EXERCISES

4.1. Consider the regression on time, $y_t = \beta_0 + \beta_1 t + \epsilon_t$, with $t = 1, 2, \ldots, n$. Here, the regressor vector is $\boldsymbol{x}' = (1, 2, \ldots, n)$. Take $n = 10$. Write down the matrices $X'X$, $(X'X)^{-1}$, $V(\hat{\boldsymbol{\beta}})$, and the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$.

4.2. For the regression model $y_t = \beta_0 + \epsilon_t$ with $n = 2$ and $\boldsymbol{y}' = (2, 4)$, draw the data in two-dimensional space. Identify the orthogonal projection of $\boldsymbol{y}$ onto $L(X) = L(\mathbf{1})$. Explain geometrically $\hat{\beta}_0$, $\hat{\boldsymbol{\mu}}$, and $\boldsymbol{e}$.

4.3. Consider the regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, 2, 3$. With

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 2.2 \\ 3.9 \\ 3.1 \end{bmatrix}$$

draw the data in three-dimensional space and identify the orthogonal projection of $\boldsymbol{y}$ onto $L(X) = L(\mathbf{1}, \boldsymbol{x})$. Explain geometrically $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}$, and $\boldsymbol{e}$.

4.4. Consider the regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, 2, 3$. With

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$$

draw the data in three-dimensional space and identify the orthogonal projection of $\boldsymbol{y}$ onto $L(X) = L(\mathbf{1}, \boldsymbol{x})$. Explain geometrically $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}$, and $\boldsymbol{e}$.

4.5. After fitting the regression model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

on 15 cases, it is found that the mean square error $s^2 = 3$ and

$$(X'X)^{-1} = \begin{bmatrix} 0.5 & 0.3 & 0.2 & 0.6 \\ 0.3 & 6.0 & 0.5 & 0.4 \\ 0.2 & 0.5 & 0.2 & 0.7 \\ 0.6 & 0.4 & 0.7 & 3.0 \end{bmatrix}$$

Find

a. The estimate of $V(\hat{\beta}_1)$.

b. The estimate of $\text{Cov}(\hat{\beta}_1, \hat{\beta}_3)$.

c. The estimate of $\text{Corr}(\hat{\beta}_1, \hat{\beta}_3)$.

d. The estimate of $V(\hat{\beta}_1 - \hat{\beta}_3)$.

4.6. When fitting the model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

to a set of $n = 15$ cases, we obtained the least squares estimates $\hat{\beta}_0 = 10$, $\hat{\beta}_1 = 12$, $\hat{\beta}_2 = 15$, and $s^2 = 2$. It is also known that

$$(X'X)^{-1} = \begin{bmatrix} 1 & 0.25 & 0.25 \\ 0.25 & 0.5 & -0.25 \\ 0.25 & -0.25 & 2 \end{bmatrix}$$

a. Estimate $V(\hat{\beta}_2)$.

b. Test the hypothesis that $\beta_2 = 0$.

c. Estimate the covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$.

d. Test the hypothesis that $\beta_1 = \beta_2$, using both the t ratio and the 95% confidence interval.

e. The corrected total sum of squares, $\text{SST} = 120$. Construct the ANOVA table and test the hypothesis that $\beta_1 = \beta_2 = 0$. Obtain the percentage of variation in y that is explained by the model.

4.7. Consider a multiple regression model of the price of houses (y) on three explanatory variables: taxes paid (x_1), number of bathrooms (x_2), and square feet (x_3). The incomplete (Minitab) output from a regression on $n = 28$ houses is given as follows:

The regression equation is price $= -10.7 +$ 0.190 taxes $+$ 81.9 baths $+$ 0.101 sqft

Predictor	Coef	SE Coef	t	p
Constant	−10.65	24.02		
taxes	0.18966	0.05623		
baths	81.87	47.82		
sqft	0.10063	0.03125		

Analysis of variance

Source	DF	SS	MS	F	p
Regression	3	504541			
Residual Error					
Total	27	541119			

a. Calculate the coefficient of determination R^2.

b. Test the null hypothesis that all three regression coefficients are zero (H_0: $\beta_1 = \beta_2 = \beta_3 = 0$). Use significance level 0.05.

c. Obtain a 95% confidence interval of the regression coefficient for "taxes." Can you simplify the model by dropping "taxes"? Obtain a 95% confidence interval of the regression coefficient for "baths." Can you simplify the model by dropping "baths"?

4.8. Continuation of Exercise 4.7. The incomplete (Minitab) output from a multiple regression

of the price of houses on the two explanatory variables, taxes paid and square feet, is given as follows:

The regression equation is price = 4.9 + 0.242 taxes + 0.134 sqft

Predictor	Coef	SE Coef	t	p
Constant	4.89	23.08		
taxes	0.24237	0.04884		
sqft	0.13397	0.02537		

Analysis of variance

Source	DF	SS	MS	F	p
Regression	2	500074	250037		
Residual Error					
Total		541119			

a. Calculate the coefficient of determination R^2.

b. Test the null hypothesis that both regression coefficients are zero (H_0: $\beta_1 = \beta_2 = 0$). Use significance level 0.05.

c. Test whether you can omit the variable "taxes" from the regression model. Use significance level 0.05.

d. Comment on the fact that the regression coefficients for taxes and square feet are different than those shown in Exercise 4.7.

