PRINCIPLES OF ECONOMETRICS

5TH EDITION

ANSWERS TO ODD-NUMBERED EXERCISES IN CHAPTER 10

EXERCISE 10.1

(a) The price of housing and rent paid are determined by supply and demand forces in the market place. The omitted factors from this regression include macroeconomic forces, such as unemployment rates, interest rates, population growth, etc., all of which might well affect not only rent paid but also the median house value. If there is correlation between median house value and the regression error term, then median house value is endogenous.

(b) The model in column (1) contains one potentially endogenous variable, *MDHOUSE*. To carry out instrumental variables estimation we require at least one strong instrument. There are 2 potential instruments. We test for strong instruments by computing the joint *F*-test of significance of these variables in the first stage regression. Column (2) contains the first stage regression results including all instruments. Column (3) contains the first stage regression omitting *FAMINC* and *REG4*. Using the sum of squared residuals *SSE* in columns (2) and (3) we can compute the *F*-statistic as $F = 25.99$. By the Staiger-Stock rule of thumb we are satisfied because the calculated *F* is greater than 10.

A more informative answer is obtained by examining the critical values for the weak instrument tests of Stock and Yogo in Table 10E.1 and 10E.2. If we adopt the Maximum Test Size criterion, for a test of the coefficient on the endogenous variable, and are willing to accept a test size of 0.10 for a 5% test, then the critical value for the *F*-statistic is 19.93 [*B*=1, *L*=2]. The null hypothesis is that the instruments are weak, so that under this criterion we can conclude that the instruments are not weak. We cannot use the Maximum Relative Bias criterion because we have only two instruments.

(c) The regression based Hausman test for endogeneity augments the regression of interest with the least squares residuals from the first stage regression. The null hypothesis is that the variable *MDHOUSE* is exogenous, and the alternative hypothesis is that *MDHOUSE* is endogenous. The Hausman test is a *t*-test for the significance of the coefficient of *VHAT*. The 2-tail critical value of the *t*-distribution with 48 degrees of freedom is 2.01. The calculated value of the *t*-statistic is −3.44. Since −3.44 < −2.01 we reject the null hypothesis that the coefficient of *VHAT* is zero using the 0.05 level of significance. We conclude that *MDHOUSE* is endogenous.

(d) We note two important changes when we compare the least squares estimates in column (1) and the instrumental variables estimates in column (5). First, the *IV* estimate of the coefficient of *PCTURBAN* is much smaller than the corresponding least squares estimate, and its standard error is larger. The coefficient of *PCTURBAN* is now insignificant, whereas the least squares estimate's *t*-value of 1.81 is significant at the 0.10 level. Secondly, the *IV* estimate of the effect of *MDHOUSE* on *RENT* is larger in magnitude, indicating a larger effect than we first estimated. The standard error of the *IV* coefficient is larger (0.339) than the corresponding least squares estimate, but the $t = 6.43$ is very significant.

That the estimates for the structural parameters are the same in columns (4) and (5) is not an accident. The first stage least squares residuals *VHAT* are uncorrelated with *PCTURBAN*, because it is an explanatory variable in the first stage regression, and it is a property of the least squares residuals that they are uncorrelated with model explanatory variables. Also,

Copyright © 2018 Wiley

VHAT is uncorrelated with the fitted value of *MDHOUSE* that is used to compute the *2SLS/IV* estimates, as explained below equation (10D.8)

(e) The test for the validity of the 1 surplus instrument (the overidentifying restrictions) is computed as NR^2 from the artificial regression of the 2SLS/IV residuals on all available instruments. The resulting statistic, under the null hypothesis that the surplus instruments are valid (uncorrelated with the regression error) is distributed as $\chi^2_{(L-B=2-1=1)}$. The value of the test statistic is $NR^2 = 50 \times 0.1977 = 9.885$. From Statistical Table 3, the 0.95 percentile of the $\chi^2_{(1)}$ distribution is 3.841. We conclude that the surplus instrument is not valid, and therefore that the *IV* estimates in column (5) are questionable. The test does not identify which instrumental variable might be the problem.

EXERCISE 10.3

(a) Subtract $E(x) = \gamma_1 + \theta_1 E(z)$ from $x = \gamma_1 + \theta_1 z + v$ to obtain $x - E(x) = \theta_1(z - E(z)) + v$. Multiply both sides by $(z - E(z))$ to obtain

$$(z-E(z))(x-E(x)) = \theta_1(z-E(z))^2 + (z-E(z))v$$

Take the expected value of both sides to obtain

$$\begin{aligned} E\left[(z-E(z))(x-E(x))\right] &= \theta_1 E\left[(z-E(z))^2\right] + E(z-E(z))v \\ &= \theta_1 E\left[(z-E(z))^2\right] \end{aligned}$$

assuming $E(z-E(z))v = 0$. Solving for θ_1 we obtain

$$\theta_1 = \frac{E\left[(z-E(z))(x-E(x))\right]}{E\left[(z-E(z))^2\right]} = \frac{\operatorname{cov}(z,x)}{\operatorname{var}(z)}$$

This is the OLS estimator of θ_1 in the regression $x = \gamma_1 + \theta_1 z + v$.

