Chapter 3: Multiple regression analysis

Overview

This chapter introduces regression models with more than one explanatory variable. Specific topics are treated with reference to a model with just two explanatory variables, but most of the concepts and results apply straightforwardly to more general models. The chapter begins by showing how the least squares principle is employed to derive the expressions for the regression coefficients and how the coefficients should be interpreted. It continues with a discussion of the precision of the regression coefficients and tests of hypotheses relating to them. Next comes multicollinearity, the problem of discriminating between the effects of individual explanatory variables when they are closely related. The chapter concludes with a discussion of *F* tests of the joint explanatory power of the explanatory variables or subsets of them, and shows how a *t* test can be thought of as a marginal *F* test.

Learning outcomes

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

- the principles behind the derivation of multiple regression coefficients (but you are not expected to learn the expressions for them or to be able to reproduce the mathematical proofs)
- how to interpret the regression coefficients
- the Frisch–Waugh–Lovell graphical representation of the relationship between the dependent variable and one explanatory variable, controlling for the influence of the other explanatory variables
- the properties of the multiple regression coefficients
- what factors determine the population variance of the regression coefficients
- what is meant by multicollinearity
- what measures may be appropriate for alleviating multicollinearity
- what is meant by a linear restriction
- the *F* test of the joint explanatory power of the explanatory variables
- the *F* test of the explanatory power of a group of explanatory variables
- why *t* tests on the slope coefficients are equivalent to marginal *F* tests.

You should know the expression for the population variance of a slope coefficient in a multiple regression model with two explanatory variables.

Additional exercises

A3.1

The output shows the result of regressing *FDHO*, expenditure on food consumed at home, on *EXP*, total household expenditure, and *SIZE*, number of persons in the household, using the *CES* data set. Provide an interpretation of the regression coefficients and perform appropriate tests.

```
. reg FDHO EXP SIZE if FDHO>0

      Source |       SS       df       MS              Number of obs =     868
-------------+------------------------------           F(  2,   865) =  426.78
       Model |  1.4826e+09     2   741314291           Prob > F      =  0.0000
    Residual |  1.5025e+09   865  1736978.64           R-squared     =  0.4967
-------------+------------------------------           Adj R-squared =  0.4955
       Total |  2.9851e+09   867  3443039.33           Root MSE      =  1317.9

------------------------------------------------------------------------------
        FDHO |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         EXP |   .0372621   .0024547    15.18   0.000     .0324442      .04208
        SIZE |   559.7692   30.85684    18.14   0.000     499.2061    620.3322
       _cons |   884.5901   100.1537     8.83   0.000     688.0173    1081.163
------------------------------------------------------------------------------
```

A3.2

Perform a regression parallel to that in Exercise A3.1 for your *CES* category of expenditure, provide an interpretation of the regression coefficients and perform appropriate tests. Delete observations where expenditure on your category is zero.

A3.3

The output shows the result of regressing *FDHOPC*, expenditure on food consumed at home per capita, on *EXPPC*, total household expenditure per capita, and *SIZE*, number of persons in the household, using the *CES* data set. Provide an interpretation of the regression coefficients and perform appropriate tests.

```
. reg FDHOPC EXPPC SIZE if FDHO>0

      Source |       SS       df       MS              Number of obs =     868
-------------+------------------------------           F(  2,   865) =  175.68
       Model |   142127276     2  71063638.2           Prob > F      =  0.0000
    Residual |   349895173   865   404503.09           R-squared     =  0.2889
-------------+------------------------------           Adj R-squared =  0.2872
       Total |   492022449   867  567499.942           Root MSE      =  636.01

------------------------------------------------------------------------------
      FDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       EXPPC |   .0316606   .0026915    11.76   0.000      .0263779    .0369432
        SIZE |   -133.775   15.18071    -8.81   0.000     -163.5703   -103.9797
       _cons |   1430.123   67.10582    21.31   0.000      1298.413    1561.832
------------------------------------------------------------------------------
```

A3.4

Perform a regression parallel to that in Exercise A3.3 for your *CES* category of expenditure. Provide an interpretation of the regression coefficients and perform appropriate tests.

