CHAPTER 15

SOLUTIONS TO PROBLEMS

15.1 (i) It has been fairly well established that socioeconomic status affects student performance. The error term *u* contains, among other things, family income, which has a positive effect on *GPA* and is also very likely to be correlated with PC ownership.

(ii) Families with higher incomes can afford to buy computers for their children. Therefore, family income certainly satisfies the second requirement for an instrumental variable: it is correlated with the endogenous explanatory variable [see (15.5) with $x = PC$ and $z = faminc$]. But as we suggested in part (i), *faminc* has a positive affect on *GPA*, so the first requirement for a good IV, (15.4), fails for *faminc*. If we had *faminc* we would include it as an explanatory variable in the equation; if it is the only important omitted variable correlated with *PC*, we could then estimate the expanded equation by OLS.

(iii) This is a natural experiment that affects whether or not some students own computers. Some students who buy computers when given the grant would not have without the grant. (Students who did not receive the grants might still own computers.) Define a dummy variable, *grant*, equal to one if the student received a grant, and zero otherwise. Then, if *grant* was randomly assigned, it is uncorrelated with *u*. In particular, it is uncorrelated with family income and other socioeconomic factors in *u*. Further, *grant* should be correlated with *PC*: the probability of owning a PC should be significantly higher for student receiving grants. Incidentally, if the university gave grant priority to low-income students, *grant* would be negatively correlated with *u*, and IV would be inconsistent.

15.3 It is easy to use (15.10) but we droped $\bar{z}$. Remember, this is allowed because $\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x}) = \sum_{i=1}^{n} z_i(x_i - \bar{x})$ and similarly when we replace x with y. So the numerator in the formula for $\hat{\beta}_1$ is

$$\sum_{i=1}^{n} z_i(y_i - \bar{y}) = \sum_{i=1}^{n} z_i y_i - \left(\sum_{i=1}^{n} z_i\right)\bar{y} = n_1\bar{y}_1 - n_1\bar{y},$$

where $n_1 = \sum_{i=1}^{n} z_i$ is the number of observations with $z_i = 1$, and we have used the fact that $\left(\sum_{i=1}^{n} z_i y_i\right)/n_1 = \bar{y}_1$, the average of the y_i over the i with $z_i = 1$. So far, we have shown that the numerator in $\hat{\beta}_1$ is $n_1(\bar{y}_1 - \bar{y})$. Next, write $\bar{y}$ as a weighted average of the averages over the two subgroups:

$$\bar{y} = (n_0/n)\,\bar{y}_0 + (n_1/n)\,\bar{y}_1,$$

where $n_0 = n - n_1$. Therefore,

$$\bar{y}_1 - \bar{y} = [(n - n_1)/n]\,\bar{y}_1 - (n_0/n)\,\bar{y}_0 = (n_0/n)\,(\bar{y}_1 - \bar{y}_0).$$

Therefore, the numerator of $\hat{\beta}_1$ can be written as

$$(n_0 n_1/n)(\bar{y}_1 - \bar{y}_0).$$

By simply replacing y with x, the denominator in $\hat{\beta}_1$ can be expressed as $(n_0 n_1/n)(\bar{x}_1 - \bar{x}_0)$. When we take the ratio of these, the terms involving n_0, n_1, and n cancel, leaving

$$\hat{\beta}_1 = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0).$$

15.5 (i) From equation (15.19) with $\sigma_u = \sigma_x$, plim $\hat{\beta}_1 = \beta_1 + (.1/.2) = \beta_1 + .5$, where $\hat{\beta}_1$ is the IV estimator. So the asymptotic bias is .5.

(ii) From equation (15.20) with $\sigma_u = \sigma_x$, plim $\tilde{\beta}_1 = \beta_1 + \text{Corr}(x,u)$, where $\tilde{\beta}_1$ is the OLS estimator. So we would have to have $\text{Corr}(x,u) > .5$ before the asymptotic bias in OLS exceeds that of IV. This is a simple illustration of how a seemingly small correlation (.1 in this case) between the IV (z) and error (u) can still result in IV being more biased than OLS if the correlation between z and x is weak (.2).

