CHAPTER 16

SOLUTIONS TO PROBLEMS

16.1 (i) If $\alpha_1 = 0$, then $y_1 = \beta_1 z_1 + u_1$, and so the right-hand-side depends only on the exogenous variable z_1 and the error term u_1. Then this is the reduced form for y_1. If $\alpha_2 = 0$, the reduced form for y_1 is $y_1 = \beta_2 z_2 + u_2$. (Note that having both α_1 and α_2 equal zero is not interesting as it implies the bizarre condition $u_2 - u_1 = \beta_1 z_1 - \beta_2 z_2$.)

If $\alpha_1 \neq 0$ and $\alpha_2 = 0$, we can plug $y_1 = \beta_2 z_2 + u_2$ into the first equation and solve for y_2:

$$\beta_2 z_2 + u_2 = \alpha_1 y_2 + \beta_1 z_1 + u_1$$

or

$$\alpha_1 y_2 = \beta_1 z_1 - \beta_2 z_2 + u_1 - u_2.$$

Dividing by α_1 (because $\alpha_1 \neq 0$) gives

$$\begin{aligned} y_2 &= (\beta_1/\alpha_1) z_1 - (\beta_2/\alpha_1) z_2 + (u_1 - u_2)/\alpha_1 \\ &\equiv \pi_{21} z_1 + \pi_{22} z_2 + v_2, \end{aligned}$$

where $\pi_{21} = \beta_1/\alpha_1$, $\pi_{22} = -\beta_2/\alpha_1$, and $v_2 = (u_1 - u_2)/\alpha_1$. Note that the reduced form for y_2 generally depends on z_1 and z_2 (as well as on u_1 and u_2).

(ii) If we multiply the second structural equation by (α_1/α_2) and subtract it from the first structural equation, we obtain

$$\begin{aligned} y_1 - (\alpha_1/\alpha_2) y_1 &= \alpha_1 y_2 - \alpha_1 y_2 + \beta_1 z_1 - (\alpha_1/\alpha_2)\beta_2 z_2 + u_1 - (\alpha_1/\alpha_2) u_2 \\ &= \beta_1 z_1 - (\alpha_1/\alpha_2)\beta_2 z_2 + u_1 - (\alpha_1/\alpha_2) u_2 \end{aligned}$$

or

$$[1 - (\alpha_1/\alpha_2)] y_1 = \beta_1 z_1 - (\alpha_1/\alpha_2)\beta_2 z_2 + u_1 - (\alpha_1/\alpha_2) u_2.$$

Because $\alpha_1 \neq \alpha_2$, $1 - (\alpha_1/\alpha_2) \neq 0$, and so we can divide the equation by $1 - (\alpha_1/\alpha_2)$ to obtain the reduced form for y_1: $y_1 = \pi_{11} z_1 + \pi_{12} z_2 + v_1$, where $\pi_{11} = \beta_1/[1 - (\alpha_1/\alpha_2)]$, $\pi_{12} = -(\alpha_1/\alpha_2)\beta_2/[1 - (\alpha_1/\alpha_2)]$, and $v_1 = [u_1 - (\alpha_1/\alpha_2) u_2]/[1 - (\alpha_1/\alpha_2)]$.

A reduced form does exist for y_2, as can be seen by subtracting the second equation from the first:

$$0 = (\alpha_1 - \alpha_2) y_2 + \beta_1 z_1 - \beta_2 z_2 + u_1 - u_2;$$

because $\alpha_1 \neq \alpha_2$, we can rearrange and divide by $\alpha_1 - \alpha_2$ to obtain the reduced form.

(iii) In supply and demand examples, $\alpha_1 \neq \alpha_2$ is very reasonable. If the first equation is the supply function, we generally expect $\alpha_1 > 0$, and if the second equation is the demand function, $\alpha_2 < 0$. The reduced forms can exist even in cases where the supply function is not upward sloping and the demand function is not downward sloping, but we might question the usefulness of such models.

16.3 No. In this example, we are interested in estimating the tradeoff between sleeping and working, controlling for some other factors. OLS is perfectly suited for this, provided we have been able to control for all other relevant factors. While it is true that individuals are assumed to optimally allocate their time subject to constraints, this does not result in a system of simultaneous equations. If we write down such a system, there is no sense in which each equation could stand on its own; neither would have an interesting ceteris paribus interpretation. Besides, we could not estimate either equation because economic reasoning gives us no way of excluding exogenous variables from either equation. See Example 16.2 for a similar discussion.

