

# Summary, further reading, and keywords

---

## SUMMARY

In this chapter we considered regression models with more than one explanatory variable. The least squares coefficients measure the direct effect of an explanatory variable on the dependent variable after neutralizing for the indirect effects that run via the other explanatory variables. These estimated effects therefore depend on the set of all explanatory variables included in the model. We paid particular attention to the question of which explanatory variables should be included in the model. For reasons of efficiency it is better to exclude variables that have only a marginal effect. The statistical properties of least squares were derived under a number of assumptions on the data generating process. Under these assumptions, the  $F$ -test can be used to test for the individual and joint significance of explanatory variables.

---

## FURTHER READING

In our analysis we made intensive use of matrix methods. We give some references to econometric textbooks that also follow this approach. Chow (1983), Greene (2000), Johnston and DiNardo (1997), Stewart and Gill (1998), Verbeek (2000), and Wooldridge (2002) are on an intermediate level; the other books are on an advanced level. The handbooks edited by Griliches and Intriligator contain overviews of many topics that are treated in this and the next chapters.

- Chow, G. G. (1983). *Econometrics*. Auckland: McGraw-Hill.
- Davidson, R., and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press.
- Gourieroux, C., and Monfort, A. (1995). *Statistics and Econometric Models*. 2 vols. Cambridge: Cambridge University Press.
- Greene, W. H. (2000). *Econometric Analysis*. New York: Prentice Hall.
- Griliches, Z., and Intriligator, M. D. (1983, 1984, 1986). *Handbook of Econometrics*. 3 vols. Amsterdam: North-Holland.
- Johnston, J., and DiNardo, J. (1997). *Econometric Methods*. New York: McGraw-Hill.
- Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., and Lee, T. C. (1985). *The Theory and Practice of Econometrics*. New York: Wiley.

- Malinvaud, E. (1980). *Statistical Methods of Econometrics*. Amsterdam: North Holland.
- Mittelhammer, R. C., Judge, G. G., and Miller, D. J. (2000). *Econometric Foundations*. Cambridge: Cambridge University Press.
- Ruud, P. A. (2000). *An Introduction to Classical Econometric Theory*. New York: Oxford University Press.
- Stewart, J., and Gill, L. (1998). *Econometrics*. London: Prentice Hall.
- Theil, H. (1971). *Principles of Econometrics*. New York: Wiley.
- Verbeek, M. (2000). *A Guide to Modern Econometrics*. Chichester: Wiley.
- Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

---

## KEYWORDS

- |                                  |                                      |
|----------------------------------|--------------------------------------|
| auxiliary regressions 140        | omitted variables bias 143           |
| <i>ceteris paribus</i> 140       | partial regression 146               |
| Chow forecast test 173           | partial regression scatter 148       |
| coefficient of determination 129 | prediction interval 171              |
| covariance matrix 126            | predictive performance 169           |
| degrees of freedom 129           | projection 123                       |
| direct effect 140                | significance 153                     |
| <i>F</i> -test 161               | significance of the regression 164   |
| Frisch–Waugh 146                 | standard error 128                   |
| indirect effect 140              | standard error of the regression 128 |
| inefficient 144                  | <i>t</i> -test 153                   |
| joint significance 164           | <i>t</i> -value 153                  |
| least squares estimator 122      | total effect 140                     |
| linear restrictions 165          | true model 142                       |
| matrix form 120                  | unbiased 126                         |
| minimal variance 127             | uncontrolled 140                     |
| multicollinearity 158            | variance inflation factor 159        |
| normal equations 121             |                                      |

# Exercises

## THEORY QUESTIONS

### 3.1 (☞ Section 3.1.2)

In this exercise we study the derivatives of (3.6) and prove the result in (3.7). For convenience, we write  $X'y = p$  (a  $k \times 1$  vector) and  $X'X = Q$  (a  $k \times k$  matrix), so that we have to minimize the function  $f(b) = y'y - p'b - b'p + b'Qb = y'y - 2b'p + b'Qb$ . Check every detail of the following argument.

