Chapter 9
Simultaneous equations estimation

9.1 Overview

Until this point the analysis has been confined to the fitting of a single regression equation on its own. In practice, most economic relationships interact with others in a system of simultaneous equations, and when this is the case the application of ordinary least squares (OLS) to a single relationship in isolation yields biased estimates. Having defined what is meant by an endogenous variable, an exogenous variable, a structural equation, and a reduced form equation, the first objective of this chapter is to demonstrate this. The second is to show how it may be possible to use instrumental variables (IV) estimation, with exogenous variables acting as instruments for endogenous ones, to obtain consistent estimates of the coefficients of a relationship. The conditions for exact identification, underidentification, and overidentification are discussed. In the case of overidentification, it is shown how two-stage least squares can be used to obtain estimates that are more efficient than those obtained with simple IV estimation. The chapter concludes with a discussion of the problem of unobserved heterogeneity and the use of the Durbin–Wu–Hausman test in the context of simultaneous equations estimation.

9.2 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 what is meant by:
  - an endogenous variable
  - an exogenous variable
  - a structural equation
  - a reduced form equation
- explain why the application of OLS to a single equation in isolation is likely to yield inconsistent estimates of the coefficients if the equation is part of a simultaneous equations model
- derive an expression for the large-sample bias in the slope coefficient when OLS is used to fit a simple regression equation in a simultaneous equations model

- explain how consistent estimates of the coefficients of an equation in a simultaneous equations model might in principle be obtained using instrumental variables
- explain what is meant by exact identification, underidentification, and overidentification
- explain the principles underlying the use of two-stage least squares, and the reason why it is more efficient than simple IV estimation
- explain what is meant by the problem of unobserved heterogeneity
- perform the Durbin–Wu–Hausman test in the context of simultaneous equations estimation.

9.3 Further material

Good governance and economic development

In development economics it has long been observed that there is a positive association between economic performance, Y, and good governance, R, especially in developing countries. However, quantification of the relationship is made problematic by the fact that it is unlikely that causality is unidirectional. While good governance may contribute to economic performance, better performing countries may also develop better institutions. Hence in its simplest form one has a simultaneous equations mode:

$$Y = \beta_1 + \beta_2 R + u \qquad (1)$$

$$R = \alpha_1 + \alpha_2 Y + v \qquad (2)$$

where u and v are disturbance terms. Assuming that the latter are distributed independently, an OLS regression of the first equation will lead to an upwards biased estimate of β_2, at least in large samples. The proof is left as an exercise (Exercise A9.10). Thus to fit the first equation, one needs an instrument for R. Obviously a better-specified model would have additional explanatory variables in both equations, but there is a problem. In general any variable that influences R is also likely to influence Y and is therefore unavailable as an instrument.

In a study of 64 ex-colonial countries that is surely destined to become a classic, 'The colonial origins of comparative development: an empirical investigation', *American Economic Review* 91(5): 1369–1401, December 2001, Acemoglu, Johnson, and Robinson (henceforward AJR) argue that settler mortality rates provide a suitable instrument. Put simply, the thesis is that where mortality rates were low, European colonisers founded neo-European settlements with European institutions and good governance. Such settlements eventually prospered. Examples are the United States, Canada, Australia, and New Zealand. Where mortality rates were high, on account of malaria, yellow fever and other diseases for which Europeans had little or no immunity, settlements were not viable. In such countries the main objective of the coloniser was economic exploitation, especially of mineral wealth. Institutional development was not a consideration. Post-independence regimes have often been as predatory as their predecessors, indigenous rulers taking the place of the former colonisers. Think of the Belgian Congo, first exploited by King Leopold and more recently by Mobutu.

The study is valuable as an example of IV estimation in that it places minimal technical demands on the reader. There is nothing that would not be easily comprehensible to students in an introductory econometrics course that covers IV. Nevertheless, it gives careful attention to the important technical issues. In particular, it discusses at length the validity of the exclusion restriction. To use mortality as an instrument for R in the first equation, one must be sure that it is not a determinant of Y in its own right, either directly or indirectly (other than through R).

The conclusion of the study is surprising. According to theory (see Exercise A9.10), the OLS estimate of β_2 will be biased upwards by the endogeneity of R. The objective of the study was to demonstrate that the estimate remains positive and significant even when the upward bias has been removed by using IV. However, the IV estimate turns out to be higher than the OLS estimate. In fact it is nearly twice as large. AJR suggest that this is attributable to measurement error in the measurement of R. This would cause the OLS estimate to be biased downwards, and the bias would be removed (asymptotically) by the use of IV. AJR conclude that the downward bias in the OLS estimate caused by measurement error is greater than the upward bias caused by endogeneity.

9.4 Additional exercises

A9.1 In a certain agricultural country, aggregate consumption, C, is simply equal to 2,000 plus a random quantity z that depends upon the weather:

$$C = 2000 + z.$$

z has mean zero and standard deviation 100. Aggregate investment, I, is subject to a four-year trade cycle, starting at 200, rising to 300 at the top of the cycle, and falling to 200 in the next year and to 100 at the bottom of the cycle, rising to 200 again the year after that, and so on. Aggregate income, Y, is the sum of C and I:

$$Y = C + I.$$

Data on C and I, and hence Y, are given in the table. z was generated by taking normally distributed random numbers with mean zero and unit standard deviation and multiplying them by 100.

t	C	I	Y	t	C	I	Y
1	1,813	200	2,013	11	1,981	200	2,181
2	1,893	300	2,193	12	2,211	100	2,311
3	2,119	200	2,319	13	2,127	200	2,327
4	1,967	100	2,067	14	1,953	300	2,253
5	1,997	200	2,197	15	2,141	200	2,341
6	2,050	300	2,350	16	1,836	100	1,936
7	2,035	200	2,235	17	2,103	200	2,303
8	2,088	100	2,188	18	2,058	300	2,358
9	2,023	200	2,223	19	2,119	200	2,319
10	2,144	300	2,444	20	2,032	100	2,132

An orthodox economist regresses C on Y, using the data in the table, and obtains (standard errors in parentheses):

$$\begin{aligned} \widehat{C} &= \underset{(252)}{512} + \underset{(0.11)}{0.68}Y \qquad R^2 = 0.67 \\ & \qquad\qquad\qquad\qquad F = 36.49 \end{aligned}$$

Explain why this result was obtained, despite the fact that C does not depend on Y at all. In particular, comment on the t and F statistics.

A9.2 A small macroeconomic model of a closed economy consists of a consumption function, an investment function, and an income identity:

$$\begin{aligned} C_t &= \beta_1 + \beta_2 Y_t + u_t \\ I_t &= \alpha_1 + \alpha_2 r_t + v_t \\ Y_t &= C_t + I_t + G_t \end{aligned}$$

where C_t is aggregate consumer expenditure in year t, It is aggregate investment, G_t is aggregate current public expenditure, Y_t is aggregate output, and r_t is the rate of interest. State which variables in the model are endogenous and exogenous, and explain how you would fit the equations, if you could.

A9.3 The model is now expanded to include a demand for money equation and an equilibrium condition for the money market:

$$\begin{aligned} M_t^d &= \delta_1 + \delta_2 Y_t + \delta_3 r_t + w_t \\ M_t^d &= \overline{M}_t \end{aligned}$$

where M_t^d is the demand for money in year t and M_t is the supply of money, assumed exogenous. State which variables are endogenous and exogenous in the expanded model and explain how you would fit the equations, including those in Exercise A9.2, if you could.

A9.4 Table 9.2 reports a simulation comparing OLS and IV parameter estimates and standard errors for 10 samples. The reported R^2 (not shown in that table) for the OLS and IV regressions are shown in the table below.

Sample	OLS R^2	IV R^2
1	0.59	0.16
2	0.69	0.52
3	0.78	0.73
4	0.61	0.37
5	0.40	0.06
6	0.72	0.57
7	0.60	0.33
8	0.58	0.44
9	0.69	0.43
10	0.39	0.13

We know that, for large samples, the IV estimator is preferable to the OLS estimator because it is consistent, while the OLS estimator is inconsistent. However, do the smaller OLS standard errors in Table 9.2 and the larger OLS values of R^2 in the present table indicate that OLS is actually preferable for small samples ($n = 20$ in the simulation)?

A9.5 A researcher investigating the relationship between aggregate wages, W, aggregate profits, P, and aggregate income, Y, postulates the following model:

$$W = \beta_1 + \beta_2 Y + u \qquad (1)$$

$$P = \alpha_1 + \alpha_2 Y + \alpha_3 K + v \qquad (2)$$

$$Y = W + P \qquad (3)$$

where K is aggregate stock of capital and u and v are disturbance terms that satisfy the usual regression model assumptions and may be assumed to be distributed independently of each other. The third equation is an identity, all forms of income being classified either as wages or as profits. The researcher intends to fit the model using data from a sample of industrialised countries, with the variables measured on a per capita basis in a common currency. K may be assumed to be exogenous.

- Explain why ordinary least squares (OLS) would yield inconsistent estimates if it were used to fit (1) and derive the large-sample bias in the slope coefficient.
- Explain what can be inferred about the finite-sample properties of OLS if used to fit (1).
- Demonstrate mathematically how one might obtain a consistent estimate of β_2 in (1).
- Explain why (2) is not identified (underidentified).
- Explain whether (3) is identified.
- At a seminar, one of the participants asserts that it is possible to obtain an estimate of α_2 even though equation (2) is underidentified. Any change in income that is not a change in wages must be a change in profits, by definition, and so one can estimate α_2 as $(1 - \widehat{\beta}_2)$, where $\widehat{\beta}_2$ is the consistent estimate of β_2 found in the third part of this question. The researcher does not think that this is right but is confused and says that he will look into it after the seminar. What should he have said?

