Chapter 2: Properties of the regression coefficients and hypothesis testing

Overview

Chapter 1 introduced least squares regression analysis, a mathematical technique for fitting a relationship given suitable data on the variables involved. It is a fundamental chapter because much of the rest of the text is devoted to extending the least squares approach to handle more complex models, for example models with multiple explanatory variables, nonlinear models, and models with qualitative explanatory variables.

However, the mechanics of fitting regression equations are only part of the story. We are equally concerned with assessing the performance of our regression techniques and with developing an understanding of why they work better in some circumstances than in others. Chapter 2 is the starting point for this objective and is thus equally fundamental. In particular, it shows how two of the three main criteria for assessing the performance of estimators, unbiasedness and efficiency, are applied in the context of a regression model. The third criterion, consistency, will be considered in Chapter 8.

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 guide, you should be able to explain what is meant by:

- cross-sectional, time series, and panel data
- unbiasedness of OLS regression estimators
- variance and standard errors of regression coefficients and how they are determined
- Gauss–Markov theorem and efficiency of OLS regression estimators
- two-sided t tests of hypotheses relating to regression coefficients and one-sided t tests of hypotheses relating to regression coefficients
- F tests of goodness of fit of a regression equation

in the context of a regression model. The chapter is a long one and you should take your time over it because it is essential that you develop a perfect understanding of every detail.

Further material

Derivation of the expression for the variance of the naïve estimator in Section 2.3.

The variance of the naïve estimator in Section 2.3 and Exercise 2.9 is not of any great interest in itself but its derivation provides an example of how one obtains expressions for variances of estimators in general.

In Section 2.3 we considered the naïve estimator of the slope coefficient derived by joining the first and last observations in a sample and calculating the slope of that line:

$$b_2 = \frac{Y_n - Y_1}{X_n - X_1}.$$

It was demonstrated that the estimator could be decomposed as

$$b_2 = \beta_2 + \frac{u_n - u_1}{X_n - X_1}$$

and hence that $E(b_2) = \beta_2$.

The population variance of a random variable X is defined to be $E([X - \mu_X]^2)$ where $\mu_X = E(X)$. Hence the population variance of b_2 is given by

$$\sigma_{b_2}^2 = E\left(\left[b_2 - \beta_2\right]^2\right) = E\left(\left[\left\{\beta_2 + \frac{u_n - u_1}{X_n - X_1}\right\} - \beta_2\right]^2\right) = E\left(\left[\frac{u_n - u_1}{X_n - X_1}\right]^2\right).$$

On the assumption that X is nonstochastic, this can be written as

$$\sigma_{b_2}^2 = \left[\frac{1}{X_n - X_1}\right]^2 E\left(\left[u_n - u_1\right]^2\right).$$

Expanding the quadratic, we have

$$\begin{aligned}\sigma_{b_2}^2 &= \left[\frac{1}{X_n - X_1}\right]^2 E\left(u_n^2 + u_1^2 - 2u_n u_1\right) \\ &= \left[\frac{1}{X_n - X_1}\right]^2 \left[E\left(u_n^2\right) + E\left(u_1^2\right) - 2E\left(u_n u_1\right)\right].\end{aligned}$$

Each value of the disturbance term is drawn randomly from a distribution with mean 0 and population variance σ_u^2, so $E\left(u_n^2\right)$ and $E\left(u_1^2\right)$ are both equal to σ_u^2. u_n and u_1 are drawn independently from the distribution, so $E(u_n u_1) = E(u_n)E(u_1) = 0$. Hence

$$\sigma_{b_2}^2 = \frac{2\sigma_u^2}{\left(X_n - X_1\right)^2} = \frac{\sigma_u^2}{\frac{1}{2}\left(X_n - X_1\right)^2}.$$

Define $A = \frac{1}{2}\left(X_1 + X_n\right)$, the average of X_1 and X_n, and $D = X_n - A = A - X_1$. Then