- Let  $b$  increase to  $b + h$ , where we may choose the elements of the  $k \times 1$  vector  $h$  as small as we like. Then  $f(b+h) = f(b) + h'(-2p + (Q' + Q)b) + h'Qh$ .
- This result can be interpreted as a Taylor expansion. If the elements of  $h$  are sufficiently small, the last term can be neglected, and the central term is a linear expression containing the  $k \times 1$  vector of first order derivatives  $\frac{\partial f}{\partial b} = -2p + (Q' + Q)b$ . There are  $k$  first order derivatives and we follow the convention to arrange them in a column vector.
- If we apply this to (3.6), this shows that  $\frac{\partial S}{\partial b} = -2X'y + 2X'Xb$ .

### 3.2 (☞ Section 3.1.2)

In this exercise we prove the result in (3.10). The vector of first order derivatives in (3.7) contains one term that depends on  $b$ . For convenience we write it as  $Qb$  and we partition the  $k \times k$  matrix  $Q = 2X'X$  into its columns as  $Q = (q_1 \ q_2 \ \dots \ q_k)$ . Verify each step in the following argument.

- $Qb$  can be written as  $Qb = q_1b_1 + q_2b_2 + \dots + q_kb_k$ .
- The derivatives of the elements of  $Qb$  with respect to the scalar  $b_i$  can be written as a column  $q_i$ . To write all derivatives for  $i = 1, \dots, k$  in one formula we follow the convention to write them as a 'row of columns'—that is, we group them into a matrix, so that  $\frac{\partial Qb}{\partial b'} = Q$  (note the prime in the left-hand denominator; this indicates that the separate derivatives are arranged as a row).

- With the same conventions we get  $\frac{\partial^2 S}{\partial b \partial b'} = Q$  for the Hessian.
- Let  $X$  be an  $n \times k$  matrix with rank  $k$ ; then prove that the  $k \times k$  matrix  $X'X$  is positive definite.

### 3.3 (☞ Section 3.1.2)

The following steps show that the least squares estimator  $b = (X'X)^{-1}X'y$  minimizes (3.6) without using the first and second order derivatives. In this exercise  $b_*$  denotes any  $k \times 1$  vector.

- Let  $b_* = (X'X)^{-1}X'y + d$ ; then show that  $y - Xb_* = e - Xd$ , where  $e$  is a vector of constants that does not depend on the choice of  $d$ .
- Show that  $S(b_*) = e'e + (Xd)'(Xd)$  and that the minimum of this expression is attained if  $Xd = 0$ .
- Derive the condition for uniqueness of this minimum and show that the minimum is then given by  $d = 0$ .

### 3.4 (☞ Section 3.1.4)

- In the model  $y = X\beta + \varepsilon$ , the normal equations are given by  $X'Xb = X'y$ , the least squares estimates by  $b = (X'X)^{-1}X'y$ , and the variance by  $\text{var}(b) = \sigma^2(X'X)^{-1}$ . Work these three formulas out for the special case of the simple regression model  $y_i = \alpha + \beta x_i + \varepsilon_i$  and prove that these results are respectively equal to the normal equations, the estimates  $a$  and  $b$ , and the variances of  $a$  and  $b$  obtained in Sections 2.1.2 and 2.2.4.
- Suppose that the  $k$  random variables  $y, x_2, x_3, \dots, x_k$  are jointly normally distributed with mean  $\mu$  and (non-singular) covariance matrix  $\Sigma$ . Let the observations be obtained by a random sample of size  $n$  from this distribution  $N(\mu, \Sigma)$ . Define the random variable  $y_c = y | \{x_2, \dots, x_k\}$ —that is,  $y$  conditional on the values of  $x_2, \dots, x_k$ . Show that the  $n$  observations  $y_c$  satisfy Assumptions 1–7 of Section 3.1.4.

## 3.5 (☞ Section 3.1.5)

In some software packages the user is asked to specify the variable to be explained and the explanatory variables, while an intercept is added automatically. Now suppose that you wish to compute the least squares estimates  $b$  in a regression of the type  $y = X\beta + \varepsilon$  where the  $n \times k$  matrix  $X$  does *not* contain an 'intercept column' consisting of unit elements. Define

$$y_* = \begin{pmatrix} y \\ -y \end{pmatrix}, \quad X_* = \begin{pmatrix} \iota & X \\ \iota & -X \end{pmatrix},$$

where the  $\iota$  columns, consisting of unit elements only, are added by the computer package and the user specifies the other data.