A3.5

The output shows the result of regressing *FDHOPC*, expenditure on food consumed at home per capita, on *EXPPC*, total household expenditure per capita, and *SIZEAM*, *SIZEAF*, *SIZEJM*, *SIZEJF*, and *SIZEIN*, numbers of adult males, adult females, junior males, junior females, and infants, respectively, in the household, using the *CES* data set. Provide an interpretation of the regression coefficients and perform appropriate tests.

```
. reg FDHOPC EXPPC SIZEAM SIZEAF SIZEJM SIZEJF SIZEIN if FDHO>0

      Source |       SS       df       MS              Number of obs =     868
-------------+------------------------------           F(  6,   861) =   59.11
       Model |   143547989     6  23924664.9           Prob > F      =  0.0000
    Residual |   348474460   861  404732.242           R-squared     =  0.2918
-------------+------------------------------           Adj R-squared =  0.2868
       Total |   492022449   867  567499.942           Root MSE      =  636.19

------------------------------------------------------------------------------
      FDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       EXPPC |   .0319472   .0027125    11.78   0.000     .0266234     .037271
      SIZEAM |  -159.6329   32.79751    -4.87   0.000    -224.0053   -95.26049
      SIZEAF |  -94.88238   37.98996    -2.50   0.013    -169.4462   -20.31861
      SIZEJM |  -101.5105   36.45485    -2.78   0.005    -173.0613    -29.9597
      SIZEJF |  -155.5774   37.49424    -4.15   0.000    -229.1682   -81.98661
      SIZEIN |  -220.7865   85.70005    -2.58   0.010     -388.992   -52.58108
       _cons |   1411.313   73.13575    19.30   0.000     1267.768    1554.859
------------------------------------------------------------------------------
```

A3.6

Perform a regression parallel to that in Exercise A3.5 for your *CES* category of expenditure. Provide an interpretation of the regression coefficients and perform appropriate tests.

A3.7

A researcher hypothesises that, for a typical enterprise, V, the logarithm of value added per worker, is related to K, the logarithm of capital per worker, and S, the logarithm of the average years of schooling of the workers, the relationship being

$$V = \beta_1 + \beta_2 K + \beta_3 S + u$$

where u is a disturbance term that satisfies the usual regression model assumptions. She fits the relationship (1) for a sample of 25 manufacturing enterprises, and (2) for a sample of 100 services enterprises. The table provides some data on the samples.

	(1) Manufacturing sample	(2) Services sample
Number of enterprises	25	100
Estimate of variance of u	0.16	0.64
Mean square deviation of K	4.00	16.00
Correlation between K and S	0.60	0.60

The mean square deviation of K is defined as $\frac{1}{n}\sum_i \left(K_i - \overline{K}\right)^2$, where n is the number of enterprises in the sample and $\overline{K}$ is the average value of K in the sample.

The researcher finds that the standard error of the coefficient of K is 0.050 for the manufacturing sample and 0.025 for the services sample. Explain the difference quantitatively, given the data in the table.

A3.8

A researcher is fitting earnings functions using a sample of data relating to individuals born in the same week in 1958. He decides to relate *Y*, gross hourly earnings in 2001, to *S*, years of schooling, and *PWE*, potential work experience, using the semilogarithmic specification

$$\log Y = \beta_1 + \beta_2 S + \beta_3 PWE + u$$

where *u* is a disturbance term assumed to satisfy the regression model assumptions. *PWE* is defined as age – years of schooling – 5. Since the respondents were all aged 43 in 2001, this becomes:

$$PWE = 43 - S - 5 = 38 - S.$$

The researcher finds that it is impossible to fit the model as specified. Stata output for his regression is reproduced below:

```
. reg LGY S PWE

      Source |       SS       df       MS              Number of obs =    5660
-------------+------------------------------           F(  1,  5658) = 1232.62
       Model |  237.170265     1  237.170265           Prob > F      =  0.0000
    Residual |  1088.66373  5658  .192411405           R-squared     =  0.1789
-------------+------------------------------           Adj R-squared =  0.1787
       Total |    1325.834  5659  .234287682           Root MSE      =  .43865

------------------------------------------------------------------------------
         LGY |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   .1038011   .0029566    35.11   0.000     .0980051    .1095971
         PWE |  (dropped)
       _cons |   .5000033   .0373785    13.38   0.000     .4267271    .5732795
------------------------------------------------------------------------------
```

Explain why the researcher was unable to fit his specification.

Explain how the coefficient of *S* might be interpreted.

Answers to the starred exercises in the textbook

3.5

Explain why the intercept in the regression of *EEARN* on *ES* is equal to zero.

Answer:

The intercept is calculated as $\overline{EEARN} - b_2\,\overline{ES}$. However, since the mean of the residuals from an OLS regression is zero, both $\overline{EEARN}$ and $\overline{ES}$ are zero, and hence the intercept is zero.

3.11

Demonstrate that $\bar{e} = 0$ in multiple regression analysis. (**Note:** The proof is a generalisation of the proof for the simple regression model, given in Section 1.5.)

Answer:

If the model is

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u,$$

$$b_1 = \overline{Y} - b_2 \overline{X}_2 - \ldots - b_k \overline{X}_k .$$

For observation i,

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_{2i} - \ldots - b_k X_{ki}.$$

Hence

$$\begin{aligned}\bar{e} &= \bar{Y} - b_1 - b_2 \bar{X}_2 - \ldots - b_k \bar{X}_k \\ &= \bar{Y} - \left[\bar{Y} - b_2 \bar{X}_2 - \ldots - b_k \bar{X}_k\right] - b_2 \bar{X}_2 - \ldots - b_k \bar{X}_k = 0.\end{aligned}$$

3.16

A researcher investigating the determinants of the demand for public transport in a certain city has the following data for 100 residents for the previous calendar year: expenditure on public transport, E, measured in dollars; number of days worked, W; and number of days not worked, NW. By definition NW is equal to 365 – W. He attempts to fit the following model

$$E = \beta_1 + \beta_2 W + \beta_3 NW + u.$$

Explain why he is unable to fit this equation. (Give both intuitive and technical explanations.) How might he resolve the problem?

Answer:

There is exact multicollinearity since there is an exact linear relationship between W, NW and the constant term. As a consequence it is not possible to tell whether variations in E are attributable to variations in W or variations in NW, or both. Noting that $NW_i - \overline{NW} = -W_i + \overline{W}$,

$$\begin{aligned}b_2 &= \frac{\sum\left(E_i - \bar{E}\right)\left(W_i - \bar{W}\right)\sum\left(NW_i - \overline{NW}\right)^2 - \sum\left(E_i - \bar{E}\right)\left(NW_i - \overline{NW}\right)\sum\left(W_i - \bar{W}\right)\left(NW_i - \overline{NW}\right)}{\sum\left(W_i - \bar{W}\right)^2 \sum\left(NW_i - \overline{NW}\right)^2 - \left(\sum\left(W_i - \bar{W}\right)\left(NW_i - \overline{NW}\right)\right)^2} \\ &= \frac{\sum\left(E_i - \bar{E}\right)\left(W_i - \bar{W}\right)\sum\left(-W_i - \bar{W}\right)^2 - \sum\left(E_i - \bar{E}\right)\left(-W_i + \bar{W}\right)\sum\left(W_i - \bar{W}\right)\left(-W_i + \bar{W}\right)}{\sum\left(W_i - \bar{W}\right)^2 \sum\left(W_i - \bar{W}\right)^2 - \left(\sum\left(W_i - \bar{W}\right)\left(-W_i + \bar{W}\right)\right)^2} \\ &= \frac{0}{0}.\end{aligned}$$

One way of dealing with the problem would be to drop NW from the regression. The interpretation of b_2 now is that it is an estimate of the *extra* expenditure on transport per day worked, compared with expenditure per day not worked.