A9.6 A researcher has data on e, the annual average rate of growth of employment, x the annual average rate of growth of output, and p, the annual average rate of growth of productivity, for a sample of 25 countries, the average rates being calculated for the period 1995–2005 and expressed as percentages. The researcher hypothesises that the variables are related by the following model:

$$e = \beta_1 + \beta_2 x + u \qquad (1)$$

$$x = e + p. \qquad (2)$$

The second equation is an identity because p is defined as the difference between x and e. The researcher believes that p is exogenous. The correlation coefficient for x and p is 0.79.

- Explain why the OLS estimator of β_2 would be inconsistent, if the researcher's model is correctly specified. Derive analytically the large-sample bias, and state whether it is possible to determine its sign.
- Explain how the researcher might use p to construct an IV estimator of β_2, that is consistent if p is exogenous. Demonstrate analytically that the estimator is consistent.
- The OLS and IV regressions are summarised below (standard errors in parentheses). Comment on them, making use of your answers to the first two parts of this question.

$$\text{OLS} \quad \hat{e} = \underset{(0.27)}{-0.52} + \underset{(0.08)}{0.48}x \qquad (3)$$

$$\text{IV} \quad \hat{e} = \underset{(0.42)}{0.37} + \underset{(0.14)}{0.17}x \qquad (4)$$

- A second researcher hypothesises that both x and p are exogenous and that equation (2) should be written:

$$e = x - p. \qquad (5)$$

  On the assumption that this is correct, explain why the slope coefficients in (3) and (4) are both biased and determine the direction of the bias in each case.
- Explain what would be the result of fitting (5), regressing e on x and p.

A9.7 A researcher has data from the World Bank *World Development Report 2000* on F, average fertility (average number of children born to each woman during her life), M, under-five mortality (number of children, per 100, dying before reaching the age of 5), and S, average years of female schooling, for a sample of 54 countries. She hypothesises that fertility is inversely related to schooling and positively related to mortality, and that mortality is inversely related to schooling:

$$F = \beta_1 + \beta_2 S + \beta_3 M + u \qquad (1)$$

$$M = \alpha_1 + \alpha_2 S + v \qquad (2)$$

where u and v are disturbance terms that may be assumed to be distributed independently of each other. S may be assumed to be exogenous.

- Derive the reduced form equations for F and M.
- Explain what would be the most appropriate method to fit equation (1).
- Explain what would be the most appropriate method to fit equation (2).

The researcher decides to fit (1) using ordinary least squares, and she decides also to perform a simple regression of F on S, again using ordinary least squares, with the following results (standard errors in parentheses):

$$\hat{F} = \underset{(0.61)}{4.08} - \underset{(0.04)}{0.17}S + \underset{(0.003)}{0.015}M \qquad R^2 = 0.83 \qquad (3)$$

$$\hat{F} = \underset{(0.39)}{6.99} - \underset{(0.03)}{0.36}S \qquad R^2 = 0.71 \qquad (4)$$

- Explain why the coefficient of S differs in the two equations.
- Explain whether one may validly perform t tests on the coefficients of (4).

At a seminar someone hypothesises that female schooling may be negatively influenced by fertility, especially in the poorer developing countries in the sample, and this would affect (4). To investigate this, the researcher adds the following equation to the model:

$$S = \delta_1 + \delta_2 F + \delta_3 G + w \qquad (5)$$

where G is GNP per capita and w is a disturbance term. She regresses F on S (1) instrumenting for S with G (column (b) in the output below), and (2) using ordinary least squares, as in equation (4) (column (B) in the output below). The correlation between S and G was 0.70. She performs a Durbin–Wu–Hausman test to compare the coefficients.

```
             ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      IV           OLS         Difference          S.E.
-------------+----------------------------------------------------------------
           S |  -.2965323    -.3637397         .0672074        .0347484
       _cons |   6.162605     6.992907        -.8303019        .4194891
------------------------------------------------------------------------------
     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(  1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =     3.31
                          Prob>chi2 =      0.1158
```

- Discuss whether G is likely to be a valid instrument.
- What should the researcher's conclusions be with regard to the test?

A9.8 Aggregate demand Q_D for a certain commodity is determined by its price, P, aggregate income, Y, and population, POP:

$$Q_D = \beta_1 + \beta_2 P + \beta_3 Y + \beta_4 POP + u_D$$

and aggregate supply is given by:

$$Q_S = \alpha_1 + \alpha_2 P + u_S$$

where u_D and u_S are independently distributed disturbance terms.

- Demonstrate that the estimator of α_2 will be inconsistent if ordinary least squares (OLS) is used to fit the supply equation, showing that the large-sample bias is likely to be negative.
- Demonstrate that a consistent estimator of α_2 will be obtained if the supply equation is fitted using instrumental variables (IV), using Y as an instrument.

The model is used for a Monte Carlo experiment, with α_2 set equal to 0.2 and suitable values chosen for the other parameters. The table shows the estimates of

α_2 obtained in 10 samples using OLS, using IV with Y as an instrument, using IV with POP as an instrument, and using two-stage least squares (TSLS) with Y and POP. s.e. is standard error. The correlation between P and Y averaged 0.50 across the samples. The correlation between P and POP averaged 0.63 across the samples. Discuss the results obtained.

	OLS		IV with Y		IV with POP		TSLS	
	coef.	s.e.	coef.	s.e.	coef.	s.e.	coef.	s.e.
1	0.15	0.03	0.22	0.05	0.21	0.05	0.21	0.03
2	0.08	0.04	0.24	0.11	0.19	0.08	0.21	0.06
3	0.11	0.02	0.18	0.06	0.19	0.05	0.19	0.04
4	0.16	0.02	0.20	0.04	0.19	0.03	0.19	0.02
5	0.15	0.02	0.27	0.09	0.18	0.04	0.20	0.03
6	0.14	0.03	0.24	0.08	0.18	0.05	0.20	0.04
7	0.20	0.03	0.22	0.05	0.26	0.04	0.25	0.03
8	0.15	0.03	0.21	0.06	0.24	0.05	0.23	0.04
9	0.11	0.02	0.17	0.05	0.14	0.03	0.15	0.03
10	0.17	0.03	0.16	0.05	0.24	0.05	0.20	0.03

A9.9 A researcher has the following data for a sample of 1,000 manufacturing enterprises on the following variables, each measured as an annual average for the period 2001–2005: G, average annual percentage rate of growth of sales; R, expenditure on research and development; and A, expenditure on advertising. R and A are measured as a proportion of sales revenue. He hypothesises the following model:

$$G = \beta_1 + \beta_2 R + \beta_3 A + u_G \qquad (1)$$

$$R = \alpha_1 + \alpha_2 G + u_R \qquad (2)$$

where u_G and u_R are disturbance terms distributed independently of each other.

A second researcher believes that expenditure on quality control, Q, measured as a proportion of sales revenue, also influences the growth of sales, and hence that the first equation should be written:

$$G = \beta_1 + \beta_2 R + \beta_3 A + \beta_4 Q + u_G. \qquad (1*)$$

A and Q may be assumed to be exogenous variables.

- Derive the reduced form equation for G for the first researcher.
- Explain why ordinary least squares (OLS) would be an inconsistent estimator of the parameters of equation (2).
- The first researcher uses instrumental variables (IV) to estimate α_2 in (2). Explain the procedure and demonstrate that the IV estimator of α_2 is consistent.
- The second researcher uses two stage least squares (TSLS) to estimate α_2 in (2). Explain the procedure and demonstrate that the TSLS estimator is consistent.

- Explain why the TSLS estimator used by the second researcher ought to produce 'better' results than the IV estimator used by the first researcher, if the growth equation is given by (1*). Be specific about what you mean by 'better'.

- Suppose that the first researcher is correct and the growth equation is actually given by (1), not (1*). Compare the properties of the two estimators in this case.

- Suppose that the second researcher is correct and the model is given by (1*) and (2), but A is not exogenous after all. Suppose that A is influenced by G:

$$A = \gamma_1 + \gamma_2 G + u_A \qquad (3)$$

where u_A is a disturbance term distributed independently of u_G and u_R. How would this affect the properties of the IV estimator of α_2 used by the first researcher?

A9.10 A researcher has data for 100 workers in a large organisation on hourly earnings, *EARNINGS*, skill level of the worker, *SKILL*, and a measure of the intelligence of the worker, *IQ*. She hypothesises that *LGEARN*, the natural logarithm of *EARNINGS*, depends on *SKILL*, and that *SKILL* depends on *IQ*.

$$\begin{aligned} LGEARN &= \beta_1 + \beta_2 SKILL + u \qquad (1) \\ SKILL &= \alpha_1 + \alpha_2 IQ + v \qquad (2) \end{aligned}$$

where u and v are disturbance terms. The researcher is not sure whether u and v are distributed independently of each other.

- State, with a brief explanation, whether each variable is endogenous or exogenous, and derive the reduced form equations for the endogenous variables.

- Explain why the researcher could use ordinary least squares (OLS) to fit equation (1) if u and v are distributed independently of each other.

- Show that the OLS estimator of β_2 is inconsistent if u and v are positively correlated and determine the direction of the large-sample bias.