(b) Subtract $E(y) = \pi_0 + \pi_1 E(z)$ from $y = \pi_0 + \pi_1 z + u$ to obtain

$$y - E(y) = \pi_0 + \pi_1(z - E(z)) + u$$

Multiply both sides by $(z - E(z))$ to obtain

$$(z-E(z))(y-E(y)) = \pi_1(z-E(z))^2 + (z-E(z))u$$

Assuming $E(z-E(z))u = 0$, take the expected value of both sides to obtain

$$E(z-E(z))(y-E(y)) = \pi_1 E(z-E(z))^2$$

Copyright © 2018 Wiley

Solving for π_1 we have

$$\pi_1 = \frac{E\left(z - E(z)\right)\left(y - E(y)\right)}{E\left(z - E(z)\right)^2}$$

This is the OLS estimator of π_1 in the regression $y = \pi_0 + \pi_1 z + u$.

(c) The substitution leaves

$$\begin{aligned} y &= \beta_1 + \beta_2 x + e = \beta_1 + \beta_2\left(\gamma_1 + \theta_1 z + v\right) + e \\ &= \left(\beta_1 + \beta_2\gamma_1\right) + \beta_2\theta_1 z + \left(\beta_2 v + e\right) \\ &= \pi_0 + \pi_1 z + u \end{aligned}$$

Thus $\pi_0 = \left(\beta_1 + \beta_2\gamma_1\right)$, $\pi_1 = \beta_2\theta_1$ and $u = \left(\beta_2 v + e\right)$

(d) Solving $\pi_1 = \beta_2\theta_1$ for β_2 we have $\beta_2 = \pi_1/\theta_1$.

(e) From (a),

$$\hat{\theta}_1 = \frac{\widehat{\text{cov}}(z,x)}{\widehat{\text{var}}(z)} = \frac{\sum(z_i - \bar{z})(x_i - \bar{x})/N}{\sum(z_i - \bar{z})^2/N} = \frac{\sum(z_i - \bar{z})(x_i - \bar{x})}{\sum(z_i - \bar{z})^2}$$

This estimator is consistent if *z* is uncorrelated with *v*. Similarly,

$$\hat{\pi}_1 = \frac{\widehat{\text{cov}}(z,y)}{\widehat{\text{var}}(z)} = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})/N}{\sum(z_i - \bar{z})^2/N} = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})^2}$$

is a consistent estimator if *z* is uncorrelated with *u*.

Then

$$\begin{aligned} \hat{\beta}_2 = \hat{\pi}_1/\hat{\theta}_1 &= \frac{\left[\dfrac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})^2}\right]}{\left[\dfrac{\sum(z_i - \bar{z})(x_i - \bar{x})}{\sum(z_i - \bar{z})^2}\right]} = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})(x_i - \bar{x})} = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})/N}{\sum(z_i - \bar{z})(x_i - \bar{x})/N} \\ &= \frac{\widehat{\text{cov}}(z,y)}{\widehat{\text{cov}}(z,x)} \end{aligned}$$

This is the IV estimator given in equation (10.17). The consistency of this estimator is established using the fact that sample moments converge to population moments, so that $\widehat{\text{cov}}(z,y) \xrightarrow{p} \text{cov}(z,y)$ and $\widehat{\text{cov}}(z,x) \xrightarrow{p} \text{cov}(z,x)$. It follows that

$$\hat{\beta}_2 = \hat{\pi}_1/\hat{\theta}_1 = \frac{\widehat{\text{cov}}(z,y)}{\widehat{\text{cov}}(z,x)} \xrightarrow{p} \frac{\text{cov}(z,y)}{\text{cov}(z,x)} = \beta_2$$

Copyright © 2018 Wiley

EXERCISE 10.5

(a) For the discrete random variable z_i, its expected value is

$$E(z_i) = \sum_{z_i} z_i f(z_i) = (1 \times p) + 0(1-p) = p$$

(b) Use the law of iterated expectations.

$$\begin{aligned} E(y_i z_i) &= E_{z_i} E(y_i z_i \mid z_i) = \sum_{z_i} E(y_i z_i \mid z_i) f(z_i) \\ &= E(y_i \times 1 \mid z_i = 1) p(z_i = 1) + E(y_i \times 0 \mid z_i = 0) p(z_i = 0) \\ &= E(y_i \mid z_i = 1) p(z_i = 1) = E(y_i \mid z_i = 1) p \end{aligned}$$

(c) $E(y_i) = E_{z_i} E(y_i \mid z_i) = \sum_{z_i} E(y_i \mid z_i) f(z_i) = E(y_i \mid z_i = 1) p + E(y_i \mid z_i = 0)(1-p)$

(d)

$$\begin{aligned} \operatorname{cov}(y_i, z_i) &= E(y_i z_i) - E(y_i) E(z_i) \\ &= E(y_i \mid z_i = 1) p - \left[E(y_i \mid z_i = 1) p + E(y_i \mid z_i = 0)(1-p)\right] p \\ &= E(y_i \mid z_i = 1) p - E(y_i \mid z_i = 1) p^2 - pE(y_i \mid z_i = 0)(1-p) \\ &= E(y_i \mid z_i = 1)(p - p^2) - p(1-p) E(y_i \mid z_i = 0) \\ &= p(1-p) E(y_i \mid z_i = 1) - p(1-p) E(y_i \mid z_i = 0) \\ &= p(1-p)\left[E(y_i \mid z_i = 1) - E(y_i \mid z_i = 0)\right] \end{aligned}$$

(e) First, using (b), $E(x_i z_i) = pE(x_i \mid z_i = 1)$. Second, using (c),

$$E(x_i) = pE(x_i \mid z_i = 1) + (1-p) E(x_i \mid z_i = 0)$$

Then from part (d), $\operatorname{cov}(x_i, z_i) = p(1-p)\left[E(x_i \mid z_i = 1) - E(x_i \mid z_i = 0)\right]$.