3.21

The researcher in Exercise 3.16 decides to divide the number of days not worked into the number of days not worked because of illness, I, and the number of days not worked for other reasons, O. The mean value of I in the sample is 2.1 and the mean value of O is 120.2. He fits the regression (standard errors in parentheses):

$$\underset{}{\hat{E}} = \underset{(8.3)}{-9.6} + \underset{(1.98)}{2.10}W + \underset{(1.77)}{0.45}O \qquad R^2 = 0.72$$

Perform t tests on the regression coefficients and an F test on the goodness of fit of the equation. Explain why the t tests and F test have different outcomes.

Answer:

Although there is not an exact linear relationship between *W* and *O*, they must have a very high negative correlation because the mean value of *I* is so small. Hence one would expect the regression to be subject to multicollinearity, and this is confirmed by the results. The *t* statistics for the coefficients of *W* and *O* are only 1.06 and 0.25, respectively, but the *F* statistic,

$$F(2,97) = \frac{0.72/2}{(1-0.72)/97} = 124.7$$

is greater than the critical value of *F* at the 0.1 per cent level, 7.41.

Answers to the additional exercises

A3.1

The regression indicates that 3.7 cents out of the marginal expenditure dollar is spent on food consumed at home, and that expenditure on this category increases by $560 for each individual in the household, keeping total expenditure constant. Both of these effects are very highly significant, and almost half of the variance in *FDHO* is explained by *EXP* and *SIZE*. The intercept has no plausible interpretation.

A3.2

With the exception of *LOCT*, all of the categories have positive coefficients for *EXP*, with high significance levels, but the *SIZE* effect varies:

- Positive, significant at the 1 per cent level: *FDHO*, *TELE*, *CLOT*, *FOOT*, *GASO*.
- Positive, significant at the 5 per cent level: *LOCT*.
- Negative, significant at the 1 per cent level: *TEXT*, *FEES*, *READ*.
- Negative, significant at the 5 per cent level: *SHEL*, *EDUC*.
- Not significant: *FDAW*, *DOM*, *FURN*, *MAPP*, *SAPP*, *TRIP*, *HEAL*, *ENT*, *TOYS*, *TOB*.

At first sight it may seem surprising that *SIZE* has a significant negative effect for some categories. The reason for this is that an increase in *SIZE* means a reduction in expenditure per capita, if total household expenditure is kept constant, and thus *SIZE* has a (negative) income effect in addition to any direct effect. Effectively poorer, the larger household has to spend more on basics and less on luxuries. To determine the true direct effect, we need to eliminate the income effect, and that is the point of the re-specification of the model in the next exercise.