$$\begin{aligned}\frac{1}{2}\left(X_n - X_1\right)^2 &= \frac{1}{2}\left(X_n - A + A - X_1\right)^2 \\ &= \frac{1}{2}\left[\left(X_n - A\right)^2 + \left(A - X_1\right)^2 + 2\left(X_n - A\right)\left(A - X_1\right)\right] \\ &= \frac{1}{2}\left[D^2 + D^2 + 2(D)(D)\right] = 2D^2 \\ &= (X_n - A)^2 + (A - X_1)^2 \\ &= \left(X_n - A\right)^2 + \left(X_1 - A\right)^2 \\ &= \left(X_n - \overline{X} + \overline{X} - A\right)^2 + \left(X_1 - \overline{X} + \overline{X} - A\right)^2 \\ &= \left(X_n - \overline{X}\right)^2 + \left(\overline{X} - A\right)^2 + 2\left(X_n - \overline{X}\right)\left(\overline{X} - A\right) \\ &\quad + \left(X_1 - \overline{X}\right)^2 + \left(\overline{X} - A\right)^2 + 2\left(X_1 - \overline{X}\right)\left(\overline{X} - A\right) \\ &= \left(X_1 - \overline{X}\right)^2 + \left(X_n - \overline{X}\right)^2 + 2\left(\overline{X} - A\right)^2 + 2\left(X_1 + X_n - 2\overline{X}\right)\left(\overline{X} - A\right) \\ &= \left(X_1 - \overline{X}\right)^2 + \left(X_n - \overline{X}\right)^2 + 2\left(\overline{X} - A\right)^2 + 2\left(2A - 2\overline{X}\right)\left(\overline{X} - A\right) \\ &= \left(X_1 - \overline{X}\right)^2 + \left(X_n - \overline{X}\right)^2 - 2\left(\overline{X} - A\right)^2 = \left(X_1 - \overline{X}\right)^2 + \left(X_n - \overline{X}\right)^2 - 2\left(A - \overline{X}\right)^2 \\ &= \left(X_1 - \overline{X}\right)^2 + \left(X_n - \overline{X}\right)^2 - \frac{1}{2}\left(X_1 + X_n - 2\overline{X}\right)^2\end{aligned}$$

Hence we obtain the expression in Exercise 2.9. There must be a shorter proof.

Additional exercises

A2.1

A variable Y_i is generated as

$$Y_i = \beta_1 + u_i$$

where β_1 is a fixed parameter and u_i is a disturbance term that is independently and identically distributed with expected value 0 and population variance σ_u^2. The least squares estimator of β_1 is $\overline{Y}$, the sample mean of *Y*. However a researcher believes that *Y* is a linear function of another variable *X* and uses ordinary least squares to fit the relationship

$$\hat{Y} = b_1 + b_2 X$$

calculating b_1 as $\overline{Y} - b_2\overline{X}$, where $\overline{X}$ is the sample mean of *X*. *X* may be assumed to be a nonstochastic variable. Determine whether the researcher's estimator b_1 is biased or unbiased, and if biased, determine the direction of the bias.

A2.2

With the model described in Exercise A2.1, standard theory states that the population variance of the researcher's estimator b_1 is $\sigma_u^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum\left(X_i - \overline{X}\right)^2}\right]$. In general, this is larger than the population variance of $\overline{Y}$, which is $\frac{\sigma_u^2}{n}$. Explain the implications of the difference in the variances.

In the special case where $\overline{X} = 0$, the variances are the same. Give an intuitive explanation.

A2.3

Using the output for the regression in Exercise A1.9 in the text, reproduced below, perform appropriate statistical tests.

```
. reg CHILDREN SM

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =   63.60
       Model |   272.69684     1   272.69684           Prob > F      =  0.0000
    Residual |   2306.7402   538  4.28762118           R-squared     =  0.1057
-------------+------------------------------           Adj R-squared =  0.1041
       Total |  2579.43704   539  4.78559747           Root MSE      =  2.0707

------------------------------------------------------------------------------
    CHILDREN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          SM |  -.2525473   .0316673    -7.98   0.000     -.314754   -.1903406
       _cons |   7.198478   .3773667    19.08   0.000     6.457186    7.939771
------------------------------------------------------------------------------
```

A2.4

Using the output for the regression in Exercise A1.1, reproduced below, perform appropriate statistical tests.

```
. reg FDHO EXP if FDHO>0

      Source |       SS       df       MS              Number of obs =     868
-------------+------------------------------           F(  1,   866) =  380.37
       Model |   911005795     1   911005795           Prob > F      =  0.0000
    Residual |  2.0741e+09   866  2395045.39           R-squared     =  0.3052
-------------+------------------------------           Adj R-squared =  0.3044
       Total |  2.9851e+09   867  3443039.33           Root MSE      =  1547.6

------------------------------------------------------------------------------
        FDHO |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         EXP |   .0527204   .0027032    19.50   0.000     .0474149     .058026
       _cons |   1922.939   96.50688    19.93   0.000     1733.525    2112.354
------------------------------------------------------------------------------
```

A2.5

Using the output for your regression in Exercise A1.2, perform appropriate statistical tests.

A2.6

Using the output for the regression in Exercise A1.3, reproduced below, perform appropriate statistical tests.

```
. reg WEIGHT02 WEIGHT85

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) = 1149.83
       Model |   620662.43     1   620662.43           Prob > F      =  0.0000
    Residual |  290406.035   538  539.788169           R-squared     =  0.6812
-------------+------------------------------           Adj R-squared =  0.6807
       Total |  911068.465   539    1690.294           Root MSE      =  23.233

------------------------------------------------------------------------------
    WEIGHT02 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    WEIGHT85 |   1.013353   .0298844    33.91   0.000     .9546483    1.072057
       _cons |   23.61869   4.760179     4.96   0.000     14.26788    32.96951
------------------------------------------------------------------------------
```