(f) Assuming, $y_i = \beta_1 + \beta_2 x_i + e_i$, and taking the expected values of both sides, $E(y_i) = \beta_1 + \beta_2 E(x_i) + E(e_i) = \beta_1 + \beta_2 E(x_i)$. Subtract the expected value expression from the assumed model to obtain $\left[y_i - E(y_i)\right] = \beta_2 \left[x_i - E(x_i)\right] + e_i$.

(g)

$$\begin{aligned} &\left[z_i - E(z_i)\right]\left\{\left[y_i - E(y_i)\right] = \beta_2\left[x_i - E(x_i)\right] + e_i\right\} \\ &\Rightarrow \left[z_i - E(z_i)\right]\left[y_i - E(y_i)\right] = \beta_2\left[z_i - E(z_i)\right]\left[x_i - E(x_i)\right] + \left[z_i - E(z_i)\right] e_i \\ &\Rightarrow E\left[z_i - E(z_i)\right]\left[y_i - E(y_i)\right] = \beta_2 E\left[z_i - E(z_i)\right]\left[x_i - E(x_i)\right] + E\left[z_i - E(z_i)\right] e_i \\ &\Rightarrow \operatorname{cov}(z_i, y_i) = \beta_2 \operatorname{cov}(z_i, x_i) + \operatorname{cov}(z_i, e_i) \\ &\Rightarrow \operatorname{cov}(y_i, z_i) = \beta_2 \operatorname{cov}(x_i, z_i) \text{ if } \operatorname{cov}(z_i, e_i) = 0 \end{aligned}$$

(h) Using $\operatorname{cov}(y_i, z_i) = \beta_2 \operatorname{cov}(x_i, z_i)$, so that $\beta_2 = \operatorname{cov}(y_i, z_i) / \operatorname{cov}(x_i, z_i)$, then

Copyright © 2018 Wiley

$$\beta_2 = \frac{\operatorname{cov}(y_i, z_i)}{\operatorname{cov}(x_i, z_i)} = \frac{p(1-p)\left[E(y_i \mid z_i = 1) - E(y_i \mid z_i = 0)\right]}{p(1-p)\left[E(x_i \mid z_i = 1) - E(x_i \mid z_i = 0)\right]}$$
$$= \frac{\left[E(y_i \mid z_i = 1) - E(y_i \mid z_i = 0)\right]}{\left[E(x_i \mid z_i = 1) - E(x_i \mid z_i = 0)\right]}$$

(i) Replace $E(y_i \mid z_i = 1)$ by its consistent estimator $\bar{y}_1$, the sample mean of the y_i values when $z_i = 1$, and so on, to obtain $\hat{\beta}_{WALD} = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0)$.

EXERCISE 10.7

(a) The approximate percentage difference is 100(5.9027 − 5.8916) = 1.11%.

(b) The *t*-value is *t* = (5.9027 − 5.8916) ÷ 0.00274 = 4.0510949. The two-tail *p*-value, using the normal distribution for the calculation, is 0.00005098. Thus, the difference in wage is statistically significant at any relevant level.

(c) The difference in years of schooling is 12.7969 − 12.6881 = 0.0988. The percentage difference is 100(12.7969 − 12.6881) ÷ 12.6881 = 0.77868239. That is, on average, those born in the fourth quarter of the year have slightly more than ¾% additional years of education than those born in the first quarter. Being born in the fourth quarter means that a child satisfies entrance age requirements for primary school earlier than someone born in the first quarter.

(d) The *t*-value is *t* = (12.7969 − 12.6881) ÷ 0.0132 = 7.4848485. The corresponding *p*-value is approximately 7.172×10^{-14}, or nearly zero. The difference is statistically very significant.

(e) The instrument being used is $Z = 1$ if an individual is born in the fourth quarter and $Z = 0$ if an individual is born in the first quarter. $\hat{\beta}_{2,WALD} = 0.11234818$. That is, we estimate approximately an additional 11% in expected wage for each additional year of schooling.

(f) Intuitively, the significant difference in average years of education makes the line connecting the two points more well defined.

EXERCISE 10.9

(a) Yes. In Example 10.5 years of education, experience, and its square are included the model. A notable omission is ability. If the omitted variable is positively correlated with the potentially endogenous variable, then the OLS estimator of the coefficient of education, β_4 in Example 10.5, will be overestimated. In the estimation results, education is shown to be strongly, and significantly, related to the ability variable. Consequently, we conclude that the OLS estimator of the effect of education is biased upwards. The OLS estimates are shown in Example 10.1. The estimated return to education is 10.75% whereas using IV estimation in Example 10.5, the estimated return to education falls to 6.14%. The OLS estimates attribute to education some of the positive effect of ability on wages.

Copyright © 2018 Wiley

(b) While the variable *ABILITY* may not be perfectly measure "ability", it is a measure related to whatever true ability means. The reduction in the coefficient of education occurs because the estimator of its effect no longer suffers from omitted variables bias.

(c) The first stage equation is

$$EDUC = \gamma_1 + \gamma_2 EXPER + \gamma_3 EXPER^2 + \gamma_4 ABILITY \\ + \theta_1 MOTHEREDUC + \theta_2 FATHEREDUC + v$$

(d) Yes. Both instruments are significant, which is good. The *F*-statistic of 33.82 is greater than the rule of thumb *F*-value = 10 for instruments that are not weak. Using the "test size" criterion in Appendix 10A.1 with $L = 2$ and $B = 1$, if we can tolerate a test size of 10% for a test at the 5% level of significance, the *F*-test value should exceed 19.93, which it does. Under this criterion we reject the null hypothesis that the instruments are weak.