		EXP		*SIZE*			
	n	b_2	$s.e.(b_2)$	b_3	$s.e.(b_3)$	R^2	F
FDHO	868	0.0373	0.0025	559.77	30.86	0.4967	426.8
FDAW	827	0.0454	0.0022	–53.06	27.50	0.3559	227.6
SHEL	867	0.1983	0.0067	–174.40	83.96	0.5263	479.9
TELE	858	0.0091	0.0010	36.10	12.08	0.1360	67.3
DOM	454	0.0217	0.0047	26.10	64.14	0.0585	14.0
TEXT	482	0.0057	0.0007	–33.15	9.11	0.1358	37.7
FURN	329	0.0138	0.0024	–47.52	35.18	0.0895	16.0
MAPP	244	0.0083	0.0019	25.35	24.33	0.0954	12.7
SAPP	467	0.0014	0.0003	–5.63	3.73	0.0539	13.2
CLOT	847	0.0371	0.0019	87.98	24.39	0.3621	239.5
FOOT	686	0.0028	0.0003	21.24	4.01	0.1908	80.5
GASO	797	0.0205	0.0015	94.58	18.67	0.2762	151.5
TRIP	309	0.0273	0.0042	–110.11	56.17	0.1238	21.6
LOCT	172	–0.0012	0.0021	54.97	23.06	0.0335	2.9
HEAL	821	0.0231	0.0032	–18.60	40.56	0.0674	29.6
ENT	824	0.0726	0.0042	–98.94	52.61	0.2774	157.6
FEES	676	0.0335	0.0028	–114.71	36.04	0.1790	73.4
TOYS	592	0.0089	0.0011	5.03	13.33	0.1145	38.1
READ	764	0.0043	0.0003	–15.86	4.06	0.1960	92.8
EDUC	288	0.0295	0.0055	–168.13	74.57	0.0937	14.7
TOB	368	0.0068	0.0014	14.44	16.29	0.0726	14.3

A3.3

Another surprise, perhaps. The purpose of this specification is to test whether household size has an effect on expenditure per capita on food consumed at home, controlling for the income effect of variations in household size mentioned in the answer to Exercise A3.2. Expenditure per capita on food consumed at home increases by 3.2 cents out of the marginal dollar of total household expenditure per capita. Now *SIZE* has a very significant negative effect. Expenditure per capita on *FDHO* decreases by $134 per year for each extra person in the household, suggesting that larger households are more efficient than smaller ones with regard to expenditure on this category, the effect being highly significant. R^2 is much lower than in Exercise A3.1, but a comparison is invalidated by the fact that the dependent variable is different.

A3.4

Several categories have significant negative *SIZE* effects. None has a significant positive effect.

- Negative, significant at the 1 per cent level: *FDHO*, *SHEL*, *TELE*, *SAPP*, *GASO*, *HEAL*, *READ*, *TOB*.
- Negative, significant at the 5 per cent level: *FURN*, *FOOT*, *LOCT*, *EDUC*.
- Not significant: *FDAW*, *DOM*, *TEXT*, *MAPP*, *CLOT*, *TRIP*, *ENT*, *FEES*, *TOYS*.

One explanation of the negative effects could be economies of scale, but this is not plausible in the case of some, most obviously *TOB*. Another might be family composition – larger families having more children. This possibility is investigated in the next exercise.

		EXPPC		*SIZE*			
	n	b_2	$s.e.(b_2)$	b_3	$s.e.(b_3)$	R^2	F
FDHO	868	0.0317	0.0027	–133.78	15.18	0.2889	175.7
FDAW	827	0.0476	0.0027	–59.89	68.15	0.3214	195.2
SHEL	867	0.2017	0.0075	–113.68	42.38	0.5178	463.9
TELE	858	0.0145	0.0014	–43.07	7.83	0.2029	108.8
DOM	454	0.0243	0.0060	–1.33	35.58	0.0404	9.5
TEXT	482	0.0115	0.0011	5.01	6.43	0.2191	67.2
FURN	329	0.0198	0.0033	–43.12	21.23	0.1621	31.5
MAPP	244	0.0124	0.0022	–25.96	13.98	0.1962	29.4
SAPP	467	0.0017	0.0004	–7.76	2.01	0.1265	33.6
CLOT	847	0.0414	0.0021	21.83	12.07	0.3327	210.4
FOOT	686	0.0034	0.0003	–3.87	1.89	0.1939	82.2
GASO	797	0.0183	0.0015	–42.49	8.73	0.2553	136.1
TRIP	309	0.0263	0.0044	–13.06	27.15	0.1447	25.9
LOCT	172	–0.0005	0.0018	–23.84	9.16	0.0415	3.7
HEAL	821	0.0181	0.0036	–178.20	20.80	0.1587	77.1
ENT	824	0.0743	0.0046	–392.86	118.53	0.2623	146.0
FEES	676	0.0337	0.0032	23.97	19.33	0.1594	63.8
TOYS	592	0.0095	0.0011	–5.89	6.20	0.1446	49.8
READ	764	0.0050	0.0004	–12.49	2.21	0.2906	155.9
EDUC	288	0.0235	0.0088	–108.18	47.45	0.0791	12.2
TOB	368	0.0057	0.0016	–48.87	37.92	0.1890	42.5