- Demonstrate mathematically how the researcher could use instrumental variables (IV) estimation to obtain a consistent estimate of β_2.

- Explain the advantages and disadvantages of using IV, rather than OLS, to estimate β_2, given that the researcher is not sure whether u and v are distributed independently of each other.

- Describe in general terms a test that might help the researcher decide whether to use OLS or IV. What are the limitations of the test?

- Explain whether it is possible for the researcher to fit equation (2) and obtain consistent estimates.

A9.11 This exercise relates to the Further material section.

In general in an introductory econometrics course, issues and problems are treated separately, one at a time. In practice in empirical work, it is common for multiple problems to be encountered simultaneously. When this is the case, the one-at-a-time analysis may no longer be valid. In the case of the AJR study, both endogeneity and measurement error seem to be issues. This exercise looks at both together, within the context of that model.

Let S be the correct good governance variable and let R be the measured variable, with measurement error w. Thus the model may be written:

$$\begin{aligned} Y &= \beta_1 + \beta_2 S + u \\ S &= \alpha_1 + \alpha_2 Y + v \\ R &= S + w. \end{aligned}$$

It may be assumed that w has zero expectation and constant variance σ_w^2 across observations, and that it is distributed independently of S and the disturbance terms in the equations in the model. Investigate the likely direction of the bias in the OLS estimator of β_2 in large samples.

9.5 Answers to the starred exercises in the textbook

9.1 A simple macroeconomic model consists of a consumption function and an income identity:

$$\begin{aligned} C &= \beta_1 + \beta_2 Y + u \\ Y &= C + I \end{aligned}$$

where C is aggregate consumption, I is aggregate investment, Y is aggregate income, and u is a disturbance term. On the assumption that I is exogenous, derive the reduced form equations for C and Y.

Answer:

Substituting for Y in the first equation:

$$C = \beta_1 + \beta_2(C + I) + u.$$

Hence:

$$C = \frac{\beta_1}{1-\beta_2} + \frac{\beta_2 I}{1-\beta_2} + \frac{u}{1-\beta_2}$$

and:

$$Y = C + I = \frac{\beta_1}{1-\beta_2} + \frac{I}{1-\beta_2} + \frac{u}{1-\beta_2}.$$

9.2 It is common to write an earnings function with the logarithm of the hourly wage as the dependent variable and characteristics such as years of schooling, cognitive ability, years of work experience, etc as the explanatory variables. Explain whether such an equation should be regarded as a reduced form equation or a structural equation.

Answer:

In the conventional model of the labour market, the wage rate and the quantity of labour employed are both endogenous variables jointly determined by the interaction of demand and supply. According to this model, the wage equation is a reduced form equation.

9.3 In the simple macroeconomic model:

$$\begin{aligned} C &= \beta_1 + \beta_2 Y + u \\ Y &= C + I \end{aligned}$$

described in Exercise 9.1, demonstrate that OLS would yield inconsistent results if used to fit the consumption function, and investigate the direction of the bias in the slope coefficient.

Answer:

The first step in the analysis of the OLS slope coefficient is to break it down into the true value and error component in the usual way:

$$\widehat{\beta}_2^{\text{OLS}} = \frac{\sum\left(Y_i - \overline{Y}\right)\left(C_i - \overline{C}\right)}{\sum\left(Y_i - \overline{Y}\right)^2} = \beta_2 + \frac{\sum\left(Y_i - \overline{Y}\right)\left(u_i - \bar{u}\right)}{\sum\left(Y_i - \overline{Y}\right)^2}.$$

From the reduced form equation in Exercise 9.1 we see that Y depends on u and hence we will not be able to obtain a closed-form expression for the expectation of the error term. Instead we take plims, having first divided the numerator and the denominator of the error term by n so that they will possess limits as n goes to infinity.

$$\text{plim } \widehat{\beta}_2^{\text{OLS}} = \beta_2 + \frac{\text{plim } \frac{1}{n}\sum\left(Y_i - \overline{Y}\right)\left(u_i - \bar{u}\right)}{\text{plim } \frac{1}{n}\sum\left(Y_i - \overline{Y}\right)^2} = \beta_2 + \frac{\text{cov}(Y, u)}{\text{var(Y)}}.$$

We next substitute for Y since it is an endogenous variable. We have two choices: we could substitute from the structural equation, or we could substitute from the reduced form. If we substituted from the structural equation, in this case the income identity, we would introduce another endogenous variable, C, and we would find ourselves going round in circles. So we must choose the reduced form.

$$\begin{aligned} \text{plim } \widehat{\beta}_2^{\text{OLS}} &= \beta_2 + \frac{\text{cov}\left(\left[\frac{\beta_1}{1-\beta_2} + \frac{I}{1-\beta_2} + \frac{u}{1-\beta_2}\right], u\right)}{\text{var}\left(\frac{\beta_1}{1-\beta_2} + \frac{I}{1-\beta_2} + \frac{u}{1-\beta_2}\right)} \\ &= \beta_2 + \frac{\frac{1}{1-\beta_2}(\text{cov}(I, u) + \text{cov}(u, u))}{\left(\frac{1}{1-\beta_2}\right)^2 \text{var}(I + u)} \\ &= \beta_2 + (1 - \beta_2)\frac{\text{cov}(I, u) + \text{var}(u)}{\text{var}(I) + \text{var}(u) + 2\text{cov}(I, u)}. \end{aligned}$$

On the assumption that I is exogenous, it is distributed independently of u and $\text{cov}(I, u) = 0$. So:

$$\text{plim } \widehat{\beta}_2^{\text{OLS}} = \beta_2 + (1 - \beta_2)\frac{\sigma_u^2}{\sigma_I^2 + \sigma_u^2}$$

since the sample variances tend to the population variances as the sample becomes large. Since the variances are positive, the sign of the bias depends on the sign of $(1 - \beta_2)$. It is reasonable to assume that the marginal propensity to consume is positive and less than 1, in which case this term will be positive and the large-sample bias in $\widehat{\beta}_2^{\text{OLS}}$ will be upwards.

The OLS estimate of the intercept is also inconsistent:

$$\widehat{\beta}_1^{\text{OLS}} = \overline{C} - \widehat{\beta}_2^{\text{OLS}}\overline{Y} = \beta_1 + \beta_2\overline{Y} + \bar{u} - \widehat{\beta}_2^{\text{OLS}}\overline{Y}.$$

Hence:

$$\begin{aligned}\text{plim } \widehat{\beta}_1^{\text{OLS}} &= \beta_1 + (\beta_2 - \text{ plim } \widehat{\beta}_2^{\text{OLS}}) \text{ plim } \overline{Y} \\ &= \beta_1 - (1 - \beta_2)\frac{\sigma_u^2}{\sigma_I^2 + \sigma_u^2} \text{ plim } \overline{Y}.\end{aligned}$$

This is evidently biased downwards, as one might expect, given that the slope coefficient was biased upwards.

9.6 The table gives consumption per capita, C, gross fixed capital formation per capita, I, and gross domestic product per capita, Y, all measured in US$, for 33 countries in 1998. The output from an OLS regression of C on Y, and an IV regression using I as an instrument for Y, are shown. Comment on the differences in the results.

	C	I	Y		C	I	Y
Australia	15,024	4,749	19,461	South Korea	4,596	1,448	6,829
Austria	19,813	6,787	26,104	Luxembourg	26,400	9,767	42,650
Belgium	18,367	5,174	24,522	Malaysia	1,683	873	3,268
Canada	15,786	4,017	20,085	Mexico	3,359	1,056	4,328
China–PR	446	293	768	Netherlands	17,558	4,865	24,086
China–HK	17,067	7,262	24,452	New Zealand	11,236	2,658	13,992
Denmark	25,199	6,947	32,769	Norway	23,415	9,221	32,933
Finland	17,991	4,741	24,952	Pakistan	389	79	463
France	19,178	4,622	24,587	Philippines	760	176	868
Germany	20,058	5,716	26,219	Portugal	8,579	2,644	9,976
Greece	9,991	2,460	11,551	Spain	11,255	3,415	14,052
Iceland	25,294	6,706	30,622	Sweden	20,687	4487	26,866
India	291	84	385	Switzerland	27,648	7,815	36,864
Indonesia	351	216	613	Thailand	1,226	479	1,997
Ireland	13,045	4,791	20,132	UK	19,743	4,316	23,844
Italy	16,134	4,075	20,580	USA	26,387	6,540	32,377
Japan	21,478	7,923	30,124				

```
. reg C Y
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =      33
-----------+------------------------------           F(  1,    31) = 1331.29
     Model |  2.5686e+09     1  2.5686e+09           Prob> F       =  0.0000
  Residual |  59810749.2    31  1929379.01           R-squared     =  0.9772
-----------+------------------------------           Adj R-squared =  0.9765
     Total |  2.6284e+09    32  82136829.4           Root MSE      =    1389
------------------------------------------------------------------------------
         C |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
         Y |   .7303066   .0200156    36.49   0.000     .6894845    .7711287
     _cons |   379.4871   443.6764     0.86   0.399     -525.397    1284.371
------------------------------------------------------------------------------


. ivregress 2sls C (Y=I)
-------------------------------------------------------------------------------
Instrumental variables (2SLS) regression             Number of obs =      33
                                                     Wald chi2(1)  = 1269.09
Prob> chi2    =  0.0000
                                                     R-squared     =  0.9770
                                                     Root MSE      =  1353.9
------------------------------------------------------------------------------
         C |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
         Y |   .7183909   .0201658    35.62   0.000     .6788667    .7579151
     _cons |    600.946   442.7386     1.36   0.175    -266.8057    1468.698
------------------------------------------------------------------------------
Instrumented:  Y
Instruments:   I
------------------------------------------------------------------------------
```

Answer:

Assuming the simple macroeconomic model:

$$C = \beta_1 + \beta_2 Y + u$$
$$Y = C + I$$

where C is consumption per capita, I is investment per capita, and Y is income per capita, and I is assumed exogenous, the OLS estimator of the marginal propensity to consume will be biased upwards. As was shown in Exercise 9.3:

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

Hence the IV estimate should be expected to be lower, but only by a small amount, given the data. With $\widehat{\beta}_2$ estimated at 0.72, $(1-\widehat{\beta}_2)$ is 0.28. σ_u^2 is estimated at 1.95 million and σ_I^2 is 7.74 million. Hence, on the basis of these estimates, the bias should be about 0.06. The actual difference in the OLS and IV estimates is smaller still. However, the actual difference would depend on the purely random sampling error as well as the bias, and it is possible that in this case the sampling error happens to have offset the bias to some extent.