- Prove that the least squares estimator obtained by regressing  $y_*$  on  $X_*$  gives the desired result.
- Prove that the standard errors of the regression coefficients of this regression must be corrected by a factor  $\sqrt{(2n - k - 1)/(n - k)}$ .

## 3.6 (☞ Section 3.1.6)

Suppose we wish to explain a variable  $y$  and that the number of possible explanatory variables is so large that it is tempting to take a subset. In such a situation some researchers apply the so-called Theil criterion and maximize the adjusted  $R^2$  defined by  $\bar{R}^2 = 1 - \frac{n-1}{n-k}(1 - R^2)$ , where  $n$  is the number of observations and  $k$  the number of explanatory variables.

- Prove that  $R^2$  never decreases by including an additional regressor in the model.
- Prove that the Theil criterion is equivalent with minimizing  $s$ , the standard error of regression.
- Prove that the Theil criterion implies that an explanatory variable  $x_j$  will be maintained if and only if the  $F$ -test statistic for the null hypothesis  $\beta_j = 0$  is larger than one.
- Show that the size (significance level) of such a test is larger than 0.05.

## 3.7 (☞ Sections 3.1.5, 3.1.6, 3.2.4, 3.4.1, 3.4.3)

Some of the following questions and arguments were mentioned in this chapter.

- Prove the result stated in Section 3.1.5 that  $b_i > 0$  if the  $n \times k$  matrix  $X$  contains a column of unit elements and  $\text{rank}(X) = k$ .
- Prove that  $R^2$  (in the model with constant term) is the square of the sample correlation coefficient between  $y$  and  $\hat{y} = Xb$ .

c. If a regression model contains no constant term so that the matrix  $X$  contains no column of ones, then show that  $1 - (SSR/SST)$  (and hence  $R^2$  when it is computed in this way) may be negative.

d. Let  $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$  and let  $\beta_1$  be estimated by regressing  $y$  on  $X_1$  alone (the 'omitted variables' case of Section 3.2.3). Show that  $\text{var}(b_R) \leq \text{var}(b_1)$  in the sense that  $\text{var}(b_1) - \text{var}(b_R)$  is a positive semidefinite matrix. When are the two variances equal?

e. Show that the  $F$ -test for a single restriction  $\beta_j = 0$  is equal to the square of the  $t$ -value of  $b_j$ . Show also that both tests lead to the same conclusion, irrespective of the chosen significance level.

f\*. Consider the expression (3.49) of the  $F$ -test in terms of the random variables  $b'_2 X'_2 M_1 X_2 b_2$  and  $e'e$ . Prove that, under the null hypothesis that  $\beta_2 = 0$ , these two random variables are independently distributed as  $\chi^2(g)$  and  $\chi^2(n - k)$  respectively by showing that (i) they can be expressed as  $e'Q_1\varepsilon$  and  $e'Q_2\varepsilon$ , with (ii)  $Q_1 = M_1 - M$  and  $Q_2 = M$ , where  $M$  is the  $M$ -matrix corresponding to  $X$  and  $M_1$  is the  $M$ -matrix corresponding to  $X_1$ , so that (iii)  $Q_1$  is idempotent with rank  $g$  and  $Q_2$  is idempotent with rank  $(n - k)$ , and (iv)  $Q_1 Q_2 = 0$ .