A2.7

Using the output for the regression in Exercise A1.4, reproduced below, perform appropriate statistical tests.

```
. reg EARNINGS HEIGHT

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =   29.39
       Model |  6242.37244     1  6242.37244           Prob > F      =  0.0000
    Residual |  114276.055   538  212.409025           R-squared     =  0.0518
-------------+------------------------------           Adj R-squared =  0.0500
       Total |  120518.428   539  223.596341           Root MSE      =  14.574

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      HEIGHT |   .8583752   .1583393     5.42   0.000     .5473361    1.169414
       _cons |  -38.05756   10.70014    -3.56   0.000    -59.07674   -17.03837
------------------------------------------------------------------------------
```

A2.8

With the information given in Exercise A1.5, how would the change in the measurement of GDP affect

- the standard error of the coefficient of GDP
- the *F* statistic for the equation?

A2.9

With the information given in Exercise A1.6, how would the change in the measurement of GDP affect

- the standard error of the coefficient of GDP
- the *F* statistic for the equation?

A2.10

[This is a continuation of Exercise 1.15 in the text.] A sample of data consists of n observations on two variables, Y and X. The true model is

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

where β_1 and β_2 are parameters and u is a disturbance term that satisfies the usual regression model assumptions. In view of the true model,

$$\overline{Y} = \beta_1 + \beta_2 \overline{X} + \overline{u}$$

where $\overline{Y}$, $\overline{X}$, and $\overline{u}$ are the sample means of Y, X, and u. Subtracting the second equation from the first, one obtains

$$Y_i^* = \beta_2 X_i^* + u_i^*$$

where $Y_i^* = Y_i - \overline{Y}$, $X_i^* = X_i - \overline{X}$, and $u_i^* = u_i - \overline{u}$. Note that, by construction, the sample means of Y^*, X^*, and u^* are all equal to zero.

One researcher fits

$$\hat{Y} = b_1 + b_2 X. \quad (1)$$

A second researcher fits

$$\hat{Y}^* = b_1^* + b_2^* X^*. \quad (2)$$

[Note: The second researcher included an intercept in the specification.]

- Comparing regressions (1) and (2), demonstrate that $\hat{Y}_i^* = \hat{Y}_i - \overline{Y}$.
- Demonstrate that the residuals in (2) are identical to the residuals in (1).
- Demonstrate that the OLS estimator of the variance of the disturbance term in (2) is equal to that in (1).
- Explain how the standard error of the slope coefficient in (2) is related to that in (1).
- Explain how R^2 in (2) is related to R^2 in (1).
- Explain why, theoretically, the specification (2) of the second researcher is incorrect and he should have fitted

  $$\hat{Y}^* = b_2^* X^* \quad (3)$$

  not including a constant in his specification.
- If the second researcher had fitted (3) instead of (2), how would this have affected his estimator of β_2? Would dropping the unnecessary intercept lead to a gain in efficiency?

A2.11

A variable Y depends on a nonstochastic variable X with the relationship

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

where u is a disturbance term that satisfies the regression model assumptions. Given a sample of n observations, a researcher decides to estimate β_2 using the expression

$$b_2 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.$$

(This is the OLS estimator of β_2 for the model $Y = \beta_2 X + u$). It can be shown that the population variance of this estimator is $\dfrac{\sigma_u^2}{\sum_{i=1}^{n} X_i^2}$.

- Demonstrate that b_2 is in general a biased estimator of β_2.
- Discuss whether it is possible to determine the sign of the bias.
- Demonstrate that b_2 is unbiased if $\beta_1 = 0$.
- What can be said in this case about the efficiency of b_2, comparing it with the estimator $\dfrac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$?
- Demonstrate that b_2 is unbiased if $\bar{X} = 0$.
  What can be said in this case about the efficiency of b_2, comparing it with the estimator

  $$\frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}?$$

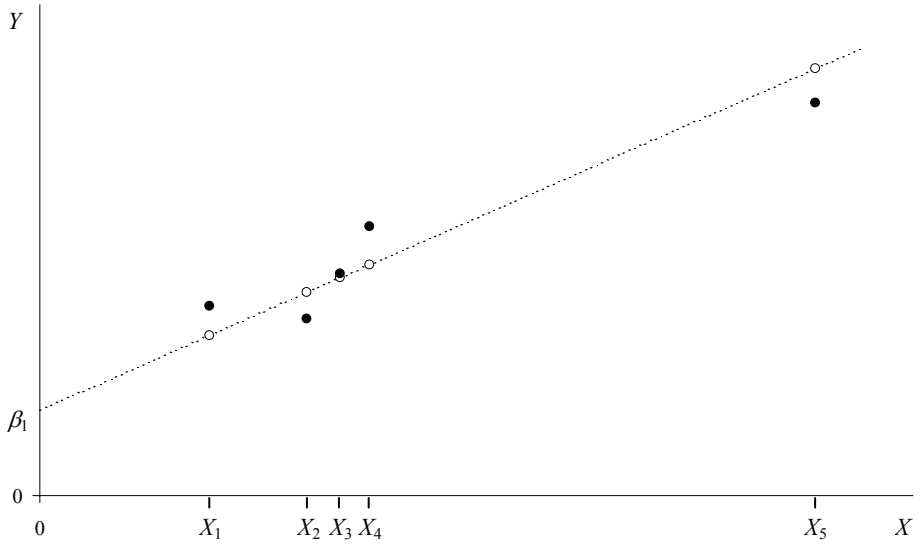
  Explain the underlying reason for this conclusion.
- Returning to the general case where $\beta_1 \neq 0$ and $\bar{X} \neq 0$, suppose that there is very little variation in X in the sample. Is it possible that b_2 might be a better estimator than the OLS estimator?

