Chapter 8

Stochastic regressors and measurement errors

8.1 Overview

Until this point it has been assumed that the only random element in a regression model is the disturbance term. This chapter extends the analysis to the case where the variables themselves have random components. The initial analysis shows that in general OLS estimators retain their desirable properties. A random component attributable to measurement error, the subject of the rest of the chapter, is however another matter. While measurement error in the dependent variable merely inflates the variances of the regression coefficients, measurement error in the explanatory variables causes OLS estimates of the coefficients to be biased and invalidates standard errors, t tests, and F tests. The analysis is illustrated with reference to the Friedman permanent income hypothesis, the most celebrated application of measurement error analysis in the economic literature. The chapter then introduces instrumental variables (IV) estimation and gives an example of its use to fit the Friedman model. The chapter concludes with a description of the Durbin–Wu–Hausman test for investigating whether measurement errors are serious enough to warrant using IV instead of OLS.

8.2 Learning outcomes

After working through the corresponding chapter in the text, studying the corresponding slideshows, and doing the starred exercises in the text and the additional exercises in this subject guide, you should be able to:

- explain the conditions under which OLS estimators remain unbiased when the variables in the regression model possess random components
- derive the large-sample expression for the bias in the slope coefficient in a simple regression model with measurement error in the explanatory variable
- demonstrate, within the context of the same model, that measurement error in the dependent variable does not cause the regression coefficients to be biased but does increase their standard errors
- describe the Friedman permanent income hypothesis and explain why OLS estimates of a conventional consumption function will be biased if it is correct
- explain what is meant by an instrumental variables estimator and state the conditions required for its use

- demonstrate that the IV estimator of the slope coefficient in a simple regression model is consistent, provided that the conditions required for its use are satisfied
- explain the factors responsible for the population variance of the IV estimator of the slope coefficient in a simple regression model
- perform the Durbin–Wu–Hausman test in the context of measurement error.

8.3 Additional exercises

A8.1 A researcher believes that a variable Y is determined by the simple regression model:

$$Y = \beta_1 + \beta_2 X + u.$$

She thinks that X is not distributed independently of u but thinks that another variable, Z, would be a suitable instrument. The instrumental estimator of the intercept, $\widehat{\beta}_1^{\mathrm{IV}}$, is given by:

$$\widehat{\beta}_1^{\mathrm{IV}} = \overline{Y} - \widehat{\beta}_2^{\mathrm{IV}}\overline{X}$$

where $\widehat{\beta}_2^{\mathrm{IV}}$ is the IV estimator of the slope coefficient. [Exercise 8.12 in the textbook asks for a proof that $\widehat{\beta}_1^{\mathrm{IV}}$ is a consistent estimator of β_1.]

Explain, with a brief mathematical proof, why $\widehat{\beta}_1^{\mathrm{OLS}}$, the ordinary least squares estimator of β_1, would be inconsistent, if the researcher is correct in believing that X is not distributed independently of u.

The researcher has only 20 observations in her sample. Does the fact that $\widehat{\beta}_1^{\mathrm{IV}}$ is consistent guarantee that it has desirable small-sample properties? If not, explain how the researcher might investigate the small-sample properties.

A8.2 Suppose that the researcher in Exercise A8.1 is wrong and X is in fact distributed independently of u. Explain the consequences of using $\widehat{\beta}_1^{\mathrm{IV}}$ instead of $\widehat{\beta}_1^{\mathrm{OLS}}$ to estimate β_1.

Note: The population variance of $\widehat{\beta}_1^{\mathrm{IV}}$ is given by:

$$\sigma^2_{\widehat{\beta}_1^{\mathrm{IV}}} = \left(1 + \frac{\mu_X^2}{\sigma_X^2} \times \frac{1}{r_{XZ}^2}\right)\frac{\sigma_u^2}{n}$$

where μ_X is the population mean of X, σ_X^2 is its population variance, r_{XZ} is the correlation between X and Z, and σ_u^2 is the population variance of the disturbance term, u. For comparison, the population variance of the OLS estimator is:

$$\sigma^2_{\widehat{\beta}_1^{\mathrm{OLS}}} = \left(1 + \frac{\mu_X^2}{\sigma_X^2}\right)\frac{\sigma_u^2}{n}$$

when the model is correctly specified and the regression model assumptions are satisfied.

A8.3 A researcher investigating the incidence of teenage knife crime has the following data for each of 35 cities for 2008:

- K = number of knife crimes per 1,000 population in 2008
- N = number of teenagers per 1,000 population living in social deprivation in 2008.

The researcher hypothesises that the relationship between K and N is given by:

$$K = \beta_1 + \beta_2 N + u \qquad (1)$$

where u is a disturbance term that satisfies the usual regression model assumptions. However, knife crime tends to be under-reported, with the degree of under-reporting worst in the most heavily afflicted boroughs, so that:

$$R = K + w \qquad (2)$$

where R = number of reported knife crimes per 1,000 population in 2008 and w is a random variable with $E(w) < 0$ and $\text{cov}(w, K) < 0$. w may be assumed to be distributed independently of u. Note that $\text{cov}(w, K) < 0$ implies $\text{cov}(w, N) < 0$. Derive analytically the sign of the bias in the estimator of β_2 if the researcher regresses R on N using ordinary least squares.

