hat{\theta}_{\tau 1}\hat{e}_{i2} - \hat{\eta}_{\tau 1}\hat{e}_{i2}^2)].\end{aligned}$$

where  $\hat{\eta}_{\tau 1}$  is the estimate on the quadratic term.

**16.6.** This problem is similar to that treated in Papke and Wooldridge (2008) for a binary or fractional response variable. Using the expression for  $c_{i1}$  we can write

$$\begin{aligned}y_{it1}^* &= \mathbf{z}_{it1}\boldsymbol{\delta}_1 + \gamma_1 y_{it2} + \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + a_{i1} + u_{it1} \\ &\equiv \mathbf{z}_{it1}\boldsymbol{\delta}_1 + \gamma_1 y_{it2} + \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + v_{it1}\end{aligned}$$

where  $v_{it1} \equiv a_{i1} + u_{it1}$ . Now we need to make some joint distributional assumptions concerning  $v_{it1}$  and  $v_{it2}$ , where

$$y_{it2} = \mathbf{z}_{it}\boldsymbol{\delta}_2 + \psi_2 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2 + v_{it2}$$

Given the marginal normal distributions assumed in the problem, it is a small step to assuming

$$v_{it1} = \theta_1 v_{it2} + e_{it1}$$

where

$$D(e_{it1}|v_{it2}, \mathbf{z}_i) = \text{Normal}(0, \tau_1^2).$$

We could allow  $\theta_1$ , and even  $\tau_1^2$ , to depend on  $t$ .

Now, we can write a control function equation (in latent variable form) as

$$y_{it1}^* = \mathbf{z}_{it1} \boldsymbol{\delta}_1 + \gamma_1 y_{it2} + \psi_1 + \bar{\mathbf{z}}_i \boldsymbol{\xi}_1 + \theta_1 v_{it2} + e_{it1};$$

given the conditional normality assumption for  $e_{it1}$ , and so using pooled probit of

$$y_{it1} \text{ on } \mathbf{z}_{it1}, y_{it2}, 1, \bar{\mathbf{z}}_i, v_{it2}, \quad t = 1, \dots, T; i = 1, \dots, N$$

consistently estimates all parameters – including the cut parameters – multiplied by  $1/\tau_1$ . The

two-step method is then (1) Estimate  $\boldsymbol{\delta}_2$ ,  $\psi_2$ , and  $\boldsymbol{\xi}_2$  by pooled OLS of

$$y_{it2} \text{ on } \mathbf{z}_{it}, 1, \bar{\mathbf{z}}_i, \quad t = 1, \dots, T; i = 1, \dots, N.$$

This is equivalent to fixed effects estimation of  $\boldsymbol{\delta}_2$ . Obtain the residuals,  $\hat{v}_{it2}$ . (2) Do pooled OP of

$$y_{it1} \text{ on } \mathbf{z}_{it1}, y_{it2}, 1, \bar{\mathbf{z}}_i, \hat{v}_{it2}, \quad t = 1, \dots, T; i = 1, \dots, N$$

to obtain  $\hat{\boldsymbol{\delta}}_{g1}$ ,  $\hat{\gamma}_{g1}$ , and so on. A simple extension is to interact  $\hat{v}_{it2}$  with time dummies to allow the regression of  $u_{it1}$  on  $v_{it2}$  to change over time.

b. Define a dummy variable  $w_{ij} = 1[y_{i1} = j]$ . Then

$$\begin{aligned} w_{ij} &= 1[\alpha_j < y_{i1}^* \leq \alpha_{j+1}] \\ &= 1[\alpha_j < \mathbf{z}_{it1} \boldsymbol{\delta}_1 + \gamma_1 y_{it2} + c_{i1} + u_{it1} \leq \alpha_{j+1}]. \end{aligned}$$

The ASF for  $w_{ij}$  is obtained by computing the expected value of the right hand side with respect to the unobservable  $c_{i1} + u_{it1}$  at specific values  $(\mathbf{z}_{i1}, y_{i2})$ :

$$\text{ASF}(\mathbf{z}_{t1}, y_{t2}) = E_{r_{it1}} \{1[\alpha_j < \mathbf{z}_{t1}\boldsymbol{\delta}_1 + \gamma_1 y_{t2} + r_{it1} \leq \alpha_{j+1}]\}$$

where  $r_{it1} = c_{it1} + u_{it1}$ . Note that  $r_{it1} = \psi_1 + \bar{\mathbf{z}}_i \boldsymbol{\xi}_1 + v_{it1}$  and so we can compute the ASF by taking the average over  $(\bar{\mathbf{z}}_i, v_{it1})$ :

$$\begin{aligned} \text{ASF}(\mathbf{z}_{t1}, y_{t2}) &= E_{(\bar{\mathbf{z}}_i, v_{it1})} \{1[\alpha_j < \mathbf{z}_{t1}\boldsymbol{\delta}_1 + \gamma_1 y_{t2} + \psi_1 + \bar{\mathbf{z}}_i \boldsymbol{\xi}_1 + v_{it1} \leq \alpha_{j+1}]\} \\ &= E_{(\bar{\mathbf{z}}_i, v_{it2}, e_{it1})} \{1[\alpha_j < \mathbf{z}_{t1}\boldsymbol{\delta}_1 + \gamma_1 y_{t2} + \psi_1 + \bar{\mathbf{z}}_i \boldsymbol{\xi}_1 + \theta_1 v_{it2} + e_{it1} \leq \alpha_{j+1}]\} \end{aligned}$$

Now we can apply iterated expectations. First find  $E(\cdot | \bar{\mathbf{z}}_i, v_{it2})$  and then average out  $(\bar{\mathbf{z}}_i, v_{it2})$ .

Now

$$\begin{aligned} E(1[\alpha_j < \mathbf{z}_{t1}\boldsymbol{\delta}_1 + \gamma_1 y_{t2} + \psi_1 + \bar{\mathbf{z}}_i \boldsymbol{\xi}_1 + \theta_1 v_{it2} + e_{it1} \leq \alpha_{j+1}] | \bar{\mathbf{z}}_i, v_{it2}) \\ = \Phi(\alpha_{g,j+1} - \mathbf{z}_{t1}\boldsymbol{\delta}_{g1} - \gamma_{g1} y_{t2} - \psi_{g1} - \bar{\mathbf{z}}_i \boldsymbol{\xi}_{g1} - \theta_{g1} v_{it2}) \\ - \Phi(\alpha_{g,j} - \mathbf{z}_{t1}\boldsymbol{\delta}_{g1} - \gamma_{g1} y_{t2} - \psi_{g1} - \bar{\mathbf{z}}_i \boldsymbol{\xi}_{g1} - \theta_{g1} v_{it2}) \end{aligned}$$

where “g” denotes divided by  $\tau_1$ . We use the fact that  $D(e_{it1} | v_{it2}, \mathbf{z}_i) = \text{Normal}(0, \tau_1^2)$ . It follows now by iterated expectations that

$$\begin{aligned} \text{ASF}(\mathbf{z}_{t1}, y_{t2}) &= E_{(\bar{\mathbf{z}}_i, v_{it2})} [\Phi(\alpha_{g,j+1} - \mathbf{z}_{t1}\boldsymbol{\delta}_{g1} - \gamma_{g1} y_{t2} - \psi_{g1} - \bar{\mathbf{z}}_i \boldsymbol{\xi}_{g1} - \theta_{g1} v_{it2}) \\ &\quad - \Phi(\alpha_{g,j} - \mathbf{z}_{t1}\boldsymbol{\delta}_{g1} - \gamma_{g1} y_{t2} - \psi_{g1} - \bar{\mathbf{z}}_i \boldsymbol{\xi}_{g1} - \theta_{g1} v_{it2})]. \end{aligned}$$

c. To estimate the ASF, we plug in estimates and use a sample average:

$$\begin{aligned} \widehat{\text{ASF}}(\mathbf{z}_{t1}, y_{t2}) &= N^{-1} \sum_{i=1}^n [\Phi(\hat{\alpha}_{g,j+1} - \mathbf{z}_{t1}\hat{\boldsymbol{\delta}}_{g1} - \hat{\gamma}_{g1} y_{t2} - \hat{\psi}_{g1} - \bar{\mathbf{z}}_i \hat{\boldsymbol{\xi}}_{g1} - \hat{\theta}_{g1} \hat{v}_{it2}) \\ &\quad - \Phi(\hat{\alpha}_{g,j} - \mathbf{z}_{t1}\hat{\boldsymbol{\delta}}_{g1} - \hat{\gamma}_{g1} y_{t2} - \hat{\psi}_{g1} - \bar{\mathbf{z}}_i \hat{\boldsymbol{\xi}}_{g1} - \hat{\theta}_{g1} \hat{v}_{it2})]. \end{aligned}$$

As usual, the estimated APEs are derivatives or changes with respect to  $(\mathbf{z}_{t1}, y_{t2})$ . To get valid standard errors, we can use Problem 12.17 or the panel bootstrap – where both estimation steps are carried out with each resampling.

d. The two-step control function procedure does not require any assumptions about the relationship between  $u_{it1}$  and  $v_{it2}$  for  $t \neq r$ . In other words, while adding  $v_{it2}$  as a control

function renders  $y_{it2}$  contemporaneously exogenous in the estimating equation –  $e_{it1}$  is independent of  $y_{it2}$  (and  $\mathbf{z}_i$ ) –  $\{y_{it2}\}$  is not generally strictly exogenous. An important implication is that we should not apply a method such as generalized estimating equations in the second stage.

A method that would render  $\{y_{it2}\}$  strictly exogenous would be to project  $v_{it1}$  on the entire history  $\{v_{ir2}, r = 1, \dots, T\}$ . There are assumptions under which the projection depends only on  $v_{it2}$  and the time average,  $\bar{v}_{i2}$ . So, we could write

$$v_{it1} = \theta_1 v_{it2} + \eta_1 \bar{v}_{i2} + e_{it1}$$

and assume  $e_{it1}$  is independent of  $\mathbf{v}_{i2} = (v_{i12}, \dots, v_{iT2})'$ . Then  $e_{it1}$  would be uncorrelated independent of  $(\mathbf{z}_i, \mathbf{y}_{i2})$  (under the other maintained assumptions). So, at each time period,  $(\hat{v}_{it2}, \hat{\bar{v}}_{i2})$  can be added to the ordered probit – that is, we apply the Mundlak device to the reduced form residuals. In addition to pooled OP, one could use a GEE-like procedure in the second stage.

More flexibility would be gotten by using the more general Chamberlain formulation:

$$v_{it1} = \mathbf{v}_{i2}' \boldsymbol{\theta}_{t1} + e_{it1}$$

where  $\boldsymbol{\theta}_{t1}$  is  $T \times 1$  for each  $t$ . Then in each time period include  $\hat{\mathbf{v}}_{i2}'$  as a set of regressors interacted with time-period dummies.

## Solutions to Chapter 17 Problems

17.1. a. No. Because  $\log(1) = 0$  and  $\log(\cdot)$  is strictly increasing,

$P[\log(1 + y) = 0] = P(y = 0) > 0$ . Of course,  $\log(1 + y)$  increases much more slowly than  $y$ , and so one could use  $\log(1 + y)$  to reduce the influence of “unusually” large observations  $y_i$  in linear regression. Also, remembering that the type I Tobit can be obtained from a latent variable model, the transformation  $\log(1 + y)$  might make the normality and homoskedasticity assumptions in the latent variable formulation more plausible.

b. We can just use ordinary least squares. OLS will be consistent for  $\beta$  (and even conditionally unbiased). Our inference should be made robust to heteroskedasticity because the restriction  $r \geq -\mathbf{x}\beta$  needs to hold, meaning  $r$  cannot be independent of  $\mathbf{x}$  (unless we restrict the range of  $r$  or  $\mathbf{x}$  somewhat arbitrarily).

c. Exponentiate and subtract one to get

$$y = \exp(\mathbf{x}\beta + r) - 1 = y = \exp(\mathbf{x}\beta) \exp(r) - 1$$

Now take the expectation conditional on  $\mathbf{x}$ :

$$E(y|\mathbf{x}) = \exp(\mathbf{x}\beta)E[\exp(r)|\mathbf{x}] - 1.$$

If we assume  $r$  is independent of  $\mathbf{x}$  then  $E[\exp(r)|\mathbf{x}] = E[\exp(r)] = \eta$ , and so

$$E(y|\mathbf{x}) = \eta \exp(\mathbf{x}\beta) - 1.$$

d. Because  $\eta = E[\exp(r)]$ , an unbiased and consistent estimator of  $\eta$  would be

$$N^{-1} \sum_{i=1}^N \exp(r_i),$$

if we observed the random sample of errors,  $\{r_i : i = 1, 2, \dots, N\}$ . Instead, we follow Duan’s (1983) “smearing” approach (which is really just a method of moments approach) and replace

the errors with the OLS residuals,  $\hat{r}_i$ , from the regression  $\log(1 + y_i)$  on  $\mathbf{x}_i$ . Then a consistent estimator of  $\eta$  is

$$\hat{\eta} = N^{-1} \sum_{i=1}^N \exp(\hat{r}_i),$$

which is guaranteed to be greater than one by Jensen's inequality. Note  $\eta$  is also greater than unity by Jensen's:

$$\eta = E[\exp(r)] > \exp[E(r)] = \exp(0) = 1.$$

e. The estimated conditional mean function is simply

$$\hat{E}(y|\mathbf{x}) = \hat{\eta} \exp(\mathbf{x}\hat{\boldsymbol{\beta}}) - 1.$$

It is not guaranteed to be nonnegative because the estimates  $\hat{\boldsymbol{\beta}}$  and  $\hat{\eta}$  have not been chosen to ensure nonnegativity. It is possible that, for some vectors  $\mathbf{x}$ ,  $\hat{\eta} \exp(\mathbf{x}\hat{\boldsymbol{\beta}}) < 1$ .

f. The Stata output follows. The estimated  $\eta$  is  $\hat{\eta} = 17.18$ , which is much higher than unity. None of the fitted values are negative; they range from about .061 to 45,202. The largest prediction is almost 10 times above the largest observed hours in the data set, and the average of the fitted values, 3,166, is much too high: the average of actual hours is 740.6. Therefore, for predicting *hours*, using  $\log(1 + \text{hours})$  in a linear regression is not very appealing.

```
. gen lhourspl = log(1 + hours)
```

```
. reg lhourspl nwifeinc educ exper expersq age kidslt6 kidsge6, robust
```

Linear regression

```
Number of obs =      753
F(   7,   745) =    73.12
Prob > F       =   0.0000
R-squared      =   0.2950
Root MSE     =   2.9367
```

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
lhourspl						
nwifeinc	-.0228321	.0098273	-2.32	0.020	-.0421247	-.0035395

educ		.2271644	.0507032	4.48	0.000	.1276262	.3267027
exper		.2968677	.0407256	7.29	0.000	.2169171	.3768182
expersq		-.0043383	.0013579	-3.19	0.001	-.007004	-.0016726
age		-.122754	.0163732	-7.50	0.000	-.1548971	-.0906109
kidslt6		-1.991432	.2110337	-9.44	0.000	-2.405724	-1.577141
kidsge6		.0372724	.0917873	0.41	0.685	-.1429201	.2174649
_cons		4.833966	1.050092	4.60	0.000	2.772473	6.895458

```
. predict xbhat
(option xb assumed; fitted values)
```

```
. predict rhat, resid
```

```
. gen exprhat = exp(rhat)
```

```
. sum exprhat
```

Variable		Obs	Mean	Std. Dev.	Min	Max
exprhat		753	17.17622	69.2013	.0012194	1045.183

```
. gen hourshat = 17.17622*exp(xbhat) - 1
```

```
. sum hours hourshat
```

Variable		Obs	Mean	Std. Dev.	Min	Max
hours		753	740.5764	871.3142	0	4950
hourshat		753	3166.422	5164.107	.061139	45202.41

g. The  $R$ -squared is computed in the Stata output that follows. It is about .159, which is substantially below not only the Tobit  $R$ -squared, .275, but also the linear regression  $R$ -squared, .266. For this data set, using  $\log(1 + y)$  in a linear regression does not work well.

```
. corr hours hourshat
(obs=753)
```

		hours	hourshat
hours		1.0000	
hourshat		0.3984	1.0000

```
. di .3984^2
.15872256
```

h. Under the null of independence between  $r_i$  and  $\mathbf{x}_i$ , we should find no significant relationship between  $r^2$  and any function of  $\mathbf{x}$ . Yet the  $F$  (that is, modified Wald) statistic for heteroskedasticity has a  $p$ -value of zero to more than four decimal places. Clearly,  $r_i$  is not



independent of  $\mathbf{x}_i$ .

```
. gen rhatsq = rhat^2
. gen xbhatsq = xbhat^2
. reg rhatsq xbhat xbhatsq
```

Source	SS	df	MS	Number of obs = 753			
Model	5304.11081	2	2652.05541	F( 2, 750) = 37.49			
Residual	53062.278	750	70.749704	Prob > F = 0.0000			
Total	58366.3888	752	77.6148787	R-squared = 0.0909			
				Adj R-squared = 0.0885			
				Root MSE = 8.4113			

rhatsq	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
xbhat	3.840246	.4952455	7.75	0.000	2.868013	4.812478
xbhatsq	-.571263	.0665162	-8.59	0.000	-.7018431	-.4406829
_cons	4.286994	.9089565	4.72	0.000	2.502592	6.071395

**17.2.** a. No. The two-limit Tobit only makes sense if there is a corner at both endpoints.

With  $P(y = 0) = 0$  the two-limit model becomes a one-limit model at unity, which means the model does not imply a zero density for  $y < 0$ . The estimates will be identical to the Tobit model with an upper corner at unity.

b. Over the range  $(0, 1]$ ,  $w \equiv -\log(y)$  takes values in  $[0, \infty)$ , with

$P(w = 0) = P(y = 1) > 0$ . Assuming that  $y$  is continuous on  $(0, 1)$ ,  $w$  is continuous over  $(0, \infty)$ . So  $w$  is nonnegative, has a pile up at zero, and is continuously distributed over strictly positive values. A type I Tobit model makes logical sense.

c. It takes some work, but it is tractable. We can write  $y = \exp(-w)$  but we cannot just pass the expected value through the exponential function. One way to proceed is to write

$w = \max(0, \mathbf{x}\boldsymbol{\beta} + u)$  where  $u|\mathbf{x} \sim \text{Normal}(0, \sigma^2)$  so  $y = \exp[-\max(0, \mathbf{x}\boldsymbol{\beta} + u)]$ . Then, using  $\exp(0) = 1$  and splitting the integral over  $u < -\mathbf{x}\boldsymbol{\beta}$  and  $u \geq -\mathbf{x}\boldsymbol{\beta}$ ,

$$\begin{aligned}
E(y|\mathbf{x}) &= \int_{-\infty}^{\infty} \exp[-\max(0, \mathbf{x}\boldsymbol{\beta} + u)](1/\sigma)\phi(u/\sigma)du \\
&= \int_{-\infty}^{-\mathbf{x}\boldsymbol{\beta}} (1/\sigma)\phi(u/\sigma)du + \int_{-\mathbf{x}\boldsymbol{\beta}}^{\infty} \exp[-(\mathbf{x}\boldsymbol{\beta} + u)](1/\sigma)\phi(u/\sigma)du \\
&= \Phi(-\mathbf{x}\boldsymbol{\beta}/\sigma) + \exp(-\mathbf{x}\boldsymbol{\beta})[1 - \Phi(-\mathbf{x}\boldsymbol{\beta}/\sigma)] \int_{-\mathbf{x}\boldsymbol{\beta}}^{\infty} \exp(-u)(1/\sigma)\phi(u/\sigma)du \\
&= \Phi(-\mathbf{x}\boldsymbol{\beta}/\sigma) + \exp(-\mathbf{x}\boldsymbol{\beta})[1 - \Phi(-\mathbf{x}\boldsymbol{\beta}/\sigma)] \{\exp(-1)\Phi[(\mathbf{x}\boldsymbol{\beta} + 1)/\sigma]\} \\
&= \Phi(-\mathbf{x}\boldsymbol{\beta}/\sigma) + \exp(-\mathbf{x}\boldsymbol{\beta} - 1)\Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\Phi[(\mathbf{x}\boldsymbol{\beta} + 1)/\sigma]
\end{aligned}$$

Although it is not obvious, this conditional mean function is bounded between zero and one.

**17.3.** a. Because  $y = a_1$  if and only if  $y^* \leq a_1$  we have

$$\begin{aligned}
P(y = a_1|\mathbf{x}) &= P(y^* \leq a_1|\mathbf{x}) = P(\mathbf{x}\boldsymbol{\beta} + u \leq a_1|\mathbf{x}) \\
&= P[(u/\sigma) \leq (a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma|\mathbf{x}] \\
&= \Phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma].
\end{aligned}$$

Similarly,

$$\begin{aligned}
P(y = a_2|\mathbf{x}) &= P(y^* = a_2|\mathbf{x}) = P(\mathbf{x}\boldsymbol{\beta} + u \geq a_2|\mathbf{x}) \\
&= P[(u/\sigma) \geq (a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] = 1 - \Phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] \\
&= \Phi[-(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma].
\end{aligned}$$

Next, for  $a_1 < \eta < a_2$ ,  $P(y \leq \eta|\mathbf{x}) = P(y^* \leq \eta|\mathbf{x}) = \Phi[(\eta - \mathbf{x}\boldsymbol{\beta})/\sigma]$ . Taking the derivative of this cdf with respect to  $y$  gives the pdf of  $y$  conditional on  $\mathbf{x}$  for values  $\eta$  strictly between  $a_1$  and  $a_2$ :  $(1/\sigma)\phi[(\eta - \mathbf{x}\boldsymbol{\beta})/\sigma]$ .

b. Because  $y = y^*$  when  $a_1 < y^* < a_2$ ,  $E(y^*|\mathbf{x}, a_1 < y_i < a_2) = E(y^*|\mathbf{x}, a_1 < y^* < a_2)$ . But  $y^* = \mathbf{x}\boldsymbol{\beta} + u$  and  $a_1 < y^* < a_2$  if and only if  $a_1 - \mathbf{x}\boldsymbol{\beta} < u < a_2 - \mathbf{x}\boldsymbol{\beta}$ . Therefore, using the hint,

$$\begin{aligned}
E(y^*|\mathbf{x}, a_1 < y^* < a_2) &= \mathbf{x}\boldsymbol{\beta} + E(u|\mathbf{x}, a_1 - \mathbf{x}\boldsymbol{\beta} < u < a_2 - \mathbf{x}\boldsymbol{\beta}) \\
&= \mathbf{x}\boldsymbol{\beta} + \sigma E[(u/\sigma)|\mathbf{x}, (a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma < u/\sigma < (a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] \\
&= \mathbf{x}\boldsymbol{\beta} + \frac{\sigma\{\phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma] - \phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma]\}}{\{\Phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] - \Phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma]\}} \\
&= E(y|\mathbf{x}, a_1 < y < a_2).
\end{aligned}$$

Now, we can easily get  $E(y|\mathbf{x})$  by using the following:

$$\begin{aligned}
E(y|\mathbf{x}) &= a_1 P(y = a_1|\mathbf{x}) + E(y|\mathbf{x}, a_1 < y < a_2) \cdot P(a_1 < y < a_2|\mathbf{x}) + a_2 P(y_2 = a_2|\mathbf{x}) \\
&= a_1 \Phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma] \\
&\quad + E(y|\mathbf{x}, a_1 < y < a_2) \cdot \{\Phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] - \Phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma]\} \\
&\quad + a_2 \Phi[(\mathbf{x}\boldsymbol{\beta} - a_2)/\sigma] \\
&= a_1 \Phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma] + (\mathbf{x}\boldsymbol{\beta}) \cdot \{\Phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] - \Phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma]\} \\
&\quad + \sigma \{\phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma] - \phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma]\} \\
&\quad + a_2 \Phi[(\mathbf{x}\boldsymbol{\beta} - a_2)/\sigma].
\end{aligned}$$

c. From part b it is clear that  $E(y^*|\mathbf{x}, a_1 < y^* < a_2) \neq \mathbf{x}\boldsymbol{\beta}$ , and so it would be a fluke if OLS on the restricted sample consistently estimated  $\boldsymbol{\beta}$ . The linear regression of  $y_i$  on  $\mathbf{x}_i$  using only those  $y_i$  such that  $a_1 < y_i < a_2$  consistently estimates the linear projection of  $y^*$  on  $\mathbf{x}$  in the subpopulation for which  $a_1 < y^* < a_2$ . Generally, there is no reason to think that this will have any simple relationship to the parameter vector  $\boldsymbol{\beta}$ . [In some restrictive cases, the regression on the restricted subsample could consistently estimate  $\boldsymbol{\beta}$  up to a common scale coefficient.]

d. We get log-likelihood immediately from part a:

$$\begin{aligned}
\ell_i(\theta) &= 1[y_i = a_1] \log\{\Phi[(a_1 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\} \\
&\quad + 1[y_i = a_2] \log\{\Phi[(\mathbf{x}_i\boldsymbol{\beta} - a_2)/\sigma]\} \\
&\quad + 1[a_1 < y_i < a_2] \log\{(1/\sigma)\phi[(y_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\}.
\end{aligned}$$

Note how the indicator function selects out the appropriate density for each of the three possible cases: at the left endpoint, at the right endpoint, or strictly between the endpoints.

e. After obtaining the maximum likelihood estimates  $\hat{\boldsymbol{\beta}}$  and  $\hat{\sigma}^2$ , just plug these into the formulas in part b. The expressions can be evaluated at interesting values of  $\mathbf{x}$ .

f. We can show this by brute-force differentiation of the expression in part b for  $E(y|\mathbf{x})$ . As a shorthand, write

$$\phi_1 \equiv \phi[(a_1 - \mathbf{x}\beta)/\sigma], \phi_2 \equiv \phi[(a_2 - \mathbf{x}\beta)/\sigma] = \phi[(\mathbf{x}\beta - a_2)/\sigma],$$

$$\Phi_1 \equiv \Phi[(a_1 - \mathbf{x}\beta)/\sigma], \text{ and } \Phi_2 \equiv \Phi[(a_2 - \mathbf{x}\beta)/\sigma]$$

Then

$$\begin{aligned} \frac{\partial E(y|\mathbf{x})}{\partial x_j} = & -(a_1/\sigma)\phi_1\beta_j + (a_2/\sigma)\phi_2\beta_j \\ & + (\Phi_2 - \Phi_1)\beta_j + [(\mathbf{x}\beta/\sigma)(\phi_1 - \phi_2)]\beta_j \\ & + \{[(a_1 - \mathbf{x}\beta)/\sigma]\phi_1\}\beta_j - \{[(a_2 - \mathbf{x}\beta)/\sigma]\phi_2\}\beta_j \end{aligned}$$

where the first two parts are the derivatives of the first and third terms, respectively, in  $E(y|\mathbf{x})$ , and the last two lines are obtained from differentiating the second term in  $E(y|\mathbf{x})$ . Careful inspection shows that all terms cancel except  $(\Phi_2 - \Phi_1)\beta_j$ , which is the expression we wanted to be left with.

The scale factor,

$$\Phi\left(\frac{a_2 - \mathbf{x}\beta}{\sigma}\right) - \Phi\left(\frac{a_1 - \mathbf{x}\beta}{\sigma}\right)$$

is simply the probability that a standard normal random variable falls in the interval  $[(a_1 - \mathbf{x}\beta)/\sigma, (a_2 - \mathbf{x}\beta)/\sigma]$ , which is necessarily between zero and one.

g. The partial effects on  $E(y|\mathbf{x})$  are given in f. These are estimated as

$$\Phi\left(\frac{a_2 - \mathbf{x}\hat{\beta}}{\hat{\sigma}}\right) - \Phi\left(\frac{a_1 - \mathbf{x}\hat{\beta}}{\hat{\sigma}}\right)$$

where the estimates are the MLEs. We could evaluate these partial effects at, say,  $\bar{\mathbf{x}}$  to estimate the PEA (partial effect at the average). Or, we can estimate the scale factor for the APE of continuous  $x_j$  as

$$\hat{\rho} \equiv N^{-1} \sum_{i=1}^N \left[ \Phi\left(\frac{a_2 - \mathbf{x}_i\hat{\beta}}{\hat{\sigma}}\right) - \Phi\left(\frac{a_1 - \mathbf{x}_i\hat{\beta}}{\hat{\sigma}}\right) \right].$$

Particularly for the APE, the scaled Tobit coefficients can be compared with the OLS coefficients (the  $\hat{\gamma}_j$ ). Generally, we expect

$$\hat{\gamma}_j \approx \hat{\rho} \cdot \hat{\beta}_j,$$

where  $0 < \hat{\rho} < 1$ . Of course, this approximation need not be very good in a particular application often it is. It does not make sense to directly compare the magnitude of  $\hat{\beta}_j$  with that of  $\hat{\gamma}_j$ . By the way, note that  $\hat{\sigma}$  appears in the partial effects along with the  $\hat{\beta}_j$ .

**17.4.** The Stata output is below. The heteroskedasticity-robust standard error on *grant* is quite a bit bigger, but the robust *t* statistic is above four. (Interestingly, the heteroskedasticity-robust standard error for *union* is substantially smaller than the usual standard error.) The coefficient on *grant* implies that a firm receiving a job training grant in 1988 is estimated to provide about 27.2 more hours of job training per worker, holding firm size and union status fixed. This effect is very large considering the average hours of annual training over all 127 firms is about 16.

```
. use jtrain1
```

```
. des hrsemp grant
```

variable name	storage type	display format	value label	variable label
hrsemp	float	%9.0g		tothrs/totrain
grant	byte	%9.0g		= 1 if received grant

```
. reg hrsemp grant lemploy union if d88
```

Source	SS	df	MS	Number of obs =	127
Model	23232.2579	3	7744.08598	F( 3, 123) =	14.58
Residual	65346.8909	123	531.275536	Prob > F =	0.0000
Total	88579.1488	126	703.009118	R-squared =	0.2623
				Adj R-squared =	0.2443
				Root MSE =	23.049

hrsemp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
grant	27.17647	4.769283	5.70	0.000	17.73597	36.61698
lemploy	-5.511867	2.012923	-2.74	0.007	-9.496324	-1.527409

union	-8.924901	5.392118	-1.66	0.100	-19.59827	1.748465
_cons	30.76978	7.345811	4.19	0.000	16.2292	45.31037

```
. reg hrsemp grant lemploy union if d88, robust
```

Linear regression

Number of obs = 127  
F( 3, 123) = 7.40  
Prob > F = 0.0001  
R-squared = 0.2623  
Root MSE = 23.049

hrsemp	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
grant	27.17647	6.525922	4.16	0.000	14.25881	40.09414
lemploy	-5.511867	2.17685	-2.53	0.013	-9.820807	-1.202926
union	-8.924901	3.181306	-2.81	0.006	-15.2221	-2.627702
_cons	30.76978	8.558935	3.60	0.000	13.8279	47.71167

b. The Tobit results are below. Out of 127 firms in 1988, 38 provide no job training.

```
. count if hrsemp == 0 & d88
      38
```

```
. tobit hrsemp grant lemploy union if d88, ll(0)
```

Tobit regression

Number of obs = 127  
LR chi2(3) = 37.46  
Prob > chi2 = 0.0000  
Pseudo R2 = 0.0398

Log likelihood = -451.88026

hrsemp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
grant	36.34335	6.121823	5.94	0.000	24.22655	48.46016
lemploy	-4.928542	2.656817	-1.86	0.066	-10.18713	.330044
union	-12.63617	7.286913	-1.73	0.085	-27.05901	1.786677
_cons	20.32933	9.769517	2.08	0.040	.9927198	39.66594
/sigma	28.70726	2.229537			24.29438	33.12014

Obs. summary: 38 left-censored observations at hrsemp<=0  
89 uncensored observations  
0 right-censored observations

The language “left censored at hrsemp <= 0” is misleading for corner solution applications, but it does tell us that 38 of the 127 firms have  $hrsemp = 0$ . The estimate of  $\sigma$  is  $\hat{\sigma} = 28.71$ .

To get the effect of *grant* on  $E(hrsemp|grant, employ, union, hrsemp > 0)$ , we must compute the inverse Mills ratio with *grant* = 1 and *grant* = 0. We set

$employ = \overline{employ} = 60.87$  and  $union = 1$ . Below is the Stata session.

```
. gen xb1 = _b[_cons] + _b[grant] + _b[lemploy]*log(60.87) + _b[union]
. gen xb0 = _b[_cons] + _b[lemploy]*log(60.87) + _b[union]
. gen prob1 = normal(xb1/_b[/sigma])
. gen prob0 = normal(xb0/_b[/sigma])
. gen imr1 = normalden(xb1/_b[/sigma])/prob1
. gen imr0 = normalden(xb0/_b[/sigma])/prob0
. gen cm1 = xb1 + _b[/sigma]*imr1
. gen cm0 = xb0 + _b[/sigma]*imr0
. gen dcm = cm1 - cm0
. list dcm in 1
```

```

+-----+
|               |
|               | dcm |
|-----|
1. | 15.09413 |
|-----|
+-----+
```

```
. gen um1 = prob1*cm1
. gen um0 = prob0*cm0
. gen dum = um1 - um0
. list dum in 1
```

```

+-----+
|               |
|               | dum |
|-----|
1. | 20.81422 |
|-----|
+-----+
```

For firms already doing some job training, the grant is estimated to increase training by about 15.1 hours per employee. When we add in the effects of firms that go from no training to positive hours, the expected change is about 20.8 hours at  $union = 1$  and the average value of  $employ$  in the sample. This is somewhat less than the OLS estimate we obtained earlier, 27.2.

The estimated APE is on the unconditional mean is computed below as 26.2, which is pretty close to the OLS estimate of 27.2. Bootstrapping can be used to obtain a valid standard

error.

```
. predict xb, xb
(31 missing values generated)
```

```
. sum xb if d88
```

Variable	Obs	Mean	Std. Dev.	Min	Max
xb	146	9.095312	16.80602	-18.41981	48.7405

```
. replace xb = . if ~d88
(294 real changes made, 294 to missing)
```

```
. replace xb = . if hrsemp == . | lemploy == . | union == .
(19 real changes made, 19 to missing)
```

```
. sum xb
```

Variable	Obs	Mean	Std. Dev.	Min	Max
xb	127	9.265182	17.20874	-18.41981	48.7405

```
. gen xb0 = xb - _b[grant]*grant
(344 missing values generated)
```

```
. gen xb1 = xb0 + _b[grant]
(344 missing values generated)
```

```
. gen prob0 = normal(xb0/_b[/sigma])
(344 missing values generated)
```

```
. gen prob1 = normal(xb1/_b[/sigma])
(344 missing values generated)
```

```
. gen imr0 = normalden(xb0/_b[/sigma])/prob0
(344 missing values generated)
```

```
. gen imr1 = normalden(xb1/_b[/sigma])/prob1
(344 missing values generated)
```

```
. gen cm0 = xb0 + _b[/sigma]*imr0
(344 missing values generated)
```

```
. gen cm1 = xb1 + _b[/sigma]*imr1
(344 missing values generated)
```

```
. gen um0 = prob0*cm0
(344 missing values generated)
```

```
. gen um1 = prob1*cm1
(344 missing values generated)
```

```
. gen pe = um1 - um0
(344 missing values generated)
```

```
. sum pe
```



Variable	Obs	Mean	Std. Dev.	Min	Max
pe	127	26.23	3.272887	16.88082	30.56553

c. They are jointly significant at the 1.5% level, as the Stata “test” command shows.

```
. test lemploy union

( 1) [model]lemploy = 0
( 2) [model]union = 0

      F( 2, 124) =    4.34
      Prob > F =    0.0151
```

d. For the Tobit model, I use the square of the correlation between  $y_i = hrsemp_i$  and  $\hat{E}(y_i|x_i)$  as an  $R$ -squared that can be compared with the linear model  $R$ -squared. After the `tobit` command in Stata,  $\hat{E}(y_i|x_i)$  can be gotten using the `ystar` option for predicted values. (Unfortunately, Stata’s naming convention conflicts with the notation used in the text, as  $y^*$  is used to denote the underlying latent variable, not the actual outcome.)

```
. predict hrsemp if d88 & hrsemp != ., ystar(0,.)
(344 missing values generated)

. corr hrsemp hrsemp
(obs=127)

-----+-----
      |   hrsemp   hrsemp
-----+-----
      |   1.0000
hrsemp |   0.5206   1.0000
hrsemp |
-----+-----

. di (.5206)^2
.27102436
```

This  $R$ -squared is slightly above that for the linear model (.262), and so the Tobit does provide a better fit. And remember, the Tobit estimates are not chosen to maximize an  $R$ -squared, so the improvement in fit is effectively better.

**17.5.** a. The results from OLS estimation of the linear model are given below.

```
. use fringe

. reg hrbens exper age educ tenure married male white nrtheast nrthcen south
  union, robust

Linear regression                                Number of obs =    616
```

hrbens	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
exper	.0029862	.0042485	0.70	0.482	-.0053574	.0113298
age	-.0022495	.0041519	-0.54	0.588	-.0104034	.0059043
educ	.082204	.0085122	9.66	0.000	.065487	.0989211
tenure	.0281931	.0037053	7.61	0.000	.0209164	.0354699
married	.0899016	.0499158	1.80	0.072	-.0081281	.1879312
male	.251898	.0496953	5.07	0.000	.1543015	.3494946
white	.098923	.0721337	1.37	0.171	-.0427402	.2405862
nrtheast	-.0834306	.0723545	-1.15	0.249	-.2255277	.0586664
nrthcen	-.0492621	.0626967	-0.79	0.432	-.1723922	.073868
south	-.0284978	.0653108	-0.44	0.663	-.1567617	.0997662
union	.3768401	.0535136	7.04	0.000	.2717448	.4819354
_cons	-.6999244	.1803555	-3.88	0.000	-1.054125	-.3457242

```
. tobit hrbens exper age educ tenure married male white nrtheast nrthcen south
      union, ll(0)
```

hrbens	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
exper	.0040631	.0046627	0.87	0.384	-.0050939	.0132201
age	-.0025859	.0044362	-0.58	0.560	-.0112981	.0061263
educ	.0869168	.0088168	9.86	0.000	.0696015	.1042321
tenure	.0287099	.0037237	7.71	0.000	.021397	.0360227
married	.1027574	.0538339	1.91	0.057	-.0029666	.2084814
male	.2556765	.0551672	4.63	0.000	.1473341	.364019
white	.0994408	.078604	1.27	0.206	-.054929	.2538106
nrtheast	-.0778461	.0775035	-1.00	0.316	-.2300547	.0743625
nrthcen	-.0489422	.0713965	-0.69	0.493	-.1891572	.0912729
south	-.0246854	.0709243	-0.35	0.728	-.1639731	.1146022
union	.4033519	.0522697	7.72	0.000	.3006999	.506004
_cons	-.8137158	.1880725	-4.33	0.000	-1.18307	-.4443616
/sigma	.5551027	.0165773			.5225467	.5876588

370

6.7% of the sample, have *hrbens* = 0. As expected, the Tobit estimates are all slightly larger in magnitude; this reflects that the scale factor is always less than unity.

c. Here is what happens when *exper*<sup>2</sup> and *tenure*<sup>2</sup> are included:

```
. tobit hrbens exper age educ tenure married male white nrtheast nrthcen south
    union expersq tenuresq, ll(0)
```

```
Tobit regression                                Number of obs   =          616
                                                LR chi2(13)    =        315.95
                                                Prob > chi2    =         0.0000
Log likelihood = -503.62108                    Pseudo R2      =         0.2388
```

hrbens	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
exper	.0306652	.0085253	3.60	0.000	.0139224	.047408
age	-.0040294	.0043428	-0.93	0.354	-.0125583	.0044995
educ	.0802587	.0086957	9.23	0.000	.0631812	.0973362
tenure	.0581357	.0104947	5.54	0.000	.037525	.0787463
married	.0714831	.0528969	1.35	0.177	-.0324014	.1753675
male	.2562597	.0539178	4.75	0.000	.1503703	.3621491
white	.0906783	.0768576	1.18	0.239	-.0602628	.2416193
nrtheast	-.0480194	.0760238	-0.63	0.528	-.197323	.1012841
nrthcen	-.033717	.0698213	-0.48	0.629	-.1708394	.1034053
south	-.017479	.0693418	-0.25	0.801	-.1536597	.1187017
union	.3874497	.051105	7.58	0.000	.2870843	.4878151
expersq	-.0005524	.0001487	-3.71	0.000	-.0008445	-.0002604
tenuresq	-.0013291	.0004098	-3.24	0.001	-.002134	-.0005242
_cons	-.9436572	.1853532	-5.09	0.000	-1.307673	-.5796409
/sigma	.5418171	.0161572			.5100859	.5735484

```
Obs. summary:      41  left-censored observations at hrbens<=0
                   575  uncensored observations
                   0  right-censored observations
```

```
. test expersq tenuresq
```

```
( 1)  [model]expersq = 0
( 2)  [model]tenuresq = 0
```

```
F( 2, 603) = 16.34
Prob > F = 0.0000
```

Both squared terms are very statistically significant as well as jointly significant. What is not clear is whether their presence would change the estimated partial effects in important ways.

d. There are nine industries, and we use *ind1* as the base industry:

```
. tobit hrbens exper age educ tenure married male white nrtheast nrthcen south
    union expersq tenuresq ind2-ind9, ll(0)
```

```
Tobit regression                                Number of obs   =          616
                                                LR chi2(21)      =          388.99
                                                Prob > chi2       =          0.0000
Log likelihood = -467.09766                    Pseudo R2        =          0.2940
```

hrbens	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
exper	.0267869	.0081297	3.29	0.001	.0108205	.0427534
age	-.0034182	.0041306	-0.83	0.408	-.0115306	.0046942
educ	.0789402	.0088598	8.91	0.000	.06154	.0963403
tenure	.053115	.0099413	5.34	0.000	.0335907	.0726393
married	.0547462	.0501776	1.09	0.276	-.0438005	.1532928
male	.2411059	.0556864	4.33	0.000	.1317401	.3504717
white	.1188029	.0735678	1.61	0.107	-.0256812	.2632871
nrtheast	-.1016799	.0721422	-1.41	0.159	-.2433643	.0400045
nrthcen	-.0724782	.0667174	-1.09	0.278	-.2035085	.0585521
south	-.0379854	.0655859	-0.58	0.563	-.1667934	.0908226
union	.3143174	.0506381	6.21	0.000	.2148662	.4137686
expersq	-.0004405	.0001417	-3.11	0.002	-.0007188	-.0001623
tenuresq	-.0013026	.0003863	-3.37	0.001	-.0020613	-.000544
ind2	-.3731778	.3742017	-1.00	0.319	-1.108095	.3617389
ind3	-.0963657	.368639	-0.26	0.794	-.8203575	.6276261
ind4	-.2351539	.3716415	-0.63	0.527	-.9650425	.4947348
ind5	.0209362	.373072	0.06	0.955	-.7117618	.7536342
ind6	-.5083107	.3682535	-1.38	0.168	-1.231545	.214924
ind7	.0033643	.3739442	0.01	0.993	-.7310468	.7377754
ind8	-.6107854	.376006	-1.62	0.105	-1.349246	.127675
ind9	-.3257878	.3669437	-0.89	0.375	-1.04645	.3948746
_cons	-.5750527	.4137824	-1.39	0.165	-1.387704	.2375989
/sigma	.5099298	.0151907			.4800959	.5397637

```
Obs. summary:      41  left-censored observations at hrbens<=0
                   575  uncensored observations
                   0   right-censored observations
```

```
. testparm ind2-ind9
```

```
( 1)  [model]ind2 = 0
( 2)  [model]ind3 = 0
( 3)  [model]ind4 = 0
( 4)  [model]ind5 = 0
( 5)  [model]ind6 = 0
( 6)  [model]ind7 = 0
( 7)  [model]ind8 = 0
( 8)  [model]ind9 = 0
```

```
F(   8,   595) =    9.66
Prob > F =    0.0000
```

Each industry dummy variable is individually insignificant at even the 10% level, but the joint Wald test says that they are jointly very significant. This is somewhat unusual for dummy

variables that are necessarily orthogonal (so that there is not a multicollinearity problem among them). The likelihood ratio statistic is  $LR = 2(503.621 - 467.098) = 73.046$ , which is roughly comparable with  $Q \cdot F = 8 \cdot 9.66 = 77.28$ . The  $p$ -values in both cases are essentially zero.

Several estimates on the industry dummies are economically significant, with a worker in, say, industry eight earning about 61 cents less per hour in benefits than a comparable worker in industry one. [In this example, with so few observations at zero, it is roughly legitimate to use the parameter estimates as the partial effects.]

**17.6.** a. First, we can write  $u_1 = \rho_1 v_2 + e_1$ , where  $\rho_1 = \text{Cov}(v_2, u_1)$ , and we use the fact that  $\text{Var}(v_2) = 1$ . Also,  $\sigma_1^2 = \rho_1^2 + \tau_1^2$  where  $\tau_1^2 = \text{Var}(e_1)$ . The distribution of  $y_1$  given  $(\mathbf{z}, v_2)$  can be written as  $g(\eta_1 | \mathbf{z}_1 \delta_1 + \alpha_1 y_2 + \rho_1 v_2, \sigma_1^2 - \rho_1^2)$ , where  $\eta_1$  is the generic argument. Next, we need the density of  $v_2$  given  $(\mathbf{z}, y_2)$ , which is given in equations (15.55) and (15.56). To obtain the density of  $y_1$  given  $(\mathbf{z}, y_2)$ , we can apply Property CD.3 in Appendix 2A. The density of  $v_2 | (\mathbf{z}, y_2 = 1)$  is  $\phi(v_2) / \Phi(\mathbf{z} \delta_2)$  for  $v_2 > -\mathbf{z} \delta_2$ . So the density of  $y_1$  given  $(\mathbf{z}, y_2 = 1)$  is

$$\frac{1}{\Phi(\mathbf{z} \delta_2)} \int_{-\mathbf{z} \delta_2}^{\infty} g(\eta_1 | \mathbf{z}_1 \delta_1 + \alpha_1 y_2 + \rho_1 v_2, \sigma_1^2 - \rho_1^2) \phi(v_2) dv_2$$

(where  $v_2$  is just the dummy argument in the integration) and the density given  $(\mathbf{z}, y_2 = 0)$  is

$$\frac{1}{[1 - \Phi(\mathbf{z} \delta_2)]} \int_{-\infty}^{-\mathbf{z} \delta_2} g(\eta_1 | \mathbf{z}_1 \delta_1 + \alpha_1 y_2 + \rho_1 v_2, \sigma_1^2 - \rho_1^2) \phi(v_2) dv_2.$$

b. We need to combine the density obtained from part a – called it  $f(\eta_1 | y_2, \mathbf{z}; \delta_1, \alpha_1, \rho_1, \sigma_1^2, \delta_2)$ , and let  $h(\eta_2 | \mathbf{z}; \delta_2)$  be the probit density of  $y_2$  given  $\mathbf{z}$ . Actually, it is easier to work with  $\tau_1^2 = \sigma_1^2 - \rho_1^2$ . Then the log-likelihood for observation  $i$  is

$$\begin{aligned}
& \log[f(y_{i1}|y_{i2}, \mathbf{z}_i; \boldsymbol{\delta}_1, \alpha_1, \rho_1, \tau_1^2, \boldsymbol{\delta}_2)] + \log[h(y_{i2}|\mathbf{z}_i; \boldsymbol{\delta}_2)] \\
&= y_{i2} \log\left(\frac{1}{\Phi(\mathbf{z}_i \boldsymbol{\delta}_2)} \int_{-\mathbf{z}_i \boldsymbol{\delta}_2}^{\infty} g(y_{i1}|\mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2} + \rho_1 v_2, \tau_1^2) \phi(v_2) dv_2\right) \\
&\quad + (1 - y_{i2}) \log\left(\frac{1}{[1 - \Phi(\mathbf{z}_i \boldsymbol{\delta}_2)]} \int_{-\infty}^{-\mathbf{z}_i \boldsymbol{\delta}_2} g(y_{i1}|\mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2} + \rho_1 v_2, \tau_1^2) \phi(v_2) dv_2\right) \\
&\quad + y_{i2} \log[\Phi(\mathbf{z}_i \boldsymbol{\delta}_2)] + (1 - y_{i2}) \log[1 - \Phi(\mathbf{z}_i \boldsymbol{\delta}_2)]
\end{aligned}$$

which simplifies to

$$\begin{aligned}
\ell_i(\boldsymbol{\theta}) &= y_{i2} \log\left(\int_{-\mathbf{z}_i \boldsymbol{\delta}_2}^{\infty} g(y_{i1}|\mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2} + \rho_1 v_2, \tau_1^2) \phi(v_2) dv_2\right) \\
&\quad + (1 - y_{i2}) \log\left(\int_{-\infty}^{-\mathbf{z}_i \boldsymbol{\delta}_2} g(y_{i1}|\mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2} + \rho_1 v_2, \tau_1^2) \phi(v_2) dv_2\right).
\end{aligned}$$

If  $\rho_1 = 0$ , the log likelihood becomes

$$\begin{aligned}
\ell_i(\boldsymbol{\theta}) &= y_{i2} \log[g(y_{i1}|\mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2}, \tau_1^2) \Phi(\mathbf{z}_i \boldsymbol{\delta}_2)] \\
&\quad + (1 - y_{i2}) \log\{g(y_{i1}|\mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2}, \tau_1^2) [1 - \Phi(\mathbf{z}_i \boldsymbol{\delta}_2)]\} \\
&= \log[g(y_{i1}|\mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2}, \tau_1^2)] + y_{i2} \log[\Phi(\mathbf{z}_i \boldsymbol{\delta}_2)] + (1 - y_{i2}) \log[1 - \Phi(\mathbf{z}_i \boldsymbol{\delta}_2)],
\end{aligned}$$

which is two separate log-likelihoods, one the standard Tobit for  $y_{i1}$  given  $(\mathbf{z}_{i1}, y_{i2})$  and the second for probit of  $y_{i2}$  given  $\mathbf{z}_i$ .

c. As in the probit case (Section 15.7.3), this is another example of a forbidden regression.

There is no way that  $E(y_1|\mathbf{z})$  has the Tobit form with  $\mathbf{z}_1$  and  $\Phi(\mathbf{z} \boldsymbol{\delta}_2) = E(y_2|\mathbf{z})$  as the explanatory variables. In fact, because  $y_1 = \max(0, \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1)$ ,  $E(y_1|\mathbf{z})$  has no simple form – although it could be computed in principle.

d. As given in the hint, it is easiest to work with the parameterization in terms of  $\tau_1^2$ , as shown in part b. Passing the derivative through the integral gives

$$\begin{aligned} \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \rho_1} = & y_{i2} \frac{\int_{-\mathbf{z}_i \boldsymbol{\delta}_2}^{\infty} v_2 g^{(1)}(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2} + \rho_1 v_2, \tau_1^2) \phi(v_2) dv_2}{\int_{-\mathbf{z}_i \boldsymbol{\delta}_2}^{\infty} g(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2} + \rho_1 v_2, \tau_1^2) \phi(v_2) dv_2} \\ & + (1 - y_{i2}) \frac{\int_{-\infty}^{-\mathbf{z}_i \boldsymbol{\delta}_2} v_2 g^{(1)}(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2} + \rho_1 v_2, \tau_1^2) \phi(v_2) dv_2}{\int_{-\infty}^{-\mathbf{z}_i \boldsymbol{\delta}_2} g(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2} + \rho_1 v_2, \tau_1^2) \phi(v_2) dv_2}. \end{aligned}$$

where  $g^{(1)}$  denotes the first derivative. When we set  $\rho_1 = 0$  the first term becomes

$$\begin{aligned} & y_{i2} \frac{g^{(1)}(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2}, \tau_1^2) \int_{-\mathbf{z}_i \boldsymbol{\delta}_2}^{\infty} v_2 \phi(v_2) dv_2}{g(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2}, \tau_1^2) \int_{-\mathbf{z}_i \boldsymbol{\delta}_2}^{\infty} \phi(v_2) dv_2} \\ = & y_{i2} \frac{g^{(1)}(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2}, \tau_1^2)}{g(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2}, \tau_1^2)} \cdot \frac{\phi(\mathbf{z}_i \boldsymbol{\delta}_2)}{[1 - \Phi(-\mathbf{z}_i \boldsymbol{\delta}_2)]} \\ = & \frac{g^{(1)}(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2}, \tau_1^2)}{g(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2}, \tau_1^2)} \cdot y_{i2} \lambda(\mathbf{z}_i \boldsymbol{\delta}_2) \end{aligned}$$

where  $\lambda(a) = \phi(a)/\Phi(a)$  is the inverse Mills ratio and we use the fact that  $\int_a^{\infty} v \phi(v) dv = \phi(a)$

for any  $a \in \mathbb{R}$ . Similarly, using  $\int_{-\infty}^a v \phi(v) dv = -\phi(a)$ , the second term is

$$\frac{g^{(1)}(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2}, \tau_1^2)}{g(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2}, \tau_1^2)} \cdot [-(1 - y_{i2}) \lambda(-\mathbf{z}_i \boldsymbol{\delta}_2)]$$

and so the partial derivative evaluated at  $\rho_1 = 0$  is

$$\begin{aligned} & \frac{g^{(1)}(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2}, \tau_1^2)}{g(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2}, \tau_1^2)} [y_{i2} \lambda(\mathbf{z}_i \boldsymbol{\delta}_2) - (1 - y_{i2}) \lambda(-\mathbf{z}_i \boldsymbol{\delta}_2)] \\ \equiv & \frac{g^{(1)}(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2}, \tau_1^2)}{g(y_{i1} | \mathbf{z}_{i1} \boldsymbol{\delta}_1 + \alpha_1 y_{i2}, \tau_1^2)} gr(y_{i2}, \mathbf{z}_i \boldsymbol{\delta}_2) \end{aligned}$$

where  $gr(y_{i2}, \mathbf{z}_i \boldsymbol{\delta}_2) \equiv y_{i2} \lambda(\mathbf{z}_i \boldsymbol{\delta}_2) - (1 - y_{i2}) \lambda(-\mathbf{z}_i \boldsymbol{\delta}_2)$  is the generalized residual. The key is that this is the same partial derivative we would obtain by simply adding  $gr_{i2} \equiv gr(y_{i2}, \mathbf{z}_i \boldsymbol{\delta}_2)$  as an explanatory variable and giving it a coefficient, say  $\eta_1$ . In other words, form the artificial model

$$y_{i1} = \max(0, \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + \eta_1 gr_{i2} + e_{i1})$$

$$e_{i1} | \mathbf{z}_{i1}, y_{i2}, gr_{i2} \sim \text{Normal}(0, \tau_1^2).$$

Of course, under the give assumptions this “model” cannot be true when  $\eta_1 \neq 0$ . But if we act as if it is true and compute the score for testing  $H_0 : \eta_1 = 0$ , we get exactly the score derived above. So we are led to a simple variable addition test. In the first stage estimate probit of  $y_{i2}$  on  $\mathbf{z}_i$  to get  $\hat{\boldsymbol{\delta}}_2$ . Construct the generalized residuals,

$$\widehat{gr}_{i2} = y_{i2}\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) - (1 - y_{i2})\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2).$$

In the second step, estimate a Tobit model of  $y_{i1}$  on  $\mathbf{z}_{i1}, y_{i2}, \widehat{gr}_{i2}$  and use a  $t$  test for the coefficient  $\hat{\eta}_1$  on  $\widehat{gr}_{i2}$ . Under the null, the statistic has an asymptotic  $\text{Normal}(0, 1)$  distribution, with no need to adjust the standard error for estimation of  $\boldsymbol{\delta}_2$ .

Incidentally, while adding the generalized residual – which acts as a kind of control function – does not generally solve the endogeneity of  $y_2$  under the assumptions of this problem, it might be a decent approximation. It is likely to do well when  $\rho_1$  is “close” to zero (although we then must wonder how much of a problem endogeneity is in the first case). There is some evidence that it can work well as an approximation more general, where focus would be on average partial effects. Putting in flexible functions of  $\widehat{gr}_{i2}$  – such as low-order polynomials – can help even more.

If we simply assert that  $D(y_1 | \mathbf{z}_1, y_2, gr_2)$  follows the Tobit model given above then adding  $\widehat{gr}_{i2}$  does produce consistent estimators of all parameters and average partial effects (by averaging out  $\widehat{gr}_{i2}$  in the partial effect formulas for the standard Tobit). This idea is nontraditional but is in the spirit of viewing all models simply as approximations.

**17.7.** Let  $s = 1[y > 0]$  and use Property CV.3 about conditional variances (see Appendix



2.A.2):

$$\text{Var}(y|\mathbf{x}) = E[\text{Var}(y|\mathbf{x}, s)|\mathbf{x}] + \text{Var}[E(y|\mathbf{x}, s)|\mathbf{x}]$$

Now because  $y = s \cdot w^*$ ,

$$\begin{aligned} E(y|\mathbf{x}, s) &= s \cdot E(w^*|\mathbf{x}, s) = s \cdot E(w^*|\mathbf{x}) = s \cdot \exp(\mathbf{x}\boldsymbol{\beta}) \\ \text{Var}(y|\mathbf{x}, s) &= s^2 \cdot \text{Var}(w^*|\mathbf{x}, s) = s \cdot \text{Var}(w^*|\mathbf{x}) = s \cdot \eta^2 [\exp(\mathbf{x}\boldsymbol{\beta})]^2 \end{aligned}$$

and so

$$\begin{aligned} \text{Var}(y|\mathbf{x}) &= E[s \cdot \eta^2 \exp(2\mathbf{x}\boldsymbol{\beta})|\mathbf{x}] + \text{Var}[s \cdot \exp(\mathbf{x}\boldsymbol{\beta})|\mathbf{x}] \\ &= P(s = 1|\mathbf{x})\eta^2 \exp(2\mathbf{x}\boldsymbol{\beta}) + \text{Var}(s|\mathbf{x}) \exp(2\mathbf{x}\boldsymbol{\beta}) \\ &= \eta^2 \Phi(\mathbf{x}\boldsymbol{\gamma}) \exp(2\mathbf{x}\boldsymbol{\beta}) + \Phi(\mathbf{x}\boldsymbol{\gamma})[1 - \Phi(\mathbf{x}\boldsymbol{\gamma})] \exp(2\mathbf{x}\boldsymbol{\beta}) \end{aligned}$$

**17.8.** a. For model (1) simply use ordinary least squares. Under the conditional mean assumption, we could use a weighted least squares procedure if we suspect heteroskedasticity, as we might, and have a particular form in mind. However, we should probably not think of the linear model as a model of  $E(y|\mathbf{x})$ ; rather, it is simply the linear projection. If we use a WLS procedure, we are effectively estimating a linear predictor in weighted variables.

For model (2) we could use nonlinear regression, or weighted nonlinear regression. The latter is attractive because of probable heteroskedasticity in  $\text{Var}(y|\mathbf{x})$ . We might use a variance function proportional to  $\exp(\mathbf{x}\boldsymbol{\beta})$  or  $[\exp(\mathbf{x}\boldsymbol{\beta})]^2$ , or a quadratic in the mean function:

$\text{Var}(y|\mathbf{x}) = \delta_0 + \delta_1 \exp(\mathbf{x}\boldsymbol{\beta}) + \delta_2 [\exp(\mathbf{x}\boldsymbol{\beta})]^2$  which contains the previous two as a special case.

We can estimate the  $\delta_j$  from the OLS regression  $\hat{u}_i^2$  on  $1, \hat{y}_i$ , and  $\hat{y}_i^2$ , where the hatted quantities are from a first stage NLS estimation. The fitted values are the estimated conditional variances (and we might have to worry about whether they are all strictly positive). Other attractive options are Poisson regression – see Chapter 18 for a description of its robustness properties for estimating  $E(y|\mathbf{x})$  – or regression using the Exponential quasi-log-likelihood (see Chapter

18).