9.11 Consider the price inflation/wage inflation model given by equations (9.1) and (9.2):

$$\begin{aligned} p &= \beta_1 + \beta_2 w + u_p \\ w &= \alpha_1 + \alpha_2 p + \alpha_3 U + u_w. \end{aligned}$$

We have seen that the first equation is exactly identified, U being used as an instrument for w. Suppose that TSLS is applied to this model, despite the fact that it is exactly identified, rather than overidentified. How will the results differ?

Answer:

If we fit the reduced form, we obtain a fitted equation:

$$\widehat{w} = h_1 + h_2 U.$$

The TSLS estimator is then given by

$$\begin{aligned} \widehat{\beta}_2^{\text{TSLS}} &= \frac{\sum \left(\widehat{w}_i - \overline{\widehat{w}} \right) (p_i - \bar{p})}{\sum \left(\widehat{w}_i - \overline{\widehat{w}} \right) (w_i - \overline{w})} = \frac{\sum \left(h_1 + h_2 U_i - h_1 - h_2 \overline{U} \right) (p_i - \bar{p})}{\sum \left(h_1 + h_2 U_i - h_1 - h_2 \overline{U} \right) (w_i - \overline{w})} \\ &= \frac{\sum h_2 \left(U_i - \overline{U} \right) (p_i - \bar{p})}{\sum h_2 \left(U_i - \overline{U} \right) (w_i - \overline{w})} = \widehat{\beta}_2^{\text{IV}} \end{aligned}$$

where $\widehat{\beta}_2^{\text{IV}}$ is the IV estimator using U. Hence the estimator is exactly the same. [Note: This is a special case of Exercise 8.18.]

9.15 Suppose the first equation in the model in Box 9.2 is fitted, with Q used as an instrument for Y. Describe the likely properties of the estimator of α_2.

Answer:

The first equation in Box 9.2 is:

$$X = \alpha_1 + \alpha_2 Y + u$$

The reduced form equation for Y is:

$$Y = \frac{1}{1 - \alpha_2 \beta_2} (\beta_1 + \alpha_1 \beta_2 + \beta_2 u + v).$$

Q is not a valid instrument for Y because it is not a determinant of Y. Mathematically, it can be shown that:

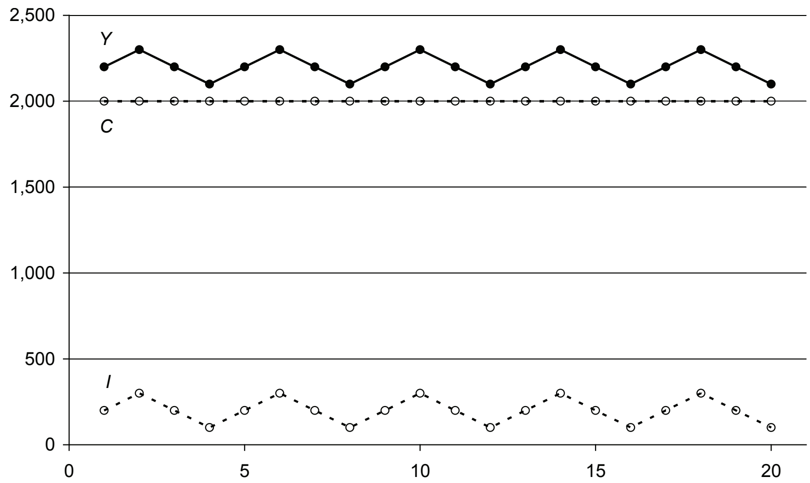
$$\text{plim } \widehat{\alpha}_2^{\text{IV}} = \alpha_2 + \frac{\text{cov}(Q, u)}{\text{cov}(Q, Y)}.$$

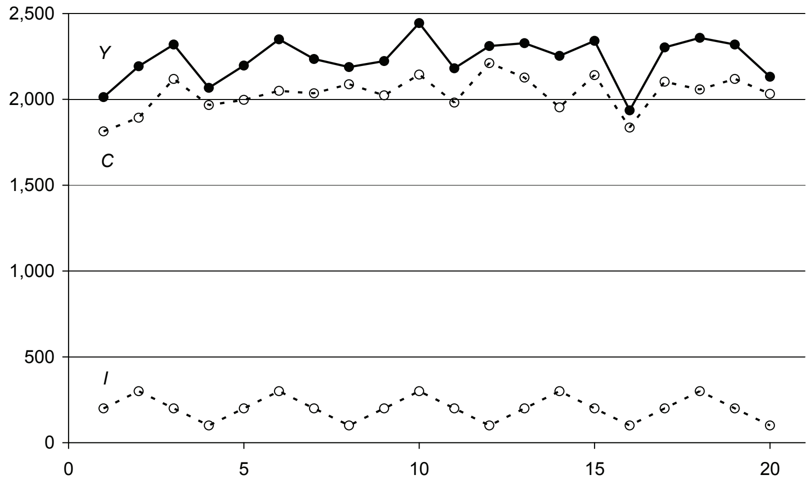
The numerator of the second term is zero, but so is its denominator and therefore the expression is undefined.

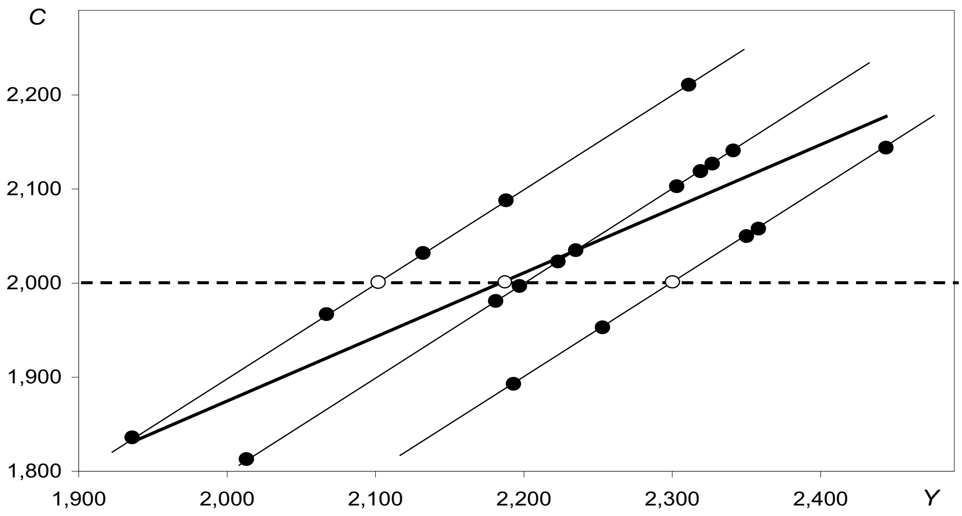
9.6 Answers to the additional exercises

A9.1 The positive coefficient of Y_t in the regression is attributable wholly to simultaneous equations bias. The three figures show this graphically.

The first diagram shows what the time series for C_t, I_t, and Y_t would look like if there were no random component of consumption. The series for C_t is constant at 2,000. That for I_t is a wave form, and that for Y_t is the same wave form shifted upward by 2,000. The second diagram shows the effect of adding the random component to consumption. Y_t still has a wave form, but there is a clear correlation between it and C_t.







In the third diagram, C_t is plotted against Y_t, with and without the random component. The three large circles represent the data when there is no random component. One circle represents the five data points [C = 2,000, Y = 2,100]; the middle circle represents the ten data points [C = 2,000, Y = 2,200]; and the other circle represents the five data points [C = 2,000, Y = 2,300]. A regression line based on these three points would be horizontal (the dashed line). The solid circles represent the 20 data points when the random component is affecting C_t and Y_t, and the solid line is the regression line for these points. Note that these 20 data points fall into three groups: five which lie on a 45 degree line through the left large circle, 10 which lie on the 45 degree line through the middle circle (actually, you can only see nine), and five on the 45 degree line through the right circle.