(e) The *t*-value is the basis for the Hausman regression based test of endogeneity. Because the *t*-value is small, with a *p*-value of 0.347, we would fail to reject the null hypothesis that the variable $\hat{v}$ has no effect on $\ln(WAGE)$ and we would conclude that *EDUC* is exogenous, after controlling for ability.

EXERCISE 10.11

(a) Valid IV should (i) not have a direct effect on the outcome variable, here wage; (ii) they should not be correlated with the random error term, and (iii) they should be correlated with the potentially endogenous variable. Conditions (i) and (ii) are plausible for *NEAR4C* and *NEAR2C*. It is hard to imagine how where one lived at age 10 has a direct effect on wage. Similarly, if the omitted variable resulting in endogeneity is ability, then it is hard to imagine that simply living near a 2 or 4-year college as a 10 year old is correlated with ability. The final condition is one that we can check.

(b) Estimate the first stage equation with dependent variable *EDUC* and explanatory variables *EXPER* and its square, along with *NEARC4*, or *NEARC2*, or both. Calculate the least squares regression of the regression of interest, shown in Example 10.5, including the residuals from the first stage regression, $\hat{v}$, as an "explanatory" variable. Test the significance of this added variable using a standard *t*-test.

(c) Using only the IV *NEARC4* we fail to reject the null hypothesis that education is exogenous. Using *NEARC2*, or both *NEARC4* and *NEARC2* as IV, the test for endogeneity is weak. If we use the 5% level of significance we cannot reject the exogeneity of education. What could lead us to such a finding? First, the IV may be weak, and not strongly related to the potentially endogenous variable, *EDUC*. Secondly, perhaps they are not valid IV, in the sense that they are correlated with the regression error.

(d) The Sargan test for the validity of the surplus instruments, or over-identifying restrictions, uses NR^2 from the regression of the IV residuals on all exogenous variables including the IV. $NR^2 \sim \chi^2_{(1)}$ in this case if the IV are valid, because there are 2 IV but only one potentially endogenous variable. The value of $NR^2 = 1.2481535$ while the 5% critical value is 3.841.

Thus, we fail to reject the null hypothesis that the extra instrument is valid. If in fact the instruments are valid, and the exogenous variables are exogenous, then we should find no significance in this regression of the IV residuals on all exogenous variables.

(e) The statement is false. The OLS estimator is designed to obtain the best fitting regression line, by using the least squares principle.

(f) The *F*-statistic 7.89 falls below the rule of thumb threshold of 10, meaning that we cannot reject the null hypothesis that *NEARC4* and *NEARC2* are weak IV. When there is one potentially endogenous variable with two IV, the Stock-Yogo critical value is 7.25, if we are willing to accept a test size of 25% when carrying out a 5% test. That seems extreme. When IV are weak the standard error of the IV estimator tends to be large, as discussed on pp. 493-494. This will result in wider confidence intervals than those from the OLS estimator, which, while biased, is more precise.

EXERCISE 10.13

(a) This is a measurement error problem. We know that when the explanatory variable is measured with error that

$$\operatorname{cov}(x_i, e_i) = E(x_i e_i) = E\left[\left(x_i^* + u_i\right)\left(v_i - \beta_2 u_i\right)\right] = E\left(-\beta_2 u_i^2\right) = -\beta_2 \sigma_u^2 \neq 0$$

For the savings equation we expect the coefficient to be positive. So that $\operatorname{cov}(x_i, e_i) < 0$. Therefore,

$$b_2 \rightarrow \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} = \beta_2 + \frac{\operatorname{cov}(x, e)}{\operatorname{var}(x)} < \beta_2$$

The least squares estimator is downward biased and thus we anticipate the OLS estimate might understate the true marginal propensity to save. The negative, but insignificant, coefficient is not expected.

(b) This variable is likely to be correlated with current income, which is good. However, a concern might be that it is correlated with the measurement error. Current income (x_i) has measurement error u_i, where $x_i = x_i^* + u_i$. For *AVGINC* to be a valid IV it must be uncorrelated with u_i. Because it is an average value this may be defensible. We might wonder if it would have been better to use *AVGINC* in the equation rather than current income. The OLS bias would likely be smaller.

(c) The first stage *F*-test statistic for the significance of *AVGINC* is $5.8^2 = 33.64$. This is larger than the rule of thumb threshold 10 for an IV that is not weak. It is also larger than the Stock-Yogo critical value 16.38 using the 10% maximum test size criterion.

(d) This is the regression based Hausman test. The *t*-statistic for the coefficient of $\hat{v}$ is 3.75, which is greater than the approximate 0.01 critical value 2.58. We reject the null hypothesis that income is exogenous, and conclude that it is correlated with the regression error term. This means that the OLS estimator is biased and may not be a reliable basis for inference in this problem.

Copyright © 2018 Wiley

(e) The estimated coefficient is the two-stage least squares estimate. In this case, if the IV is valid, then the 2SLS estimator is consistent. The estimate of the marginal propensity to save is close to 4%. First, it is positive which is what we expected, and the magnitude is reasonable.

(f) No, the OLS standard errors are not correct. The proper estimator of the error variance is

$$\hat{\sigma}_{IV}^2 = \frac{\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right)^2}{N-2}$$

The two OLS standard errors use

$$\hat{\sigma}_{WRONG}^2 = \frac{\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 \hat{x}_i\right)^2}{N-2}$$

(g) Using the approximate critical value 1.96, the IV interval estimate is $[0, 0.0784]$. The OLS interval estimate is $[-0.02715, 0.01675]$. The IV interval estimate contains only positive values, so that we would be surprised if true marginal propensity to save was negative. The OLS estimated interval covers zero, and has a large range of negative values, which is not what we expect.