g. In Section 3.4 we considered the prediction of  $y_2$  for given values of  $X_2$  under the assumptions that  $y_1 = X_1\beta + \varepsilon_1$  and  $y_2 = X_2\beta + \varepsilon_2$  where  $E[\varepsilon_1] = 0$ ,  $E[\varepsilon_2] = 0$ ,  $E[\varepsilon_1\varepsilon'_1] = \sigma^2 I$ ,  $E[\varepsilon_2\varepsilon'_2] = \sigma^2 I$ , and  $E[\varepsilon_1\varepsilon'_2] = 0$ . Prove that under Assumptions 1–6 the predictor  $X_2 b$  with  $b = (X'_1 X_1)^{-1} X'_1 y_1$  is best linear unbiased. That is, among all predictors of the form  $\hat{y}_2 = L y_1$  (with  $L$  a given matrix) with the property that  $E[y_2 - \hat{y}_2] = 0$ , it minimizes the variance of the forecast error  $y_2 - \hat{y}_2$ .

h. Using the notation introduced in Section 3.4.3, show that a  $(1 - \alpha)$  prediction interval for  $y_{2j}$  is given by  $X'_{2j} b \pm cs\sqrt{d_{jj}}$ .

## 3.8 (☞ Section 3.4.1)

Consider the model  $y = X\beta + \varepsilon$  with the null hypothesis that  $R\beta = r$  where  $R$  is a given  $g \times k$  matrix of rank  $g$  and  $r$  is a given  $g \times 1$  vector. Use the following steps to show that the expression (3.54) for the  $F$ -test can be written in terms of residual sums of squares as in (3.50).

- The restricted least squares estimator  $b_R$  minimizes the sum of squares  $(y - X\hat{\beta})'(y - X\hat{\beta})$  under the restriction that  $R\hat{\beta} = r$ . Show that

$b_R = b - A(Rb - r)$ , where  $b$  is the unrestricted least squares estimator and  $A = (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}$ .

- b. Let  $e = y - Xb$  and  $e_R = y - Xb_R$ ; then show that  $e'_R e_R = e'e + (Rb - r)'[R(X'X)^{-1}R']^{-1}(Rb - r)$ .
- c. Show that the  $F$ -test in (3.54) can be written as in (3.50).
- d. In Section 3.4.2 we tested the null hypothesis that  $\beta_4 + \beta_5 = 0$  in the model with  $k = 5$  explanatory variables. Describe a method to determine the restricted sum of squared residuals  $e'_R e_R$  in this case.

**3.9** (☞ Section 3.2.5)

This exercise serves to clarify a remark on standard errors in partial regressions that was made in Example 3.3 (p. 150). We use the notation of Section 3.2.5, in particular the estimated regressions

- (1)  $y = X_1 b_1 + X_2 b_2 + e$ , and
- (2)  $M_2 y = (M_2 X_1) b_* + e_*$

in the result of Frisch–Waugh. Here  $X_1$  and  $M_2 X_1$  are  $n \times (k - g)$  matrices and  $X_2$  is an  $n \times g$  matrix.

- a. Prove that  $\text{var}(b_1) = \text{var}(b_*) = \sigma^2(X_1' M_2 X_1)^{-1}$ .
- b. Derive expressions for the estimated variance  $s^2$  in regression (1) and  $s_*^2$  in regression (2), both in terms of  $e'e$ .
- c. Prove that the standard errors of the coefficients  $b_1$  in (1) can be obtained by multiplying the standard errors of the coefficients  $b_*$  in (2) by the factor  $\sqrt{(n - k + g)/(n - k)}$ .
- d. Check this result by considering the standard errors of the variable education in the second regression in Exhibit 3.7 and the last regression in Exhibit 3.10. (These values are rounded; a more precise result is obtained when higher precision values from a regression package are used).
- e. Derive the relation between the  $t$ -values of (1) and (2).

**3.10** (☞ Section 3.4.1)

In Section 1.4.2 we mentioned the situation of two independent random samples, one of size  $n_1$  from  $N(\mu_1, \sigma^2)$  and a second one of size  $n_2$  from  $N(\mu_2, \sigma^2)$ . We want to test the null hypothesis  $H_0: \mu_1 = \mu_2$  against the alternative  $H_1: \mu_1 \neq \mu_2$ . The pooled  $t$ -test is based on the difference between the sample means  $\bar{y}_1$  and  $\bar{y}_2$  of the two

sub-samples. Let  $e'_1 e_1 = \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2$  and  $e'_2 e_2 = \sum_{i=n_1+1}^{n_1+n_2} (y_i - \bar{y}_2)^2$  be the total sum of squares in the first and second sub-sample respectively; then the pooled estimator of the variance is defined by  $s_p^2 = (e'_1 e_1 + e'_2 e_2)/(n_1 + n_2 - 2)$  and the pooled  $t$ -test is defined by

$$t_p = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\bar{y}_2 - \bar{y}_1}{s_p}.$$