A8.4 Suppose that in the model:

$$Y = \beta_1 + \beta_2 X + u$$

where the disturbance term u satisfies the regression model assumptions, the variable X is subject to measurement error, being underestimated by a fixed amount α in all observations.

- Discuss whether it is true that the ordinary least squares estimator of β_2 will be biased downwards by an amount proportional to both α and β_2.
- Discuss whether it is true that the fitted values of Y from the regression will be reduced by an amount $\alpha\beta_2$.
- Discuss whether it is true that R^2 will be reduced by an amount proportional to α.

A8.5 A researcher believes that the rate of migration from Country B to Country A, M_t, measured in thousands of persons per year, is a linear function of the relative average wage, RW_t, defined as the average wage in Country A divided by the average wage in Country B, both measured in terms of the currency of Country A:

$$M_t = \beta_1 + \beta_2 RW_t + u_t. \qquad (1)$$

u_t is a disturbance term that satisfies the regression model assumptions. However, Country B is a developing country with limited resources for statistical surveys and the wage data for that country, derived from a small sample of social security records, are widely considered to be unrepresentative, with a tendency to overstate the true average wage because those working in the informal sector are excluded. As a consequence the measured relative wage, MRW_t, is given by

$$MRW_t = RW_t + w_t \qquad (2)$$

where w_t is a random quantity with expected value less than 0. It may be assumed to be distributed independently of u_t and RW_t.

The researcher also has data on relative GDP per capita, $RGDP_t$, defined as the ratio of GDP per capita in countries A and B, respectively, both measured in terms of the currency of Country A. He has annual observations on M_t, MRW_t, and $RGDP_t$ for a 30-year period. The correlation between MRW_t, and $RGDP_t$ in the sample period is 0.8. Analyse mathematically the consequences for the estimates of the intercept and the slope coefficient, the standard errors and the t statistics, if the migration equation (1) is fitted:

- using ordinary least squares with MRW_t as the explanatory variable.
- using OLS, with $RGDP_t$ as a proxy for RW_t.
- using instrumental variables, with $RGDP_t$ as an instrument for MRW_t.

A8.6 Suppose that in Exercise A8.5 $RGDP_t$ is subject to the same kind of measurement error as RW_t, and that as a consequence there is an exact linear relationship between $RGDP_t$ and MRW_t. Demonstrate mathematically how this would affect the IV estimator of β_2 in part (3) of Exercise A8.5 and give a verbal explanation of your result.

8.4 Answers to the starred exercises in the textbook

8.5 A variable Q is determined by the model:

$$Q = \beta_1 + \beta_2 X + v$$

where X is a variable and v is a disturbance term that satisfies the regression model assumptions. The dependent variable is subject to measurement error and is measured as Y where:

$$Y = Q + r$$

and r is the measurement error, distributed independently of v. Describe analytically the consequences of using OLS to fit this model if:

1. The expected value of r is not equal to zero (but r is distributed independently of Q).
2. r is not distributed independently of Q (but its expected value is zero).

Answer:

Substituting for Q, the model may be rewritten:

$$\begin{aligned} Y &= \beta_1 + \beta_2 X + v + r \\ &= \beta_1 + \beta_2 X + u \end{aligned}$$

where $u = v + r$. Then:

$$\widehat{\beta}_2 = \beta_2 + \frac{\left(X_i - \overline{X}\right)(u_i - \bar{u})}{\sum \left(X_i - \overline{X}\right)^2} = \beta_2 + \frac{\sum \left(X_i - \overline{X}\right)(v_i - \bar{v}) + \sum \left(X_i - \overline{X}\right)(r_i - \bar{r})}{\sum \left(X_i - \overline{X}\right)^2}$$

and:

$$\begin{aligned}
E(\widehat{\beta}_2) &= E\left(\beta_2 + \frac{\sum\left(X_i-\overline{X}\right)(v_i-\bar{v})+\sum\left(X_i-\overline{X}\right)(r_i-\bar{r})}{\sum\left(X_i-\overline{X}\right)^2}\right)\\
&= \beta_2 + \frac{1}{\sum\left(X_i-\overline{X}\right)^2}E\left(\sum\left(X_i-\overline{X}\right)(v_i-\bar{v})+\sum\left(X_i-\overline{X}\right)(r_i-\bar{r})\right)\\
&= \beta_2 + \frac{1}{\sum\left(X_i-\overline{X}\right)^2}\left(\sum\left(X_i-\overline{X}\right)E(v_i-\bar{v})+\sum\left(X_i-\overline{X}\right)E(r_i-\bar{r})\right)\\
&= \beta_2
\end{aligned}$$

provided that X is nonstochastic. (If X is stochastic, the proof that the expected value of the error term is zero is parallel to that in Section 8.2 of the text.) Thus $\widehat{\beta}_2$ remains an unbiased estimator of β_2.