(h) It is not an accident. As noted the estimates in part (d) are the 2SLS estimates, and these are identical to those in (g) based on 2SLS/IV software. In the Hausman regression the inclusion of $\hat{v}$ into the regression serves to purpose of testing for endogeneity, and as noted in the discussion of the Hausman test logic, yields the IV estimates for the regression parameters.

(i) Unfortunately, the test for valid IV requires there to be surplus instruments. Here we have no surplus IV, so the test for validity cannot be carried out.

EXERCISE 10.15

(a) This result is shown on page 494, 3 lines from the bottom.

(b) The estimated covariance between the IV and the outcome variable is $\widehat{\text{cov}}\left(MOTHEREDUC_i, \ln\left(WAGE_i\right)\right) = 0.11279583$. The estimated covariance between the IV and the explanatory variable is $\widehat{\text{cov}}\left(MOTHEREDUC_i, EDUC_i\right) = 2.9259669$. The ratio is 0.03854993. This the same as the IV estimate, which, to more decimals than given in Example 10.2, is 0.03854993.

(c) The partial regressions are given in the table below in columns (1)-(3). Saving the residuals, we find

$$\widehat{\text{cov}}\left(RMOTHEREDUC_i, RLWAGE_i\right) = 0.14203353$$

$$\widehat{\text{cov}}\left(RMOTHEREDUC_i, REDUC_i\right) = 2.8831715$$

Their ratio is 0.04926295. The IV estimate of the coefficient of *EDUC*, in column (4) of the table above, to more decimal places, is 0.04926295.

Partial Regressions: 10.15(c)

	(1) *EDUC*	(2) *LWAGE*	(3) *MOTHEREDUC*	(4) *IV*
C	12.37***	0.808***	9.691***	0.198
	(0.322)	(0.100)	(0.464)	(0.471)
EXPER	0.0565	0.0476***	0.0285	0.0449***
	(0.0451)	(0.0140)	(0.0649)	(0.0135)
EXPER2	-0.00190	-0.00102*	-0.00233	-0.000922*
	(0.00135)	(0.000418)	(0.00194)	(0.000404)
EDUC				0.0493
				(0.0373)
N	428	428	428	428

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(d) The estimated values of β_2 is 0.0492629 with standard error 0.0372607, the same as the IV results when estimating the model in Example 10.5 using only *MOTHEREDUC* as an instrument.

EXERCISE 10.17

(a) The equation estimates are in Table XR10.17 column (1). This first stage equation is important because it shows that the instrument *MOTHEREDUC* is very significant 8.60, which corresponds to an F = 73.96, well above the rule-of-thumb value of 10 for an instrument that is not weak. $\sum REDUCHAT_i^2 = 1889.65843$.

Table XR10.17

	(1) part(a)	(2) part (b)	(3) part (c)	(4) part (d)
C	9.7751	12.3694	9.6912	
	(0.4239)	(0.1242)	(0.4640)	
EXPER	0.0489	0.0565	0.0285	
	(0.0417)	(0.0174)	(0.0649)	
EXPER2	-0.0013	-0.0019	-0.0023	
	(0.0012)	(0.0005)	(0.0019)	
MOTHEREDUC	0.2677			
	(0.0311)			
RMOM				0.2677
				(0.0000)
N	428	428	428	428
R^2	.1526941	.0322427	.0158082	1
SSE	1889.658	329.5579	4599.016	3.91e-11

Standard errors in parentheses

Copyright © 2018 Wiley

(b) These estimates are in column (2) of Table XR10.17. In this equation we have "partialed" out the included exogenous variables from the fitted value of *EDUC* from the first stage regression. The sum of squared errors is $\sum REDUC_i^2 = 329.5579$. This value is the amount of variation in $\widehat{EDUC}$ that is not explained by the included exogenous variables.

(c) These estimates are in column (3) of Table XR10.17. In this equation we have "partialed" out the included exogenous variables from the instrument *MOTHEREDUC*. The sum of squared residuals is $\sum RMOM_i^2 = 4599.01563$. This is the variation in *MOTHEREDUC* not explained by the included exogenous variables.

(d) These estimates are in column (4) of Table XR17.10. Note that the estimate of θ_1 is identical to the estimate from the first stage equation. The value of R^2 = 1.0, indicating that this relationship is a perfect fit. The sum of squared residuals is 0. Your computer software may report a number such as 3.9128×10^{-11}.

(e) As noted $\sum RMOM_i^2 = 4599.01563$. The value $\hat{\theta}_1 = 0.2676908$. The value of $\sum REDUC_i^2 = 329.5579$. Thus,

$$\hat{\theta}_1^2 \sum RMOM_i^2 = \left(0.2676908^2\right)\left(4599.01563\right) = \ 329.5579 = \sum REDUC_i^2$$

(f) Equation (10.25) makes the important point that the variance of the IV estimator equals the estimated error variance divided by the portion of the variation in $\widehat{EDUC}$ that is not explained by the included exogenous variables.

EXERCISE 10.19

(a) *COLLSUM* can take the values 0, 1, 2. In this sample of 428, 348 of the values are 0, 58 of the values are 1, and 22 of the values are 2. *COLLBOTH* can take the values 0 or 1, with 406 of the sample values zero and 22 of the values 1.