15.7 (i) Even at a given income level, some students are more motivated and more able than others, and their families are more supportive (say, in terms of providing transportation) and enthusiastic about education. Therefore, there is likely to be a self-selection problem: students that would do better anyway are also more likely to attend a choice school.

(ii) Assuming we have the functional form for *faminc* correct, the answer is yes. Since u_1 does not contain income, random assignment of grants within income class means that grant designation is not correlated with unobservables such as student ability, motivation, and family support.

(iii) The reduced form is

$$choice = \pi_0 + \pi_1 faminc + \pi_2 grant + v_2,$$

and we need $\pi_2 \neq 0$. In other words, after accounting for income, the grant amount must have some effect on *choice*. This seems reasonable, provided the grant amounts differ within each income class.

(iv) The reduced form for score is just a linear function of the exogenous variables (see Problem 15.6):

$$score = \alpha_0 + \alpha_1 faminc + \alpha_2 grant + v_1.$$

This equation allows us to directly estimate the effect of increasing the grant amount on the test score, holding family income fixed. From a policy perspective, this is itself of some interest.

15.9 Just use OLS on an expanded equation, where *SAT* and *cumGPA* are added as proxy variables for student ability and motivation; see Chapter 9.

15.11 (i) We plug $x_t^* = x_t - e_t$ into $y_t = \beta_0 + \beta_1 x_t^* + u_t$:

$$\begin{aligned} y_t &= \beta_0 + \beta_1(x_t - e_t) + u_t = \beta_0 + \beta_1 x_t + u_t - \beta_1 e_t \\ &\equiv \beta_0 + \beta_1 x_t + v_t, \end{aligned}$$

where $v_t \equiv u_t - \beta_1 e_t$. By assumption, u_t is uncorrelated with x_t^* and e_t; therefore, u_t is uncorrelated with x_t. Since e_t is uncorrelated with x_t^*, $E(x_t e_t) = E[(x_t^* + e_t)e_t] = E(x_t^* e_t) + E(e_t^2) = \sigma_e^2$. Therefore, with v_t defined as above, $Cov(x_t,v_t) = Cov(x_t,u_t) - \beta_1 Cov(x_t,e_t) = -\beta_1 \sigma_e^2 < 0$ when $\beta_1 > 0$. Because the explanatory variable and the error have negative covariance, the OLS estimator of β_1 has a downward bias [see equation (5.4)].

(ii) By assumption $E(x_{t-1}^* u_t) = E(e_{t-1}u_t) = E(x_{t-1}^* e_t) = E(e_{t-1}e_t) = 0$, and so $E(x_{t-1}u_t) = E(x_{t-1}e_t) = 0$ because $x_t = x_t^* + e_t$. Therefore, $E(x_{t-1}v_t) = E(x_{t-1}u_t) - \beta_1 E(x_{t-1}e_t) = 0$.

(iii) Most economic time series, unless they represent the first difference of a series or the percentage change, are positively correlated over time. If the initial equation is in levels or logs, x_t and x_{t-1} are likely to be positively correlated. If the model is for first differences or percentage changes, there still may be positive or negative correlation between x_t and x_{t-1}.

(iv) Under the assumptions made, x_{t-1} is exogenous in

$$y_t = \beta_0 + \beta_1 x_t + v_t,$$

as we showed in part (ii): $Cov(x_{t-1},v_t) = E(x_{t-1}v_t) = 0$. Second, x_{t-1} will often be correlated with x_t, and we can check this easily enough by running a regression of x_t of x_{t-1}. This suggests estimating the equation by instrumental variables, where x_{t-1} is the IV for x_t. The IV estimator will be consistent for β_1 (and β_0) and asymptotically normally distributed.

SOLUTIONS TO COMPUTER EXERCISES

C15.1 (i) The regression of log(*wage*) on *sibs* gives

$$\widehat{\log(wage)} = 6.861 - .0279\, sibs$$
$$(0.022) \quad (.0059)$$
$$n = 935, \quad R^2 = .023.$$

This is a reduced form simple regression equation. It shows that, controlling for no other factors, one more sibling in the family is associated with monthly salary that is about 2.8% lower. The *t* statistic on *sibs* is about –4.73. Of course *sibs* can be correlated with many things that should have a bearing on wage including, as we already saw, years of education.