Naturally, for model (3) we would use MLE.

b. We can compute an  $R$ -squared type measure any time we directly model  $E(y|\mathbf{x})$  or we have an implied model for  $E(y|\mathbf{x})$  (such as in the Tobit case). In each case, we obtain the fitted values,  $\hat{E}(y_i|\mathbf{x}_i), i = 1, \dots, N$ . Once we have fitted values we can obtain the squared correlation between  $y_i$  and  $\hat{E}(y_i|\mathbf{x}_i)$ . These can be compared across different models and even estimation methods. Alternatively, one can use a sum-of-squared residuals form:

$$R^2 = 1 - \frac{\sum_{i=1}^N [y_i - \hat{E}(y_i|\mathbf{x}_i)]^2}{\sum_{i=1}^N (y_i - \bar{y})^2}.$$

In the linear regression case with an intercept, the two ways of computing  $R$ -squared are identical, but the equivalence does not hold in general. In fact, the SSR version of  $R$ -squared can be negative in some cases. One can always compute an “adjusted”  $R$ -squared, too:

$$\bar{R}^2 = 1 - \frac{(N - P)^{-1} \sum_{i=1}^N [y_i - \hat{E}(y_i|\mathbf{x}_i)]^2}{(N - 1)^{-1} \sum_{i=1}^N (y_i - \bar{y})^2}$$

where  $P$  is the number of estimated parameters in the mean function.

c. This is clear from equation (17.20). If  $y_i > 0$  for  $i = 1, \dots, N$ , then only the second term in the log likelihood appears. But that is just the log likelihood for the classical linear regression model where  $y_i|\mathbf{x}_i \sim \text{Normal}(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$ . It is well known that the MLE of  $\boldsymbol{\beta}$  in this case is the OLS estimator.

It may seem a bit odd, but if we truly believe the population follows a Tobit model – and just happen to obtain a sample where  $y_i > 0$  for all  $i$  – then the appropriate estimate of  $E(y|\mathbf{x})$  is gotten from (17.14), where we plug in the usual OLS estimators for  $\boldsymbol{\beta}$  and  $\sigma^2$ . Estimates of  $E(y|\mathbf{x})$  computed in this way would ensure that fitted values in the sample are all positive, even

though  $\mathbf{x}_i \hat{\boldsymbol{\beta}}$  could be negative for some  $i$ .

d. If  $y > 0$  in the population, a Tobit model makes no sense because  $P(y = 0) > 0$  for a Tobit model. Instead, we could assume  $E[\log(y)|\mathbf{x}] = \mathbf{x}\boldsymbol{\gamma}$ , or, equivalently,  $\log(y) = \mathbf{x}\boldsymbol{\gamma} + v$ ,  $E(v|\mathbf{x}) = 0$ . If we make the stronger assumption that  $v$  is independent of  $\mathbf{x}$ , then  $E(y|\mathbf{x}) = \eta \cdot \exp(\mathbf{x}\boldsymbol{\gamma})$ , where  $\eta \equiv E[\exp(v)] > 1$ . After estimating  $\boldsymbol{\gamma}$  from the OLS regression  $\log(y_i)$  on  $\mathbf{x}_i, i = 1, \dots, N$ , we can estimate  $\eta$  using Duan's (1983) estimator, as in Problem 17.1:

$$\hat{\eta} = N^{-1} \sum_{i=1}^N \exp(\hat{v}_i),$$

where the  $\hat{v}_i$  are the OLS residuals.

**17.9.** a. A two-limit Tobit model, of the kind analyzed in Problem 17.3, is appropriate, with  $a_1 = 0, a_2 = 10$ .

b. The lower limit at zero is logically necessary considering the kind of response: the smallest percentage of one's income that can be invested in a pension plan is zero. On the other hand, the upper limit of 10 is an arbitrary corner imposed by law. One can imagine that some people at the corner  $y = 10$  would choose  $y > 10$  if they could. So, we can think of an underlying variable, which would be the percentage invested in the absence of any restrictions. Then, there would be no upper bound required (since we would not have to worry about 100 percent of income being invested in a pension plan).

**17.10.** A more general version of this problem is done in Problem 17.3, part f: set  $a_1 = 0$  and let  $a_2 \rightarrow \infty$ .

**17.11.** No. OLS always consistently estimates the parameters of a linear projection provided the second moments of  $y$  and the  $x_j$  are finite and  $\text{Var}(\mathbf{x})$  has full rank  $K$  – regardless

of the nature of  $y$  or  $x$  (discrete, continuous, some mixture). The fact that we can always consistently estimate a linear projection by OLS is why linear regression analysis is always a reasonable step for discrete outcomes (provided there is no data censoring problem of the type we discuss in Chapter 19). As discussed in Chapters 15 and 17, the linear regression coefficients often are close to estimated average partial effects from more complicated models. See Problem 17.4 part b for an example.

**17.12.** a. 248 out of 660, or about 37.6%, have  $ecolbs_i = 0$ . The positive responses range from .333 to a high of 42, but with focal points at integer values, especially one pound and two pounds. Therefore, a Tobit model cannot literally be true, but it can still lead to good estimates of the conditional mean and partial effects.

b. The linear model results are given below:

```
. use apple
. gen lecoprc = log(ecoprc)
. gen lregprc = log(regprc)
. gen lfaminc = log(faminc)
. reg ecolbs lecoprc lregprc lfaminc educ hhsiz num5_17
```

Source	SS	df	MS	Number of obs = 660		
Model	155.149478	6	25.8582463	F( 6, 653) = 4.17		
Residual	4048.98735	653	6.20059318	Prob > F = 0.0004		
Total	4204.13682	659	6.3795703	R-squared = 0.0369		
				Adj R-squared = 0.0281		
				Root MSE = 2.4901		

ecolbs	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
lecoprc	-2.56959	.5865181	-4.38	0.000	-3.721279	-1.417901
lregprc	2.204184	.5903005	3.73	0.000	1.045068	3.3633
lfaminc	.203861	.155967	1.31	0.192	-.1023964	.5101184
educ	.0251628	.0457977	0.55	0.583	-.0647657	.1150913
hhsiz	.0015866	.0886932	0.02	0.986	-.1725717	.1757449
num5_17	.1111276	.1343768	0.83	0.409	-.1527351	.3749903
_cons	.7307278	.7610805	0.96	0.337	-.7637326	2.225188

The price coefficients are of the expected sign: there is a negative own price effect, and a positive price effect for the substitute good, regular apples. The coefficient on  $\log(ecoprc)$  implies that a 10% increase in  $ecoprc$  leads to a fall in estimated demand of about .26 lbs. At the mean value of  $ecolbs$ , about 1.47 lbs, this is an estimated own price elasticity of  $-2.57/1.47 = -1.75$ , which is very large in magnitude.

c. The test for heteroskedasticity is given below. The  $F$  statistic, which is asymptotically valid as a test for heteroskedasticity, gives a pretty large  $p$ -value, .362, so this test does not find much evidence of heteroskedasticity.

```
. predict ecolbsh
(option xb assumed; fitted values)

. gen ecolbshsq = ecolbsh^2

. predict uh, resid

. gen uhsq = uh^2

. reg uhsq ecolbsh ecolbshsq
```

Source	SS	df	MS	Number of obs = 660		
Model	8923.31842	2	4461.65921	F( 2, 657) = 1.02		
Residual	2880416.28	657	4384.19525	Prob > F = 0.3620		
Total	2889339.6	659	4384.43034	R-squared = 0.0031		
				Adj R-squared = 0.0001		
				Root MSE = 66.213		

uhsq	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
ecolbsh	32.61476	31.09945	1.05	0.295	-28.45153	93.68105
ecolbshsq	-8.9604	10.32346	-0.87	0.386	-29.23136	11.31056
_cons	-20.36486	21.92073	-0.93	0.353	-63.40798	22.67827

d. The fitted values were already gotten from part c. The summary statistics are

```
. sum ecolbs ecolbsh
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ecolbs	660	1.47399	2.525781	0	42
ecolbsh	660	1.47399	.485213	.2251952	2.598743

```
. count if ecolbs < 2.6
541
```

```
. di 541/660
.81969697
```

The smallest fitted value is .225, and so none are negative. The largest fitted value is only about 2.6, but about 82 percent of the observations have *ecolbs<sub>i</sub>* below 2.6. Generally, it is difficult to find models that will track such a wide range in actual outcomes. Further, one might suspect the largest value, 42, is a mistake or an outlier. (The estimates with this one observation dropped give a similar story, but the price coefficients shrink in magnitude.)

e. The Tobit results are given below. The signs are the same as for the linear model, with the price and income variables being more statistically significant for Tobit. We know that the coefficients need to be scaled down in order to obtain the partial effects. That the Tobit coefficient on  $\log(ecoprc)$  is about double the OLS estimate is not surprising, and we need to compute a scale factor. The scale factor for the APEs (of continuous explanatory variables) is about .547. If we multiply each Tobit coefficient by .547, we get fairly close to the OLS estimates.

```
. tobit ecolbs lecoprc lregprc lfaminc educ hhsize num5_17, ll(0)
```

Tobit regression	Number of obs	=	660
	LR chi2(6)	=	50.79
	Prob > chi2	=	0.0000
Log likelihood = -1265.7088	Pseudo R2	=	0.0197

ecolbs	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
lecoprc	-5.238074	.8748606	-5.99	0.000	-6.955949	-3.5202
lregprc	4.261536	.8890055	4.79	0.000	2.515887	6.007185
lfaminc	.4149175	.2363235	1.76	0.080	-.0491269	.8789619
educ	.1005481	.068439	1.47	0.142	-.0338386	.2349348
hhsize	.0330173	.1325415	0.25	0.803	-.2272409	.2932756
num5_17	.2260429	.1970926	1.15	0.252	-.1609678	.6130535
_cons	-1.917668	1.160126	-1.65	0.099	-4.195689	.3603525
/sigma	3.445719	.1268015			3.196732	3.694706

Obs. summary:      248 left-censored observations at ecolbs<=0  
                      412 uncensored observations  
                      0 right-censored observations

```
. predict xbh, xb
. gen prob = normal(xbh/_b[/sigma])
. sum prob
```

Variable	Obs	Mean	Std. Dev.	Min	Max
prob	660	.5472506	.1152633	.2610003	.8191264

f. This question is a bit ambiguous. I will evaluate the partial effect at the mean value of *ecoprc*, *regprc*, and *faminc*, and then take the log, rather than averaging the logs. The scale factor for the APEs is given in part e: .547. The scale factor for the partial effects at the mean is .539, which is fairly close. The PAE of *lecoprc* is about  $-2.82$ , which is somewhat bigger in magnitude than the OLS estimate,  $-2.57$ .

To get the estimated elasticity, we need to estimate  $E(ecolbs|x)$  at the mean values of the covariates; we get about 1.55. So the estimated elasticity at the mean values of the covariates is about  $-2.82/1.55 \approx -1.82$ . This is slightly larger in magnitude than that computed for the linear model,  $-1.75$ .

```
. sum ecoprc regprc faminc educ hhsize num5_17
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ecoprc	660	1.081515	.295573	.59	1.59
regprc	660	.8827273	.2444687	.59	1.19
faminc	660	53.40909	35.74122	5	250
educ	660	14.38182	2.274014	8	20
hhsize	660	2.940909	1.526049	1	9
num5_17	660	.6212121	.994143	0	6

```
. di normal((_b[_cons] + _b[lecoprc]*log( 1.081515 )
+ _b[lregprc]*log(.8827273) + _b[lfaminc]*log(53.40909)
+ _b[educ]*14.38182 + _b[hhsize]*2.940909)/_b[/sigma])
```

```
.53860761
```

```
. di .53860761*_b[lecoprc]
-2.8212668
```

```
. di _b[_cons] + _b[lecoprc]*log( 1.081515 )
+ _b[lregprc]*log(.8827273) + _b[lfaminc]*log(53.40909)
+ _b[educ]* 14.38182 + _b[hhsize]*2.940909
```

```
.33398136
```

```
. di normalden((_b[_cons] + _b[lecoprc]*log( 1.081515 )
+ _b[lregprc]* log(.8827273) + _b[lfaminc]*log(53.40909)
+ _b[educ]* 14.38182 + _b[hhsize]*2.940909)/_b[/sigma])

.3970727

. di .33398136*.53860761 + _b[/sigma]*.3970727
1.5480857

. di -2.82/1.55
-1.8193548
```

g. Dropping  $\log(\text{regprc})$  greatly reduces the magnitude of the coefficient on  $\log(\text{ecoprc})$ : from  $-5.24$  to  $-1.82$ . A standard omitted variable analysis in a linear context suggests a positive correlation between  $\text{lecoprc}$  and  $\text{lregprc}$ . In fact, they are very highly positively correlated, with a correlation of about .82. This high correlation was built in as part of the experimental design.

```
. tobit ecolbs lecoprc lfaminc educ hhsize num5_17, ll(0)
```

```
Tobit regression                                Number of obs   =          660
                                                LR chi2(5)      =          27.60
                                                Prob > chi2     =          0.0000
Log likelihood = -1277.3043                    Pseudo R2      =          0.0107
```

ecolbs	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
lecoprc	-1.822712	.5044411	-3.61	0.000	-2.813229	-.8321952
lfaminc	.3931692	.2395441	1.64	0.101	-.0771978	.8635362
educ	.1169085	.0692025	1.69	0.092	-.0189769	.252794
hhsize	.0222283	.1340901	0.17	0.868	-.2410699	.2855266
num5_17	.2474529	.1996317	1.24	0.216	-.1445424	.6394483
_cons	-2.873156	1.161745	-2.47	0.014	-5.154351	-.5919615
/sigma	3.499092	.1291121			3.245569	3.752616

```
Obs. summary:      248 left-censored observations at ecolbs<=0
                   412 uncensored observations
                   0 right-censored observations
```

```
. corr lecoprc lregprc
(obs=660)
```

	lecoprc	lregprc
lecoprc	1.0000	
lregprc	0.8205	1.0000

h. In fact, the Tobit model with prices in level form, rather than logarithms, fits a bit better



(log-likelihood = -1,263.37 versus -1,265.71).

```
. tobit ecolbs ecoprc regprc lfaminc educ hhsize num5_17, ll(0)
```

Tobit regression	Number of obs	=	660
	LR chi2(6)	=	55.47
	Prob > chi2	=	0.0000
Log likelihood = -1263.3702	Pseudo R2	=	0.0215

ecolbs	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
ecoprc	-5.649516	.887358	-6.37	0.000	-7.391931	-3.907102
regprc	5.575299	1.063999	5.24	0.000	3.486032	7.664566
lfaminc	.4195658	.2354371	1.78	0.075	-.0427381	.8818696
educ	.1002944	.0681569	1.47	0.142	-.0335384	.2341271
hhsize	.0264861	.1320183	0.20	0.841	-.2327448	.2857171
num5_17	.2351291	.1963111	1.20	0.231	-.1503469	.6206051
_cons	-1.632596	1.314633	-1.24	0.215	-4.214007	.9488146
/sigma	3.431504	.1262031			3.183692	3.679316
Obs. summary:						
	248	left-censored observations at ecolbs<=0				
	412	uncensored observations				
	0	right-censored observations				

**17.13.** This extension has no practical effect on how we estimate an unobserved effects

Tobit or probit model, or how we estimate a variety of unobserved effects panel data models with conditional normal heterogeneity. We simply have

$$c_i = -\left(T^{-1} \sum_{t=1}^T \pi_t\right) \xi + \bar{\mathbf{x}}_i \xi + a_i \equiv \psi + \bar{\mathbf{x}}_i \xi + a_i,$$

where  $\psi \equiv -(T^{-1} \sum_{t=1}^T \pi_t \xi)$ . Of course, any aggregate time dummies explicitly get swept out of  $\bar{\mathbf{x}}_i$  but they would usually be included in the equation.

An interesting follow-up question is: What if we standardize each  $\mathbf{x}_{it}$  by its cross-sectional mean *and* variance at time  $t$ , and assume  $c_i$  is related to the mean and variance of the standardized vectors? In other words, let  $\mathbf{z}_{it} \equiv (\mathbf{x}_{it} - \pi_t) \Omega_t^{-1/2}$ ,  $t = 1, \dots, T$ , for each random draw  $i$  from the population, where  $\Omega_t \equiv \text{Var}(\mathbf{x}_{it})$ . Then, we might assume

$$c_i | \mathbf{x}_i \sim \text{Normal}(\psi + \bar{\mathbf{z}}_i \xi, \sigma_a^2)$$

(where, again,  $\mathbf{z}_{it}$  would not contain aggregate time dummies). This is the kind of scenario that is handled by Chamberlain's more general assumption concerning the relationship between  $c_i$  and  $\mathbf{x}_i$ :  $c_i = \psi + \sum_{t=1}^T \mathbf{x}_{it} \boldsymbol{\lambda}_t + a_i$ , where  $\boldsymbol{\lambda}_t = \boldsymbol{\Omega}_t^{-1/2} \boldsymbol{\xi}/T$ ,  $t = 1, 2, \dots, T$ . Alternatively, one could estimate  $\boldsymbol{\pi}_t$  and  $\boldsymbol{\Omega}_t$  for each  $t$  using the cross section observations  $\{\mathbf{x}_{it} : i = 1, 2, \dots, N\}$ . The usual sample means and sample variance matrices, say  $\hat{\boldsymbol{\pi}}_t$  and  $\hat{\boldsymbol{\Omega}}_t$ , are consistent and  $\sqrt{N}$ -asymptotically normal. Then, form  $\hat{\mathbf{z}}_{it} \equiv (\mathbf{x}_{it} - \hat{\boldsymbol{\pi}}_t) \hat{\boldsymbol{\Omega}}_t^{-1/2}$ , and proceed with the usual Tobit (or probit) unobserved effects analysis that includes the time averages  $\bar{\mathbf{z}}_i = T^{-1} \sum_{t=1}^T \hat{\mathbf{z}}_{it}$ . This is a simple two-step estimation method, but accounting for the sample variation in  $\hat{\boldsymbol{\pi}}_t$  and  $\hat{\boldsymbol{\Omega}}_t$  analytically would be cumbersome. The panel bootstrap is an attractive alternative. Or, it may be possible to use a much larger sample to obtain  $\hat{\boldsymbol{\pi}}_t$  and  $\hat{\boldsymbol{\Omega}}_t$ , in which case one might ignore the sampling error in the first-stage estimates.

**17.14.** a. Because heteroskedasticity is only in the distribution of  $a_i$  given  $\mathbf{x}_i$ , the density of  $y_{it}$  given  $(\mathbf{x}_i, a_i)$  is the same as that implied by (17.75) and (17.76), namely,

$$y_{it} | \mathbf{x}_i, a_i \sim \text{Tobit}(\psi + \mathbf{x}_{it} \boldsymbol{\beta} + \bar{\mathbf{x}}_i \boldsymbol{\xi} + a_i, \sigma_u^2).$$

b. Let  $f(y_{it} | \mathbf{x}_i, a_i; \boldsymbol{\eta})$  denote the Tobit density of  $y_{it} | \mathbf{x}_i, a_i$  implied by part a, where  $\boldsymbol{\eta}$  contains  $\boldsymbol{\beta}$ ,  $\psi$ ,  $\boldsymbol{\xi}$ , and  $\sigma_u^2$ . Then, under (17.78),

$$f(y_1, \dots, y_T | \mathbf{x}_i, a_i; \boldsymbol{\eta}) = \prod_{t=1}^T f_t(y_{it} | \mathbf{x}_i, a_i; \boldsymbol{\eta}).$$

Therefore, to obtain  $f(y_1, \dots, y_T | \mathbf{x}_i; \boldsymbol{\theta})$ , we integrate out  $a_i$ :

$$f(y_1, \dots, y_T | \mathbf{x}_i; \boldsymbol{\theta}) = \int_{-\infty}^{\infty} \prod_{t=1}^T f_t(y_{it} | \mathbf{x}_i, a_i; \boldsymbol{\eta}) h(a | \mathbf{x}_i; \boldsymbol{\lambda}, \sigma_a^2) da, \quad (17.91)$$

where  $h(a | \mathbf{x}_i; \boldsymbol{\lambda}, \sigma_a^2)$  denotes the normal density with mean zero and variance  $\sigma_a^2 \exp(\bar{\mathbf{x}}_i \boldsymbol{\lambda})$ . The

log-likelihood is obtained by plugging the  $y_{it}$  into (17.91) and taking the log.

c. The starting point is still equation (17.79), but the calculation of

$E[m(\psi + \mathbf{x}_i\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i, \sigma_u^2)|\mathbf{x}_i]$  is complicated by the heteroskedasticity in  $\text{Var}(a_i|\bar{\mathbf{x}}_i)$ .

Nevertheless, essentially the same argument used on page 542 shows that

$$E[m(\psi + \mathbf{x}_i\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i, \sigma_u^2)|\mathbf{x}_i] = m[\psi + \mathbf{x}_i\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi}, \sigma_a^2 \exp(\bar{\mathbf{x}}_i\boldsymbol{\lambda}) + \sigma_u^2].$$

Given the MLEs, we can estimate the APEs from the average structural function:

$$\widehat{\text{ASF}}(\mathbf{x}_t) = N^{-1} \sum_{i=1}^N m[\hat{\psi} + \mathbf{x}_t\hat{\boldsymbol{\beta}} + \bar{\mathbf{x}}_t\hat{\boldsymbol{\xi}}, \hat{\sigma}_a^2 \exp(\bar{\mathbf{x}}_t\hat{\boldsymbol{\lambda}}) + \hat{\sigma}_u^2];$$

we would compute changes or derivatives with respect to the elements of  $\mathbf{x}_t$ . Incidentally, if we

drop assumption (17.78), we could use a pooled heteroskedastic Tobit procedure, and still

consistently estimate the APEs.

**17.15.** a. The Stata output is given below. The value of the log-likelihood is  $-17,599.96$ .

```
. use cps91
```

```
. tobit hours nwifeinc educ exper expersq age kidlt6 kidge6, ll(0)
```

Tobit regression	Number of obs	=	5634
	LR chi2(7)	=	645.55
	Prob > chi2	=	0.0000
Log likelihood = -17599.958	Pseudo R2	=	0.0180

hours	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
nwifeinc	-.2444726	.0165886	-14.74	0.000	-.2769926	-.2119525
educ	-6.064707	22.73817	-0.27	0.790	-50.64029	38.51087
exper	-8.234015	22.74967	-0.36	0.717	-52.83214	36.36411
expersq	-.0178206	.0041379	-4.31	0.000	-.0259325	-.0097087
age	8.53901	22.73703	0.38	0.707	-36.03435	53.11237
kidlt6	-14.0809	1.21084	-11.63	0.000	-16.45461	-11.70719
kidge6	-1.593786	1.09917	-1.45	0.147	-3.748583	.5610116
_cons	-56.32579	136.3411	-0.41	0.680	-323.6069	210.9553
/sigma	28.90194	.3998526			28.11807	29.6858
Obs. summary:	2348	left-censored observations at hours<=0				
	3286	uncensored observations				
	0	right-censored observations				

b. The lognormal hurdle model – which has eight more parameters than the Tobit model – does fit better in this application. The log likelihood – which properly account for the fact that the linear regression for  $\log(hours_i)$  is to be viewed as MLE for  $D(hours|x, hours > 0)$  – is about  $-16,987.50$ . The contribution of the probit is about  $-3,538.41$  and the contribution of the lognormal distribution conditional on  $hours > 0$  is  $-13,449.09$ . The log likelihood for the Tobit is  $-17,599.96$ .

```
. probit inlf nwifeinc educ exper expersq age kidlt6 kidge6
```

Probit regression	Number of obs	=	5634
	LR chi2(7)	=	576.67
	Prob > chi2	=	0.0000
Log likelihood = -3538.4086	Pseudo R2	=	0.0753

inlf	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
nwifeinc	-.0091475	.0006759	-13.53	0.000	-.0104722	-.0078227
educ	-.0626136	.9045369	-0.07	0.945	-1.835473	1.710246
exper	-.157161	.9050879	-0.17	0.862	-1.931101	1.616779
expersq	-.0005574	.0001713	-3.25	0.001	-.000893	-.0002217
age	.1631286	.9044966	0.18	0.857	-1.609652	1.935909
kidlt6	-.4810832	.051688	-9.31	0.000	-.5823897	-.3797767
kidge6	.0409155	.0471194	0.87	0.385	-.0514367	.1332678
_cons	-1.489209	5.422855	-0.27	0.784	-12.11781	9.139393

```
. gen lhours = log(hours)
(2348 missing values generated)
```

```
. glm lhours nwifeinc educ exper expersq age kidlt6 kidge6
```

```
Iteration 0: log likelihood = -1954.9002
```

Generalized linear models	No. of obs	=	3286
Optimization : ML	Residual df	=	3278
	Scale parameter	=	.1928961
Deviance	= 632.3133243	(1/df) Deviance	= .1928961
Pearson	= 632.3133243	(1/df) Pearson	= .1928961

Variance function: $V(u) = 1$	[Gaussian]
Link function : $g(u) = u$	[Identity]

Log likelihood	= -1954.900228	AIC	= 1.194705
		BIC	= -25911.05

lhours	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval	
--------	-------	------------------	---	------	---------------------	--

nwifeinc	-.0018706	.0003327	-5.62	0.000	-.0025227	-.0012185
educ	-.2022625	.4403476	-0.46	0.646	-1.065328	.660803
exper	-.2074679	.4405366	-0.47	0.638	-1.070904	.655968
expersq	-.0001549	.0000812	-1.91	0.057	-.000314	4.33e-06
age	.2112264	.4403162	0.48	0.631	-.6517775	1.07423
kidlt6	-.1944414	.0222299	-8.75	0.000	-.2380113	-.1508715
kidge6	-.1256763	.0199962	-6.29	0.000	-.1648681	-.0864845
_cons	2.252439	2.640541	0.85	0.394	-2.922926	7.427805

```
. sum lhours
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lhours	3286	3.497927	.4467688	0	4.787492

```
. di 3286*r(mean)
11494.189
```

```
. di -1954.9002 - 11494.189
-13449.089
```

```
. di -3538.4086 - 13449.089
-16987.498
```

c. The ET2T model is given below. Again, to properly compare its log likelihood, we must subtract  $\sum_{i=1}^{N_1} \log(hours_i)$  to obtain the final log likelihood. As must be the case, the ET2T model fits better than the lognormal hurdle model, but the improvement is very slight. In fact, the estimate of  $\rho$  is very small – about .018 – and not statistically different from zero. The likelihood ratio statistic gives the same result, producing  $p$ -value = .862. Fortunately the estimated coefficients are very similar across the two approaches, as we would hope with  $\hat{\rho}$  so close to zero.

These findings are very different from what we found using the data in MROZ.RAW – see Table 17.2. There, the estimate of  $\rho$  is an implausible  $-.972$ . Without an exclusion restriction (in either application) it is hard to be confident of the results. But with the current data set, we are led to the lognormal hurdle model with all explanatory variables in the selection and amount equations.

```
. heckman lhours nwifeinc educ exper expersq age kidlt6 kidge6,
      select(inlf = nwifeinc educ exper expersq age kidlt6 kidge6)
```

```

Heckman selection model
(regression model with sample selection)
Number of obs      =      5634
Censored obs       =      2348
Uncensored obs     =      3286

Log likelihood = -5493.294
Wald chi2(7)      =      93.35
Prob > chi2       =      0.0000

```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
lhours						
nwifeinc	-.001911	.0003968	-4.82	0.000	-.0026886	-.0011333
educ	-.2023362	.4398277	-0.46	0.645	-1.064383	.6597103
exper	-.2079298	.4400233	-0.47	0.637	-1.07036	.6545001
expersq	-.0001578	.0000827	-1.91	0.056	-.0003198	4.21e-06
age	.2117355	.4398047	0.48	0.630	-.6502658	1.073737
kidlt6	-.1964925	.0247921	-7.93	0.000	-.245084	-.1479009
kidge6	-.1255154	.0199914	-6.28	0.000	-.1646978	-.086333
_cons	2.240756	2.63817	0.85	0.396	-2.929962	7.411474
inlf						
nwifeinc	-.0091462	.000676	-13.53	0.000	-.0104711	-.0078214
educ	-.0628333	.9045172	-0.07	0.945	-1.835654	1.709988
exper	-.1574369	.9050687	-0.17	0.862	-1.931339	1.616465
expersq	-.0005568	.0001713	-3.25	0.001	-.0008925	-.0002211
age	.1633826	.9044773	0.18	0.857	-1.60936	1.936125
kidlt6	-.4810173	.0516912	-9.31	0.000	-.5823302	-.3797043
kidge6	.0410785	.0471309	0.87	0.383	-.0512964	.1334534
_cons	-1.491111	5.422743	-0.27	0.783	-12.11949	9.13727
/athrho	.0178479	.0959507	0.19	0.852	-.1702121	.2059078
/lnsigma	-.8239347	.0123732	-66.59	0.000	-.8481857	-.7996837
rho	.017846	.0959202			-.1685871	.2030463
sigma	.4387021	.0054281			.4281911	.4494711
lambda	.0078291	.0420881			-.074662	.0903202
LR test of indep. eqns. (rho = 0):    chi2(1) =      0.03    Prob > chi2 = 0.8615						

```

. di -5493.294 - 11494.189
-16987.483

. di 2*(16987.498 - 16987.483)
.03

```

d. The estimates for the amount part of the truncated normal hurdle model are given below.

Because the participation equation is still the probit model we estimated earlier, we can compare the log likelihood for the truncated normal regression to that from the lognormal estimation in part b. The former is -12,445.76 and we already computed the latter as -13,449.09. Thus, in this example the TNH model fits substantially better than the LH model.

The full log likelihood for the TNH model is  $-15,984.17$  and this is much larger than that for the Tobit (a special case),  $-17,599.96$ .

```
. truncreg hours nwifeinc educ exper expersq age kidlt6 kidge6, ll(0)
(note: 2348 obs. truncated)
```

Fitting full model:

```
Limit:   lower =          0          Number of obs =    3286
         upper =        +inf        Wald chi2(7)  =   132.08
Log likelihood = -12445.76        Prob > chi2   =   0.0000
```

hours	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
nwifeinc	-.0439736	.0081584	-5.39	0.000	-.0599638	-.0279835
educ	-9.183178	11.12374	-0.83	0.409	-30.98531	12.61896
exper	-9.426741	11.12822	-0.85	0.397	-31.23765	12.38417
expersq	-.0024584	.0019888	-1.24	0.216	-.0063564	.0014396
age	9.470886	11.12299	0.85	0.395	-12.32978	31.27155
kidlt6	-4.779305	.5444546	-8.78	0.000	-5.846417	-3.712194
kidge6	-3.370223	.4896076	-6.88	0.000	-4.329837	-2.41061
_cons	-21.34309	66.70579	-0.32	0.749	-152.084	109.3979
/sigma	10.72244	.1347352	79.58	0.000	10.45836	10.98651

```
. di -3538.4086 - 12445.76
-15984.169
```

**17.16. a.** Write  $c_i = \psi + \bar{\mathbf{x}}_i \boldsymbol{\xi} + a_i$  and substitute to get

$$y_{it}^* = \mathbf{x}_{it} \boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i \boldsymbol{\xi} + a_i + u_{it}.$$

Now, conditional on  $(\mathbf{x}_i, a_i)$ ,  $y_{it}$  follows a standard two-limit Tobit model. Therefore, the density is

$$\begin{aligned} f(y_t | \mathbf{x}_i, a_i; \boldsymbol{\gamma}) &= [\Phi((q_1 - \mathbf{x}_{it} \boldsymbol{\beta} - \psi - \bar{\mathbf{x}}_i \boldsymbol{\xi} - a_i)/\sigma_u)]^{1[y_t=q_1]} \\ &\quad \cdot \{\sigma_u^{-1} \phi[(y_t - \mathbf{x}_{it} \boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i \boldsymbol{\xi} + a_i)/\sigma_u]\}^{1[q_1 < y_t < q_2]} \\ &\quad \cdot [\Phi((-q_2 + \mathbf{x}_{it} \boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i \boldsymbol{\xi} + a_i)/\sigma_u)]^{1[y_t=q_2]} \end{aligned}$$

Byt the conditional independence assumption, the joint density of  $(y_{i1}, y_{i2}, \dots, y_{iT})$  conditional on  $(\mathbf{x}_i, a_i)$  is

$$\prod_{t=1}^T f(y_t | \mathbf{x}_i, a_i; \boldsymbol{\gamma}).$$

Now we integrate out  $a_i$  to get the joint density of  $(y_{i1}, y_{i2}, \dots, y_{iT})$  given  $\mathbf{x}_i$ :

$$\int_{-\infty}^{\infty} \left[ \prod_{t=1}^T f(y_t | \mathbf{x}_i, a; \boldsymbol{\gamma}) \right] \sigma_a^{-1} \phi(a/\sigma_a) da.$$

Now the log likelihood for a random draw  $i$  is just

$$\ell_i(\boldsymbol{\theta}) \equiv \log \left\{ \int_{-\infty}^{\infty} \left[ \prod_{t=1}^T f(y_{it} | \mathbf{x}_i, a; \boldsymbol{\gamma}) \right] \sigma_a^{-1} \phi(a/\sigma_a) da \right\}$$

where  $\boldsymbol{\theta}$  is the vector of all parameters, including  $\sigma_a^2$ . As usual, we sum across all  $i$  to get the log likelihood for the entire cross section.

b. This is no different from any of the other CRE models that we have covered. We can easily find  $E(c_i)$  and  $\text{Var}(c_i)$  from  $c_i = \psi + \bar{\mathbf{x}}_i \boldsymbol{\xi} + a_i$  because  $E(a_i | \mathbf{x}_i) = 0$  and  $\text{Var}(a_i | \mathbf{x}_i) = \sigma_a^2$ . In fact,

$$E(c_i) = E(\psi + \bar{\mathbf{x}}_i \boldsymbol{\xi}) = \psi + E(\bar{\mathbf{x}}_i) \boldsymbol{\xi}$$

and so a consistent estimator of  $E(c_i)$  is

$$\hat{\mu}_c = \hat{\psi} + \left( N^{-1} \sum_{i=1}^N \bar{\mathbf{x}}_i \right) \hat{\boldsymbol{\xi}}$$

where  $\hat{\psi}$  and  $\hat{\boldsymbol{\xi}}$  are the MLEs.

Next,

$$\begin{aligned} \text{Var}(c_i) &= \text{Var}(\psi + \bar{\mathbf{x}}_i \boldsymbol{\xi}) + \text{Var}(a_i) \\ &= \boldsymbol{\xi}' \text{Var}(\bar{\mathbf{x}}_i) \boldsymbol{\xi} + \sigma_a^2, \end{aligned}$$

where we use the fact that  $a_i$  and  $\bar{\mathbf{x}}_i$  are uncorrelated. So a consistent estimator is



$$\hat{\sigma}_c^2 = \hat{\xi}' \left[ N^{-1} \sum_{i=1}^N (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}) \right] \hat{\xi} + \hat{\sigma}_a^2$$

where  $\bar{\mathbf{x}} = N^{-1} \sum_{i=1}^N \bar{\mathbf{x}}_i$ .

c. We can get the average structural function by slightly modifying equation (17.66). First, the conditional mean is

$$\begin{aligned} E(y_{it} | \mathbf{x}_{it}, c_i) &= q_1 [\Phi((q_1 - \mathbf{x}_{it}\boldsymbol{\beta} - c_i)/\sigma_u)] \\ &\quad + [\Phi((-q_2 + \mathbf{x}_{it}\boldsymbol{\beta} + c_i)/\sigma_u) - \Phi((q_1 - \mathbf{x}_{it}\boldsymbol{\beta} - c_i)/\sigma_u)] \cdot g(q_1, q_2, \mathbf{x}_{it}\boldsymbol{\beta} + c_i, \sigma_u^2) \\ &\quad + q_2 [\Phi((-q_2 + \mathbf{x}_{it}\boldsymbol{\beta} + c_i)/\sigma_u)] \end{aligned}$$

where

$$g(q_1, q_2, z, \sigma^2) \equiv z + \frac{\{\phi[(q_2 - z)/\sigma_u] - \phi[(q_1 - z)/\sigma_u]\}}{\{\Phi[(q_2 - z)/\sigma_u] - \Phi[(q_1 - z)/\sigma_u]\}}.$$

The ASF is obtained as a function of  $\mathbf{x}_t$  by averaging out  $c_i$ . But we can use iterated expectations, as usual, by first conditioning on  $\bar{\mathbf{x}}_i$  and then averaging out  $\bar{\mathbf{x}}_i$ :

$$E_{c_i}[m(\mathbf{x}_t, c_i)] = E_{\bar{\mathbf{x}}_i}\{E[m(\mathbf{x}_t, c_i) | \bar{\mathbf{x}}_i]\}$$

where  $m(\mathbf{x}_t, c) = E(y_{it} | \mathbf{x}_{it} = \mathbf{x}_t, c_i = c)$ . Using the same argument from the one-limit CRE

Tobit model,

$$\begin{aligned} h(\mathbf{x}_t, \bar{\mathbf{x}}_i) &\equiv E[m(\mathbf{x}_t, c_i) | \bar{\mathbf{x}}_i] = q_1 [\Phi((q_1 - \mathbf{x}_t\boldsymbol{\beta} - \psi - \bar{\mathbf{x}}_i\xi)/\sigma_v)] \\ &\quad + [\Phi((-q_2 + \mathbf{x}_t\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\xi)/\sigma_v) - \Phi((q_1 - \mathbf{x}_t\boldsymbol{\beta} - \psi - \bar{\mathbf{x}}_i\xi)/\sigma_v)] \\ &\quad \cdot g(q_1, q_2, \mathbf{x}_t\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\xi, \sigma_v^2) \\ &\quad + q_2 [\Phi((-q_2 + \mathbf{x}_t\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\xi)/\sigma_v)] \end{aligned}$$

where  $\sigma_v^2 = \sigma_a^2 + \sigma_u^2$ . The ASF is consistently estimated as

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1} \sum_{i=1}^N \hat{h}(\mathbf{x}_t, \bar{\mathbf{x}}_i)$$

where  $\hat{h}(\cdot, \cdot)$  denotes plugging in the MLEs of all estimates. Now take derivatives and changes with respect to  $\mathbf{x}_t$ .

d. Without assumption (17.78), we can just use a pooled two-limit Tobit analysis to estimate  $\beta$ ,  $\psi$ ,  $\xi$ , and  $\sigma_v^2$ . As in the standard Tobit case, we cannot separately estimate  $\sigma_a^2$  and  $\sigma_u^2$ . But the APEs are still identified as they depend only on  $\sigma_v^2$ , as shown in part c.

**17.17.** a. Plug in the expressions for  $c_{i1}$  and  $c_{i2}$  to get

$$\begin{aligned} y_{it1} &= \max(0, \alpha_1 y_{it2} + \mathbf{z}_{it1} \boldsymbol{\delta}_1 + \psi_1 + \bar{\mathbf{z}}_i \boldsymbol{\xi}_1 + a_{i1} + u_{it1}) \\ y_{it2} &= \mathbf{z}_{it} \boldsymbol{\pi}_2 + \psi_2 + \bar{\mathbf{z}}_i \boldsymbol{\xi}_2 + a_{i2} + u_{it2} \end{aligned}$$

or

$$\begin{aligned} y_{it1} &= \max(0, \alpha_1 y_{it2} + \mathbf{z}_{it1} \boldsymbol{\delta}_1 + \psi_1 + \bar{\mathbf{z}}_i \boldsymbol{\xi}_1 + v_{it1}) \\ y_{it2} &= \mathbf{z}_{it} \boldsymbol{\pi}_2 + \psi_2 + \bar{\mathbf{z}}_i \boldsymbol{\xi}_2 + v_{it2} \end{aligned}$$

where  $v_{it} = a_{i1} + u_{it1}$  and  $v_{it2} = a_{i2} + u_{it2}$ . Given the assumptions on  $D(u_{it1}, u_{it2} | \mathbf{z}_i, \mathbf{a}_i)$  and  $D(a_{i1}, a_{i2} | \mathbf{z}_i)$ , it follows that  $D(v_{it1}, v_{it2} | \mathbf{z}_i) = D(v_{it1}, v_{it2})$  is bivariate normal with mean zero. Therefore, we can write

$$\begin{aligned} v_{it1} &= \rho_1 v_{it2} + e_{it1} \\ D(e_{it1} | \mathbf{z}_i, v_{it2}) &= E(e_{it1}) = \text{Normal}(0, \sigma_{e_1}^2). \end{aligned}$$

It follows we can write

$$\begin{aligned} y_{it1} &= \max(0, \alpha_1 y_{it2} + \mathbf{z}_{it1} \boldsymbol{\delta}_1 + \psi_1 + \bar{\mathbf{z}}_i \boldsymbol{\xi}_1 + \rho_1 v_{it2} + e_{it1}) \\ D(e_{it1} | \mathbf{z}_i, y_{it2}, v_{it2}) &= \text{Normal}(0, \sigma_{e_1}^2) \end{aligned}$$

and now a pooled two-step method is immediate. First, obtain the residuals  $\hat{v}_{it2}$  from the pooled regression

$$y_{it2} \text{ on } \mathbf{z}_{it}, 1, \bar{\mathbf{z}}_i, t = 1, \dots, T; i = 1, \dots, N.$$

Then use pooled Tobit of

$$y_{it1} \text{ on } y_{it2}, \mathbf{z}_{it1}, 1, \bar{\mathbf{z}}_i, \hat{v}_{it2}$$

to estimate  $\alpha_1, \delta_1, \psi_1, \xi_1, \rho_1$ , and  $\sigma_{e_1}^2$ .

Incidentally, the statement of the problem said that  $\{y_{it2}\}$  will not be strictly exogenous in the estimable equation. While that is true of the previously proposed solution, in other approaches  $\{y_{it2}\}$  can be rendered strictly exogenous. Here is one possibility. Let  $\mathbf{v}_{i2} = (v_{i12}, \dots, v_{iT2})$  be the entire history on the reduced form errors. Then, given the previous assumptions,  $D(v_{it1} | \mathbf{z}_i, \mathbf{v}_{i2}) = D(v_{it1} | \mathbf{v}_{i2})$ . Because  $v_{it1} = a_{i1} + u_{it1}$ , it is reasonable to assume a Chamberlain-Mundlak representation, for example,

$$v_{it1} = \rho_1 v_{it2} + \gamma_1 \bar{v}_{i2} + e_{it1}$$

where now  $e_{it1}$  is independent of  $(\mathbf{z}_i, \mathbf{v}_{i2})$  and therefore of  $(\mathbf{z}_i, \mathbf{v}_{i2}, \mathbf{y}_{i2})$ , where

$\mathbf{y}_{i2} = (y_{i12}, \dots, y_{iT2})$ . This means that in the equation

$$y_{it1} = \max(0, \alpha_1 y_{it2} + \mathbf{z}_{it1} \delta_1 + \psi_1 + \bar{\mathbf{z}}_i \xi_1 + \rho_1 v_{it2} + \gamma_1 \bar{v}_{i2} + e_{it1}),$$

$\{y_{ir2}, \mathbf{z}_{ir}, v_{ir2} : r = 1, \dots, T\}$  is strictly exogenous with respect to  $e_{it1}$ . The CF approach changes in that we add  $\hat{v}_{i2}$  as an additional explanatory variable (along with  $\hat{v}_{it2}$ ) in using pooled Tobit. Because of strict exogeneity, approaches that attempt to exploit the serial dependence in the scores are now possible.

b. As usual, the two-step nature of the estimation needs to be accounted for by using either the delta method or the panel bootstrap. In using the delta method, the serial dependence in the scores should be accounted for. It is automatically accounted for with the panel bootstrap because the cross section units are resampled.

c. We have used this approach several times. Let  $m(z, \sigma^2)$  denote the unconditional mean

function for the standard Tobit model. Then

$$\text{ASF}(y_{t2}, \mathbf{z}_{t1}) = E_{(\bar{\mathbf{z}}_i, v_{it2})} [m(\alpha_1 y_{t2} + \mathbf{z}_{t1} \boldsymbol{\delta}_1 + \psi_1 + \bar{\mathbf{z}}_i \boldsymbol{\xi}_1 + \rho_1 v_{it2}, \sigma_{e1}^2)]$$

and so a consistent estimator is

$$\widehat{\text{ASF}}(y_{t2}, \mathbf{z}_{t1}) = N^{-1} \sum_{i=1}^N m(\hat{\alpha}_1 y_{t2} + \mathbf{z}_{t1} \hat{\boldsymbol{\delta}}_1 + \hat{\psi}_1 + \bar{\mathbf{z}}_i \hat{\boldsymbol{\xi}}_1 + \hat{\rho}_1 \hat{v}_{it2}, \hat{\sigma}_{e1}^2).$$

As usual, the estimated APEs are obtained by taking derivatives or changes with respect to  $(y_{t2}, \mathbf{z}_{t1})$ .

**17.18.** a. Once we assume  $\mathbf{z}$  is exogenous in the structural equation – and  $E(u_1|\mathbf{z}) = 0$  ensures exogeneity – then we only need the rank condition. The assumption that  $E(\mathbf{z}'\mathbf{z})$  is nonsingular is not usually restrictive. The important condition with a single endogenous explanatory variable is

$$L(y_2|\mathbf{z}) \neq L(y_2|\mathbf{z}_1),$$

so there is at least one element of  $\mathbf{z}$  not in  $\mathbf{z}_1$  that explains variation in  $y_2$ .

b. We can draw on the optimal instrument variables results from Section 8.6. The condition  $E(u_1|\mathbf{z}) = 0$  ensures that any function of  $\mathbf{z}$  is a valid instrumental variable candidate, and also implies that  $E(u_1^2|\mathbf{z}) = \text{Var}(u_1|\mathbf{z})$ . Because  $E(u_1^2|\mathbf{z})$  is constant, from Theorem 8.5 the optimal IVs are

$$[E(y_2|\mathbf{z}), \mathbf{z}_1].$$

If we think  $D(y_2|\mathbf{z})$  follows a standard Tobit then we should obtain  $E(y_2|\mathbf{z})$  from the Tobit model. Recall that if

$$D(y_2|\mathbf{z}) = \text{Tobit}(\mathbf{z}\boldsymbol{\delta}_2, \tau_2^2)$$

then

$$E(y_2|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\delta}_2/\tau_2)\mathbf{z}\boldsymbol{\delta}_2 + \tau_2\phi(\mathbf{z}\boldsymbol{\delta}_2/\tau_2)$$

Therefore, if we run Tobit in a first stage, we get

$$\hat{m}_{i2} \equiv \hat{E}(y_{i2}|\mathbf{z}_i) = \Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2/\hat{\tau}_2)\mathbf{z}_i\hat{\boldsymbol{\delta}}_2 + \hat{\tau}_2\phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2/\hat{\tau}_2)$$

and then use IVs  $(\hat{m}_{i2}, \mathbf{z}_{i1})$  in the equation

$$y_{i1} = \alpha_1 y_{i2} + \mathbf{z}_{i1}\boldsymbol{\delta}_1 + u_{i1}$$

by IV. This approach just identifies the parameters. We can get overidentification (if we have enough elements of  $\mathbf{z}_i$ ) by using all of  $\mathbf{z}_i$  in place of  $\mathbf{z}_{i1}$ .

Provided we maintain  $E(u_1|\mathbf{z}) = 0$ , using Tobit fitted values as instruments is no less (and no more) robust than using 2SLS. As mentioned previously, *any* function of  $\mathbf{z}_i$  is valid as a potential instrument. Even if the Tobit model is incorrect, we know the quasi-MLEs converge very generally. Call the plims  $\boldsymbol{\delta}_2^*$  and  $\tau_2^*$  and define

$$m_{i2}^* \equiv \Phi(\mathbf{z}_i\boldsymbol{\delta}_2^*/\tau_2^*)\mathbf{z}_i\boldsymbol{\delta}_2^* + \tau_2^*\phi(\mathbf{z}_i\boldsymbol{\delta}_2^*/\tau_2^*),$$

which is just a function of  $\mathbf{z}_i$ . Ruling out perfect collinearity in  $(m_{i2}^*, \mathbf{z}_{i1})$ , the rank condition is

$$L(y_{i2}|m_{i2}^*, \mathbf{z}_{i1}) \neq L(y_{i2}|\mathbf{z}_{i1}),$$

which simply means that  $m_{i2}^*$  should have some partial correlation with  $\mathbf{z}_{i1}$ , something we would expect quite generally if  $\mathbf{z}_{i2}$  is partially correlated with  $y_{i2}$ .

Using  $\hat{m}_{i2}$  as an instrument for  $y_{i2}$  is preferred to using it as a regressor in place of  $\hat{m}_{i2}$ . If we use  $\hat{m}_{i2}$  as a regressor then we are effectively assuming

$E(y_2|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\delta}_2/\tau_2)\mathbf{z}\boldsymbol{\delta}_2 + \tau_2\phi(\mathbf{z}\boldsymbol{\delta}_2/\tau_2)$  (and that we have consistent estimators of the parameters in this mean). Generally, the estimates of  $\boldsymbol{\delta}_1$  and  $\alpha_1$  would be inconsistent of the Tobit model for  $y_2$  is misspecified. When  $\hat{m}_{i2}$  as an instrument, the reduced implicit reduced

form for the IV estimation is

$$L(y_2|m_{i2}^*, \mathbf{z}_{i1}) = \psi_2 m_{i2}^* + \mathbf{z}_{i1} \boldsymbol{\eta}_1$$

and we do not need  $\psi_2 = 1$  and  $\boldsymbol{\eta}_1 = \mathbf{0}$ , as the plug-in-regressor method essentially does.

c. We can write

$$\begin{aligned} y_1 &= \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 v_2 + e_1 \\ E(e_1|\mathbf{z}, y_2, v_2) &= 0 \end{aligned}$$

It follows that

$$E(y_1|\mathbf{z}, y_2) = \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 E(v_2|\mathbf{z}, y_2)$$

and we can compute  $E(v_2|\mathbf{z}, y_2)$  given that  $D(y_2|\mathbf{z}) = \text{Tobit}(\mathbf{z}\boldsymbol{\delta}_2, \tau_2^2)$ . In fact, as shown in Vella (1993, *International Economic Review*),

$$\begin{aligned} E(v_2|\mathbf{z}, y_2) &= 1[y_2 > 0]v_2 - 1[y_2 = 0]\tau_2 \left[ \frac{\phi(\mathbf{z}\boldsymbol{\delta}_2/\tau_2)}{1 - \Phi(\mathbf{z}\boldsymbol{\delta}_2/\tau_2)} \right] \\ &= 1[y_2 > 0]v_2 - 1[y_2 = 0]\tau_2 \lambda(-\mathbf{z}\boldsymbol{\delta}_2/\tau_2) \end{aligned}$$

where  $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$  is the inverse Mills ratio. (This is an example of a *generalized residual*.)

Given the Tobit MLEs, we can easily construct

$$\hat{E}(v_{i2}|\mathbf{z}_i, y_{i2}) = 1[y_{i2} > 0]\hat{v}_{i2} - 1[y_{i2} = 0]\hat{\tau}_2 \lambda(-\mathbf{z}_i \hat{\boldsymbol{\delta}}_2 / \hat{\tau}_2)$$

in a first stage, and then in a second stage run the OLS regression

$$y_{i1} \text{ on } \mathbf{z}_{i1}, y_{i2}, 1[y_{i2} > 0]\hat{v}_{i2} - 1[y_{i2} = 0]\hat{\tau}_2 \lambda(-\mathbf{z}_i \hat{\boldsymbol{\delta}}_2 / \hat{\tau}_2)$$

to consistently estimate  $\boldsymbol{\delta}_1$ ,  $\alpha_1$ , and  $\rho_1$ .

Because the CF approach is based on  $E(y_1|\mathbf{z}, y_2)$ , nothing important changes if we start with

$$y_1 = \mathbf{g}_1(\mathbf{z}_1, y_2)\boldsymbol{\beta}_1 + u_1.$$

The same reasoning as before gets us to

$$E(y_1|\mathbf{z}, y_2) = \mathbf{g}_1(\mathbf{z}_1, y_2)\boldsymbol{\beta}_1 + \rho_1 \{1[y_2 > 0]v_2 - 1[y_2 = 0]\tau_2\lambda(-\mathbf{z}\boldsymbol{\delta}_2/\tau_2)\}$$

and so adding the same CF as before works for consistently estimating  $\boldsymbol{\beta}_1$ . Of course, the interpretation of  $\boldsymbol{\beta}_1$  depends on the nature of the functions in  $\mathbf{g}_1(\mathbf{z}_1, y_2)$ .

d. The 2SLS estimator that effectively ignores the nature of  $y_2$  is simple and fully robust – assuming we have at least one valid instrument for  $y_2$ . Standard errors (robust to heteroskedasticity) are easy to obtain. Its primary drawback is that it may be (asymptotically) inefficient compared with the other methods. An additional shortcoming is that if we use general functions  $\mathbf{g}_1(\mathbf{z}_1, y_2)$  we need to decide on instruments for any function that includes  $y_2$ . [Remember we are generally *not* allowed to plug in a fitted value to obtain  $\mathbf{g}_1(\mathbf{z}_{i1}, \hat{y}_{i2})$  and then regress  $y_{i1}$  on  $\mathbf{g}_1(\mathbf{z}_{i1}, \hat{y}_{i2})$ .]

The method of using the Tobit fitted value as the IV for  $y_2$  is just as robust as 2SLS estimator yet it exploits the corner solution nature of  $y_2$ . It need not be more (asymptotically) efficient than 2SLS, but it could be even if the Tobit model for  $y_2$  is misspecified [or  $\text{Var}(u_1|\mathbf{z})$  is homoskedastic, or both]. That we have estimated the instruments in a first stage can be ignored in the  $\sqrt{N}$ -asymptotic distribution of the IV estimator. Like the 2SLS estimator, having general functions  $\mathbf{g}_1(\mathbf{z}_1, y_2)$  means we would have to obtain IVs for all endogenous functions. This is almost always possible but is not always obvious.

The CF method is simple to compute but the standard errors generally have to account for the two-step estimation unless  $\rho_1 = 0$ . (The CF method provides a simple test of the null that  $y_2$  is endogenous: just use a heteroskedasticity-robust  $t$  statistic for  $\hat{\rho}_1$ .) Another drawback to

the CF method is that it is derived assuming the Tobit model for  $y_2$  holds. Generally, it is inconsistent if the Tobit model fails (just like using Tobit fitted values as regressors rather than instruments). An advantage of the CF method is that, as discussed in part c, it is easily applied for general functions  $\mathbf{g}_1(\mathbf{z}_1, y_2)$ . In such cases, the CF method is likely to be more efficient asymptotically than 2SLS or the IV method described in part b.

e. If we assume joint normality of  $(u_1, v_2)$  (and independence from  $\mathbf{z}$ ) then MLE becomes attractive. It will give the asymptotically efficient estimators and there is no two-step estimation issue to deal with (as in the CF case). The log likelihood is a bit tricky to obtain because it depends on  $D(y_1|y_2, \mathbf{z})$ . We already know that  $D(y_2|\mathbf{z})$  follows a Tobit. We also know  $D(y_1|v_2, \mathbf{z})$  follows a classical linear regression model with mean  $\mathbf{z}_1\delta_1 + \alpha_1 y_2 + \rho_1 v_2$  and variance  $\sigma_{\varepsilon_1}^2$ . For  $y_2 > 0$ ,  $D(y_1|y_2, \mathbf{z}) = D(y_1|v_2, \mathbf{z})$ . For  $y_2 = 0$ , we have to integrate over  $v_2 \leq -\mathbf{z}\delta_2$ , just like in Problem 17.6. When we have to two densities, we use, for each  $i$ ,

$$\log[f_1(y_{i1}|y_{i2}, \mathbf{z}_i; \boldsymbol{\theta})] + \log[f_2(y_{i2}|\mathbf{z}_i; \boldsymbol{\delta}_2, \tau_2^2)]$$

as the log likelihood.

**17.19.** a. Because of the conditional independence assumption we have

$$f(\eta_1, \dots, \eta_G | \mathbf{x}, \mathbf{c}; \boldsymbol{\gamma}_o) = \prod_{g=1}^G f_g(\eta_g | \mathbf{x}, \mathbf{c}; \boldsymbol{\gamma}_o^g)$$

for dummy arguments  $(\eta_1, \dots, \eta_G)$ .

b. To obtain the density of  $(y_1, \dots, y_G)$  given  $\mathbf{x}$  we integrate out  $\mathbf{c}$ :

$$g(\eta_1, \dots, \eta_G | \mathbf{x}; \boldsymbol{\gamma}_o, \boldsymbol{\delta}_o) = \int \left[ \prod_{g=1}^G f_g(\eta_g | \mathbf{x}, \mathbf{c}; \boldsymbol{\gamma}_o^g) \right] h(\mathbf{c} | \mathbf{x}; \boldsymbol{\delta}_o) d\mathbf{c}$$

where, in general, the integral is a multiple integral. Also, we have indicated  $\mathbf{c}$  as a continuous random vector but it need not be.



c. The log likelihood for a random draw  $i$  is simply

$$\ell_i(\boldsymbol{\theta}) = \log \int \left[ \prod_{g=1}^G f_g(y_{ig} | \mathbf{x}_i, \mathbf{c}; \boldsymbol{\gamma}^g) \right] h(\mathbf{c} | \mathbf{x}_i; \boldsymbol{\delta}) d\mathbf{c}$$

where  $\boldsymbol{\theta}$  contains all parameters.

**17.20.** a. The Stata output is given below. The signs of the coefficients are generally what we expect: lagged hours has a positive coefficient, as does initial hours in 1980. Thus, unobserved heterogeneity that positively affects hours worked in 1980 also positively affects hours contemporaneously. The variables *nwifeinc*, *ch0\_2*, and *ch3\_5* all have negative and statistically significant coefficients. The one slight puzzle is that the older children variable has a positive and just statistically significant coefficient.