16.5 (i) Other things equal, a higher rate of condom usage should reduce the rate of sexually transmitted diseases (STDs). So $\beta_1 < 0$.

(ii) If students having sex behave rationally and condom usage does prevent STDs, then condom usage should increase as the rate of infection increases.

(iii) If we plug the structural equation for *infrate* into $conuse = \gamma_0 + \gamma_1 infrate + \ldots$, we see that *conuse* depends on γ_1 and u_1. Because $\gamma_1 > 0$, *conuse* is positively related to u_1. In fact, if the structural error (u_2) in the *conuse* equation is uncorrelated with u_1, $\text{Cov}(conuse,u_1) = \gamma_1 \text{Var}(u_1) > 0$. If we ignore the other explanatory variables in the *infrate* equation, we can use equation (5.4) to obtain the direction of bias: $\text{plim}(\hat{\beta}_1) - \beta_1 > 0$ because $\text{Cov}(conuse,u_1) > 0$, where $\hat{\beta}_1$ denotes the OLS estimator. Since we think $\beta_1 < 0$, OLS is biased towards zero. In other words, if we use OLS on the *infrate* equation, we are likely to underestimate the importance of condom use in reducing STDs. (Remember, the more negative is β_1, the more effective is condom usage.)

(iv) We would have to assume that *condis* does not appear, in addition to *conuse*, in the *infrate* equation. This seems reasonable, as its usage should directly affect STDs, and not those just having a distribution program. But we must also assume *condis* is exogenous in the *infrate*: it cannot be correlated with unobserved factors (in u_1) that also affect *infrate*.

We must also assume that *condis* has some partial effect on *conuse*, something that can be tested by estimating the reduced form for *conuse*. It seems likely that this requirement for an IV – see equations (15.30) and (15.31) – is satisfied.

16.7 (i) Attendance at women's basketball may grow in ways that are unrelated to factors that we can observe and control for. The taste for women's basketball may increase over time, and this would be captured by the time trend.

(ii) No. The university sets the price, and it may change price based on expectations of next year's attendance; if the university uses factors that we cannot observe, these are necessarily in

the error term u_t. So even though the supply is fixed, it does not mean that price is uncorrelated with the unobservables affecting demand.

(iii) If people only care about how this year's team is doing, $SEASPERC_{t-1}$ can be excluded from the equation once $WINPERC_t$ has been controlled for. Of course, this is not a very good assumption for all games, as attendance early in the season is likely to be related to how the team did last year. We would also need to check that $1PRICE_t$ is partially correlated with $SEASPERC_{t-1}$ by estimating the reduced form for $1PRICE_t$.

(iv) It does make sense to include a measure of men's basketball ticket prices, as attending a women's basketball game is a substitute for attending a men's game. The coefficient on $1MPRICE_t$ would be expected to be positive: an increase in the price of men's tickets should increase the demand for women's tickets. The winning percentage of the men's team is another good candidate for an explanatory variable in the women's demand equation.

(v) It might be better to use first differences of the logs, which are then growth rates. We would then drop the observation for the first game in each season.

(vi) If a game is sold out, we cannot observe true demand for that game. We only know that desired attendance is some number above capacity. If we just plug in capacity, we are understating the actual demand for tickets. (Chapter 17 discusses censored regression methods that can be used in such cases.)

SOLUTIONS TO COMPUTER EXERCISES

C16.1 (i) Assuming the structural equation represents a causal relationship, $100 \cdot \beta_1$ is the approximate percentage change in income if a person smokes one more cigarette per day.

(ii) Since consumption and price are, ceteris paribus, negatively related, we expect $\gamma_5 \leq 0$ (allowing for $\gamma_5 = 0$). Similarly, everything else equal, restaurant smoking restrictions should reduce cigarette smoking, so $\gamma_6 \leq 0$.

(iii) We need γ_5 or γ_6 to be different from zero. That is, we need at least one exogenous variable in the *cigs* equation that is not also in the log(*income*) equation.