(ii) It could be that older children are given priority for higher education, and families may hit budget constraints and may not be able to afford as much education for children born later. The simple regression of *educ* on *brthord* gives

$$\widehat{educ} = 14.15 - .283\, brthord$$
$$(0.13) \quad (.046)$$
$$n = 852, \quad R^2 = .042.$$

(Note that *brthord* is missing for 83 observations.) The equation predicts that every one-unit increase in *brthord* reduces predicted education by about .28 years. In particular, the difference in predicted education for a first-born child and a fourth-born child is about .85 years.

(iii) When *brthord* is used as an IV for *educ* in the simple wage equation, we get

$$\widehat{\log(wage)} = 5.03 + .131\, educ$$
$$(0.43) \quad (.032)$$
$$n = 852.$$

(The *R*-squared is negative.) This is much higher than the OLS estimate (.060) and even above the estimate when *sibs* is used as an IV for *educ* (.122). Because of missing data on *brthord*, we are using fewer observations than in the previous analyses.

(iv) In the reduced form equation

$$educ = \pi_0 + \pi_1 sibs + \pi_2 brthord + v,$$

we need $\pi_2 \neq 0$ in order for the β_j to be identified. We take the null to be H_0: $\pi_2 = 0$ and look to reject H_0 at a small significance level. The regression of *educ* on *sibs* and *brthord* (using 852 observations) yields $\hat{\pi}_2 = -.153$ and $se(\hat{\pi}_2) = .057$. The *t* statistic is about –2.68, which rejects H_0 fairly strongly. Therefore, the identification assumptions appear to hold.

(v) The equation estimated by IV is

$$\widehat{\log(wage)} = \underset{(1.06)}{4.94} + \underset{(.075)}{.137\ educ} + \underset{(.0174)}{.0021\ sibs}$$

$$n = 852.$$

The standard error on $\hat{\beta}_{educ}$ is much larger than we obtained in part (iii). The 95% CI for β_{educ} is roughly –.010 to .284, which is very wide and includes the value zero. The standard error of $\hat{\beta}_{sibs}$ is very large relative to the coefficient estimate, rendering *sibs* very insignificant.

(vi) Letting $\widehat{educ}_i$ be the first-stage fitted values, the correlation between $\widehat{educ}_i$ and $sibs_i$ is about –.930, which is a very strong negative correlation. This means that, for the purposes of using IV, multicollinearity is a serious problem here and is not allowing us to estimate β_{educ} with much precision.

C15.3 (i) IQ scores are known to vary by geographic region and so does the availability of four year colleges. It could be that, for a variety of reasons, people with higher abilities grow up in areas with four-year colleges nearby.

(ii) The simple regression of *IQ* on *nearc4* gives

$$\widehat{IQ} = \underset{(0.63)}{100.61} + \underset{(0.74)}{2.60\ nearc4}$$

$$n = 2{,}061,\ R^2 = .0059,$$

which shows that predicted *IQ* score is about 2.6 points higher for a man who grew up near a four-year college. The difference is statistically significant (*t* statistic $\approx$ 3.51).

(iii) When we add *smsa66*, *reg662*, …, *reg669* to the regression in part (ii), we obtain

$$\widehat{IQ} = \underset{(1.62)}{104.77} + \underset{(.814)}{.348\ nearc4} + \underset{(0.81)}{1.09\ smsa66} + \ldots$$

$$n = 2{,}061,\ R^2 = .0626,$$

where, for brevity, the coefficients on the regional dummies are not reported. Now, the relationship between *IQ* and *nearc4* is much weaker and statistically insignificant. In other words, once we control for region and environment while growing up, there is no apparent link between IQ score and living near a four-year college.

(iv) The findings from parts (ii) and (iii) show that it is important to include *smsa66*, *reg662*, …, *reg669* in the wage equation to control for differences in access to colleges that might also be correlated with ability.