```
use \mitbook1_2e\statafiles\psid80_92, clear
tsset id year
* Lagged dependent variable:
bysort id (year): gen hours_1 = L.hours
* Put initial condition in years 81-92:
by id: gen hours80 = hours[1]
* Create exogenous variables for years 81-92:
forv i=81/92 {
by id: gen nwifeinc`i' = nwifeinc[`i'-80]
}
forv i=81/92 {
by id: gen ch0_2`i' = ch0_2[`i'-80]
}
forv i=81/92 {
by id: gen ch3_5`i' = ch3_5[`i'-80]
}
forv i=81/92 {
by id: gen ch6_17`i' = ch6_17[`i'-80]
}
forv i=81/92 {
by id: gen marr`i' = marr[`i'-80]
}

xttobit hours hours_1 hours80 nwifeinc nwifeinc81-nwifeinc92
      ch0_2 ch0_2_81-ch0_2_92 ch3_5 ch3_5_81-ch3_5_92
      ch6_17 ch6_17_81-ch6_17_92 marr marr81-marr92 y82-y92, ll(0) re

note: marr86 omitted because of collinearity
note: marr89 omitted because of collinearity
note: marr90 omitted because of collinearity
note: marr91 omitted because of collinearity
note: marr92 omitted because of collinearity
```

Random-effects tobit regression  
Group variable: id

Number of obs = 10776  
Number of groups = 898

Random effects u\_i ~Gaussian

Obs per group: min = 12  
avg = 12.  
max = 12

Log likelihood = -62882.574

Wald chi2(73) = 7997.27  
Prob > chi2 = 0.0000

hours	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
hours_1	.7292676	.0119746	60.90	0.000	.7057978	.7527375
hours80	.2943114	.0181831	16.19	0.000	.2586731	.3299496
nwifeinc	-1.286033	.3221528	-3.99	0.000	-1.917441	-.6546256
nwifeinc81	.6329715	1.08601	0.58	0.560	-1.49557	2.761513
nwifeinc82	.1812886	1.914677	0.09	0.925	-3.57141	3.933987
nwifeinc83	-.6567582	1.822493	-0.36	0.719	-4.228778	2.915262
nwifeinc84	-.9568491	1.344172	-0.71	0.477	-3.591379	1.677681
nwifeinc85	-1.169828	1.186202	-0.99	0.324	-3.494742	1.155085
nwifeinc86	.437133	1.142004	0.38	0.702	-1.801153	2.675419
nwifeinc87	-2.53217	1.067478	-2.37	0.018	-4.624388	-.4399525
nwifeinc88	-.8224415	.6884551	-1.19	0.232	-2.171789	.5269057
nwifeinc89	1.325135	.792212	1.67	0.094	-.2275717	2.877842
nwifeinc90	.0811052	.5898146	0.14	0.891	-1.07491	1.237121
nwifeinc91	1.550942	.861745	1.80	0.072	-.1380467	3.239932
nwifeinc92	-.6307469	.7778347	-0.81	0.417	-2.155275	.893781
ch0_2	-146.0974	21.04471	-6.94	0.000	-187.3443	-104.8506
ch0_2_81	170.7185	93.39678	1.83	0.068	-12.33577	353.7729
ch0_2_82	89.33674	97.13967	0.92	0.358	-101.0535	279.727
ch0_2_83	86.23234	100.9511	0.85	0.393	-111.6281	284.0928
ch0_2_84	-68.59569	100.3924	-0.68	0.494	-265.3612	128.1699
ch0_2_85	-11.17728	96.67042	-0.12	0.908	-200.6478	178.2933
ch0_2_86	83.53639	108.6359	0.77	0.442	-129.386	296.4588
ch0_2_87	-77.82159	110.9494	-0.70	0.483	-295.2784	139.6352
ch0_2_88	-84.03353	141.8912	-0.59	0.554	-362.1352	194.0681
ch0_2_89	48.15522	211.0668	0.23	0.820	-365.528	461.8385
ch0_2_90	17.49295	102.1897	0.17	0.864	-182.7952	217.7811
ch0_2_91	123.7578	96.82641	1.28	0.201	-66.01852	313.534
ch0_2_92	-48.17428	84.53394	-0.57	0.569	-213.8578	117.5092
ch3_5	-80.13216	17.85232	-4.49	0.000	-115.1221	-45.14226
ch3_5_81	39.44899	62.21243	0.63	0.526	-82.48514	161.3831
ch3_5_82	102.3494	72.50917	1.41	0.158	-39.766	244.4647
ch3_5_83	-38.86165	74.09378	-0.52	0.600	-184.0828	106.3595
ch3_5_84	-101.8966	94.20263	-1.08	0.279	-286.5304	82.73714
ch3_5_85	-4.967801	99.28115	-0.05	0.960	-199.5553	189.6197
ch3_5_86	-25.96859	100.6704	-0.26	0.796	-223.279	171.3418
ch3_5_87	5.59682	98.3939	0.06	0.955	-187.2517	198.4453
ch3_5_88	46.38591	93.80288	0.49	0.621	-137.4644	230.2362
ch3_5_89	-95.69263	129.6341	-0.74	0.460	-349.7709	158.3856
ch3_5_90	43.70922	129.4244	0.34	0.736	-209.9579	297.3763
ch3_5_91	147.7391	143.4973	1.03	0.303	-133.5105	428.9886
ch3_5_92	-166.5773	214.5918	-0.78	0.438	-587.1694	254.0149
ch6_17	22.18895	10.08538	2.20	0.028	2.421976	41.95593
ch6_17_81	4.64258	37.68437	0.12	0.902	-69.21744	78.5026
ch6_17_82	64.27872	55.13925	1.17	0.244	-43.79223	172.3497
ch6_17_83	-66.82245	57.25136	-1.17	0.243	-179.033	45.38815

ch6_17_84	1.173452	56.00241	0.02	0.983	-108.5893	110.9362
ch6_17_85	6.738214	54.27217	0.12	0.901	-99.63328	113.1097
ch6_17_86	85.64549	57.28103	1.50	0.135	-26.62327	197.9142
ch6_17_87	-65.96152	62.6244	-1.05	0.292	-188.7031	56.78006
ch6_17_88	19.1112	56.21565	0.34	0.734	-91.06945	129.2918
ch6_17_89	4.85883	61.37184	0.08	0.937	-115.4278	125.1454
ch6_17_90	16.18911	60.09357	0.27	0.788	-101.5921	133.9703
ch6_17_91	-21.25498	55.55783	-0.38	0.702	-130.1463	87.63636
ch6_17_92	-7.632119	53.88032	-0.14	0.887	-113.2356	97.97137
marr	-199.1315	144.719	-1.38	0.169	-482.7755	84.51247
marr81	127.5178	356.477	0.36	0.721	-571.1642	826.1998
marr82	-13.59679	491.133	-0.03	0.978	-976.1997	949.0062
marr83	-507.5586	434.9469	-1.17	0.243	-1360.039	344.9217
marr84	1318.284	564.6247	2.33	0.020	211.6404	2424.928
marr85	-326.1983	585.7084	-0.56	0.578	-1474.166	821.769
marr86	(omitted)					
marr87	131.824	331.72	0.40	0.691	-518.3353	781.9832
marr88	-491.7295	306.7196	-1.60	0.109	-1092.889	109.4299
marr89	(omitted)					
marr90	(omitted)					
marr91	(omitted)					
marr92	(omitted)					
y82	-32.79071	25.88734	-1.27	0.205	-83.52896	17.94755
y83	20.40184	25.84829	0.79	0.430	-30.25988	71.06355
y84	105.7757	25.7722	4.10	0.000	55.2631	156.2883
y85	26.36698	25.95325	1.02	0.310	-24.50046	77.23441
y86	26.82807	25.99402	1.03	0.302	-24.11928	77.77542
y87	-.1477861	26.16878	-0.01	0.995	-51.43764	51.14207
y88	21.84475	26.28302	0.83	0.406	-29.66903	73.35853
y89	33.76287	26.39745	1.28	0.201	-17.97518	85.50092
y90	30.54594	26.52445	1.15	0.249	-21.44102	82.5329
y91	29.17601	26.64107	1.10	0.273	-23.03953	81.39155
y92	-27.66915	26.97277	-1.03	0.305	-80.53481	25.19651
_cons	-165.6397	47.85094	-3.46	0.001	-259.4258	-71.85356
-----						
/sigma_u	310.4876	12.44431	24.95	0.000	286.0972	334.878
/sigma_e	508.4561	4.327479	117.49	0.000	499.9744	516.9378
-----						
rho	.2716099	.0164159			.2404141	.3046996
-----						

Observation summary:      2835   left-censored observations  
                                  7941   uncensored observations  
                                  0   right-censored observations

b. The Stata commands below produce the scale factor for the APE of a continuous explanatory variable, evaluated at  $hours_{t-1} = 0$ . All other variables are averaged out, and the scale factor is for 1992. The APE for *nwifeinc* in 1992 is about  $-.742$ . Because *nwifeinc* is in \$1,000s, the coefficient implies that a \$10,000 increase in other sources of income decreases estimated annual hours by about 7.4. This is a small economic effect given that the average

hours in 1992 is about 1,155, and a \$10,000 increase is fairly large.

```
. predict xbh, xb
(898 missing values generated)

. gen xbh_h0 = xbh - _b[hours_1]*hours_1
(898 missing values generated)

. gen scale = normal(xbh_h0/sqrt(_b[/sigma_u]^2 + _b[/sigma_e]^2))
(898 missing values generated)

. sum scale if y92
```

Variable	Obs	Mean	Std. Dev.	Min	Max
scale	898	.5769774	.1692471	.0649065	.9402357

```
. di .5769774*_b[nwifeinc]
-.74201219
```

```
. sum hours nwifeinc if y92
```

Variable	Obs	Mean	Std. Dev.	Min	Max
hours	898	1155.318	899.656	0	3916
nwifeinc	898	43.57829	44.2727	-7.249999	601.504

c. Because *ch0\_2* is a discrete variable, we compute the difference in the conditional mean

function and then average. The APE in 1992 in moving from zero to one small children is

about -116.47, which means average annual hours fall by about 116.5 hours.

```
. gen xbh_c0 = xbh - _b[ch0_2]*ch0_2
(898 missing values generated)

. gen xbh_c1 = xbh_c0 + _b[ch0_2]
(898 missing values generated)

. gen mean0 = normal(xbh_c0/sqrt(_b[/sigma_u]^2 + _b[/sigma_e]^2))*xbh_c0 + sqrt
> en(xbh_c0/sqrt(_b[/sigma_u]^2 + _b[/sigma_e]^2))
(898 missing values generated)

. gen mean1 = normal(xbh_c1/sqrt(_b[/sigma_u]^2 + _b[/sigma_e]^2))*xbh_c1 + sqrt
> en(xbh_c1/sqrt(_b[/sigma_u]^2 + _b[/sigma_e]^2))
(898 missing values generated)

. gen diff = mean1 - mean0
(898 missing values generated)

. sum diff if y92
```

Variable	Obs	Mean	Std. Dev.	Min	Max
----------	-----	------	-----------	-----	-----

diff		898	-116.4689	38.50869	-146.0974	-7.479618
------	--	-----	-----------	----------	-----------	-----------

## Solutions to Chapter 18 Problems

**18.1.** a. This is a simple problem in univariate calculus. Write  $q(\mu) \equiv \mu_o \log(\mu) - \mu$  for  $\mu > 0$ . Then  $dq(\mu)/d\mu \equiv \mu_o/\mu - 1$ , so  $\mu = \mu_o$  uniquely sets the derivative to zero. The second derivative of  $q(\mu)$  is  $-\mu_o\mu^{-2} < 0$  for all  $\mu > 0$ , so the sufficient second order condition for a maximum is satisfied.

b. For the exponential case,  $q(\mu) \equiv E[\ell_i(\mu)] = -\mu_o/\mu - \log(\mu)$ . The first order condition is  $\mu_o\mu^{-2} - \mu^{-1} = 0$ , which is uniquely solved by  $\mu = \mu_o$ . The second derivative is  $-2\mu_o\mu^{-3} + \mu^{-2}$ , which, when evaluated at  $\mu_o$ , gives  $-2\mu_o^{-2} + \mu_o^{-2} = -\mu_o^{-2} < 0$ .

**18.2.** When  $m(\mathbf{x}, \boldsymbol{\beta}) = \exp(\mathbf{x}\boldsymbol{\beta})$ , we have  $\mathbf{s}_i(\hat{\boldsymbol{\beta}}) = \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})\mathbf{x}_i'\hat{u}_i/\exp(\mathbf{x}_i\hat{\boldsymbol{\beta}}) = \mathbf{x}_i'\hat{u}_i$ , where  $\hat{u}_i = y_i - \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})$ . Further, the Hessian  $\mathbf{H}_i(\boldsymbol{\beta})$  does not depend on  $y_i$ , and

$$\mathbf{A}_i(\hat{\boldsymbol{\beta}}) = -\mathbf{H}_i(\hat{\boldsymbol{\beta}}) = \left[ \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}}) \right]^2 \mathbf{x}_i'\mathbf{x}_i / \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}}) = \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})\mathbf{x}_i'\mathbf{x}_i.$$

Therefore, we can write equation (18.14) as

$$\left( \sum_{i=1}^N \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})\mathbf{x}_i'\mathbf{x}_i \right)^{-1} \left( \sum_{i=1}^N \hat{u}_i^2 \mathbf{x}_i'\mathbf{x}_i \right) \left( \sum_{i=1}^N \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})\mathbf{x}_i'\mathbf{x}_i \right)^{-1}.$$

**18.3.** a. The Stata output is below. Neither the price nor income variable is significant at any reasonable significance level, although the coefficient estimates are the expected sign. It does not matter whether we use the usual or robust standard errors. The two variables are jointly insignificant, too, using the usual and heteroskedasticity-robust tests ( $p$ -values = .490, .344, respectively).

```
. use smoke
```

```
. reg cigs lcigpric lincome restaurn white educ age agesq
```

Source	SS	df	MS	Number of obs =	807
-----+-----				F( 7, 799) =	6.38
Model	8029.43631	7	1147.06233	Prob > F =	0.0000

Residual		143724.246	799	179.880158	R-squared	=	0.0529
-----							
Total		151753.683	806	188.280003	Adj R-squared	=	0.0446
					Root MSE	=	13.412

cigs	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
lcigpric	-.8509044	5.782321	-0.15	0.883	-12.20124	10.49943
lincome	.8690144	.7287636	1.19	0.233	-.561503	2.299532
restaurn	-2.865621	1.117406	-2.56	0.011	-5.059019	-.6722234
white	-.5592363	1.459461	-0.38	0.702	-3.424067	2.305594
educ	-.5017533	.1671677	-3.00	0.003	-.829893	-.1736135
age	.7745021	.1605158	4.83	0.000	.4594197	1.089585
agesq	-.0090686	.0017481	-5.19	0.000	-.0124999	-.0056373
_cons	-2.682435	24.22073	-0.11	0.912	-50.22621	44.86134

```
. test lcigpric lincome
```

```
( 1)  lcigpric = 0
( 2)  lincome = 0
```

```
F( 2, 799) = 0.71
Prob > F = 0.4899
```

```
. reg cigs lcigpric lincome restaurn white educ age agesq, robust
```

Linear regression	Number of obs =	807
	F( 7, 799) =	9.38
	Prob > F =	0.0000
	R-squared =	0.0529
	Root MSE =	13.412

cigs	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
lcigpric	-.8509044	6.054396	-0.14	0.888	-12.7353	11.0335
lincome	.8690144	.597972	1.45	0.147	-.3047672	2.042796
restaurn	-2.865621	1.017275	-2.82	0.005	-4.862469	-.868774
white	-.5592363	1.378283	-0.41	0.685	-3.26472	2.146247
educ	-.5017533	.1624097	-3.09	0.002	-.8205533	-.1829532
age	.7745021	.1380317	5.61	0.000	.5035545	1.04545
agesq	-.0090686	.0014589	-6.22	0.000	-.0119324	-.0062048
_cons	-2.682435	25.90194	-0.10	0.918	-53.52632	48.16145

```
. test lcigpric lincome
```

```
( 1)  lcigpric = 0
( 2)  lincome = 0
```

```
F( 2, 799) = 1.07
Prob > F = 0.3441
```

b. While the price variable is still highly insignificant ( $p$ -value = .46), the income variable,

based on the usual Poisson standard errors, is very significant:  $t = 5.11$ . Both estimates are elasticities: the estimate price elasticity is  $-.106$  and the estimated income elasticity is  $.104$ . Incidentally, if you drop *restaurn* – a binary indicator for restaurant smoking restrictions at the state level – then *lcigpric* becomes much more significant (using the MLE standard errors). In this data set, both *cigpric* and *restaurn* vary only at the state level, and, not surprisingly, they are significantly correlated. (States that have restaurant smoking restrictions also have higher average cigarette prices, on the order of 2.9%.)

```
. poisson cigs lcigpric lincome restaurn white educ age agesq
```

```
Poisson regression                                Number of obs   =          807
                                                    LR chi2(7)      =       1068.70
                                                    Prob > chi2     =         0.0000
Log likelihood = -8111.519                        Pseudo R2      =         0.0618
```

cigs	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
lcigpric	-.1059607	.1433932	-0.74	0.460	-.3870061	.1750847
lincome	.1037275	.0202811	5.11	0.000	.0639772	.1434779
restaurn	-.3636059	.0312231	-11.65	0.000	-.4248021	-.3024098
white	-.0552012	.0374207	-1.48	0.140	-.1285444	.0181421
educ	-.0594225	.0042564	-13.96	0.000	-.0677648	-.0510802
age	.1142571	.0049694	22.99	0.000	.1045172	.1239969
agesq	-.0013708	.000057	-24.07	0.000	-.0014825	-.0012592
_cons	.3964494	.6139626	0.65	0.518	-.8068952	1.599794

c. The GLM estimate of  $\sigma$  is about  $\hat{\sigma} = 4.51$ . This means all of the Poisson standard errors should be multiplied by this factor, as is done using the `glm` command in Stata, with the `sca(x2)` option. The  $t$  statistic on *lcigpric* is now very small ( $-.16$ ), and that on *lincome* falls to 1.13 – much more in line with the linear model  $t$  statistic (1.19 with the usual standard errors). Clearly, using the maximum likelihood standard errors is very misleading in this example. With the GLM standard errors, the restaurant restriction variable, education, and the age variables are still significant. (There is no race effect, conditional on the other covariates.)

```
. glm cigs lcigpric lincome restaurn white educ age agesq, family(poisson)
    sca(x2)
```



```

Generalized linear models
Optimization      : ML

Deviance          = 14752.46933
Pearson           = 16232.70987

Variance function: V(u) = u
Link function     : g(u) = ln(u)

No. of obs       = 807
Residual df      = 799
Scale parameter  =
(1/df) Deviance = 18.46367
(1/df) Pearson  = 20.31628

[Poisson]
[Log]

AIC               = 20.12272
BIC               = 9404.504

Log likelihood    = -8111.519022

```

cigs	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
lcigpric	-.1059607	.6463244	-0.16	0.870	-1.372733	1.160812
lincome	.1037275	.0914144	1.13	0.257	-.0754414	.2828965
restaurn	-.3636059	.1407338	-2.58	0.010	-.6394391	-.0877728
white	-.0552011	.1686685	-0.33	0.743	-.3857854	.2753831
educ	-.0594225	.0191849	-3.10	0.002	-.0970243	-.0218208
age	.1142571	.0223989	5.10	0.000	.0703561	.158158
agesq	-.0013708	.0002567	-5.34	0.000	-.001874	-.0008677
_cons	.3964493	2.76735	0.14	0.886	-5.027457	5.820355

(Standard errors scaled using square root of Pearson X2-based dispersion.)

```

. di sqrt(20.31628)
4.5073584

```

d. The usual LR statistic is about  $LR = 2 \cdot (8125.291 - 8111.519) = 27.54$ , which is a very large value in a  $\chi^2_2$  distribution ( $p$ -value  $\approx 0$ ). The QLR statistic divides the usual LR statistic by  $\hat{\sigma}^2 = 20.32$ , so  $QLR = 1.36$  ( $p$ -value  $\approx .51$ ). As expected, the QLR statistic shows that the variables are jointly insignificant, while the  $LR$  statistic shows strong statistical significance.

```

. poisson cigs restaurn white educ age agesq

```

```

Iteration 0:  log likelihood = -8125.618
Iteration 1:  log likelihood = -8125.2907
Iteration 2:  log likelihood = -8125.2906

```

```

Poisson regression
Log likelihood = -8125.2906

Number of obs   = 807
LR chi2(5)      = 1041.16
Prob > chi2     = 0.0000
Pseudo R2      = 0.0602

```

cigs	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
restaurn	-.3545336	.0308796	-11.48	0.000	-.4150564	-.2940107

white	-.0618025	.037371	-1.65	0.098	-.1350483	.0114433
educ	-.0532166	.0040652	-13.09	0.000	-.0611842	-.0452489
age	.1211174	.0048175	25.14	0.000	.1116754	.1305594
agesq	-.0014458	.0000553	-26.14	0.000	-.0015543	-.0013374
_cons	.7617484	.1095991	6.95	0.000	.5469381	.9765587

```
. di 2*(8125.291 - 8111.519)
27.544
```

```
. di 27.54/20.32
1.355315
```

```
. di chi2tail(2,1.36)
.50661699
```

e. Using the robust standard errors does not change any conclusions; in fact, most explanatory variables become slightly more significant than when we use the GLM standard errors. In this example, it is the adjustment by  $\hat{\sigma} > 1$  that makes the most difference. Having fully robust standard errors has no additional effect once we account for the severe overdispersion.

```
. glm cigs lcigpric lincome restaurn white educ age agesq, family(poisson)
robust
```

Generalized linear models	No. of obs	=	807
Optimization : ML	Residual df	=	799
	Scale parameter	=	
Deviance	=	14752.46933	(1/df) Deviance = 18.46367
Pearson	=	16232.70987	(1/df) Pearson = 20.31628
Variance function: V(u) = u	[Poisson]		
Link function : g(u) = ln(u)	[Log]		
	AIC	=	20.12272
Log pseudolikelihood = -8111.519022	BIC	=	9404.504

cigs	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
lcigpric	-.1059607	.6681827	-0.16	0.874	-1.415575	1.203653
lincome	.1037275	.083299	1.25	0.213	-.0595355	.2669906
restaurn	-.3636059	.140366	-2.59	0.010	-.6387182	-.0884937
white	-.0552011	.1632959	-0.34	0.735	-.3752553	.264853
educ	-.0594225	.0192058	-3.09	0.002	-.0970653	-.0217798
age	.1142571	.0212322	5.38	0.000	.0726427	.1558715
agesq	-.0013708	.0002446	-5.60	0.000	-.0018503	-.0008914
_cons	.3964493	2.97704	0.13	0.894	-5.438442	6.23134

f. We simply compute the turning point for the quadratic:

$$\hat{\beta}_{age}/(-2\hat{\beta}_{age^2}) = .1143/[2(.00137)] \approx 41.72, \text{ or at about 42 years of age.}$$

g. A double-hurdle model – which separates the initial decision to smoke at all from the decision of how much to smoke – seems like a good idea. Variables such as level of education, income, and age could have very different effects on the decision to smoke versus how much to smoke. It is certainly worth investigating. One approach is to model  $D(y|\mathbf{x}, y \geq 1)$  as, say, a truncated Poisson distribution, and then to model  $P(y = 0|\mathbf{x})$  as a logit or probit (with parameters free to vary from the truncated Poisson distribution).

**18.4.** In the notation of Section 14.5.3,  $r(\mathbf{w}_i, \boldsymbol{\theta}) = r(\mathbf{w}_i, \boldsymbol{\beta}) = y_i - m(\mathbf{x}_i, \boldsymbol{\beta})$ , and so  $R_o(\mathbf{x}_i) = -\nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta}_o)$ . Further,  $\Omega_o(\mathbf{x}_i) = \text{Var}[y_i - m(\mathbf{x}_i, \boldsymbol{\beta}_o)|\mathbf{x}_i] = \text{Var}(y_i|\mathbf{x}_i) = \sigma_o^2 m(\mathbf{x}_i, \boldsymbol{\beta}_o)$  under the GLM assumption. From equation (14.60), the asymptotic variance lower bound is

$$E\{[R_o(\mathbf{x}_i)' \Omega_o(\mathbf{x}_i)^{-1} R_o(\mathbf{x}_i)]\}^{-1} = \sigma_o^2 E[\nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta}_o)' \nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta}_o) / m(\mathbf{x}_i, \boldsymbol{\beta}_o)],$$

which is the same asymptotic variance for the Poisson QMLE under the GLM assumption.

**18.5. a.** We just use iterated expectations:

$$\begin{aligned} E(y_{it}|\mathbf{x}_i) &= E[E(y_{it}|\mathbf{x}_i, c_i)|\mathbf{x}_i] = E[c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})|\mathbf{x}_i, c_i] \\ &= E(c_i|\mathbf{x}_i) \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \\ &= \exp(\alpha + \bar{\mathbf{x}}_i\boldsymbol{\gamma}) \exp(\mathbf{x}_{it}\boldsymbol{\beta}) = \exp(\alpha + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma}). \end{aligned}$$

b. We are explicitly testing  $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ , but we are maintaining full independence of  $c_i$  and  $\mathbf{x}_i$  under  $H_0$ . We have enough assumptions to derive  $\text{Var}(\mathbf{y}_i|\mathbf{x}_i)$ , the  $T \times T$  conditional variance matrix of  $\mathbf{y}_i$  given  $\mathbf{x}_i$  under  $H_0$ . First,

$$\begin{aligned} \text{Var}(y_{it}|\mathbf{x}_i) &= E[\text{Var}(y_{it}|\mathbf{x}_i, c_i)|\mathbf{x}_i] + \text{Var}[E(y_{it}|\mathbf{x}_i, c_i)|\mathbf{x}_i] \\ &= E[c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})|\mathbf{x}_i] + \text{Var}[c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})|\mathbf{x}_i] \\ &= \exp(\alpha + \mathbf{x}_{it}\boldsymbol{\beta}) + \tau^2 [\exp(\mathbf{x}_{it}\boldsymbol{\beta})]^2, \end{aligned}$$

where  $\tau^2 \equiv \text{Var}(c_i)$  and we have used  $E(c_i|\mathbf{x}_i) = \exp(\alpha)$  under  $H_0$ . A similar, general expression holds for conditional covariances:

$$\begin{aligned}\text{Cov}(y_{it}, y_{ir}|\mathbf{x}_i) &= E[\text{Cov}(y_{it}, y_{ir}|\mathbf{x}_i, c_i)|\mathbf{x}_i] + \text{Cov}[E(y_{it}|\mathbf{x}_i, c_i), E(y_{ir}|\mathbf{x}_i, c_i)|\mathbf{x}_i] \\ &= 0 + \text{Cov}[c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}), c_i \exp(\mathbf{x}_{ir}\boldsymbol{\beta})|\mathbf{x}_i] \\ &= \tau^2 \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \exp(\mathbf{x}_{ir}\boldsymbol{\beta}).\end{aligned}$$

So, under  $H_0$ ,  $\text{Var}(y_i|\mathbf{x}_i)$  depends on  $\alpha$ ,  $\boldsymbol{\beta}$ , and  $\tau^2$ , all of which we can estimate. It is natural to use a score test – actually, its variable addition counterpart – to test  $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ . First, obtain consistent estimators  $\check{\alpha}$ ,  $\check{\boldsymbol{\beta}}$  by, say, pooled Poisson QMLE. Let  $\check{y}_{it} = \exp(\check{\alpha} + \mathbf{x}_{it}\check{\boldsymbol{\beta}})$  and  $\check{u}_{it} = y_{it} - \check{y}_{it}$ . A consistent estimator of  $\tau^2$  can be obtained from a simple pooled regression, through the origin, of

$$\check{u}_{it}^2 - \check{y}_{it} \text{ on } \exp(2\mathbf{x}_{it}\check{\boldsymbol{\beta}}), \quad t = 1, \dots, T; \quad i = 1, \dots, N.$$

Let  $\tilde{\tau}^2$  be the coefficient on  $\exp(2\mathbf{x}_{it}\check{\boldsymbol{\beta}})$ . It is consistent for  $\tau^2$  because, under  $H_0$ ,

$$E(u_{it}^2|\mathbf{x}_i) = \exp(\alpha + \mathbf{x}_{it}\boldsymbol{\beta}) + \tau^2[\exp(\mathbf{x}_{it}\boldsymbol{\beta})]^2,$$

where  $u_{it} \equiv y_{it} - E(y_{it}|\mathbf{x}_i)$ . We could also use the many covariance terms in estimating  $\tau^2$  because  $E(u_{it}u_{ir}|\mathbf{x}_i) = \tau^2 \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \exp(\mathbf{x}_{ir}\boldsymbol{\beta})$ ,  $t \neq r$ . So for all  $t, r = 1, \dots, T$ , we can write

$$u_{it}u_{ir} - d_{tr} \exp(\alpha + \mathbf{x}_{it}\boldsymbol{\beta}) = \tau^2 \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \exp(\mathbf{x}_{ir}\boldsymbol{\beta}) + v_{itr}$$

where  $E(v_{itr}|\mathbf{x}_i) = 0$  and  $d_{tr} = 1[t = r]$  is a dummy variable. The pooled regression would be

$$\check{u}_{it}\check{u}_{ir} - d_{tr}\check{y}_{it} \text{ on } \exp(\mathbf{x}_{it}\check{\boldsymbol{\beta}}) \exp(\mathbf{x}_{ir}\check{\boldsymbol{\beta}})$$

Next, we construct the  $T \times T$  weighting matrix for observation  $i$ , as in Section 18.7.3. The matrix  $\mathbf{W}_i(\check{\boldsymbol{\delta}}) = \mathbf{W}(\mathbf{x}_i, \check{\boldsymbol{\delta}})$  has diagonal elements

$$\exp(\check{\alpha} + \mathbf{x}_{it}\check{\boldsymbol{\beta}}) + \tilde{\tau}^2 \exp(2\mathbf{x}_{it}\check{\boldsymbol{\beta}}), \quad t = 1, \dots, T$$

and off-diagonal elements

$$\tilde{\tau}^2 \exp(\mathbf{x}_{it}\tilde{\boldsymbol{\beta}}) \exp(\mathbf{x}_{ir}\tilde{\boldsymbol{\beta}}), t \neq r.$$

Using this weighting matrix in a MWNLS estimation problem we can simply add the time averages,  $\bar{\mathbf{x}}_i$ , as an additional set of explanatory variables, and test their joint significance. This is the VAT version of the score test.

In practice, we might want a robust form of the test that does not require  $\text{Var}(\mathbf{y}_i|\mathbf{x}_i) = \mathbf{W}(\mathbf{x}_i, \boldsymbol{\delta})$  under  $H_0$ , where  $\mathbf{W}(\mathbf{x}_i, \boldsymbol{\delta})$  is the matrix described above. We can just use the fully robust variance matrix reported at the bottom of page 761.

Using modern software that supports MWNLS a simpler approach is to estimate the model under the alternative and obtain a Wald test of  $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ , where it is valid to act as if  $\text{Var}(c_i|\mathbf{x}_i) = \tau^2$  because this is true under the null. This would differ from the score approach in that  $\tau^2$  would be estimated using a first stage where  $\boldsymbol{\gamma}$  is also estimated. A fully robust Wald test is easy to obtain if we have any doubts about the variance-covariance structure.

Incidentally, this variance-covariance structure is different from the one used in the GEE literature for Poisson regression. With GEE and an exchangeable correlation structure, the nominal variance would be

$$\text{Var}(y_{it}|\mathbf{x}_i) = \exp(\alpha + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma})$$

and the nominal covariances

$$\text{Cov}(y_{it}, y_{ir}|\mathbf{x}_i) = \rho \sqrt{\exp(\mathbf{x}_{it}\boldsymbol{\beta}) \exp(\mathbf{x}_{ir}\boldsymbol{\beta})}.$$

c. If we assume (18.83), (18.84) and  $c_i = a_i \exp(\alpha + \bar{\mathbf{x}}_i\boldsymbol{\gamma})$  where  $a_i|\mathbf{x}_i \sim \text{Gamma}(\delta, \delta)$ , then testing involves estimation of a Poisson panel data model under random effects assumptions. Under these assumptions, we have

$$\begin{aligned}
y_{it}|\mathbf{x}_i, a_i &\sim \text{Poisson}[a_i \exp(\alpha + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma})] \\
y_{it}, y_{ir} &\text{ are independent conditional on } (\mathbf{x}_i, a_i) \\
a_i|\mathbf{x}_i &\sim \text{Gamma}(\delta, \delta).
\end{aligned}$$

In other words, the full set of random effect Poisson assumptions holds, but where the mean function in the Poisson distribution is  $a_i \exp(\alpha + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma})$ . In practice, we just add the (nonredundant elements of)  $\bar{\mathbf{x}}_i$  in each time period, along with a constant and  $\mathbf{x}_{it}$ , and carry out a random effects Poisson analysis. We can test  $H_0 : \boldsymbol{\gamma} = 0$  using the LR, Wald, or score approaches. Any of these would be asymptotically efficient. None is robust to misspecification of the Poisson distribution or the conditional independence assumption because we have used a full distribution for  $\mathbf{y}_i$  given  $\mathbf{x}_i$  in the MLE analysis.

**18.6. a.** We know from Problem 12.6 that pooled nonlinear least squares consistently estimates  $\boldsymbol{\beta}_o$  when  $\boldsymbol{\beta}_o$  appears in correctly specified conditional means for each  $t$ . Because  $\ddot{m}_{it}(\boldsymbol{\beta})$  depends on  $\mathbf{x}_i$  we should show

$$E(\ddot{y}_{it}|\mathbf{x}_i) = \ddot{m}_{it}(\boldsymbol{\beta}_o), \quad t = 1, \dots, T,$$

as suggested in the hint. To this end, write

$$y_{it} = c_i + m_{it}(\mathbf{x}_i, \boldsymbol{\beta}_o) + u_{it}, \quad E(u_{it}|\mathbf{x}_i, c_i) = 0, \quad t = 1, \dots, T.$$

Then subtracting off time averages gives

$$\begin{aligned}
\ddot{y}_{it} &= \ddot{m}_{it}(\boldsymbol{\beta}_o) + \ddot{u}_{it}, \\
\ddot{u}_{it} &\equiv u_{it} - T^{-1} \sum_{r=1}^T u_{ir}.
\end{aligned}$$

Because  $E(\ddot{u}_{it}|\mathbf{x}_i) = 0, t = 1, \dots, T$ , consistency follows generally by Problem 12.6. We do have to make an assumption that ensures that  $\boldsymbol{\beta}_o$  is identified, which restricts the way that time-constant variables can appear in  $m(\mathbf{x}_{it}, \boldsymbol{\beta})$ . (For example, additive time-constant variables

get swept away by the time demeaning.)

b. By the general theory of M-estimation, or by adapting either Problem 12.6 or 12.7, we can show

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) = \mathbf{A}_o^{-1} N^{-1/2} \sum_{i=1}^N \sum_{t=1}^T \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o)' \ddot{u}_{it} + o_p(1),$$

where

$$\mathbf{A}_o = T^{-1} \sum_{t=1}^T \mathbf{E} \left[ \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o)' \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o) \right]$$

is  $P \times P$  and  $P$  is the dimension of  $\boldsymbol{\beta}$ . (As part of the identification assumption, we would assume that  $\mathbf{A}_o$  is nonsingular.) As in the linear case, we can write, for each  $t$ ,

$$\sum_{t=1}^T \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o)' \ddot{u}_{it} = \sum_{t=1}^T \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o)' u_{it}.$$

Further,  $\text{Var}(\mathbf{y}_i | \mathbf{x}_i, c_i) = \sigma_o^2 \mathbf{I}_T$  is the same as  $\mathbf{E}(\mathbf{u}_i \mathbf{u}_i' | \mathbf{x}_i, c_i) = \sigma_o^2 \mathbf{I}_T$ , which implies

$$\begin{aligned} \mathbf{E}(u_{it}^2 | \mathbf{x}_i) &= \sigma_o^2 \\ \mathbf{E}(u_{it} u_{ir} | \mathbf{x}_i) &= 0, \quad t \neq r \end{aligned}$$

Therefore, by the usual iterated expectations argument,

$$\mathbf{E} \left[ u_{it}^2 \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o)' \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o) \right] = \sigma_o^2 \mathbf{E} \left[ \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o)' \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o) \right], \quad t = 1, \dots, T$$

and

$$\mathbf{E} \left[ u_{it} u_{ir} \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o)' \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o) \right] = 0, \quad t \neq r.$$

It follows that

$$\text{Var} \left( \sum_{t=1}^T \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o)' u_{it} \right) = \sigma_o^2 \left( \sum_{t=1}^T \mathbf{E} \left[ \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o)' \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o) \right] \right).$$

Therefore, under the given assumptions,

$$\text{Avar}\left[\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o)\right] = \sigma_o^2 \left( \sum_{t=1}^T \text{E}\left[\nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o)' \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\boldsymbol{\beta}_o)\right] \right)^{-1}.$$

As in the linear case, the tricky part is in estimating  $\sigma_o^2$ . We can apply virtually the same argument. Let  $\hat{u}_{it} = \ddot{y}_{it} - \ddot{m}_{it}(\hat{\boldsymbol{\beta}})$  for all  $i$  and  $t$ . Then a consistent estimator of  $\sigma_o^2$  is

$$\hat{\sigma}^2 = \frac{1}{[N(T-1) - P]} \sum_{i=1}^N \sum_{t=1}^T \hat{u}_{it}^2,$$

where the subtraction of  $P$  is not needed but is often used as an adjustment for estimation of  $\boldsymbol{\beta}_o$ . Estimation of  $\mathbf{A}_o$  gives

$$\hat{\mathbf{A}} = N^{-1} \sum_{i=1}^N \sum_{t=1}^T \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\hat{\boldsymbol{\beta}})' \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\hat{\boldsymbol{\beta}}).$$

Then,

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 \left( \sum_{i=1}^N \sum_{t=1}^T \left[ \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\hat{\boldsymbol{\beta}})' \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\hat{\boldsymbol{\beta}}) \right] \right)^{-1}.$$

c. A fully robust variance matrix estimator uses  $\hat{\mathbf{A}}$  and

$$\hat{\mathbf{B}} = N^{-1} \sum_{i=1}^N \sum_{t=1}^T \sum_{r=1}^T \hat{u}_{it} \hat{u}_{ir} \nabla_{\boldsymbol{\beta}} \ddot{m}_{it}(\hat{\boldsymbol{\beta}})' \nabla_{\boldsymbol{\beta}} \ddot{m}_{ir}(\hat{\boldsymbol{\beta}}),$$

which allows for arbitrary heteroskedasticity and serial correlation in  $\{u_{it} : t = 1, \dots, T\}$ . Then

$\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) = \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{-1} / N$ , as usual.

Remember that the estimator of  $\mathbf{A}_o$  relies on correct specification of the conditional mean; in its weakest form,  $\text{E}(\ddot{y}_{it} | \mathbf{x}_i) = \ddot{m}_{it}(\boldsymbol{\beta}_o)$ , which is implied by the model we started with,  $\text{E}(y_{it} | \mathbf{x}_i, c_i) = c_i + m(\mathbf{x}_{it}, \boldsymbol{\beta}_o)$ . If we want to allow the model to be misspecified we should use



the full Hessian,  $\nabla_{\beta} \ddot{m}_{it}(\hat{\beta})' \nabla_{\beta} \ddot{m}_{it}(\hat{\beta}) - \hat{u}_{it} \nabla_{\beta}^2 \ddot{m}_{it}(\hat{\beta})$  in place of  $\nabla_{\beta} \ddot{m}_{it}(\hat{\beta})' \nabla_{\beta} \ddot{m}_{it}(\hat{\beta})$ .

d. This is easy following the hint. For each  $i$  and given  $\beta$ ,  $\hat{c}_i(\beta)$  is just the intercept in the simple regression of  $y_{it}$  on 1,  $m_{it}(\beta)$ ,  $t = 1, \dots, T$ . Therefore,  $\hat{c}_i(\beta) = \bar{y}_i - \bar{m}_i(\beta)$ . Therefore, we can write problem (18.105), after concentrating out the  $c_i$ , as

$$\min_{\beta} \sum_{i=1}^N \sum_{t=1}^T \{y_{it} - [\bar{y}_i - \bar{m}_i(\beta)] - m_{it}(\beta)\}^2 = \min_{\beta} \sum_{i=1}^N \sum_{t=1}^T [y_{it} - \ddot{m}_{it}(\beta)]^2,$$

which is what we wanted to show. Note by treating the  $c_i$  as  $N$  parameters to estimate and using a standard degrees-of-freedom adjustment, solving (18.105) does yield the estimate  $\hat{\sigma}^2$  from part b when we use the sum of squared residuals over the degrees of freedom,  $NT - N - P = N(T - 1) - P$ .

**18.7.** a. First, for each  $t$ , the density of  $y_{it}$  given  $(\mathbf{x}_i = \mathbf{x}, c_i = c)$  is

$$f(y_t | \mathbf{x}, c; \beta_o) = \exp[-c \cdot m(\mathbf{x}_t, \beta_o)] [c \cdot m(\mathbf{x}_t, \beta_o)]^{y_t} / y_t!, \quad y_t = 0, 1, 2, \dots$$

Multiplying these together gives the joint density of  $(y_{i1}, \dots, y_{iT})$  given  $(\mathbf{x}_i = \mathbf{x}, c_i = c)$ .

Taking the log, plugging in the observed data for observation  $i$ , and dropping the factorial term gives

$$\sum_{t=1}^T \{-c_i m(\mathbf{x}_{it}, \beta) + y_{it} [\log(c_i) + \log(m(\mathbf{x}_{it}, \beta))]\}.$$

b. Taking the derivative of  $\ell_i(c_i, \beta)$  with respect to  $c_i$ , setting the result to zero, and rearranging gives

$$(n_i / c_i) = \sum_{t=1}^T m(\mathbf{x}_{it}, \beta).$$

Letting  $c_i(\beta)$  denote the solution as a function of  $\beta$ , we have  $c_i(\beta) = n_i / M_i(\beta)$ , where

$M_i(\boldsymbol{\beta}) \equiv \sum_{t=1}^T m(\mathbf{x}_{it}, \boldsymbol{\beta})$ . The second order sufficient condition for a maximum is easily seen to hold.

c. Plugging the solution from part b into  $\ell_i(c_i, \boldsymbol{\beta})$  gives

$$\begin{aligned}\ell_i[c_i(\boldsymbol{\beta}), \boldsymbol{\beta}] &= -[n_i/M_i(\boldsymbol{\beta})]M_i(\boldsymbol{\beta}) + \sum_{t=1}^T y_{it} \{\log[n_i/M_i(\boldsymbol{\beta})] + \log[m(\mathbf{x}_{it}, \boldsymbol{\beta})]\} \\ &= -n_i + n_i \log(n_i) + \sum_{t=1}^T y_{it} \{\log[m(\mathbf{x}_{it}, \boldsymbol{\beta})/M_i(\boldsymbol{\beta})]\} \\ &= \sum_{t=1}^T y_{it} \log[p_t(\mathbf{x}_{it}, \boldsymbol{\beta})] + (n_i - 1) \log(n_i),\end{aligned}$$

because  $p_t(\mathbf{x}_{it}, \boldsymbol{\beta}) \equiv m(\mathbf{x}_{it}, \boldsymbol{\beta})/M_i(\boldsymbol{\beta})$ ; see equation (18.89).

d. From part c it follows that if we maximize  $\sum_{i=1}^N \ell_i(c_i, \boldsymbol{\beta})$  with respect to  $(c_i, \dots, c_N)$  – that is, we concentrate out these parameters – we get exactly  $\sum_{i=1}^N \ell_i[c_i(\boldsymbol{\beta}), \boldsymbol{\beta}]$ . Except for the term  $\sum_{i=1}^N (n_i - 1) \log(n_i)$  – which does not depend on  $\boldsymbol{\beta}$  – this is exactly the conditional log-likelihood for the conditional multinomial distribution obtained in Section 18.7.4. Therefore, this is another case where treating the  $c_i$  as parameters to be estimated leads us to a  $\sqrt{N}$ -consistent, asymptotically normal estimator of  $\boldsymbol{\beta}_o$ .

**18.8.** a. Generally, there is no simple way to recover  $E(y|\mathbf{x})$  from  $E\{\log[y/(1-y)]|\mathbf{x}\}$ . In particular, if  $E(w|\mathbf{x}) = \mathbf{x}\boldsymbol{\alpha}$ , it is *not* true that  $E(y|\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\alpha})/[1 + \exp(\mathbf{x}\boldsymbol{\alpha})]$ . In other words, we cannot simply “undo” the log-odds transformation any more than we can undo any nonlinear transformation when trying to recover conditional means.

If we make stronger assumptions, we can recover  $E(y|\mathbf{x})$  from  $E(w|\mathbf{x})$ . Suppose we write  $w = \mathbf{x}\boldsymbol{\alpha} + v$  and *assume* that  $v$  is independent of  $\mathbf{x}$ . Assume for simplicity that  $v$  is continuous with density  $g(\cdot)$ . Then

$$E(y|\mathbf{x} = \mathbf{x}^o) = \int_{-\infty}^{\infty} \exp(\mathbf{x}^o \boldsymbol{\alpha} + v) / [1 + \exp(\mathbf{x}^o \boldsymbol{\alpha} + v)] g(v) dv.$$

If we parameterize  $g(\cdot)$  – say,  $g(\cdot; \boldsymbol{\rho})$  – and we have a consistent estimator of  $\boldsymbol{\rho}$ , then

$$\hat{E}(y|\mathbf{x} = \mathbf{x}^o) = \int_{-\infty}^{\infty} \exp(\mathbf{x}^o \hat{\boldsymbol{\alpha}} + v) / [1 + \exp(\mathbf{x}^o \hat{\boldsymbol{\alpha}} + v)] g(v; \hat{\boldsymbol{\rho}}) dv,$$

where  $\hat{\boldsymbol{\alpha}}$  could be the OLS estimator from regressing  $w_i$  on  $\mathbf{x}_i$  or the maximum likelihood estimator based on  $D(w|\mathbf{x})$ . If  $D(v|\mathbf{x})$  is assumed to be normal then OLS is MLE. Even if we specify  $g(\cdot; \boldsymbol{\rho})$  to be a mean-zero normal distribution, obtaining the integral is cumbersome.

There is a simpler approach that is also more robust. If we just maintain that  $v$  and  $\mathbf{x}$  are independent then, by the law of large numbers,  $E(y|\mathbf{x} = \mathbf{x}^o)$  for a given vector  $\mathbf{x}^o$  is consistently estimated by

$$N^{-1} \sum_{i=1}^N \exp(\mathbf{x}^o \boldsymbol{\alpha} + v_i) / [1 + \exp(\mathbf{x}^o \boldsymbol{\alpha} + v_i)],$$

where we can think of drawing random samples  $\{(\mathbf{x}_i, v_i) : i = 1, 2, \dots, N\}$ . Because we cannot observe  $v_i$ , and we do not know  $\boldsymbol{\alpha}$ , we operationalize this formula by replacing  $\boldsymbol{\alpha}$  with  $\hat{\boldsymbol{\alpha}}$ , including computing residuals  $\hat{v}_i = w_i - \mathbf{x}_i \hat{\boldsymbol{\alpha}}$ ,  $i = 1, \dots, N$ . Then

$$\hat{E}(y|\mathbf{x} = \mathbf{x}^o) = N^{-1} \sum_{i=1}^N \exp(\mathbf{x}^o \hat{\boldsymbol{\alpha}} + \hat{v}_i) / [1 + \exp(\mathbf{x}^o \hat{\boldsymbol{\alpha}} + \hat{v}_i)].$$

This is an example of Duan's (1983) “smearing estimate.” This estimator is consistent under the assumptions given – which do not require a full distribution, but do include independence between  $v$  and  $\mathbf{x}$ . Obtaining analytical standard errors can be done by following Problem 12.17. Bootstrapping is also valid. Unfortunately, this approach does not work if  $y$  can take on the boundary values zero or one.

The above integral for  $E(y|\mathbf{x} = \mathbf{x}^o)$  can be written as  $E(y|\mathbf{x} = \mathbf{x}^o) = r(\mathbf{x}^o\boldsymbol{\alpha})$  and so, if  $v$  and  $\mathbf{x}$  are independent, then  $[\partial E(y|\mathbf{x})/\partial x_j]/[\partial E(y|\mathbf{x})/\partial x_h] = \alpha_j/\alpha_h$ : for continuous explanatory variables, the ratio of the partial effects equals the ratio of the parameters in the linear model for  $w = \log[y/(1 - y)]$ .

b. The functional form  $E(y|\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta})/[1 + \exp(\mathbf{x}\boldsymbol{\beta})]$ , and that implied by the assumptions in part a, are generally incompatible. However, as mentioned above, independence between  $v$  and  $\mathbf{x}$  in  $\log[y/(1 - y)] = \mathbf{x}\boldsymbol{\alpha} + v$  implies that  $\alpha_j/\alpha_h$  is the ratio of the partial effects of continuous explanatory variables  $x_j$  and  $x_h$ . In the fractional logit model, the ratio of partial effects is  $\beta_j/\beta_h$ . Therefore, it can make sense to compare ratios of coefficients on continuous explanatory variables across the two procedures. But the magnitudes themselves are not generally comparable.

c. Because we have a full distribution of  $y$  given  $\mathbf{x}$ , we should use maximum likelihood, just as described in Section 17.7.

d. The functional form for  $E(y|\mathbf{x})$  – as a function of the parameters  $\boldsymbol{\gamma}$  and  $\sigma^2$  – is given in equation (17.66) where we set  $a_1 = 0$ ,  $a_2 = 1$ , with the obvious change in notation:

$$\begin{aligned} E(y|\mathbf{x}) &= \{\Phi[(1 - \mathbf{x}\boldsymbol{\gamma})/\sigma] - \Phi[-(\mathbf{x}\boldsymbol{\gamma})/\sigma]\}\mathbf{x}\boldsymbol{\gamma} + \sigma\{\phi[((1 - \mathbf{x}\boldsymbol{\gamma})/\sigma] - \phi[-(\mathbf{x}\boldsymbol{\gamma})/\sigma]\} \\ &\quad + \Phi[-(1 - \mathbf{x}\boldsymbol{\gamma})/\sigma] \\ &= \{\Phi[(1 - \mathbf{x}\boldsymbol{\gamma})/\sigma] - \Phi[-(\mathbf{x}\boldsymbol{\gamma})/\sigma]\}\mathbf{x}\boldsymbol{\gamma} + \sigma\{\phi[((1 - \mathbf{x}\boldsymbol{\gamma})/\sigma] - \phi[-(\mathbf{x}\boldsymbol{\gamma})/\sigma]\} \\ &\quad + 1 - \Phi[(1 - \mathbf{x}\boldsymbol{\gamma})/\sigma]. \end{aligned}$$

This gives yet a different functional form. Nevertheless, it is easily seen from equation (17.67) that

$$[\partial E(y|\mathbf{x})/\partial x_j]/[\partial E(y|\mathbf{x})/\partial x_h] = \gamma_j/\gamma_h,$$

so that ratios of the coefficients on continuous variables can be compared with those from part

b.

e. Because part b only specified a conditional mean, it does not make much sense to compare the Bernoulli quasi-log-likelihood with the Tobit log-likelihood. If we are mainly interested in  $E(y|\mathbf{x})$  – which part b essentially maintains – it makes sense to base comparisons on goodness-of-fit for  $E(y|\mathbf{x})$ . For each approach, we can compute a squared correlation between the  $y_i$  and the  $\hat{E}(y_i|\mathbf{x}_i)$ , where the conditional expectations are estimated using each approach. Or, we can use a sum-of-squared residuals version (and possibly adjust for degrees-of-freedom because the Tobit model has an extra mean parameter,  $\sigma$ ).

f. We would not expect to get similar answers for the full sample – which includes observations with  $y_i = 0$  – and the subsample that excludes  $y_i = 0$  (unless the fraction of excluded observations is small). Clearly, we cannot have both

$E(y|\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta})/[1 + \exp(\mathbf{x}\boldsymbol{\beta})]$  and  $E(y|\mathbf{x}, y > 0) = \exp(\mathbf{x}\boldsymbol{\delta})/[1 + \exp(\mathbf{x}\boldsymbol{\delta})]$ . Moreover, there is no reason to expect the best fits to yield roughly the same parameter estimates.

g. Because we have assumed  $E(y|\mathbf{x}, y > 0) = \exp(\mathbf{x}\boldsymbol{\delta})/[1 + \exp(\mathbf{x}\boldsymbol{\delta})]$ , we consistently estimate  $\boldsymbol{\delta}$  using the sample for which  $0 < y_i < 1$ , provided we use the Bernoulli QMLE (or NLS or weighted NLS). There is no bias from excluding the  $y_i = 0$  observations because we have specified the mean for the subpopulation with  $y > 0$ . (We discuss sample selection issues in Chapter 19.)

h. We would use a two-part model. Let  $\hat{\boldsymbol{\delta}}$  be the Bernoulli QMLE from part g, using observations for which  $y_i > 0$ . To estimate  $\boldsymbol{\eta}$ , we run a binary response model using the binary variable  $r_i = 1[y_i > 0]$ . Then  $P(r_i = 1|\mathbf{x}_i) = G(\mathbf{x}_i\boldsymbol{\eta})$ . Probably we would use a probit or logit model. Then,

$$\begin{aligned}\hat{E}(y_i|\mathbf{x}_i) &= \hat{P}(y_i > 0|\mathbf{x}_i) \cdot \hat{E}(y_i|\mathbf{x}_i, y_i > 0) \\ &= G(\mathbf{x}_i\hat{\boldsymbol{\eta}}) \cdot \{\exp(\mathbf{x}_i\hat{\boldsymbol{\delta}})/[1 + \exp(\mathbf{x}_i\hat{\boldsymbol{\delta}})]\}.\end{aligned}$$

**18.9.** a. The Stata output follows. I first convert the dependent variable to be in  $[0, 1]$ , rather than  $[0, 100]$ ; this is needed to estimate a fractional response model.

The coefficient on *ACT* means that five more points on the ACT test, other things equal, is associated with a lower attendance rate of about  $.017(5) = .085$ , or 8.5 percentage points. For *priGPA*, another point on the GPA (a large change) is associated with an attendance rate roughly 18.2 percentage points higher.

Twelve of the fitted values are bigger than one. This is not surprising because almost 10 percent of the students have perfect attendance rates.

```
. use attend
. sum atndrte
```

Variable	Obs	Mean	Std. Dev.	Min	Max
atndrte	680	81.70956	17.04699	6.25	100

```
. replace atndrte = atndrte/100
(680 real changes made)
```

```
. reg atndrte ACT priGPA frosh soph
```

Source	SS	df	MS	Number of obs =	680
Model	5.95396289	4	1.48849072	F( 4, 675) =	72.92
Residual	13.7777696	675	.020411511	Prob > F =	0.0000
Total	19.7317325	679	.029059989	R-squared =	0.3017
				Adj R-squared =	0.2976
				Root MSE =	.14287

atndrte	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
ACT	-.0169202	.001681	-10.07	0.000	-.0202207	-.0136196
priGPA	.1820163	.0112156	16.23	0.000	.1599947	.2040379
frosh	.0517097	.0173019	2.99	0.003	.0177377	.0856818
soph	.0110085	.014485	0.76	0.448	-.0174327	.0394496
_cons	.7087769	.0417257	16.99	0.000	.6268492	.7907046

```
. predict atndrteh_lin
(option xb assumed; fitted values)
```

```
. sum atndrteh_lin
```

Variable	Obs	Mean	Std. Dev.	Min	Max
atndrteh_lin	680	.8170956	.0936415	.4846666	1.086443

```
. count if atndrteh_lin > 1
12
```

```
. count if atndrte == 1
66
```

b. The GLM standard errors are given in the output. Note that  $\hat{\sigma} \approx .0161$ . In other words, the usual MLE standard errors, obtained, say, from the expected Hessian of the quasi-log-likelihood, are much too *large*. The standard errors that account for  $\sigma^2 < 1$  are given by the GLM output. (If you omit the `sca(x2)` option in the `glm` command, you get the usual MLE standard errors.)

```
. glm atndrte ACT priGPA frosh soph, family(binomial) link(logit) sca(x2)
note: atndrte has noninteger values
```

Generalized linear models		No. of obs	=	680
Optimization : ML		Residual df	=	675
		Scale parameter	=	
Deviance	= 87.81698799	(1/df) Deviance	=	.1300992
Pearson	= 85.57283238	(1/df) Pearson	=	.1267746

Variance function:	V(u) = u*(1-u/1)	[Binomial]
Link function :	g(u) = ln(u/(1-u))	[Logit]

Log likelihood	= -223.6493665	AIC	= .6724981
		BIC	= -4314.596

atndrte	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval
ACT	-.1113802	.0113217	-9.84	0.000	-.1335703 -.0891901
priGPA	1.244375	.0771321	16.13	0.000	1.093199 1.395552
frosh	.3899318	.113436	3.44	0.001	.1676013 .6122622
soph	.0928127	.0944066	0.98	0.326	-.0922209 .2778463
_cons	.7621699	.2859966	2.66	0.008	.201627 1.322713

(Standard errors scaled using square root of Pearson X2-based dispersion.)

```
. di (.1268)^2
.01607824
```

c. Because the coefficient on *ACT* is negative, we know that an increase in ACT score,

holding year and prior GPA fixed, actually reduces predicted attendance rate. The calculation below shows that for  $priGPA = 3.0$  and  $frosh = soph = 0$ , when  $ACT$  increases from 25 to 30, the estimated fall in  $atndrte$  is about .087, or 8.7 percentage points. This is very similar to the estimate using the linear model – 8.5 percentage points – which is the same for any values of the explanatory variables.

```
. di exp(_b[_cons] + _b[ACT]*30 + _b[priGPA]*3)/(1 + exp(_b[_cons] + _b[ACT]*
+ _b[priGPA]*3)) - exp(_b[_cons] + _b[ACT]*25 + _b[priGPA]*3)
/(1 + exp(_b[_cons] + _b[ACT]*25 + _b[priGPA]*3))
-.08671822
```

d. The  $R$ -squared for the linear model is about .302. For the logistic functional form, I computed the squared correlation between  $atndrte_i$  and  $\hat{E}(atndrte_i | \mathbf{x}_i)$ . This  $R$ -squared is about .328, and so the logistic functional form does fit better than the linear model. And, remember that the parameters in the logistic functional form are not chosen to maximize an  $R$ -squared; the linear model coefficients are chosen to maximize  $R$ -squared given the set of explanatory variables.

```
. predict atndrteh_log
(option mu assumed; predicted mean atndrte)

. corr atndrte atndrteh_log
(obs=680)
```

	atndrte	atndrteh_log
atndrte	1.0000	
atndrteh_log	0.5725	1.0000

```
. di .5725^2
.32775625
```

**18.10.** a. The pooled Poisson estimates, with the usual pooled standard errors that assume a unit variance-mean ratio and dynamic completeness of the conditional mean, are given below. Using these nonrobust standard errors, all lags except the first are significantly different from zero.



```
. use patent
```

```
. poisson patents y77-y81 lrnd lrnd_1 lrnd_2 lrnd_3 lrnd_4
```

```
Poisson regression                                Number of obs   =      1356
                                                    LR chi2(10)    =    68767.04
                                                    Prob > chi2     =      0.0000
Log likelihood = -12194.868                        Pseudo R2      =      0.7382
```

patents	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	Interval
y77	-.0732934	.0190128	-3.85	0.000	-.1105578	-.0360291
y78	-.227293	.0196925	-11.54	0.000	-.2658896	-.1886965
y79	-.36251	.0196912	-18.41	0.000	-.4011041	-.3239159
y80	-.7066175	.0211325	-33.44	0.000	-.7480365	-.6651985
y81	-2.115567	.0331249	-63.87	0.000	-2.18049	-2.050643
lrnd	.4406223	.0425948	10.34	0.000	.357138	.5241066
lrnd_1	.0767312	.0635969	1.21	0.228	-.0479165	.2013788
lrnd_2	.2452529	.0622048	3.94	0.000	.1233337	.3671721
lrnd_3	-.1557527	.0630881	-2.47	0.014	-.2794031	-.0321023
lrnd_4	.1619174	.0469008	3.45	0.001	.0699936	.2538412
_cons	1.157326	.0191835	60.33	0.000	1.119727	1.194925

b. The standard errors computed in part a can be wrong for at least two reasons. The first is that the conditional variance,  $\text{Var}(y_{it}|\mathbf{x}_{it})$ , may not equal the conditional mean,  $E(y_{it}|\mathbf{x}_{it})$ , where  $\mathbf{x}_{it}$  contains the current and lagged R&D spending variables. The second is that the mean may not be dynamically complete in the sense that

$$E(y_{it}|\mathbf{x}_{it}) \neq E(y_{it}|\mathbf{x}_{it}, y_{i,t-1}, \mathbf{x}_{i,t-1}, \dots).$$

A failure of dynamic completeness generally leads to serial correlation in the implied error terms, and cause the score of the partial quasi-log-likelihood function to be serially correlated.

A third reason the standard errors might not be valid is they use the expected Hessian form of the asymptotic variance. This form is incorrect if the conditional mean is misspecified.

c. The estimates below give  $\hat{\sigma} \approx 4.14$ , which shows that, even if we assume a constant variance-mean ratio and dynamic completeness of the conditional mean, we need to multiply all Poisson standard errors by just over four.

Now only the contemporaneous R&D variable is significant; none of the lags has a  $t$

statistic above one.