If OLS is used to fit the equation:

$$\widehat{\beta}_2^{\text{OLS}} = \frac{\sum\left(Y_i - \overline{Y}\right)\left(C_i - \overline{C}\right)}{\sum\left(Y_i - \overline{Y}\right)^2} = \frac{\sum\left(Y_i - \overline{Y}\right)\left([2000 + z_i] - [2000 + \bar{z}]\right)}{\sum\left(Y_i - \overline{Y}\right)^2} = \frac{\sum\left(Y_i - \overline{Y}\right)(z_i - \bar{z})}{\sum\left(Y_i - \overline{Y}\right)^2}.$$

Note that at this stage we have broken down the slope coefficient into its true value plus an error term. The true value does not appear explicitly because it is zero, so we only have the error term. We cannot take expectations because both the numerator and the denominator are functions of z:

$$Y = C + I = 2000 + I + z.$$

z is a component of C and hence of Y. As a second-best procedure, we investigate the large-sample properties of the estimator by taking plims. We must first divide the numerator and denominator by n so that they tend to finite limits:

$$\text{plim}\ \widehat{\beta}_2^{\text{OLS}} = \frac{\text{plim}\ \frac{1}{n}\sum\left(Y_i - \overline{Y}\right)(z_i - \bar{z})}{\text{plim}\ \frac{1}{n}\sum\left(Y_i - \overline{Y}\right)^2} = \frac{\text{cov}(Y, z)}{\text{var}(Y)}.$$

Substituting for Y from its reduced form equation:

$$\text{plim}\ \widehat{\beta}_2^{\text{OLS}} = \frac{\text{cov}([2000 + i + z], z)}{\text{var}(2000 + I + z)} = \frac{\text{cov}(I, z) + \text{var}(z)}{\text{var}(I) + \text{var}(z) + 2\text{cov}(I, z)} = \frac{\sigma_z^2}{\sigma_I^2 + \sigma_z^2}.$$

$\text{cov}(I, z) = 0$ because I is distributed independently of z. σ_z^2 is equal to 10,000 (since we are told that σ_z is equal to 100). Over a four-year cycle, the mean value of I is 200 and hence its population variance is given by:

$$\sigma_I^2 = \frac{1}{4}\left[0 + 100^2 + 0 + (-100)^2\right] = 5000.$$

Hence:

$$\text{plim } \widehat{\beta}_2^{\text{OLS}} = \frac{10000}{15000} = 0.67.$$

The actual coefficient in the 20-observation sample, 0.68, is very close to this (probably atypically close for such a model).

The estimator of the intercept, whose true value is 2,000, is biased downwards because $\widehat{\beta}_2^{\text{OLS}}$ is biased upwards. The standard errors of the coefficients are invalid because the regression model assumption B.7 is violated, and hence t tests would be invalid.

By virtue of the fact that $Y = C + I$, C is being regressed against a variable which is largely composed of itself. Hence the high R^2 is inevitable, despite the fact that there is no behavioural relationship between C and Y. Mathematically, R^2 is equal to the square of the sample correlation between the actual and fitted values of C. Since the fitted values of C are a linear function of the values of Y, R^2 is equal to the square of the sample correlation between C and Y. The population correlation coefficient is given by

$$\begin{aligned} \rho_{C,Y} &= \frac{\text{cov}(C,Y)}{\sqrt{\text{var}(C)\text{var}(Y)}} = \frac{\text{cov}\left([2000+z],[2000+I+z]\right)}{\sqrt{\text{var}\left([2000+z]\right)\text{var}\left([2000+I+z]\right)}} \\ &= \frac{\text{var}(z)}{\sqrt{\text{var}(z)\text{var}[I+z]}} = \frac{\sigma_z^2}{\sqrt{\sigma_z^2\left(\sigma_I^2+\sigma_z^2\right)}}. \end{aligned}$$

Hence in large samples:

$$R^2 = \frac{10000^2}{10000[10000+5000]} = 0.67.$$

R^2 in the regression is exactly equal to this, the closeness probably being something of a coincidence.

Since regression model assumption B.7 is violated, the F statistic cannot be used to perform an F test of goodness of fit.

A9.2 C_t, I_t, and Y_t are endogenous, the first two being the dependent variables of the behavioural relationships and the third being defined by an identity. G_t and r_t are exogenous.

Either I_t or r_t could be used as an instrument for Y_t in the consumption function. If it can be assumed that u_t and v_t are distributed independently, It can also be regarded as exogenous as far as the determination of C_t and Y_t are concerned. It would be preferable to r_t since it is more highly correlated with Y_t. One's first thought, then, would be to use TSLS, with the first stage fitting the equation:

$$Y_t = \frac{\beta_1}{1-\beta_2} + \frac{I_t}{1-\beta_2} + \frac{G_t}{1-\beta_2} + \frac{u_t}{1-\beta_2}.$$

Note, however, that the equation implies the restriction that the coefficients of I_t and G_t are equal. Hence all one has to do is to define a variable:

$$Z_t = I_t + G_t$$

and use Z_t as an instrument for Y_t in the consumption function.

The investment function would be fitted using OLS since r_t is exogenous. The income identity does not need to be fitted.

A9.3 M_t^d is endogenous because it is determined by the second of the two new relationships. The addition of the first of these relationships makes r_t endogenous. To see this, substituting for C_t and I_t in the income identity, using the consumption function and the investment function, one obtains:

$$Y_t = \frac{(\alpha_1 + \beta_1) + \alpha_2 r_t + u_t + v_t}{1 - \beta_2}.$$

This is usually known as the IS curve. Substituting for M_t^d in the first of the two new relationships, using the second, one has:

$$\overline{M}_t = \delta_1 + \delta_2 Y_t + \delta_3 r_t + w_t.$$

This is usually known as the LM curve. The equilibrium values of both Y_t and r_t are determined by the intersection of these two curves and hence r_t is endogenous as well as Y_t. G_t remains exogenous, as before, and M_t is also exogenous.

The consumption and investment functions are overidentified and one would use TSLS to fit them, the exogenous variables being government expenditure and the supply of money. The demand for money equation is exactly identified, two of the explanatory variables, r_t and Y_t, being endogenous, and the two exogenous variables being available to act as instruments for them.

A9.4 The OLS standard errors are invalid so a comparison is illegitimate. They are not of any great interest anyway because the OLS estimator is biased. Figure 9.3 in the text shows that the variance of the OLS estimator is smaller than that of the IV estimator, but, using a criterion such as the mean square error, there is no doubt that the IV estimator should be preferred. The comment about R^2 is irrelevant. OLS has a better fit but we have had to abandon the least squares principle because it yields inconsistent estimates.

A9.5 *Explain why ordinary least squares (OLS) would yield inconsistent estimates if it were used to fit (1) and derive the large-sample bias in the slope coefficient.*

At some point we will need the reduced form equation for Y. Substituting into the third equation from the first two, and rearranging, it is:

$$Y = \frac{1}{1 - \alpha_2 - \beta_2}(\alpha_1 + \beta_1 + \alpha_3 K + u + v).$$

Since Y depends on u, the assumption that the disturbance term be distributed independently of the regressors is violated in (1).

$$\widehat{\beta}_2^{\text{OLS}} = \frac{\sum \left(Y_i - \overline{Y}\right)\left(W_i - \overline{W}\right)}{\sum \left(Y_i - \overline{Y}\right)^2} = \beta_2 + \frac{\sum \left(Y_i - \overline{Y}\right)(u_i - \overline{u})}{\sum \left(Y_i - \overline{Y}\right)^2}$$

after substituting for W from (1) and simplifying. We are not able to obtain a closed-form expression for the expectation of the error term because u influences both its numerator and denominator, directly and by virtue of being a component of Y, as seen in the reduced form. Dividing both the numerator and denominator by n, and noting that:

$$\text{plim}\ \frac{1}{n}\sum\left(Y_i-\overline{Y}\right)^2=\text{var}(Y)$$

as a consequence of a law of large numbers, and that it can also be shown that:

$$\text{plim}\ \frac{1}{n}\sum\left(Y_i-\overline{Y}\right)(u_i-\bar{u})=\text{cov}(Y,u)$$

we can write

$$\text{plim}\ \widehat{\beta}_2^{\text{OLS}}=\beta_2+\frac{\text{plim}\ \frac{1}{n}\sum\left(Y_i-\overline{Y}\right)(u_i-\bar{u})}{\text{plim}\ \frac{1}{n}\sum\left(Y_i-\overline{Y}\right)^2}=\beta_2+\frac{\text{cov}(Y,u)}{\text{var}(Y)}.$$

Now:

$$\begin{aligned}\text{cov}(Y,u)&=\text{cov}\left(\frac{1}{1-\alpha_2-\beta_2}(\alpha_1+\beta_1+\alpha_3K+u+v),u\right)\\&=\frac{1}{1-\alpha_2-\beta_2}(\alpha_3\text{cov}(K,u)+\text{var}(u)+\text{cov}(v,u))\end{aligned}$$

the covariance of u with the constants being zero. Since K is exogenous, $\text{cov}(K,u)=0$. We are told that u and v are distributed independently of each other, and so $\text{cov}(u,v)=0$. Hence:

$$\text{plim}\ \widehat{\beta}_2^{\text{OLS}}=\beta_2+\frac{1}{1-\alpha_2-\beta_2}\frac{\sigma_u^2}{\text{plim var}(Y)}.$$

From the reduced form equation for Y it is evident that $(1-\alpha_2-\beta_2)>0$, and so the large-sample bias will be positive.

Explain what can be inferred about the finite-sample properties of OLS if used to fit (1).

It is not possible for an estimator that is unbiased in a finite sample to develop a bias if the sample size increases. Therefore, since the estimator is biased in large samples, it must also be biased in finite ones. The plim may well be a guide to the mean of the estimator in a finite sample, but this is not guaranteed and it is unlikely to be exactly equal to the mean.