(b) The correlation between *EDUC* and *COLLSUM* is 0.4603. The correlation between *EDUC* and *COLLBOTH* is 0.3454. It might be the case that parents who have attended some college have an appreciation of college education that would make them more inclined to encourage their daughter to attend college. *COLLBOTH* might have a stronger effect because both parents have some college.

(c) These estimates are in column (1) of Table XR10.19. The 95% interval estimate of the coefficient of *EDUC* is [0.0264707, 0.1469456].

(d) The first stage equation is in column (2) of Table XR10.19. The *F*-test statistic for the null hypothesis that the coefficient of *COLLSUM* is zero is 113.357. This is far greater than the rule-of-thumb value of 10 for a strong instrument, and far greater than the Stock-Yogo critical value of 16.38 using the test size criterion, given that we are willing to accept a Type I error of 10% for a 5% test. We conclude that the instrument *COLLSUM* is not weak.

(e) The estimated regression is in column (3) of Table XR10.19. The coefficients are "similar" in that the coefficient of *MOTHERCOLL* is 1.749947 and the coefficient of *FATHERCOLL*

Copyright © 2018 Wiley

is 2.126612. The *t*-test of the equality of these coefficients is −0.81 with a *p*-value of 0.417. The *F*-statistic is 0.66 which has the same *p*-value. We cannot reject the null hypothesis that the coefficients of the two variables are equal, at the 5% level.

Table XR10.19

	(1) part(c)	(2) part (d)	(3) part (e)	(4) part (f)
C	-0.2650	11.8937	11.8903	-0.2791
	(0.3916)	(0.2901)	(0.2903)	(0.3904)
EDUC	0.0867			0.0878
	(0.0307)			(0.0306)
EXPER	0.0427	0.0499	0.0491	0.0427
	(0.0132)	(0.0401)	(0.0401)	(0.0132)
EXPER2	-0.0009	-0.0015	-0.0014	-0.0008
	(0.0004)	(0.0012)	(0.0012)	(0.0004)
MOTHERCOLL			1.7499	
			(0.3223)	
FATHERCOLL			2.1866	
			(0.3299)	
COLLSUM		1.9646		
		(0.1845)		
N	428	428	428	428
R^2	.1525289	.2148381	.2160603	.1529866
SSE	189.2636	1751.065	1748.339	189.1613
RMSE	.6649846	2.032208	2.033025	.664805

Standard errors in parentheses

(f) The model estimates using both *MOTHERCOLL* and *FATHERCOLL* as IV are in column (4) of Table XR10.19. The 95% interval estimate of the coefficient of *EDUC* is [0.027801, 0.1478943]. This interval estimate is slightly narrower than the one using *COLLSUM* only as the IV. The first stage *F*–value is 56.9629. This is far greater than the rule-of-thumb value of 10 for a strong instrument, and far greater than the Stock-Yogo critical value of 19.93 using the test size criterion, given that we are willing to accept a Type I error of 10% for a 5% test. We conclude that the instruments are not weak.

One advantage of having two rather than one IV is that we can test the validity of the surplus IV. The value of the Sargan $NR^2 = 0.237585$. The $\chi^2_{(1)}$ critical value is 3.841, thus we fail to reject the validity of the surplus IV. The question of whether we are better off using one or two instruments is open to debate. There are those who say that if you have only one potentially endogenous variable find the best one IV and use that. Others say it is useful to have a surplus IV, or two, so that the validity of the surplus IV can be tested. Here it makes little difference.

EXERCISE 10.21

The baseline model results appear in column (1) in Tables XR10.21a and XR10.21b.

Table XR10.21a

	(1) baseline	(2) part (a)	(3) part (c)	(4) red form
C	0.0481	9.1026	0.0599	9.0825
	(0.12)	(21.34)	(0.15)	(21.40)
EDUC	0.0614		0.0604	
	(1.96)		(1.93)	
EXPER	0.0442	0.0452	0.0442	0.0453
	(3.30)	(1.12)	(3.31)	(1.13)
EXPER2	-0.0009	-0.0010	-0.0009	-0.0010
	(-2.25)	(-0.84)	(-2.25)	(-0.84)
MOTHEREDUC		0.1576		
		(4.39)		
FATHEREDUC		0.1895		
		(5.62)		
PARENTSUM				0.1742
				(10.52)
N	428	428	428	428
R^2	.1357085	.2114706	.1348249	.2109634
SSE	193.02	1758.575	193.2173	1759.707
RMSE	.6715514	2.038967	.6718946	2.037217

t statistics in parentheses

Table XR10.21b

	(1) baseline	(2) IV	(3) red form
C	0.0481	-0.1722	11.9071
	(0.12)	(-0.48)	(16.25)
EDUC	0.0614	0.0792	
	(1.96)	(2.81)	
EXPER	0.0442	0.0432	0.0461
	(3.30)	(3.26)	(1.17)
EXPER2	-0.0009	-0.0009	-0.0011
	(-2.25)	(-2.18)	(-0.94)
MOTHEREDUC			-0.0768
			(-0.54)
MOMED2			0.0125
			(1.71)
FATHEREDUC			-0.2539
			(-2.05)
DADED2			0.0232
			(3.71)
N	428	428	428
R^2	.1357085	.1488718	.2550482
SSE	193.02	190.0803	1661.389
RMSE	.6715514	.6664179	1.986528

t statistics in parentheses

(a) The model is

$$EDUC = \gamma_1 + \gamma_2 EXPER + \gamma_3 EXPER^2 + \theta_1 MOTHEREDUC + \theta_2 FATHEREDUC + v$$

The estimated equation is in column (2) of Table XR10.21a. The test of the null hypothesis $\theta_1 = \theta_2$, against the two-tail alternative, yields a *t*-value of −0.52. We fail to reject the null hypothesis at the 5% level. We cannot conclude that the effects of mother's education and father's education on their daughter's education are different.