However, the estimator of the intercept is affected if $E(r)$ is not zero.

$$\widehat{\beta}_1 = \overline{Y} - \widehat{\beta}_2\overline{X} = \beta_1 + \beta_2\overline{X} + \bar{u} - \widehat{\beta}_2\overline{X} = \beta_1 + \beta_2\overline{X} + \bar{v} + \bar{r} - \widehat{\beta}_2\overline{X}.$$

Hence:

$$\begin{aligned}
E(\widehat{\beta}_1) &= \beta_1 + \beta_2\overline{X} + E(\bar{v}) + E(\bar{r}) - E(\widehat{\beta}_2\overline{X})\\
&= \beta_1 + \beta_2\overline{X} + E(\bar{v}) + E(\bar{r}) - \overline{X}E(\widehat{\beta}_2)\\
&= \beta_1 + E(\bar{r}).
\end{aligned}$$

Thus the intercept is biased if $E(r)$ is not equal to zero, for then $E(\bar{r})$ is not equal to 0.

If r is not distributed independently of Q, the situation is a little bit more complicated. For it to be distributed independently of Q, it must be distributed independently of both X and v, since these are the determinants of Q. Thus if it is not distributed independently of Q, one of these two conditions must be violated. We will consider each in turn.

(a) r not distributed independently of X. We now have:

$$\begin{aligned}
\operatorname{plim}\widehat{\beta}_2 &= \beta_2 + \frac{\operatorname{plim}\frac{1}{n}\sum\left(X_i-\overline{X}\right)(v_i-\bar{v})+\operatorname{plim}\frac{1}{n}\sum\left(X_i-\overline{X}\right)(r_i-\bar{r})}{\operatorname{plim}\frac{1}{n}\sum\left(X_i-\overline{X}\right)^2}\\
&= \beta_2 + \frac{\sigma_{Xr}}{\sigma_X^2}.
\end{aligned}$$

Since $\sigma_{Xr} \neq 0$, $\widehat{\beta}_2$ is an inconsistent estimator of β_2. It follows that $\widehat{\beta}_1$ will also be an inconsistent estimator of β_1:

$$\widehat{\beta}_1 = \beta_1 + \beta_2\overline{X} + \bar{v} + \bar{r} - \widehat{\beta}_2\overline{X}.$$

Hence:

$$\begin{aligned}\text{plim } \widehat{\beta}_1 &= \beta_1 + \beta_2\overline{X} + \text{ plim } \bar{v} + \text{ plim } \bar{r} - \overline{X} \text{ plim } \widehat{\beta}_2 \\ &= \beta_1 + \overline{X}(\beta_2 - \text{ plim } \widehat{\beta}_2)\end{aligned}$$

and this is different from β_1 if plim $\widehat{\beta}_2$ is not equal to β_2.

(b) r is not distributed independently of v. This condition is not required in the proof of the unbiasedness of either $\widehat{\beta}_1$ or $\widehat{\beta}_2$ and so both remain unbiased.

8.6 A variable Y is determined by the model:

$$Y = \beta_1 + \beta_2 Z + v$$

where Z is a variable and v is a disturbance term that satisfies the regression model conditions. The explanatory variable is subject to measurement error and is measured as X where:

$$X = Z + w$$

and w is the measurement error, distributed independently of v. Describe analytically the consequences of using OLS to fit this model if:

(1) the expected value of w is not equal to zero (but w is distributed independently of Z)

(2) w is not distributed independently of Z (but its expected value is zero).

Answer:

Substituting for Z, we have:

$$Y = \beta_1 + \beta_2(X - w) + v = \beta_1 + \beta_2 X + u$$

where $u = v - \beta_2 w$.

$$\widehat{\beta}_2 = \beta_2 + \frac{\sum \left(X_i - \overline{X}\right)(u_i - \bar{u})}{\sum \left(X_i - \overline{X}\right)^2}.$$

It is not possible to obtain a closed-form expression for the expectation of the error term since both its numerator and its denominator depend on w. Instead we take plims, having first divided the numerator and the denominator of the error term by n so that they have limits:

$$\begin{aligned}\text{plim } \widehat{\beta}_2 &= \beta_2 + \frac{\text{plim } \frac{1}{n}\sum \left(X_i - \overline{X}\right)(u_i - \bar{u})}{\text{plim } \frac{1}{n}\sum \left(X_i - \overline{X}\right)^2} \\ &= \beta_2 + \frac{\text{cov}(X, u)}{\text{var}(X)} = \beta_2 + \frac{\text{cov}([Z + w], [v - \beta_2 w])}{\text{var}(X)} \\ &= \beta_2 + \frac{\text{cov}(Z, v) - \beta_2\text{cov}(Z, w) + \text{cov}(w, v) - \beta_2\text{cov}(w, w)}{\text{var}(X)}.\end{aligned}$$