- a. Formulate the testing problem of  $\mu_1 = \mu_2$  against  $\mu_1 \neq \mu_2$  in terms of a parameter restriction in a multivariate regression model (with parameters  $\mu_1$  and  $\mu_2$ ).
- b. Derive the  $F$ -test for  $H_0: \mu_1 = \mu_2$  in the form (3.50).
- c. Prove that  $t_p^2$  is equal to the  $F$ -test in **b** and that  $t_p$  follows the  $t(n_1 + n_2 - 2)$  distribution if the null hypothesis of equal means holds true.
- d. In Example 1.12 (p. 62) we considered the FGPA scores of  $n_1 = 373$  male students and  $n_2 = 236$  female students. Use the results reported in Exhibit 1.6 to perform a test of the null hypothesis of equal means for male and female students against the alternative that female students have on average higher scores than male students.

**3.11** (☞ Section 3.4.3)

We consider the Chow forecast test (3.58) for the case  $g = 1$  of a single new observation  $(x_{n+1}, y_{n+1})$ . The  $n$  preceding observations are used in the model  $y_1 = X_1 \beta + \varepsilon$  with least squares estimator  $b$ . We assume that Assumptions 1–4 and 7 are satisfied for the full sample  $i = 1, \dots, n + 1$ , and Assumptions 5 and 6 for the estimation sample  $i = 1, \dots, n$ , whereas for the  $(n + 1)$ st observation we write

$$y_{n+1} = x'_{n+1} \beta + \gamma + \varepsilon_{n+1}$$

with  $\gamma$  an unknown scalar parameter. We consider the null hypothesis that  $\gamma = 0$  against the alternative that  $\gamma \neq 0$ .

- a. Prove that the least squares estimators of  $\beta$  and  $\gamma$  over the full sample  $i = 1, \dots, n + 1$ , are given by  $b$  and  $c = y_{n+1} - x'_{n+1} b$ . Show that the residual for the  $(n + 1)$ st observation is equal to zero. Provide an intuitive explanation for this result.

- b. Derive the residual sum of squares over the full sample  $i = 1, \dots, n + 1$  under the alternative hypothesis.
- c. Derive the  $F$ -test for the hypothesis that  $\gamma = 0$ .

## EMPIRICAL AND SIMULATION QUESTIONS

### 3.12 (☞ Section 3.3.3)

In this simulation exercise we consider five variables ( $y$ ,  $z$ ,  $x_1$ ,  $x_2$ , and  $x_3$ ) that are generated as follows. Let  $n = 100$  and let  $\varepsilon_i, \omega_i, \eta_i \sim \text{NID}(0, 1)$  be independent random samples from the standard normal distribution,  $i = 1, \dots, n$ . Define

$$x_{1i} = 5 + \omega_i + 0.3\eta_i$$

$$x_{2i} = 10 + \omega_i$$

$$x_{3i} = 5 + \eta_i$$

$$y_i = x_{1i} + x_{2i} + \varepsilon_i$$

$$z_i = x_{2i} + x_{3i} + \varepsilon_i$$

- What is the correlation between  $x_1$  and  $x_3$ ? And what is the correlation between  $x_2$  and  $x_3$ ?
- Perform the regression of  $y$  on a constant,  $x_1$  and  $x_2$ . Compute the regression coefficients and their  $t$ -values. Comment on the outcomes.
- Answer the questions of **b** for the regression of  $z$  on a constant,  $x_2$  and  $x_3$ .
- Perform also regressions of  $y$  on a constant and  $x_1$ , and of  $z$  on a constant and  $x_3$ . Discuss the differences that arise between these two cases.