(b) The restricted model is

$$EDUC = \gamma_1 + \gamma_2 EXPER + \gamma_3 EXPER^2 + \theta\left(MOTHEREDUC + FATHEREDUC\right) + v$$

(c) The estimated equation is in column (3) of Table XR10.21a. There is only a very slight change in the IV estimates. The first stage equation is in column (4) of Table XR10.21a. The *F*-statistic value for the significance of *PARENTSUM* is 110.719, a value much larger than the rule of thumb threshold of 10 for a weak IV. It is larger than the Stock-Yogo critical value 16.38 using the test size criteria. If we will accept a 10% probability of a Type I error for a 5% test, then we can reject the null hypothesis the IV is weak.

(d) The estimation results for this equation are in column (3) of Table XR10.21b. The coefficient of *MOTHEREDUC* is not significant, and its square is significant at the 10% level. *FATHEREDUC* and its square are, however, significant. Multicollinearity is suspected for some of these results, as the squares are often highly correlated with the original variables. The correlation between *MOTHEREDUC* and its square is 0.975 and the correlation between *FATHEREDUC* and its square is 0.971. Such highly correlated variables will not add much as IV.

The joint *F*-test statistic for these 4 IV is 35.34. We reject the null hypothesis that the IV are weak using the rule-of-thumb threshold of 10. The Stock-Yogo critical value is 24.58 for the test size criterion (10% rejection rate of a 5% test) and is 16.85 for the relative bias criterion, in which the relative bias of the IV estimator is 5% of the OLS bias.

(e) The estimation results for this equation are in column (2) of Table XR10.21b. The estimated coefficient of *EDUC* is very slightly larger and has a slightly larger *t*-value. It is interesting that *MOTHEREDUC* and *FATHEREDUC* have negative coefficients while their squares have positive coefficients. This means that their marginal effects change. For example, the marginal effects of *FATHEREDUC* and *MOTHEREDUC* show an increasing marginal effect on daughter's education. In the baseline model the marginal effects of *MOTHEREDUC* and *FATHEREDUC* are constant.

(f) For the model with more instruments, the Sargan test statistic $NR^2 = 2.72416$, while $\chi^2_{(.95,3)} = 7.815$; thus, we fail to reject the validity of the surplus IV. In the model with more instruments there is collinearity and the IV estimates do not differ much from the baseline model. In both models the instruments are not weak. For IV estimation the baseline model is simpler than therefore preferred. If, however, the first stage equation is of independent interest, then an expanded specification might be useful to explore.

Copyright © 2018 Wiley

EXERCISE 10.23

(a)

$$E\left[\ln\left(WAGE_i\right)-\beta_1-\beta_2 EDUC_i-\beta_3 EXPER_i\right]=0$$

$$E\left[MOTHEREDUC_i\left(\ln\left(WAGE_i\right)-\beta_1-\beta_2 EDUC_i-\beta_3 EXPER_i\right)\right]=0$$

$$E\left[EXPER_i\left(\ln\left(WAGE_i\right)-\beta_1-\beta_2 EDUC_i-\beta_3 EXPER_i\right)\right]=0$$

(b) The IV estimates are in Table XR10.23.

Table XR 10.23

	(1) IV
C	0.3023
	(0.64)
EDUC	0.0542
	(1.46)
EXPER	0.0154
	(3.78)
N	428
R^2	.1178864
SSE	197.0002
SSR	26.32728

t statistics in parentheses

The empirical sums are

stats	$\hat{e}_{IV}$	$MOTHEREDUC_i \times \hat{e}_{IV}$	$EXPER_i \times \hat{e}_{IV}$
sum	2.18e-14	2.03e-13	-1.35e-13

The sums are zero because they are the empirical counterparts of the moment conditions, which replace the sample averages for the expected values, and the unknown parameters by the IV estimates.

$$\frac{1}{N}\sum\left[\ln\left(WAGE_i\right)-b_{1,IV}-b_{2,IV}EDUC_i-b_{3,IV}EXPER_i\right]=0$$

$$\frac{1}{N}\sum\left[MOTHEREDUC_i\left(\ln\left(WAGE_i\right)-b_{1,IV}-b_{2,IV}EDUC_i-b_{3,IV}EXPER_i\right)\right]=0$$

$$\frac{1}{N}\sum\left[EXPER_i\left(\ln\left(WAGE_i\right)-b_{1,IV}-b_{2,IV}EDUC_i-b_{3,IV}EXPER_i\right)\right]=0$$

(c) The empirical value of $\sum EDUC_i \times \hat{e}_{IV,i} = 123.1803$. The sum of squared IV residuals is $\sum \hat{e}_{IV,i}^2 = 197.0002$. In the OLS model $\sum EDUC_i \times \hat{e}_i = 0$ and $\sum \hat{e}^2 =$ 190.194984 . The OLS sum of squared residuals is smaller because that is the way OLS estimates are chosen, to minimize the sum of squared residuals.