If $E(w)$ is not equal to zero, $\widehat{\beta}_2$ is not affected. The first three terms in the numerator are zero and:

$$\text{plim } \widehat{\beta}_2 = \beta_2 + \frac{-\beta_2\sigma_w^2}{\sigma_X^2}$$

so $\widehat{\beta}_2$ remains inconsistent as in the standard case. If w is not distributed independently of Z, then the second term in the numerator is not 0. $\widehat{\beta}_2$ remains inconsistent, but the expression is now:

$$\text{plim } \widehat{\beta}_2 = \beta_2 + \frac{-\beta_2(\sigma_{Zw} + \sigma_w^2)}{\sigma_X^2}.$$

The OLS estimator of the intercept is affected in both cases, but like the slope coefficient, it was inconsistent anyway.

$$\widehat{\beta}_1 = \overline{Y} - \widehat{\beta}_2\overline{X} = \beta_1 + \beta_2\overline{X} + \bar{u} - \widehat{\beta}_2\overline{X} = \beta_1 + \beta_2\overline{X} + \bar{v} - \beta_2\overline{w} - \widehat{\beta}_2\overline{X}.$$

Hence:

$$\text{plim } \widehat{\beta}_1 = \beta_1 + (\beta_2 - \text{ plim } \widehat{\beta}_2)\overline{X} + \text{ plim } \bar{v} - \beta_2 \text{ plim } \overline{w}.$$

In the standard case this would reduce to:

$$\begin{aligned}\text{plim } \widehat{\beta}_1 &= \beta_1 + (\beta_2 - \text{ plim } \widehat{\beta}_2)\overline{X} \\ &= \beta_1 + \beta_2\frac{\sigma_w^2}{\sigma_X^2}\overline{X}.\end{aligned}$$

If w has expected value μ_w, not equal to zero:

$$\text{plim } \widehat{\beta}_1 = \beta_1 + \beta_2\left(\frac{\sigma_w^2}{\sigma_X^2}\overline{X} - \mu_w\right).$$

If w is not distributed independently of Z:

$$\text{plim } \widehat{\beta}_1 = \beta_1 + \beta_2\frac{\sigma_{Zw} + \sigma_w^2}{\sigma_X^2}\overline{X}.$$

8.10 A researcher investigating the shadow economy using international crosssectional data for 25 countries hypothesises that consumer expenditure on shadow goods and services, Q, is related to total consumer expenditure, Z, by the relationship:

$$Q = \beta_1 + \beta_2 Z + v$$

where v is a disturbance term that satisfies the regression model assumptions. Q is part of Z and any error in the estimation of Q affects the estimate of Z by the same amount. Hence:

$$Y_i = Q_i + w_i$$

and:

$$X_i = Z_i + w_i$$

where Y_i is the estimated value of Q_i, X_i is the estimated value of Z_i, and w_i is the measurement error affecting both variables in observation i. It is assumed that the expected value of w is 0 and that v and w are distributed independently of Z and of each other.

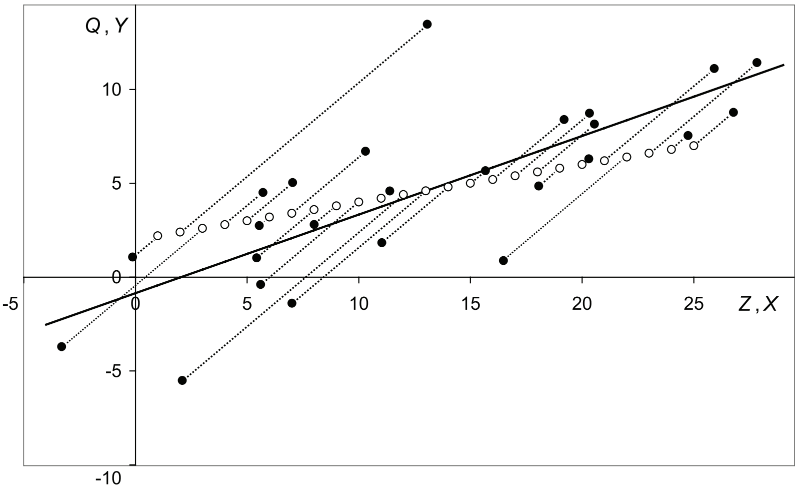
1. Derive an expression for the large-sample bias in the estimate of β_2 when OLS is used to regress Y on X, and determine its sign if this is possible. [Note: The standard expression for measurement error bias is not valid in this case.]
2. In a Monte Carlo experiment based on the model above, the true relationship between Q and Z is:

$$Q = 2.0 + 0.2Z.$$

A sample of 25 observations is generated using the integers 1, 2,..., 25 as data for Z. The variance of Z is 52.0. A normally distributed random variable with mean 0 and variance 25 is used to generate the values of the measurement error in the dependent and explanatory variables. The results with 10 samples are summarised in the table below. Comment on the results, stating whether or not they support your theoretical analysis.