(iv) OLS estimation of the log(*income*) equation gives

$$\widehat{\log(income)} = \underset{(0.17)}{7.80} + \underset{(.0017)}{.0017\, cigs} + \underset{(.008)}{.060\, educ} + \underset{(.008)}{.058\, age} - \underset{(.00008)}{.00063\, age^2}$$

$n = 807$, $R^2 = .165$.

The coefficient on *cigs* implies that cigarette smoking causes income to increase, although the coefficient is not statistically different from zero. Remember, OLS ignores potential simultaneity between income and cigarette smoking.

(v) The estimated reduced form for *cigs* is

$$\widehat{cigs} = \underset{(23.70)}{1.58} - \underset{(.162)}{.450\,educ} + \underset{(.154)}{.823\,age} - \underset{(.0017)}{.0096\,age^2} - \underset{(5.766)}{.351\log(cigpric)} - \underset{(1.11)}{2.74\,restaurn}$$

$n = 807,\ R^2 = .051.$

While log(*cigpric*) is very insignificant, *restaurn* had the expected negative sign and a *t* statistic of about –2.47. (People living in states with restaurant smoking restrictions smoke almost three fewer cigarettes, on average, given education and age.) We could drop log(*cigpric*) from the analysis but we leave it in. (Incidentally, the *F* test for joint significance of log(*cigpric*) and *restaurn* yields *p*-value ≈ .044.)

(vi) Estimating the log(*income*) equation by 2SLS gives

$$\widehat{\log(income)} = \underset{(0.23)}{7.78} - \underset{(.026)}{.042\,cigs} + \underset{(.016)}{.040\,educ} + \underset{(.023)}{.094\,age} - \underset{(.00027)}{.00105\,age^2}$$

$n = 807.$

Now the coefficient on *cigs* is negative and almost significant at the 10% level against a two-sided alternative. The estimated effect is very large: each additional cigarette someone smokes lowers predicted income by about 4.2%. Of course, the 95% CI for β_{cigs} is very wide.

(vii) Assuming that state level cigarette prices and restaurant smoking restrictions are exogenous in the income equation is problematical. Incomes are known to vary by region as do restaurant smoking restrictions. It could be that in states where income is lower (after controlling for education and age), restaurant smoking restrictions are less likely to be in place.

C16.3 (i) The OLS estimates are

$$\widehat{inf} = \underset{(4.10)}{25.23} - \underset{(.093)}{.215\,open}$$

$n = 114,\ R^2 = .045.$

The IV estimates are

$$\widehat{inf} = \underset{(5.66)}{29.61} - \underset{(.140)}{.333\,open}$$

$n = 114,\ R^2 = .032.$

The OLS coefficient is the same, to three decimal places, when log(*pcinc*) is included in the model. The IV estimate with log(*pcinc*) in the equation is –.337, which is very close to –.333. Therefore, dropping log(*pcinc*) makes little difference.

(ii) Subject to the requirement that an IV be exogenous, we want an IV that is as highly correlated as possible with the endogenous explanatory variable. If we regress *open* on *land* we obtain $R^2 = .095$. The simple regression of *open* on log(*land*) gives $R^2 = .448$. Therefore, log(*land*) is much more highly correlated with *open*. Further, if we regress open on log(*land*) and *land* we get

$$\widehat{open} = \underset{(10.47)}{129.22} - \underset{(0.98)}{8.40}\log(land) + \underset{(.0000031)}{.0000043}\,land$$

$$n = 114, \quad R^2 = .457.$$

While log(*land*) is very significant, *land* is not, so we might as well use only log(*land*) as the IV for open.

(iii) When we add *oil* to the original model and assume *oil* is exogenous, the IV estimates are

$$\widehat{inf} = \underset{(15.75)}{24.01} - \underset{(.1424)}{.337}\,open + \underset{(2.08)}{.803}\log(pcinc) - \underset{(9.80)}{6.56}\,oil$$

$$n = 114, \quad R^2 = .035.$$

Being an oil producer, it is estimated to reduce average annual inflation by over 6.5 percentage points, but the effect is not statistically significant. This is not too surprising, as there are only seven oil producers in the sample.