Copyright © 2018 Wiley

(d) The sample averages are $\overline{\widehat{\ln(WAGE)}} =$ 1.190173 and $\overline{\ln(WAGE)} =$ 1.190173. They are the same because of the first empirical moment condition.

$$\frac{1}{N}\sum \ln(WAGE_i) - b_{1,IV} - b_{2,IV}\frac{1}{N}\sum EDUC_i - b_{3,IV}\frac{1}{N}\sum EXPER_i = 0$$
$$\Rightarrow$$
$$\overline{\ln(WAGE_i)} = b_{1,IV} + b_{2,IV}\overline{EDUC_i} + b_{3,IV}\overline{EXPER_i}$$

(e) $$SST = \sum\left[\ln(WAGE_i) - \overline{\ln(WAGE)}\right]^2 = 223.32744$$

$$SSE\_IV = \sum \hat{e}^2_{IV,i} = 197.00017$$

$$SSR\_IV = \sum\left[\widehat{\ln(WAGE)} - \overline{\ln(WAGE)}\right]^2 = 12.963922$$

$$SSR\_IV + SSE\_IV = \ 209.96409$$

$$R^2_{IV,1} = SSR\_IV\,/\,SST = 0.05804894$$

$$R^2_{IV,2} = 1 - SSE\_IV/SST = .11788644$$

The *SST* is the same as in Exercise 10.22. However, the sum of squared residuals is larger. Furthermore, *SST* computed directly does not equal the sum of the unexplained and explained sums of squares $SSR\_IV + SSE\_IV =$ 209.96409. The usual calculation of R^2 is different depending on how it is calculated because the decomposition of the total sum of squares into *SSE* + *SSR* no longer holds.

(f) Stata 15.1 reports $R^2 = 0.1179$, which is the second definition from the previous part. The usual calculation of R^2 is different depending on how it is calculated because the decomposition of the total sum of squares into *SSE* + *SSR* no longer holds.

EXERCISE 10.25

(a) The estimates are stored in Table XR10.25a column (a). The intercept is not statistically significant, supporting the strong hypothesis. The coefficient of *MONEY* is close to one, consistent with the quantity theory. The coefficient of *OUTPUT* is negative and somewhat close to negative one, and the 95% interval estimate covers −1, which is also consistent with the quantity theory. However, individual *t*-tests are not the same as a joint test.

Table XR10.25a Estimates

	Part (a) IV estimates		Part (d) IV estimates		Part (f) Robust IV	
C	-1.0940	(-0.5887)	-1.4977	(-1.1459)	-1.4977	(-1.3279)
MONEY	1.0351	(105.9139)	1.0570	(145.8307)	1.0570	(187.4520)
OUTPUT	-1.3942	(-2.5280)	-1.3496	(-3.4741)	-1.3496	(-4.0075)
N	76		75		75	

t statistics in parentheses

Copyright © 2018 Wiley

(b) The *F*-value for the strong hypothesis is 8.23. The test statistic has an approximate *F*-distribution in large samples if the null hypothesis is true. There are $J = 3$ hypotheses, the numerator degrees of freedom, and $N - K = 73$, the denominator degrees of freedom. The *p*-value is 0.0001. The 0.95 percentile of the *F*-distribution is 2.73. Thus, we reject the strong hypothesis.

The *F*-value for the weak hypothesis is 9.26. The test statistic has an approximate *F*-distribution if the null hypothesis is true. There are $J = 2$ hypotheses, the numerator degrees of freedom, and $N - K = 73$, the denominator degrees of freedom. The *p*-value is 0.0003. The 0.95 percentile of the *F*-distribution is 3.12. Thus, we reject the weak hypothesis.

(c) The largest absolute residual is 24.6332. The next largest is 9.5417.

(d) The estimates are given in Table XR10.25a. The estimates do not change a great deal, except for the intercept. The *t*-values increase, but the intercept is still not significantly different from zero. The *F*-statistics for the strong and weak hypotheses are 28.54 and 38.79, respectively. The critical values change only a small amount due to the loss of one denominator degree of freedom. Again, we reject both hypotheses.

(e) $NR^2 = 0.127$. The test critical value is 3.84. We find no evidence of heteroskedasticity using this test.

(f) The estimates and robust standard errors are in Table XR10.25a. The *F*-statistics for the strong and weak hypotheses are 48.63 and 65.22, respectively. Again, we reject both hypotheses.

(g) The estimates of the first stage are in Table XR10.25b. Using conventional standard errors and covariance matrix, the *F*-test of the joint significance of the IV is 4.60, with *p*-value 0.0024. Using robust standard errors, it is 3.23, with *p*-value 0.0173. In both cases we fail to reject the null hypothesis that the IV are weak.

Table XR10.25b

	First Stage Estimates		Robust First Stage Estimates		Part (h) Sargan test	
C	2.5959	(2.0881)	2.5959	(1.9621)	-0.9040	(-0.4397)
INITIAL	-0.2744	(-1.7381)	-0.2744	(-1.9311)	0.4040	(1.5479)
SCHOOL	-1.1900	(-1.3343)	-1.1900	(-1.1854)	-2.1791	(-1.4775)
INV	12.8730	(3.5520)	12.8730	(2.8231)	5.7067	(0.9522)
POPRATE	-0.4461	(-1.5821)	-0.4461	(-1.6596)	0.3374	(0.7237)
MONEY	-0.0041	(-0.9963)	-0.0041	(-1.1892)	0.0020	(0.2922)
N	75		75		75	
R^2	0.234		0.234		0.056	

t-statistics in parentheses

(h) The estimates are in the final column of Table XR10.25b. All the *t*-values are less than 1.645, the approximate critical value for the 10% level of significance. If the IV are valid we would expect none of the coefficients to be significant. The Sargan test statistic is $NR^2 = 4.20$. The 0.05 critical value for a chi-square distribution with 3 degrees of freedom is 7.81. Thus, we fail to reject the null hypothesis that the surplus IV are valid.

Copyright © 2018 Wiley