Sample	$\widehat{\beta_1}$	s.e.$(\widehat{\beta_1})$	$\widehat{\beta_2}$	s.e.$(\widehat{\beta_2})$	R^2
1	−0.85	1.09	0.42	0.07	0.61
2	−0.37	1.45	0.36	0.10	0.36
3	−2.85	0.88	0.49	0.06	0.75
4	−2.21	1.59	0.54	0.10	0.57
5	−1.08	1.43	0.47	0.09	0.55
6	−1.32	1.39	0.51	0.08	0.64
7	−3.12	1.12	0.54	0.07	0.71
8	−0.64	0.95	0.45	0.06	0.74
9	0.57	0.89	0.38	0.05	0.69
10	−0.54	1.26	0.40	0.08	0.50

3. The figure below plots the points (Q, Z), represented as circles, and (Y, X), represented as solid markers, for the first sample, with each (Q, Z) point linked to the corresponding (Y, X) point. Comment on this graph, given your answers to parts 1 and 2.



Answer:

1. Substituting for Q and Z in the first equation:

$$(Y - w) = \beta_1 + \beta_2(X - w) + v.$$

Hence:

$$\begin{aligned} Y &= \beta_1 + \beta_2 X + v + (1-\beta_2)w \\ &= \beta_1 + \beta_2 X + u \end{aligned}$$

where $u = v + (1-\beta_2)w$. So:

$$\widehat{\beta}_2 = \beta_2 + \frac{\sum\left(X_i - \overline{X}\right)(u_i - \bar{u})}{\sum\left(X_i - \overline{X}\right)^2}.$$

It is not possible to obtain a closed-form expression for the expectation of the error term since both its numerator and its denominator depend on w. Instead we take plims, having first divided the numerator and the denominator of the error term by n so that they have limits:

$$\begin{aligned} \text{plim } \widehat{\beta}_2 &= \beta_2 + \frac{\text{plim } \frac{1}{n}\sum\left(X_i - \overline{X}\right)(u_i - \bar{u})}{\text{plim } \frac{1}{n}\sum\left(X_i - \overline{X}\right)^2} \\ &= \beta_2 + \frac{\text{cov}(X,u)}{\text{var}(u)} = \beta_2 + \frac{\text{cov}([Z+w],[v+(1-\beta_2)w])}{\text{var}(X)} \\ &= \beta_2 + \frac{\text{cov}(Z,v) + (1-\beta_2)\text{cov}(Z,w) + \text{cov}(w,v) + (1-\beta_2)\text{cov}(w,w)}{\text{var}(X)}. \end{aligned}$$

Since v and w are distributed independently of Z and of each other, $\text{cov}(Z,v) = \text{cov}(Z,w) = \text{cov}(w,v) = 0$, and so:

$$\text{plim } \widehat{\beta}_2 = \beta_2 + (1-\beta_2)\frac{\sigma_w^2}{\sigma_X^2}.$$

β_2 clearly should be positive and less than 1, so the bias is positive.

2. $\sigma_X^2 = \sigma_Z^2 + \sigma_w^2$, given that w is distributed independently of Z, and hence $\sigma_X^2 = 52 + 25 = 77$. Thus:

$$\text{plim } \widehat{\beta}_2 = 0.2 + \frac{(1-0.2)\times 25}{77} = 0.46.$$

The estimates of the slope coefficient do indeed appear to be distributed around this number.

As a consequence of the slope coefficient being overestimated, the intercept is underestimated, negative estimates being obtained in each case despite the fact that the true value is positive. The standard errors are invalid, given the severe problem of measurement error.

3. The diagram shows how the measurement error causes the observations to be displaced along 45° lines. Hence the slope of the regression line will be a compromise between the true slope, β_2 and 1. More specifically, plim $\widehat{\beta}_2$ is a

weighted average of β_2 and 1, the weights being proportional to the variances of Z and w:

$$\begin{aligned}\operatorname{plim} \widehat{\beta}_2 &= \beta_1 + (1-\beta_2)\frac{\sigma_w^2}{\sigma_Z^2+\sigma_w^2} \\ &= \frac{\sigma_Z^2}{\sigma_Z^2+\sigma_w^2}\beta_2 + \frac{\sigma_w^2}{\sigma_Z^2+\sigma_w^2}.\end{aligned}$$

8.16 It is possible that the *ASVABC* test score is a poor measure of the kind of ability relevant for earnings. Accordingly, perform an OLS regression of the logarithm of hourly earnings on *S*, *EXP*, *ASVABC*, *MALE*, *ETHBLACK*, and *ETHHISP* using your *EAWE* data set and an IV regression using *SM*, *SF*, and *SIBLINGS* as instruments for *ASVABC*. Perform a Durbin–Wu–Hausman test to evaluate whether *ASVABC* appears to be subject to measurement error.