C16.5 This is an open-ended question without a single answer. Even if we settle on extending the data through a particular year, we might want to change the disposable income and nondurable consumption numbers in earlier years, as these are often recalculated. For example, the value for real disposable personal income in 1995, as reported in Table B-29 of the 1997 *Economic Report of the President* (*ERP*), is $4,945.8 billions. In the 1999 *ERP*, this value has been changed to $4,906.0 billions (see Table B-31). All series can be updated using the latest edition of the *ERP*. The key is to use real values and make them per capita by dividing by population. Make sure that you use nondurable consumption.

C16.7 (i) If county administrators can predict when crime rates will increase, they may hire more police to counteract crime. This would explain the estimated positive relationship between $\Delta\log(crmrte)$ and $\Delta\log(polpc)$ in equation (13.33).

(ii) This may be reasonable, although tax collections depend in part on income and sales taxes and revenues from these depend on the state of the economy, which can also influence crime rates.

(iii) The reduced form for $\Delta\log(polpc_{it})$, for each *i* and *t*, is

$$\begin{aligned}\Delta\log(polpc_{it}) = {} & \pi_0 + \pi_1 d83_t + \pi_2 d84_t + \pi_3 d85_t + \pi_4 d86_t + \pi_5 d87_t \\ & + \pi_6 \Delta\log(prbarr_{it}) + \pi_7 \Delta\log(prbconv_{it}) + \pi_8 \Delta\log(prbpris_{it}) \\ & + \pi_9 \Delta\log(avgsen_{it}) + \pi_{10} \Delta\log(taxpc_{it}) + v_{it}.\end{aligned}$$

We need $\pi_{10} \neq 0$ for $\Delta\log(taxpc_{it})$ to be a reasonable IV candidate for $\Delta\log(polpc_{it})$. When we estimate this equation by pooled OLS ($N = 90$, $T = 6$ for $n = 540$), we obtain $\hat{\pi}_{10} = .0052$ with a *t* statistic of only .080. Therefore, $\Delta\log(taxpc_{it})$ is not a good IV for $\Delta\log(polpc_{it})$.

(iv) If the grants were awarded randomly, then the grant amounts, say $grant_{it}$ for the dollar amount for county *i* and year *t*, would be uncorrelated with Δu_{it}, the changes in unobservables that affect county crime rates. By definition, $grant_{it}$ should be correlated with $\Delta\log(polpc_{it})$ across *i* and *t*. This means we have an exogenous variable that can be omitted from the crime equation and that is (partially) correlated with the endogenous explanatory variable. We could reestimate (13.33) by IV.

C16.9 (i) The demand function should be downward sloping, so $\alpha_1 < 0$: as price increases, quantity demanded for air travel decreases.

(ii) The estimated price elasticity is –.391 (*t* statistic = –5.82).

(iii) We must assume that passenger demand depends only on air fare, so that, once price is controlled for, passengers are indifferent about the fraction of travel accounted for by the largest carrier.

(iv) The reduced form equation for log(*fare*) is

$$\begin{aligned}\widehat{\log(fare)} = {} & 6.19 + .395\, concen - .936 \log(dist) + .108\, [\log(dist)]^2 \\ & (0.89) \quad (.063) \qquad (.272) \qquad (.021)\end{aligned}$$

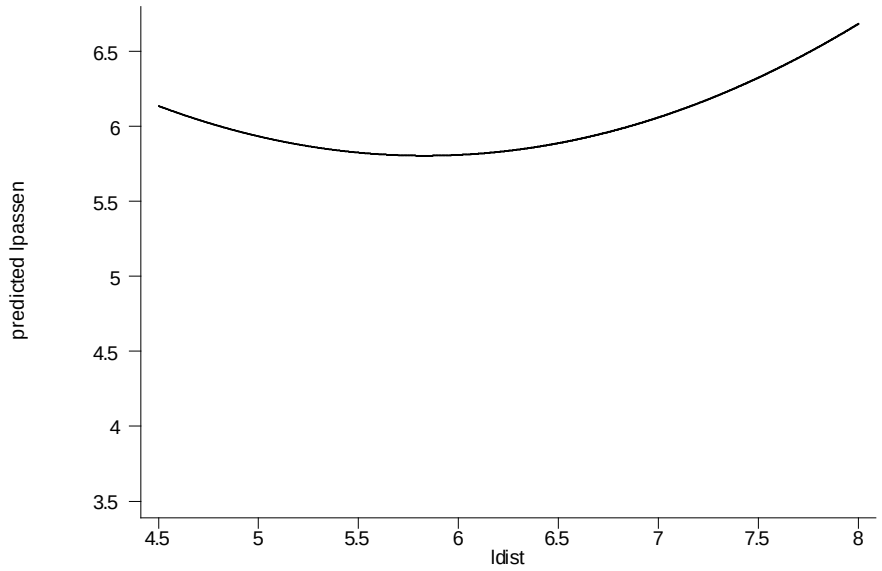
$n = 1{,}149$, $R^2 = .408$.