A2.12

A variable Y_i is generated as

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad (1)$$

where β_1 and β_2 are fixed parameters and u_i is a disturbance term that satisfies the regression model assumptions. The values of X are fixed and are as shown in the figure opposite. Four of them, X_1 to X_4, are close together. The fifth, X_5, is much larger. The corresponding values that Y would take, if there were no disturbance term, are given by the circles on the line. The presence of the disturbance term in the model causes the actual values of Y in a sample to be different. The solid black circles depict a typical sample of observations.



Discuss the advantages and disadvantages of dropping the observation corresponding to X_5 when regressing Y on X. If you keep the observation in the sample, will this cause the regression estimates to be biased?

Answers to the starred exercises in the textbook

2.1

Demonstrate that $b_1 = \beta_1 + \sum_{i=1}^{n} c_i u_i$, where $c_i = \frac{1}{n} - a_i \overline{X}$ and a_i is defined in equation (2.21).

Answer:

$$b_1 = \overline{Y} - b_2 \overline{X} = \left(\beta_1 + \beta_2 \overline{X} + \overline{u}\right) - \overline{X}\left(\beta_2 + \sum_{i=1}^{n} a_i u_i\right)$$

$$= \beta_1 + \frac{1}{n}\sum_{i=1}^{n} u_i - \overline{X}\sum_{i=1}^{n} a_i u_i = \beta_1 + \sum_{i=1}^{n} c_i u_i \,.$$

2.5

An investigator correctly believes that the relationship between two variables X and Y is given by

$$Y_i = \beta_1 + \beta_2 X_i + u_i.$$

Given a sample of observations on Y, X, and a third variable Z (which is not a determinant of Y), the investigator estimates β_2 as

$$\frac{\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)\left(X_i - \overline{X}\right)}.$$

Demonstrate that this estimator is unbiased.

Answer: Noting that $Y_i - \overline{Y} = \beta_2\left(X_i - \overline{X}\right) + u_i - \overline{u}$,

$$b_2 = \frac{\sum\left(Z_i - \overline{Z}\right)\left(Y_i - \overline{Y}\right)}{\sum\left(Z_i - \overline{Z}\right)\left(X_i - \overline{X}\right)} = \frac{\sum\left(Z_i - \overline{Z}\right)\beta_2\left(X_i - \overline{X}\right) + \sum\left(Z_i - \overline{Z}\right)\left(u_i - \overline{u}\right)}{\sum\left(Z_i - \overline{Z}\right)\left(X_i - \overline{X}\right)}$$

$$= \beta_2 + \frac{\sum\left(Z_i - \overline{Z}\right)\left(u_i - \overline{u}\right)}{\sum\left(Z_i - \overline{Z}\right)\left(X_i - \overline{X}\right)}.$$

Hence

$$E(b_2) = \beta_2 + \frac{\sum (Z_i - \bar{Z}) E(u_i - \bar{u})}{\sum (Z_i - \bar{Z})(X_i - \bar{X})} = \beta_2.$$

2.6

Using the decomposition of b_1 obtained in Exercise 2.1, derive the expression for $\sigma_{b_1}^2$ given in equation (2.38).

Answer: $b_1 = \beta_1 + \sum_{i=1}^{n} c_i u_i$, where $c_i = \frac{1}{n} - a_i \bar{X}$. Hence

$$\sigma_{b_1}^2 = E\left[\left(\sum_{i=1}^{n} c_i u_i\right)^2\right] = \sigma_u^2 \sum_{i=1}^{n} c_i^2 = \sigma_u^2 \left(n\frac{1}{n^2} - 2\frac{\bar{X}}{n}\sum_{i=1}^{n} a_i + \bar{X}^2 \sum_{i=1}^{n} a_i^2\right).$$

From Box 2.2, $\sum_{i=1}^{n} a_i = 0$ and $\sum_{i=1}^{n} a_i^2 = \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$.