Answer:

Contrary to expectations, the coefficient of *ASVABC* is lower in the IV regression. It is 0.048 in the OLS regression and −0.094 in the IV regression. The chi-squared statistic, 1.21, is low. One might therefore conclude that there is no serious measurement error and the change in the coefficient is random. Another possibility is that the instruments are too weak. *ASVABC* is not highly correlated with any of the instruments and the standard error of the coefficient rises from 0.028 in the OLS regression to 0.132 in the IV regression.

```
. ivreg LGEARN S EXP MALE ETHBLACK ETHHISP (ASVABC=SM SF SIBLINGS)
Instrumental variables (2SLS) regression
----------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =     500
-----------+------------------------------           F(  6,   493) =   22.29
     Model |   27.631679     6  4.60527983            Prob> F      =  0.0000
  Residual |  121.501359   493  .246453061           R-squared     =  0.1853
-----------+------------------------------           Adj R-squared =  0.1754
     Total |  149.133038   499  .298863804           Root MSE      =  .49644
----------------------------------------------------------------------------
    LGEARN |      Coef.   Std. Err.      t    P>|t|      [95% Conf. Interval]
-----------+----------------------------------------------------------------
    ASVABC |  -.0938253   .1319694    -0.71   0.477    -.3531172    .1654666
         S |   .1203265   .0251596     4.78   0.000     .0708931    .1697599
       EXP |   .0444094   .0092246     4.81   0.000      .026285    .0625338
      MALE |   .1909863   .0456252     4.19   0.000     .1013424    .2806302
  ETHBLACK |  -.1678914   .1355897    -1.24   0.216    -.4342963    .0985136
   ETHHISP |    .075698   .0828383     0.91   0.361    -.0870617    .2384576
     _cons |   .6503199   .3570741     1.82   0.069    -.0512548    1.351895
----------------------------------------------------------------------------
Instrumented:  ASVABC
Instruments:   S EXP MALE ETHBLACK ETHHISP SM SF SIBLINGS
----------------------------------------------------------------------------
```

```
. estimates store IV1

. reg LGEARN S EXP ASVABC MALE ETHBLACK ETHHISP
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =     500
-----------+------------------------------           F(  6,   493) =   23.81
     Model |  33.5095496     6  5.58492493            Prob> F      =  0.0000
  Residual |  115.623489   493  .234530403           R-squared     =  0.2247
-----------+------------------------------           Adj R-squared =  0.2153
     Total |  149.133038   499  .298863804           Root MSE      =  .48428
------------------------------------------------------------------------------
    LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
         S |   .0953713   .0106101     8.99   0.000     .0745246    .1162179
       EXP |    .043139   .0089279     4.83   0.000     .0255976    .0606805
    ASVABC |   .0477892   .0282877     1.69   0.092      -.00779    .1033685
      MALE |   .1954406   .0443323     4.41   0.000     .1083371    .2825441
  ETHBLACK |  -.0448382    .074738    -0.60   0.549    -.1916824     .102006
   ETHHISP |   .1226463   .0692577     1.77   0.077    -.0134303     .258723
     _cons |   .9766376   .1938648     5.04   0.000     .5957345    1.357541
------------------------------------------------------------------------------
```

```
. estimates store OLS1

. hausman IV1 OLS1, constant

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      IV1          OLS1        Difference          S.E.
-------------+----------------------------------------------------------------
      ASVABC |   -.0938253     .0477892       -.1416145        .1289021
           S |    .1203265     .0953713        .0249552         .022813
         EXP |    .0444094      .043139        .0012704        .0023208
        MALE |    .1909863     .1954406       -.0044543        .0107847
    ETHBLACK |   -.1678914    -.0448382       -.1230532        .1131318
     ETHHISP |     .075698     .1226463       -.0469484        .0454484
       _cons |    .6503199     .9766376       -.3263177        .2998639
------------------------------------------------------------------------------
                          b = consistent under Ho and Ha; obtained from ivreg
         B = inconsistent under Ha, efficient under Ho; obtained from regress
    Test:  Ho:  difference in coefficients not systematic
chi2(7) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        1.21
Prob>chi2 =      0.9908
```

```
. cor ASVABC SM SF SIBLINGS
(obs=500)
             |   ASVABC       SM       SF SIBLINGS
-------------+------------------------------------
      ASVABC |   1.0000
          SM |   0.3426   1.0000
          SF |   0.3613   0.5622   1.0000
    SIBLINGS |  -0.2360  -0.3038  -0.2516   1.0000
```

8.17 What is the difference between an instrumental variable and a proxy variable (as described in Section 6.4)? When would you use one and when would you use the other?

Answer:

An instrumental variable estimator is used when one has data on an explanatory variable in the regression model but OLS would give inconsistent estimates because the explanatory variable is not distributed independently of the disturbance term. The instrumental variable partially replaces the original explanatory variable in the estimator and the estimator is consistent.

A proxy variable is used when one has no data on an explanatory variable in a regression model. The proxy variable is used as a straight substitute for the original variable. The interpretation of the regression coefficients will depend on the relationship between the proxy and the original variable, and the properties of the other estimators in the model and the tests and diagnostic statistics will depend on the degree of correlation between the proxy and the original variable.