C15.5 (i) When we add $\hat{v}_2$ to the original equation and estimate it by OLS, the coefficient on $\hat{v}_2$ is about –.057 with a *t* statistic of about –1.08. Therefore, while the difference in the estimates of the return to education is practically large, it is not statistically significant.

(ii) We now add *nearc2* as an IV along with *nearc4*. (Although, in the reduced form for *educ*, *nearc2* is not significant.) The 2SLS estimate of β_1 is now .157, se($\hat{\beta}_1$) = .053. So the estimate is even larger.

(iii) Let $\hat{u}_i$ be the 2SLS residuals. We regress these on all exogenous variables, including *nearc2* and *nearc4*. The *n*-*R*-squared statistic is (3,010)(.0004) ≈ 1.20. There is one over-identifying restriction, so we compute the *p*-value from the χ_1^2 distribution: *p*-value = P(χ_1^2 > 1.20) ≈ .55, so the overidentifying restriction is not rejected.

C15.7 (i) As usual, if $unem_t$ is correlated with e_t, OLS will be biased and inconsistent for estimating β_1.

(ii) If E(e_t|inf_{t-1},$unem_{t-1}$, …) = 0, then $unem_{t-1}$ is uncorrelated with e_t, which means $unem_{t-1}$ satisfies the first requirement for an IV in

$$\Delta inf_t = \beta_0 + \beta_1 unem_t + e_t.$$

(iii) The second requirement for $unem_{t-1}$ to be a valid IV for $unem_t$ is that $unem_{t-1}$ must be sufficiently correlated. The regression $unem_t$ on $unem_{t-1}$ yields

$$\widehat{unem}_t = 1.57 + .732\, unem_{t-1}$$
$$(0.58) \quad (.097)$$
$$n = 48, \quad R^2 = .554.$$

Therefore, there is a strong, positive correlation between $unem_t$ and $unem_{t-1}$.

(iv) The expectations-augmented Phillips curve estimated by IV is

$$\widehat{\Delta inf}_t = .694 - .138\, unem_t$$
$$(1.883) \quad (.319)$$
$$n = 48, \quad R^2 = .048.$$

The IV estimate of β_1 is much lower in magnitude than the OLS estimate (–.543), and $\hat{\beta}_1$ is not statistically different from zero. The OLS estimate had a *t* statistic of about –2.36 [see equation (11.19)].

C15.9 (i) The IV (2SLS) estimates are

$$\widehat{\log(wage)} = \underset{(.54)}{5.22} + \underset{(.0337)}{.0936\,educ} + \underset{(.0084)}{.0209\,exper} + \underset{(.0027)}{.0115\,tenure} - \underset{(.050)}{.183\,black}$$

$n = 935,\ R^2 = .169.$

(ii) The coefficient on $\widehat{educ}_i$ in the second stage regression is, naturally, .0936. But the reported standard error is .0353, which is slightly too large.

(iii) When instead we (incorrectly) use $\widetilde{educ}_i$ in the second stage regression, its coefficient is .0700 and the corresponding standard error is .0264. Both are too low. The reduction in the estimated return to education from about 9.4% to 7.0% is not trivial. This illustrates that it is best to avoid doing 2SLS manually.

C15.11 (i) We look at the variables *selectyrs* and *choiceyrs*. From *selectyrs*, out of 990 students, 468 were never awarded a voucher and 108 were selected in the voucher system for all four years. From *choiceyrs*, only 56 actually attended a choice school for four years.

(ii) The estimated equation, with usual OLS standard errors in parentheses, is

$$\widehat{choiceyrs} = \underset{(.025)}{.020} + \underset{(.013)}{.767\,selectyrs}$$

$n = 990,\ R^2 = .790.$

This shows, as we would expect, a very strong relationship between the number of years attending a choice school and the number of years eligible. The *t* statistic is about 60, making *selectyrs* a very good candidate as an IV. Remember, unlike *choiceyrs*, *selectyrs* is randomly assigned.