Hence

$$\sigma_{b_1}^2 = \sigma_u^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\right).$$

2.7

Given the decomposition in Exercise 2.2 of the OLS estimator of β_2 in the model $Y_i = \beta_2 X_i + u_i$, demonstrate that the variance of the slope coefficient is given by

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum_{j=1}^{n} X_j^2}.$$

Answer:

$b_2 = \beta_2 + \sum_{i=1}^{n} d_i u_i$, where $d_i = \frac{X_i}{\sum_{j=1}^{n} X_j^2}$, and $E(b_2) = \beta_2$. Hence

$$\sigma_{b_2}^2 = E\left[\left(\sum_{i=1}^{n} d_i u_i\right)^2\right] = \sigma_u^2 \sum_{i=1}^{n} d_i^2 = \sigma_u^2 \sum_{i=1}^{n} \left(\frac{X_i^2}{\left(\sum_{j=1}^{n} X_j^2\right)^2}\right) = \frac{\sigma_u^2}{\left(\sum_{j=1}^{n} X_j^2\right)^2} \sum_{i=1}^{n} X_i^2 = \frac{\sigma_u^2}{\sum_{j=1}^{n} X_j^2}.$$

2.10

It can be shown that the variance of the estimator of the slope coefficient in Exercise 2.4,

$$\frac{\sum_{i=1}^{n} (Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (Z_i - \bar{Z})(X_i - \bar{X})}$$

is given by

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \times \frac{1}{r_{XZ}^2}$$

where r_{XZ} is the correlation between X and Z. What are the implications for the efficiency of the estimator?

Answer:

If Z happens to be an exact linear function of X, the population variance will be the same as that of the OLS estimator. Otherwise $\frac{1}{r_{XZ}^2}$ will be greater than 1, the variance will be larger, and so the estimator will be less efficient.

2.13

Suppose that the true relationship between Y and X is $Y_i = \beta_1 + \beta_2 X_i + u_i$ and that the fitted model is $\hat{Y}_i = b_1 + b_2 X_i$. In Exercise 1.12 it was shown that if $X_i^* = \mu_1 + \mu_2 X_i$, and Y is regressed on X^*, the slope coefficient $b_2^* = b_2 / \mu_2$. How will the standard error of b_2^* be related to the standard error of b_2?

Answer:

In Exercise 1.22 it was demonstrated that the fitted values of Y would be the same. This means that the residuals are the same, and hence s_u^2, the estimator of the variance of the disturbance term, is the same. The standard error of b_2^* is then given by

$$\text{s.e.}\left(b_2^*\right) = \sqrt{\frac{s_u^2}{\sum\left(X_i^* - \overline{X}^*\right)^2}} = \sqrt{\frac{s_u^2}{\sum\left(\mu_1 + \mu_2 X_i - \mu_1 - \mu_2 \overline{X}\right)^2}}$$

$$= \sqrt{\frac{s_u^2}{\mu_2^2 \sum\left(X_i - \overline{X}\right)^2}} = \frac{1}{\mu_2}\text{s.e.}\left(b_2\right).$$

2.15

A researcher with a sample of 50 individuals with similar education but differing amounts of training hypothesises that hourly earnings, *EARNINGS*, may be related to hours of training, *TRAINING*, according to the relationship

$$EARNINGS = \beta_1 + \beta_2 TRAINING + u .$$

He is prepared to test the null hypothesis H_0: $\beta_2 = 0$ against the alternative hypothesis H_1: $\beta_2 \neq 0$ at the 5 per cent and 1 per cent levels. What should he report

1. if $b_2 = 0.30$, s.e.(b_2) = 0.12?
2. if $b_2 = 0.55$, s.e.(b_2) = 0.12?
3. if $b_2 = 0.10$, s.e.(b_2) = 0.12?
4. if $b_2 = -0.27$, s.e.(b_2) = 0.12?

Answer:

There are 48 degrees of freedom, and hence the critical values of t at the 5 per cent, 1 per cent, and 0.1 per cent levels are 2.01, 2.68, and 3.51, respectively.

1. The *t* statistic is 2.50. Reject H_0 at the 5 per cent level but not at the 1 per cent level.
2. $t = 4.58$. Reject at the 0.1 per cent level.
3. $t = 0.83$. Fail to reject at the 5 per cent level.
4. $t = -2.25$. Reject H_0 at the 5 per cent level but not at the 1 per cent level.

2.20

Explain whether it would have been possible to perform one-sided tests instead of two-sided tests in Exercise 2.15. If you think that one-sided tests are justified, perform them and state whether the use of a one-sided test makes any difference.

Answer:

First, there should be a discussion of whether the parameter β_2 in

$$EARNINGS = \beta_1 + \beta_2 TRAINING + u$$

can be assumed not to be negative. The objective of training is to impart skills. It would be illogical for an individual with greater skills to be paid less on that account, and so we can argue that we can rule out $\beta_2 < 0$. We can then perform a one-sided test. With 48 degrees of freedom, the critical values of *t* at the 5 per cent, 1 per cent, and 0.1 per cent levels are 1.68, 2.40, and 3.26, respectively.