8.5 Answers to the additional exercises

A8.1

$$\begin{aligned}\widehat{\beta}_1^{\text{OLS}} &= \overline{Y} - \widehat{\beta}_2^{\text{OLS}}\overline{X} \\ &= \beta_1 + \beta_2\overline{X} + \bar{u} - \widehat{\beta}_2^{\text{OLS}}\overline{X}.\end{aligned}$$

Therefore:

$$\begin{aligned}\text{plim } \widehat{\beta}_1^{\text{OLS}} &= \beta_1 - (\text{plim } \widehat{\beta}_2^{\text{OLS}} - \beta_2) \text{ plim } \overline{X} \\ &\neq \beta_1.\end{aligned}$$

However:

$$\begin{aligned}\widehat{\beta}_1^{\text{IV}} &= \overline{Y} - \widehat{\beta}_2^{\text{IV}}\overline{X} \\ &= \beta_1 + \beta_2\overline{X} + \bar{u} - \widehat{\beta}_2^{\text{IV}}\overline{X} \\ &= \beta_1 - (\widehat{\beta}_2^{\text{IV}} - \beta_2)\overline{X} + \bar{u}.\end{aligned}$$

Therefore:

$$\begin{aligned}\text{plim } \widehat{\beta}_1^{\text{IV}} &= \beta_1 - (\text{plim } \widehat{\beta}_2^{\text{IV}} - \beta_2) \text{ plim } \overline{X} \\ &= \beta_1.\end{aligned}$$

Consistency does not guarantee desirable small-sample properties. The latter could be investigated with a Monte Carlo experiment.

A8.2 Both estimators will be consistent (actually, unbiased) but the IV estimator will be less efficient than the OLS estimator, as can be seen from a comparison of the expressions for the population variances.

A8.3 The regression model is:

$$R = \beta_1 + \beta_2 N + u + w.$$

Hence:

$$\widehat{\beta}_2^{\text{OLS}} = \beta_2 + \frac{\sum \left(N_i - \overline{N}\right)(u_i + w_i - \bar{u} - \overline{w})}{\sum \left(N_i - \overline{N}\right)^2}.$$

It is not possible to obtain a closed-form expression for the expectation since N and w are correlated. Hence, instead, we investigate the plim:

$$\begin{aligned}\text{plim } \widehat{\beta}_2^{\text{OLS}} &= \beta_2 + \text{ plim } \frac{\frac{1}{n}\sum \left(N_i - \overline{N}\right)(u_i + w_i - \bar{u} - \overline{w})}{\frac{1}{n}\sum \left(N_i - \overline{N}\right)^2} \\ &= \beta_2 + \frac{\text{cov}(N,u) + \text{cov}(N,w)}{\text{var}(N)} < \beta_2\end{aligned}$$

since $\text{cov}(N,u) = 0$ and $\text{cov}(N,w) < 0$.

A8.4 *Discuss whether it is true that the ordinary least squares estimator of β_2 will be biased downwards by an amount proportional to both α and β_2.*

It is not true. Let the measured X be X', where $X' = X - \alpha$. Then:

$$\widehat{\beta}_2^{\text{OLS}} = \frac{\sum (X_i' - X')\left(Y_i - \overline{Y}\right)}{\sum (X_i' - X')^2} = \frac{\sum \left(X_i - \alpha - [\overline{X} - \alpha]\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \alpha - [\overline{X} - \alpha]\right)^2} = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2}.$$

Thus the measurement error has no effect on the estimate of the slope coefficient.

Discuss whether it is true that the fitted values of Y from the regression will be reduced by an amount $\alpha\beta_2$.

The estimator of the intercept will be $\overline{Y} - \widehat{\beta}_2\overline{X}' = \overline{Y} - \widehat{\beta}_2(\overline{X} - \alpha)$. Hence the fitted value in observation i will be:

$$\overline{Y} - \widehat{\beta}_2(\overline{X} - \alpha) + \widehat{\beta}_2 X_i' = \overline{Y} - \widehat{\beta}_2(\overline{X} - \alpha) + \widehat{\beta}_2(X_i - \alpha) = \overline{Y} - \widehat{\beta}_2\overline{X} + \widehat{\beta}_2 X_i$$

which is what it would be in the absence of the measurement error.

Discuss whether it is true that R^2 will be reduced by an amount proportional to α.

Since R^2 is the variance of the fitted values of Y divided by the variance of the actual values, it will be unaffected.