```
. glm patents y77-y81 lrnd lrnd_1 lrnd_2 lrnd_3 lrnd_4, family(poisson)
    link(log) sca(x2)
```

Generalized linear models		No. of obs	=	1356
Optimization	: ML	Residual df	=	1345
		Scale parameter	=	
Deviance	= 20618.28952	(1/df) Deviance	=	15.32958
Pearson	= 23082.45413	(1/df) Pearson	=	17.16168
Variance function:	V(u) = u	[Poisson]		
Link function	: g(u) = ln(u)	[Log]		
		AIC	=	18.00276
Log likelihood	= -12194.86797	BIC	=	10917.75

patents	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
y77	-.0732934	.0787636	-0.93	0.352	-.2276672	.0810803
y78	-.227293	.0815794	-2.79	0.005	-.3871858	-.0674003
y79	-.36251	.0815742	-4.44	0.000	-.5223925	-.2026275
y80	-.7066175	.087545	-8.07	0.000	-.8782025	-.5350325
y81	-2.115567	.1372256	-15.42	0.000	-2.384524	-1.846609
lrnd	.4406223	.176456	2.50	0.013	.0947748	.7864698
lrnd_1	.0767312	.2634608	0.29	0.771	-.4396424	.5931048
lrnd_2	.2452529	.2576938	0.95	0.341	-.2598177	.7503235
lrnd_3	-.1557527	.2613529	-0.60	0.551	-.6679949	.3564895
lrnd_4	.1619174	.1942941	0.83	0.405	-.2188921	.5427269
_cons	1.157326	.0794708	14.56	0.000	1.001566	1.313086

(Standard errors scaled using square root of Pearson X2-based dispersion.)

```
. di sqrt(17.16)
4.142463
```

d. The QLR statistic is just the usual LR statistic divided by  $\hat{\sigma}^2 = 17.17$ . The value of the unrestricted log-likelihood is  $\mathcal{L}_{ur} = -12,194.87$ . The value of the restricted log-likelihood (without any of the lags), using the same set of years in estimation (1976 to 1981), is  $\mathcal{L}_r = -12,252.37$ . Therefore,

$$QLR = 2 \cdot (12,252.37 - 12,194.87)/17.17 = 6.70.$$

With four degrees of freedom in a chi-square distribution, this leads to  $p$ -value = .153. The lags are jointly insignificant at the usual 5% level. The usual LR statistic is 115, which (incorrectly)

implies very strong statistical significance for the lags.

e. The Stata results are blow. With the fully robust standard errors, the contemporaneous term and the second lag are marginally significantly. The robust Wald test for the exclusion of the four lags gives  $p$ -value = .494. The fully robust standard errors are clearly smaller than the Poisson MLE standard errors, but they are actually smaller in some cases than the GLM standard errors from part c. The four lags are joint insignificant.

```
. glm patents y77-y81 lrnd lrnd_1 lrnd_2 lrnd_3 lrnd_4, family(poisson)
    link(log) robust cluster(cusip)
```

Generalized linear models		No. of obs	=	1356
Optimization	: ML	Residual df	=	1345
		Scale parameter	=	
Deviance	= 20618.28952	(1/df) Deviance	=	15.32958
Pearson	= 23082.45413	(1/df) Pearson	=	17.16168
Variance function: V(u) = u		[Poisson]		
Link function : g(u) = ln(u)		[Log]		
		AIC	=	18.00276
Log pseudolikelihood = -12194.86797		BIC	=	10917.75

(Std. Err. adjusted for 226 clusters in cusip)

patents	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
y77	-.0732934	.0317955	-2.31	0.021	-.1356115	-.0109754
y78	-.227293	.0499251	-4.55	0.000	-.3251445	-.1294416
y79	-.36251	.0681543	-5.32	0.000	-.49609	-.22893
y80	-.7066175	.0667816	-10.58	0.000	-.837507	-.575728
y81	-2.115567	.1140381	-18.55	0.000	-2.339077	-1.892056
lrnd	.4406223	.2409156	1.83	0.067	-.0315637	.9128083
lrnd_1	.0767312	.1228435	0.62	0.532	-.1640376	.3175
lrnd_2	.2452529	.1411443	1.74	0.082	-.0313848	.5218906
lrnd_3	-.1557527	.2160959	-0.72	0.471	-.579293	.2677875
lrnd_4	.1619174	.2679931	0.60	0.546	-.3633395	.6871743
_cons	1.157326	.2061445	5.61	0.000	.7532903	1.561362

```
. test lrnd_1 lrnd_2 lrnd_3 lrnd_4
```

```
( 1) [patents]lrnd_1 = 0
( 2) [patents]lrnd_2 = 0
( 3) [patents]lrnd_3 = 0
( 4) [patents]lrnd_4 = 0
```

```
      chi2( 4) =      3.40
Prob > chi2 =      0.4937
```

f. The estimated long run elasticity is about  $.441 + .077 + .245 - .156 + .162 = .769$ . The `lincom` command in Stata provides a simple way to obtain a fully robust standard error. Its fully robust standard error is about  $.072$ , which gives a 95% confidence interval from about  $.627$  to  $.910$ . As is often the case in distributed lag models, we cannot estimate the lag distribution very precisely but we can get a fairly precise estimate of the long run effect.

```
. lincom lrnd + lrnd_1 + lrnd_2 + lrnd_3 + lrnd_4
```

```
( 1)  [patents]lrnd + [patents]lrnd_1 + [patents]lrnd_2 + [patents]lrnd_3
      + [patents]lrnd_4 = 0
```

patents	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
(1)	.768771	.0722693	10.64	0.000	.6271258	.9104162

g. The fixed effects Poisson estimates are given below. The contemporaneous spending term and second lag have much smaller effects now, while lags three and four become larger and even statistically significant – but with the third lag still having a large, negative coefficient. When we use the fully robust standard errors, only the second lag is statistically significant at conventional levels, although the third and fourth lags are close.

The estimated long-run elasticity is now only  $.261$  and it is, at best, marginally significant with  $t = 1.60$ .

```
. xtpqml patents y77-y81 lrnd lrnd_1 lrnd_2 lrnd_3 lrnd_4, fe
note: 8 groups (48 obs) dropped because of all zero outcomes
```

```
Conditional fixed-effects Poisson regression    Number of obs    =    1308
Group variable: cusip                          Number of groups  =     218

                                                Obs per group:  min =
                                                avg =    6.
                                                max =

Log likelihood = -2423.7694                    Wald chi2(10)     =   3002.51
                                                Prob > chi2       =    0.0000
```

patents	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
-----+-----						

```

      y77 | -.0210069   .0204558   -1.03   0.304   -.0610995   .0190856
      y78 | -.108368   .0251005   -4.32   0.000   -.157564   -.059172
      y79 | -.1721306   .0306902   -5.61   0.000   -.2322822   -.1119789
      y80 | -.4468227   .039243   -11.39   0.000   -.5237375   -.3699079
      y81 | -1.797958   .0547882   -32.82   0.000   -1.905341   -1.690575
      lrnd | .0492403   .0558275   0.88   0.378   -.0601795   .15866
      lrnd_1 | .0512096   .0666844   0.77   0.443   -.0794894   .1819086
      lrnd_2 | .130944   .0662164   1.98   0.048   .0011622   .2607259
      lrnd_3 | -.1909907   .0714669   -2.67   0.008   -.3310632   -.0509182
      lrnd_4 | .2201799   .0703992   3.13   0.002   .0821999   .3581599
-----
Calculating Robust Standard Errors...
-----
      patents |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
patents
      y77 | -.0210069   .026186   -0.80   0.422   -.0723306   .0303168
      y78 | -.108368   .055447   -1.95   0.051   -.2170422   .0003062
      y79 | -.1721306   .071949   -2.39   0.017   -.313148   -.0311131
      y80 | -.4468227   .0829316   -5.39   0.000   -.6093657   -.2842797
      y81 | -1.797958   .1380887  -13.02   0.000   -2.068607   -1.527309
      lrnd | .0492403   .0868099   0.57   0.571   -.120904   .2193845
      lrnd_1 | .0512096   .0600491   0.85   0.394   -.0664845   .1689038
      lrnd_2 | .130944   .0592739   2.21   0.027   .0147694   .2471187
      lrnd_3 | -.1909907   .1066283   -1.79   0.073   -.3999783   .0179968
      lrnd_4 | .2201799   .1431273   1.54   0.124   -.0603446   .5007043
-----
Wald chi2(10) =    366.83                                Prob > chi2 =    0.0000

. lincom lrnd + lrnd_1 + lrnd_2 + lrnd_3 + lrnd_4

( 1)  [patents]lrnd + [patents]lrnd_1 + [patents]lrnd_2 + [patents]lrnd_3
      + [patents]lrnd_4 = 0
-----
      patents |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      (1) | .2605831   .1632377   1.60   0.110   -.059357   .5805231
-----

```

**18.11.** a. For each  $t$ , the density is

$$f_t(y_t|\mathbf{x}_i, c_i) = \exp(-c_i m_{it}) m_{it}^{y_t} / y_t!, \quad y_t = 0, 1, 2, \dots$$

Under the conditional independence assumption, the joint density of  $(y_{i1}, \dots, y_{iT})$  given  $(\mathbf{x}_i, c_i)$

is

$$\begin{aligned}
f(y_1, \dots, y_T | \mathbf{x}_i, c_i) &= \prod_{t=1}^T [\exp(-c_i m_{it}) (c_i m_{it})^{y_t} / y_t!] \\
&= \left( \prod_{t=1}^T m_{it}^{y_t} / y_t! \right) c_i^s \exp(-c_i M_i),
\end{aligned}$$

where  $M_i \equiv m_{i1} + \dots + m_{iT}$  and  $s = y_1 + \dots + y_T$ , for all nonnegative integers  $\{y_t : t = 1, \dots, T\}$ .

b. To obtain the density of  $(y_{i1}, \dots, y_{iT})$  given  $\mathbf{x}_i$  – say  $g(y_{i1}, \dots, y_{iT} | \mathbf{x}_i)$  – we integrate out with respect to the distribution of  $c_i$  (because  $c_i$  is independent of  $\mathbf{x}_i$ ). Therefore,

$$g(y_1, \dots, y_T | \mathbf{x}_i) = \left( \prod_{t=1}^T m_{it}^{y_t} / y_t! \right) \int_0^\infty c^s \exp(-c M_i) [\delta^\delta / \Gamma(\delta)] c^{\delta-1} \exp(-\delta c) dc.$$

Next, we follow the hint, noting that the general  $\text{Gamma}(\alpha, \beta)$  density has the form

$h(c) = [\beta^\alpha / \Gamma(\alpha)] c^{\alpha-1} \exp(-\beta c)$ . Now

$$\begin{aligned}
\int_0^\infty c^s \exp(-c M_i) [\delta^\delta / \Gamma(\delta)] c^{\delta-1} \exp(-\delta c) dc &= \int_0^\infty [\delta^\delta / \Gamma(\delta)] c^{(s+\delta-1)} \exp[-(M_i + \delta)c] dc \\
&= [\delta^\delta / \Gamma(\delta)] [\Gamma(s + \delta) / (M_i + \delta)^{(s+\delta)}] \\
&\quad \cdot \int_0^\infty [(M_i + \delta)^{(s+\delta)} / \Gamma(s + \delta)] c^{(s+\delta-1)} \exp[-(M_i + \delta)c] dc,
\end{aligned}$$

and the integrand is easily seen to be the  $\text{Gamma}(s + \delta, M_i + \delta)$  density, and so it integrates to unity. Therefore, we have shown

$$g(y_1, \dots, y_T | \mathbf{x}_i) = \left( \prod_{t=1}^T m_{it}^{y_t} / y_t! \right) [\delta^\delta / \Gamma(\delta)] [\Gamma(s + \delta) / (M_i + \delta)^{(s+\delta)}]$$

for all nonnegative integers  $\{y_t : t = 1, \dots, T\}$ .

**18.12.** a. First, the density of  $y_{it}$  given  $(\mathbf{x}_i, c_i)$  is

$$f_t(y_t | \mathbf{x}_i, c_i) = [(1/c_i)^{m_{it}} / \Gamma(m_{it})] y_t^{(m_{it}-1)} \exp[-(1/c_i)y_t], \quad y_t > 0.$$

Following the hint, the density of the sum,  $s_i$ , given  $(\mathbf{x}_i, c_i)$  is

$$g(s|\mathbf{x}_i, c_i) = [(1/c_i)^{M_i}/\Gamma(M_i)]s^{(M_i-1)} \exp[-(1/c_i)s], \quad s > 0,$$

where  $M_i = m_{i1} + \dots + m_{iT}$ . Therefore, the density of  $(y_{i1}, \dots, y_{iT})$  given  $(s_i = s, \mathbf{x}_i, c_i)$  is

$$\begin{aligned} & \left( [(1/c_i)^{m_{i1}}/\Gamma(m_{i1})] y_1^{(m_{i1}-1)} \exp[-(1/c_i)y_1] \right) \cdots \left( [(1/c_i)^{m_{iT-1}}/\Gamma(m_{iT-1})] y_{T-1}^{(m_{iT-1}-1)} \exp[-(1/c_i)y_{T-1}] \right) \\ & \cdot [(1/c_i)^{m_{iT}}/\Gamma(m_{iT})] (s - y_1 - \dots - y_{T-1})^{m_{iT}-1} \exp[-(1/c_i)(s - y_1 - \dots - y_{T-1})] / g(s|\mathbf{x}_i, c_i) \\ & = (1/c_i)^{M_i} \left( \prod_{t=1}^T \Gamma(m_{it}) \right)^{-1} \left( \prod_{t=1}^T y_t^{(m_{it}-1)} \right) \exp[-(1/c_i)s] / [(1/c_i)^{M_i}/\Gamma(M_i)] s^{(M_i-1)} \exp[-(1/c_i)s] \\ & = \Gamma(M_i) \left( \prod_{t=1}^T \Gamma(m_{it}) \right)^{-1} \left( \prod_{t=1}^T y_t^{(m_{it}-1)} \right) / s^{(M_i-1)}, \end{aligned}$$

which is what we wanted to show. Note how  $c_i$  has dropped out of the density.

b. The conditional log-likelihood for observation  $i$  is

$$\begin{aligned} \ell(\boldsymbol{\beta}; y_{i1}, \dots, y_{iT}, \mathbf{x}_i) &= \log\{\Gamma[M_i(\boldsymbol{\beta})]\} - \sum_{t=1}^T \log\{\Gamma[m_{it}(\boldsymbol{\beta})]\} \\ &\quad + \sum_{t=1}^T [m_{it}(\boldsymbol{\beta}) - 1] \log(y_{it}) - [M_i(\boldsymbol{\beta}) - 1] \log(y_{i1} \cdots y_{iT}), \end{aligned}$$

where  $m_{it}(\boldsymbol{\beta}) = m_t(\mathbf{x}_i, \boldsymbol{\beta})$  and  $M_i = \sum_{t=1}^T m_{it}(\boldsymbol{\beta})$ . We can sum across all  $i$  and maximize the resulting log-likelihood with respect to  $\boldsymbol{\beta}$  to obtain the fixed effects gamma estimator. The asymptotic theory is standard, provided the regression functions are smooth functions of  $\boldsymbol{\beta}$  and depend on the covariates in such a way that  $\boldsymbol{\beta}_o$  is identified.

**18.13.** a. Plug in the data, a generic value  $\boldsymbol{\beta}$ , and take the natural log:

$$\ell_i(\boldsymbol{\beta}) = \mathbf{x}_i \boldsymbol{\beta} + \log(y_i) [\exp(\mathbf{x}_i \boldsymbol{\beta}) - 1].$$

Notice that this is not a member of the linear exponential family (because it is  $\log(y_i)$ , not  $y_i$ , that appears).

b. The gradient is

$$\nabla_{\beta} \ell_i(\beta) = \mathbf{x}_i + \log(y_i) \mathbf{x}_i \exp(\mathbf{x}_i \beta)$$

and taking the transpose gives

$$\begin{aligned} \mathbf{s}_i(\beta) &= \mathbf{x}_i' + \log(y_i) \mathbf{x}_i' \exp(\mathbf{x}_i \beta) \\ &= \mathbf{x}_i' [1 + \log(y_i) \exp(\mathbf{x}_i \beta)]. \end{aligned}$$

c. Because  $0 < y_i < 1$ ,  $\log(y_i) < 0$  for all  $i$ . Therefore, we know  $E[\log(y_i)|\mathbf{x}_i] < 0$  for any outcome  $\mathbf{x}_i$ .

d. We use part c:

$$\begin{aligned} E[\mathbf{s}_i(\beta_o)|\mathbf{x}_i] &= \mathbf{x}_i' \{1 + E[\log(y_i)|\mathbf{x}_i] \exp(\mathbf{x}_i \beta_o)\} \\ &= \mathbf{x}_i' [1 - \exp(-\mathbf{x}_i \beta_o) \exp(\mathbf{x}_i \beta_o)] = \mathbf{0}. \end{aligned}$$

e. Using  $\mathbf{s}_i(\beta) = \mathbf{x}_i' [1 + \log(y_i) \exp(\mathbf{x}_i \beta)]$ , the Hessian is

$$\mathbf{H}_i(\beta) = \nabla_{\beta} \mathbf{s}_i(\beta) = \mathbf{x}_i' \mathbf{x}_i \log(y_i) \exp(\mathbf{x}_i \beta)$$

and so

$$\begin{aligned} E[\mathbf{H}_i(\beta_o)|\mathbf{x}_i] &= \mathbf{x}_i' \mathbf{x}_i E[\log(y_i)|\mathbf{x}_i] \exp(\mathbf{x}_i \beta_o) \\ &= \mathbf{x}_i' \mathbf{x}_i E[\log(y_i)|\mathbf{x}_i] \exp(\mathbf{x}_i \beta_o) \\ &= -\mathbf{x}_i' \mathbf{x}_i \exp(-\mathbf{x}_i \beta_o) \exp(\mathbf{x}_i \beta_o) = -\mathbf{x}_i' \mathbf{x}_i, \end{aligned}$$

and so

$$-E[\mathbf{H}_i(\beta_o)|\mathbf{x}_i] = \mathbf{x}_i' \mathbf{x}_i.$$

f. Given part e, the formula based on the expected Hessian is easiest:

$$\text{Avar}[\sqrt{N}(\hat{\beta} - \beta_o)] = [E(\mathbf{x}_i' \mathbf{x}_i)]^{-1}$$

and so

$$\widehat{\text{Avar}}[\sqrt{N}(\hat{\beta} - \beta_o)] = \left( N^{-1} \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \right)^{-1} = (\mathbf{X}' \mathbf{X} / N)^{-1}.$$



g. From part d, we see that the key condition for Fisher consistency, that is, to make  $E[\mathbf{s}_i(\boldsymbol{\beta}_o)|\mathbf{x}_i] = \mathbf{0}$ , is that

$$E[\log(y_i)|\mathbf{x}_i] = -\exp(-\mathbf{x}_i\boldsymbol{\beta}_o).$$

In other words, the implied model for  $E[\log(y_i)|\mathbf{x}_i]$  must be correct. Unfortunately, having  $E(y_i|\mathbf{x}_i)$  correctly specified, that is,  $E(y_i|\mathbf{x}_i) = \exp(\mathbf{x}_i\boldsymbol{\beta}_o)/[1 + \exp(\mathbf{x}_i\boldsymbol{\beta}_o)]$ , generally says nothing about  $E[\log(y_i)|\mathbf{x}_i]$ .

h. We could use the Bernoulli QMLE to estimate the parameters in  $E(y_i|\mathbf{x}_i) = \exp(\mathbf{x}_i\boldsymbol{\beta}_o)/[1 + \exp(\mathbf{x}_i\boldsymbol{\beta}_o)]$  directly, without extra assumptions about  $D(y_i|\mathbf{x}_i)$ .

**18.14.** a. The Stata output is given below with the three sets of standard errors asked for in the problem. The inference starts from the least robust and ends with the most robust.

The difference in standard errors is striking. The standard errors that effectively maintain a binomial distribution – at least its first two moments – lead to huge  $t$  statistics. When we allow for a scale factor – the so-called GLM variance assumption, (18.34) – the standard errors increase by at least a factor of 20. The fully robust standard errors, which allow unrestricted  $\text{Var}(\text{partic}_i|\text{employ}_i, \mathbf{x}_i)$  are still larger – more than three times the standard errors produced under the GLM assumption. It seems pretty clear the binomial distribution does not hold in this application and that the actual conditional variance is not proportional to the nominal variance in the binomial distribution. It is pretty clear that we should use the fully robust standard errors, which lead to much more modest (but still quite significant)  $t$  statistics.

```
. glm partic mrate ltotemp age agesq sole, fam(bin employ) link(logit)
```

Generalized linear models		No. of obs	=	4075
Optimization	: ML	Residual df	=	4069
		Scale parameter	=	
Deviance	= 2199795.239	(1/df) Deviance	=	540.6231
Pearson	= 2021563.356	(1/df) Pearson	=	496.8207

[Binomial]  
[Logit]

```
AIC      = 544.0016
BIC      = 2165971
```

```
. glm partic mrate ltotemp age agesq sole, fam(bin employ) link(logit) sca(x2
```

```
No. of obs      =      4075
Residual df     =      4069
Scale parameter =
(1/df) Deviance =  540.6231
(1/df) Pearson  =  496.8207
```

[Binomial]  
[Logit]

```
AIC          = 544.0016
BIC          = 2165971
```

(Standard errors scaled using square root of Pearson X2-based dispersion)

```
No. of obs      =      4075
Residual df     =      4069
Scale parameter =
(1/df) Deviance =  540.6231
(1/df) Pearson  =  496.8207
```

[Binomial]  
[Logit]

```
AIC      = 544.0016
BIC      = 2165971
```

partic	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval	
mrate	.9871354	.2622177	3.76	0.000	.4731982	1.501073
ltotemp	-.1386562	.0546138	-2.54	0.011	-.2456972	-.0316151
age	.0718575	.0142656	5.04	0.000	.0438974	.0998176
agesq	-.0005512	.0001746	-3.16	0.002	-.0008934	-.000209
sole	.3419834	.1145195	2.99	0.003	.1175294	.5664375
_cons	1.442014	.4368904	3.30	0.001	.5857248	2.298303

b. The fractional logit results for *prate* are given below – with the same kinds of standard errors in part a. In this case the usual MLE standard errors that are too large: they treat  $\sigma^2 = 1$  in (18.58), which is true in the binary case but not generally. With a fractional variable,  $\sigma^2 < 1$ . In fact, the estimate for this data set is  $\hat{\sigma}^2 = .214$ .

The GLM and fully robust standard errors are much closer now, with the fully robust ones typically (but not always) being slightly larger.

```
. glm prate mrate ltotemp age agesq sole, fam(bin) link(logit)
note: prate has noninteger values
```

Generalized linear models		No. of obs	=	4075
Optimization	: ML	Residual df	=	4069
Deviance	= 883.051611	Scale parameter	=	
Pearson	= 871.5810654	(1/df) Deviance	=	.2170193
		(1/df) Pearson	=	.2142003
Variance function:	V(u) = u*(1-u/1)	[Binomial]		
Link function	: g(u) = ln(u/(1-u))	[Logit]		
		AIC	=	.6350527
Log likelihood	= -1287.919784	BIC	=	-32941.02

prate	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval	
mrate	1.147984	.1468736	7.82	0.000	.8601167	1.43585
ltotemp	-.2075898	.0290032	-7.16	0.000	-.264435	-.1507446
age	.0481773	.0145566	3.31	0.001	.0196469	.0767077
agesq	-.0004519	.0004301	-1.05	0.293	-.0012948	.000391
sole	.1652908	.10408	1.59	0.112	-.0387022	.3692838
_cons	2.355715	.2299685	10.24	0.000	1.904985	2.806445

```
. glm prate mrate ltotemp age agesq sole, fam(bin) link(logit) sca(x2)
note: prate has noninteger values
```

Generalized linear models		No. of obs	=	4075
---------------------------	--	------------	---	------

```

Optimization      : ML
Deviance          = 883.051611
Pearson           = 871.5810654

Residual df      = 4069
Scale parameter  =
(1/df) Deviance = .2170193
(1/df) Pearson  = .2142003

Variance function: V(u) = u*(1-u/1)
Link function     : g(u) = ln(u/(1-u))

[Binomial]
[Logit]

AIC               = .6350527
BIC               = -32941.02
Log likelihood    = -1287.919784

```

prate	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
mrate	1.147984	.0679757	16.89	0.000	1.014754	1.281213
ltotemp	-.2075898	.0134232	-15.47	0.000	-.2338988	-.1812808
age	.0481773	.006737	7.15	0.000	.0349729	.0613817
agesq	-.0004519	.000199	-2.27	0.023	-.000842	-.0000618
sole	.1652908	.0481701	3.43	0.001	.0708792	.2597024
_cons	2.355715	.1064335	22.13	0.000	2.147109	2.564321

(Standard errors scaled using square root of Pearson X2-based dispersion)

```

. glm prate mrate ltotemp age agesq sole, fam(bin) link(logit) robust
note: prate has noninteger values

```

```

Generalized linear models
Optimization      : ML
Deviance          = 883.051611
Pearson           = 871.5810654

No. of obs       = 4075
Residual df      = 4069
Scale parameter  =
(1/df) Deviance = .2170193
(1/df) Pearson  = .2142003

Variance function: V(u) = u*(1-u/1)
Link function     : g(u) = ln(u/(1-u))

[Binomial]
[Logit]

AIC               = .6350527
BIC               = -32941.02
Log pseudolikelihood = -1287.919784

```

prate	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
mrate	1.147984	.0747331	15.36	0.000	1.001509	1.294458
ltotemp	-.2075898	.0141209	-14.70	0.000	-.2352662	-.1799134
age	.0481773	.0061543	7.83	0.000	.036115	.0602396
agesq	-.0004519	.0001764	-2.56	0.010	-.0007976	-.0001063
sole	.1652908	.0505915	3.27	0.001	.0661334	.2644483
_cons	2.355715	.1066441	22.09	0.000	2.146696	2.564734

c. It makes sense to compare the coefficients in parts a and b because both approaches could be estimating the same conditional mean function for  $prate_i = partic_i/employ_i$ .

Generally, the binomial approach starts with

$$E(y_i|\mathbf{x}_i, n_i) = n_i \Lambda(\mathbf{x}_i \boldsymbol{\beta}).$$

If we divide both sides by  $n_i$  we get

$$\frac{E(y_i|\mathbf{x}_i, n_i)}{n_i} = \Lambda(\mathbf{x}_i \boldsymbol{\beta})$$

or

$$E\left(\frac{y_i}{n_i} \mid \mathbf{x}_i, n_i\right) = \Lambda(\mathbf{x}_i \boldsymbol{\beta})$$

which, of course, implies

$$E\left(\frac{y_i}{n_i} \mid \mathbf{x}_i\right) = \Lambda(\mathbf{x}_i \boldsymbol{\beta})$$

In other words, the fractional variable  $w_i \equiv y_i/n_i$  follows a fractional response model with a logistic response function. So if we start with  $E(y_i|\mathbf{x}_i, n_i) = n_i \Lambda(\mathbf{x}_i \boldsymbol{\beta})$  then both methods consistently estimate  $\boldsymbol{\beta}$ .

d. The Stata output is given below. Because we want the APE on *prate*, we compute

$$\hat{\beta}_{mrate} \left( N^{-1} \sum_{i=1}^N \frac{\exp(\mathbf{x}_i \hat{\boldsymbol{\beta}})}{[1 + \exp(\mathbf{x}_i \hat{\boldsymbol{\beta}})]^2} \right)$$

for both set of estimates. For the binomial QMLE the estimate is about .147. For the Bernoulli QMLE, the estimate is about .130. Incidentally, the linear regression estimate – coefficient on *mrate* – is about .106, so quite a bit below the other two.

```
. qui glm partic mrate ltotemp age agesq sole, fam(bin employ) link(logit)
. predict xb_bin, xb
. gen sca_bin = exp(xb_bin)/((1 + exp(xb_bin))^2)
. sum sca_bin
```

Variable	Obs	Mean	Std. Dev.	Min	Max
sca_bin	4075	.1492273	.0602467	.0091082	.2499969

```

. di .1492273*_b[mrate]
.14730755

. qui glm prate mrate ltotemp age agesq sole, fam(bin) link(logit)

. predict xb_ber, xb

. gen sca_ber = exp(xb_ber)/((1 + exp(xb_ber))^2)

. di sca_ber*_b[mrate]
.13000441

```

e. The Stata output is given below. The APE is about .038. If we use the linear model, we would get  $.106(.25) \approx .027$ , so somewhat less.

```

. qui glm prate mrate ltotemp age agesq sole, fam(bin) link(logit)

. gen xb_p50 = xb_ber - _b[mrate]*mrate + _b[mrate]*.5

. gen xb_p25 = xb_ber - _b[mrate]*mrate + _b[mrate]*.25

. gen phat_p50 = exp(xb_p50)/(1 + exp(xb_p50))

. gen phat_p25 = exp(xb_p25)/(1 + exp(xb_p25))

. gen diff = phat_p50 - phat_p25

. sum diff

```

Variable	Obs	Mean	Std. Dev.	Min	Max
diff	4075	.0375776	.0107886	.0116919	.0689509

**18.15.** a. We can just use the usual fixed effects or first-differencing estimators. If we define  $w_i \equiv \log[y_{it}/(1 - y_{it})]$  then we have

$$w_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}$$

$$E(u_{it}|\mathbf{x}_i, c_i) = 0, t = 1, \dots, T,$$

which means the key strict exogeneity assumption on  $\{\mathbf{x}_{it} : t = 1, \dots, T\}$  holds. Of course, we could use a GLS version of FE or FD, or use Chamberlain's approach.

b. Because  $\log[y_{it}/(1 - y_{it})] = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}$ ,

$$y_{it}/(1 - y_{it}) = \exp(\mathbf{x}_{it}\boldsymbol{\beta} + v_{it})$$

and so

$$\frac{(1 - y_{it})}{y_{it}} = \frac{1}{\exp(\mathbf{x}_{it}\boldsymbol{\beta} + v_{it})}$$

or

$$\frac{1}{y_{it}} = \frac{1}{\exp(\mathbf{x}_{it}\boldsymbol{\beta} + v_{it})} + 1 = \frac{1 + \exp(\mathbf{x}_{it}\boldsymbol{\beta} + v_{it})}{\exp(\mathbf{x}_{it}\boldsymbol{\beta} + v_{it})}$$

which implies

$$y_{it} = \frac{\exp(\mathbf{x}_{it}\boldsymbol{\beta} + v_{it})}{1 + \exp(\mathbf{x}_{it}\boldsymbol{\beta} + v_{it})}.$$

The ASF is defined, for each  $t$ , as

$$\text{ASF}_t(\mathbf{x}_t) = \int_{-\infty}^{\infty} \left[ \frac{\exp(\mathbf{x}_t\boldsymbol{\beta} + v)}{1 + \exp(\mathbf{x}_t\boldsymbol{\beta} + v)} g_t(v) dv \right]$$

where  $g_t(\cdot)$  is the density of  $v_{it}$ . (Of course, allowing this density to be discrete changes the integral to a sum.) We can also write

$$\text{ASF}_t(\mathbf{x}_t) = E_{v_{it}} \left[ \frac{\exp(\mathbf{x}_t\boldsymbol{\beta} + v_{it})}{1 + \exp(\mathbf{x}_t\boldsymbol{\beta} + v_{it})} \right],$$

that is, we fix the covariates at values  $\mathbf{x}_t$  and average across the distribution of the unobservables,  $v_{it}$ .

c. The ASF cannot be estimated without further assumptions because we cannot estimate the expected value of  $\exp(\mathbf{x}_t\boldsymbol{\beta} + v_{it})/[1 + \exp(\mathbf{x}_t\boldsymbol{\beta} + v_{it})]$  for given  $\mathbf{x}_t$  without further assumptions.

d. By the law of iterated expectations, we have

$$\begin{aligned} \text{ASF}_t(\mathbf{x}_t) &= E_{\bar{\mathbf{x}}_i} \left\{ E \left[ \frac{\exp(\mathbf{x}_t \boldsymbol{\beta} + v_{it})}{1 + \exp(\mathbf{x}_t \boldsymbol{\beta} + v_{it})} \mid \bar{\mathbf{x}}_i \right] \right\} \\ &E_{\bar{\mathbf{x}}_i} \left\{ E \left[ \frac{\exp(\mathbf{x}_t \boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i \boldsymbol{\xi} + r_{it})}{1 + \exp(\mathbf{x}_t \boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i \boldsymbol{\xi} + r_{it})} \mid \bar{\mathbf{x}}_i \right] \right\}. \end{aligned}$$

With  $r_{it}$  is independent of  $\mathbf{x}_i$  – and we can assume  $E(a_i) = 0$  due to the presence of  $\psi$  – then we can consistently estimate  $\boldsymbol{\beta}$ ,  $\psi$ , and  $\boldsymbol{\xi}$  by pooled OLS:

$$w_{it} = \text{on } \mathbf{x}_{it}, 1, \bar{\mathbf{x}}_i, t = 1, \dots, T; i = 1, \dots, N.$$

(Recall this produces the FE estimator of  $\boldsymbol{\beta}$ .) Further, by independence,

$$E \left[ \frac{\exp(\mathbf{x}_t \boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i \boldsymbol{\xi} + r_{it})}{1 + \exp(\mathbf{x}_t \boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i \boldsymbol{\xi} + r_{it})} \mid \bar{\mathbf{x}}_i \right] = \int_{-\infty}^{\infty} \frac{\exp(\mathbf{x}_t \boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i \boldsymbol{\xi} + \tau)}{1 + \exp(\mathbf{x}_t \boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i \boldsymbol{\xi} + \tau)} f_t(\tau) d\tau$$

For fixed  $\bar{\mathbf{x}}_i = \bar{\mathbf{x}}$ , we can consistently estimate this expression as

$$N^{-1} \sum_{i=1}^N \frac{\exp(\mathbf{x}_t \hat{\boldsymbol{\beta}} + \hat{\psi} + \bar{\mathbf{x}} \hat{\boldsymbol{\xi}} + \hat{r}_{it})}{1 + \exp(\mathbf{x}_t \hat{\boldsymbol{\beta}} + \hat{\psi} + \bar{\mathbf{x}} \hat{\boldsymbol{\xi}} + \hat{r}_{it})},$$

where  $\hat{r}_{it} \equiv w_{it} - \mathbf{x}_{it} \hat{\boldsymbol{\beta}} - \hat{\psi} - \bar{\mathbf{x}}_i \hat{\boldsymbol{\xi}}$  are the pooled OLS residuals. To get the ASF, we need to further average out over the distribution of  $\bar{\mathbf{x}}_i$ , which gives

$$\widehat{\text{ASF}}_t(\mathbf{x}_t) = N^{-1} \sum_{i=1}^N \frac{\exp(\mathbf{x}_t \hat{\boldsymbol{\beta}} + \hat{\psi} + \bar{\mathbf{x}}_i \hat{\boldsymbol{\xi}} + \hat{r}_{it})}{1 + \exp(\mathbf{x}_t \hat{\boldsymbol{\beta}} + \hat{\psi} + \bar{\mathbf{x}}_i \hat{\boldsymbol{\xi}} + \hat{r}_{it})}.$$

We use this as usual: take derivatives and changes with respect to the elements of  $\mathbf{x}_t$ .

**18.16. (Bonus Question)** Consider a panel data mode for  $y_{it} \geq 0$  with multiplicative heterogeneity and a multiplicative idiosyncratic error:

$$y_{it} = c_i \exp(\mathbf{x}_{it} \boldsymbol{\beta}) r_{it}.$$

If we assume  $\{\mathbf{x}_{it} : t = 1, \dots, T\}$  is strictly exogenous then we can estimate  $\boldsymbol{\beta}$  using the fixed effects Poisson QMLE. Instead, assume we have instruments,  $\{\mathbf{z}_{it} : t = 1, 2, \dots, T\}$  that satisfy



a sequential exogeneity assumption:

$$E(r_{it} | \mathbf{z}_{it}, \dots, \mathbf{z}_{i1}, c_i) = E(r_{it}) = 1,$$

where setting the expected value to unity is a normalization. (As usual,  $\mathbf{x}_{it}$  should probably include a full set of time period dummies.)

a. Show that we can write

$$\frac{y_{it}}{\exp(\mathbf{x}_{it}\boldsymbol{\beta})} - \frac{y_{i,t+1}}{\exp(\mathbf{x}_{i,t+1}\boldsymbol{\beta})} = c_i(r_{it} - r_{i,t+1}) \equiv e_{i,t+1}$$

where

$$E(e_{it} | \mathbf{z}_{it}, \dots, \mathbf{z}_{i1}) = 0, \quad t = 1, \dots, T-1.$$

b. Part a implies that we can use the moment conditions

$$E \left[ \frac{y_{it}}{\exp(\mathbf{x}_{it}\boldsymbol{\beta})} - \frac{y_{i,t+1}}{\exp(\mathbf{x}_{i,t+1}\boldsymbol{\beta})} \middle| \mathbf{z}_{it}, \dots, \mathbf{z}_{i1} \right] = 0, \quad t = 1, \dots, T-1$$

to estimate  $\boldsymbol{\beta}$ . Explain why using these moments directly can cause computational problems.

(Hint: Suppose For example, if  $x_{itj} > 0$  for some  $j$  and all  $i$  and  $t$ . What would happen if  $\beta_j$  is made larger and larger?)

c. Define the average of the population means across time as

$$\boldsymbol{\mu}_{\mathbf{x}} \equiv T^{-1} \sum_{r=1}^T E(\mathbf{x}_{ir}).$$

Show that if you multiply the moment conditions in part b by  $\exp(\boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\beta})$ , the resulting moment conditions are

$$E \left[ \frac{y_{it}}{\exp[(\mathbf{x}_{it} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\beta}]} - \frac{y_{i,t+1}}{\exp[(\mathbf{x}_{i,t+1} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\beta}]} \middle| \mathbf{z}_{it}, \dots, \mathbf{z}_{i1} \right] = 0.$$

[See Windmeijer (2002, Economics Letters).] How does this help with the computational

problem in part b?

d. What would you use in place of  $\mu_x$  given that  $\mu_x$  is unknown?

e. Suppose that  $\{\mathbf{x}_{it} : t = 1, 2, \dots, T\}$  is sequentially exogenous, so that we can take  $\mathbf{z}_{it} = \mathbf{x}_{it}$ . Show that

$$E\left[y_{it} - \frac{y_{i,t+1}}{\exp[(\mathbf{x}_{i,t+1} - \mathbf{x}_{it})\boldsymbol{\beta}]} \mid \mathbf{x}_{it}, \dots, \mathbf{x}_{i1}\right] = 0, t = 1, \dots, T-1.$$

In other words, we can write moment conditions in terms of the first difference of the explanatory variables.

### Solution

a. From  $y_{it} = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})r_{it}$  for all  $t = 1, \dots, T$  we have

$$\begin{aligned} \frac{y_{it}}{\exp(\mathbf{x}_{it}\boldsymbol{\beta})} &= c_i r_{it} \\ \frac{y_{i,t+1}}{\exp(\mathbf{x}_{i,t+1}\boldsymbol{\beta})} &= c_i r_{i,t+1}, \end{aligned}$$

and subtracting the first equation from the second gives

$$\frac{y_{it}}{\exp(\mathbf{x}_{it}\boldsymbol{\beta})} - \frac{y_{i,t+1}}{\exp(\mathbf{x}_{i,t+1}\boldsymbol{\beta})} = c_i(r_{it} - r_{i,t+1}).$$

Now

$$\begin{aligned} E[c_i(r_{it} - r_{i,t+1}) \mid \mathbf{z}_{it}, \dots, \mathbf{z}_{i1}, c_i] &= c_i[E(r_{it} \mid \mathbf{z}_{it}, \dots, \mathbf{z}_{i1}, c_i) \\ &\quad - E(r_{i,t+1} \mid \mathbf{z}_{it}, \dots, \mathbf{z}_{i1}, c_i)] \\ &= c_i(1 - 1) = 0 \end{aligned}$$

b. Suppose  $x_{it1} > 0$  for all  $i$  and  $t$ . Then  $\beta_1 x_{it1} \rightarrow \infty$  as  $\beta_1 \rightarrow \infty$ , which means

$$\exp(\beta_1 x_{it1} + \beta_2 x_{it2} + \dots + \beta_K x_{itK}) \rightarrow \infty$$

for all  $i$  and  $t$ , for any values of  $\beta_2, \dots, \beta_K$ . Then

$$\frac{y_{it}}{\exp(\mathbf{x}_{it}\boldsymbol{\beta})} - \frac{y_{i,t+1}}{\exp(\mathbf{x}_{i,t+1}\boldsymbol{\beta})} \rightarrow 0$$

as  $\beta_1 \rightarrow \infty$ , and so the residual function can be made closer and closer to zero by increasing  $\beta_1$  without bound.

c. Multiplying the original moment conditions by the  $\exp(\boldsymbol{\mu}_x\boldsymbol{\beta})$  clearly does not change that they still hold:

$$\exp(\boldsymbol{\mu}_x\boldsymbol{\beta})\mathbb{E}\left[\frac{y_{it}}{\exp(\mathbf{x}_{it}\boldsymbol{\beta})} - \frac{y_{i,t+1}}{\exp(\mathbf{x}_{i,t+1}\boldsymbol{\beta})} \middle| \mathbf{z}_{it}, \dots, \mathbf{z}_{i1}\right] = 0.$$

The left hand side is simply

$$\frac{\exp(\boldsymbol{\mu}_x\boldsymbol{\beta})y_{it}}{\exp(\mathbf{x}_{it}\boldsymbol{\beta})} - \frac{\exp(\boldsymbol{\mu}_x\boldsymbol{\beta})y_{i,t+1}}{\exp(\mathbf{x}_{i,t+1}\boldsymbol{\beta})} = \frac{y_{it}}{\exp[(\mathbf{x}_{it} - \boldsymbol{\mu}_x)\boldsymbol{\beta}]} - \frac{y_{i,t+1}}{\exp[(\mathbf{x}_{i,t+1} - \boldsymbol{\mu}_x)\boldsymbol{\beta}]}.$$

Using these new moment conditions does not lead to the problem discussed in part b because the deviated covariates,  $\mathbf{x}_{it} - \boldsymbol{\mu}_x$ , can take on both negative and positive values.

d. We would use the sample counterpart,

$$\bar{\mathbf{x}} \equiv T^{-1} \sum_{r=1}^T \left( N^{-1} \sum_{i=1}^N \mathbf{x}_{ir} \right) = (NT)^{-1} \sum_{i=1}^N \sum_{r=1}^T \mathbf{x}_{ir}.$$

In the sample, the deviated variables,  $\mathbf{x}_{it} - \bar{\mathbf{x}}$ , will always take on positive and negative values.

Technically we should account for the estimation error in  $\bar{\mathbf{x}}$  but it likely has a minor effect.

The sample moments we would like to make close to zero have the form

$$\sum_{i=1}^N \sum_{t=1}^{T-1} \mathbf{g}'_{it} \left[ \frac{y_{it}}{\exp[(\mathbf{x}_{it} - \bar{\mathbf{x}})\boldsymbol{\beta}]} - \frac{y_{i,t+1}}{\exp[(\mathbf{x}_{i,t+1} - \bar{\mathbf{x}})\boldsymbol{\beta}]} \right]$$

where  $\mathbf{g}_{it} \equiv \mathbf{g}_t(\mathbf{z}_{it}, \dots, \mathbf{z}_{i1})$  is a function of the instruments up through time  $t$ . Or, stack the moments over time rather than sum them up to enhance efficiency. In either case, we would

use GMM with an optimal weighting matrix to set the sample moments as close to zero as possible.

e. If we can take  $\mathbf{z}_{it} = \mathbf{x}_{it}$  then we know from part a that

$$E[c_i(r_{it} - r_{i,t+1})|\mathbf{x}_{it}, \dots, \mathbf{x}_{i1}, c_i)].$$

That means any function of  $(\mathbf{x}_{it}, \dots, \mathbf{x}_{i1})$  can multiply the moment conditions and we are still left with a zero conditional mean. In particular,

$$E\left\{\exp(\mathbf{x}_{it}\boldsymbol{\beta})\left[\frac{y_{it}}{\exp(\mathbf{x}_{it}\boldsymbol{\beta})} - \frac{y_{i,t+1}}{\exp(\mathbf{x}_{i,t+1}\boldsymbol{\beta})}\right]\middle|\mathbf{x}_{it}, \dots, \mathbf{x}_{i1}\right\} = 0, t = 1, \dots, T-1$$

and simple algebra shows

$$\exp(\mathbf{x}_{it}\boldsymbol{\beta})\left[\frac{y_{it}}{\exp(\mathbf{x}_{it}\boldsymbol{\beta})} - \frac{y_{i,t+1}}{\exp(\mathbf{x}_{i,t+1}\boldsymbol{\beta})}\right] = y_{it} - \frac{y_{i,t+1}}{\exp[(\mathbf{x}_{i,t+1} - \mathbf{x}_{it})\boldsymbol{\beta}]}.$$

## Solutions to Chapter 19 Problems

**19.1.** If  $r_i$  is the same for any random draw  $i$ , then it is nonrandom. From equation (19.9), we can write

$$\begin{aligned} P(w_i = 1|\mathbf{x}_i) &= \Phi[(\beta_1 + \beta_2 x_{i2} + \dots + \beta_K x_{iK} - \log(r))/\sigma] \\ &= \Phi[(\beta_1 - \log(r))/\sigma + (\beta_2/\sigma)x_{i2} + \dots + (\beta_K/\sigma)x_{iK}], \end{aligned}$$

where it is helpful to separate the intercept from the slopes. From this equation, it is clear that probit of  $w_i$  on  $(1, x_{i2}, \dots, x_{iK})$  consistently estimates  $(\beta_1 - \log(r))/\sigma, \beta_2/\sigma, \dots, \beta_K/\sigma$ . Let  $\alpha_1^* = (\beta_1 - \log(r))/\sigma$  and define  $\beta_j^* = \beta_j/\sigma, j = 1, \dots, K$ . Unfortunately, we cannot recover the original parameters because, for example,  $\beta_1 = \sigma\alpha_1^* + \log(r)$ , and we do not know  $\sigma$ . Although  $\alpha_1^*$  is identified, and  $\log(r)$  is known, we can not recover the scaled intercept  $\beta_1^* \equiv \beta_1/\sigma = \alpha_1^* + \log(r)/\sigma$ . Of course, we directly estimate the scaled slopes,  $\beta_j/\sigma$ , and so we can estimate the direction of the effects on  $E(y|\mathbf{x})$ . But we cannot estimate the original intercepts or slopes. Assuming  $\beta_h \neq 0$ , we can estimate  $\beta_j/\beta_h$  for  $j \neq h$ , which means we can estimate the relative effects. Unlike in the case where the  $r_i$  vary, we cannot estimate the magnitudes of the partial effects on  $E(y|\mathbf{x})$ .

**19.2.** a. It suffices to find the density of  $\log(w_i)$  conditional on  $\mathbf{x}_i$ ; of course we arrive at the same place for the MLEs if we work with  $D(w_i|\mathbf{x}_i)$ . Now

$$\log(w_i) = \max[\log(f), \log(y_i)]$$

and  $\log(y_i) = \mathbf{x}_i\boldsymbol{\beta} + u_i$ , where

$$D(u_i|\mathbf{x}_i) = \text{Normal}(0, \sigma^2).$$

Let  $\tilde{w}_i = \log(w_i)$ ,  $\tilde{f} = \log(f)$ , and  $\tilde{y}_i = \log(y_i)$ , so that  $D(\tilde{y}_i|\mathbf{x}_i) = \text{Normal}(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$ . Now

$$\begin{aligned}
P(\tilde{w}_i = \tilde{f}|\mathbf{x}_i) &= P(\tilde{y}_i \leq \tilde{f}|\mathbf{x}_i) = P\left(\frac{\mathbf{x}_i\boldsymbol{\beta} + u_i}{\sigma} \leq \frac{\tilde{f}}{\sigma} \middle| \mathbf{x}_i\right) \\
&= P\left(\frac{u_i}{\sigma} \leq \frac{\tilde{f} - \mathbf{x}_i\boldsymbol{\beta}}{\sigma} \middle| \mathbf{x}_i\right) = \Phi\left(\frac{\tilde{f} - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)
\end{aligned}$$

The conditional density for  $\tilde{w} > \tilde{f}$  is simply the conditional density for  $\tilde{y}_i$ , that is,

$$\left(\frac{1}{\sigma}\right)\phi\left(\frac{\tilde{w} - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right).$$

Therefore, the density for  $\tilde{w}_i$  conditional on  $\mathbf{x}_i$  can be written as

$$\left[\left(\frac{1}{\sigma}\right)\phi\left(\frac{\tilde{w} - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)\right]^{1[\tilde{w} > \tilde{f}]} \left[\Phi\left(\frac{\tilde{f} - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)\right]^{1[\tilde{w} = \tilde{f}]}$$

It follows that the log likelihood for a random draw  $i$  is

$$1[\tilde{w}_i > \tilde{f}] \log\left[\left(\frac{1}{\sigma}\right)\phi\left(\frac{\tilde{w}_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)\right] + 1[\tilde{w}_i = \tilde{f}] \log\left[\Phi\left(\frac{\tilde{f} - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)\right].$$

Notice that when  $\tilde{f} = 0$  we get the same log likelihood as for the Type I Tobit model for corner solutions, which we covered in Chapter 17.

b. Because  $u$  is independent of  $\mathbf{x}$  with a  $Normal(0, \sigma^2)$  distribution,

$$E[\exp(u)|\mathbf{x}] = E[\exp(u)] = \exp(\sigma^2/2),$$

where the second inequality follows from the moments of a lognormal distribution. Therefore,

$$\begin{aligned}
E(y|\mathbf{x}) &= \exp(\mathbf{x}\boldsymbol{\beta})E[\exp(u)|\mathbf{x}] = \exp(\mathbf{x}\boldsymbol{\beta})\exp(\sigma^2/2) \\
&= \exp(\mathbf{x}\boldsymbol{\beta} + \sigma^2/2).
\end{aligned}$$

After using the MLE on the censored data to obtain  $\hat{\boldsymbol{\beta}}$  and  $\hat{\sigma}^2$ , we can use

$$\hat{E}(y|\mathbf{x}) = \exp(\mathbf{x}\hat{\boldsymbol{\beta}} + \hat{\sigma}^2/2).$$

c. It is hard to see why  $E(w|\mathbf{x})$  would be of much interest. In most cases the floor,  $f$ , is

arbitrary, and so it is unclear why we would be interested in how the mean of the censored variable changes with the  $x_j$ . One could imagine that, if  $f$  is a minimum wage and  $w_i$  represents the observed wage for worker  $i$ , one might be interested to know how a change in a policy variable affects observed wage, on average.

**19.3.** a. The two-limit Tobit model from Section 17.7 could be used with limits at 0 and 10.

b. The lower bound of zero reflects the fact that pension contribution cannot be a negative percentage of income. But the upper bound of 10 percent is imposed by law, and is essentially arbitrary. If we defined a variable as the desired percentage put into the pension plan, then it could range from 0 to 100. So the upper bound of 10 can be viewed as a data censoring problem because some individuals presumably would contribute  $y > 10$  if the limit were raised. But it depends on the purpose of the study: to estimate the effects within the current institutional setting or to estimate effects on pension contributions in the absence of constraints.

c. From Problem 17.3 part b, with  $a_1 = 0$ , we have

$$\begin{aligned} E(y|\mathbf{x}) &= (\mathbf{x}\boldsymbol{\beta}) \cdot \{\Phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] - \Phi(-\mathbf{x}\boldsymbol{\beta}/\sigma)\} \\ &\quad + \sigma\{\phi(\mathbf{x}\boldsymbol{\beta}/\sigma) - \phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma]\} + a_2\Phi[(\mathbf{x}\boldsymbol{\beta} - a_2)/\sigma]. \end{aligned}$$

Taking the derivative of this function with respect to  $a_2$  gives

$$\begin{aligned} \partial E(y|\mathbf{x})/\partial a_2 &= (\mathbf{x}\boldsymbol{\beta}/\sigma) \cdot \phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] + [(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] \cdot \phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] \\ &\quad + \Phi[(\mathbf{x}\boldsymbol{\beta} - a_2)/\sigma] - (a_2/\sigma)\phi[(\mathbf{x}\boldsymbol{\beta} - a_2)/\sigma] \\ &= \Phi[(\mathbf{x}\boldsymbol{\beta} - a_2)/\sigma]. \end{aligned}$$

We can plug in  $a_2 = 10$  to obtain the approximate effect of increasing the cap from 10 to 11.

For a given value of  $\mathbf{x}$ , we would compute  $\Phi[(\mathbf{x}\hat{\boldsymbol{\beta}} - 10)/\hat{\sigma}]$ , where  $\hat{\boldsymbol{\beta}}$  and  $\hat{\sigma}$  are the MLEs. We might evaluate this expression at the sample average of  $\mathbf{x}$  or at other interesting values (such as

across gender or race).

d. If  $y_i < 10$  for  $i = 1, \dots, N$ ,  $\hat{\beta}$  and  $\hat{\sigma}$  are just the usual type I Tobit estimates with lower limit at zero: there are no observations that contribute to the third piece in the log likelihood.

**19.4.** a. If you are interested in the effects of things like age of the building and neighborhood demographics on fire damage, given that a fire has occurred, then there is no problem. We simply need a random sample of buildings that actually caught on fire. You might want to supplement this with an analysis of the probability that buildings catch fire, given building and neighborhood characteristics. But then a two-part (hurdle) model is appropriate.

b. The issue in this case is a bit subtle because it depends on the population of interest. One possibility is, at a given point in time, to define the population of interest to be workers currently enrolled in a 401(k) plan. Then using a random sample of workers already in a 401(k) plan is appropriate. But workers currently enrolled in a plan may not represent those that may be enrolled in the future. In fact, we might think of being interested in a scenario where *all* workers are enrolled. It makes sense to think about the sensitivity of contributions to the match rate for the population of all workers. Of course, in general, using a random sample of those already enrolled leads to a sample selection problem for estimating the parameters for the larger population – much like the problem of estimating a wage offer equation (except that, in addition to not observing contributions, we would not observe a match rate for those not enrolled).

**19.5.** Because  $IQ$  and  $KWW$  are both indicators of *abil* we can write

$$IQ = \xi_1 abil + a_i, \quad KWW = \xi_2 abil + a_i,$$

where  $\xi_1, \xi_2 > 0$ . For simplicity, I set the intercepts to zero, as this does not affect the



conclusions of the problem. The structural equation is  $\log(wage) = \mathbf{z}_1\delta_1 + abil + v$ . Now, given the selection mechanism described in Example 19.4 ( $IQ$  is observed if  $IQ + r \geq 0$ ), we can assume that

$$E(v|\mathbf{z}_1, abil, IQ, KWW, r) = 0, \quad (19.124)$$

which is the standard ignorability assumption with the added assumption that  $v$  is unrelated to  $r$  in the conditional mean sense. To see what else we need, write  $abil$  in terms of  $IQ$  and  $a_1$  and plug into the structural equation to get

$$\log(wage) = \mathbf{z}_1\delta_1 + \xi_1^{-1}IQ + v + \xi_1^{-1}a_1.$$

Now, we want to use  $KWW$  as an instrument for  $IQ$  in this equation, and use 2SLS on the selected sample. The full set of instruments is  $(\mathbf{z}_1, KWW)$ . From Theorem 19.1 we need the error  $u = v + \xi_1^{-1}a_1$  to satisfy  $E(u|\mathbf{z}_1, KWW, s) = 0$ . Now, because  $s$  is a function of  $IQ$  and  $r$ , from (19.124) we have  $E(v|\mathbf{z}_1, KWW, s) = 0$ . To ensure  $E(a_1|\mathbf{z}_1, KWW, s) = 0$  we can assume  $E(a_1|\mathbf{z}_1, KWW, r) = 0$  or, equivalently,  $E(a_1|\mathbf{z}_1, a_2, r) = 0$ . The symmetrical assumption on  $a_2$  is  $E(a_2|\mathbf{z}_1, a_1, r) = 0$ . Loosely, in addition to the errors in the indicator equations being uncorrelated, they are also uncorrelated with the selection error. But for all of this to work we need to make zero conditional mean assumptions.

**19.6.** This is essentially given in equation (19.45), but were we allow the truncation points to depend on  $\mathbf{x}_i$ . Let  $y_i$  given  $\mathbf{x}_i$  have density  $f(y|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})$ , where  $\boldsymbol{\beta}$  is the vector indexing  $E(y_i|\mathbf{x}_i)$  and  $\boldsymbol{\gamma}$  is another set of parameters (often a single variance parameter). Then the density of  $y_i$  given  $\mathbf{x}_i, s_i = 1$ , when  $s_i = 1[a_1(\mathbf{x}_i) < y_i < a_2(\mathbf{x}_i)]$ , is

$$p(y|\mathbf{x}_i, s_i = 1) = \frac{f(y|\mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\gamma})}{F(a_2(\mathbf{x}_i)|\mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}) - F(a_1(\mathbf{x}_i)|\mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\gamma})}, a_1(\mathbf{x}_i) < y < a_2(\mathbf{x}_i).$$

In the Hausman and Wise (1977) study,  $y_i = \log(\text{income}_i)$ ,  $a_1(\mathbf{x}_i) = -\infty$ , and  $a_2(\mathbf{x}_i)$  was a function of family size (which determines the official poverty level).

**19.7.** a. If  $E(u_1|v_2) = \gamma_1 v_2 + \gamma_2(v_2^2 - 1)$  then, because  $(u_1, v_2)$  is independent of  $\mathbf{x}$ ,

$$E(y_1|\mathbf{x}, v_2) = \mathbf{x}_1\boldsymbol{\beta}_1 + E(u_1|v_2) = \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1 v_2 + \gamma_2(v_2^2 - 1).$$

Now, using iterated expectations (since  $y_2$  is a function of  $(\mathbf{x}, v_2)$ ), we have

$$\begin{aligned} E(y_1|\mathbf{x}, y_2) &= \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1 E(v_2|\mathbf{x}, y_2) + \gamma_2 \{E(v_2^2|\mathbf{x}, y_2) - 1\} \\ &= \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1 E(v_2|\mathbf{x}, y_2) + \gamma_2 \{\text{Var}(v_2|\mathbf{x}, y_2) + [E(v_2|\mathbf{x}, y_2)]^2 - 1\}. \end{aligned}$$

We only need these expressions for  $y_2 = 1$ . Using  $E(v_2|v_2 > -\mathbf{x}\boldsymbol{\delta}_2) = \lambda(\mathbf{x}\boldsymbol{\delta}_2)$  and

$$\text{Var}(v_2|v_2 > -\mathbf{x}\boldsymbol{\delta}_2) = 1 - \lambda(\mathbf{x}\boldsymbol{\delta}_2)[\lambda(\mathbf{x}\boldsymbol{\delta}_2) + \mathbf{x}\boldsymbol{\delta}_2],$$

we have

$$\begin{aligned} E(y_1|\mathbf{x}, y_2 = 1) &= \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1 E(v_2|v_2 > -\mathbf{x}\boldsymbol{\delta}_2) + \gamma_2 \text{Var}(v_2|v_2 > -\mathbf{x}\boldsymbol{\delta}_2) \\ &= \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1 \lambda(\mathbf{x}\boldsymbol{\delta}_2) + \gamma_2 \{1 - \lambda(\mathbf{x}\boldsymbol{\delta}_2)[\lambda(\mathbf{x}\boldsymbol{\delta}_2) + \mathbf{x}\boldsymbol{\delta}_2] + [\lambda(\mathbf{x}\boldsymbol{\delta}_2)]^2 - 1\} \\ &= \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1 \lambda(\mathbf{x}\boldsymbol{\delta}_2) - \gamma_2 \lambda(\mathbf{x}\boldsymbol{\delta}_2) \mathbf{x}\boldsymbol{\delta}_2. \end{aligned}$$

b. Now, we obtain  $\mathbf{x}_i\hat{\boldsymbol{\delta}}_2$  and  $\hat{\lambda}_{i2}$  after first-stage probit and then run the regression

$$y_{i1} \text{ on } \mathbf{x}_{i1}, \hat{\lambda}_{i2}, \hat{\lambda}_{i2} \cdot (\mathbf{x}_i\hat{\boldsymbol{\delta}}_2)$$

using the selected sample. We get consistent estimators of  $\boldsymbol{\beta}_1, \gamma_1$ , and  $-\gamma_2$ .

c. A standard  $F$  test of joint significance of  $\hat{\lambda}_{i2}$  and  $\hat{\lambda}_{i2} \cdot (\mathbf{x}_i\hat{\boldsymbol{\delta}}_2)$  (two restrictions) in the regression from part b is a valid test, assuming homoskedasticity in the population structural model. As usual, the null is no sample selection bias.

**19.8.** If we replace  $y_2$  with  $\hat{y}_2$ , we need to see what happens when  $y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2$  is plugged into the structural mode:

$$\begin{aligned} y_1 &= \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 \cdot (\mathbf{z} \boldsymbol{\delta}_2 + v_2) + u_1 \\ &= \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 \cdot (\mathbf{z} \boldsymbol{\delta}_2) + (u_1 + \alpha_1 v_2). \end{aligned} \quad (19.125)$$

So, the procedure is to replace  $\boldsymbol{\delta}_2$  in (19.125) its  $\sqrt{N}$ -consistent estimator,  $\hat{\boldsymbol{\delta}}_2$ . The key is to note that the error term in (19.125) is  $u_1 + \alpha_1 v_2$ . If the selection correction is going to work when the fitted value is plugged in for  $y_2$ , we need the expected value of  $u_1 + \alpha_1 v_2$  given  $(\mathbf{z}, v_3)$  to be linear in  $v_3$  (in particular, it cannot depend on  $\mathbf{z}$ ). Then we can write

$$E(y_1 | \mathbf{z}, v_3) = \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 \cdot (\mathbf{z} \boldsymbol{\delta}_2) + \gamma_1 v_3,$$

where  $E[(u_1 + \alpha_1 v_2) | v_3] = \gamma_1 v_3$  by normality. Conditioning on  $y_3 = 1$  gives

$$E(y_1 | \mathbf{z}, y_3 = 1) = \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 \cdot (\mathbf{z} \boldsymbol{\delta}_2) + \gamma_1 \lambda(\mathbf{z} \boldsymbol{\delta}_3). \quad (19.126)$$

A sufficient condition for (19.126) is that  $(u_1, v_2, v_3)$  is independent of  $\mathbf{z}$  with a trivariate normal distribution. We can get by with less than this, but the nature of  $v_2$  is restricted. If we use the IV approach – rather than plugging in fitted values – we need assume nothing about  $v_2$ ;  $y_2 = \mathbf{z} \boldsymbol{\delta}_2 + v_2$  is just a linear projection.

As a practical matter, if we cannot write  $y_2 = \mathbf{z} \boldsymbol{\delta}_2 + v_2$ , where  $v_2$  is independent of  $\mathbf{z}$  and approximately normal, then the OLS alternative will not be consistent. Thus, equations where  $y_2$  is binary, or is some other variable that exhibits nonnormality, cannot be consistently estimated using the OLS procedure. This is why 2SLS is generally preferred.

**19.9.** Here is the Stata session I used to implement Procedure 19.4, although the standard errors in the second step are not adjusted to account for the first-stage Tobit estimation. Still,  $\hat{v}_3$  is not statistically significant, and adding it is not really necessary in this application.