1. The *t* statistic is 2.50. We can now reject H_0 at the 1 per cent level (but not at the 0.1 per cent level).
2. $t = 4.58$. Not affected by the change. Reject at the 0.1 per cent level.
3. $t = 0.83$. Not affected by the change. Fail to reject at the 5 per cent level.
4. $t = -2.25$. Fail to reject H_0 at the 5 per cent level. Here there is a problem because the coefficient has an unexpected sign and is large enough to reject H_0 at the 5 per cent level with a two-sided test.

In principle we should ignore this and fail to reject H_0. Admittedly, the likelihood of such a large negative *t* statistic occurring under H_0 is very small, but it would be smaller still under the alternative hypothesis $H_1: \beta_2 > 0$.

However we should consider two further possibilities. One is that the justification for a one-sided test is incorrect. For example, some jobs pay relatively low wages because they offer training that is valued by the employee. Apprenticeships are the classic example. Alternatively, workers in some low-paid occupations may, for technical reasons, receive a relatively large amount of training. In either case, the correlation between training and earnings might be negative instead of positive.

Another possible reason for a coefficient having an unexpected sign is that the model is misspecified in some way. For example, the coefficient might be distorted by omitted variable bias, to be discussed in Chapter 6.

2.25

Suppose that the true relationship between *Y* and *X* is $Y_i = \beta_1 + \beta_2 X_i + u_i$ and that the fitted model is $\hat{Y}_i = b_1 + b_2 X_i$. In Exercise 1.12 it was shown that if $X_i^* = \mu_1 + \mu_2 X_i$, and *Y* is regressed on X^*, the slope coefficient $b_2^* = b_2 / \mu_2$. How will the *t* statistic for b_2^* be related to the *t* statistic for b_2? (See also Exercise 2.13.)

Answer:

From Exercise 2.13, we have $\text{s.e.}(b_2^*) = \text{s.e.}(b_2)/\mu_2$. Since $b_2^* = b_2/\mu_2$, it follows that the *t* statistic must be the same.

Alternatively, since we saw in Exercise 1.22 that R^2 must be the same, it follows that the *F* statistic for the equation must be the same. For a simple regression the *F* statistic is the square of the *t* statistic on the slope coefficient, so the *t* statistic must be the same.

2.28

Calculate the 95 per cent confidence interval for β_2 in the price inflation/wage inflation example:

$$\begin{matrix}\hat{p} = & -1.21 + & 0.82w \\ & (0.05) & (0.10)\end{matrix}.$$

What can you conclude from this calculation?

Answer:

With *n* equal to 20, there are 18 degrees of freedom and the critical value of *t* at the 5 per cent level is 2.10. The 95 per cent confidence interval is therefore

$$0.82 - 0.10 \times 2.10 \leq \beta_2 \leq 0.82 + 0.10 \times 2.10$$

that is,

$$0.61 \leq \beta_2 \leq 1.03.$$

2.34

Suppose that the true relationship between *Y* and *X* is $Y_i = \beta_1 + \beta_2 X_i + u_i$ and that the fitted model is $\hat{Y}_i = b_1 + b_2 X_i$. Suppose that $X_i^* = \mu_1 + \mu_2 X_i$, and *Y* is regressed on X^*. How will the *F* statistic for this regression be related to the *F* statistic for the original regression? (See also Exercises 1.22, 2.13, and 2.24.)

Answer:

We saw in Exercise 1.22 that R^2 would be the same, and it follows that *F* must also be the same.

Answers to the additional exercises

Note:

Each of the exercises below relates to a simple regression. Accordingly, the *F* test is equivalent to a two-sided *t* test on the slope coefficient and there is no point in performing both tests. The *F* statistic is equal to the square of the *t* statistic and, for any significance level, the critical value of *F* is equal to the critical value of *t*. Obviously a one-sided *t* test, when justified, is preferable to either in that it has greater power for any given significance level.

A2.1

First we need to show that $E(b_2) = 0$.

$$b_2 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{\sum_i (X_i - \bar{X})(\beta_1 + u_i - \beta_1 - \bar{u})}{\sum_i (X_i - \bar{X})^2} = \frac{\sum_i (X_i - \bar{X})(u_i - \bar{u})}{\sum_i (X_i - \bar{X})^2}.$$

Hence, given that we are told that X is nonstochastic,

$$E(b_2) = E\left(\frac{\sum_i (X_i - \bar{X})(u_i - \bar{u})}{\sum_i (X_i - \bar{X})^2}\right) = \frac{1}{\sum_i (X_i - \bar{X})^2} E\left(\sum_i (X_i - \bar{X})(u_i - \bar{u})\right)$$

$$= \frac{1}{\sum_i (X_i - \bar{X})^2} \sum_i (X_i - \bar{X}) E(u_i - \bar{u}) = 0$$

since $E(u) = 0$. Thus

$$E(b_1) = E(\bar{Y} - b_2 \bar{X}) = \beta_1 - \bar{X}E(b_2) = \beta_1$$

and the estimator is unbiased.