### 3.13 (☞ Section 3.4.1)

In Section 3.4.2 we tested four different hypotheses—that is, (i)  $\beta_5 = 0$ , (ii)  $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ , (iii)  $\beta_4 = \beta_5 = 0$ , and (iv)  $\beta_4 + \beta_5 = 0$ . As data set we considered the data on all 474 employees (see Exhibit 3.16). Use a significance level of 5 per cent in all tests below.

- Test these four hypotheses also for the subset of employees working in management (job category 3), using the results in the last two columns in Exhibit 3.16.
- Now consider the hypothesis (iii) that gender and minority have no effect on salary for employees in management. We mention that of the eighty-four employees in management, seventy are male non-minority, ten are female-non-minority, four are male-minority, and no one is female-

minority. Discuss the relevance of this information with respect to the power of the test for hypothesis (iii).

- Finally consider the subset of employees with custodial jobs (job category 2, where all employees are male). Use the results in Exhibit 3.16 to test the hypothesis that  $\beta_5 = 0$ . Test also the hypothesis that  $\beta_2 = \beta_3 = \beta_5 = 0$ .

### 3.14 (☞ Sections 3.2.2, 3.3.3)

In this exercise we consider the data set on student learning of Example 1.1 (p. 12) for 609 students. The dependent variable ( $y$ ) is the FGPA score of a student, and the explanatory variables are  $x_1$  (constant term),  $x_2$  (SATM score),  $x_3$  (SATV score), and  $x_4$  (FEM, with  $x_4 = 1$  for females and  $x_4 = 0$  for males).

- Compute the  $4 \times 4$  correlation matrix for the variables ( $y$ ,  $x_2$ ,  $x_3$ ,  $x_4$ ).
- Estimate a model for FGPA in terms of SATV by regressing  $y$  on  $x_1$  and  $x_3$ . Estimate also a model by regressing  $y$  on  $x_1$ ,  $x_2$ ,  $x_3$ , and  $x_4$ .
- Comment on the differences between the two models in **b** for the effect of SATV on FGPA.
- Investigate the presence of collinearity between the explanatory variables by computing  $R_j^2$  in (3.47) and the square root of the variance inflation factors,  $1/\sqrt{1 - R_j^2}$ , for  $j = 2, 3, 4$ .

### 3.15 (☞ Section 3.4.1)

In this exercise we consider production data for the year 1994 of  $n = 26$  US firms in the sector of primary metal industries (SIC33). The data are taken from E. J. Bartelsman and W. Gray, National Bureau of Economic Research, NBER Technical Working Paper 205, 1996. For each firm, values are given of production ( $Y$ , value added in millions of dollars), labour ( $L$ , total payroll in millions of dollars), and capital ( $K$ , real capital stock in millions of 1987 dollars). A log-linear production function is estimated with the following result (standard errors are in parentheses).



$$\log(Y) = 0.701 + 0.756 \log(L) + 0.242 \log(K) + e$$

(0.415) (0.091) (0.110)

The model is also estimated under two alternative restrictions, the first with equal coefficients for  $\log(L)$  and  $\log(K)$  and the second with the sum of the coefficients of  $\log(L)$  and  $\log(K)$  equal to one ('constant returns to scale'). For this purpose the following two regressions are performed.

$$\log(Y) = 0.010 + 0.524(\log(L) + \log(K)) + e_1$$

(0.358) (0.026)

$$\log(Y) - \log(K) = 0.686 + 0.756(\log(L) - \log(K)) + e_2$$

(0.132) (0.089)

The residual sums of squares are respectively  $e'e = 1.825544$ ,  $e_1'e_1 = 2.371989$ , and  $e_2'e_2 = 1.825652$ , and the  $R^2$  are respectively equal to  $R^2 = 0.956888$ ,  $R_1^2 = 0.943984$ , and  $R_2^2 = 0.751397$ . In the following tests use a significance level of 5%.

- Test for the individual significance of  $\log(L)$  and  $\log(K)$  in the first regression. Test also for the joint significance of these two variables.
- Test the restriction of equal coefficients by means of an  $F$ -test based on the residual sums of squares.
- Test this restriction also by means of the  $R^2$ .
- Test the restriction of constant returns to scale also in two ways, one with the  $F$ -test based on the residual sums of squares and the other with the  $F$ -test based on the  $R^2$ .
- Explain why the outcomes of **b** and **c** are the same but the two outcomes in **d** are different. Which of the two tests in **d** is the correct one?