A3.5

It is not completely obvious how to interpret these regression results and possibly this is not the most appropriate specification for investigating composition effects. The coefficient of *SIZEAF* suggests that for each additional adult female in the household, expenditure falls by $95 per year, probably as a consequence of economies of scale. For each infant, there is an extra reduction, relative to adult females, of $126 per year, because infants consume less food. Similar interpretations might be given to the coefficients of the other composition variables.

A3.6

The regression results for this specification are summarised in the table below. In the case of *SHEL*, the regression indicates that the *SIZE* effect is attributable to *SIZEAM*. To investigate this further, the regression was repeated: (1) restricting the sample to households with at least one adult male, and (2) restricting the sample to households with either no adult male or just 1 adult male. The first regression produces a negative effect for *SIZEAM*, but it is smaller than with the whole sample and not significant. In the second regression the coefficient of *SIZEAM* jumps dramatically, from –$424 to –$793, suggesting very strong economies of scale for this particular comparison.

As might be expected, the *SIZE* composition variables on the whole do not appear to have significant effects if the *SIZE* variable does not in Exercise A3.4. The results for *TOB* are puzzling, in that the apparent economies of scale do not appear to be related to household composition.

Category	*FDHOPC*	*FDAWPC*	*SHELPC*	*TELEPC*	*DOMPC*	*TEXTPC*	*FURNPC*	*MAPPPC*
EXP	0.0319	0.0473	0.2052	0.0146	0.0262	0.0116	0.0203	0.0125
	(0.0027)	(0.0027)	(0.0075)	(0.0014)	(0.0061)	(0.0011)	(0.0034)	(0.0022)
SIZEAM	–159.63	29.32	–423.85	–48.79	–133.37	2.36	–69.54	–46.54
	(32.80)	(32.48)	(90.57)	(16.99)	(83.47)	(13.07)	(42.20)	(28.26)
SIZEAF	–94.88	–22.82	–222.96	–56.23	–71.36	–15.66	–79.52	–19.74
	(37.99)	(37.59)	(105.22)	(19.80)	(95.81)	(17.36)	(54.43)	(32.49)
SIZEJM	–101.51	1.85	53.70	–39.65	84.39	10.02	0.26	–22.34
	(36.45)	(35.61)	(100.60)	(18.80)	(84.30)	(14.59)	(47.01)	(32.84)
SIZEJF	–155.58	–19.48	–6.32	–38.01	23.95	11.83	–36.24	–12.48
	(37.49)	(36.67)	(103.52)	(19.33)	(82.18)	(14.05)	(48.41)	(29.21)
SIZEIN	–220.79	–24.44	469.75	–5.40	176.93	17.34	–25.96	–35.46
	(85.70)	(83.05)	(236.44)	(44.12)	(183.84)	(34.47)	(87.82)	(78.95)
R^2	0.2918	0.3227	0.5297	0.2041	0.0503	0.2224	0.1667	0.1988
F	59.1	65.1	161.4	36.4	4.0	22.6	10.7	9.8
n	868	827	867	858	454	482	329	244