```
. tobit hours exper expersq age kidslt6 kidsge6 nwifeinc motheduc fatheduc
    huseduc, ll(0)
```

Tobit regression	Number of obs	=	753
	LR chi2(9)	=	261.82
	Prob > chi2	=	0.0000

Log likelihood = -3823.9826

Pseudo R2 = 0.0331

hours	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
exper	136.9463	17.27271	7.93	0.000	103.0373	170.8554
expersq	-1.947776	.5388933	-3.61	0.000	-3.005708	-.8898433
age	-54.78118	7.568762	-7.24	0.000	-69.63985	-39.9225
kidslt6	-864.3263	111.6246	-7.74	0.000	-1083.463	-645.1896
kidsge6	-24.68934	38.77122	-0.64	0.524	-100.8034	51.42468
nwifeinc	-5.312478	4.572888	-1.16	0.246	-14.28978	3.664822
motheduc	24.28791	16.74349	1.45	0.147	-8.582209	57.15803
fatheduc	6.566355	16.00859	0.41	0.682	-24.86103	37.99374
huseduc	3.129767	17.46452	0.18	0.858	-31.15583	37.41537
_cons	1548.141	437.1192	3.54	0.000	690.0075	2406.275
/sigma	1126.282	41.77533			1044.271	1208.294

Obs. summary: 325 left-censored observations at hours<=0  
 428 uncensored observations  
 0 right-censored observations

. predict zd3hat  
 (option xb assumed; fitted values)

. sum zd3hat

Variable	Obs	Mean	Std. Dev.	Min	Max
zd3hat	753	302.7538	814.8662	-2486.756	1933.23

. gen v3hat = hours - zd3hat if hours > 0  
 (325 missing values generated)

. ivreg lwage exper expersq v3hat (educ = age kidslt6 kidsge6 nwifeinc  
 motheduc fatheduc huseduc)

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs = 428	
Model	34.6676357	4	8.66690893	F( 4, 423) =	9.97
Residual	188.659805	423	.446004268	Prob > F =	0.0000
Total	223.327441	427	.523015084	R-squared =	0.1552
				Adj R-squared =	0.1472
				Root MSE =	.66784

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
educ	.085618	.0213955	4.00	0.000	.0435633	.1276726
exper	.0378509	.0137757	2.75	0.006	.0107734	.0649283
expersq	-.0007453	.0004036	-1.85	0.065	-.0015386	.0000479
v3hat	-.0000515	.0000412	-1.25	0.211	-.0001325	.0000294
_cons	-.1786154	.2925231	-0.61	0.542	-.7535954	.3963645

Instrumented: educ  
 Instruments: exper expersq v3hat age kidslt6 kidsge6 nwifeinc motheduc  
 fatheduc huseduc

If we just use 2SLS on the selected sample without including  $\hat{v}_3$ , and the IVs for *educ* are *motheduc*, *fatheduc*, and *huseduc*, then the estimated return to education is about 8.0%:

```
. ivreg lwage exper expersq (educ = motheduc fatheduc huseduc)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs = 428		
Model	33.3927368	3	11.1309123	F( 3, 424) = 11.52		
Residual	189.934704	424	.447959208	Prob > F = 0.0000		
Total	223.327441	427	.523015084	R-squared = 0.1495		
				Adj R-squared = 0.1435		
				Root MSE = .6693		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
educ	.0803918	.021774	3.69	0.000	.0375934	.1231901
exper	.0430973	.0132649	3.25	0.001	.0170242	.0691704
expersq	-.0008628	.0003962	-2.18	0.030	-.0016415	-.0000841
_cons	-.1868572	.2853959	-0.65	0.513	-.7478242	.3741097

Instrumented: educ  
Instruments: exper expersq motheduc fatheduc huseduc

**19.10. a.** Substitute the reduced forms for  $y_1$  and  $y_2$  into the third equation:

$$y_3 = \max(0, \alpha_{31}(\mathbf{z}\delta_1) + \alpha_{32}(\mathbf{z}\delta_2) + \mathbf{z}_3\delta_3 + v_3) \\ \equiv \max(0, \mathbf{z}\pi_3 + v_3),$$

where  $v_3 \equiv u_3 + \alpha_{31}v_1 + \alpha_{32}v_2$ . Under the assumptions given,  $v_3$  is independent of  $\mathbf{z}$  and normally distributed. Thus, if we knew  $\delta_1$  and  $\delta_2$ , we could consistently estimate  $\alpha_{31}$ ,  $\alpha_{32}$ , and  $\delta_3$  from a Tobit of  $y_3$  on  $\mathbf{z}\delta_1$ ,  $\mathbf{z}\delta_2$ , and  $\mathbf{z}_3$ . From the usual argument, consistent estimators are obtained by using initial consistent estimators of  $\delta_1$  and  $\delta_2$ . Estimation of  $\delta_2$  is simple: just use OLS using the entire sample. Estimation of  $\delta_1$  follows exactly as in Procedure 19.3 using the system

$$y_1 = \mathbf{z}\delta_1 + v_1$$

$$y_3 = \max(0, \mathbf{z}\pi_3 + v_3),$$

where  $y_1$  is observed only when  $y_3 > 0$ .

Given  $\hat{\delta}_1$  and  $\hat{\delta}_2$ , form  $\mathbf{z}_i\hat{\delta}_1$  and  $\mathbf{z}_i\hat{\delta}_2$  for each observation  $i$  in the sample. Then, obtain  $\hat{\alpha}_{31}$ ,  $\hat{\alpha}_{32}$ , and  $\hat{\delta}_3$  from the Tobit

$$y_{i3} \text{ on } (\mathbf{z}_i\hat{\delta}_1), (\mathbf{z}_i\hat{\delta}_2), \mathbf{z}_{i3}$$

using all observations.

For identification,  $(\mathbf{z}\delta_1, \mathbf{z}\delta_2, \mathbf{z}_3)$  can contain no exact linear dependencies. Necessary is that there must be at least two elements in  $\mathbf{z}$  not also in  $\mathbf{z}_3$ .

Obtaining the correct asymptotic variance matrix is complicated. It is most easily done in a generalized method of moments framework. Alternatively, it is easy to use bootstrap resampling on both steps of the estimation procedure.

b. This is not very different from part a. The only difference is that  $\delta_2$  must be estimated using Procedure 19.3. Then follow the steps from part a.

c. We need to estimate the variance of  $u_3$ ,  $\sigma_3^2$ , and then use the standard formula for the mean of a Tobit model. This gives the ASF as a function of  $(y_2, y_3, \mathbf{z}_3)$  and the parameters  $(\alpha_{31}, \alpha_{32}, \delta_3, \sigma_3^2)$ .

**19.11.** a. This follows from the usual iterated expectations argument, because  $\mathbf{Z}_i$  is a function of  $\mathbf{x}_i$ :

$$\begin{aligned} E[s_i \mathbf{Z}_i' \mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o)] &= E\{E[s_i \mathbf{Z}_i' \mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o) | \mathbf{x}_i, s_i]\} \\ &= E\{s_i \mathbf{Z}_i' [\mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o) | \mathbf{x}_i, s_i]\} = \mathbf{0} \end{aligned}$$

because  $E[\mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o) | \mathbf{x}_i, s_i] = \mathbf{0}$ .

b. We modify equation (14.24) from Chapter 14 to allow for selection:

$$\min_{\boldsymbol{\theta} \in \Theta} \left( \sum_{i=1}^N s_i \mathbf{Z}_i' \mathbf{r}_i(\boldsymbol{\theta}) \right)' \left( N^{-1} \sum_{i=1}^N s_i \mathbf{Z}_i' \mathbf{Z}_i \right)^{-1} \left( \sum_{i=1}^N s_i \mathbf{Z}_i' \mathbf{r}_i(\boldsymbol{\theta}) \right).$$

For consistency, we would have to assume that  $\text{rank } E(s_i \mathbf{Z}_i' \mathbf{Z}_i) = L$  – which means, that in the selected sample, the instrument matrix is not perfectly collinear – and we have to assume that  $\boldsymbol{\theta}_o$  is the unique solution to  $E[s_i \mathbf{Z}_i' \mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o)] = \mathbf{0}$ . For  $\sqrt{N}$ -asymptotic normality, we would also have to assume that  $\text{rank } E[s_i \mathbf{Z}_i' \nabla_{\boldsymbol{\theta}} \mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta}_o)] = P$ , the dimension of  $\boldsymbol{\theta}$ . None of the conditions can be true unless  $P(s_i = 1) > 0$ , that is, we observe a randomly drawn observation with positive probability. But  $P(s_i = 1) > 0$  is not nearly sufficient, as we might not have identification in the selected population even if we have identification in the full population. (For example, we might have an instrument that varies sufficiently in the full population but not in the  $s = 1$  subpopulation.)

c. Let  $\check{\boldsymbol{\theta}}$  denote the (system) nonlinear 2SLS estimator on the selected sample. For the minimum chi-square estimator, we would compute

$$\hat{\Lambda} = N^{-1} \sum_{i=1}^N s_i \mathbf{Z}_i' \mathbf{r}_i(\check{\boldsymbol{\theta}}) \mathbf{r}_i(\check{\boldsymbol{\theta}})' \mathbf{Z}_i$$

and then solve

$$\min_{\boldsymbol{\theta} \in \Theta} \left( \sum_{i=1}^N s_i \mathbf{Z}_i' \mathbf{r}_i(\boldsymbol{\theta}) \right)' \hat{\Lambda}^{-1} \left( \sum_{i=1}^N s_i \mathbf{Z}_i' \mathbf{r}_i(\boldsymbol{\theta}) \right).$$

**19.12.** a. Take the expected value of (19.56) conditional on  $(\mathbf{z}, y_3)$  :

$$\begin{aligned} E(y_1 | \mathbf{z}, y_3) &= \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 E(y_2 | \mathbf{z}, y_3) + E(u_1 | \mathbf{z}, y_3) \\ &= \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 E(y_2 | \mathbf{z}, y_3) \end{aligned}$$

because  $E(u_1 | \mathbf{z}, y_3) = 0$  follows from  $E(u_1 | \mathbf{z}, y_3) = 0$ .

b. Now take the expected value of (19.56) conditional on  $(\mathbf{z}, v_3)$ :

$$\begin{aligned} E(y_1|\mathbf{z}, v_3) &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 E(y_2|\mathbf{z}, v_3) + E(u_1|\mathbf{z}, v_3) \\ &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 [\mathbf{z}\boldsymbol{\delta}_2 + E(v_2|\mathbf{z}, v_3)] \\ &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 (\mathbf{z}\boldsymbol{\delta}_2 + \gamma_2 v_3). \end{aligned}$$

Therefore,

$$E(y_1|\mathbf{z}, y_3) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 [\mathbf{z}\boldsymbol{\delta}_2 + \gamma_2 E(v_3|\mathbf{z}, y_3)],$$

and when  $y_3 = 1$  we get the usual inverse Mill's ratio:  $E(v_3|\mathbf{z}, y_3 = 1) = \lambda(\mathbf{z}\boldsymbol{\delta}_3)$ . So

$$E(y_1|\mathbf{z}, y_3 = 1) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 [\mathbf{z}\boldsymbol{\delta}_2 + \gamma_2 \lambda(\mathbf{z}\boldsymbol{\delta}_3)].$$

c. We can view it as a three-step estimation method. The first step is to obtain  $\hat{\boldsymbol{\delta}}_3$  from probit of  $y_{i3}$  on  $\mathbf{z}_i$ , using all of the observations. Then, we can estimate  $\boldsymbol{\delta}_2$  and  $\gamma_2$  from standard Heckit applied to  $y_{i2}$  using the selection sample. (My initial thought was that the two steps in the Heckit method are treated as one, as it could be carried out by partial MLE.) Given  $\hat{\boldsymbol{\delta}}_3$ ,  $\hat{\boldsymbol{\delta}}_2$ , and  $\hat{\gamma}_2$ , the final stage is the OLS regression

$$y_{i1} \text{ on } \mathbf{z}_{i1}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2 + \hat{\gamma}_2 \lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_3)$$

using the  $s_{i1} = 1$  sample. Note that the final regressor,  $\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_3)$ , is simply our estimate of

$E(y_2|\mathbf{z}, y_3 = 1)$ . Intuitively, if there is one relevant element in  $\mathbf{z}_i$  not in  $\mathbf{z}_{i1}$ , then

$E(y_{i2}|\mathbf{z}_i, y_{i3} = 1)$  has sufficient variation apart from  $\mathbf{z}_{i1}$  to identify  $\boldsymbol{\delta}_1$  and  $\alpha_1$ . However, I did

overlook one issue when I wrote this problem: we cannot get a very good estimate of  $\boldsymbol{\delta}_2$ , or  $\gamma_2$

for that matter, in the preliminary Heckit unless we can set an element of  $\boldsymbol{\delta}_2$  equal to zero. In

other words, we would really need an exclusion restriction in the reduced form of  $y_2$  in order to

get a good Heckit estimate of  $\boldsymbol{\delta}_2$ . Thus, this procedure seems no better – and perhaps even

worse – than Procedure 19.2, even when we assume  $E(u_1|\mathbf{z}, v_3) = 0$ .



If  $y_2$  is always observed, then we can estimate  $\delta_2$  by a first-stage OLS regression, and we could then estimate  $\gamma_2$  precisely, also, without resorting to an exclusion restriction in the reduced form of  $y_2$ .

d. Unlike Procedure 19.2, the method in part c does not work if  $E(u_1|\mathbf{z}, y_3) \neq 0$ . Therefore, there is little to recommend it.

e. If  $E(u_1|\mathbf{z}, y_2, y_3) = 0$ , we would just use OLS on the selected sample:  $y_{i1}$  on  $\mathbf{z}_i, y_{i2}$ .

**19.13.** a. There is no sample selection problem because, by definition, you have specified the distribution of  $y$  given  $\mathbf{x}$  and  $y > 0$ . We only need to obtain a random sample from the subpopulation with  $y > 0$ .

b. Again, there is no sample selection bias because we have specified the conditional expectation for the population of interest. If we have a random sample from that population, NLS is generally consistent and  $\sqrt{N}$ -asymptotically normal.

c. We would use a standard probit model. Let  $w = 1[y > 0]$ . Then  $w$  given  $\mathbf{x}$  follows a probit model with  $P(w = 1|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\gamma})$ .

d.  $E(y|\mathbf{x}) = P(y > 0|\mathbf{x}) \cdot E(y|\mathbf{x}, y > 0) = \Phi(\mathbf{x}\boldsymbol{\gamma}) \cdot \exp(\mathbf{x}\boldsymbol{\beta})$ . So we would plug in the NLS estimator of  $\boldsymbol{\beta}$  and the probit estimator of  $\boldsymbol{\gamma}$ .

e. By definition, there is no sample selection problem when you specify the conditional distribution – conditional means – for the second part. As discussed in Section 17.6.3, confusion can arise when two part models are specified with unobservables that may be correlated, as in equation (17.50):

$$\begin{aligned} y &= s \cdot \exp(\mathbf{x}\boldsymbol{\beta} + u), \\ s &= 1[\mathbf{x}\boldsymbol{\gamma} + v > 0], \end{aligned}$$

so that  $s = 0 \Leftrightarrow y = 0$ . As shown in Section 17.6.3, if  $u$  and  $v$  are correlated then estimation of

$\beta$  does use methods that are closely related to the Heckman sample selection correction. But  $\beta$  does not tell us what we need to know because both  $E(y|\mathbf{x})$  and  $E(y|\mathbf{x}, y > 0)$  are much more complicated than in the truncated normal or lognormal hurdle cases. See Section 17.6.3 for further discussion.

**19.14.** a. Write  $u = \alpha_0(1 - s) + \alpha_1 s + e$  where, by assumption,  $E(e|\mathbf{z}, s) = 0$ . Plugging this expression for  $u$  into (19.30) gives

$$\begin{aligned} y &= \beta_1 + \beta_2 x_2 + \dots + \beta_K x_K + \alpha_0(1 - s) + \alpha_1 s + e \\ E(e|\mathbf{z}, s) &= 0. \end{aligned}$$

Using the selected sample and applying IV corresponds to multiplying the equation through by  $s$ , and then applying 2SLS. We have

$$\begin{aligned} s \cdot y &= \beta_1 s + \beta_2 (s \cdot x_2) + \dots + \beta_K (s \cdot x_K) + \alpha_1 s + s \cdot e \\ &= (\alpha_1 + \beta_1) s + \beta_2 (s \cdot x_2) + \dots + \beta_K (s \cdot x_K) + s \cdot e, \end{aligned}$$

where we use  $s(1 - s) = 0$  and  $s^2 = s$ . Because  $E(s \cdot e|\mathbf{z}, s) = 0$ , it follows that, under the rank conditions in Theorem 19.1, 2SLS applied to the selected sample consistently estimates  $(\alpha_1 + \beta_1), \beta_2, \dots, \beta_K$ .

b. This is not so much a “show” question as it is just recognizing a basic property of conditional expectations: if  $(u, s)$  is independent of  $\mathbf{z}$ , then  $E(u|\mathbf{z}, s) = E(u|s)$ . Because we are willing to assume something like independence between  $u$  and  $\mathbf{z}$  (or, at least, a zero conditional mean), the important assumption would be independence between  $s$  and  $\mathbf{z}$ . But if the mean of the unobservable,  $u$ , changes with  $s$ , why would we assume that the mean of the exogenous observables,  $E(\mathbf{z}|s)$ , does not? Even  $E(\mathbf{z}|s) = E(\mathbf{z})$  is a strong assumption, let alone full independence between  $\mathbf{z}$  and  $s$ .

**19.15.** a. We cannot use censored Tobit because that requires observing  $\mathbf{x}$  when whatever

the value of  $y$ . Instead, we can use truncated Tobit: we use the distribution of  $y$  given  $\mathbf{x}$  and  $y > 0$ . If we observed  $\mathbf{x}$  always then using the truncated Normal regression model would be inefficient, but censored Tobit for  $D(y|\mathbf{x})$  implies truncated Tobit for  $D(y|\mathbf{x}, y > 0)$ .

b. Because we have assumed  $y$  given  $\mathbf{x}$  follows a standard Tobit,  $E(y|\mathbf{x})$  is the parametric function

$$E(y|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\mathbf{x}\boldsymbol{\beta} + \sigma\phi(\mathbf{x}\boldsymbol{\beta}/\sigma).$$

Therefore, even though we never observe some elements of  $\mathbf{x}$  when  $y = 0$ , we can still estimate  $E(y|\mathbf{x})$  because we can estimate  $\boldsymbol{\beta}$  and  $\sigma$  and we have an expression for  $E(y|\mathbf{x})$  that (we assume) holds for all  $\mathbf{x}$ . To estimate  $\boldsymbol{\beta}$  and  $\sigma^2$  We do have to assume that  $\mathbf{x}$  varies enough in the subpopulation where  $y > 0$ , namely,  $\text{rank } E(\mathbf{x}'\mathbf{x}|y > 0) = K$ . In the case where an element of  $\mathbf{x}$  is a derived price, we need sufficient price variation for the population that consumes some of the good.

**19.16.** a. To obtain the expected value of

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1$$

conditional on  $(\mathbf{z}, r_2, v_2)$ , use the fact that  $y_2$  is a function of  $(\mathbf{z}, v_2)$ , and use independence of  $(u_1, v_2)$  and  $\mathbf{z}$ :

$$\begin{aligned} E(y_1|\mathbf{z}, r_2, v_2) &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + E(u_1|\mathbf{z}, r_2, v_2) \\ &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + E(u_1|v_2). \end{aligned}$$

Now use the linearity assumption  $E(u_1|v_2) = \rho_1 v_2$  to get

$$E(y_1|\mathbf{z}, r_2, v_2) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 v_2.$$

b. With  $s_2 = 1[y_2 < w_2]$ ,  $s_2$  is clearly a function of  $(\mathbf{z}, r_2, v_2)$ , and so  $s_2$  is redundant in  $E(y_1|\mathbf{z}, r_2, v_2, s_2)$ :

$$E(y_1|\mathbf{z}, r_2, v_2, s_2) = E(y_1|\mathbf{z}, r_2, v_2) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 v_2.$$

c. Because of part b, if we could observe  $v_{i2}$  whenever  $s_{i2} = 1$  we could consistently estimate  $\boldsymbol{\delta}_1$ ,  $\alpha_1$ , and  $\rho_1$  by running the regression

$$y_{i1} \text{ on } \mathbf{z}_{i1}, y_{i2}, v_{i2} \text{ if } s_{i2} = 1.$$

Naturally, we can replace  $v_{i2} = y_{i2} - \mathbf{z}_i\boldsymbol{\delta}_2$  with  $\hat{v}_{i2} \equiv y_{i2} - \mathbf{z}_i\hat{\boldsymbol{\delta}}_2$  for a consistent estimator  $\hat{\boldsymbol{\delta}}_2$  of  $\boldsymbol{\delta}_2$ . That estimator should be from a censored normal regression using

$$w_{i2} = \min(r_{i2}, \mathbf{z}_i\boldsymbol{\delta}_2 + v_{i2})$$

and then defining

$$\hat{v}_{i2} \equiv y_{i2} - \mathbf{z}_i\hat{\boldsymbol{\delta}}_2 \text{ if } y_{i2} < r_{i2}.$$

Then run the regression

$$y_{i1} \text{ on } \mathbf{z}_{i1}, y_{i2}, \hat{v}_{i2} \text{ if } s_{i2} = 1.$$

We can use the delta method to obtain valid standard errors, or bootstrap both steps of the procedure. A simple test of

**19.17.** a. The assumption is that, conditional on  $(\mathbf{x}_i, c_i)$ ,  $u_{it}$  is independent of the entire history of censoring values,  $(r_{i1}, r_{i2}, \dots, r_{iT})$ . This is a kind of strict exogeneity assumption on the censoring, which rules out the censoring values being related to current or past shocks to  $y$ . It does allow censoring to be arbitrarily correlated with heterogeneity  $c_i$ .

b. Substitute for  $y_{it}$  to get

$$w_{it} = 1[\mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} > r_{it}]$$

and then substitute for  $c_i$  to get

$$\begin{aligned}
w_{it} &= 1[\mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\xi + \eta\bar{r}_i + a_i + u_{it} > r_{it}] \\
&= 1[(a_i + u_{it}) > r_{it} - (\mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\xi + \eta\bar{r}_i)] \\
&= 1\left[\frac{(a_i + u_{it})}{(\sigma_a^2 + \sigma_u^2)^{1/2}} > \frac{r_{it} - (\mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\xi + \eta\bar{r}_i)}{(\sigma_a^2 + \sigma_u^2)^{1/2}}\right].
\end{aligned}$$

Now use the fact that  $D(a_i + u_{it}|\mathbf{x}_i, \mathbf{r}_i) \sim \text{Normal}(0, \sigma_a^2 + \sigma_u^2)$ :

$$\begin{aligned}
P(w_{it} = 1|\mathbf{x}_i, \mathbf{r}_i) &= 1 - \Phi\left[\frac{r_{it} - (\mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\xi + \eta\bar{r}_i)}{(\sigma_a^2 + \sigma_u^2)^{1/2}}\right] \\
&= \Phi\left[\frac{\mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\xi + \eta\bar{r}_i - r_{it}}{(\sigma_a^2 + \sigma_u^2)^{1/2}}\right] \\
&\equiv \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_{au} + \psi_{au} + \bar{\mathbf{x}}_i\xi_{au} + \eta_{au}\bar{r}_i + \gamma_{au}r_{it})
\end{aligned}$$

where  $\boldsymbol{\beta}_{au} = \boldsymbol{\beta}/(\sigma_a^2 + \sigma_u^2)^{1/2}$ ,  $\psi_{au} = \psi/(\sigma_a^2 + \sigma_u^2)^{1/2}$ ,  $\xi_{au} = \xi/(\sigma_a^2 + \sigma_u^2)^{1/2}$ , and

$$\gamma_{au} = -1/(\sigma_a^2 + \sigma_u^2).$$

c. From part b, we can estimate all of the scaled coefficients, including  $\gamma_{au}$ , by pooled probit, provided  $\{\mathbf{x}_{it}\}$  and  $\{r_{it}\}$  have time variation for at least some units. But

$$\boldsymbol{\beta} = -\boldsymbol{\beta}_{au}/\gamma_{au}$$

and so we just use

$$\hat{\boldsymbol{\beta}} = -\hat{\boldsymbol{\beta}}_{au}/\hat{\gamma}_{au}.$$

d. The pooled estimation from part c only allows us to estimate  $\sigma_a^2 + \sigma_u^2$  and the unscaled parameters. If we add the assumption that  $\{u_{it} : t = 1, 2, \dots, T\}$  are independent then  $\text{Cov}(a_i + u_{it}, a_i + u_{is}) = \text{Var}(a_i) = \sigma_a^2$  for all  $t \neq s$ . We can use a slight modification of correlated random effects probit, which takes the idiosyncratic error to have unit variance. To this end, write

$$\begin{aligned}
w_{it} &= 1[\mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\xi + \eta\bar{r}_i + a_i + u_{it} > r_{it}] \\
&= 1\left[\frac{(\mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\xi + \eta\bar{r}_i - r_{it})}{\sigma_u} > \frac{(a_i + u_{it})}{\sigma_u}\right] \\
&\equiv 1\left[\frac{(\mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\xi + \eta\bar{r}_i - r_{it})}{\sigma_u} > g_i + e_{it}\right]
\end{aligned}$$

where  $e_{it} = u_{it}/\sigma_u$  and  $g_i = a_i/\sigma_u$ . This shows that if we apply the CRE probit model to  $w_{it}$  on  $(\mathbf{x}_{it}, 1, \bar{\mathbf{x}}_i, \bar{r}_i, r_{it})$  we consistently estimate  $\boldsymbol{\beta}_u = \boldsymbol{\beta}/\sigma_u$ ,  $\psi_u = \psi/\sigma_u$ ,  $\xi_u = \xi/\sigma_u$ , and  $\gamma_u = -1/\sigma_u$  as the coefficients and

$$\text{Var}(g_i) = \sigma_a^2/\sigma_u^2$$

as the heterogeneity variance. Thus, we can recover the original unscaled coefficients,

$\sigma_u^2 = 1/\gamma_u^2$ , and

$$\sigma_a^2 = \sigma_g^2/\gamma_u^2.$$

**19.18.** a. Conditional on  $y > 0$ ,  $y$  follows a truncated normal distribution. So truncated normal regression would consistently estimate  $\boldsymbol{\beta}$  and  $\sigma^2$ .

b. Because we are claiming that  $D(y|\mathbf{x})$  follows a type I Tobit in the population, we use the expected value derived from that assumption. Namely,

$$E(y|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\mathbf{x}\boldsymbol{\beta} + \sigma\phi(\mathbf{x}\boldsymbol{\beta}/\sigma).$$

and then we compute derivatives and changes with respect to  $x_j$ , as usual.

This differs from a hurdle model because we do not have a separate model for  $P(y = 0|\mathbf{x})$ ; we assume this is also governed by the Tobit model, so  $P(y = 0|\mathbf{x}) = 1 - \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)$ .

c. We could not estimate a hurdle model in this case because we have no data when  $y = 0$ . We have not sampled from that part of the population, and so we cannot estimate a model for  $P(s = 1|\mathbf{x})$  where  $s = 1[y > 0]$ .

**19.19.** a. First, if  $r_i = 0$  the observation contains no information for estimating the distribution  $D(y_i|\mathbf{x}_i)$  because then  $P(w_i = 0|\mathbf{x}_i) = 1$  regardless of  $D(y_i|\mathbf{x}_i)$ . So what follows is only relevant for  $r_i \in \{1, 2, 3, \dots\}$ .

For  $0 \leq w < r_i$ ,

$$P(w_i = w|\mathbf{x}_i) = P(y_i = w|\mathbf{x}_i) = f_y(w|\mathbf{x}_i).$$

Next,  $w_i = r_i$  if and only if  $y_i \geq r_i$ , and so

$$\begin{aligned} P(w_i = r_i|\mathbf{x}_i) &= 1 - P(y_i < r_i|\mathbf{x}_i) = 1 - P(y_i \leq r_i - 1|\mathbf{x}_i) \\ &= 1 - F_y(r_i - 1|\mathbf{x}_i). \end{aligned}$$

We can write the conditional density of  $w_i$  as

$$f_w(w|\mathbf{x}_i, r_i) = [f_y(w|\mathbf{x}_i)]^{1[w < r_i]} [1 - F_y(r_i - 1|\mathbf{x}_i)]^{1[w = r_i]}, \quad w = 0, \dots, r_i.$$

b. In the Poisson case with an exponential mean, the conditional density of  $y_i$  is

$$f_y(y|\mathbf{x}_i; \boldsymbol{\beta}) = \frac{\exp[-\exp(\mathbf{x}_i\boldsymbol{\beta})][\exp(\mathbf{x}_i\boldsymbol{\beta})]^y}{y!}$$

and the cdf is

$$F_y(y|\mathbf{x}_i) = \exp[-\exp(\mathbf{x}_i\boldsymbol{\beta})] \sum_{h=0}^y \frac{[\exp(\mathbf{x}_i\boldsymbol{\beta})]^h}{h!}$$

Now just plug this into the general formula in part a.

c. Maximum likelihood estimators based on censored data are generally not robust to misspecification of the underlying population – even when that distribution is in the linear exponential family. (The log likelihood for the censored variable is not in the linear exponential family; even if it were,  $E(w_i|\mathbf{x}_i)$  depends on the underlying distribution.) Just like censored regression with a normal distribution is not robust for estimating the mean parameters

under nonnormality, neither is censored regression with a Poisson distribution. One way to see this is to write down the score for the general case and observe that just having  $E(y_i|\mathbf{x}_i)$  correctly specified will not imply that the score has zero expectation.

d. As we know from earlier chapters, nonlinear least squares, Poisson QMLE, and other QMLEs in the LEF are robust for estimating the mean parameters. Thus, if there were no data censoring, we could use the Poisson QMLE to estimate  $\beta$ . With data censoring,  $E(w_i|\mathbf{x}_i)$  always depends on the underlying population distribution. Thus, in general we need to specify  $D(y_i|\mathbf{x}_i)$  even if we are primarily interested in  $E(y_i|\mathbf{x}_i)$ .

e. Because data censoring requires us to have  $D(y_i|\mathbf{x}_i)$  correctly specified, a strong case can be made for specifying flexible models for  $D(y_i|\mathbf{x}_i)$  – even if we are primarily interested in  $E(y_i|\mathbf{x}_i)$ . For example, if we use a NegBin I or NegBin II model, these at least include the Poisson as (limiting) special cases. So, if we are pretty sure the underlying population has overdispersion, we can use one of these distributions in accounting for the right censoring. Ideally we would have a distribution that allows underdispersion, too.



## Solutions to Chapter 20 Problems

**20.1.** a. Just use calculus and set the derivative to zero:

$$\sum_{i=1}^{N_0} p_{j_i}^{-1} (w_i - \hat{\mu}_w) = 0$$

or

$$\sum_{i=1}^{N_0} p_{j_i}^{-1} w_i = \sum_{i=1}^{N_0} p_{j_i}^{-1} \hat{\mu}_w = \left( \sum_{i=1}^{N_0} p_{j_i}^{-1} \right) \hat{\mu}_w.$$

Solving for  $\hat{\mu}_w$  gives

$$\hat{\mu}_w = \left( \sum_{i=1}^{N_0} p_{j_i}^{-1} \right)^{-1} \sum_{i=1}^{N_0} p_{j_i}^{-1} w_i = \sum_{i=1}^{N_0} v_{ji} w_i$$

where

$$v_{ji} = \left( \sum_{h=1}^{N_0} p_{j_h}^{-1} \right)^{-1} p_{j_i}^{-1}.$$

b. From equation (20.7),

$$P(s_i = 1 | \mathbf{z}_i, w_i) = P(s_i = 1 | \mathbf{z}_i) = p_{1z_{i1}} + \dots + p_{Jz_{iJ}}$$

which implies

$$E(s_i | j_i, w_i) = p_{j_i}$$

because the stratum for observation  $i$  is  $j_i$  if and only if  $z_{ij} = 1$ . Now

$$E \left[ N^{-1} \sum_{i=1}^N (s_i / p_{j_i}) w_i \right] = N^{-1} \sum_{i=1}^N E[(s_i / p_{j_i}) w_i]$$

and, by iterated expectations,

$$\begin{aligned}
E[(s_i/p_{j_i})w_i] &= E\{E[(s_i/p_{j_i})w_i|j_i, w_i]\} \\
&= E\{[E(s_i|j_i, w_i)/p_{j_i}]]w_i\} \\
&= E[(p_{j_i}/p_{j_i})w_i] = E(w_i) = \mu_o.
\end{aligned}$$

This shows  $E(\tilde{\mu}_w) = \mu_o$ .

c. Notice that  $\tilde{\mu}_w$  depends on  $N$ , the number of times we sampled the population, including when we did not record the observation. By contrast, to obtain  $\hat{\mu}_w$  we need only need information on the sampling weights and data on the units actually kept. Therefore, in addition to knowing the sampling probabilities,  $\tilde{\mu}_w$  requires the extra information that we know how many observations were discarded by the VP sampling scheme.

**20.2.** Write the log likelihood for all  $N$  observations as

$$\sum_{i=1}^N \sum_{h=1}^J z_{ih} [s_i \log(p_h) + (1 - s_i) \log(1 - p_h)].$$

For a given  $j \in \{1, 2, \dots, J\}$ , take the derivative with respect to  $p_j$ , and set the result to zero:

$$\sum_{i=1}^N z_{ij} \left[ \frac{s_i}{\hat{p}_j} - \frac{(1 - s_i)}{(1 - \hat{p}_j)} \right] \equiv 0$$

or, by obtaining a common denominator,

$$\sum_{i=1}^N z_{ij} \left[ \frac{(1 - \hat{p}_j)s_i - \hat{p}_j(1 - s_i)}{\hat{p}_j(1 - \hat{p}_j)} \right] \equiv 0.$$

Of course, the problem only makes sense for interior solutions  $0 < \hat{p}_j < 1$  so the first order condition is equivalent to

$$\sum_{i=1}^N [(1 - \hat{p}_j)z_{ij}s_i - \hat{p}_jz_{ij}(1 - s_i)] \equiv 0.$$

Simple algebra gives

$$\sum_{i=1}^N z_{ij} s_i = \sum_{i=1}^N \hat{p}_j z_{ij}$$

or

$$\hat{p}_j = \frac{\sum_{i=1}^N z_{ij} s_i}{\sum_{i=1}^N z_{ij}} \equiv \frac{M_j}{N_j}.$$

**20.3.** a. To be specific, consider the case of variable probability sampling, where the probability weights are

$$p(\mathbf{z}_i) = p_1 z_{i1} + \dots + p_J z_{iJ} = P(s_i = 1 | \mathbf{z}_i, \mathbf{w}_i)$$

where  $\mathbf{w}_i = (\mathbf{x}_i, y_i)$ . We can write the IPW nonlinear least squares objective function as

$$\min_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^N \frac{s_i}{p(\mathbf{z}_i)} [y_i - m(\mathbf{x}_i, \boldsymbol{\theta})]^2,$$

which is for form useful for studying asymptotic properties. (For the asymptotic distribution theory, we divide the objective function by two to make the notation easier.)

b. For VP sampling, we have already assumed that each  $p_j > 0$ , and, because we can write

$$|s_i/p(\mathbf{z}_i)| \leq \max(p_1^{-1}, \dots, p_J^{-1})$$

it follows that the regularity conditions sufficient for consistency of NLS on a random sample are also sufficient for NLS on a VP sample: the objective function is still continuous and the moment conditions do not need to be changed because

$$\frac{s_i}{p(\mathbf{z}_i)} [y_i - m(\mathbf{x}_i, \boldsymbol{\theta})]^2 \leq \max(p_1^{-1}, \dots, p_J^{-1}) [y_i - m(\mathbf{x}_i, \boldsymbol{\theta})]^2.$$

Further, we know generally that

$$E \left\{ \frac{s_i}{p(\mathbf{z}_i)} [y_i - m(\mathbf{x}_i, \boldsymbol{\theta})]^2 \right\} = E \{ [y_i - m(\mathbf{x}_i, \boldsymbol{\theta})]^2 \}$$

and so if  $\boldsymbol{\theta}_o$  uniquely minimizes the right hand side, it uniquely minimizes the left hand side, too.

c. The theory in Section 19.8 can be applied directly. In particular, we can use equation (19.90) because the probabilities are known, not estimated. In the formula,

$$\begin{aligned}\mathbf{A}_o &= E[\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}_o)] = E[\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o)] \\ &= E[\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta}_o)]\end{aligned}$$

and

$$\begin{aligned}\mathbf{B}_o &= E\left\{ \frac{s_i}{[p(\mathbf{z}_i)]^2} \nabla_{\boldsymbol{\theta}} q(\mathbf{x}_i, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} q(\mathbf{x}_i, \boldsymbol{\theta}_o) \right\} \\ &= E\left\{ \frac{s_i u_i^2}{[p(\mathbf{z}_i)]^2} \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta}_o) \right\}\end{aligned}$$

where  $u_i = y_i - m(\mathbf{x}_i, \boldsymbol{\theta}_o)$ . We can consistently estimate  $\mathbf{A}_o$  as

$$\hat{\mathbf{A}} = N^{-1} \sum_{i=1}^N \left[ \frac{s_i}{p(\mathbf{z}_i)} \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w) \right]$$

and  $\mathbf{B}_o$  as

$$\hat{\mathbf{B}} = N^{-1} \sum_{i=1}^N \left[ \frac{s_i \hat{u}_i^2}{[p(\mathbf{z}_i)]^2} \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w) \right]$$

where  $\hat{u}_i = y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w)$  are the residuals. Then

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}_w) = \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{-1} / N$$

(which does not actually require knowing  $N$ , as it cancels everywhere).

d. The formula does not generally simplify because  $E(u_i^2 | \mathbf{x}_i, \mathbf{z}_i)$  might depend on  $\mathbf{z}_i$  even if  $\text{Var}(u_i | \mathbf{x}_i) = \sigma_o^2$ . [In fact, we do not even assume that  $E(u_i | \mathbf{x}_i, \mathbf{z}_i) = 0$  in this problem because the stratification may be endogenous.]

e. If  $m(\mathbf{x}, \boldsymbol{\theta})$  is misspecified we must use a more general estimator for  $\mathbf{A}^*$  based on

$$\mathbf{A}^* = E[\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}^*)] = E[\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta}^*)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta}^*) - u_i^* \nabla_{\boldsymbol{\theta}}^2 m(\mathbf{x}_i, \boldsymbol{\theta}^*)]$$

where  $\boldsymbol{\theta}^*$  is the pseudo-true value of  $\boldsymbol{\theta}$  that solves the population minimization problem and

$u_i^* = y_i - m(\mathbf{x}_i, \boldsymbol{\theta}^*)$ . Our estimator of  $\mathbf{A}^*$  is

$$\hat{\mathbf{A}} = N^{-1} \sum_{i=1}^N \left\{ \frac{S_i}{p(\mathbf{z}_i)} [\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w) - \hat{u}_i \nabla_{\boldsymbol{\theta}}^2 m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w)] \right\}.$$

The estimator of  $\mathbf{B}^*$  can be the same as in part c.

**20.4.** First, we can write the unweighted objective function as

$$\begin{aligned} N^{-1} \sum_{j=1}^J \sum_{i=1}^{N_j} q(\mathbf{w}_{ij}, \boldsymbol{\theta}) &= \sum_{j=1}^J (N_j/N) N_j^{-1} \sum_{i=1}^{N_j} q(\mathbf{w}_{ij}, \boldsymbol{\theta}) \\ &= \sum_{j=1}^J H_j \left( N_j^{-1} \sum_{i=1}^{N_j} q(\mathbf{w}_{ij}, \boldsymbol{\theta}) \right), \end{aligned}$$

as suggested in the hint. Further, by the same argument as on page 860,  $N_j^{-1} \sum_{i=1}^{N_j} q(\mathbf{w}_{ij}, \boldsymbol{\theta})$

converges (uniformly) to  $E[q(\mathbf{w}, \boldsymbol{\theta}) | \mathbf{w} \in \mathcal{W}_j] = E[q(\mathbf{w}, \boldsymbol{\theta}) | \mathbf{x} \in \mathcal{X}_j]$ , where we use the fact that

the strata are determined by the conditioning variables and given by  $\mathcal{X}_1, \dots, \mathcal{X}_J$ . Therefore, if

$H_j \rightarrow \bar{H}_j$  as  $N \rightarrow \infty$  the unweighted objective function converges uniformly to

$$\bar{H}_1 E[q(\mathbf{w}, \boldsymbol{\theta}) | \mathbf{x} \in \mathcal{X}_1] + \dots + \bar{H}_J E[q(\mathbf{w}, \boldsymbol{\theta}) | \mathbf{x} \in \mathcal{X}_J] \quad (20.97)$$

Given that  $\boldsymbol{\theta}_o$  solves (20.15) for each  $\mathbf{x}$ , we can also show  $\boldsymbol{\theta}_o$  minimizes  $E[q(\mathbf{w}, \boldsymbol{\theta}) | \mathbf{x} \in \mathcal{X}_j]$

over  $\Theta$  for each  $j$ : by iterated expectations (since the indicator  $1[\mathbf{x} \in \mathcal{X}_j]$  is a function of  $\mathbf{x}$ ),

$$E[q(\mathbf{w}, \boldsymbol{\theta}) | \mathbf{x} \in \mathcal{X}_j] = E\{E[q(\mathbf{w}, \boldsymbol{\theta}) | \mathbf{x}] | \mathbf{x} \in \mathcal{X}_j\},$$

and if  $\boldsymbol{\theta}_o$  minimizes  $E[q(\mathbf{w}, \boldsymbol{\theta}) | \mathbf{x}]$ , it must also minimize  $E\{E[q(\mathbf{w}, \boldsymbol{\theta}) | \mathbf{x}] | \mathbf{x} \in \mathcal{X}_j\}$ . Therefore,  $\boldsymbol{\theta}_o$

is one minimizer of (20.97) over  $\Theta$ . Now we just have to show it is the unique minimizer if it

uniquely minimizes  $E[q(\mathbf{w}, \boldsymbol{\theta})]$ . Without the assumption  $\bar{H}_j > 0$ ,  $\boldsymbol{\theta}_o$  need not be the unique

minimizer of (20.97). To show uniqueness when each  $\bar{H}_j$  is strictly positive, let  $s_j = 1[\mathbf{x} \in \mathcal{X}_j]$ .

Then we can write, for any  $\boldsymbol{\theta}$ ,

$$E[q(\mathbf{w}, \boldsymbol{\theta})] - E[q(\mathbf{w}, \boldsymbol{\theta}_o)] = \sum_{j=1}^J Q_j \{E[q(\mathbf{w}, \boldsymbol{\theta})|s_j] - E[q(\mathbf{w}, \boldsymbol{\theta}_o)|s_j]\},$$

where the  $Q_j$  are the population frequencies. By assumption, the left hand side is strictly positive when  $\boldsymbol{\theta} \neq \boldsymbol{\theta}_o$ , which means, because  $Q_j > 0$  for all  $j$ ,  $E[q(\mathbf{w}, \boldsymbol{\theta})|s_j] - E[q(\mathbf{w}, \boldsymbol{\theta}_o)|s_j]$  must be strictly positive for at least one  $j$ ; we already know that each difference is nonnegative. This, along with the fact that  $\bar{H}_j > 0, j = 1, \dots, J$ , implies that (20.97) is uniquely minimized at  $\boldsymbol{\theta}_o$ .

**20.5.** a. The Stata output is given below. The variables with “bar” added on denote the district-level averages. Note that we can still use `xtreg` even though this is a cluster sample, not a panel data set. An alternative for obtaining the FE estimates is the `areg` command in Stata. The pooled OLS and FE estimates are identical on all explanatory variables. The pooled OLS standard errors reported below are almost certainly incorrect because they assume no within-district correlation in the unobservables.

```
. reg lavgsal bs lstaff lenroll lunch bsbar lstaffbar lenrollbar lunchbar
```

Source	SS	df	MS	Number of obs =	1848
Model	49.9510474	8	6.24388093	F( 8, 1839) =	228.60
Residual	50.2303314	1839	.027313938	Prob > F =	0.0000
Total	100.181379	1847	.054240054	R-squared =	0.4986
				Adj R-squared =	0.4964
				Root MSE =	.16527

	lavgsal	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
bs		-.4948449	.2199466	-2.25	0.025	-.9262162	-.0634736
lstaff		-.6218901	.0277027	-22.45	0.000	-.6762221	-.5675581
lenroll		-.0515063	.0155411	-3.31	0.001	-.0819865	-.0210262
lunch		.0005138	.0003452	1.49	0.137	-.0001632	.0011908
bsbar		.441438	.2630336	1.68	0.093	-.074438	.9573139
lstaffbar		-.1493942	.0370985	-4.03	0.000	-.2221538	-.0766346
lenrollbar		.0315714	.0184565	1.71	0.087	-.0046266	.0677694

lunchbar		-.0016765	.0003903	-4.30	0.000	-.0024419	-.000911
_cons		13.98544	.141118	99.10	0.000	13.70867	14.26221

---

```
. xtreg lavgsal bs lstaff lenroll lunch, fe
```

Fixed-effects (within) regression	Number of obs	=	1848
Group variable: distid	Number of groups	=	537
R-sq: within	=	0.5486	
between	=	0.3544	
overall	=	0.4567	
	Obs per group: min	=	
	avg	=	3.
	max	=	162
	F(4,1307)	=	397.05
corr(u_i, Xb) = 0.1433	Prob > F	=	0.0000

lavgsal		Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
bs		-.4948449	.133039	-3.72	0.000	-.7558382	-.2338515
lstaff		-.6218901	.0167565	-37.11	0.000	-.6547627	-.5890175
lenroll		-.0515063	.0094004	-5.48	0.000	-.0699478	-.0330648
lunch		.0005138	.0002088	2.46	0.014	.0001042	.0009234
_cons		13.61783	.1133406	120.15	0.000	13.39548	13.84018
sigma_u		.15491886					
sigma_e		.09996638					
rho		.70602068	(fraction of variance due to u_i)				

---

F test that all u\_i=0: F(536, 1307) = 7.24 Prob > F = 0.0000

b. The RE estimates are given below, with and without cluster-robust standard errors. Also, the cluster-robust standard errors for FE are provided. The fully robust standard errors are bigger than the nonrobust ones, suggesting there might be additional within-district correlation even after accounting for an additive district effect.

```
. xtreg lavgsal bs lstaff lenroll lunch bsbar lstaffbar lenrollbar lunchbar, re
```

Random-effects GLS regression	Number of obs	=	1848
Group variable: distid	Number of groups	=	537
R-sq: within	=	0.5486	
between	=	0.4006	
overall	=	0.4831	
	Obs per group: min	=	
	avg	=	3.
	max	=	162
Random effects u_i ~Gaussian	Wald chi2(8)	=	1943.89
corr(u_i, X) = 0 (assumed)	Prob > chi2	=	0.0000

lavgsal		Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
---------	--	-------	-----------	---	------	---------------------	--

---

bs	-.4948449	.1334822	-3.71	0.000	-.7564652	-.2332245
lstaff	-.6218901	.0168123	-36.99	0.000	-.6548417	-.5889385
lenroll	-.0515063	.0094317	-5.46	0.000	-.0699921	-.0330205
lunch	.0005138	.0002095	2.45	0.014	.0001032	.0009244
bsbar	.2998553	.2437798	1.23	0.219	-.1779443	.777655
lstaffbar	-.0255493	.0418946	-0.61	0.542	-.1076613	.0565627
lenrollbar	.0657286	.0157977	4.16	0.000	.0347657	.0966914
lunchbar	-.0007259	.0004022	-1.80	0.071	-.0015143	.0000625
_cons	13.22003	.2136208	61.89	0.000	12.80135	13.63872
-----						
sigma_u	.12627558					
sigma_e	.09996638					
rho	.61473634	(fraction of variance due to u_i)				
-----						

```
. xtreg lavgsal bs lstaff lenroll lunch bsbar lstaffbar lenrollbar lunchbar,
    re cluster(distid)
```

Random-effects GLS regression	Number of obs	=	1848
Group variable: distid	Number of groups	=	537

R-sq: within = 0.5486	Obs per group: min =	
between = 0.4006	avg =	3.
overall = 0.4831	max =	162

Random effects u_i ~Gaussian	Wald chi2(8)	=	556.49
corr(u_i, X) = 0 (assumed)	Prob > chi2	=	0.0000

(Std. Err. adjusted for 537 clusters in distid)

lavgsal	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval	
bs	-.4948449	.1939422	-2.55	0.011	-.8749646	-.1147252
lstaff	-.6218901	.0432281	-14.39	0.000	-.7066157	-.5371645
lenroll	-.0515063	.013103	-3.93	0.000	-.0771876	-.025825
lunch	.0005138	.000213	2.41	0.016	.0000964	.0009312
bsbar	.2998553	.3031961	0.99	0.323	-.2943981	.8941087
lstaffbar	-.0255493	.0651932	-0.39	0.695	-.1533256	.1022269
lenrollbar	.0657286	.020655	3.18	0.001	.0252455	.1062116
lunchbar	-.0007259	.0004378	-1.66	0.097	-.0015839	.0001322
_cons	13.22003	.2556139	51.72	0.000	12.71904	13.72103
sigma_u	.12627558					
sigma_e	.09996638					
rho	.61473634	(fraction of variance due to u_i)				

```
. xtreg lavgsal bs lstaff lenroll lunch, fe cluster(distid)
```

Fixed-effects (within) regression	Number of obs	=	1848
Group variable: distid	Number of groups	=	537

R-sq: within = 0.5486	Obs per group: min =	
between = 0.3544	avg =	3.
overall = 0.4567	max =	162

corr(u_i, Xb) = 0.1433	F(4,536)	=	57.84
	Prob > F	=	0.0000



(Std. Err. adjusted for 537 clusters in distid)						
lavgsal	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
bs	-.4948449	.1937316	-2.55	0.011	-.8754112	-.1142785
lstaff	-.6218901	.0431812	-14.40	0.000	-.7067152	-.5370649
lenroll	-.0515063	.0130887	-3.94	0.000	-.0772178	-.0257948
lunch	.0005138	.0002127	2.42	0.016	.0000959	.0009317
_cons	13.61783	.2413169	56.43	0.000	13.14379	14.09187
sigma_u	.15491886					
sigma_e	.09996638					
rho	.70602068	(fraction of variance due to u_i)				

c. The robust Wald test for joint significance of the four district-level averages gives a strong rejection of the null, with  $p$ -value = .0004. Therefore, we conclude that at least some of the variables are correlated with unobserved district effects.

```
. qui xtreg lavgsal bs lstaff lenroll lunch bsbar lstaffbar lenrollbar lunchbar
    re cluster(distid)

. test  bsbar lstaffbar lenrollbar lunchbar

( 1)  bsbar = 0
( 2)  lstaffbar = 0
( 3)  lenrollbar = 0
( 4)  lunchbar = 0

      chi2( 4) =    20.70
    Prob > chi2 =    0.0004
```

**20.6.** a. Only three schools in the sample have reported benefits/salary ratios of at least .5.

The highest of these is about .66.

```
. count if bs >= .5
    3

. list distid bs if bs >= .5
```

	distid	bs
68.	9030	.6594882
1127.	63160	.5747756
1670.	82040	.5022581



lunch	.0005538	.0003954	1.40	0.162	-.0002217	.0013293
bsbar	.3679283	.3003398	1.23	0.221	-.2211145	.9569712
lstaffbar	-.1374073	.0421226	-3.26	0.001	-.2200204	-.0547941
lenrollbar	.0075581	.0210143	0.36	0.719	-.0336564	.0487726
lunchbar	-.0014894	.0004477	-3.33	0.001	-.0023675	-.0006113
_cons	14.23874	.1612496	88.30	0.000	13.92249	14.55499

20.7. a. Out of 1,683 schools, 922 have all five years of data. The fewest number of years is three. Note that the `tab` command gives includes many more observations than schools because there are multiple years per school.

```
. xtsum math4
```

Variable		Mean	Std. Dev.	Min	Max	Observations
math4	overall	63.57726	20.19047	2.9	100	N = 7150
	between		16.08074	11.75	98.94	n = 1683
	within		12.37335	13.71059	122.3439	T-bar = 4.24837

```
. egen tobs = sum(1), by(schid)
```

```
. count if tobs == 5 & y98
922
```

```
. tab tobs
```

tobs	Freq.	Percent	Cum.
3	1,512	21.15	21.15
4	1,028	14.38	35.52
5	4,610	64.48	100.00
Total	7,150	100.00	

b. The pooled OLS estimates, with all time averages included, and the fixed effects estimates – with so-called “school fixed effects” – are given below. Variables with a “b” at the end are the within-school time averages. As expected, they are identical, including the coefficients on the year dummies.