A2.2

If $\bar{X} = 0$, the estimators are identical. $\bar{Y} - b_2\bar{X}$ reduces to $\bar{Y}$.

A2.3

The *t* statistic for the coefficient of *SM* is –7.98, very highly significant. The *t* statistic for the intercept is even higher, but it is of no interest. All the mothers in the sample must have had at least one child (the respondent), for otherwise they would not be in the sample. The *F* statistic is 63.60, very highly significant.

A2.4

The *t* statistic for the coefficient of *EXP* is 19.50, very highly significant. There is little point performing a *t* test on the intercept, given that it has no plausible meaning. The *F* statistic is 380.4, very highly significant.

A2.5

The slope coefficient for every category is significantly different from zero at a very high significance level, with the exception of local public transportation. The coefficient for the latter is virtually equal to zero and the *t* statistic is only 0.40. Evidently this category is on the verge of being an inferior good.

A2.6

A straight *t* test on the coefficient of *WEIGHT85* is not very interesting since we know that those who were relatively heavy in 1985 will also be relatively heavy in 2002. The *t* statistic confirms this obvious fact. Of more interest would be a test investigating whether those relatively heavy in 1985 became even heavier in 2002. We set up the null hypothesis that they did not, H_0: $\beta_2 = 1$, and see if we can reject it. The *t* statistic for this test is

$$t = \frac{1.0134 - 1}{0.0299} = 0.45$$

and hence the null hypothesis is not rejected. The constant indicates that the respondents have tended to put on 23.6 pounds over the interval, irrespective of their 1985 weight. The null hypothesis that the general increase is zero is rejected at the 0.1 per cent level.

A2.7

The *t* statistic for height, 5.42, suggests that the effect of height on earnings is highly significant, despite the very low R^2. In principle the estimate of an extra 86 cents of hourly earnings for every extra inch of height could have been a purely random result of the kind that one obtains with nonsense models. However, the fact that it is apparently highly significant causes us to look for other explanations, the most likely one being that suggested in the answer to Exercise A1.4. Of course we would not attempt to test the negative constant.

A2.8

The standard error of the coefficient of GDP. This is given by

$$\frac{s_u^*}{\sqrt{\sum\left(G_i^* - \overline{G}^*\right)^2}},$$

where s_u^*, the standard error of the regression, is estimated as $\frac{\sum e_i^{*2}}{n-2}$. Since *RSS* is unchanged, $s_u^* = s_u$.

We saw in Exercise A1.5 that $G_i^* - \overline{G}^* = G_i - \overline{G}$ for all *i*. Hence the new standard error is given

by $\frac{s_u}{\sqrt{\sum\left(G_i - \overline{G}\right)^2}}$ and is unchanged.

$F = \frac{ESS}{RSS/n-2}$ where *ESS* = explained sum of squares = $\sum\left(\hat{Y}_i^* - \overline{\hat{Y}}^*\right)^2$.

Since $e_i^* = e_i$, $\hat{Y}_i^* = \hat{Y}_i$ and *ESS* is unchanged. We saw in Exercise A1.5 that *RSS* is unchanged. Hence *F* is unchanged.

A2.9

The standard error of the coefficient of GDP. This is given by

$$\frac{s_u^*}{\sqrt{\sum\left(G_i^* - \overline{G}^*\right)^2}},$$

where s_u^*, the standard error of the regression, is estimated as $\frac{\sum e_i^{*2}}{n-2}$,

where *n* is the number of observations. We saw in Exercise 1.6 that $e_i^* = e_i$ and so *RSS* is unchanged. Hence $s_u^* = s_u$. Thus the new standard error is given by

$$\frac{s_u}{\sqrt{\sum\left(2G_i - 2\overline{G}\right)^2}} = \frac{1}{2}\frac{s_u}{\sqrt{\sum\left(G_i - \overline{G}\right)^2}} = 0.005.$$

$F = \frac{ESS}{RSS/n-2}$ where *ESS* = explained sum of squares = $\sum\left(\hat{Y}_i^* - \overline{\hat{Y}}^*\right)^2$.

Since $e_i^* = e_i$, $\hat{Y}_i^* = \hat{Y}_i$ and *ESS* is unchanged. Hence *F* is unchanged.