### 3.16 (☞ Section 3.2.5)

Consider the data on bank wages of the example in Section 3.1.7. To test for the possible effect of gender on wage, someone proposes to estimate the model  $y = \beta_1 + \beta_4 x_4 + \varepsilon$ , where  $y$  is the yearly wage (in logarithms) and  $x_4$  is the variable gender (with  $x_4 = 0$  for females and  $x_4 = 1$  for males). As an alternative we consider the model with  $x_2$  (education) as an additional explanatory variable.

- Use the data to perform the two regressions.

- Comment on the differences between the conclusions that could be drawn (without further thinking) from each of these two regressions.
- Draw a partial regression scatter plot (with regression line) for salary (in logarithms) against gender after correction for the variable education (see Case 3 in Section 3.2.5). Draw also a scatter plot (with regression line) for the original (uncorrected) data on salary (in logarithms) and gender. Discuss how these plots help in clarifying the differences in **b**.
- Check the results on regression coefficients and residuals in the result of Frisch–Waugh (3.39) for these data, where  $X_1$  refers to the variable  $x_4$ , and  $X_2$  refers to the constant term and the variable  $x_2$ .

### 3.17 (☞ Section 3.4.3)

In this exercise we consider data on weekly coffee sales of a certain brand of coffee. These data come from the same marketing experiment as discussed in Example 2.3 (p. 78), but for another brand of coffee and for another selection of weeks. The data provide for  $n = 18$  weeks the values of the coffee sales in that week ( $Q$ , in units), the applied deal rate ( $D = 1$  for the usual price,  $D = 1.05$  in weeks with 5% price reduction, and  $D = 1.15$  in weeks with 15% price reduction), and advertisement ( $A = 1$  in weeks with advertisement,  $A = 0$  otherwise). We postulate the model

$$\log(Q) = \beta_1 + \beta_2 \log(D) + \beta_3 A + \varepsilon.$$

For all tests below use a significance level of 5%.

- Test whether advertisement has a significant effect on sales, both by a  $t$ -test and by an  $F$ -test.
- Test the null hypothesis that  $\beta_2 = 1$  against the alternative that  $\beta_2 > 1$ .
- Construct 95% interval estimates for the parameters  $\beta_2$  and  $\beta_3$ .
- Estimate the model using only observations in weeks without advertisement. Test whether this model produces acceptable forecasts for the sales (in logarithms) in the weeks with advertisement. Note: take special care of the fact that the estimated model can not predict the effect of advertisement.
- Make two scatter plots, one of the actual values of  $\log(Q)$  against the fitted values of **d** for the



twelve observations in the estimation sample, and a second one of  $\log(Q)$  against the predicted values for the six observations in the prediction sample. Relate these graphs to your conclusions in d.

### 3.18 (👉 Section 3.2.5)

In this exercise we consider yearly data (from 1970 to 1999) related to motor gasoline consumption in the USA. The data are taken from different sources (see the table). Here ‘rp’ refers to data in the Economic Report of the President (see [w3.access.gpo.gov](http://w3.access.gpo.gov)), ‘ecocb’ to data of the Census Bureau, and ‘ecode’ to data of the Department of Energy (see [www.economagic.com](http://www.economagic.com)). The price indices are defined so that the average value over the years 1982–4 is equal to 100. We define the variables  $y = \log(SGAS/PGAS)$ ,  $x_2 = \log(INC/PALL)$ ,  $x_3 = \log(PGAS/PALL)$ ,  $x_4 = \log(PPUB/PALL)$ ,  $x_5 = \log(PNCAR/PALL)$ , and  $x_6 = \log(PUCAR/PALL)$ . We are interested in the price elasticity of gasoline consumption — that is, the marginal relative increase in sold quantity due to a marginal relative price increase.