The coefficient on *lunchb* is  $-.426$ , and its fully robust *t* statistic is  $-11.76$ . Therefore, the average poverty level over the available years has a very large effect on the math pass rate: a ten percentage point increase in the average poverty rate predicts a pass rate that is about 4.3 percentage points lower.

```
. reg math4 lavgrexp lunch lenrol y95 y96 y97 y98 lavgrexpb lunchb lenrolb
y95b y96b y97b y98b, cluster(distid)
```

Linear regression

Number of obs = 7150  
F( 14, 466) = 182.55  
Prob > F = 0.0000  
R-squared = 0.4147  
Root MSE = 15.462

(Std. Err. adjusted for 467 clusters in distid)

math4	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
lavgrexp	6.288376	3.13387	2.01	0.045	.1301085	12.44664
lunch	-.0215072	.0399402	-0.54	0.590	-.0999924	.056978
lenrol	-2.038461	2.099636	-0.97	0.332	-6.164387	2.087466
y95	11.6192	.7213934	16.11	0.000	10.20162	13.03679
y96	13.05561	.9331425	13.99	0.000	11.22192	14.8893
y97	10.14771	.9581113	10.59	0.000	8.264957	12.03046
y98	23.41404	1.027817	22.78	0.000	21.39431	25.43377
lavgrexpb	2.7178	4.04162	0.67	0.502	-5.224258	10.65986
lunchb	-.4256461	.0361912	-11.76	0.000	-.4967642	-.3545279
lenrolb	.2880016	2.17219	0.13	0.895	-3.9805	4.556503
y95b	21.26329	15.95857	1.33	0.183	-10.09639	52.62297
y96b	15.69885	6.523566	2.41	0.016	2.879602	28.5181
y97b	20.66597	15.71006	1.32	0.189	-10.20536	51.5373
y98b	-8.501184	18.89568	-0.45	0.653	-45.63248	28.63011
_cons	-6.616139	25.07553	-0.26	0.792	-55.89125	42.65897

```
. xtreg math4 lavgrexp lunch lenrol y95 y96 y97 y98, fe cluster(distid)
```

Fixed-effects (within) regression  
Group variable: schid

Number of obs = 7150  
Number of groups = 1683

R-sq: within = 0.3602  
between = 0.0292  
overall = 0.1514

Obs per group: min =  
avg = 4.  
max =

corr(u\_i, Xb) = 0.0073  
F(7,466) = 259.90  
Prob > F = 0.0000

(Std. Err. adjusted for 467 clusters in distid)

math4	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
lavgrexp	6.288376	3.132334	2.01	0.045	.1331271	12.44363
lunch	-.0215072	.0399206	-0.54	0.590	-.0999539	.0569395
lenrol	-2.038461	2.098607	-0.97	0.332	-6.162365	2.085443
y95	11.6192	.7210398	16.11	0.000	10.20231	13.0361
y96	13.05561	.9326851	14.00	0.000	11.22282	14.8884
y97	10.14771	.9576417	10.60	0.000	8.26588	12.02954
y98	23.41404	1.027313	22.79	0.000	21.3953	25.43278
_cons	11.84422	32.68429	0.36	0.717	-52.38262	76.07107
sigma_u	15.84958					
sigma_e	11.325028					

rho | .66200804 (fraction of variance due to u\_i)

c. The RE estimates are given below, and they are identical to the FE estimates. The RE coefficients on the time averages are not identical to those for POLS. In particular, on *lunchb*, the RE coefficient is  $-.415$ , just slightly smaller in magnitude than the POLS estimate. It has a slightly smaller fully robust *t* statistic (in absolute value).

```
. xtreg math4 lavgrexp lunch lenrol y95 y96 y97 y98 lavgrexpb lunchb lenrolb
    y95b y96b y97b y98b, re cluster(distid)
```

Random-effects GLS regression                      Number of obs        =        7150  
Group variable: schid                              Number of groups     =        1683

R-sq:    within    = 0.3602                      Obs per group: min =  
              between   = 0.4366    avg    =        4.  
              overall   = 0.4146    max   =

Random effects u\_i ~Gaussian                      Wald chi2(14)        =        2532.10  
corr(u\_i, X)        = 0 (assumed)                      Prob > chi2        =        0.0000

(Std. Err. adjusted for 467 clusters in distid)

math4	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval	
lavgrexp	6.288376	3.13387	2.01	0.045	.1461029	12.43065
lunch	-.0215072	.0399402	-0.54	0.590	-.0997886	.0567741
lenrol	-2.038461	2.099636	-0.97	0.332	-6.153671	2.07675
y95	11.6192	.7213934	16.11	0.000	10.2053	13.03311
y96	13.05561	.9331425	13.99	0.000	11.22668	14.88453
y97	10.14771	.9581113	10.59	0.000	8.269847	12.02557
y98	23.41404	1.027817	22.78	0.000	21.39956	25.42852
lavgrexpb	2.569862	3.99586	0.64	0.520	-5.261881	10.4016
lunchb	-.4153413	.0363218	-11.44	0.000	-.4865308	-.3441518
lenrolb	.3829623	2.157847	0.18	0.859	-3.84634	4.612264
y95b	18.96418	15.24131	1.24	0.213	-10.90824	48.83659
y96b	16.16473	6.628049	2.44	0.015	3.173993	29.15547
y97b	17.50964	15.42539	1.14	0.256	-12.72357	47.74285
y98b	-9.420143	18.25294	-0.52	0.606	-45.19524	26.35495
_cons	-5.159784	24.08649	-0.21	0.830	-52.36844	42.04887
sigma_u	10.702446					
sigma_e	11.325028					
rho	.47175866	(fraction of variance due to u_i)				

d. When we drop the time averages of the year dummies the RE estimates are slightly different from the FE estimates. That is because we must now recognize that, with an unbalanced panel, the time averages of the year dummies are no longer constant. With a

balanced panel, the time average are  $1/T$  in each case. Now, the average is either zero – if the unit does not appear in the appropriate year – or  $1/T_i$  where  $T_i$  is the total number of years for unit (school)  $i$ . For example, the `list` command below shows that the school with identifier number 557 has data for the years 1994, 1997, and 1998. Therefore,  $y95b$  and  $y96b$  are both zero, while  $y97b$  and  $y98b$  are both  $1/3$ . With an unbalanced panel, we should include the time averages of the year dummies. In effect, this is allowing certain forms of sample selection to be correlated with the unobserved school heterogeneity.

```
. xtreg math4 lavgrexp lunch lenrol y95 y96 y97 y98 lavgrexpb lunchb lenrolb,
    re cluster(distid)
```

Random-effects GLS regression                      Number of obs        =        7150  
Group variable: schid                              Number of groups     =        1683

R-sq:    within    = 0.3602                      Obs per group: min =  
          between   = 0.4291    avg =        4.  
          overall    = 0.4105    max =

Random effects u\_i ~Gaussian                      Wald chi2(10)        =    2073.48  
corr(u\_i, X)        = 0 (assumed)                      Prob > chi2        =    0.0000

(Std. Err. adjusted for 467 clusters in distid)

math4	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval	
lavgrexp	6.222429	3.121881	1.99	0.046	.1036546	12.3412
lunch	-.0209812	.0402425	-0.52	0.602	-.099855	.0578926
lenrol	-2.06064	2.070938	-1.00	0.320	-6.119604	1.998325
y95	11.78595	.7084874	16.64	0.000	10.39734	13.17456
y96	13.16626	.91466	14.39	0.000	11.37356	14.95896
y97	10.21612	.9441691	10.82	0.000	8.365579	12.06665
y98	23.46409	1.055457	22.23	0.000	21.39544	25.53275
lavgrexpb	2.417603	3.887099	0.62	0.534	-5.20097	10.03618
lunchb	-.4088571	.0365525	-11.19	0.000	-.4804986	-.3372155
lenrolb	.7979708	2.109349	0.38	0.705	-3.336278	4.93222
_cons	2.619295	24.78096	0.11	0.916	-45.95049	51.18908
sigma_u	10.702446					
sigma_e	11.325028					
rho	.47175866	(fraction of variance due to u_i)				

```
. list schid year y95b y96b y97b y98b if schid == 557
```

schid	year	y95b	y96b	y97b	y98b
557	1994	0	0	0	0
557	1997	0	0	1	1
557	1998	0	0	1	1

740.		557	1994	0	0	.33333333	.33333333	
741.		557	1997	0	0	.33333333	.33333333	
742.		557	1998	0	0	.33333333	.33333333	
+-----+-----+-----+-----+								

e. The FE estimates without the year dummies are given below. The coefficient on the spending variable is more than seven times larger than when the year dummies are included. The estimate without the year dummies is very misleading. During this period in Michigan, spending was increasing and, at the same time, the definition of a passing score was changed so that more students passed the exam. Thus, without controlling for time dummies, most of the relationship between pass rates and spending is spurious.

```
. xtreg math4 lavgrexp lunch lenrol, fe cluster(distid)
```

```
Fixed-effects (within) regression           Number of obs   =       7150
Group variable: schid                     Number of groups =       1683

R-sq:  within = 0.1632                     Obs per group:  min =
        between = 0.0001                             avg =       4.
        overall = 0.0233                             max =

corr(u_i, Xb) = -0.3272                     F(3,466)        =      136.54
                                           Prob > F         =      0.0000
```

(Std. Err. adjusted for 467 clusters in distid)

math4	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
lavgrexp	45.00103	2.452645	18.35	0.000	40.18141	49.82064
lunch	.0179948	.0377204	0.48	0.634	-.0561284	.092118
lenrol	-2.372125	3.403866	-0.70	0.486	-9.060952	4.316701
_cons	-294.8467	32.11083	-9.18	0.000	-357.9467	-231.7468
sigma_u	17.573721					
sigma_e	12.9465					
rho	.64820501	(fraction of variance due to u_i)				

f. The POLS and RE estimates, without the time averages, are given below. The spending effects are larger than FE and the effect of the *lunch* variable are much larger. If we do not remove the school effect – of which a large component is demographics that do not change over time – then the poverty measure *lunch* becomes very important. From the POLS/RE

estimates with the time averages included, it is really the average poverty level over several years that has the most predictive power. Of course, the *lunch* variable does not vary across time nearly as much as it does across school. Therefore, using FE, it is difficult to separate the effect of  $lunch_{it}$  from  $c_i$ .

```
. reg math4 lavgrexp lunch lenrol y95 y96 y97 y98, cluster(distid)
```

```
Linear regression                               Number of obs =      7150
                                                F(   7,   466) =   256.84
                                                Prob > F       =    0.0000
                                                R-squared      =    0.4029
                                                Root MSE      =   15.609
```

(Std. Err. adjusted for 467 clusters in distid)

math4	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
lavgrexp	8.628338	2.488897	3.47	0.001	3.737487	13.51919
lunch	-.4255479	.0391249	-10.88	0.000	-.5024309	-.3486648
lenrol	-1.294046	1.149539	-1.13	0.261	-3.552969	.9648762
y95	12.09916	.8909378	13.58	0.000	10.34841	13.84992
y96	13.06982	1.128072	11.59	0.000	10.85308	15.28655
y97	10.29535	1.114853	9.23	0.000	8.104584	12.48611
y98	23.57121	1.29055	18.26	0.000	21.03519	26.10723
_cons	2.758117	23.09242	0.12	0.905	-42.62005	48.13628

```
. xtreg math4 lavgrexp lunch lenrol y95 y96 y97 y98, re cluster(distid)
```

```
Random-effects GLS regression                Number of obs      =      7150
Group variable: schid                       Number of groups   =     1683
```

```
R-sq:  within = 0.3455                      Obs per group: min =
        between = 0.4288                      avg =      4.
        overall = 0.4016                      max =
```

```
Random effects u_i ~Gaussian                Wald chi2(7)       =   1886.18
corr(u_i, X) = 0 (assumed)                  Prob > chi2        =    0.0000
```

(Std. Err. adjusted for 467 clusters in distid)

math4	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval	
lavgrexp	7.838068	2.157833	3.63	0.000	3.608793	12.06734
lunch	-.3785643	.0400361	-9.46	0.000	-.4570336	-.3000949
lenrol	-1.391074	.9449022	-1.47	0.141	-3.243048	.4609008
y95	11.66598	.7704663	15.14	0.000	10.1559	13.17607
y96	12.88762	.9420724	13.68	0.000	11.04119	14.73404
y97	10.18776	.896855	11.36	0.000	8.429958	11.94557
y98	23.53236	1.029968	22.85	0.000	21.51366	25.55106



_cons	8.166742	20.06401	0.41	0.684	-31.158	47.49148
sigma_u	10.702446					
sigma_e	11.325028					
rho	.47175866	(fraction of variance due to u_i)				

g. It seems pretty clear we need to go with the FE estimate and its standard error robust to serial correlation within school and cluster correlation within district. Removing a school effect most likely gives us the least biased estimator of school spending. Clustering at the district level, rather than just at the school level, increases the standard error to 3.13 from about 2.43, and so it seems prudent to use the standard error clustered at the district level.

```
. xtreg math4 lavgrexp lunch lenrol y95 y96 y97 y98, fe cluster(schid)
```

```
Fixed-effects (within) regression      Number of obs   =      7150
Group variable: schid                 Number of groups =      1683

R-sq:  within = 0.3602                  Obs per group:  min =
        between = 0.0292                                avg   =      4.
        overall  = 0.1514                                max   =

                                         F(7,1682)        =      431.08
corr(u_i, Xb)  = 0.0073                 Prob > F         =      0.0000
```

(Std. Err. adjusted for 1683 clusters in schid)

math4	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
lavgrexp	6.288376	2.431317	2.59	0.010	1.519651	11.0571
lunch	-.0215072	.0390732	-0.55	0.582	-.0981445	.05513
lenrol	-2.038461	1.789094	-1.14	0.255	-5.547545	1.470623
y95	11.6192	.5358469	21.68	0.000	10.56821	12.6702
y96	13.05561	.6910815	18.89	0.000	11.70014	14.41108
y97	10.14771	.7326314	13.85	0.000	8.710745	11.58468
y98	23.41404	.7669553	30.53	0.000	21.90975	24.91833
_cons	11.84422	25.16643	0.47	0.638	-37.51659	61.20503
sigma_u	15.84958					
sigma_e	11.325028					
rho	.66200804	(fraction of variance due to u_i)				

**20.8.** a. The information contained in  $(\mathbf{x}_g, \mathbf{Z}_g, c_g)$  and  $(\mathbf{x}_g, \mathbf{Z}_g, a_g)$  is the same, and so if we substitute for  $c_g$  we have

$$\begin{aligned}
E(y_{gm}|\mathbf{x}_g, \mathbf{Z}_g, c_g) &= \Phi(\alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + c_g) \\
&= \Phi(\alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + \eta_g + \bar{\mathbf{z}}_g\boldsymbol{\xi}_g + a_g) \\
&= E(y_{gm}|\mathbf{x}_g, \mathbf{Z}_g, a_g).
\end{aligned}$$

b. Mechanically, we can get  $E(y_{gm}|\mathbf{x}_g, \mathbf{Z}_g, a_g) = E(y_{gm}|\mathbf{x}_g, \mathbf{Z}_g, a_g)$  from

$$\int_{-\infty}^{\infty} 1[\alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + \eta_g + \bar{\mathbf{z}}_g\boldsymbol{\xi}_g + a_g + u > 0] \phi(u) du$$

where  $\phi(\cdot)$  is the standard normal distribution. If we want  $E(y_{gm}|\mathbf{x}_g, \mathbf{Z}_g)$  then we integrate out  $a_g$  with respect to the  $\text{Normal}(0, \tau_g^2)$  distribution. Just as in the probit case this the same as computing

$$E(1[\alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + \eta_g + \bar{\mathbf{z}}_g\boldsymbol{\xi}_g + a_g + u_{gm} > 0]|\mathbf{x}_g, \mathbf{Z}_g)$$

where  $(a_g + u_{gm})$  is  $\text{Normal}(0, 1 + \tau_g^2)$  and independent of  $(\mathbf{x}_g, \mathbf{Z}_g)$ . Therefore,

$$E(y_{gm}|\mathbf{x}_g, \mathbf{Z}_g) = \Phi\left[\frac{(\alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + \eta_g + \bar{\mathbf{z}}_g\boldsymbol{\xi}_g)}{(1 + \tau_g^2)^{1/2}}\right].$$

Notice that  $\alpha$  can just be absorbed into  $\eta_g$ .

c. Under the asymptotic scheme where  $G \rightarrow \infty$  and the  $M_g$  are fixed, there is an upper bound, say  $M$ , with  $M_g \leq M$  for all  $g$ . If we see relatively few group sizes – and lots of data per group size – we can allow the parameters to be different for each  $M_g$ , with an appropriate normalization. For example, we can have

$$E(y_{gm}|\mathbf{x}_g, \mathbf{Z}_g) = \Phi\left[\frac{(\eta_{M_g} + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + \bar{\mathbf{z}}_g\boldsymbol{\xi}_{M_g})}{(1 + \tau_{M_g}^2)^{1/2}}\right].$$

where  $\tau_{M_g}^2$  is set to zero for one value, such as  $\tau_M^2 = 0$ . We can easily estimate all of the parameters using the quasi-log likelihood associated with a “heteroskedastic probit,” where we include in the heteroskedasticity function dummy variables for all but one outcome on  $M_g$ .

And, of course, we include an intercept and dummy variables in the index as well as  $\bar{z}_g$  and interactions with the group-size dummies.

d. If we use the Bernoulli QMLE with the mean function discussed in part c, we need to be sure that the inference is robust both to the true distribution not being Bernoulli and the within-cluster correlation.

## Solutions to Chapter 21 Problems

**21.1.** a. We use equation (21.5). First, because we have a random sample from the treatment and control groups,  $E(\bar{y}_1) = E(y|w = 1)$  and  $E(\bar{y}_0) = E(y|w = 0)$ . Therefore, by equation (21.5),

$$E(\bar{y}_1 - \bar{y}_0) = [E(y_0|w = 1) - E(y_0|w = 0)] + \tau_{att}.$$

It follows that the bias term for estimating  $\tau_{att}$  is given by the first term.

b. If  $E(y_0|w = 1) < E(y_0|w = 0)$ , those who participate in the program would have had lower average earnings without training than those who chose not to participate. This is a form of self-selection, and, on average, leads to an underestimate of the impact of the program.

**21.2.** Let  $k \equiv [w - p(\mathbf{x})]y/\{p(\mathbf{x})[1 - p(\mathbf{x})]\}$ . Then we know from equation (21.21) that

$$E(k|\mathbf{x}) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) = \tau_{ate}(\mathbf{x}).$$

Define a dummy variable as  $d \equiv 1[\mathbf{x} \in \mathcal{R}]$ . Then, by iterated expectations and the fact that  $d$  is a function of  $\mathbf{x}$ ,

$$\begin{aligned} E(y_1 - y_0|d) &= E[E(y_1 - y_0|\mathbf{x}, d)|d] = E[E(y_1 - y_0|\mathbf{x})|d] \\ &= E[\tau_{ate}(\mathbf{x})|d] = E[E(k|\mathbf{x})|d] = E(k|d) \end{aligned}$$

It follows that  $\tau_{ate, \mathcal{R}} \equiv \tau_{ate, \mathcal{R}} = E(y_1 - y_0|\mathbf{x} \in \mathcal{R}) = E(y_1 - y_0|d = 1) = E(k|d = 1)$ . Now use the simple relationship

$$E(d \cdot k) = P(d = 1)E(k|d = 1)$$

and so

$$E(k|d = 1) = \frac{E(d \cdot k)}{P(d = 1)} = \frac{E(d \cdot k)}{P(\mathbf{x} \in \mathcal{R})}.$$

If we know the propensity score, a consistent estimator of  $E(d \cdot k)$  would be

$$N^{-1} \sum_{i=1}^N 1[\mathbf{x}_i \in \mathcal{R}]k_i,$$

and a consistent estimator of  $P(\mathbf{x} \in \mathcal{R})$  is just the fraction of observations with  $\mathbf{x}_i \in \mathcal{R}$ , call this  $N_{\mathcal{R}}/N$ . Combining these two estimators and using the expression for  $k_i$  gives

$$\tilde{\tau}_{ate, \mathcal{R}} = N_{\mathcal{R}}^{-1} \sum_{i=1}^N 1[\mathbf{x}_i \in \mathcal{R}]k_i,$$

which is simply the average of  $k_i$  over the subset of observations with  $\mathbf{x}_i \in \mathcal{R}$ .

**21.3.** a. The simple regression estimate is  $\hat{\tau}_{ate} = .128$ , which means that those participating in the job training program are about .128 *more* likely of being unemployed after completing the program. Further, its heteroskedasticity-robust  $t$  statistic is about four. This appears to be a case of self-selection into training: those who would have a higher chance of being unemployed are also more likely to participate in job training.

```
. reg unem78 train, robust
```

Linear regression

```
Number of obs =    2675
F(   1,   2673) =    15.90
Prob > F       =    0.0001
R-squared      =    0.0098
Root MSE     =    .32779
```

unem78	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
train	.1283838	.0321964	3.99	0.000	.0652514	.1915162
_cons	.1148594	.0063922	17.97	0.000	.1023252	.1273936

b. Adding the controls listed in the problem changes the picture considerable. The estimate of  $\tau_{ate}$  is now  $-.199$ , so participating in the job training program is estimated to reduce the unemployment probability by about .20. The 95% confidence interval for  $\tau_{ate}$  is  $[-.288, -.111]$ , which clearly excludes zero.

```
. reg unem78 train age educ black hisp married re74 re75 unem75 unem74, robust
```

Linear regression

Number of obs = 2675  
F( 10, 2664) = 64.36  
Prob > F = 0.0000  
R-squared = 0.3141  
Root MSE = .27327

unem78	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
train	-.1993525	.045185	-4.41	0.000	-.2879538	-.1107512
age	.0028579	.0006397	4.47	0.000	.0016036	.0041123
educ	.0002969	.0020983	0.14	0.887	-.0038176	.0044114
black	-.0179975	.0122695	-1.47	0.143	-.0420563	.0060613
hisp	-.0625543	.0250947	-2.49	0.013	-.1117613	-.0133474
married	-.0136721	.0173229	-0.79	0.430	-.0476399	.0202957
re74	.0008451	.001004	0.84	0.400	-.0011236	.0028138
re75	-.0042097	.0010084	-4.17	0.000	-.006187	-.0022325
unem75	.2994134	.0395227	7.58	0.000	.2219151	.3769118
unem74	.2385391	.0419072	5.69	0.000	.1563652	.3207131
_cons	.0433446	.0358278	1.21	0.226	-.0269085	.1135978

c. After running the regressions for the untrained and trained groups separately, we obtain a fitted value (fitted probability) in each state for all 2,675 men in the sample. For each  $i$  we estimate the treatment effect conditional on  $\mathbf{x}$  as

$$\hat{\tau}(\mathbf{x}_i) = (\hat{\alpha}_1 + \mathbf{x}_i \hat{\boldsymbol{\beta}}_1) - (\hat{\alpha}_0 + \mathbf{x}_i \hat{\boldsymbol{\beta}}_0).$$

Then

$$\hat{\tau}_{ate} = N^{-1} \sum_{i=1}^N \hat{\tau}(\mathbf{x}_i)$$

$$\hat{\tau}_{att} = N_1^{-1} \sum_{i=1}^N \text{train}_i \cdot \hat{\tau}(\mathbf{x}_i)$$

We get  $\hat{\tau}_{ate} = -.203$ , which is very close to the estimate when we assume  $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0$ . The estimate of  $\tau_{att}$  is somewhat larger in magnitude:  $\hat{\tau}_{att} = -.270$ .

```
. reg unem78 age educ black hisp married re74 re75 unem75 unem74 if ~train
```

Source	SS	df	MS	Number of obs =
Model	100.075847	9	11.1195386	2490
				F( 9, 2480) = 180.15
				Prob > F = 0.0000

Residual		153.074354	2480	.06172353		R-squared	=	0.3953
-----								
Total		253.150201	2489	.101707594		Adj R-squared	=	0.3931
						Root MSE	=	.24844

unem78		Coef.	Std. Err.	t	P> t	[95% Conf. Interval
-----						
age		.0021732	.0005579	3.90	0.000	.0010791 .0032673
educ		-.0014064	.0019407	-0.72	0.469	-.005212 .0023992
black		-.0173876	.0125968	-1.38	0.168	-.0420889 .0073136
hisp		-.0517355	.0285084	-1.81	0.070	-.1076382 .0041672
married		-.0149914	.0151672	-0.99	0.323	-.0447332 .0147503
re74		.0014736	.0007966	1.85	0.064	-.0000884 .0030356
re75		-.0035097	.0007814	-4.49	0.000	-.0050419 -.0019774
unem75		.3435381	.0257242	13.35	0.000	.293095 .3939813
unem74		.3363692	.0275345	12.22	0.000	.2823763 .3903622
_cons		.0500675	.0349642	1.43	0.152	-.0184946 .1186296

```
. predict unem78_0
(option xb assumed; fitted values)
```

```
. reg unem78 age educ black hisp married re74 re75 unem75 unem74 if train
```

Source		SS	df	MS		Number of obs =	185
-----							
Model		2.71236085	9	.301373428		F( 9, 175) =	1.68
Residual		31.3416932	175	.17909539		Prob > F =	0.0962
-----							
Total		34.0540541	184	.185076381		R-squared =	0.0796
						Adj R-squared =	0.0323
						Root MSE =	.4232

unem78		Coef.	Std. Err.	t	P> t	[95% Conf. Interval
-----						
age		-.0022981	.0046702	-0.49	0.623	-.0115153 .0069192
educ		-.008484	.0158595	-0.53	0.593	-.0397845 .0228166
black		.1374346	.1067107	1.29	0.199	-.073171 .3480401
hisp		-.1412636	.1655747	-0.85	0.395	-.468044 .1855168
married		-.0761776	.0855254	-0.89	0.374	-.2449717 .0926165
re74		-.0019756	.0098056	-0.20	0.841	-.0213281 .017377
re75		-.010362	.014196	-0.73	0.466	-.0383794 .0176553
unem75		.1822138	.1020566	1.79	0.076	-.0192063 .3836338
unem74		-.233911	.1194775	-1.96	0.052	-.4697132 .0018912
_cons		.3735869	.2407415	1.55	0.123	-.1015435 .8487174

```
. predict unem78_1
(option xb assumed; fitted values)
```

```
. gen te = unem78_1 - unem78_0
```

```
. sum te
```

Variable		Obs	Mean	Std. Dev.	Min	Max
-----						
te		2675	-.2031515	.2448774	-1.5703	.3241221

```
. sum te if train
```

Variable	Obs	Mean	Std. Dev.	Min	Max
te	185	-.2698234	.309953	-.7017545	.3241221

d. Using the subsample of men who were unemployed in 1974, 1975, or both gives

$\hat{\tau}_{ate} = -.625$  and  $\hat{\tau}_{att} = -.194$ . The estimate of  $\tau_{ate}$  is much larger in magnitude than on the full sample and  $\hat{\tau}_{att}$  is somewhat smaller.

```
. keep if unem74 | unem75
(2240 observations deleted)
```

```
. reg unem78 age educ black hisp married re74 re75 if ~train
```

Source	SS	df	MS	Number of obs =	302
Model	17.2414134	7	2.46305906	F( 7, 294) =	12.93
Residual	56.020176	294	.190544816	Prob > F =	0.0000
				R-squared =	0.2353
				Adj R-squared =	0.2171
Total	73.2615894	301	.243393985	Root MSE =	.43651

unem78	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
age	.0121428	.0025998	4.67	0.000	.0070262	.0172594
educ	-.0000954	.0090746	-0.01	0.992	-.0179548	.017764
black	-.0713435	.0713164	-1.00	0.318	-.2116989	.0690119
hisp	-.1965901	.1220144	-1.61	0.108	-.4367224	.0435422
married	.0610631	.075997	0.80	0.422	-.088504	.2106302
re74	-.0094196	.0029031	-3.24	0.001	-.0151331	-.0037061
re75	-.0190763	.0029208	-6.53	0.000	-.0248247	-.013328
_cons	.1819096	.1810088	1.00	0.316	-.1743278	.5381469

```
. predict unem78_0
(option xb assumed; fitted values)
```

```
. reg unem78 age educ black hisp married re74 re75 if train
```

Source	SS	df	MS	Number of obs =	133
Model	2.33329022	7	.333327175	F( 7, 125) =	1.90
Residual	21.9674617	125	.175739693	Prob > F =	0.0754
				R-squared =	0.0960
				Adj R-squared =	0.0454
Total	24.3007519	132	.184096605	Root MSE =	.41921

unem78	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
age	-.0058054	.0049952	-1.16	0.247	-.0156914	.0040807
educ	-.0267626	.0175847	-1.52	0.131	-.0615649	.0080397
black	.1754782	.1201604	1.46	0.147	-.0623342	.4132906
hisp	-.1106474	.2078183	-0.53	0.595	-.5219455	.3006508
married	-.1606594	.1015391	-1.58	0.116	-.3616179	.040299
re74	-.0150277	.066169	-0.23	0.821	-.1459844	.1159289
re75	-.0269891	.0282243	-0.96	0.341	-.0828484	.0288702



_cons		.5632464	.253897	2.22	0.028	.0607527	1.06574
-------	--	----------	---------	------	-------	----------	---------

```
. predict unem78_1
(option xb assumed; fitted values)
```

```
. gen te = unem78_1 - unem78_0
```

```
. sum te
```

Variable	Obs	Mean	Std. Dev.	Min	Max
te	435	-.625014	.3867973	-1.62662	.1450891

```
. sum te if train
```

Variable	Obs	Mean	Std. Dev.	Min	Max
te	133	-.1935882	.2039181	-.7526801	.1450891

e. We use the entire set of data for this part. The logit model for *train* is estimated below.

Of the 2,675 observations, 78 failures are completely determined. This means that the overlap assumption fails because for some values of *x* the probability of being in the training group is zero. If we are interested in the ATE then our only recourse is to redefine the population so that each unit has a nonzero chance of being in the treated group (and a nonzero chance of being in the control group, which is not a problem in this example).

```
. logit train age educ black hisp married re74 re75 unem74 unem75
```

Logistic regression	Number of obs	=	2675
	LR chi2(9)	=	926.52
	Prob > chi2	=	0.0000
Log likelihood = -209.38931	Pseudo R2	=	0.6887

train	Coef.	Std. Err.	z	P> z	[95% Conf. Interval
age	-.1109206	.0177106	-6.26	0.000	-.1456327 - .0762084
educ	-.1008807	.0561133	-1.80	0.072	-.2108608 .0090994
black	2.650097	.3605668	7.35	0.000	1.943399 3.356795
hisp	2.247747	.5908963	3.80	0.000	1.089611 3.405882
married	-1.560628	.2817913	-5.54	0.000	-2.112928 -1.008327
re74	.0201797	.0313149	0.64	0.519	-.0411963 .0815557
re75	-.2743162	.0477066	-5.75	0.000	-.3678194 -.1808129
unem74	3.272456	.4887585	6.70	0.000	2.314507 4.230405
unem75	-1.371405	.4545789	-3.02	0.003	-2.262363 -.4804465
_cons	1.794543	.979261	1.83	0.067	-.1247735 3.713859

Note: 78 failures and 0 successes completely determined.

f. The State session is below. The IPW estimate is  $\hat{\tau}_{ate,psw} = -.132$ . The standard error that adjusts for the first-step estimation is about .0504. If we do not take advantage of the smaller asymptotic variance due to estimating the propensity score, the standard error is .0580, which is about 15% larger. The estimate of  $\tau_{att}$  is similar, about  $-.124$ .

If we assume a constant treatment effect in using regression adjustment,  $\hat{\tau}_{ate,reg} = \hat{\tau}_{att,reg} = -.235$  and its standard error is .0509. Interestingly, this is very close to the standard error for  $\hat{\tau}_{ate,psw}$ , but the estimate is much larger in magnitude, leading to a large  $t$  statistic. Unfortunately, it appears separate regression are warranted, and this changes  $\hat{\tau}_{ate,reg}$  to  $-.119$  (although  $\hat{\tau}_{att,reg} = -.294$ ). The standard error for  $\hat{\tau}_{ate,reg}$  that does not even account for the randomness in the sample averages is quite large, .0911, and so  $\hat{\tau}_{ate,reg}$  is barely statistically different from zero at the 10% level if we use a one-sided alternative. The IPW estimator appears to be more efficient for this application. (It could have something to do with using linear regression adjustment rather than, say, probit or logit.) The joint test of the interaction terms shows separate regressions are warranted.

```
. keep if avgre <= 15
(1513 observations deleted)
```

```
. logit train age educ black hisp married re74 re75 unem74 unem75
```

Logistic regression	Number of obs	=	1162
	LR chi2(9)	=	641.37
	Prob > chi2	=	0.0000
Log likelihood = -180.28028	Pseudo R2	=	0.6401

train	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
age	-.1155512	.0187215	-6.17	0.000	-.1522447	-.0788577
educ	-.1049275	.0591078	-1.78	0.076	-.2207766	.0109217
black	2.608068	.3772016	6.91	0.000	1.868767	3.34737
hisp	2.395905	.6292337	3.81	0.000	1.162629	3.62918
married	-1.631159	.3038189	-5.37	0.000	-2.226633	-1.035685
re74	-.0290672	.04281	-0.68	0.497	-.1129732	.0548387
re75	-.3794923	.0682029	-5.56	0.000	-.5131676	-.245817
unem74	3.009282	.5221746	5.76	0.000	1.985839	4.032726

unem75		-1.751808	.4995608	-3.51	0.000	-2.730929	-.7726867
_cons		2.695208	1.053604	2.56	0.011	.6301819	4.760234

---

```
. predict phat
(option pr assumed; Pr(train))
```

```
. tab train
```

=1 if in job training	Freq.	Percent	Cum.
0	982	84.51	84.51
1	180	15.49	100.00
Total	1,162	100.00	

```
. sum train
```

Variable	Obs	Mean	Std. Dev.	Min	Max
train	1162	.1549053	.3619702	0	1

```
. gen rhohat = r(mean)
```

```
. gen kate = ((train - phat)*unem78)/(phat*(1 - phat))
```

```
. gen katt = ((train - phat)*unem78)/(rhohat*(1 - phat))
```

```
. sum kate katt
```

Variable	Obs	Mean	Std. Dev.	Min	Max
kate	1162	-.1319506	1.977683	-16.62496	56.51032
katt	1162	-.1243131	4.922131	-100.8678	6.455555

```
. * Get the correct standard error for the ATE estimate.
```

```
. gen uh = train - phat
```

```
. gen ageuh = age*uh
```

```
. gen educuh = educ*uh
```

```
. gen blackuh = black*uh
```

```
. gen hispuh = hisp*uh
```

```
. gen marrieduh = married*uh
```

```
. gen re74uh = re74*uh
```

```
. gen re75uh = re75*uh
```

```
. gen unem74uh = unem74*uh
```

```
. gen unem75uh = unem75*uh
```

```
. reg kate uh ageuh educuh blackuh hispuh marrieduh re74uh re75uh unem74uh
unem75uh
```

Source	SS	df	MS	Number of obs =	1162
Model	1138.33705	10	113.833705	F( 10, 1151) =	38.51
Residual	3402.59957	1151	2.95621161	Prob > F =	0.0000
				R-squared =	0.2507
				Adj R-squared =	0.2442
Total	4540.93661	1161	3.91122878	Root MSE =	1.7194

kate	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
uh	3.525428	1.973887	1.79	0.074	-.3473914	7.398247
ageuh	.0016821	.0350983	0.05	0.962	-.0671816	.0705458
educuh	.2945194	.1191697	2.47	0.014	.0607052	.5283336
blackuh	-3.176048	.6611273	-4.80	0.000	-4.473198	-1.878898
hispuh	-5.475508	1.012662	-5.41	0.000	-7.462378	-3.488638
marrieduh	4.005544	.5475872	7.31	0.000	2.931163	5.079926
re74uh	.3468368	.075946	4.57	0.000	.1978287	.495845
re75uh	-.8364872	.1060216	-7.89	0.000	-1.044504	-.62847
unem74uh	-2.607257	.818097	-3.19	0.001	-4.212386	-1.002129
unem75uh	.2278527	.796608	0.29	0.775	-1.335114	1.790819
_cons	-.1319506	.0504388	-2.62	0.009	-.2309129	-.0329883

```
. di e(rmse)/sqrt(e(N))
.05043879
```

```
. di -.1320/.0504
-2.6190476
```

```
. reg kate
```

Source	SS	df	MS	Number of obs =	1162
Model	0	0	.	F( 0, 1161) =	0.00
Residual	4540.93661	1161	3.91122878	Prob > F =	
				R-squared =	0.0000
				Adj R-squared =	0.0000
Total	4540.93661	1161	3.91122878	Root MSE =	1.9777

kate	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
_cons	-.1319506	.0580168	-2.27	0.023	-.24578	-.0181211

```
. reg unem78 train, robust
```

Linear regression

```
Number of obs = 1162
F( 1, 1160) = 0.18
Prob > F = 0.6734
R-squared = 0.0002
Root MSE = .4259
```

unem78	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	

train		.0147658	.0350282	0.42	0.673	-.0539599	.0834915
_cons		.2352342	.0135467	17.36	0.000	.2086555	.2618129

```
. reg unem78 train age educ black hisp married re74 re75 unem74 unem75, robust
```

Linear regression	Number of obs =	1162
	F( 10, 1151) =	61.27
	Prob > F =	0.0000
	R-squared =	0.3312
	Root MSE =	.34968

unem78	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval
train	-.2349689	.0509218	-4.61	0.000	-.3348787 -.135059
age	.0059358	.0012367	4.80	0.000	.0035094 .0083622
educ	.0022623	.0042076	0.54	0.591	-.005993 .0105177
black	-.0202408	.022745	-0.89	0.374	-.0648671 .0243855
hisp	-.100478	.0399462	-2.52	0.012	-.1788536 -.0221024
married	-.0352163	.0272463	-1.29	0.196	-.0886743 .0182417
re74	-.0010355	.002876	-0.36	0.719	-.0066783 .0046073
re75	-.0177354	.0024155	-7.34	0.000	-.0224746 -.0129961
unem74	.2220472	.051956	4.27	0.000	.1201081 .3239863
unem75	.1439644	.048573	2.96	0.003	.0486629 .2392658
_cons	.1103197	.0759773	1.45	0.147	-.0387499 .2593893

```
. reg unem78 age educ black hisp married re74 re75 unem74 unem75 if ~train
```

Source	SS	df	MS	Number of obs =	982
Model	78.510332	9	8.72337022	F( 9, 972) =	86.39
Residual	98.1505642	972	.100977947	Prob > F =	0.0000
				R-squared =	0.4444
				Adj R-squared =	0.4393
Total	176.660896	981	.180082463	Root MSE =	.31777

unem78	Coef.	Std. Err.	t	P> t	[95% Conf. Interval
age	.0050777	.0011449	4.43	0.000	.0028309 .0073245
educ	-.0002579	.0039421	-0.07	0.948	-.0079938 .0074781
black	-.0146538	.0238818	-0.61	0.540	-.0615196 .0322119
hisp	-.0862098	.0524883	-1.64	0.101	-.1892132 .0167936
married	-.0424904	.0262258	-1.62	0.106	-.093956 .0089752
re74	.0022784	.0024006	0.95	0.343	-.0024325 .0069892
re75	-.0143134	.0025479	-5.62	0.000	-.0193134 -.0093133
unem74	.3521536	.0435278	8.09	0.000	.2667344 .4375729
unem75	.1965244	.0423339	4.64	0.000	.1134481 .2796007
_cons	.0770668	.0757616	1.02	0.309	-.0716084 .2257419

```
. predict unem78_0
(option xb assumed; fitted values)
```

```
. reg unem78 age educ black hisp married re74 re75 unem74 unem75 if train
```

Source	SS	df	MS	Number of obs =	180
--------	----	----	----	-----------------	-----

Model	2.58861704	9	.287624115	F( 9, 170) = 1.57
Residual	31.161383	170	.183302253	Prob > F = 0.1281
				R-squared = 0.0767
				Adj R-squared = 0.0278
Total	33.75	179	.188547486	Root MSE = .42814

unem78	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	-.002544	.0047571	-0.53	0.593	-.0119347	.0068466
educ	-.0086994	.0162153	-0.54	0.592	-.0407086	.0233098
black	.1402344	.108018	1.30	0.196	-.072995	.3534638
hisp	-.1480334	.1683835	-0.88	0.381	-.4804252	.1843585
married	-.0713415	.0879005	-0.81	0.418	-.2448585	.1021756
re74	.0073134	.0145599	0.50	0.616	-.021428	.0360549
re75	-.0064075	.0214837	-0.30	0.766	-.0488166	.0360016
unem74	-.1885821	.1321929	-1.43	0.156	-.4495331	.0723688
unem75	.1935779	.1115475	1.74	0.084	-.0266186	.4137745
_cons	.3229791	.2550769	1.27	0.207	-.1805469	.8265051

```
. predict unem78_1
(option xb assumed; fitted values)
```

```
. gen te = unem78_1 - unem78_0
```

```
. sum te
```

Variable	Obs	Mean	Std. Dev.	Min	Max
te	1162	-.1193285	.3326819	-.9173806	.3599507

```
. sum te if train
```

Variable	Obs	Mean	Std. Dev.	Min	Max
te	180	-.2941826	.2835388	-.728443	.2494144

```
. egen mage = mean(age)
. gen trainage = train*(age - mage)
. egen meduc = mean(educ)
. gen traineduc = train*(educ - meduc)
. egen mblack = mean(black)
. gen trainblack = train*(black - mblack)
. egen mhisp = mean(hisp)
. gen trainhisp = train*(hisp - mhisp)
. egen mmarried = mean(married)
. gen trainmarried = train*(married - mmarried)
. egen mre74 = mean(re74)
. gen trainre74 = train*(re74 - mre74)
. egen mre75 = mean(re75)
. gen trainre75 = train*(re75 - mre75)
. egen munem74 = mean(unem74)
. gen trainunem74 = train*(unem74 - munem74)
. egen munem75 = mean(unem75)
. gen trainunem75 = train*(unem75 - munem75)

. reg unem78 train age educ black hisp married re74 re75 unem74 unem75
    trainage traineduc trainblack trainhisp trainmarried trainre74
```

```
trainre75 trainunem74 trainunem75, robust
```

Linear regression

```
Number of obs =    1162
F( 19, 1142) =    42.62
Prob > F      =    0.0000
R-squared     =    0.3855
Root MSE     =    .3365
```

unem78	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
train	-.1193284	.0910893	-1.31	0.190	-.2980497	.0593928
age	.0050777	.0012214	4.16	0.000	.0026812	.0074742
educ	-.0002579	.0041835	-0.06	0.951	-.008466	.0079503
black	-.0146538	.0231254	-0.63	0.526	-.0600269	.0307192
hisp	-.0862098	.0424936	-2.03	0.043	-.169584	-.0028356
married	-.0424904	.0277233	-1.53	0.126	-.0968847	.0119039
re74	.0022784	.0028885	0.79	0.430	-.003389	.0079457
re75	-.0143134	.0025811	-5.55	0.000	-.0193776	-.0092491
unem74	.3521536	.0566374	6.22	0.000	.2410286	.4632787
unem75	.1965244	.0570442	3.45	0.001	.0846013	.3084475
trainage	-.0076217	.0050195	-1.52	0.129	-.0174702	.0022267
traineduc	-.0084415	.0152788	-0.55	0.581	-.0384192	.0215361
trainblack	.1548883	.0871611	1.78	0.076	-.0161256	.3259022
trainhisp	-.0618236	.0975772	-0.63	0.526	-.2532742	.1296271
trainmarried	-.0288511	.0822415	-0.35	0.726	-.1902126	.1325104
trainre74	.0050351	.0166047	0.30	0.762	-.027544	.0376142
trainre75	.0079059	.0185161	0.43	0.669	-.0284236	.0442353
trainunem74	-.5407358	.1357835	-3.98	0.000	-.807149	-.2743226
trainunem75	-.0029465	.0975097	-0.03	0.976	-.1942647	.1883717
_cons	.0770668	.0760186	1.01	0.311	-.0720851	.2262186

```
. test  trainage traineduc trainblack trainhisp trainmarried trainre74
        trainre75 trainunem74 trainunem75
```

- ( 1) trainage = 0
- ( 2) traineduc = 0
- ( 3) trainblack = 0
- ( 4) trainhisp = 0
- ( 5) trainmarried = 0
- ( 6) trainre74 = 0
- ( 7) trainre75 = 0
- ( 8) trainunem74 = 0
- ( 9) trainunem75 = 0

```
F( 9, 1142) =    8.61
Prob > F =    0.0000
```

21.4. The integral is equivalent to

$$\int_{-(\theta_0 + \mathbf{x}\theta_1 + \mathbf{z}\theta_2)}^{\infty} a\phi(a)da.$$

Because  $d\phi(a)/da = -a\phi(a)$ , the antiderivative of  $a\phi(a)$  is simply  $-\phi(a)$ . Now

$$\begin{aligned}\int_{-(\theta_0 + \mathbf{x}\theta_1 + \mathbf{z}\theta_2)}^m a\phi(a)da &= -\phi(a)]_{-(\theta_0 + \mathbf{x}\theta_1 + \mathbf{z}\theta_2)}^m = -\phi(m) + \phi[-(\theta_0 + \mathbf{x}\theta_1 + \mathbf{z}\theta_2)] \\ &= -\phi(m) + \phi(\theta_0 + \mathbf{x}\theta_1 + \mathbf{z}\theta_2)\end{aligned}$$

where we use the symmetry of  $\phi(\cdot)$ . As  $m \rightarrow \infty$ ,  $\phi(m) \rightarrow 0$ . Therefore,

$$\int_{-(\theta_0 + \mathbf{x}\theta_1 + \mathbf{z}\theta_2)}^{\infty} a\phi(a)da = \phi(\theta_0 + \mathbf{x}\theta_1 + \mathbf{z}\theta_2).$$

**21.5.** The Stata output to answer all parts follows.

a. The first two Stata commands are used to obtain the probit fitted values, called *PHIhat*.

b. The IV estimate of  $\tau$  is  $-43.27$  and its standard error is huge,  $585.78$ . Clearly we can learn nothing of value from this estimate.

c. The collinearity suspected in part b is confirmed by regressing  $\hat{\Phi}_i$  on the  $\mathbf{x}_i$ : the R-squared is .9989, which means there is almost no separate variation in  $\hat{\Phi}_i$  that cannot be explained by  $\mathbf{x}_i$ .

d. This example illustrates why trying to achieve identification off of a nonlinearity can be fraught with problems. In cases with larger sample sizes the estimates may seem more reasonable, but we are only able to compute estimates at all because of the presumed functional form for  $P(w|\mathbf{x})$ . A good general rule is that if a linear IV approach does not identify  $\tau$  then we should not hope to learn anything useful by introducing nonlinearity in  $P(w|\mathbf{x})$ .

```
. probit train age educ black hisp married re74 re75
```

```
Probit regression               Number of obs   =          445
                                LR chi2(7)         =           8.60
                                Prob > chi2         =          0.2829
Log likelihood = -297.80166      Pseudo R2       =          0.0142
```

train	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.0066826	.0087391	0.76	0.444	-.0104458	.0238109
educ	.0387341	.0341574	1.13	0.257	-.0282132	.1056815



black	-.2216642	.2242952	-0.99	0.323	-.6612747	.2179463
hisp	-.5753033	.3062908	-1.88	0.060	-1.175622	.0250157
married	.0900855	.1703412	0.53	0.597	-.2437771	.4239482
re74	-.0138226	.0155792	-0.89	0.375	-.0443572	.016712
re75	.028755	.0267469	1.08	0.282	-.0236679	.0811779
_cons	-.5715372	.475416	-1.20	0.229	-1.503335	.3602609

```
. predict PHIhat
(option pr assumed; Pr(train))
```

```
. ivreg re78 age educ black hisp married re74 re75 (train = PHIhat)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs =	445
Model	-213187.422	8	-26648.4277	F( 8, 436) =	0.18
Residual	232713.078	436	533.745593	Prob > F =	0.9934
				R-squared =	
				Adj R-squared =	
Total	19525.6566	444	43.9767041	Root MSE =	23.103

re78	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
train	-43.26513	585.7793	-0.07	0.941	-1194.567	1108.037
age	.1717735	1.520127	0.11	0.910	-2.815914	3.159461
educ	1.067645	8.646843	0.12	0.902	-15.92703	18.06232
black	-6.114187	51.56931	-0.12	0.906	-107.4695	95.24116
hisp	-9.523185	126.287	-0.08	0.940	-257.7302	238.6838
married	1.432202	20.72909	0.07	0.945	-39.30917	42.17357
re74	-.1443703	2.973787	-0.05	0.961	-5.98911	5.70037
re75	.5327602	6.2896	0.08	0.933	-11.82894	12.89447
_cons	13.30468	165.517	0.08	0.936	-312.0058	338.6151

Instrumented: train

Instruments: age educ black hisp married re74 re75 PHIhat

```
. reg PHIhat age educ black hisp married re74 re75
```

Source	SS	df	MS	Number of obs =	445
Model	2.04859095	7	.292655851	F( 7, 437) =	55245.15
Residual	.002314965	437	5.2974e-06	Prob > F =	0.0000
				R-squared =	0.9989
				Adj R-squared =	0.9989
Total	2.05090592	444	.004619157	Root MSE =	.0023

PHIhat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
age	.0025883	.0000158	163.53	0.000	.0025572	.0026194
educ	.0146708	.000062	236.45	0.000	.0145488	.0147927
black	-.0875955	.0004094	-213.96	0.000	-.0884002	-.0867909
hisp	-.2156441	.0005445	-396.02	0.000	-.2167143	-.2145739
married	.0351309	.0003107	113.08	0.000	.0345203	.0357415
re74	-.0051274	.0000271	-189.22	0.000	-.0051807	-.0050742
re75	.0108521	.0000474	228.89	0.000	.0107589	.0109453
_cons	.2823687	.0008635	326.99	0.000	.2806715	.2840659

-----

**21.6.** As in Procedure 21.1, the IV estimator is consistent whether or not  $G(\mathbf{x}, \mathbf{z}; \boldsymbol{\gamma})$  is correctly specified for  $P(w = 1|\mathbf{x}, \mathbf{z})$ . The OLS estimator from  $y_i$  on  $1, \hat{G}_i, \mathbf{x}_i, \hat{G}_i \cdot (\mathbf{x}_i - \bar{\mathbf{x}})$ ,  $i = 1, \dots, N$  generally requires the model for  $P(w = 1|\mathbf{x}, \mathbf{z})$  to be correctly specified. This can be seen by writing

$$E(y|\mathbf{x}, \mathbf{z}) = \gamma + \tau E(w|\mathbf{x}, \mathbf{z}) + \mathbf{x}\boldsymbol{\beta}_0 + E(w|\mathbf{x}, \mathbf{z}) \cdot (\mathbf{x} - \boldsymbol{\psi})\boldsymbol{\delta},$$

which is the estimating equation underlying the OLS regression on probit fitted values and the interactions. If  $E(w|\mathbf{x}, \mathbf{z}) = P(w = 1|\mathbf{x}, \mathbf{z}) \neq G(\mathbf{x}, \mathbf{z}; \boldsymbol{\gamma})$  for all  $\boldsymbol{\gamma}$  then plugging in  $\hat{G}_i$  generally produces inconsistent estimators.

Even if  $G(\mathbf{x}, \mathbf{z}; \boldsymbol{\gamma})$  is correctly specified, the standard errors for the two-step OLS estimator are harder to obtain. One must use the material on generated regressors in Chapter 6 or apply the bootstrap.

**21.7. a.** There are several options. To estimate the mean parameters, we can use Poisson regression (especially in the case where  $w$  is a count variable) or gamma regression (if  $w$  is nonnegative and continuous). Of course, we can use NLS, too (which is also a QMLE in the LEF).

As stated in the hint, if we define  $r = w - E(w|\mathbf{x})$  then  $E(r^2|\mathbf{x}) = \text{Var}(w|\mathbf{x}) = \exp(\delta_0 + \mathbf{x}\boldsymbol{\delta}_1)$ . Therefore, if we observed,  $r^2$ , we could use it as the dependent variable in, say, a gamma or negative binomial QMLE. In practice, we use  $\hat{r}_i = w_i - \exp(\hat{\gamma}_0 + \mathbf{x}_i\hat{\boldsymbol{\gamma}}_1)$ , the residuals from estimating the mean parameters.

b. By the law of large numbers,

$$N^{-1} \sum_{i=1}^N \left\{ \frac{[w_i - \psi(\mathbf{x}_i)]y_i}{\omega(\mathbf{x}_i)} \right\} \xrightarrow{p} \boldsymbol{\beta}.$$

By the usual argument, we can replace  $\psi(\mathbf{x}_i)$  and  $\omega(\mathbf{x}_i)$  with consistent estimators; more precisely, in a parametric context replace the unknown parameters with consistent estimators.

In the case of exponential mean and variance functions,

$$\hat{\beta} = N^{-1} \sum_{i=1}^N \left\{ \frac{[w_i - \exp(\hat{\gamma}_0 + \mathbf{x}_i \hat{\gamma}_1)] y_i}{\exp(\hat{\delta}_0 + \mathbf{x}_i \hat{\delta}_1)} \right\} \equiv N^{-1} \sum_{i=1}^N \left\{ \frac{[w_i - \hat{\psi}(\mathbf{x}_i)] y_i}{\hat{\omega}(\mathbf{x}_i)} \right\}.$$

We can use Problem 12.17 to get a standard error for  $\hat{\beta}$  or use the bootstrap.

c. I use Poisson regression to estimate the mean parameters and then gamm regression to estimate the variance parameters. The resulting estimate of  $\beta$  is about .102; the standard error is not reported. If we ignore estimation of the parameters in  $E(w|\mathbf{x})$  and  $\text{Var}(w|\mathbf{x})$  then the standard error is about .050.

There is not much reason to compute a standard error for  $\hat{\beta}$  because the standard regression adjustment estimate,  $\tilde{\beta}$ , is very close, and provides a valid standard error. Namely, running the regression

*re78<sub>i</sub> on 1, mostrn<sub>i</sub>, age<sub>i</sub>, educ<sub>i</sub>, black<sub>i</sub>, hisp<sub>i</sub>, married<sub>i</sub>, re74<sub>i</sub>, re75*

gives  $\tilde{\beta} = .103$  (*se* = .038). With random assignment to the job training program it is perhaps not too surprising to see the methods give similar estimates. In fact, the simple regression estimate is .112 (*se* = .038).