A2.10

One way of demonstrating that $\hat{Y}_i^* = \hat{Y}_i - \overline{Y}$:

$$\hat{Y}_i^* = b_1^* + b_2^* X_i^* = b_2\left(X_i - \overline{X}\right)$$

$$\hat{Y}_i - \overline{Y} = \left(b_1 + b_2 X_i\right) - \overline{Y} = \left(\overline{Y} - b_2\overline{X}\right) + b_2 X_i - \overline{Y} = b_2\left(X_i - \overline{X}\right).$$

Demonstration that the residuals are the same:

$$e_i^* = Y_i^* - \hat{Y}_i^* = \left(Y_i - \overline{Y}\right) - \left(\hat{Y}_i - \overline{Y}\right) = e_i \,.$$

Demonstration that the OLS estimator of the variance of the disturbance term in (2) is equal to that in (1):

$$s_u^{*2} = \frac{\sum e_i^{*2}}{n-2} = \frac{\sum e_i^2}{n-2} = s_u^2 \,.$$

The standard error of the slope coefficient in (2) is equal to that in (1).

$$\hat{\sigma}_{b_2^*}^2 = \frac{s_u^{*2}}{\sum\left(X_i^* - \overline{X}^*\right)^2} = \frac{s_u^2}{\sum X_i^{*2}} = \frac{s_u^2}{\sum\left(X_i - \overline{X}\right)^2} = \hat{\sigma}_{b_2}^2 \,.$$

Hence the standard errors are the same.

Demonstration that R^2 in (2) is equal to R^2 in (1):

$$R^{2^*} = \frac{\sum\left(\hat{Y}_i^* - \overline{\hat{Y}^*}\right)^2}{\sum\left(Y_i^* - \overline{Y}^*\right)^2}$$

$\hat{Y}_i^* = \hat{Y}_i - \overline{Y}$ and $\overline{\hat{Y}} = \overline{Y}$. Hence $\overline{\hat{Y}^*} = 0$. $\overline{Y}^* = \overline{Y} - \overline{Y} = 0$. Hence

$$R^{2^*} = \frac{\sum\left(\hat{Y}_i^*\right)^2}{\sum\left(Y_i^*\right)^2} = \frac{\sum\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2} = R^2 \,.$$

The reason that specification (2) of the second researcher is incorrect is that the model does not include an intercept.

If the second researcher had fitted (3) instead of (2), this would not in fact have affected his estimator of β_2. Using (3), the researcher should have estimated β_2 as $b_2^* = \dfrac{\sum X_i^* Y_i^*}{\sum X_i^{*2}}$. However, Exercise 1.15 demonstrates

that, effectively, he has done exactly this. Hence the estimator will be the same. It follows that dropping the unnecessary intercept would not have led to a gain in efficiency.

A2.11

$$b_2 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} = \frac{\sum_{i=1}^n X_i\left(\beta_1 + \beta_2 X_i + u_i\right)}{\sum_{i=1}^n X_i^2} = \frac{\beta_1 \sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2} + \beta_2 + \frac{\sum_{i=1}^n X_i u_i}{\sum_{i=1}^n X_i^2} \,.$$

Hence

$$E\left(b_2\right) = \frac{\beta_1 \sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2} + \beta_2 + E\left(\frac{\sum_{i=1}^n X_i u_i}{\sum_{i=1}^n X_i^2}\right) = \frac{\beta_1 \sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2} + \beta_2 + \frac{\sum_{i=1}^n X_i E\left(u_i\right)}{\sum_{i=1}^n X_i^2}$$

assuming that X is nonstochastic. Since $E(u_i) = 0$,

$$E(b_2) = \frac{\beta_1 \sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2} + \beta_2 .$$

Thus b_2 will in general be a biased estimator. The sign of the bias depends on the signs of β_1 and $\sum_{i=1}^{n} X_i$.

We have no information about either of these.

b_2 is more efficient than $\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ unless $\bar{X} = 0$ since its

population variance is $\dfrac{\sigma_u^2}{\sum_{i=1}^{n} X_i^2}$, whereas that of $\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ is

$$\frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sigma_u^2}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} .$$

The expression for the variance of b_2 has a smaller denominator if $\bar{X} \neq 0$.

If $\bar{X} = 0$, the estimators are equally efficient because the population variance expressions are identical. The reason for this is that the estimators are now identical:

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i(Y_i - \bar{Y})}{\sum_{i=1}^{n} X_i^2} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} - \frac{\bar{Y}\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

since $\sum_{i=1}^{n} X_i = n\bar{X} = 0$.

If there is little variation in X in the sample, $\sum_{i=1}^{n}(X_i - \bar{X})^2$

may be small and hence the population variance of $\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

may be large. Thus using a criterion such as mean square error, b_2 may be preferable if the bias is small.

A2.12

The inclusion of the fifth observation does not cause the model to be misspecified or the regression model assumptions to be violated, so retaining it in the sample will not give rise to biased estimates. There would be no advantages in dropping it and there would be one major disadvantage. $\sum_{i=1}^{n}(X_i - \bar{X})^2$ would be greatly reduced and hence the

variances of the coefficients would be increased, adversely affecting the precision of the estimates.

This said, in practice one would wish to check whether it is sensible to assume that the model relating Y to X for the other observations really does apply to the observation corresponding to X_5 as well. This question can be answered only by being familiar with the context and having some intuitive understanding of the relationship between Y and X.