Demonstrate mathematically how one might obtain a consistent estimate of β_2 in (1).

Use K as an instrument for Y:

$$\widehat{\beta}_2^{\text{IV}}=\frac{\sum\left(K_i-\overline{K}\right)\left(W_i-\overline{W}\right)}{\sum\left(K_i-\overline{K}\right)\left(Y_i-\overline{Y}\right)}=\beta_2+\frac{\sum\left(K_i-\overline{K}\right)(u_i-\bar{u})}{\sum\left(K_i-\overline{K}\right)\left(Y_i-\overline{Y}\right)}$$

after substituting for W from (1) and simplifying. We are not able to obtain a closed-form expression for the expectation of the error term because u influences both its numerator and denominator, directly and by virtue of being a component of Y, as seen in the reduced form. Dividing both the numerator and denominator by n, and noting that it can be shown that:

$$\operatorname{plim} \frac{1}{n} \sum \left(K_i - \overline{K}\right)(u_i - \bar{u}) = \operatorname{cov}(K, u) = 0$$

since K is exogenous, and that:

$$\operatorname{plim} \frac{1}{n} \sum \left(K_i - \overline{K}\right)\left(Y_i - \overline{Y}\right) = \operatorname{cov}(K, Y)$$

we can write:

$$\operatorname{plim} \widehat{\beta}_2^{\text{IV}} = \beta_2 + \frac{\operatorname{cov}(K, u)}{\operatorname{cov}(K, Y)} = \beta_2.$$

$\operatorname{cov}(K, Y)$ is non-zero since the reduced form equation for Y reveals that K is a determinant of Y. Hence the instrumental variable estimator is consistent.

Explain why (2) is not identified (underidentified).

(2) is underidentified because the endogenous variable Y is a regressor and there is no valid instrument to use with it. The only potential instrument is the exogenous variable K and it is already a regressor in its own right.

Explain whether (3) is identified.

(3) is an identity so the issue of identification does not arise.

At a seminar, one of the participants asserts that it is possible to obtain an estimate of α_2 even though equation (2) is underidentified. Any change in income that is not a change in wages must be a change in profits, by definition, and so one can estimate α_2 as $(1 - \widehat{\beta}_2)$, where $\widehat{\beta}_2$ is the consistent estimate of β_2 found in the third part of this question. The researcher does not think that this is right but is confused and says that he will look into it after the seminar. What should he have said?

The argument would be valid if Y were exogenous, in which case one could characterise β_2 and α_2 as being the effects of Y on W and P, holding other variables constant. But Y is endogenous, and so the coefficients represent only part of an adjustment process. Y cannot change autonomously, only in response to variations in K, u, or v.

The reduced form equations for W and P are:

$$\begin{aligned}
W &= \beta_1 + \frac{\beta_2}{1 - \alpha_2 - \beta_2}(\alpha_1 + \beta_1 + \alpha_3 K + u + v) + u \\
&= \frac{1}{1 - \alpha_2 - \beta_2}(\beta_1 + \alpha_1\beta_2 - \alpha_2\beta_1 + \alpha_3\beta_2 K + (1 - \alpha_2)u + \beta_2 v)
\end{aligned}$$

$$\begin{aligned}
P &= \alpha_1 + \frac{\alpha_2}{1 - \alpha_2 - \beta_2}(\alpha_1 + \beta_1 + \alpha_3 K + u + v) + \alpha_3 K + v \\
&= \frac{1}{1 - \alpha_2 - \beta_2}(\alpha_1 - \alpha_1\beta_2 + \alpha_2\beta_1 + \alpha_3(1 - \beta_2)K + \alpha_2 u + (1 - \beta_2)v).
\end{aligned}$$

Thus, for example, a change in K will lead to changes in W and P in the proportions $\beta_2 : (1 - \beta_2)$, not $\beta_2 : \alpha_2$. The same is true of changes caused by a variation in v. For a variation in u, the proportions would be $(1 - \alpha_2) : \alpha_2$.

A9.6 *Explain why the OLS estimator of β_2 would be inconsistent, if the researcher's model is correctly specified. Derive analytically the largesample bias, and state whether it is possible to determine its sign.*

The reduced form equation for x is:

$$x = \frac{\beta_1 + p + u}{1 - \beta_2}.$$

Thus:

$$\begin{aligned} \widehat{\beta}_2^{\text{OLS}} &= \frac{\sum(x_i - \bar{x})(e_i - \bar{e})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})(\beta_1 + \beta_2 x_i + u_i - \beta_1 - \beta_2\bar{x} - \bar{u})}{\sum(x_i - \bar{x})^2} \\ &= \beta_2 + \frac{\sum(x_i - \bar{x})(u_i - \bar{u})}{\sum(x_i - \bar{x})^2}. \end{aligned}$$

It is not possible to obtain a closed-form expression for the expectation of the estimator because the error term is a nonlinear function of u. Instead we investigate whether the estimator is consistent, first dividing the numerator and the denominator of the error term by n so that they tend to limits as the sample size becomes large.

$$\begin{aligned} \text{plim } \widehat{\beta}_2^{\text{OLS}} &= \beta_2 + \frac{\text{plim } \frac{1}{n}\sum\left(\frac{1}{1-\beta_2}[\beta_1 + p_i + u_i - \beta_1 - \bar{p} - \bar{u}]\right)(u_i - \bar{u})}{\text{plim } \frac{1}{n}\sum(x_i - \bar{x})^2} \\ &= \beta_2 + \frac{1}{1-\beta_2}\frac{\text{plim } \frac{1}{n}\sum(p_i - \bar{p})(u_i - \bar{u}) + \text{ plim } \frac{1}{n}\sum(u_i - \bar{u})^2}{\text{plim } \frac{1}{n}\sum(x_i - \bar{x})^2} \\ &= \beta_2 + \frac{1}{1-\beta_2}\frac{\text{cov}(p, u) + \text{var}(u)}{\text{var}(x)} = \beta_2 + \frac{1}{1-\beta_2}\frac{\sigma_u^2}{\sigma_x^2} \end{aligned}$$

since $\text{cov}(p, u) = 0$, p being exogenous. It is reasonable to assume that employment grows less rapidly than output, and hence β_2, and so $(1 - \beta_2)$, are less than 1. The bias is therefore likely to be positive.

Explain how the researcher might use p to construct an IV estimator of β_2 that is consistent if p is exogenous. Demonstrate analytically that the estimator is consistent.

p is available as an instrument, being exogenous, and therefore independent of u, being correlated with x, and not being in the equation in its own right.

$$\begin{aligned} \widehat{\beta}_2^{\text{IV}} &= \frac{\sum(p_i - \bar{p})(e_i - \bar{e})}{\sum(p_i - \bar{p})(x_i - \bar{x})} = \frac{\sum(p_i - \bar{p})(\beta_1 + \beta_2 x_i + u_i - \beta_1 - \beta_2\bar{x} - \bar{u})}{\sum(p_i - \bar{p})(x_i - \bar{x})} \\ &= \beta_2 + \frac{\sum(p_i - \bar{p})(u_i - \bar{u})}{\sum(p_i - \bar{p})(x_i - \bar{x})}. \end{aligned}$$

Hence, dividing the numerator and the denominator of the error term by n so that they tend to limits as the sample size becomes large,

$$\text{plim } \widehat{\beta}_2^{\text{IV}} = \beta_2 + \frac{\text{plim } \frac{1}{n}\sum(p_i - \bar{p})(u_i - \bar{u})}{\text{plim } \frac{1}{n}\sum(p_i - \bar{p})(x_i - \bar{x})} = \beta_2 + \frac{\text{cov}(p, u)}{\text{cov}(p, x)} = \beta_2$$

since $\text{cov}(p, u) = 0$, p being exogenous, and $\text{cov}(p, x) \neq 0$, x being determined partly by p.

The OLS and IV regressions are summarised below (standard errors in parentheses). Comment on them, making use of your answers to the first two parts of this question.

$$\begin{array}{lcll} \text{OLS} & \widehat{e} = & -0.52 + 0.48x & (3) \\ & & (0.27) \quad (0.08) & \\ \text{IV} & \widehat{e} = & 0.37 + 0.17x & (4) \\ & & (0.42) \quad (0.14) & \end{array}$$

The IV estimate of the slope coefficient is lower than the OLS estimate, as expected. The standard errors are not comparable because the OLS ones are invalid.

A second researcher hypothesises that both x and p are exogenous and that equation (2) should be written:

$$e = x - p. \qquad (5)$$

On the assumption that this is correct, explain why the slope coefficients in (3) and (4) are both biased and determine the direction of the bias in each case.

If (5) is correct, (3) is a misspecification that omits p and includes a redundant intercept. From the identity, the true values of the coefficients of x and p are 1 and -1, respectively. For (3):

$$E(\widehat{\beta}_2^{\text{OLS}}) = 1 - 1 \times \frac{\sum(x_i - \bar{x})(p_i - \bar{p})}{\sum(x_i - \bar{x})^2}.$$

x and p are positively correlated, so the bias will be downwards.