```
. glm mostrn age educ black hisp married re74 re75, fam(poisson) link(log)
  robust
```

Generalized linear models	No. of obs	=	445
Optimization : ML	Residual df	=	437
	Scale parameter =		
Deviance = 6136.777616	(1/df) Deviance =	14.04297	
Pearson = 5296.290311	(1/df) Pearson =	12.11966	
Variance function: V(u) = u	[Poisson]		
Link function : g(u) = ln(u)	[Log]		
	AIC	=	15.78862
Log pseudolikelihood = -3504.968642	BIC	=	3471.919

mostrn	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval	
age	.0037486	.0081575	0.46	0.646	-.0122398	.0197369
educ	.0448349	.0344975	1.30	0.194	-.0227789	.1124487
black	-.1809126	.1906097	-0.95	0.343	-.5545006	.1926755
hisp	-.4907343	.3198765	-1.53	0.125	-1.117681	.1362121
married	.0824876	.1620227	0.51	0.611	-.2350709	.4000462
re74	-.0048997	.0140984	-0.35	0.728	-.0325322	.0227327
re75	.0388417	.0197161	1.97	0.049	.0001989	.0774846
_cons	1.605453	.4551975	3.53	0.000	.7132821	2.497624

```
. predict mostnrh
(option mu assumed; predicted mean mostnrh)
```

```
. gen rh = mostnrh - mostnrh
```

```
. gen rhsq = rh^2
```

```
. glm rhsq age educ black hisp married re74 re75, fam(gamma) link(log) robust
```

Generalized linear models		No. of obs	=	445
Optimization	: ML	Residual df	=	437
		Scale parameter	=	.6888918
Deviance	= 251.4433046	(1/df) Deviance	=	.5753851
Pearson	= 301.0457257	(1/df) Pearson	=	.6888918

Variance function:	V(u) = u^2	[Gamma]
Link function	: g(u) = ln(u)	[Log]

	AIC	=	11.01458
Log pseudolikelihood = -2442.743803	BIC	=	-2413.415

rhsq	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval	
age	-.0020813	.0054739	-0.38	0.704	-.01281	.0086475
educ	.0206816	.0254745	0.81	0.417	-.0292474	.0706107
black	-.0424931	.1083251	-0.39	0.695	-.2548063	.1698202
hisp	-.2107907	.2173269	-0.97	0.332	-.6367437	.2151623
married	.0391702	.0937328	0.42	0.676	-.1445427	.2228831
re74	.0094508	.0107463	0.88	0.379	-.0116116	.0305132
re75	.051212	.0190328	2.69	0.007	.0139084	.0885156
_cons	4.288161	.3201553	13.39	0.000	3.660668	4.915654

```
. predict omegah
(option mu assumed; predicted mean rhsq)
```

```
. sum omegah
```

Variable	Obs	Mean	Std. Dev.	Min	Max
omegah	445	91.62743	28.73968	60.9556	369.0591

```
. gen kh = ( mostrn - mostrnh)*re78/omegah
```

```
. sum kh
```

Variable	Obs	Mean	Std. Dev.	Min	Max
kh	445	.1024405	1.047671	-3.030556	10.28323

```
. reg kh
```

Source	SS	df	MS	Number of obs =	445
Model	0	0	.	F( 0, 444) =	0.00
Residual	487.34086	444	1.09761455	Prob > F =	
Total	487.34086	444	1.09761455	R-squared =	0.0000
				Adj R-squared =	0.0000
				Root MSE =	1.0477

kh	Coef.	Std. Err.	t	P> t	[95% Conf. Interval
_cons	.1024405	.0496644	2.06	0.040	.0048341 .200047

```
. reg re78 mostrn age educ black hisp married re74 re75, robust
```

Linear regression

Number of obs =	445
F( 8, 436) =	3.09
Prob > F =	0.0021
R-squared =	0.0613
Root MSE =	6.4838

re78	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval
mostrn	.102825	.0380686	2.70	0.007	.0280043 .1776458
age	.0570883	.0399249	1.43	0.153	-.0213808 .1355575
educ	.3980183	.1548109	2.57	0.010	.09375 .7022867
black	-2.150926	1.007271	-2.14	0.033	-4.130637 -.1712163
hisp	.1712523	1.365153	0.13	0.900	-2.511846 2.85435
married	-.154993	.8733899	-0.18	0.859	-1.871571 1.561585
re74	.0788359	.1071444	0.74	0.462	-.1317478 .2894197
re75	.0305561	.1266573	0.24	0.809	-.2183787 .2794909
_cons	.6004532	2.366495	0.25	0.800	-4.050703 5.25161

```
. reg re78 mostrn, robust
```

Linear regression

Number of obs =	445
F( 1, 443) =	8.66
Prob > F =	0.0034
R-squared =	0.0269
Root MSE =	6.5491

re78	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval
mostrn	.1126397	.0382802	2.94	0.003	.0374063 .1878731

_cons		4.434831	.3358041	13.21	0.000	3.774864	5.094798
-------	--	----------	----------	-------	-------	----------	----------

---

d. Because  $E(w|\mathbf{x})$  follows a logistic regression model we can use fractional logit (that is, maximize the Bernoulli QMLE). After we have estimated the mean parameters  $\gamma_0$  and  $\gamma_1$ , we form the fitted values and residuals

$$\hat{w}_i = \Lambda(\hat{\gamma}_0 + \mathbf{x}_i \hat{\gamma}_1)$$

$$\hat{r}_i = w_i - \hat{w}_i$$

and then estimate  $\delta_0$ ,  $\delta_1$ , and  $\delta_2$  from the OLS regression

$$\hat{r}_i^2 \text{ on } 1, \hat{w}_i, \hat{w}_i^2$$

to get the variance estimates

$$\hat{\omega}_i = \hat{\delta}_0 + \hat{\delta}_1 \hat{w}_i + \hat{\delta}_2 \hat{w}_i^2.$$

Because the  $\hat{\omega}_i$  are fitted values from a linear regression, nothing guarantees  $\hat{\omega}_i > 0$  for all  $i$ , something we need for the method in part b to make sense. To avoid this problem, we might use  $\text{Var}(w|\mathbf{x}) = \exp\{\delta_0 + \delta_1 E(w|\mathbf{x}) + \delta_2 [E(w|\mathbf{x})]^2\}$  instead, and use the gamma QMLE with the squared residuals as the dependent variable.

e. The Stata code carries out the procedure from part d, except that, because 13 estimated variances were not positive, the exponential variance function was used instead, with a gamma QMLE. below produces the estimate  $\hat{\beta} = .689$ . The regression coefficient is not too different:

$$\tilde{\beta} = .644 \text{ (se} = .235\text{)}$$

```
. use attend
. gen ACTsq = ACT^2
. gen ACTcu = ACT^3
. gen priGPAsq = priGPA^2
. gen priGPACu = priGPA^3
. sum atndrte
```

Variable	Obs	Mean	Std. Dev.	Min	Max
atndrte	680	81.70956	17.04699	6.25	100

```
. replace atndrte = atndrte/100
(680 real changes made)
```

```
. glm atndrte priGPA priGPASq priGPACu ACT ACTsq ACTcu frosh soph, fam(bin)
    link(logit) robust
note: atndrte has noninteger values
```

Generalized linear models	No. of obs	=	680
Optimization : ML	Residual df	=	671
	Scale parameter	=	
Deviance	=	87.0709545	(1/df) Deviance = .129763
Pearson	=	85.07495268	(1/df) Pearson = .1267883

Variance function:  $V(u) = u*(1-u/1)$  [Binomial]  
Link function :  $g(u) = \ln(u/(1-u))$  [Logit]

	AIC	=	.6831657
Log pseudolikelihood = -223.2763498	BIC	=	-4289.253

atndrte	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
priGPA	-3.371154	2.195517	-1.54	0.125	-7.67429	.9319806
priGPASq	1.886443	.8972586	2.10	0.036	.1278489	3.645038
priGPACu	-.2454989	.118004	-2.08	0.037	-.4767825	-.0142153
ACT	.5538998	.6744028	0.82	0.411	-.7679054	1.875705
ACTsq	-.0280986	.0304868	-0.92	0.357	-.0878516	.0316544
ACTcu	.0003858	.0004505	0.86	0.392	-.000497	.0012687
frosh	.3939498	.1155299	3.41	0.001	.1675154	.6203841
soph	.0941678	.1006569	0.94	0.350	-.1031161	.2914517
_cons	-.7731446	5.13392	-0.15	0.880	-10.83544	9.289154

```
. predict atndrteh
(option mu assumed; predicted mean atndrte)
```

```
. gen rh = atndrte - atndrteh
```

```
. gen rhsq = rh^2
```

```
. gen atndrtehsq = atndrteh^2
```

```
. reg rhsq atndrteh atndrtehsq
```

Source	SS	df	MS	Number of obs	=	680
Model	.098172929	2	.049086465	F( 2, 677)	=	37.14
Residual	.894850267	677	.001321788	Prob > F	=	0.0000
				R-squared	=	0.0989
				Adj R-squared	=	0.0962
Total	.993023196	679	.001462479	Root MSE	=	.03636

rhsq	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
------	-------	-----------	---	------	----------------------	--

atndrteh	.1604177	.1514883	1.06	0.290	-.1370257	.457861
atndrtehsq	-.1854786	.0994129	-1.87	0.063	-.3806733	.0097161
_cons	.0137235	.0571515	0.24	0.810	-.0984919	.1259389

```
. predict omegah
(option xb assumed; fitted values)
```

```
. sum omegah
```

Variable	Obs	Mean	Std. Dev.	Min	Max
omegah	680	.0192213	.0120243	-.0049489	.0483992

```
. count if omegah < 0
13
```

```
. drop omegah
```

```
. glm rhsq atndrteh atndrtehsq, fam(gamma) link(log)
```

Generalized linear models		No. of obs	=	680
Optimization	: ML	Residual df	=	677
		Scale parameter	=	3.781574
Deviance	= 1657.02942	(1/df) Deviance	=	2.447606
Pearson	= 2560.125871	(1/df) Pearson	=	3.781574

Variance function:	$V(u) = u^2$	[Gamma]
Link function	: $g(u) = \ln(u)$	[Log]

Log likelihood	= 2144.628097	AIC	= -6.298906
		BIC	= -2758.427

rhsq	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
atndrteh	21.19375	8.816922	2.40	0.016	3.912901	38.4746
atndrtehsq	-17.96346	5.764534	-3.12	0.002	-29.26174	-6.665185
_cons	-9.308977	3.33455	-2.79	0.005	-15.84458	-2.77338

```
. predict omegah
(option mu assumed; predicted mean rhsq)
```

```
. gen kh = rh*stndfnl/omegah
```

```
. sum kh
```

Variable	Obs	Mean	Std. Dev.	Min	Max
kh	680	.6890428	8.318917	-47.34537	41.90279

```
. reg stndfnl atndrte priGPA priGPAsq priGPacu ACT ACTsq ACTcu frosh soph,
robust
```

Linear regression		Number of obs	=	680
		F( 9, 670)	=	31.01



Prob > F = 0.0000  
R-squared = 0.2356  
Root MSE = .8709

stndfml	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
atndrte	.6444118	.2345274	2.75	0.006	.1839147	1.104909
priGPA	1.987666	2.651676	0.75	0.454	-3.21893	7.194262
priGPAsq	-1.055604	1.01697	-1.04	0.300	-3.052436	.9412284
priGPAcu	.1842414	.1262839	1.46	0.145	-.0637183	.4322011
ACT	.3059699	.7236971	0.42	0.673	-1.115017	1.726957
ACTsq	-.0141693	.0319499	-0.44	0.658	-.0769033	.0485648
ACTcu	.0002633	.0004629	0.57	0.570	-.0006456	.0011722
frosh	-.1138172	.1035799	-1.10	0.272	-.3171976	.0895631
soph	-.1863224	.0870459	-2.14	0.033	-.3572381	-.0154067
_cons	-4.501184	5.629291	-0.80	0.424	-15.55436	6.55199

**21.8.** a. From (21.129) and (21.130), we are assuming

$$a = \gamma_0 + \mathbf{x}\boldsymbol{\gamma} + u$$

$$E(u|\mathbf{x}, \mathbf{z}) = 0$$

and so we can write

$$y = \gamma_0 + \beta w + \mathbf{x}\boldsymbol{\gamma} + u + e$$

$$\equiv \gamma_0 + \beta w + \mathbf{x}\boldsymbol{\gamma} + r$$

where  $E(r|\mathbf{x}, \mathbf{z}) = E(u|\mathbf{x}, \mathbf{z}) + E(e|\mathbf{x}, \mathbf{z}) = 0$ . Therefore, we need to add the usual rank condition:

$\mathbf{z}$  must appear with nonzero coefficient vector in the linear projection of  $w$  on  $(1, \mathbf{x}, \mathbf{z})$ . More precisely, if

$$L(w|1, \mathbf{x}, \mathbf{z}) = \xi_0 + \mathbf{x}\boldsymbol{\xi}_1 + \mathbf{z}\boldsymbol{\xi}_2$$

then  $\boldsymbol{\xi}_2 \neq \mathbf{0}$ .

b. If

$$w = \max(0, \pi_0 + \mathbf{x}\boldsymbol{\pi}_1 + \mathbf{z}\boldsymbol{\pi}_2 + v),$$

$$D(v|\mathbf{x}, \mathbf{z}) \sim \text{Normal}(0, \eta^2),$$

then

$$E(w|\mathbf{x}, \mathbf{z}) = \Phi(\mathbf{q}\boldsymbol{\pi}/\eta) \cdot \mathbf{q}\boldsymbol{\pi} + \eta \cdot \phi(\mathbf{q}\boldsymbol{\pi}/\eta),$$

where  $\mathbf{q}\boldsymbol{\pi} \equiv \pi_0 + \mathbf{x}\pi_1 + \mathbf{z}\pi_2$  [see equation (17.14)]. Because  $E(w|\mathbf{x}, \mathbf{z})$  is a function of  $(\mathbf{x}, \mathbf{z})$  and we have

$$\begin{aligned} y &= \gamma_0 + \beta w + \mathbf{x}\boldsymbol{\gamma} + r \\ E(r|\mathbf{x}, \mathbf{z}) &= 0, \end{aligned}$$

we can use  $\Phi(\mathbf{q}\boldsymbol{\pi}/\eta) \cdot \mathbf{q}\boldsymbol{\pi} + \eta \cdot \phi(\mathbf{q}\boldsymbol{\pi}/\eta)$  as a valid instrument for  $w$ . (Remember, any function of  $(\mathbf{x}, \mathbf{z})$  is uncorrelated with  $r$  provided the second moments exist.) Because we do not know  $\boldsymbol{\pi}$  or  $\eta$ , we replace them with estimators. In other words, use  $\Phi(\mathbf{q}_i \hat{\boldsymbol{\pi}}/\hat{\eta}) \cdot \mathbf{q}_i \hat{\boldsymbol{\pi}} + \hat{\eta} \cdot \phi(\mathbf{q}_i \hat{\boldsymbol{\pi}}/\hat{\eta})$  as the IV for  $w_i$ , where  $\hat{\boldsymbol{\pi}}$  and  $\hat{\eta}$  are the estimates from an initial Tobit MLE.

c. By equation (14.57), the optimal IV for  $w$  is

$$E(w|\mathbf{x}, \mathbf{z})/\text{Var}(r|\mathbf{x}, \mathbf{z}).$$

But  $E(e|a, \mathbf{x}, \mathbf{z}) = 0$  implies that  $e$  and  $a$  are uncorrelated, conditional on  $(\mathbf{x}, \mathbf{z})$ . Therefore,

$$\begin{aligned} \text{Var}(u + e|\mathbf{x}, \mathbf{z}) &= \text{Var}(u|\mathbf{x}, \mathbf{z}) + \text{Var}(e|\mathbf{x}, \mathbf{z}) \\ &= \sigma_a^2 + \sigma_e^2 \equiv \sigma_r^2. \end{aligned}$$

Therefore,  $\text{Var}(r|\mathbf{x}, \mathbf{z})$  is constant, and  $E(w|\mathbf{x}, \mathbf{z})$  can serve as the optimal IV for  $w$ . As usual, we replace the parameters in  $E(w|\mathbf{x}, \mathbf{z})$  with  $\sqrt{N}$ -consistent estimators. The results in Chapter 6 on generated instruments can be used to show that the resulting IV estimator has the same  $\sqrt{N}$ -asymptotic distribution as if we know  $\boldsymbol{\pi}$  and  $\eta$ .

d. An alternative method would be to run the OLS regression

$$y_i \text{ on } 1, \hat{w}_i, \mathbf{x}_i, i = 1, \dots, N$$

where

$$\hat{w}_i \equiv \Phi(\mathbf{q}_i \hat{\boldsymbol{\pi}}/\hat{\eta}) \cdot \mathbf{q}_i \hat{\boldsymbol{\pi}} + \hat{\eta} \cdot \phi(\mathbf{q}_i \hat{\boldsymbol{\pi}}/\hat{\eta})$$

are the estimated conditional means. While this “plug-in” approach may produce estimates similar to the IV approach, it is less preferred for the same reasons we covered for the probit case. First, using the  $\hat{w}_i$  as regressors rather than instruments is less robust: using them as regressors essentially requires the Tobit model to be correctly specified for  $w$  given  $(\mathbf{x}, \mathbf{z})$ . Second, valid standard errors are harder to get using  $\hat{w}_i$  as a regressor as opposed to an IV. Third, the plug-in procedure does not appear to be optimal within an interesting class of estimators. (By contrast, we know that the IV estimator is optimal in the class of IV estimators under the assumptions given for part c.)

e. Estimate  $y_i = \eta_0 + \mathbf{x}_i\boldsymbol{\gamma} + \beta w_i + w_i \cdot (\mathbf{x}_i - \bar{\mathbf{x}})\boldsymbol{\delta} + \text{error}_i$  by IV, using instruments,  $[1, \mathbf{x}_i, \hat{w}_i, \hat{w}_i \cdot (\mathbf{x}_i - \bar{\mathbf{x}})]$  as instruments, where  $\hat{w}_i$  are the Tobit fitted values. This would be generally inefficient, as the error,  $r$ , is not necessarily homoskedastic. Another drawback is that this would not generate overidentifying restrictions.

**21.9.** a. A histogram of the estimated propensity score for the untreated ( $\text{train} = 0$ ) and treated ( $\text{train} = 1$ ) cases is given below. There is a clear problem with overlap, as can be seen by studying the histogram for the control group: over 80% of units have propensity scores that are zero or practically zero. This means there are values of  $\mathbf{x}$  where  $p(\mathbf{x}) = 0$  or is barely distinguishable from zero.

The large differences in the histograms for the control and treatment groups spells trouble. Because  $p(\mathbf{x})$  is just a particular function of  $\mathbf{x}$ , ideally its distribution would be similar across the control and treatment groups, and this is clearly not the case. The problems this causes is easily reasoned when thinking of matching on the propensity score. We need to find both control and treated units with similar values of  $p(\mathbf{x})$ , but the histograms make it clear that there are very few in the control group with  $\hat{p}(\mathbf{x}_i) > .5$ , whereas this is where the bulk of the

observations lie for the treated group.

For comparison, the same histograms are plotted using the experimental data in JTRAIN2.RAW. Now the histograms are virtually indistinguishable and neither has mass at zero or one.

```
. use jtrain3
```

```
. logit train age educ black hisp married re74 re75
```

```
Logistic regression               Number of obs   =       2675
                                LR chi2(7)         =       872.82
                                Prob > chi2         =       0.0000
Log likelihood = -236.23799        Pseudo R2      =       0.6488
```

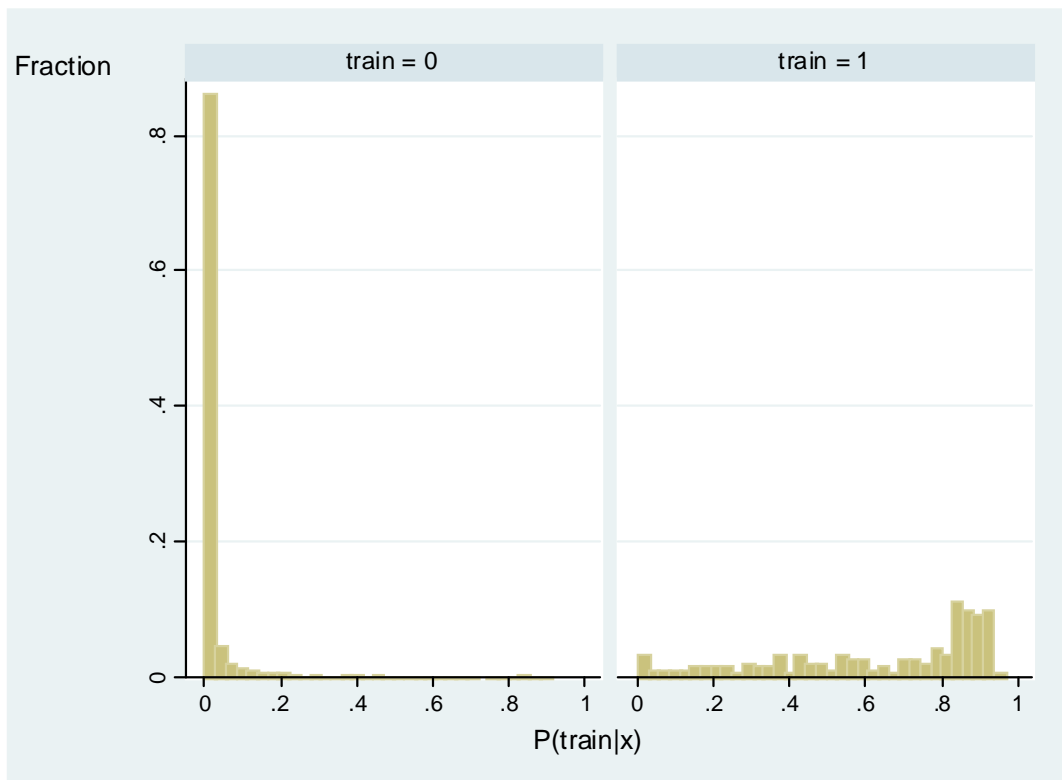
train	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
age	-.0840291	.014761	-5.69	0.000	-.1129601	-.055098
educ	-.0624764	.0513973	-1.22	0.224	-.1632134	.0382605
black	2.242955	.3176941	7.06	0.000	1.620286	2.865624
hisp	2.094338	.5584561	3.75	0.000	.9997841	3.188892
married	-1.588358	.2602448	-6.10	0.000	-2.098428	-1.078287
re74	-.117043	.0293604	-3.99	0.000	-.1745882	-.0594977
re75	-.2577589	.0394991	-6.53	0.000	-.3351758	-.1803421
_cons	2.302714	.9112559	2.53	0.012	.5166853	4.088743

Note: 158 failures and 0 successes completely determined.

```
. predict phat
```

```
(option pr assumed; Pr(train))
```

```
. histogram phat, fraction by(train)
```



```
. use jtrain2
```

```
. logit train age educ black hisp married re74 re75
```

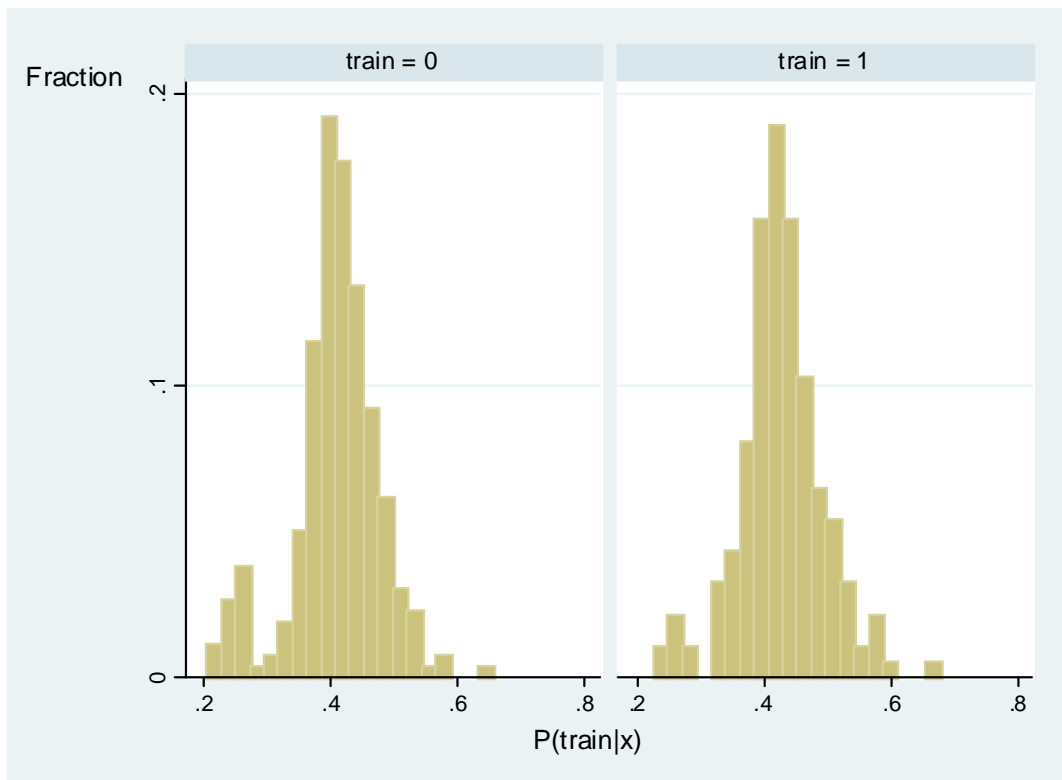
```
Logistic regression               Number of obs   =          445
                                LR chi2(7)        =           8.58
                                Prob > chi2        =          0.2840
Log likelihood = -297.80826       Pseudo R2      =          0.0142
```

train	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
age	.0107155	.014017	0.76	0.445	-.0167572	.0381882
educ	.0628366	.0558026	1.13	0.260	-.0465346	.1722077
black	-.3553063	.3577202	-0.99	0.321	-1.056425	.3458123
hisp	-.9322569	.5001292	-1.86	0.062	-1.912492	.0479784
married	.1440193	.2734583	0.53	0.598	-.3919492	.6799878
re74	-.0221324	.0252097	-0.88	0.380	-.0715425	.0272777
re75	.0459029	.0429705	1.07	0.285	-.0383177	.1301235
_cons	-.9237055	.7693924	-1.20	0.230	-2.431687	.5842759

```
. predict phat
```

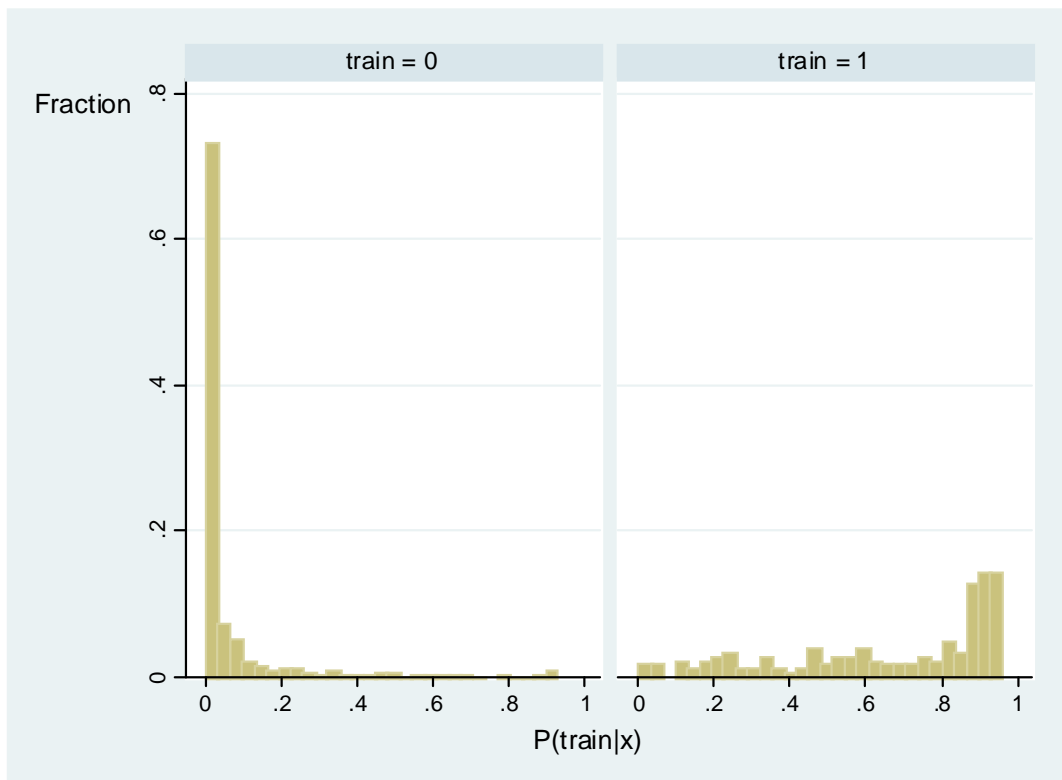
```
(option pr assumed; Pr(train))
```

```
. histogram phat, fraction by(train)
```





b. Using the sample restricted to  $avgre \leq 15$  helps a little in that there now seem to be at least some untreated units in bins with  $\hat{p}(\mathbf{x}_i) > .3$ . But there are not many. The pile-up at near zero for the control group is still present. The two histograms still look very different from the experimental data in JTRAIN2.RAW.



c. (Bonus Part) Suppose that using all 2,675 observations that, after estimating the logit model for *train*, we drop all data with  $\hat{p}(\mathbf{x}_i) < .05$ . We then reestimate the logit model using the remaining observations. How many observations are left? Obtain the resulting histograms as in part a and part b.

### **Solution**

The Stata session is given below. Only 422 observations are left after dropping those with  $\hat{p}(\mathbf{x}_i) < .05$ . The histograms look much better in terms of overlap: for the most part, it appears that for  $\hat{p}(\mathbf{x}_i)$  within a given bin, there are both treated and untreated observations. But the skew of the distributions is completely different (and not too surprising).

```
. use jtrain3

. qui logit train age educ black hisp married re74 re75

. predict phat
(option pr assumed; Pr(train))

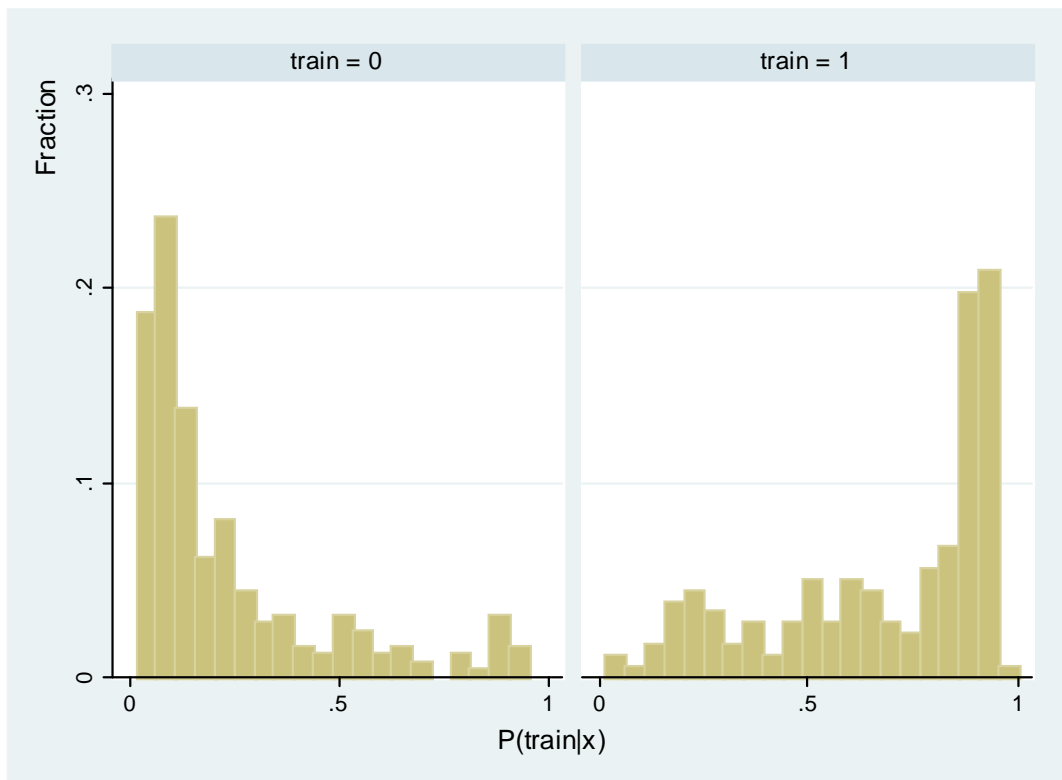
. drop if phat < .05
(2253 observations deleted)

. drop phat

. qui logit train age educ black hisp married re74 re75

. predict phat
(option pr assumed; Pr(train))

. histogram phat, fraction by(train)
```



**21.10.** a. This is just problem in the asymptotic theory of simple regression with a binary explanatory variable. With  $y_i = \mu_0 + \tau w_i + v_{i0}$  we have that  $w_i$  is independent of  $v_{i0}$ , and so there is no heteroskedasticity. It follows that (see Theorem 4.2)

$$\text{Avar}[\sqrt{N}(\hat{\tau} - \tau)] = \frac{\text{Var}(v_{i0})}{\text{Var}(w_i)} = \frac{\text{Var}(v_{i0})}{\rho(1 - \rho)}$$

because  $P(w_i = 1) = \rho$ . This means, by definition,

$$\text{Avar}(\tilde{\tau}) = \frac{\text{Var}(v_{i0})}{N\rho(1 - \rho)}.$$

b. By definition of the linear projection we can write

$$\begin{aligned} y_{i0} &= \alpha_0 + \mathbf{x}_i \boldsymbol{\beta}_0 + u_{i0} \\ E(u_{i0}) &= 0, E(\mathbf{x}_i' u_{i0}) = \mathbf{0}. \end{aligned}$$

Now we just plug this into  $y_i = y_{i0} + \tau w_i$ :

$$y_i = \alpha_0 + \mathbf{x}_i \boldsymbol{\beta}_0 + u_{i0} + \tau w_i = \alpha_0 + \tau w_i + \mathbf{x}_i \boldsymbol{\beta}_0 + u_{i0}.$$

The problem says to assume that  $w_i$  is independent of  $(y_{i0}, \mathbf{x}_i)$  and so  $w_i$  is actually independent of  $(\mathbf{x}_i, u_{i0})$  [because  $u_{i0}$  is a function of  $(y_{i0}, \mathbf{x}_i)$ ].

c. Let  $\mathbf{z}_i \equiv (w_i, \mathbf{x}_i)$  be the set of nonconstant regressors and let  $\boldsymbol{\gamma} = (\tau, \boldsymbol{\beta}_0')'$ . Then, as we showed in Chapter 4, if  $\hat{\boldsymbol{\gamma}}$  is the OLS estimator under random sampling,

$$\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) = [\text{Var}(\mathbf{z}_i)]^{-1} N^{-1/2} \sum_{i=1}^N (\mathbf{z}_i - \boldsymbol{\mu}_z)' u_{i0} + o_p(1).$$

Given that  $\text{Cov}(\mathbf{x}_i, w_i) = \mathbf{0}$ , we restrict attention to the first element,  $\sqrt{N}(\hat{\tau} - \tau)$ , we get

$$\sqrt{N}(\hat{\tau} - \tau) = [\text{Var}(w_i)]^{-1} N^{-1/2} \sum_{i=1}^N (w_i - \rho)' u_{i0} + o_p(1).$$

Therefore, using independence between  $u_{i0}$  and  $w_i$ ,

$$\text{Avar}\left[\sqrt{N}(\hat{\tau} - \tau)\right] = \frac{\text{Var}[(w_i - \rho)u_{i0}]}{[\text{Var}(w_i)]^2} = \frac{\text{Var}(u_{i0})}{\rho(1 - \rho)}$$

and so

$$\text{Avar}(\hat{\tau}) = \frac{\text{Var}(u_{i0})}{N\rho(1 - \rho)}.$$

d. Because  $y_{i0} = \mu_0 + v_{i0}$ ,  $\text{Var}(v_{i0}) = \text{Var}(y_{i0})$ . Using the linear projection representation

$$y_{i0} = \alpha_0 + \mathbf{x}_i\boldsymbol{\beta}_0 + u_{i0},$$

$$\text{Var}(y_{i0}) = \text{Var}(\mathbf{x}_i\boldsymbol{\beta}_0) + \text{Var}(u_{i0})$$

and, assuming that  $\text{Var}(\mathbf{x}_i)$  has full rank,  $\text{Var}(\mathbf{x}_i\boldsymbol{\beta}_0) > 0$  whenever  $\boldsymbol{\beta}_0 \neq \mathbf{0}$ . So

$$\text{Var}(u_{i0}) < \text{Var}(y_{i0}) = \text{Var}(v_{i0}).$$

It follows by comparing  $\text{Avar}(\hat{\tau})$  and  $\text{Avar}(\tilde{\tau})$  that  $\text{Avar}(\hat{\tau}) < \text{Avar}(\tilde{\tau})$  whenever  $\boldsymbol{\beta}_0 \neq \mathbf{0}$ , that is, whenever  $\mathbf{x}_i$  is correlated with  $y_{i0}$ .

e. Even though  $\hat{\tau}$  is asymptotically more efficient than  $\tilde{\tau}$ ,  $\hat{\tau}$  is generally biased if

$E(y_{i0}|\mathbf{x}_i) \neq \alpha_0 + \mathbf{x}_i\boldsymbol{\beta}_0$ . [If  $E(y_{i0}|\mathbf{x}_i) = \alpha_0 + \mathbf{x}_i\boldsymbol{\beta}_0$  then  $\hat{\tau}$  would be conditionally unbiased, that is,

$E(\hat{\tau}|\mathbf{W}, \mathbf{X}) = \tau$ .] The difference-in-means estimator  $\tilde{\tau}$  is unbiased conditional on  $\mathbf{W}$  because

$$E(y_i|\mathbf{W}) = E(y_i|w_i) = \mu_0 + \tau w_i.$$

**21.11.** Suppose that we allow full slope, as well as intercept, heterogeneity in a linear representation of two counterfactual outcomes,

$$y_{i0} = a_{i0} + \mathbf{x}_i\mathbf{b}_{i0}$$

$$y_{i1} = a_{i1} + \mathbf{x}_i\mathbf{b}_{i1}$$

Assume that the vector  $(\mathbf{x}_i, \mathbf{z}_i)$  is independent of  $(a_{i0}, \mathbf{b}_{i0}, a_{i1}, \mathbf{b}_{i1})$  – which makes, as we will see,  $\mathbf{z}_i$  instrumental variables candidates in a control function or correction function setting.

a. Because  $\mathbf{x}_i$  is independent of  $\mathbf{b}_{ig}$ ,  $g = 0, 1$ ,

$$\begin{aligned} E(y_{ig}) &= E(a_{ig}) + E(\mathbf{x}_i \mathbf{b}_{ig}) = \alpha_g + E(\mathbf{x}_i)E(\mathbf{b}_{ig}) \\ &= \alpha_g + \boldsymbol{\psi} \boldsymbol{\beta}_g, g = 0, 1. \end{aligned}$$

b. From part a,

$$\begin{aligned} \mu_0 &= \alpha_0 + \boldsymbol{\psi} \boldsymbol{\beta}_0 \\ \mu_1 &= \alpha_1 + \boldsymbol{\psi} \boldsymbol{\beta}_1 \end{aligned}$$

and so

$$\tau = (\alpha_1 - \alpha_0) + \boldsymbol{\psi}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) = (\alpha_1 - \alpha_0) + \boldsymbol{\psi} \boldsymbol{\delta}.$$

Also,

$$y_{ig} = \alpha_g + \mathbf{x}_i \boldsymbol{\beta}_g + c_{ig} + \mathbf{x}_i \mathbf{f}_{ig}, g = 0, 1$$

and so

$$\begin{aligned} y_i &= (1 - w_i)y_{i0} + w_i y_{i1} \\ &= (1 - w_i)(\alpha_0 + \mathbf{x}_i \boldsymbol{\beta}_0 + c_{i0} + \mathbf{x}_i \mathbf{f}_{i0}) + w_i(\alpha_1 + \mathbf{x}_i \boldsymbol{\beta}_1 + c_{i1} + \mathbf{x}_i \mathbf{f}_{i1}) \\ &= \alpha_0 + (\alpha_1 - \alpha_0)w_i + \mathbf{x}_i \boldsymbol{\beta}_0 + w_i \mathbf{x}_i (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) \\ &\quad + c_{i0} + w_i(c_{i1} - c_{i0}) + \mathbf{x}_i \mathbf{f}_{i0} + w_i \mathbf{x}_i (\mathbf{f}_{i1} - \mathbf{f}_{i0}) \\ &\equiv \alpha_0 + (\alpha_1 - \alpha_0)w_i + \mathbf{x}_i \boldsymbol{\beta}_0 + w_i \mathbf{x}_i \boldsymbol{\delta} + c_{i0} + w_i e_i + \mathbf{x}_i \mathbf{f}_{i0} + w_i \mathbf{x}_i \mathbf{d}_i. \end{aligned}$$

Next, substitute

$$\begin{aligned} \alpha_0 &= \mu_0 - \boldsymbol{\psi} \boldsymbol{\beta}_0 \\ \alpha_1 - \alpha_0 &= \tau - \boldsymbol{\psi} \boldsymbol{\delta} \end{aligned}$$

to get

$$\begin{aligned} y_i &= \mu_0 - \boldsymbol{\psi} \boldsymbol{\beta}_0 + (\tau - \boldsymbol{\psi} \boldsymbol{\delta})w_i + \mathbf{x}_i \boldsymbol{\beta}_0 + w_i \mathbf{x}_i \boldsymbol{\delta} + c_{i0} + w_i e_i + \mathbf{x}_i \mathbf{f}_{i0} + w_i \mathbf{x}_i \mathbf{d}_i \\ &= \mu_0 + \tau w_i + (\mathbf{x}_i - \boldsymbol{\psi}) \boldsymbol{\beta}_0 + w_i (\mathbf{x}_i - \boldsymbol{\psi}) \boldsymbol{\delta} + c_{i0} + \mathbf{x}_i \mathbf{f}_{i0} + w_i e_i + w_i \mathbf{x}_i \mathbf{d}_i, \end{aligned}$$

which is what we wanted to show.

c. So that there is no notational conflict, write  $E(\mathbf{d}_i | a_i, \mathbf{x}_i, \mathbf{z}_i) = \boldsymbol{\zeta} a_i$  and keep  $\boldsymbol{\theta}$  to index the



binary response model. Now take the expectation of (21.149) conditional on  $(a_i, \mathbf{x}_i, \mathbf{z}_i)$ , using the fact that  $w_i$  is a function of  $(a_i, \mathbf{x}_i, \mathbf{z}_i)$ :

$$\begin{aligned} E(y_i|a_i, \mathbf{x}_i, \mathbf{z}_i) &= \mu_0 + \tau w_i + (\mathbf{x}_i - \boldsymbol{\psi})\boldsymbol{\beta}_0 + w_i(\mathbf{x}_i - \boldsymbol{\psi})\boldsymbol{\delta} \\ &\quad + E(c_{i0}|a_i, \mathbf{x}_i, \mathbf{z}_i) + \mathbf{x}_i E(\mathbf{f}_{i0}|a_i, \mathbf{x}_i, \mathbf{z}_i) \\ &\quad + w_i E(e_i|a_i, \mathbf{x}_i, \mathbf{z}_i) + w_i \mathbf{x}_i E(\mathbf{d}_i|a_i, \mathbf{x}_i, \mathbf{z}_i) \\ &= \mu_0 + \tau w_i + (\mathbf{x}_i - \boldsymbol{\psi})\boldsymbol{\beta}_0 + w_i(\mathbf{x}_i - \boldsymbol{\psi})\boldsymbol{\delta} \\ &\quad + \rho_0 a_i + a_i \mathbf{x}_i \boldsymbol{\eta}_0 + \xi w_i a_i + w_i a_i \mathbf{x}_i \boldsymbol{\zeta} \end{aligned}$$

d. Just use iterated expectations along with

$$E(a_i|w_i, \mathbf{q}_i) = h(w_i, \mathbf{q}_i \boldsymbol{\theta}) \equiv w_i \lambda(\mathbf{q}_i \boldsymbol{\theta}) - (1 - w_i) \lambda(-\mathbf{q}_i \boldsymbol{\theta}).$$

So

$$\begin{aligned} E(y_i|w_i, \mathbf{x}_i, \mathbf{z}_i) &= \mu_0 + \tau w_i + (\mathbf{x}_i - \boldsymbol{\psi})\boldsymbol{\beta}_0 + w_i(\mathbf{x}_i - \boldsymbol{\psi})\boldsymbol{\delta} \\ &\quad + \rho_0 h(w_i, \mathbf{q}_i \boldsymbol{\theta}) + h(w_i, \mathbf{q}_i \boldsymbol{\theta}) \mathbf{x}_i \boldsymbol{\eta}_0 + \xi w_i h(w_i, \mathbf{q}_i \boldsymbol{\theta}) + w_i h(w_i, \mathbf{q}_i \boldsymbol{\theta}) \mathbf{x}_i \boldsymbol{\zeta} \end{aligned}$$

e. Given  $E(y_i|w_i, \mathbf{x}_i, \mathbf{z}_i)$ , the CF method is straightforward. In the first step, estimate probit of  $w_i$  on  $\mathbf{q}_i$  to get  $\hat{\boldsymbol{\theta}}$ , and then compute  $\hat{h}_i \equiv h(w_i, \mathbf{q}_i \hat{\boldsymbol{\theta}})$ . Then run the OLS regression

$$y_i \text{ on } 1, w_i, \mathbf{x}_i - \bar{\mathbf{x}}, w_i(\mathbf{x}_i - \bar{\mathbf{x}}), \hat{h}_i, \hat{h}_i \mathbf{x}_i, w_i \hat{h}_i, w_i \hat{h}_i \mathbf{x}_i, i = 1, \dots, N.$$

We replace  $\boldsymbol{\psi}$  with the sample average,  $\bar{\mathbf{x}}$ . The coefficient on  $w_i$  is  $\hat{\tau}$ .

Compared with the regression in equation (21.85), we have included the interactions  $\hat{h}_i \mathbf{x}$  and  $w_i \hat{h}_i \mathbf{x}_i$ . These account for the random coefficients in the counterfactual equations.

f. Of course we could work through the delta method to obtain a valid asymptotic standard error for  $\hat{\tau}$ , but bootstrapping both steps in the procedure provides a simple alternative.

g. We just compute  $E(y_{ig}|\mathbf{x})$  for  $g = 0, 1$ :

$$\begin{aligned}
E(y_{ig}|\mathbf{x}_i) &= E(a_{ig} + \mathbf{x}_i \mathbf{b}_{ig}|\mathbf{x}_i) \\
&= E(a_{ig}|\mathbf{x}_i) + \mathbf{x}_i E(\mathbf{b}_{ig}|\mathbf{x}_i) \\
&= \alpha_g + \mathbf{x}_i \boldsymbol{\beta}_g
\end{aligned}$$

where the last equality holds by the independence assumption. Therefore,

$$\begin{aligned}
\tau_{ate}(\mathbf{x}) &= (\alpha_1 - \alpha_0) + \mathbf{x}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) \\
&= \tau - \boldsymbol{\psi}\boldsymbol{\delta} + \mathbf{x}\boldsymbol{\delta} = \tau + (\mathbf{x} - \boldsymbol{\psi})\boldsymbol{\delta}.
\end{aligned}$$

So

$$\hat{\tau}_{ate}(\mathbf{x}) = \hat{\tau} + (\mathbf{x} - \bar{\mathbf{x}})\hat{\boldsymbol{\delta}}.$$

**21.12.** a. The terms  $c_{i0}$  and  $\mathbf{x}_i \mathbf{f}_{i0}$  have zero means conditional on  $(\mathbf{x}_i, \mathbf{z}_i)$  by the independence of all heterogeneity terms –  $(a_{i0}, \mathbf{b}_{i0}, a_{i1}, \mathbf{b}_{i1})$  – and  $(\mathbf{x}_i, \mathbf{z}_i)$ . Remember,  $c_{i0} = a_{i0} - \alpha_0$  and  $\mathbf{f}_{i0} = \mathbf{b}_{i0} - \boldsymbol{\beta}_0$ .

b. The correction functions in this case are  $E(w_i e_i|\mathbf{x}_i, \mathbf{z}_i)$  and  $E(w_i \mathbf{x}_i \mathbf{d}_i|\mathbf{x}_i, \mathbf{z}_i) = \mathbf{x}_i E(w_i \mathbf{d}_i|\mathbf{x}_i, \mathbf{z}_i)$ . Now we just use the formula in equation (21.80) because  $E(e_i|a_i, \mathbf{x}_i, \mathbf{z}_i) = \xi a_i$  and  $E(\mathbf{d}_i|a_i, \mathbf{x}_i, \mathbf{z}_i) = \zeta a_i$ . Therefore,

$$\begin{aligned}
E(w_i e_i|\mathbf{x}_i, \mathbf{z}_i) &= \xi \phi(\mathbf{q}_i \boldsymbol{\theta}) \\
E(w_i \mathbf{x}_i \mathbf{d}_i|\mathbf{x}_i, \mathbf{z}_i) &= \mathbf{x}_i \phi(\mathbf{q}_i \boldsymbol{\theta}) \zeta = \phi(\mathbf{q}_i \boldsymbol{\theta}) \mathbf{x}_i \zeta.
\end{aligned}$$

c. From part b we can write

$$y_i = \mu_0 + \tau w_i + (\mathbf{x}_i - \boldsymbol{\psi})\boldsymbol{\beta}_0 + w_i(\mathbf{x}_i - \boldsymbol{\psi})\boldsymbol{\delta} + \xi \phi(\mathbf{q}_i \boldsymbol{\theta}) + \phi(\mathbf{q}_i \boldsymbol{\theta}) \mathbf{x}_i \zeta + c_{i0} + \mathbf{x}_i \mathbf{f}_{i0} + r_i$$

where

$$r_i = [w_i e_i - E(w_i e_i|\mathbf{x}_i, \mathbf{z}_i)] + [w_i \mathbf{x}_i \mathbf{d}_i - E(w_i \mathbf{x}_i \mathbf{d}_i|\mathbf{x}_i, \mathbf{z}_i)]$$

and so  $E(r_i|\mathbf{x}_i, \mathbf{z}_i) = 0$ . The estimating equation, after the first-stage probit to get  $\hat{\boldsymbol{\theta}}$ , is

$$y_i = \mu_0 + \tau w_i + (\mathbf{x}_i - \bar{\mathbf{x}})\boldsymbol{\beta}_0 + w_i(\mathbf{x}_i - \bar{\mathbf{x}})\boldsymbol{\delta} + \xi \phi(\mathbf{q}_i \hat{\boldsymbol{\theta}}) + \phi(\mathbf{q}_i \hat{\boldsymbol{\theta}}) \mathbf{x}_i \zeta + error_i$$

which we can estimate using IV with instruments, say,

$$[1, \hat{\Phi}_i, (\mathbf{x}_i - \bar{\mathbf{x}}), \hat{\Phi}_i \cdot (\mathbf{x}_i - \bar{\mathbf{x}}), \hat{\phi}_i, \hat{\phi}_i \cdot \mathbf{x}_i]$$

d. Under the null hypothesis,  $\xi = 0$  and  $\zeta = \mathbf{0}$ . The conditions sufficient to ignore the first-stage estimator for IV estimators in Chapter 6 hold here, so we can use a standard Wald test (perhaps made robust to heteroskedasticity) to test joint significance of the  $K + 1$  variables  $(\hat{\phi}_i, \hat{\phi}_i \cdot \mathbf{x}_i)$ . Remember, these are acting as their own instruments in the estimation.

**21.13.** Fractional probit or logit are natural, or some other model that keeps the fitted values in the unit interval. For the treatment rule  $w_i = 1[x_i \geq c]$ , let  $G(\hat{\alpha}_0 + \hat{\beta}_0 x)$  be the estimated fractional response model using the data with  $x_i < c$  and let  $G(\hat{\alpha}_1 + \hat{\beta}_1 x)$  be the estimated model using the data with  $x_i \geq c$ . Probably the Bernoulli QMLE would be used. Then similar to equation (21.104),

$$\hat{\tau}_c = G(\hat{\alpha}_1 + \hat{\beta}_1 c) - G(\hat{\alpha}_0 + \hat{\beta}_0 c).$$

The delta method or bootstrapping can be used to obtain valid inference for

$$\tau_c = G(\alpha_1 + \beta_1 c) - G(\alpha_0 + \beta_0 c).$$

**21.14.** a. Just take the expected value of  $y_{it}(g) = a_{itg} + \mathbf{x}_{it}\boldsymbol{\beta}_g$ :

$$\mu_{gt} = E(a_{itg}) + E(\mathbf{x}_{it})\boldsymbol{\beta}_g = \alpha_{tg} + \boldsymbol{\psi}_t\boldsymbol{\beta}_g, g = 0, 1.$$

Therefore, for each  $t$ ,

$$\tau_{t,ate} = (\alpha_{t1} - \alpha_{t0}) + \boldsymbol{\psi}_t(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0).$$

b. Use  $\alpha_{t0} = \mu_{t0} - \boldsymbol{\psi}_t\boldsymbol{\beta}_0$  and  $(\alpha_{t1} - \alpha_{t0}) = \tau_t - \boldsymbol{\psi}_t(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$  and plug into the equation for  $y_{it}$ :

$$\begin{aligned}
y_{it} &= (1 - w_{it})(\alpha_{t0} + \mathbf{x}_{it}\boldsymbol{\beta}_0 + c_{it0}) + w_{it}(\alpha_{t1} + \mathbf{x}_{it}\boldsymbol{\beta}_1 + c_{it1}) \\
&= \alpha_{t0} + w_{it}(\alpha_{t1} - \alpha_{t0}) + \mathbf{x}_{it}\boldsymbol{\beta}_0 + w_{it}\mathbf{x}_{it}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + c_{it0} + w_{it}(c_{it1} - c_{it0}) \\
&= (\mu_{t0} - \boldsymbol{\psi}_t\boldsymbol{\beta}_0) + w_{it}[\tau_t - \boldsymbol{\psi}_t(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)] + \\
&\quad + \mathbf{x}_{it}\boldsymbol{\beta}_0 + w_{it}\mathbf{x}_{it}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + c_{it0} + w_{it}(c_{it1} - c_{it0}) \\
&= \mu_{t0} + \tau_t w_{it} + (\mathbf{x}_{it} - \boldsymbol{\psi}_t)\boldsymbol{\beta}_0 + w_{it}(\mathbf{x}_{it} - \boldsymbol{\psi}_t)\boldsymbol{\delta} + c_{it0} + w_{it}e_{it}
\end{aligned}$$

where  $\boldsymbol{\delta} \equiv \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$  and  $e_{it} \equiv c_{it1} - c_{it0}$ .

c. Plugging in for  $c_{it0}$  and  $e_{it}$  gives

$$\begin{aligned}
y_{it} &= \mu_{t0} + \tau_t w_{it} + (\mathbf{x}_{it} - \boldsymbol{\psi}_t)\boldsymbol{\beta}_0 + w_{it}(\mathbf{x}_{it} - \boldsymbol{\psi}_t)\boldsymbol{\delta} \\
&\quad + (\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}})\boldsymbol{\xi}_1 + (\bar{\mathbf{z}}_i - \boldsymbol{\mu}_{\bar{\mathbf{z}}})\boldsymbol{\xi}_2 + r_{it0} + w_{it}[(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}})\boldsymbol{\eta}_1 + (\bar{\mathbf{z}}_i - \boldsymbol{\mu}_{\bar{\mathbf{z}}})\boldsymbol{\eta}_2 + v_{it}] \\
&= \mu_{t0} + \tau_t w_{it} + (\mathbf{x}_{it} - \boldsymbol{\psi}_t)\boldsymbol{\beta}_0 + w_{it}(\mathbf{x}_{it} - \boldsymbol{\psi}_t)\boldsymbol{\delta} \\
&\quad + (\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}})\boldsymbol{\xi}_1 + (\bar{\mathbf{z}}_i - \boldsymbol{\mu}_{\bar{\mathbf{z}}})\boldsymbol{\xi}_2 + w_{it}(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}})\boldsymbol{\eta}_1 + w_{it}(\bar{\mathbf{z}}_i - \boldsymbol{\mu}_{\bar{\mathbf{z}}})\boldsymbol{\eta}_2 \\
&\quad + r_{it0} + w_{it}v_{it}
\end{aligned}$$

d. As usual, we first condition on  $(q_{it}, \mathbf{x}_i, \mathbf{z}_i)$ :

$$\begin{aligned}
E(y_{it}|q_{it}, \mathbf{x}_i, \mathbf{z}_i) &= \mu_{t0} + \tau_t w_{it} + (\mathbf{x}_{it} - \boldsymbol{\psi}_t)\boldsymbol{\beta}_0 + w_{it}(\mathbf{x}_{it} - \boldsymbol{\psi}_t)\boldsymbol{\delta} \\
&\quad + (\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}})\boldsymbol{\xi}_1 + (\bar{\mathbf{z}}_i - \boldsymbol{\mu}_{\bar{\mathbf{z}}})\boldsymbol{\xi}_2 + w_{it}(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}})\boldsymbol{\eta}_1 + w_{it}(\bar{\mathbf{z}}_i - \boldsymbol{\mu}_{\bar{\mathbf{z}}})\boldsymbol{\eta}_2 \\
&\quad + E(r_{it0}|q_{it}, \mathbf{x}_i, \mathbf{z}_i) + w_{it}E(v_{it}|q_{it}, \mathbf{x}_i, \mathbf{z}_i) \\
&= \mu_{t0} + \tau_t w_{it} + (\mathbf{x}_{it} - \boldsymbol{\psi}_t)\boldsymbol{\beta}_0 + w_{it}(\mathbf{x}_{it} - \boldsymbol{\psi}_t)\boldsymbol{\delta} \\
&\quad + (\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}})\boldsymbol{\xi}_1 + (\bar{\mathbf{z}}_i - \boldsymbol{\mu}_{\bar{\mathbf{z}}})\boldsymbol{\xi}_2 + w_{it}(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}})\boldsymbol{\eta}_1 + w_{it}(\bar{\mathbf{z}}_i - \boldsymbol{\mu}_{\bar{\mathbf{z}}})\boldsymbol{\eta}_2 \\
&\quad + \alpha_0 q_{it} + \rho w_{it} q_{it}.
\end{aligned}$$

Now condition on  $(w_{it}, \mathbf{x}_i, \mathbf{z}_i)$  using iterated expectations:

$$\begin{aligned}
E(y_{it}|w_{it}, \mathbf{x}_i, \mathbf{z}_i) &= \mu_{t0} + \tau_t w_{it} + (\mathbf{x}_{it} - \boldsymbol{\psi}_t)\boldsymbol{\beta}_0 + w_{it}(\mathbf{x}_{it} - \boldsymbol{\psi}_t)\boldsymbol{\delta} \\
&\quad + (\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}})\boldsymbol{\xi}_1 + (\bar{\mathbf{z}}_i - \boldsymbol{\mu}_{\bar{\mathbf{z}}})\boldsymbol{\xi}_2 + w_{it}(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_{\bar{\mathbf{x}}})\boldsymbol{\eta}_1 + w_{it}(\bar{\mathbf{z}}_i - \boldsymbol{\mu}_{\bar{\mathbf{z}}})\boldsymbol{\eta}_2 \\
&\quad + \alpha_0 h(w_{it}, \mathbf{g}_{it}\boldsymbol{\theta}) + \rho w_{it} h(w_{it}, \mathbf{g}_{it}\boldsymbol{\theta}),
\end{aligned}$$

where

$$h(w_{it}, \mathbf{g}_{it}\boldsymbol{\theta}) = w_{it}\lambda(\mathbf{g}_{it}\boldsymbol{\theta}) - (1 - w_{it})\lambda(-\mathbf{g}_{it}\boldsymbol{\theta})$$

and  $\mathbf{g}_{it}\boldsymbol{\theta} = \theta_0 + \mathbf{x}_{it}\boldsymbol{\theta}_1 + \mathbf{z}_{it}\boldsymbol{\theta}_2 + \bar{\mathbf{x}}_i\boldsymbol{\theta}_3 + \bar{\mathbf{z}}_i\boldsymbol{\theta}_4$ .

e. In the first step we can use pooled probit of  $w_{it}$  on 1,  $\mathbf{x}_{it}$ ,  $\mathbf{z}_{it}$ ,  $\bar{\mathbf{x}}_i$ ,  $\bar{\mathbf{z}}_i$  to obtain  $\hat{\boldsymbol{\theta}}$  and  $h(w_{it}, \mathbf{g}_{it}\hat{\boldsymbol{\theta}})$ . Then we can use pooled OLS in a second step:

$$y_{it} \text{ on } 1, d2_t, \dots, dT_t, w_{it}, d2_t w_{it}, \dots, dT_t w_{it}, (\mathbf{x}_{it} - \bar{\mathbf{x}}_t), w_{it} \cdot (\mathbf{x}_{it} - \bar{\mathbf{x}}_t), \\ (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}), (\bar{\mathbf{z}}_i - \bar{\mathbf{z}}), w_{it} \cdot (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}), w_{it} \cdot (\bar{\mathbf{z}}_i - \bar{\mathbf{z}}), \hat{h}_{it}, w_{it} \cdot \hat{h}_{it}$$

where  $dr_t$  is a time dummy for period  $r$ ,  $\hat{h}_{it} \equiv h(w_{it}, \mathbf{g}_{it}\hat{\boldsymbol{\theta}})$ , and overbars denote sample averages.

Note that in the conditional expectation underlying the CF method,  $E(y_{it}|w_{it}, \mathbf{x}_i, \mathbf{z}_i)$ ,  $\{w_{it} : t = 1, \dots, T\}$  is not guaranteed to be strictly exogenous. Therefore, we cannot use GLS-type methods without making extra assumptions.

**21.15.** a. The Stata output is given below. There are 5,735 observations, and 2,184 received a right-heart catheterization.

```
. tab rhc
```

=1 if received right heart catheteriza tion			
	Freq.	Percent	Cum.
0	3,551	61.92	61.92
1	2,184	38.08	100.00
Total	5,735	100.00	

b. The Stata output, using simple regression and heteroskedasticity-robust standard errors, is given below. According to the estimate, people receiving an RHR have a .051 higher probability of dying. The estimate has a robust  $t$  statistic of 3.95, so it is statistically different from zero (and practically large – in the “wrong” direction).

```
. reg death rhc, robust
```

Linear regression	Number of obs =	5735
	F( 1, 5733) =	15.56

Prob > F = 0.0001  
R-squared = 0.0027  
Root MSE = .47673

death	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
rhc	.0507212	.0128566	3.95	0.000	.0255174	.0759249
_cons	.6296818	.0081049	77.69	0.000	.6137931	.6455705

c. The Stata code and results are given below. We still obtain counterintuitive results:

$\hat{\tau}_{ate,reg} = .078$ , with a bootstrapped standard error of .013, and  $\hat{\tau}_{att,reg} = .066$  (se = .014). Thus, both estimates are statistically different from zero. we can either conclude that the controls we include do not make treatment assignment ignorable or that the RHC actually increases the probability of death.

```
clear all

capture program drop ateboot

program ateboot, eclass

* Estimate logit on treatment and control groups separately
tempvar touse
gen byte `touse' = 1
xi: logit death i.female i.race i.income i.cat1 i.cat2 i.ninsclas age if rhc
predict dlhat
xi: logit death i.female i.race i.income i.cat1 i.cat2 i.ninsclas age if ~rhc
predict d0hat
gen diff = dlhat - d0hat
sum diff
scalar ate = r(mean)
sum diff if rhc
scalar att = r(mean)
matrix b = (ate, att)
matrix colnames b = ate att
ereturn post b , esample(`touse')
ereturn display
drop dlhat d0hat diff _I*
end

use catheter

bootstrap _b[ate] _b[att], reps(1000) seed(123): ateboot

program drop ateboot

do catheter_reg

. use catheter
```

```
. bootstrap _b[ate] _b[att], reps(1000) seed(123): ateboot
(running ateboot on estimation sample)

Bootstrap replications (1000)
-----+--- 1 -----+--- 2 -----+--- 3 -----+--- 4 -----+--- 5
..... 50
..... 1000

Bootstrap results
Number of obs      =      5735
Replications       =      1000

command:  ateboot
   _bs_1:  _b[ate]
   _bs_2:  _b[att]
```

	Observed Coef.	Bootstrap Std. Err.	z	P> z	Normal-based [95% Conf. Interval]	
_bs_1	.0776176	.0129611	5.99	0.000	.0522143	.1030208
_bs_2	.0656444	.01366	4.81	0.000	.0388713	.0924175

```
.
. program drop ateboot
end of do-file
```

d. The average  $\hat{p}_i$  for the treated group is about .445 and it ranges from about .085 to .737. For the control group, the numbers are .341, .044, and .738. Though the mean propensity score is somewhat higher for the treated group, the ranges are comparable. The two histograms show that for both groups the probabilities stay away from the extremes of zero and one, and for every bin representing intervals of the estimated propensity score, there are several individuals in the control and treatment groups. Overlap appears to be good.