(iii) Regressing *mnce* on *choiceyrs* gives the following (usual OLS standard errors in parentheses):

$$\widehat{mnce} = \underset{(0.85)}{46.23} - \underset{(0.53)}{1.84\,choiceyrs}$$

$n = 990,\ R^2 = .012.$

This shows that the math percentile is negatively related to the number of years in a choice school. If we try to interpret this causally, we would have to assume the voucher program was an abysmal failure. Each additional year in a choice school reduces the percentile by almost two percentage points.

When *black*, *hispanic*, and *female* are added, the coefficient on *choiceyrs* becomes $-.57$ with a *t* of about -1.07. The magnitude of the effects has fallen substantially and it is no longer

statistically significant. However, we have not found a positive, statistically significant effect, either. Based on this analysis, the voucher program also had no positive effects. You are invited to see that the main conclusions do not change using heteroskedasticity-robust standard errors.

(iv) Even controlling for race, ethnicity, and gender, and even with the vouchers randomly assigned, students (aided by parents) could self-select into the program. In particular, it could be that students struggling in the traditional public school, or students with lower performance in math, are more likely to take the opportunity to switch to a choice school. If so, this induces a negative correlation between *choiceyrs* and the error term, u_1, resulting in a downward bias in the estimator of β_1.

(v) The IV estimates are given by

$$\widehat{mnce} = \underset{(1.66)}{57.07} - \underset{(.605)}{.241\ choiceyrs} - \underset{(1.81)}{16.32\ black} - \underset{(2.34)}{13.78\ hispanic} + \underset{(1.28)}{1.32\ female}$$

$n = 990$, $R^2 = .086$.

The IV estimate of β_1 is even smaller in magnitude with a much smaller *t* statistic, so all we can conclude is that the voucher program had no effect on math performance. The coefficients on *black* and *hispanic* are very large and statistically significant, illustrating an achievement gap that has often been noted on standardized tests.

(vi) When *mnce90* is added to the equation, and OLS is used, the estimate of β_1 becomes .411 with $t = .56$. So now the coefficient is positive but it is pretty small and, more importantly, not nearly statistically significant.

When *selectyrs* is used as an IV in the same equation, the estimate of β_1 becomes 1.80 with $t = 2.09$. (The heteroskedasticity-robust *t* is 1.90.) So controlling for *mnce90* combined with using an instrument for *choiceyrs* now shows a positive effect of the program. Each year in a choice school increases the percentile in the math NCE by about 1.8 percentage points. The estimate is marginally statistically significant. This seems like a reasonably large effect.

(vii) Unfortunately, including *mnce90* in the equation results in a severe loss of observations. The initial score is available for only 328 students, and so it is missing for 662 students. It seems likely that whether *mnce90* is missing is not entirely random. Thus, at best we can claim an effect for the subpopulation where the 1990 percentile is available. This need not be representative of the population of interest.

(viii) There is a typo in the first printing of the text. The equation should include *mnce90*, although it is useful to see what happens without it, too. The IV estimates and *t* statistics are presented below for the four *choiceyrs* dummy variables.

Without *mnce90* (*n* = 990):

*choiceyrs*1: .390 (.17)
*choiceyrs*2: .774 (.19)
*choiceyrs*3: −4.28 (−1.16)
*choiceyrs*4: 2.41 (4.16)

None of the estimates is close to statistically significant; the one with the largest *t* in absolute value has the "wrong" sign. The joint *F* test of significance gives *p*-value = .785.

With *mnce90* (*n* = 328):

*choiceyrs*1: −2.16 (−.42)
*choiceyrs*2: 1.49 (.32)
*choiceyrs*3: 1.08 (.15)
*choiceyrs*4: 13.93 (2.19)

The coefficient on *choiceyrs*4 is huge – a 14 percentage point move up in the math score distribution – but it is imprecisely estimated (with a marginal *t* statistic). It is clear from the pattern of coefficients and *t* statistics that it is students who went to a choice school all four years driving the finding in part (vi). It is troubling that being in the choice program, say, three years has a much smaller and statistically insignificant effect. Part of the problem is that the sample size is pretty small. But the pattern is nevertheless suspicious and suggests more research is warranted.