For (4):

$$\begin{aligned} \widehat{\beta}_2^{\text{IV}} &= \frac{\sum(p_i - \bar{p})(e_i - \bar{e})}{\sum(p_i - \bar{p})(x_i - \bar{x})} = \frac{\sum(p_i - \bar{p})([x_i - p_i] - [\bar{x} - \bar{p}])}{\sum(p_i - \bar{p})(x_i - \bar{x})} \\ &= 1 - \frac{\sum(p_i - \bar{p})^2}{\sum(p_i - \bar{p})(x_i - \bar{x})} = 1 - \frac{\frac{1}{n}\sum(p_i - \bar{p})^2}{\frac{1}{n}\sum(p_i - \bar{p})(x_i - \bar{x})}. \end{aligned}$$

Hence:

$$\text{plim } \widehat{\beta}_2^{\text{IV}} = 1 - \frac{\text{var}(p)}{\text{cov}(x, p)}$$

and so again the bias is downwards.

Explain what would be the result of fitting (5), regressing e on x and p.

One would obtain a perfect fit with the coefficient of x equal to 1, the coefficient of p equal to -1, and $R^2 = 1$.

A9.7 *Derive the reduced form equations for F and M.*

(2) is the reduced form equation for M. Substituting for M in (1), we have:

$$F = (\beta_1 + \alpha_1\beta_3) + (\beta_2 + \alpha_2\beta_3)S + u + \beta_3 v.$$

Explain what would be the most appropriate method to fit equation (1).

Since M does not depend on u, OLS may be used to fit (1).

Explain what would be the most appropriate method to fit equation (2).

There are no endogenous explanatory variables in (2), so again OLS may be used.

Explain why the coefficient of S differs in the two equations.

In (3), the coefficient is an estimate of the direct effect of S on fertility, controlling for M. In (4), the reduced form equation, it is an estimate of the total effect, taking account of the indirect effect via M (female education reduces mortality, and a reduction in mortality leads to a reduction in fertility).

Explain whether one may validly perform t tests on the coefficients of (4).

It is legitimate to use OLS to fit (4), so the t tests are valid.

Discuss whether G is likely to be a valid instrument.

G should be a valid instrument since it is highly correlated with S, it may reasonably be considered to be exogenous and therefore uncorrelated with the disturbance term in (4), and it does not appear in the equation in its own right (though perhaps it should).

What should the researchers conclusions be with regard to the test?

With 1 degree of freedom as indicated by the output, the critical value of chi-squared at the 5 per cent significance level is 3.84. Therefore we do not reject the null hypothesis of no significant difference between the estimates of the coefficients and conclude that there is no need to instrument for S. (4) should be preferred because OLS is more efficient than IV, when both are consistent.

A9.8 *Demonstrate that the estimate of α_2 will be inconsistent if ordinary least squares (OLS) is used to fit the supply equation, showing that the large-sample bias is likely to be negative.*

The reduced form equation for P is:

$$P = \frac{1}{\alpha_2 - \beta_2}(\beta_1 - \alpha_1 + \beta_3 Y + \beta_4 POP + u_D - u_S).$$

The OLS estimator of α_2 is:

$$\begin{aligned}\widehat{\alpha}_2^{\text{OLS}} &= \frac{\sum\left(P_i - \overline{P}\right)\left(Q_i - \overline{Q}\right)}{\sum\left(P_i - \overline{P}\right)^2} = \frac{\sum\left(P_i - \overline{P}\right)\left(\alpha_1 + \alpha_2 P_i + u_{Si} - \alpha_1 - \alpha_2\overline{P} - \bar{u}_S\right)}{\sum\left(P_i - \overline{P}\right)^2} \\ &= \alpha_2 + \frac{\sum\left(P_i - \overline{P}\right)(u_{Si} - \bar{u}_S)}{\sum\left(P_i - \overline{P}\right)^2}.\end{aligned}$$

We cannot take expectations because u_S is a determinant of both the numerator and the denominator of the error term, in view of the reduced form equation for P. Instead, we take probability limits, after first dividing the numerator and the denominator of the error term by n to ensure that limits exist.

$$\text{plim } \widehat{\alpha}_2^{\text{OLS}} = \alpha_2 + \frac{\text{plim } \frac{1}{n}\sum\left(P_i - \overline{P}\right)(u_{Si} - \bar{u}_S)}{\text{plim } \frac{1}{n}\sum\left(P_i - \overline{P}\right)^2} = \alpha_2 + \frac{\text{cov}(P, u_S)}{\text{var}(P)}.$$

Substituting from the reduced form equation for P:

$$\begin{aligned}\text{plim } \widehat{\alpha}_2^{\text{OLS}} &= \alpha_2 + \frac{\text{cov}\left(\frac{1}{\alpha_2-\beta_2}(\beta_1 - \alpha_1 + \beta_3 Y + \beta_4 POP + u_D - u_S), u_S\right)}{\text{var}(P)} \\ &= \alpha_2 - \frac{\frac{1}{\alpha_2-\beta_2}\text{var}(u_S)}{\text{var}(P)} = \alpha_2 - \frac{1}{\alpha_2 - \beta_2}\frac{\sigma_{u_S}^2}{\sigma_P^2}\end{aligned}$$

assuming that Y and POP are exogenous and so $\text{cov}(u_S, Y) = \text{cov}(u_S, POP) = 0$. We are told that u_S and u_D are distributed independently, so $\text{cov}(u_S, u_D) = 0$. Since it is reasonable to suppose that α_2 is positive and β_2 is negative, the large-sample bias will be negative.

Demonstrate that a consistent estimate of α_2 will be obtained if the supply equation is fitted using instrumental variables (IV), using Y as an instrument.

$$\begin{aligned}\widehat{\alpha}_2^{\text{IV}} &= \frac{\sum\left(Y_i - \overline{Y}\right)\left(Q_i - \overline{Q}\right)}{\sum\left(Y_i - \overline{Y}\right)\left(P_i - \overline{P}\right)} = \frac{\sum\left(Y_i - \overline{Y}\right)\left(\alpha_1 + \alpha_2 P_i + u_{Si} - \alpha_1 - \alpha_2\overline{P} - \bar{u}_S\right)}{\sum\left(Y_i - \overline{Y}\right)\left(P_i - \overline{P}\right)} \\ &= \alpha_2 + \frac{\sum\left(Y_i - \overline{Y}\right)(u_{Si} - \bar{u}_S)}{\sum\left(Y_i - \overline{Y}\right)\left(P_i - \overline{P}\right)}.\end{aligned}$$

We cannot take expectations because u_S is a determinant of both the numerator and the denominator of the error term, in view of the reduced form equation for P. Instead, we take probability limits, after first dividing the numerator and the denominator of the error term by n to ensure that limits exist.

$$\text{plim } \widehat{\alpha}_2^{\text{IV}} = \alpha_2 + \frac{\text{plim } \frac{1}{n}\sum\left(Y_i - \overline{Y}\right)(u_{Si} - \bar{u}_S)}{\text{plim } \frac{1}{n}\sum\left(Y_i - \overline{Y}\right)\left(P_i - \overline{P}\right)} = \alpha_2 + \frac{\text{cov}(Y, u)}{\text{cov}(Y, P)} = \alpha_2$$

since $\text{cov}(Y, u_s) = 0$ and $\text{cov}(P, Y) \neq 0$, Y being a determinant of P.

The model is used for a Monte Carlo experiment ... Discuss the results obtained.

- The OLS estimates are clearly biased downwards.
- The IV and TSLS estimates appear to be distributed around the true value, although one would need a much larger number of samples to be sure of this.
- The IV estimates with POP appear to be slightly closer to the true value than those with Y, as should be expected given the higher correlation, and the TSLS estimates appear to be slightly closer than either, again as should be expected.

- The OLS standard errors should be ignored. The standard errors for the IV regressions using *POP* tend to be smaller than those using Y, reflecting the fact that *POP* is a better instrument. Those for the TSLS regressions are smallest of all, reflecting its greater efficiency.

A9.9 *Derive the reduced form equation for G for the first researcher.*

$$G = \frac{1}{1-\alpha_2\beta_2}(\beta_1 + \alpha_1\beta_2 + \beta_3 A + u_G + \beta_2 u_R).$$

Explain why ordinary least squares (OLS) would be an inconsistent estimator of the parameters of equation (2).

The reduced form equation for G demonstrates that G is not distributed independently of the disturbance term u_R, a requirement for the consistency of OLS when fitting (2).

The first researcher uses instrumental variables (IV) to estimate α_2 in (2). Explain the procedure and demonstrate that the IV estimator of α_2 is consistent.

The first researcher would use A as an instrument for G. It is exogenous, so independent of u_R; correlated with G; and not in the equation in its own right. The estimator of the slope coefficient is:

$$\begin{aligned}\widehat{\alpha}_2^{\text{IV}} &= \frac{\sum\left(A_i-\overline{A}\right)\left(R_i-\overline{R}\right)}{\sum\left(A_i-\overline{A}\right)\left(G_i-\overline{G}\right)} = \frac{\sum\left(A_i-\overline{A}\right)\left([\alpha_1+\alpha_2 G_i+u_{Ri}]-[\alpha_1+\alpha_2\overline{G}=\overline{u}]\right)}{\sum\left(A_i-\overline{A}\right)\left(G_i-\overline{G}\right)} \\ &= \alpha_2 + \frac{\sum\left(A_i-\overline{A}\right)\left(u_{Ri}-\bar{u}_R\right)}{\sum\left(A_i-\overline{A}\right)\left(G_i-\overline{G}\right)}.\end{aligned}$$

Hence:

$$\text{plim}\ \widehat{\alpha}_2^{\text{IV}} = \alpha_2 + \ \text{plim}\ \frac{\frac{1}{n}\sum\left(A_i-\overline{A}\right)\left(u_{Ri}-\bar{u}_R\right)}{\frac{1}{n}\sum\left(A_i-\overline{A}\right)\left(G_i-\overline{G}\right)} = \alpha_2 + \frac{\text{cov}(A,u_R)}{\text{cov}(A,G)} = \alpha_2$$

since $\text{cov}(A,u_R)=0$, A being exogenous, and $\text{cov}(A,G)\neq 0$, A being a determinant of G.