```
. use catheter

. xi: logit rhc i.female i.race i.income i.cat1 i.cat2 i.ninsclas age
i.female      _Ifemale_0-1      (naturally coded; _Ifemale_0 omitted)
i.race        _Irace_0-2        (naturally coded; _Irace_0 omitted)
i.income      _Iincome_0-3      (naturally coded; _Iincome_0 omitted)
i.cat1        _Icat1_1-9        (naturally coded; _Icat1_1 omitted)
i.cat2        _Icat2_1-7        (naturally coded; _Icat2_1 omitted)
i.ninsclas    _Ininsclas_1-6    (naturally coded; _Ininsclas_1 omitted)

Logistic regression
Number of obs      =      5735
LR chi2(26)        =      625.48
Prob > chi2        =      0.0000
Pseudo R2          =      0.0821

Log likelihood = -3497.9617
```

rhc	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
_Ifemale_1	.1630495	.0587579	2.77	0.006	.0478861	.2782129
_Irace_1	.0424279	.0827591	0.51	0.608	-.1197771	.2046328
_Irace_2	.0393684	.1361998	0.29	0.773	-.2275784	.3063151
_Iincome_1	.0443619	.0762726	0.58	0.561	-.1051296	.1938535
_Iincome_2	.151793	.0892757	1.70	0.089	-.0231841	.3267701
_Iincome_3	.1579471	.1140752	1.38	0.166	-.0656361	.3815303
_Icat1_2	.498032	.107388	4.64	0.000	.2875553	.7085086
_Icat1_3	-1.226306	.1495545	-8.20	0.000	-1.519428	-.9331849
_Icat1_4	-.7173791	.1714465	-4.18	0.000	-1.053408	-.3813501
_Icat1_5	-1.002513	.1085305	-0.92	0.356	-3.129671	1.124645
_Icat1_6	-.6941957	.1260198	-5.51	0.000	-.94119	-.4472013
_Icat1_7	-1.258815	.4833701	-2.60	0.009	-2.206203	-.3114273
_Icat1_8	-.2076635	.1177652	-1.76	0.078	-.438479	.0231519
_Icat1_9	1.003787	.0768436	13.06	0.000	.8531766	1.154398
_Icat2_2	.9804654	1.465085	0.67	0.503	-1.891048	3.851979
_Icat2_3	-.4141065	.4428411	-0.94	0.350	-1.282059	.4538461
_Icat2_4	-.8864827	.8454718	-1.05	0.294	-2.543577	.7706116
_Icat2_5	-.195389	.3933026	-0.50	0.619	-.966248	.57547
_Icat2_6	1.034498	.369503	2.80	0.005	.3102859	1.758711
_Icat2_7	.1415088	.3649828	0.39	0.698	-.5738443	.8568619
_Ininsclas_2	.1849583	.1216214	1.52	0.128	-.0534153	.4233318
_Ininsclas_3	.1082916	.152243	0.71	0.477	-.1900992	.4066824
_Ininsclas_4	.5216726	.1495659	3.49	0.000	.2285288	.8148164
_Ininsclas_5	.468176	.1122184	4.17	0.000	.248232	.6881199
_Ininsclas_6	.3742273	.1249122	3.00	0.003	.1294038	.6190508
age	.0006419	.002252	0.29	0.776	-.0037719	.0050557
_cons	-1.36677	.3834979	-3.56	0.000	-2.118412	-.6151284

```
. predict phat
(option pr assumed; Pr(rhc))
```

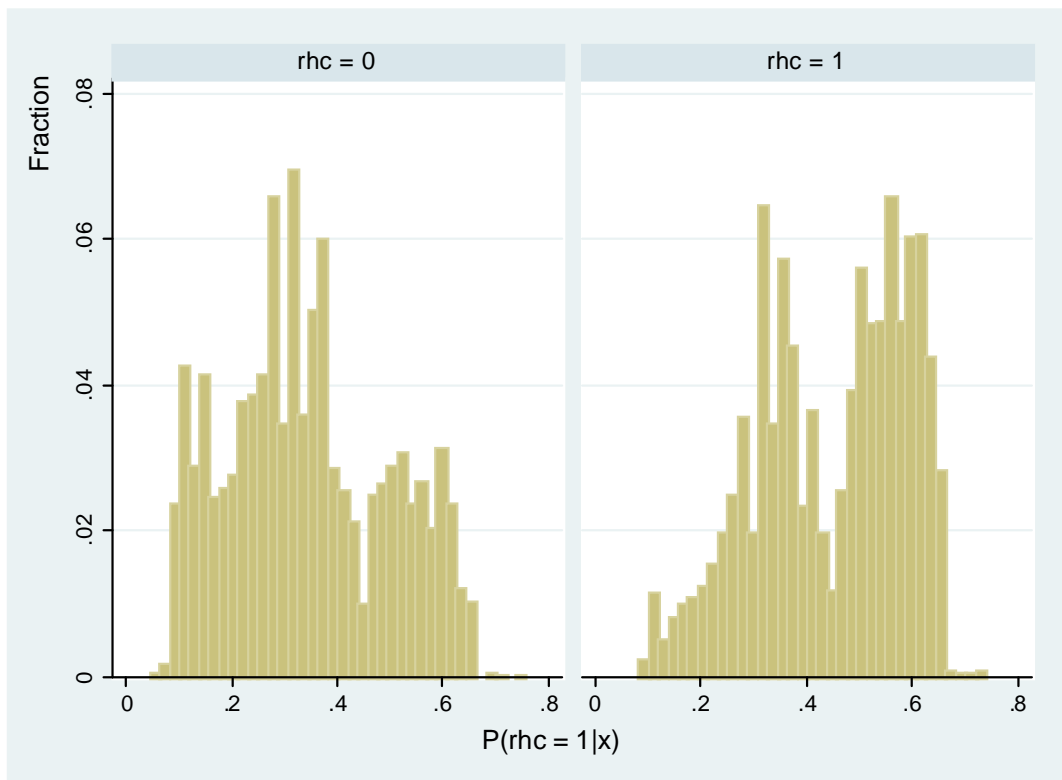
```
. sum phat if rhc
```

Variable	Obs	Mean	Std. Dev.	Min	Max
phat	2184	.4449029	.1421669	.0851523	.7369323

```
. sum phat if ~rhc
```

Variable	Obs	Mean	Std. Dev.	Min	Max
phat	3551	.3414058	.1517016	.0435625	.7379614





e. The estimates using propensity score weighting are very similar to the regression-adjustment estimates using logit models. The PSW estimates are  $\hat{\tau}_{ate,psw} = .072$  (se = .013) and  $\hat{\tau}_{att,psw} = .063$  (se = .014). Unfortunately, out of the 1,000 bootstrap replications that I ran to obtain the standard errors, only 219 produced usable results because the estimated propensity score for at least some of the draws was identically zero or identically one (when the covariates perfectly classify *rhc*). But there is clearly no evidence, based on these and the regression-adjustment estimates, that RHC reduces the probability of death. It appears that RHC is being applied to patients based on variables are not included in the data set that are associated with both mortality and a doctor recommending RHC.

```
. do catheter_psw
clear all
capture program drop ateboot
program ateboot, rclass
* Estimate propensity score
xi: logit rhc i.female i.race i.income i.cat1 i.cat2 i.ninsclas age
predict phat
gen kiate = (rhc - phat)*death/(phat*(1 - phat))
sum kiate
return scalar atew = r(mean)
sum rhc
scalar rho = r(mean)
gen kiatt = (rhc - phat)*death/(1 - phat)
sum kiatt
return scalar attw = r(mean)/rho
drop phat kiate kiatt _I*
end
use catheter
bootstrap r(atew) r(attw), reps(1000) seed(123): ateboot
program drop ateboot
.
. use catheter
.
. bootstrap r(atew) r(attw), reps(1000) seed(123): ateboot
```



probabilities: for the LPM it is about .271 and for the logit it is about .192. In fact, the logit estimate is closer to the true jump in the propensity score at  $x = 5$ , which is about .186. [The treatment probability was generated from the probit model

$P(w = 1|x) = \Phi(.1 + .5 \cdot 1[x \geq 5] + .3 \cdot (x - 5))$ , and so the jump in the probability of treatment at  $x = 5$  is  $\Phi(.6) - \Phi(.1) \approx .186$ .]

```
. reg w x if ~z
```

Source	SS	df	MS	Number of obs = 1000		
Model	17.5177744	1	17.5177744	F( 1, 998)	=	96.61
Residual	180.953226	998	.181315857	Prob > F	=	0.0000
Total	198.471	999	.19866967	R-squared	=	0.0883
				Adj R-squared	=	0.0874
				Root MSE	=	.42581

w	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
x	.0916522	.0093244	9.83	0.000	.0733545	.1099499
_cons	.043984	.0269105	1.63	0.102	-.0088237	.0967917

```
. predict wh0_lpm
(option xb assumed; fitted values)

. gen wh0_lpm_5 = _b[_cons] + _b[x]*5 in 1
(1999 missing values generated)

. reg w x if z
```

Source	SS	df	MS	Number of obs = 1000		
Model	4.4914493	1	4.4914493	F( 1, 998)	=	47.59
Residual	94.1875507	998	.094376303	Prob > F	=	0.0000
Total	98.679	999	.098777778	R-squared	=	0.0455
				Adj R-squared	=	0.0446
				Root MSE	=	.30721

w	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
x	.0464084	.0067272	6.90	0.000	.0332073	.0596095
_cons	.5408787	.0513891	10.53	0.000	.4400356	.6417218

```
. predict wh1_lpm
(option xb assumed; fitted values)

. gen wh1_lpm_5 = _b[_cons] + _b[x]*5 in 1
```

```
(1999 missing values generated)
```

```
. gen jump_lpm = wh1_lpm_5 - wh0_lpm_5  
(1999 missing values generated)
```

```
. gen what = what0 if ~z  
(1000 missing values generated)
```

```
. replace what = what1 if z  
(1000 real changes made)
```

```
. sum jump_lpm
```

Variable	Obs	Mean	Std. Dev.	Min	Max
jump_lpm	1	.2706757	.	.2706757	.2706757

```
. qui logit w x if ~z
```

```
. predict wh0_logit  
(option pr assumed; Pr(w))
```

```
. gen wh0_logit_5 = invlogit(_b[_cons] + _b[x]*5) in 1  
(1999 missing values generated)
```

```
. qui logit w x if z
```

```
. predict wh1_logit  
(option pr assumed; Pr(w))
```

```
. gen wh1_logit_5 = invlogit(_b[_cons] + _b[x]*5) in 1  
(1999 missing values generated)
```

```
. gen jump_logit = wh1_logit_5 - wh0_logit_5  
(1999 missing values generated)
```

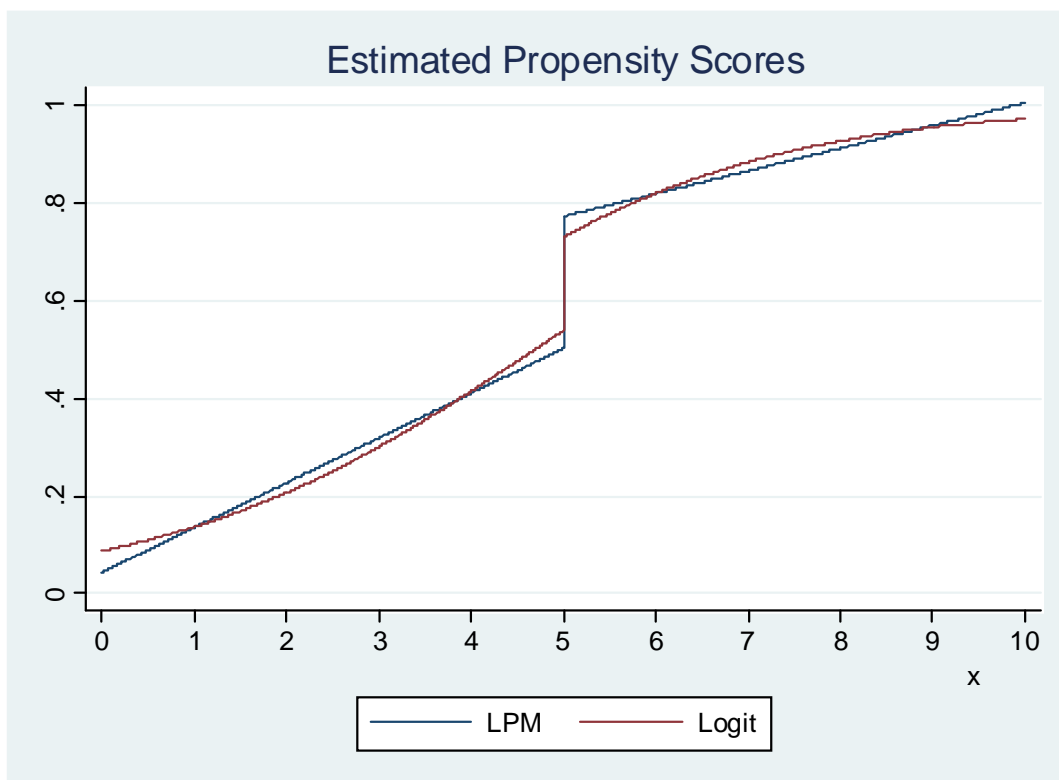
```
. sum jump_logit
```

Variable	Obs	Mean	Std. Dev.	Min	Max
jump_logit	1	.1922199	.	.1922199	.1922199

```
. gen wh_logit = wh0_logit if ~z  
(1000 missing values generated)
```

```
. replace wh_logit = wh1_logit if z  
(1000 real changes made)
```

```
. twoway (line wh_lpm x, sort) (line wh_logit x, sort)
```



c. We already computed the jump for the LPM estimates of the propensity score in part b.

Now we need to estimate the jump in  $E(y|x)$  at  $x = 5$ . Using the linear model, this turns out to be about .531. From equation (21.107) the estimate of  $\tau_c$  is the ratio of the jumps, which is about 1.96. In fact, the true effect is two, so this estimate is very close for this set of data. The data on  $y$  were generated as  $y_i = 1 + 2w_i + x_i/4 + u_i$  where  $u_i$  is independent of  $x_i$  and treatment with a  $\text{Normal}(0, .36)$  distribution.

```
. reg y x if ~z
```

Source	SS	df	MS	Number of obs = 1000		
Model	409.736838	1	409.736838	F( 1, 998)	=	368.55
Residual	1109.52773	998	1.11175123	Prob > F	=	0.0000
				R-squared	=	0.2697
				Adj R-squared	=	0.2690
Total	1519.26457	999	1.52078535	Root MSE	=	1.0544

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
x	.4432576	.0230891	19.20	0.000	.3979488	.4885664
_cons	1.066599	.0666359	16.01	0.000	.9358364	1.197361

```
. gen yh0_5 = _b[_cons] + _b[x]*5 in 1
(1999 missing values generated)
```

```
. reg y x if z
```

Source	SS	df	MS	Number of obs = 1000		
Model	230.759468	1	230.759468	F( 1, 998)	=	310.37
Residual	742.020178	998	.743507193	Prob > F	=	0.0000
				R-squared	=	0.2372
				Adj R-squared	=	0.2365
Total	972.779646	999	.973753399	Root MSE	=	.86227

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
x	.3326468	.0188819	17.62	0.000	.295594	.3696997
_cons	2.151037	.1442388	14.91	0.000	1.86799	2.434083

```
. gen yh1_5 = _b[_cons] + _b[x]*5 in 1
(1999 missing values generated)
```

```
. gen jumpy = yh1_5 - yh0_5
(1999 missing values generated)
```

```
. sum jumpy
```

Variable	Obs	Mean	Std. Dev.	Min	Max
jumpy	1	.531384	.	.531384	.531384

```
. gen ate5 = jumpy/jump_lpm
(1999 missing values generated)
```

```
. sum ate5
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ate5	1	1.963176	.	1.963176	1.963176

d. As is claimed in the text, the IV estimate from (21.108) is the same as the estimate from (21.107), with a very slight rounding error in the sixth digit after the decimal point. The heteroskedasticity-robust standard error is about .205. (The nonrobust standard error is about .197.)

Because the true equation for  $y$  is linear in  $x$  – with an expected jump at  $x = 5$  – the IV estimator (and hence the estimate from part c) is consistent even though  $w_i$  follows a logit model rather than an LPM.

```
. ivreg y x_5 zx_5 (w = z), robust
```

Instrumental variables (2SLS) regression	Number of obs =	2000
	F( 3, 1996) =	3588.42
	Prob > F =	0.0000
	R-squared =	0.8722
	Root MSE =	.5959

y	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
w	1.963177	.2046892	9.59	0.000	1.56175	2.364604
x_5	.263328	.0295197	8.92	0.000	.2054354	.3212206
zx_5	-.0217891	.0214587	-1.02	0.310	-.0638729	.0202947
_cons	2.29689	.1352279	16.99	0.000	2.031688	2.562093

```
Instrumented: w
Instruments: x_5 zx_5 z
```

e. Using only the data with  $3 < x_i < 7$  results in 800 observations, rather than 2,000. The estimate is substantially smaller than two but, more importantly, its standard error has



increased to about .327 from .205. Nevertheless, the 95% confidence interval for  $\tau_c$  easily contains the true value  $\tau_c = 2$ .

```
. ivreg y x_5 zx_5 (w = z) if x > 3 & x < 7, robust
```

```
Instrumental variables (2SLS) regression
```

Number of obs =	800
F( 3, 796) =	351.50
Prob > F	= 0.0000
R-squared	= 0.7662
Root MSE	= .61919

y	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval	
w	1.775465	.3267695	5.43	0.000	1.134033	2.416897
x_5	.3471895	.0726118	4.78	0.000	.2046563	.4897226
zx_5	-.0991082	.0772654	-1.28	0.200	-.2507762	.0525599
_cons	2.442008	.2182725	11.19	0.000	2.01355	2.870466

```
Instrumented: w
Instruments: x_5 zx_5 z
```

## Solutions to Chapter 22 Problems

**22.1. a.** In Stata, there are two possibilities for estimating a lognormal duration model: the `cnreg` command (where we use the log of the duration as a response), and the `streg` command (where we specify “lognormal” as the distribution). The `streg` command is more flexible (and I use it in the next problem), but here I give the `cnreg` output. The value of the log likelihood is  $-1,597.06$ .

```
. use recid
. cnreg ldurat workprg priors tserve d felon alcohol drugs black married educ
    age, censored(cens)
```

Censored-normal regression

Number of obs	=	1445
LR chi2(10)	=	166.74
Prob > chi2	=	0.0000
Pseudo R2	=	0.0496

Log likelihood = -1597.059

ldurat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval
workprg	-.0625715	.1200369	-0.52	0.602	-.2980382 .1728951
priors	-.1372529	.0214587	-6.40	0.000	-.1793466 -.0951592
tserve d	-.0193305	.0029779	-6.49	0.000	-.0251721 -.013489
felon	.4439947	.1450865	3.06	0.002	.1593903 .7285991
alcohol	-.6349092	.1442166	-4.40	0.000	-.9178072 -.3520113
drugs	-.2981602	.1327355	-2.25	0.025	-.5585367 -.0377837
black	-.5427179	.1174428	-4.62	0.000	-.7730958 -.31234
married	.3406837	.1398431	2.44	0.015	.066365 .6150024
educ	.0229196	.0253974	0.90	0.367	-.0269004 .0727395
age	.0039103	.0006062	6.45	0.000	.0027211 .0050994
_cons	4.099386	.347535	11.80	0.000	3.417655 4.781117
/sigma	1.81047	.0623022			1.688257 1.932683

Observation summary:

0	left-censored observations
552	uncensored observations
893	right-censored observations

**b.** I graphed the hazard at the stated values of covariates using the Stata commands below. The estimated hazard initially increases, until about  $t^* = 4.6$ , where it reaches the value of .0116 (roughly). It then falls, until it hits about about .005 at  $t = 81$ . It may make sense that there are startup costs to becoming involved in crime upon release, so that the instantaneous

probability of recidivism initially increases (for about four and one-half months). After that, the hazard falls monotonically, although it does not become zero at the largest observed duration, 81 months.

```
. di = _b[_cons] + _b[felon] + _b[alcohol] + _b[drugs] + _b[priors]* 1.431834
      + _b[tserve]*19.18201 +_b[educ]* 9.702422 +_b[age]* 345.436

4.616118

. clear

. range t .1 81 5000
obs was 0, now 5000

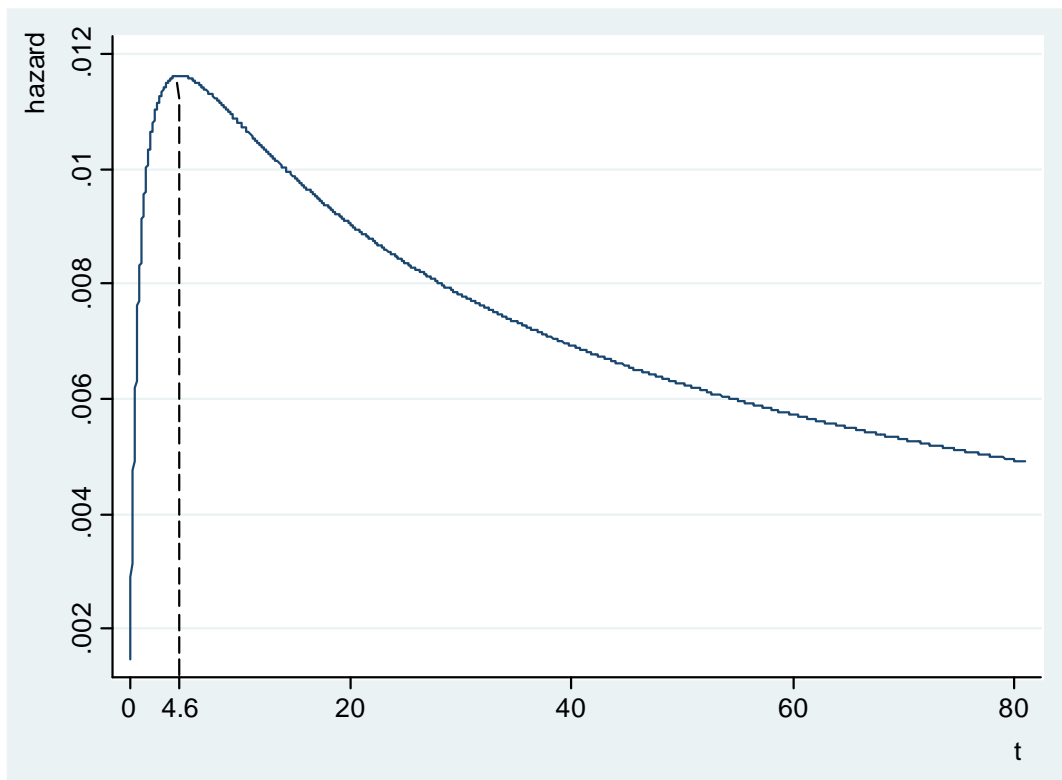
. gen hazard = (normalden((log(t) - 4.62)/1.81)/
                (1 - normal((log(t) - 4.62)/1.81)))/(1.81*t)

egen maxhazard = max(hazard)

. list t hazard if hazard >= maxhazard
```

	t	hazard
277.	4.566573	.011625

```
. twoway (line hazard t)
```



c. Using the only the uncensored in a linear regression analysis provides very different estimates. For example, the *alcohol* and *drugs* coefficients are much smaller in magnitude, with the latter actually changing sign and becoming very insignificant.

```
. reg ldurat workprg priors tserve felon alcohol drugs black married educ age
```

Source	SS	df	MS	Number of obs = 552		
Model	33.7647818	10	3.37647818	F( 10, 541) = 4.13		
Residual	442.796158	541	.818477187	Prob > F = 0.0000		
				R-squared = 0.0709		
				Adj R-squared = 0.0537		
Total	476.56094	551	.864901888	Root MSE = .9047		

ldurat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
workprg	.0923415	.0827407	1.12	0.265	-.0701909	.254874
priors	-.0483627	.0140418	-3.44	0.001	-.0759459	-.0207795
tserve	-.0067761	.001938	-3.50	0.001	-.010583	-.0029692
felon	.1187173	.103206	1.15	0.251	-.0840163	.3214509
alcohol	-.2180496	.0970583	-2.25	0.025	-.408707	-.0273923
drugs	.0177737	.0891098	0.20	0.842	-.1572699	.1928172
black	-.0008505	.0822071	-0.01	0.992	-.1623348	.1606338
married	.2388998	.0987305	2.42	0.016	.0449577	.432842
educ	-.0194548	.0189254	-1.03	0.304	-.0566312	.0177215
age	.0005345	.0004228	1.26	0.207	-.000296	.0013651
_cons	3.001025	.2438418	12.31	0.000	2.522032	3.480017

d. Treating the censored durations as if they are uncensored also gives very different estimates from the censored regression. Again, the estimated *alcohol* and *drug* effects are attenuated toward zero, although not as much as when we drop all of the censored observations. In any case, we should use censored regression analysis.

```
. reg ldurat workprg priors tserve felon alcohol drugs black married educ age
```

Source	SS	df	MS	Number of obs = 1445		
Model	134.350088	10	13.4350088	F( 10, 1434) = 17.49		
Residual	1101.29155	1434	.767985737	Prob > F = 0.0000		
				R-squared = 0.1087		
				Adj R-squared = 0.1025		
Total	1235.64163	1444	.855707503	Root MSE = .87635		

ldurat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
workprg	.008758	.0489457	0.18	0.858	-.0872548	.1047709

priors	-.0590636	.0091717	-6.44	0.000	-.077055	-.0410722
tserved	-.0094002	.0013006	-7.23	0.000	-.0119516	-.0068488
felon	.1785428	.0584077	3.06	0.002	.0639691	.2931165
alcohol	-.2628009	.0598092	-4.39	0.000	-.3801238	-.1454779
drugs	-.0907441	.0549372	-1.65	0.099	-.19851	.0170217
black	-.1791014	.0474354	-3.78	0.000	-.2721516	-.0860511
married	.1344326	.0554341	2.43	0.015	.025692	.2431732
educ	.0053914	.0099256	0.54	0.587	-.0140789	.0248618
age	.0013258	.0002249	5.90	0.000	.0008847	.0017669
_cons	3.569168	.137962	25.87	0.000	3.298539	3.839797

---

**22.2.** a. For this question, I use the `streg` command. The `nohr` option means that the  $\hat{\beta}_j$  that estimate the  $\beta_j$  in equation (22.25) are reported, rather than  $\exp(\hat{\beta}_j)$ . Whether or not a release was “supervised” has no discernible effect on the hazard, whereas, not surprisingly, a history of rules violation while in prison does increase the recidivism hazard.

```
. use recid.dta
. gen failed = ~cens
. stset durat, failure(failed)

      failure event:  failed != 0 & failed < .
obs. time interval:  (0, durat]
exit on or before:   failure
```

---

```
1445  total obs.
    0  exclusions
```

---

```
1445  obs. remaining, representing
  552  failures in single record/single failure data
80013  total analysis time at risk, at risk from t =          0
      earliest observed entry t =          0
      last observed exit t =          81
```

```
. streg super rules workprg priors tserved felon alcohol drugs black married
      educ age, dist(weibull) nohr

      failure _d:  failed
analysis time _t:  durat
```

Weibull regression -- log relative-hazard form

No. of subjects =	1445	Number of obs =	1445
No. of failures =	552		
Time at risk =	80013		
Log likelihood =	-1630.517	LR chi2(12) =	170.51
		Prob > chi2 =	0.0000

---

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
super	-.0078523	.0979703	-0.08	0.936	-.1998705	.184166
rules	.0386963	.0166936	2.32	0.020	.0059774	.0714151
workprg	.1039345	.0914158	1.14	0.256	-.0752371	.2831061
priors	.086349	.0136871	6.31	0.000	.0595227	.1131752
tserved	.0116506	.001933	6.03	0.000	.0078621	.0154392
felon	-.3111997	.1074569	-2.90	0.004	-.5218114	-.1005879
alcohol	.4510744	.1059953	4.26	0.000	.2433275	.6588214
drugs	.2623752	.0982732	2.67	0.008	.0697632	.4549872
black	.458454	.0884443	5.18	0.000	.2851063	.6318016
married	-.1563693	.10941	-1.43	0.153	-.3708088	.0580703
educ	-.0246717	.019442	-1.27	0.204	-.0627772	.0134339
age	-.0035167	.0005306	-6.63	0.000	-.0045567	-.0024767
_cons	-3.466394	.3105515	-11.16	0.000	-4.075064	-2.857724
/ln_p	-.2142514	.038881	-5.51	0.000	-.2904567	-.1380461
p	.8071455	.0313826			.7479219	.8710585
1/p	1.238934	.0481709			1.148028	1.337038

b. The lognormal estimates are given below. The estimated coefficients on *super* and *rules* are consistent with the Weibull results because a decrease in  $x\delta$  shifts up the hazard in the lognormal case.

```
. streg super rules workprg priors tserved felon alcohol drugs black married
educ age, dist(lognormal)

      failure _d:  failed
analysis time _t:  durat

Lognormal regression -- accelerated failure-time form

No. of subjects =          1445                Number of obs   =          1445
No. of failures =           552
Time at risk    =          80013

                                      LR chi2(12)    =          172.52
Log likelihood   =   -1594.1683                Prob > chi2      =          0.0000
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
super	.0328411	.1280452	0.26	0.798	-.2181229	.2838052
rules	-.0644316	.0276338	-2.33	0.020	-.1185929	-.0102703
workprg	-.0883445	.1208173	-0.73	0.465	-.325142	.148453
priors	-.1341294	.0215358	-6.23	0.000	-.1763388	-.0919199
tserved	-.0156015	.0033872	-4.61	0.000	-.0222403	-.0089628
felon	.4345115	.1460924	2.97	0.003	.1481757	.7208473
alcohol	-.6415683	.1439736	-4.46	0.000	-.9237515	-.3593851
drugs	-.2785464	.1326321	-2.10	0.036	-.5385007	-.0185922
black	-.549173	.1172236	-4.68	0.000	-.7789271	-.3194189
married	.3308228	.1399244	2.36	0.018	.056576	.6050695
educ	.0234116	.0253302	0.92	0.355	-.0262347	.073058
age	.0036626	.0006117	5.99	0.000	.0024637	.0048614

_cons		4.173851	.3580214	11.66	0.000	3.472142	4.87556
/ln_sig		.5904906	.034399	17.17	0.000	.5230698	.6579114
sigma		1.804874	.0620858			1.687199	1.930755

c. The coefficient from the lognormal model directly estimates the proportional effect of rules violations on the duration. So, one more rules violation reduces the estimated expected duration by about 6.4%. To obtain the comparable Weibull estimate, we need

$-\hat{\beta}_{rules}/\hat{\alpha} = -.0387/.807 \approx -.048$ , or about a 4.8% reduction for each rules violation – a bit smaller than the lognormal estimate.

**22.3.** a. If all durations in the sample are censored,  $d_i = 0$  for all  $i$ , and so the log-likelihood is  $\sum_{i=1}^N \log[1 - F(t_i|\mathbf{x}_i;\theta)] = \sum_{i=1}^N \log[1 - F(c_i|\mathbf{x}_i;\theta)]$ .

b. For the Weibull case,  $F(t|\mathbf{x}_i;\theta) = 1 - \exp[-\exp(\mathbf{x}_i\boldsymbol{\beta})t^\alpha]$ , and so the log-likelihood is  $-\sum_{i=1}^N \exp(\mathbf{x}_i\boldsymbol{\beta})c_i^\alpha$ .

c. Without covariates, the Weibull log-likelihood with all observations censored is  $-\exp(\beta) \sum_{i=1}^N c_i^\alpha$ . Because  $c_i > 0$ , we can choose any  $\alpha > 0$  so that  $\sum_{i=1}^N c_i^\alpha > 0$ . But then, for any  $\alpha > 0$ , the log-likelihood is maximized by minimizing  $\exp(\beta)$  across  $\beta$ . But as  $\beta \rightarrow -\infty$ ,  $\exp(\beta) \rightarrow 0$ . Therefore, plugging any value  $\alpha$  into the log-likelihood will lead to  $\beta$  getting more and more negative without bound. So no two real numbers for  $\alpha$  and  $\beta$  maximize the log likelihood.

d. It is not possible to estimate duration models from flow data when all durations are right censored.

e. To have all durations censored in a large sample we would have to have  $P(t_i^* > c_i)$  very close to one. But if  $P(t_i^* < t) > 0$  for all  $t > 0$  and  $c_i > b > 0$ ,

$$P(t_i^* > c_i) \leq P(t_i^* > b) = 1 - P(t_i^* \leq b),$$



and  $P(t_i^* \leq b) > 0$ . So with large samples we should not expect to find that all durations have been censored.

**22.4. a.** The binary response  $d_i$  is equal to one if the observation is uncensored. Because  $t_i^*$  is independent of  $c_i$  conditional on  $\mathbf{x}_i$ ,

$$P(d_i = 1 | \mathbf{x}_i, c_i) = P(t_i^* \leq c_i | \mathbf{x}_i) = F(c_i | \mathbf{x}_i; \boldsymbol{\theta}).$$

Therefore, the log-likelihood is

$$\sum_{i=1}^N \{d_i \log[F(c_i | \mathbf{x}_i; \boldsymbol{\theta})] + (1 - d_i) \log[1 - F(c_i | \mathbf{x}_i; \boldsymbol{\theta})]\},$$

which is just of the usual binary response form.

**b.** When the distribution is Weibull and  $\mathbf{x}_i = 1$ , we have (from Problem 22.3c),

$F(c_i | \boldsymbol{\theta}) = 1 - \exp[-\exp(\beta)c_i^\alpha]$ , and so the log-likelihood is

$$\mathcal{L}(\alpha, \beta) = \sum_{i=1}^N \{d_i \log[1 - \exp[-\exp(\beta)c_i^\alpha]] + (1 - d_i) \log(\exp[-\exp(\beta)c_i^\alpha])\}.$$

If  $c_i = c > 0$  for all  $i$ , we have

$$\mathcal{L}(\alpha, \beta) = \sum_{i=1}^N \{d_i \log[1 - \exp[-\exp(\beta)c^\alpha]] + (1 - d_i) \log(\exp[-\exp(\beta)c^\alpha])\}.$$

If we define  $\rho \equiv \exp[-\exp(\beta)c^\alpha]$  then  $0 < \rho < 1$ , and the log-likelihood can be written as

$$\sum_{i=1}^N [d_i \log(\rho) + (1 - d_i) \log(1 - \rho)].$$

In other words, the log-likelihood is the same for all combinations of  $\alpha$  and  $\beta$  that give the same value of  $\rho$ . While we can consistently estimate  $\rho$  – the fraction of uncensored observations is the maximum likelihood estimator – we cannot recover estimates of  $\alpha$  and  $\beta$ . In

other words,  $\alpha$  and  $\beta$  are not identified.

c. In the log-normal case, the log-likelihood is based on

$P[\log(t_i^*) \leq \log(c_i) | \mathbf{x}_i] = \phi[(1/\sigma)\log(c_i) - (1/\sigma)\mathbf{x}_i\boldsymbol{\beta}]$ , because

$\log(t_i^*) | (\mathbf{x}_i, c_i) \sim \text{Normal}(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$ . The log-likelihood is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \sigma^2) &= \sum_{i=1}^N \ell_i(\boldsymbol{\beta}, \sigma^2) \\ &= \sum_{i=1}^N \{d_i \log(\phi[(1/\sigma)\log(c_i) - (1/\sigma)\mathbf{x}_i\boldsymbol{\beta}]) + (1 - d_i) \log(1 - \phi[(1/\sigma)\log(c_i) - (1/\sigma)\mathbf{x}_i\boldsymbol{\beta}])\}.\end{aligned}$$

Even though  $x_{i1} \equiv 1$ ,  $1/\sigma$  is identified as the coefficient on  $\log(c_i)$ , provided  $c_i$  varies. Then, of course, we can identify  $\boldsymbol{\beta}$  because  $\boldsymbol{\beta}/\sigma$  is generally identified from the probit log-likelihood.

[We would need to assume the usual condition that  $\text{rank } E(\mathbf{x}_i'\mathbf{x}_i) = K$ .] If  $c_i = c$  for all  $i$  then the intercept effectively becomes  $(1/\sigma)\log(c) - \beta_1/\sigma$ ; along with  $\beta_j/\sigma$ ,  $j = 2, \dots, K$ , these are the only parameters we can identify. We cannot separately identify  $\boldsymbol{\beta}$  or  $\sigma$ . This is the same situation we faced in Section 19.2.1.

**22.5.** a. We have

$$\begin{aligned}P(t_i^* \leq t | \mathbf{x}_i, a_i, c_i, s_i = 1) &= P(t_i^* \leq t | \mathbf{x}_i, t_i^* > b - a_i) \\ &= P(t_i^* \leq t, t_i^* > b - a_i | \mathbf{x}_i) / P(t_i^* > b - a_i | \mathbf{x}_i) = P(b - a_i < t_i^* \leq t | \mathbf{x}_i) / P(t_i^* > b - a_i | \mathbf{x}_i) \\ &= [F(t | \mathbf{x}_i) - F(b - a_i | \mathbf{x}_i)] / [1 - F(b - a_i | \mathbf{x}_i)]\end{aligned}$$

where we use the fact that  $t < b - a_i$ .

b. The derivative of the cdf in part a, with respect to  $t$ , is simply  $f(t | \mathbf{x}_i) / [1 - F(b - a_i | \mathbf{x}_i)]$ .

c. Because  $s_i = 1[t_i^* > b - a_i]$  and  $t_i = c_i$  if and only if  $t_i^* \geq c_i$ , we have

$$\begin{aligned}P(t_i = c_i | \mathbf{x}_i, a_i, c_i, s_i = 1) &= P(t_i^* \geq c_i | \mathbf{x}_i, t_i^* > b - a_i) \\ &= P(t_i^* \geq c_i | \mathbf{x}_i) / P(t_i^* \geq b - a_i | \mathbf{x}_i) \\ &= [1 - F(c_i | \mathbf{x}_i)] / [1 - F(b - a_i | \mathbf{x}_i)]\end{aligned}$$

where the third equality follows because  $c_i > b - a_i$ .

d. Parts b and c show that the density of the censored variable,  $t_i$ , conditional on  $(\mathbf{x}_i, a_i, c_i, s_i = 1)$ , can be written as

$$\frac{[f(t|\mathbf{x}_i)]^{1[t < c_i]} [1 - F(c_i|\mathbf{x}_i)]^{1[t = c_i]}}{[1 - F(b - a_i|\mathbf{x}_i)]}.$$

Showing the dependence on the parameters, plugging in  $t_i$ , noting that  $d_i = 1[t_i < c_i]$ , and taking the log gives

$$d_i \log[f(t|\mathbf{x}_i; \boldsymbol{\theta})] + (1 - d_i) \log[1 - F(c_i|\mathbf{x}_i; \boldsymbol{\theta})] - \log[1 - F(b - a_i|\mathbf{x}_i; \boldsymbol{\theta})].$$

Summing across all  $N$  observations gives equation (22.30).

**22.6.** In what follows, we initially suppress dependence on the parameters.

a. Because  $s_i = 1[t_i^* \geq b - a_i]$ ,

$$\begin{aligned} P(a_i \leq a, s_i = 1|\mathbf{x}_i) &= P(a_i \leq a, t_i^* \geq b - a_i|\mathbf{x}_i) \\ &= \int_0^a \int_{b-u}^{\infty} q(u, v|\mathbf{x}_i) dv du, \end{aligned}$$

where  $q(\cdot, \cdot|\mathbf{x}_i)$  denotes the joint density of  $(a_i, t_i^*)$  given  $\mathbf{x}_i$ . By conditional independence,  $q(a, t|\mathbf{x}_i) = k(a|\mathbf{x}_i)f(t|\mathbf{x}_i)$ , and so the double integral is

$$\int_0^a \left( \int_{b-u}^{\infty} f(v|\mathbf{x}_i) dv \right) k(u|\mathbf{x}_i) du = \int_0^a [1 - F(b - u|\mathbf{x}_i)] k(u|\mathbf{x}_i) du$$

because  $\int_{b-u}^{\infty} f(v|\mathbf{x}_i) dv = [1 - F(b - u|\mathbf{x}_i)]$ .

b. From the hint, we first compute  $E(s_i = 1|a_i, \mathbf{x}_i) = P(t_i^* \geq b - a_i|\mathbf{x}_i) = 1 - F(b - a_i|\mathbf{x}_i)$ .

Next, we compute the expected value of this with respect to the distribution  $D(a_i|\mathbf{x}_i)$ , which is simply equation (22.32).

c. The conditional cdf is obtained by dividing the answer from part a by the answer from part b. The density is just the derivative of the resulting expression with respect to  $a$ ; by the

fundamental theorem of calculus, the derivative is (22.31).

d. When  $b = 1$  and  $k(a|\mathbf{x}_i) = 1$ , all  $0 < a < 1$ , the numerator of (22.31) is just  $1 - F(1 - a|\mathbf{x}_i)$ . The denominator is simply  $\int_0^1 [1 - F(1 - u|\mathbf{x}_i)] du$ .

e. In the Weibull case,  $1 - F(1 - a|\mathbf{x}_i) = \exp[-\exp(\mathbf{x}_i\beta)(1 - a)^\alpha]$  and the denominator is  $\int_0^1 \exp[-\exp(\mathbf{x}_i\beta)(1 - u)^\alpha] du$ . This integral cannot be solved in closed form unless  $\alpha = 1$ .

**22.7.** a. For notational simplicity, the parameters in the densities are suppressed. Then, by equation (22.22) and  $D(a_i|c_i, \mathbf{x}_i) = D(a_i|\mathbf{x}_i)$ , the density of  $(a_i, t_i^*)$  given  $(c_i, \mathbf{x}_i)$  does not depend on  $c_i$  and is given by  $k(a|\mathbf{x}_i)f(t|\mathbf{x}_i)$  for  $0 < a < b$  and  $0 < t < \infty$ . This is also the conditional density of  $(a_i, t_i)$  given  $(c_i, \mathbf{x}_i)$  for  $t < c_i$ , that is, for values of  $t$  corresponding to being uncensored. For  $t = c_i$ , the density is  $k(a|\mathbf{x}_i)[1 - F(c_i|\mathbf{x}_i)]$  by the usual right censoring argument. Now, the probability of observing the random draw  $(a_i, c_i, \mathbf{x}_i, t_i)$ , conditional on  $\mathbf{x}_i$ , is  $P(t_i^* \geq b - a_i, \mathbf{x}_i)$ , which is exactly (22.32). From the standard result for densities for truncated distributions, the density of  $(a_i, t_i)$  given  $(c_i, d_i, \mathbf{x}_i)$  and  $s_i = 1$  is

$$k(a|\mathbf{x}_i)[f(t|\mathbf{x}_i)]^{d_i}[1 - F(c_i|\mathbf{x}_i)]^{(1-d_i)}/P(s_i = 1|\mathbf{x}_i),$$

for all combinations  $(a, t)$  such as that  $s_i = 1$ . Putting in the observed data, inserting the parameters, and taking the log gives (22.56).

b. We have the usual trade-off between robustness and efficiency. Using the log likelihood (20.56) results in more efficient estimators provided we have the two densities correctly specified; (20.30) requires us to only specify  $f(\cdot|\mathbf{x}_i)$ .

**22.8.** a. Again I suppress the parameters in the densities. Let  $z_i \equiv t_i^* - b$ , so that, by assumption,  $a_i$  and  $z_i$  are independent conditional on  $\mathbf{x}_i$ . The conditional density of  $z_i$  is simply  $g(z|\mathbf{x}_i) \equiv f(z + b|\mathbf{x}_i)$ ,  $z > -b$ . By the usual convolution formula for the density of a sum of

independent random variables,

$$h(r|\mathbf{x}_i) = \int_0^b k(u|\mathbf{x}_i)g(r-u|\mathbf{x}_i)du = \int_0^b k(u|\mathbf{x}_i)f(r+b-u|\mathbf{x}_i)du,$$

where  $f(r+b-u|\mathbf{x}_i) = 0$  if  $r+b-u \leq 0$ . When  $r > 0$ ,  $r+b-u > 0$  for all  $0 \leq u \leq b$ , and so we need not modify the formula.

b. As usual for right censoring,  $P(r_i = q|\mathbf{x}_i) = P(r_i^* > q|\mathbf{x}_i) = 1 - H(q|\mathbf{x}_i)$ .

c. The argument is now essentially the same as the first stock sampling argument treated in Section 22.3.3, except that we cannot condition on  $a_i$ . Instead, for  $0 < r < q$ ,  $P(r_i \leq r|\mathbf{x}_i, s_i = 1) = [H(r|\mathbf{x}_i) - H(0|\mathbf{x}_i)]/P(s_i = 1|\mathbf{x}_i)$  – and so the density for  $0 < r < q$  is  $h(r|\mathbf{x}_i)/P(s_i = 1|\mathbf{x}_i)$  – and  $P(r_i = q|\mathbf{x}_i, s_i = 1) = [1 - H(q|\mathbf{x}_i)]/P(s_i = 1|\mathbf{x}_i)$ , where  $P(s_i = 1|\mathbf{x}_i)$  is given by (22.32).

d. With  $b = 1$  and a uniform distribution for  $a_i$ , the log-likelihood reduces to

$$d_i \log[h(r_i|\mathbf{x}_i; \boldsymbol{\theta}, \boldsymbol{\eta})] + (1 - d_i) \log[1 - H(r_i|\mathbf{x}_i; \boldsymbol{\theta}, \boldsymbol{\eta})] - \log \left\{ \int_0^1 [1 - F(1-u|\mathbf{x}_i; \boldsymbol{\theta})] du \right\}.$$

**22.9.** a. Let  $\omega$  be the value for type B people. Then we must have

$$\rho\eta + (1 - \rho)\omega = 1$$

or  $\omega = (1 - \rho\eta)/(1 - \rho)$ .

b. The cdf conditional on  $(\mathbf{x}, v)$  is  $F(t|\mathbf{x}, v; \alpha, \boldsymbol{\beta}) = 1 - \exp[-v \exp(\mathbf{x}\boldsymbol{\beta})t^\alpha]$ . Therefore, the cdf conditional on  $\mathbf{x}$  is obtained by averaging out  $v$ :

$$G(t|\mathbf{x}; \alpha, \boldsymbol{\beta}, \eta, \rho) = \rho \{1 - \exp[-\eta \exp(\mathbf{x}\boldsymbol{\beta})t^\alpha]\} + (1 - \rho) \{1 - \exp[-((1 - \rho\eta)/(1 - \rho)) \exp(\mathbf{x}\boldsymbol{\beta})t^\alpha]\}.$$

c. The density function is just the derivative with respect to  $t$ :

$$g(t|\mathbf{x}; \alpha, \boldsymbol{\beta}, \eta, \rho) = \rho\eta \exp(\mathbf{x}\boldsymbol{\beta})t^{\alpha-1} \exp[-\eta \exp(\mathbf{x}\boldsymbol{\beta})t^\alpha] \\ + (1 - \rho\eta) \exp(\mathbf{x}\boldsymbol{\beta})\alpha t^{\alpha-1} \exp[-((1 - \rho\eta)/(1 - \rho)) \exp(\mathbf{x}\boldsymbol{\beta})t^\alpha]\}$$

If none of the durations are censored, the log-likelihood for each observation  $i$  is obtained by taking the log of this density in plugging in  $(\mathbf{x}_i, t_i)$ . If we have right censored data with censoring values  $c_i$ , the log likelihood takes on the usual form:

$$d_i \log[g(t_i|\mathbf{x}_i; \alpha, \boldsymbol{\beta}, \eta, \rho)] + (1 - d_i) \log[G(t_i|\mathbf{x}_i; \alpha, \boldsymbol{\beta}, \eta, \rho)]$$

where  $d_i$  is the dummy variable equal to one if the observation is not censored. The log likelihood for the entire sample is a very smooth function of all parameters. As a computational device, it might be better to replace  $\eta$  with, say,  $\exp(\xi)/[1 + \exp(\xi)]$  for  $-\infty < \xi < \infty$  and similarly for  $\rho$ .

**22.10.** a. If  $P(T > a_{m-1}) = 0$  then  $P(T > a_m) = 0$  because  $a_m > a_{m-1}$  in which case the equality is trivial. So assume that  $P(T > a_{m-1}) > 0$ . Then, by definition of conditional probability,

$$P(T > a_m | T > a_{m-1}) = P(T > a_m, T > a_{m-1}) / P(T > a_{m-1}) \\ = P(T > a_m) / P(T > a_{m-1})$$

since the events  $\{T > a_m, T > a_{m-1}\}$  and  $\{T > a_m\}$  are identical when  $a_m > a_{m-1}$ .

Rearranging the equality gives the result.

b. We can use induction to obtain an algebraically simple proof. First, equation (22.48) holds trivially when  $m = 1$ :  $P(T > a_1) = P(T > a_1 | T > 0)$  because  $P(T > 0) = 1$ . Now assume that (22.48) holds for any  $m \geq 1$ . We show it holds for  $m + 1$ . By part a,

$$P(T > a_{m+1}) = P(T > a_{m+1} | T > a_m) P(T > a_m) \\ = P(T > a_{m+1} | T > a_m) \prod_{r=1}^m P(T > a_r | T > a_{r-1})$$

because  $P(T > a_m) = \prod_{r=1}^m P(T > a_r | T > a_{r-1})$  by the induction hypothesis. It follows that

$$P(T > a_{m+1}) = \prod_{r=1}^{m+1} P(T > a_r | T > a_{r-1})$$

and this completes the proof.

**22.11.** a. The estimates from the log-logistic model with gamma-distributed hazard are given below. For comparison purposes, the Stata output used to produce Table 22.2 follows.

The log likelihood for the log-logistic model is  $-1,587.92$  and that for the Weibull model (both with gamma heterogeneity) is  $-1,584.92$ . The Weibull model fits somewhat better. (A Vuong model selection statistic could be computed to see if the fit is statistically better.)

```
. streg workprg priors tserverd felon alcohol drugs black married educ age,
      d(loglogistic) fr(gamma)
```

```
      failure _d: failed
      analysis time _t: durat
```

Loglogistic regression -- accelerated failure-time form  
Gamma frailty

No. of subjects =	1445	Number of obs =	1445
No. of failures =	552		
Time at risk =	80013		
		LR chi2(10) =	132.67
Log likelihood =	-1587.9185	Prob > chi2 =	0.0000

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
workprg	.0098501	.11888	0.08	0.934	-.2231504	.2428505
priors	-.1488187	.0231592	-6.43	0.000	-.19421	-.1034275
tserverd	-.0190707	.0036147	-5.28	0.000	-.0261553	-.0119861
felon	.4219072	.1502377	2.81	0.005	.1274467	.7163677
alcohol	-.6597875	.147537	-4.47	0.000	-.9489547	-.3706203
drugs	-.2156168	.131043	-1.65	0.100	-.4724563	.0412227
black	-.4355534	.1209008	-3.60	0.000	-.6725147	-.1985921
married	.4345344	.1353415	3.21	0.001	.16927	.6997988
educ	.0243336	.0265813	0.92	0.360	-.0277648	.0764321
age	.0032531	.000633	5.14	0.000	.0020124	.0044937
_cons	3.34599	.3570848	9.37	0.000	2.646117	4.045864
/ln_gam	-.3648587	.0716481	-5.09	0.000	-.5052864	-.2244309
/ln_the	.8176437	.1998151	4.09	0.000	.4260133	1.209274
gamma	.6942948	.0497449			.6033327	.7989708

```

theta | 2.265156 .4526125 1.531141 3.351051
-----+-----
Likelihood-ratio test of theta=0: chibar2(01) = 46.05 Prob>=chibar2 = 0.000

```

```

. streg workprg priors terved felon alcohol drugs black married educ age,
  d(weibull) fr(gamma) nohr

```

```

      failure _d: nocens
analysis time _t: durat

```

```

Weibull regression -- log relative-hazard form
                    Gamma frailty

```

```

No. of subjects = 1445          Number of obs   = 1445
No. of failures = 552
Time at risk    = 80013
Log likelihood   = -1584.9172
LR chi2(10)     = 143.82
Prob > chi2     = 0.0000

```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval	
workprg	.0073827	.2038775	0.04	0.971	-.3922099	.4069753
priors	.2431142	.0421543	5.77	0.000	.1604933	.3257352
terved	.0349363	.0070177	4.98	0.000	.0211818	.0486908
felon	-.7909533	.2666084	-2.97	0.003	-1.313496	-.2684104
alcohol	1.173558	.2805222	4.18	0.000	.6237451	1.723372
drugs	.2847665	.2233072	1.28	0.202	-.1529074	.7224405
black	.7715762	.2038289	3.79	0.000	.372079	1.171073
married	-.8057042	.2578214	-3.13	0.002	-1.311025	-.3003834
educ	-.0271193	.044901	-0.60	0.546	-.1151237	.060885
age	-.0052162	.0009974	-5.23	0.000	-.0071711	-.0032613
_cons	-5.393658	.720245	-7.49	0.000	-6.805312	-3.982004
/ln_p	.5352553	.0951206	5.63	0.000	.3488225	.7216882
/ln_the	1.790243	.1788498	10.01	0.000	1.439703	2.140782
p	1.707884	.1624549			1.417398	2.057904
1/p	.5855198	.055695			.4859312	.7055184
theta	5.990906	1.071472			4.219445	8.506084

```

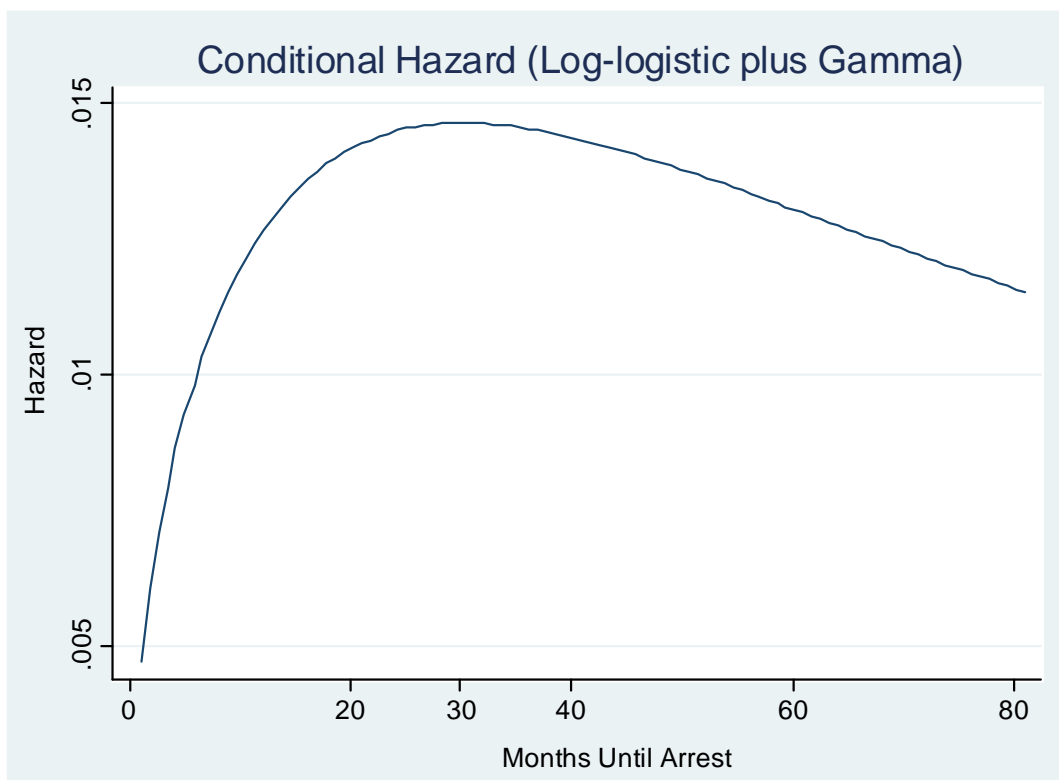
Likelihood-ratio test of theta=0: chibar2(01) = 96.23 Prob>=chibar2 = 0.000

```

b. The conditional hazard is plotted below. Its shape is very different from the Weibull case, which is always upward sloping. (See Figure 22.3.) The hazard for the log-logistic model with gamma heterogeneity has its maximum value near 30 weeks, and it falls off gradually after that.

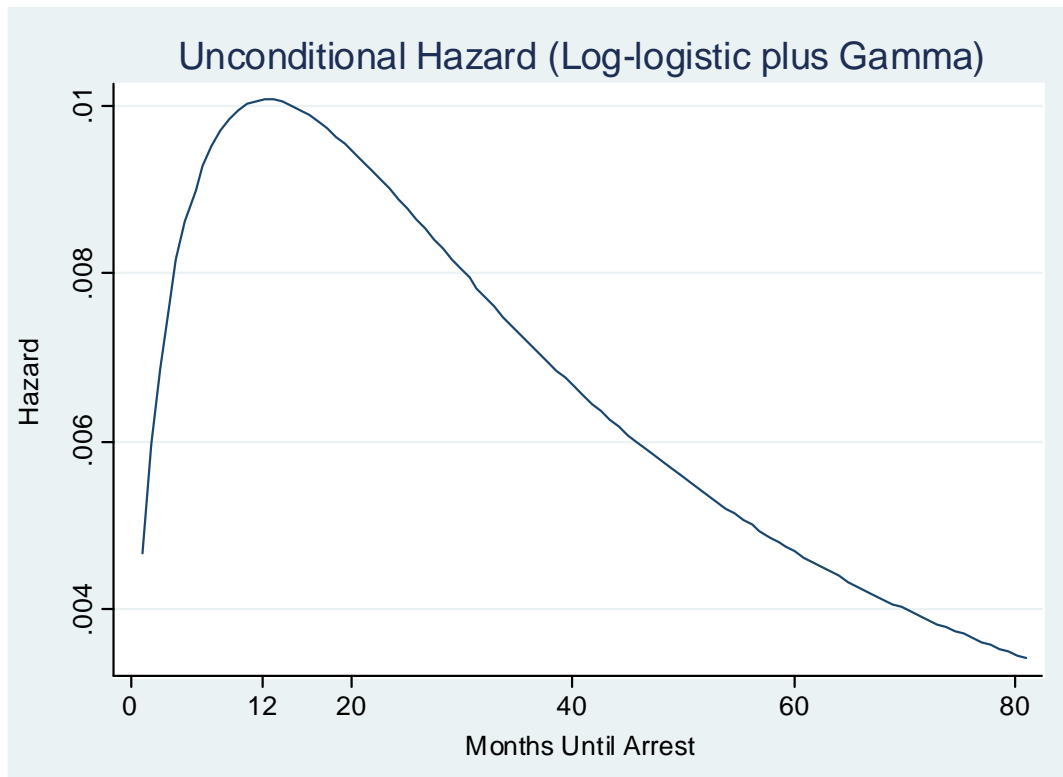


```
. stcurve, haz alpha1
```

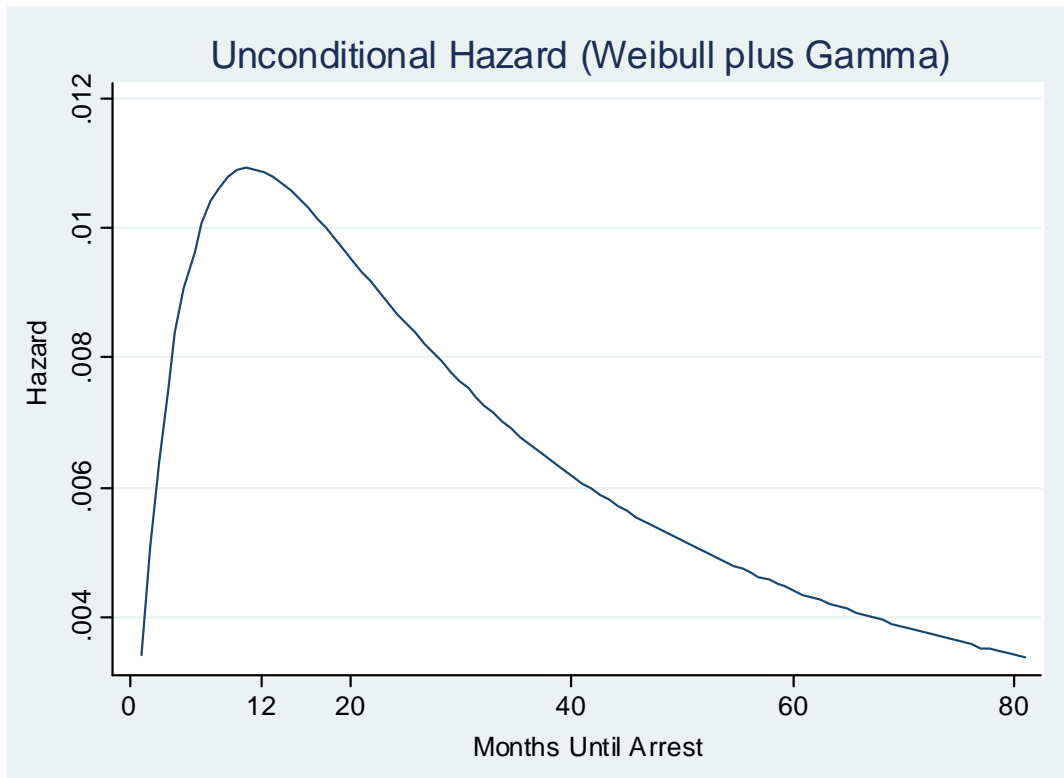


c. The unconditional hazard is plotted below. While it also has a hump shape, the maximum value of the unconditional hazard is around 12 – which is well below that for the conditional hazard.

```
. stcurve, haz  
(option unconditional assumed)
```



There is a final point worth making about this example that builds on the discussion in Section 22.3.4. Below is the graph in Figure 22.4, reproduced for convenience. It is the unconditional hazard for the Weibull model with gamma heterogeneity. While it differs somewhat from the unconditional hazard for the log-logistic model with gamma heterogeneity, it is practically very similar. Both hazards have sharp increases until about 12 months, and then fall off to zero more gradually. In other words, when we study features of the distribution  $D(t_i^*|\mathbf{x}_i)$  – which is what we can generally hope to identify – we get pretty similar findings. Yet the hazards based on  $D(t_i^*|\mathbf{x}_i, \nu_i)$  are very different. Recall that the hazard for  $D(t_i^*|\mathbf{x}_i, \nu_i)$  in the Weibull case is of the proportional hazard form while that for the log-logistic is not (which is apparent when studying the plots of the conditional hazards). Given that the models fit the data roughly equally well, and that they give similar shapes for the hazard of  $D(t_i^*|\mathbf{x}_i)$ , it seems that trying to decide whether the conditional hazard has a hump shape, as in part b, or is strictly increasing, as in Figure 22.3, seems pretty hopeless.



Incidentally, if we use a gamma distribution for  $D(t_i^*|\mathbf{x}_i, v_i)$  with gamma heterogeneity, we obtain an unconditional hazard very similar to the Weibull and log-logistic cases. The conditional hazard is similar to the log-logistic case, and the log likelihood value is  $-1,585.09$ .