The coefficient on *concen* shows a pretty strong link between concentration and fare. If *concen* increases by .10 (10 percentage points), *fare* is estimated to increase by almost 4%. The *t* statistic is about 6.3.

(v) Using *concen* as an IV for log(*fare*) [and where the distance variables act as their own IVs], the estimated price elasticity is –1.17, which shows much greater price sensitivity than did the OLS estimate. The IV estimate suggests that a one percent increase in fare leads to a slightly more than one percent increase drop in passenger demand. Of course, the standard error of the

IV estimate is much larger (about .389 compared with the OLS standard error of .067), but the IV estimate is statistically significant (*t* is about −3.0).

(vi) The coefficient on *ldist* = log(*dist*) is about −2.176 and that on *ldistsq* = [log(*dist*)]^2 is about .187. Therefore, the relationship between log(*passen*) and log(*dist*) has a U-shape, as given in the following graph:



The minimum is at about *ldist* = 2.176/(2*.187) ≈ 5.82, which, in terms of distance, is about 337 miles. About 11.3% of the routes are less than 337 miles long. If the estimated quadratic is believable, the lowest demand occurs for short, but not very short, routes (holding price fixed). It is possible, of course, that we should ignore the quadratic to the left of the turning point, but it does contain a nontrivial fraction of the observations.

C16.11 (i) Logically, *sfood* has to be between zero and one. In this sample, it ranges from a low of about .057 to a high of .789. It is not surprising to see no zeros; presumably everyone has to spend something on food (except maybe a completely self-sufficient farmer who does not put a price on his or her own food).

(ii) The coefficient on *ltotexpend* is $-.146$ with a robust standard error of .0062. If *ltotexpend* increases by, say, .10 – which means a 10% increase in total expenditure – the share of food in total expenditure falls by about .015, or 1.5 percentage points. Of course, this holds fixed age of the head of household and number of children.

(iii) The reduced form equation is

$$ltotexpend = \pi_0 + \pi_1 lincome + \pi_2 age + \pi_3 kids + v_2.$$

When we estimate this by OLS, the coefficient (robust *t* statistic) on *lincome* is .478 (16.78). Because *lincome* is such a good predictor of *ltotexpend* – with the sign we expect – *lincome* is a very good IV for *ltotexpend*, provided *lincome* is exogenous.

(iv) When (16.43) is estimated by IV, using *lincome* as an instrument for *ltotexpend*, the estimate (robust *t* statistic) on *ltotexpend* is $-.160$ (-12.41). The robust 95% CI runs from $-.185$ to $-.135$. The IV estimate is somewhat larger in magnitude, and the CI is somewhat wider than for OLS: $-.158$ to $-.134$.

(v) When we obtain the reduced form residuals, say $\hat{v}_2$, from part (iii) and add them to the OLS regression from part (ii), the robust *t* statistic on $\hat{v}_2$ is only 1.14 (two-sided *p*-value = .254). Therefore, we cannot reject the null hypothesis that *ltotexpend* is exogenous even at the 25% significance level.

There are no overidentifying restrictions to test. We have one IV, *lincome*, for the potentially endogenous explanatory variable *ltotexpend*.

(vi) The OLS estimate of the *ltotexpend* coefficient is .028 (robust $t = 6.66$), while the IV estimate is .030 (robust $t = 3.14$). The estimates are similar (and the IV estimate is less precise), and the test for endogeneity of *ltotexpend* gives a robust *t* of about $-.27$. Therefore, we can safely estimate the share equation for alcohol by OLS.