The second researcher uses two stage least squares (TSLS) to estimate α_2 in (2). Explain the procedure and demonstrate that the TSLS estimator is consistent.

The reduced form equation for G for the second researcher is:

$$G = \frac{1}{1-\alpha_2\beta_2}(\beta_1 + \alpha_1\beta_2 + \beta_3 A + \beta_4 Q + u_G + \beta_2 u_R).$$

It is fitted using TSLS. The fitted values of G are used as the instrument:

$$\widehat{\alpha}_2^{\text{TSLS}} = \frac{\sum\left(\widehat{G}_i-\overline{\widehat{G}}\right)\left(R_i-\overline{R}\right)}{\sum\left(\widehat{G}_i-\overline{\widehat{G}}\right)\left(G_i-\overline{G}\right)}.$$

Following the same method as in the third part of the question:

$$\text{plim } \widehat{\alpha}_2^{\text{TSLS}} = \alpha_2 + \frac{\text{cov}(\widehat{G}, u_R)}{\text{cov}(\widehat{G}, G)} = \alpha_2$$

$\text{cov}(\widehat{G}, u_R)$ because $\widehat{G}$ is a linear combination of the exogenous variables, and $\text{cov}(\widehat{G}, G) \neq 0$.

Explain why the TSLS estimator used by the second researcher ought to produce 'better' results than the IV estimator used by the first researcher, if the growth equation is given by (1). Be specific about what you mean by 'better'.*

The TSLS estimator of α_2 should have a smaller variance. The variance of an IV estimator is inversely proportional to the square of the correlation of G and the instrument. $\widehat{G}$ is the linear combination of A and Q that has the highest correlation. It will therefore, in general, have a lower variance than the IV estimator using A.

Suppose that the first researcher is correct and the growth equation is actually given by (1), not (1). Compare the properties of the two estimators in this case.*

If the first researcher is correct, A is the optimal instrument because it will be more highly correlated with G (in the population) than the TSLS combination of A and Q and it will therefore be more efficient.

Suppose that the second researcher is correct and the model is given by (1) and (2), but A is not exogenous after all. Suppose that A is influenced by G:*

$$A = \gamma_1 + \gamma_2 G + u_A$$

where u_A is a disturbance term distributed independently of u_G and u_R. How would this affect the properties of the IV estimator of α_2 used by the first researcher?

$\text{cov}(A, u_R)$ would not be equal to 0 and so the estimator would be inconsistent.

A9.10 *State, with a brief explanation, whether each variable is endogenous or exogenous, and derive the reduced form equations for the endogenous variables.*

In this model *LGEARN* and *SKILL* are endogenous. *IQ* is exogenous. The reduced form equation for *LGEARN* is:

$$LGEARN = \beta_1 + \alpha_1\beta_2 + \alpha_2\beta_2 IQ + u + \beta_2 v.$$

The reduced form equation for *SKILL* is the structural equation.

Explain why the researcher could use ordinary least squares (OLS) to fit equation (1) if u and v are distributed independently of each other.

SKILL is not determined either directly or indirectly by u. Thus in equation (1) there is no violation of the requirement that the regressor be distributed independently of the disturbance term.

Show that the OLS estimator of β_2 is inconsistent if u and v are positively correlated and determine the direction of the large-sample bias.

Writing L for *LGEARN*, S for *SKILL*:

$$\begin{aligned}\widehat{\beta}_2^{\text{OLS}} &= \frac{\sum\left(S_i-\overline{S}\right)\left(L_i-\overline{L}\right)}{\sum\left(S_i-\overline{S}\right)^2} = \frac{\sum\left(S_i-\overline{S}\right)\left([\beta_1+\beta_2 S_i+u_i]-[\beta_1+\beta_2\overline{S}+\overline{u}]\right)}{\sum\left(S_i-\overline{S}\right)^2}\\ &= \beta_2+\frac{\sum\left(S_i-\overline{S}\right)(u_i-\overline{u})}{\sum\left(S_i-\overline{S}\right)^2}.\end{aligned}$$

We cannot obtain a closed-form expression for the expectation of the error term since S depends on v and v is correlated with u. Hence instead we take plims, dividing the numerator and the denominator by n to ensure that the limits exist:

$$\text{plim }\widehat{\beta}_2^{\text{OLS}} = \beta_2+\frac{\text{plim }\frac{1}{n}\sum\left(S_i-\overline{S}\right)(u_i-\overline{u})}{\text{plim }\frac{1}{n}\sum\left(S_i-\overline{S}\right)^2} = \beta_2+\frac{\text{cov}(S,u)}{\text{var}(S)}.$$

Now:

$$\text{cov}(S,u)=\text{cov}([\alpha_1+\alpha_2 IQ+v],u)=\text{cov}(v,u)$$

since α_1 is a constant and IQ is exogenous. Hence the numerator of the error term is positive in large samples. The denominator, being a variance, is also positive. So the large-sample bias is positive.

Demonstrate mathematically how the researcher could use instrumental variables (IV) estimation to obtain a consistent estimate of β_2.

The researcher could use IQ as an instrument for *SKILL*:

$$\begin{aligned}\widehat{\beta}_2^{\text{IV}} &= \frac{\sum\left(I_i-\overline{I}\right)\left(L_i-\overline{L}\right)}{\sum\left(I_i-\overline{I}\right)\left(S_i-\overline{S}\right)} = \frac{\sum\left(I_i-\overline{I}\right)\left([\beta_1+\beta_2 S_i+u_i]-[\beta_1+\beta_2\overline{S}+\overline{u}]\right)}{\sum\left(I_i-\overline{I}\right)\left(S_i-\overline{S}\right)}\\ &= \beta_2+\frac{\sum\left(I_i-\overline{I}\right)(u_i-\overline{u})}{\sum\left(I_i-\overline{I}\right)\left(S_i-\overline{S}\right)}.\end{aligned}$$

We cannot obtain a closed-form expression for the expectation of the error term since S depends on v and v is correlated with u. Hence instead we take plims, dividing the numerator and the denominator by n to ensure that the limits exist:

$$\text{plim }\widehat{\beta}_2^{\text{IV}} = \beta_2+\frac{\text{plim }\frac{1}{n}\sum\left(I_i-\overline{I}\right)(u_i-\overline{u})}{\text{plim }\frac{1}{n}\sum\left(I_i-\overline{I}\right)\left(S_i-\overline{S}\right)} = \beta_2+\frac{\text{cov}(I,u)}{\text{cov}(I,s)}.$$

The numerator of the error term is zero because I is exogenous. The denominator is not zero because S is determined by I. Hence the IV estimator is consistent.

Explain the advantages and disadvantages of using IV, rather than OLS, to estimate β_2, given that the researcher is not sure whether u and v are distributed independently of each other.

The advantage of IV is that, being consistent, there will be no bias in large samples and hence one may hope that there is no serious bias in a finite sample. One disadvantage is that there is a loss of efficiency if u and v are independent. Even if they are not independent, the IV estimator may be inferior to the OLS estimator using some criterion such as the mean square error that allows a trade-off between the bias of an estimator and its variance.

Describe in general terms a test that might help the researcher decide whether to use OLS or IV. What are the limitations of the test?

Durbin–Wu–Hausman test. Also known as Hausman test. The test statistic is a chi-squared statistic based on the differences of all the coefficients in the regression. The null hypothesis is that *SKILL* is distributed independently of u and the differences in the coefficients are random. If the test statistic exceeds its critical value, given the significance level of the test, we reject the null hypothesis and conclude that we ought to use IV rather than OLS. The main limitation is lack of power if the instrument is weak.

Explain whether it is possible for the researcher to fit equation (2) and obtain consistent estimates.

There is no reason why the equation should not be fitted using OLS.

A9.11 Substituting for Y from the first equation into the second, and re-arranging, we have the reduced form equation for S:

$$S = \frac{\alpha_1 + \alpha_2\beta_1 + v + \alpha_2 u}{1 - \alpha_2\beta_2}.$$

Substituting from the third equation into the first, we have:

$$Y = \beta_1 + \beta_2 R + u - \beta_2 w.$$

If this equation is fitted using OLS, we have:

$$\begin{aligned}\text{plim } \widehat{\beta}_2^{\text{OLS}} &= \beta_2 + \frac{\text{cov}(R, [u - \beta_2 w])}{\text{var}(R)} = \beta_2 + \frac{\text{cov}([S + w], [u - \beta_2 w])}{\text{var}(S + w)} \\ &= \beta_2 + \frac{\alpha_2\gamma\sigma_u^2 - \beta_2\sigma_w^2}{\sigma_S^2 + \sigma_w^2} = \beta_2 + \frac{\alpha_2\gamma\sigma_u^2 - \beta_2\sigma_w^2}{\gamma^2(\sigma_v^2 + \alpha_2^2\sigma_u^2) + \sigma_w^2}\end{aligned}$$

where:

$$\gamma = \frac{1}{1 - \alpha_2\beta_2}.$$

The denominator of the bias term is positive. Hence the bias will be positive if (the component attributable to simultaneity) is greater than (the component attributable to measurement error), and negative if it is smaller.