Category	*SAPPPC*	*CLOTPC*	*FOOTPC*	*GASOPC*	*TRIPPC*	*LOCTPC*	*HEALPC*	*ENTPC*
EXP	0.0017	0.0420	0.0035	0.0179	0.0263	–0.0005	0.0182	0.0740
	(0.0004)	(0.0021)	(0.0003)	(0.0015)	(0.0044)	(0.0019)	(0.0037)	(0.0046)
SIZEAM	–9.13	–27.91	–6.66	13.99	4.33	–33.64	–191.60	74.58
	(4.17)	(25.90)	(3.93)	(18.49)	(54.53))	(19.53)	(44.43)	(56.32)
SIZEAF	–2.49	47.58	–9.31	–40.43	31.58	10.23	–46.92	24.53
	(4.99)	(30.29)	(5.03)	(21.37)	(66.29)	(24.15)	(52.65)	(64.94)
SIZEJM	–8.93	19.87	–2.58	–62.37	–40.20	–50.45	–230.65	38.60
	(4.63)	(28.55)	(4.28)	(20.10)	(65.07)	(21.71)	(50.63)	(61.24)
SIZEJF	–8.63	40.08	2.35	–64.07	–34.98	–21.49	–194.56	65.74
	(4.64)	(29.42)	(4.35)	(20.28)	(70.51)	(22.02)	(51.80)	(63.12)
SIZEIN	–10.55	87.53	–8.35	–112.58	–51.85	19.04	–247.58	–16.49
	(11.44)	(66.80)	(9.94)	(46.57)	(194.69)	(70.79)	(113.55)	(142.40)
R^2	0.1290	0.3373	0.1987	0.2680	0.1472	0.0636	0.1665	0.2629
F	11.4	71.3	28.1	48.2	8.7	1.9	27.1	48.6
n	467	847	686	797	309	172	821	824

Category	*FEESPC*	*TOYSPC*	*READPC*	*EDUCPC*	*TOBPC*
EXP	0.0337 (0.0032)	0.0096 (0.0012)	0.0050 (0.0004)	0.0232 (0.0090)	0.0056 (0.0016)
SIZEAM	28.62 (39.84)	–17.99 (13.16)	–21.85 (4.79)	–135.34 (88.87)	–37.24 (17.19)
SIZEAF	32.68 (46.77)	–3.68 (15.82)	–4.22 (5.51)	–46.03 (103.88)	–56.54 (17.50)
SIZEJM	15.65 (44.40)	–2.59 (13.70)	–13.28 (5.27)	–106.39 (92.25)	–44.45 (18.53)
SIZEJF	32.07 (42.92)	3.07 (13.66)	–8.61 (5.40)	–119.36 (91.60)	–52.68 (22.87)
SIZEIN	–29.86 (95.20)	–18.08 (30.40)	–15.12 (11.86)	–149.87 (262.13)	–76.25 (53.68)
R^2	0.1599	0.1468	0.2969	0.0808	0.1913
F	21.2	16.8	53.3	4.1	14.2
n	676	592	764	288	368

A3.7

The standard error is given by

$$s.e.(b_2) = s_u \times \frac{1}{\sqrt{n}} \times \frac{1}{\sqrt{\text{MSD}(K)}} \times \frac{1}{\sqrt{1 - r_{K,S}^2}} \quad .$$

	Data		Factors	
	manufacturing sample	services sample	manufacturing sample	services sample
Number of enterprises	25	100	0.20	0.10
Estimate of variance of u	0.16	0.64	0.40	0.80
Mean square deviation of K	4	16	0.50	0.25
Correlation between K and S	0.6	0.6	1.25	1.25
Standard errors			0.050	0.025

The table shows the four factors for the two sectors. Other things being equal, the larger number of enterprises and the greater MSD of K would separately cause the standard error of b_2 for the services sample to be half that in the manufacturing sample. However, the larger estimate of the variance of u would, taken in isolation, cause it to be double. The net effect, therefore, is that it is half.

A3.8

The specification is subject to exact multicollinearity since there is an exact linear relationship linking *PWE* and *S*.

The coefficient of *S* should be interpreted as providing an estimate of the proportional effect on hourly earnings of an extra year of schooling, allowing for the fact that this means one fewer year of work experience.

Notes