A8.5 *Using ordinary least squares with MRW_t as the explanatory variable.*

$$\text{plim } \widehat{\beta}_2^{\text{OLS}} = \beta_2 - \beta_2 \frac{\sigma_w^2}{\sigma_{Rw}^2 + \sigma_w^2} = \beta_2 \frac{\sigma_{Rw}^2}{\sigma_{Rw}^2 + \sigma_w^2}$$

(standard theory). Hence the bias is towards zero.

$$\begin{aligned}\widehat{\beta}_1^{\text{OLS}} &= \overline{M} - \widehat{\beta}_2^{\text{OLS}}\overline{MRW} \\ &= \beta_1 + \beta_2\overline{RW} + \bar{u} - \widehat{\beta}_2^{\text{OLS}}\left(\overline{RW} + \overline{w}\right) \\ &= \beta_1 + (\beta_2 - \widehat{\beta}_2^{\text{OLS}})\overline{RW} + \bar{u} - \widehat{\beta}_2^{\text{OLS}}\overline{w}\end{aligned}$$

and so:

$$\text{plim } \widehat{\beta}_1^{\text{OLS}} = \beta_1 + \beta_2 \frac{\sigma_w^2}{\sigma_{Rw}^2 + \sigma_w^2} \overline{RW} - \beta_2 \frac{\sigma_{Rw}^2}{\sigma_{Rw}^2 + \sigma_w^2} \mu_w$$

where μ_w is the population mean of w. The first component of the bias will be positive and the second negative, given that μ_w is negative. It is not possible without further information to predict the direction of the bias. The standard errors and t statistics will be invalidated if there is substantial measurement error in *MRW*.

Using OLS, with $RGDP_t$ as a proxy for RW.

Suppose $RW = \alpha_1 + \alpha_2 RGDP$. Then the migration equation may be rewritten:

$$\begin{aligned} M_t &= \beta_1 + \beta_2(\alpha_1 + \alpha_2 RGDP_t) + u_t \\ &= (\beta_1 + \alpha_1\beta_2) + \alpha_2\beta_2 RGDP_t + u_t. \end{aligned}$$

In general it would not be possible to derive estimates of either β_1 or β_2. Likewise one has no information on the standard errors of either $\widehat{\beta}_1$ or $\widehat{\beta}_2$. Nevertheless the t statistic for the slope coefficient would be approximately equal to the t statistic in a regression of M on *RW*, if the proxy is a good one. R^2 will be approximately the same as it would have been in a regression of M on *RW*, if the proxy is a good one. One might hypothesise that *RGDP* might be approximately equal to *RW*, in which case $\alpha_1 = 0$ and $\alpha_2 = 1$ and one can effectively fit the original model.

Using instrumental variables, with RGDP_t *as an instrument for* MRW_t.

The IV estimator of β_2 is consistent:

$$\begin{aligned} \widehat{\beta}_2^{\text{IV}} &= \frac{\sum \left(M_i - \overline{M}\right)\left(RGDP_i - \overline{RGDP}\right)}{\sum \left(MRW_i - \overline{MRW}\right)\left(RGDP_i - \overline{RGDP}\right)} \\ &= \beta_2 + \frac{\sum (u_i - \beta_2 w_i - \bar{u} + \beta_2 \overline{w})\left(RGDP_i - \overline{RGDP}\right)}{\sum \left(MRW_i - \overline{MRW}\right)\left(RGDP_i - \overline{RGDP}\right)}. \end{aligned}$$

Hence plim $\widehat{\beta}_2^{\text{IV}} = \beta_2$ if u and w are distributed independently of *RGDP*. Likewise the IV estimator of $\widehat{\beta}_1$ is consistent:

$$\widehat{\beta}_1^{\text{IV}} = \overline{M} - \widehat{\beta}_2^{\text{IV}} \overline{MRW} = \beta_1 + \beta_2 \overline{RW} + \bar{u} - \widehat{\beta}_2^{\text{IV}} \overline{RW} - \widehat{\beta}_2^{\text{IV}} \overline{w}.$$

Hence:

$$\begin{aligned} \text{plim } \widehat{\beta}_1^{\text{IV}} &= \beta_1 + \beta_2 \overline{RW} + \text{ plim } \bar{u} - \text{ plim } \widehat{\beta}_2^{\text{IV}} \overline{RW} - \text{ plim } \widehat{\beta}_2^{\text{IV}} \text{ plim } \overline{w} \\ &= \beta_1 \end{aligned}$$

since plim $\widehat{\beta}_2^{\text{IV}} = \beta_2$ and plim $u =$ plim $w = 0$. The standard errors will be higher, and hence t statistics lower, than they would have been if it had been possible to run the original regression using OLS.

A8.6 Suppose $RGDP = \theta + \phi MRW$. Then:

$$\begin{aligned}
\widehat{\beta}_2^{\mathrm{IV}} &= \frac{\sum \left(M_i - \overline{M}\right)\left(RGDP_i - \overline{RGDP}\right)}{\sum \left(MRW_i - \overline{MRW}\right)\left(RGDP_i - \overline{RGDP}\right)} \\
&= \frac{\sum \left(M_i - \overline{M}\right)\left(\phi MRW_i - \phi\overline{MRW}\right)}{\sum \left(MRW_i - \overline{MRW}\right)\left(\phi MRW_i - \phi\overline{MRW}\right)} \\
&= \widehat{\beta}_2^{\mathrm{OLS}}.
\end{aligned}$$

The instrument is no longer valid because it is correlated with the measurement error.