Variable	Definition	Units	Source
SGAS	Retail sales gasoline service stations	$10^6$ dollars	ecocb
PGAS	Motor gasoline retail price, US city average	cts/gallon	ecode
INC	Nominal personal disposable income	$10^9$ dollars	rp
PALL	Consumer price index	$(1982 - 4)/3 = 100$	rp
PPUB	Consumer price index of public transport	idem	rp
PNCAR	Consumer price index of new cars	idem	rp
PUCAR	Consumer price index of used cars	idem	rp

- Estimate this price elasticity by regressing  $\log(SGAS)$  on a constant and  $\log(PGAS)$ . Comment on the outcome, and explain why this outcome is misleading.
- Estimate the price elasticity now by regressing  $y$  on a constant and  $\log(PGAS)$ . Explain the precise relation with the results in a. Why is this outcome still misleading?

- Now estimate the price elasticity by regressing  $y$  on a constant and the variables  $x_2$  and  $x_3$ . Provide a motivation for this choice of explained and explanatory variables and comment on the outcomes.
- If  $y$  is regressed on a constant and the variable  $x_3$  then the estimated elasticity is more negative than in c. Check this result and give an explanation in terms of partial regressions. Use the fact that, in the period 1970–99, real income has mostly gone up and the price of gasoline (as compared with other prices) has mostly gone down.
- Perform the partial regressions needed to remove the effect of income ( $x_2$ ) on the consumption ( $y$ ) and on the relative price ( $x_3$ ). Make a partial regression scatter plot of the ‘cleaned’ variables and check the validity of the result of Frisch–Waugh in this case.
- Estimate the price elasticity by regressing  $y$  on a constant and the variables  $x_2, x_3, x_4, x_5$ , and  $x_6$ . Comment on the outcomes and compare them with the ones in c.
- Transform the four price indices ( $PALL, PPUB, PNCAR$ , and  $PUCAR$ ) so that they all have the value 100 in 1970. Perform the regression of f for the transformed data (taking logarithms again) and compare the outcomes with the ones in f. Which regression statistics remain the same, and which ones have changed? Explain these results.

### 3.19 (👉 Sections 3.4.1, 3.4.3)

We consider the same data on motor gasoline consumption as in Exercise 3.18 and we use the same notation as introduced there. For all tests below, compute sums of squared residuals of appropriate regressions, determine the degrees of freedom of the test statistic, and use a significance level of 5%.

- Regress  $y$  on a constant and the variables  $x_2, x_3, x_4, x_5$ , and  $x_6$ . Test for the joint significance of the prices of new and used cars.
- Regress  $y$  on a constant and the four explanatory variables  $\log(PGAS), \log(PALL), \log(INC)$ , and  $\log(PPUB)$ . Use the results to construct a 95% interval estimate for the price elasticity of gasoline consumption.



- c. Test the null hypothesis that the sum of the coefficients of the four regressors in the model in **b** (except the constant) is equal to zero. Explain why this restriction is of interest by relating this regression model to the restricted regression in **a**.
- d. Show that the following null hypothesis is not rejected: the sum of the coefficients of  $\log(PALL)$ ,  $\log(INC)$ , and  $\log(PPUB)$  in the model of **b** is equal to zero. Show that the restricted model has regressors  $\log(PGAS)$ ,  $x_2$  and  $x_4$  (and a constant term), and estimate this model.
- e. Use the model of **d** (with the constant,  $\log(PGAS)$ ,  $x_2$  and  $x_4$  as regressors) to construct a 95% interval estimate for the price elasticity of gasoline consumption. Compare this with the result in **b** and comment.
- f. Search the Internet to find the most recent year with values of the variables  $SGAS$ ,  $PGAS$ ,  $PALL$ ,  $INC$ , and  $PPUB$  (make sure to use the same units as the ones mentioned in Exercise 3.18). Use the models in **b** and **d** to construct 95% forecast intervals of  $y = \log(SGAS/PGAS)$  for the given most recent values of the regressors.
- g. Compare the most recent value of  $y$  with the two forecast intervals of part **f**. For the two models in **b** and **d**, perform Chow forecast tests for the most recent value of  $y$ .