4.9. Fitting the regression $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ on $n = 30$ cases leads to the following results:

$$X'X = \begin{bmatrix} 30 & 2{,}108 & 5{,}414 \\ 2{,}108 & 152{,}422 & 376{,}562 \\ 5{,}414 & 376{,}562 & 1{,}015{,}780 \end{bmatrix}$$

$$X'\boldsymbol{y} = \begin{bmatrix} 5{,}263 \\ 346{,}867 \\ 921{,}939 \end{bmatrix} \quad \text{and} \quad \boldsymbol{y}'\boldsymbol{y} = 1{,}148{,}317$$

a. Use computer software to find $(X'X)^{-1}$. Obtain the least squares estimates and their standard errors.

b. Compute the t statistics to test the simple hypotheses that each regression coefficient is zero.

c. Determine the coefficient of variation R^2. (The complete data are given in the file **abrasion**.)

4.10. The following matrices were computed for a certain regression problem:

$$X'X = \begin{bmatrix} 15 & 3{,}626 & 44{,}428 \\ 3{,}626 & 1{,}067{,}614 & 11{,}419{,}181 \\ 44{,}428 & 11{,}419{,}181 & 139{,}063{,}428 \end{bmatrix},$$

$$X'\boldsymbol{y} = \begin{bmatrix} 2{,}259 \\ 647{,}107 \\ 7{,}096{,}619 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 1.2463484 & 2.1296642 \times 10^{-4} & -4.1567125 \times 10^{-4} \\ & 7.7329030 \times 10^{-6} & -7.0302518 \times 10^{-7} \\ & & 1.9771851 \times 10^{-7} \end{bmatrix},$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 3.452613 \\ 0.496005 \\ 0.009191 \end{bmatrix}$$

$$\boldsymbol{y}'\boldsymbol{y} = 394{,}107$$

a. Write down the estimated regression equation. Obtain the standard errors of the regression coefficients.

b. Compute the t statistics to test the simple hypotheses that each regression coefficient is equal to zero. Carry out these tests. State your conclusions.

4.11. A study was conducted to investigate the determinants of survival size of nonprofit U.S. hospitals. Survival size, y, was defined to be the largest U.S. hospital (in terms of the number of beds) exhibiting growth in market share. For the investigation, 10 states were selected at random, and the survival size for nonprofit hospitals in each of the selected states was determined for two time periods t: 1981–1982 and 1984–1985.

Furthermore, the following characteristics were collected on each selected state for each of the two time periods:

x_1 = Percentage of beds that are in for-profit hospitals.

x_2 = Number of people enrolled in health maintenance organizations as a fraction

of the number of people covered by hospital insurance.

x_3 = State population in thousands.

x_4 = Percentage of state that is urban.

The data are given in the file **hospital.**

a. Fit the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

b. The influence of the percentage of beds in for-profit hospitals was of particular interest to the investigators. What does the analysis tell us?

c. What further investigation might you do with this data set. Give reasons?

d. Rather than selecting 10 states at random, how else might you collect the data on survival size? Would your approach be an improvement over the random selection?

4.12. The amount of water used by the production facilities of a plant varies. Observations on water usage and other, possibly related, variables were collected for 17 months. The data are given in the file **water**. The explanatory variables are

TEMP = average monthly temperature(°F)

PROD = amount of production

DAYS = number of operating days in the month

PAYR = number of people on the monthly plant payroll

HOUR = number of hours shut down for maintenance

The response variable is USAGE = monthly water usage (gallons/100).

a. Fit the model containing all five independent variables,

$$y = \beta_0 + \beta_1 \text{TEMP} + \beta_2 \text{PROD} + \beta_3 \text{DAYS} + \beta_4 \text{PAYR} + \beta_5 \text{HOUR} + \epsilon$$

Plot residuals against fitted values and residuals against the case index, and comment about model adequacy.

b. Test the hypothesis that $\beta_1 = \beta_3 = \beta_5 = 0$.

c. Which model or set of models would you suggest for predictive purposes? Briefly justify.

d. Which independent variable seems to be the most important one in determining the amount of water used?

e. Write a **nontechnical** paragraph that summarizes your conclusions about plant water usage that is supported by the data.

4.13. Data on last year's sales (y, in 100,000s of dollars) in 15 sales districts are given in the file **sales**. This file also contains promotional expenditures (x_1, in thousands of dollars), the number of active accounts (x_2), the number of competing brands (x_3), and the district potential (x_4, coded) for each of the districts.

a. A model with all four regressors is proposed:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Interpret the parameters β_0, β_1, and β_4.

b. Fit the proposed model in (a) and calculate estimates of β_i, $i = 0, 1, \ldots, 4$, and σ^2.

c. Test the following hypotheses:

(i) $\beta_4 = 0$; (ii) $\beta_3 = \beta_4 = 0$;
(iii) $\beta_2 = \beta_3$; (iv) $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

d. Consider the reduced (restricted) model with $\beta_4 = 0$. Estimate its coefficients and give an expression for the expected sales.

e. Using the model in (d), obtain a prediction for the sales in a district where $x_1 = 3.0$, $x_2 = 45$, and $x_3 = 10$. Obtain the corresponding 95% prediction interval.

4.14. The survival rate (in percentage) of bull semen after storage is measured at various combinations of concentrations of three materials (additives) that are thought to increase the chance of survival. The data listed below are given in the file **bsemen**.

% Survival (y)	% Weight 1 (x_1)	% Weight 2 (x_2)	% Weight 3 (x_3)
25.5	1.74	5.30	10.80
31.2	6.32	5.42	9.40
25.9	6.22	8.41	7.20
38.4	10.52	4.63	8.50
18.4	1.19	11.60	9.40
26.7	1.22	5.85	9.90

% Survival (y)	% Weight 1 (x_1)	% Weight 2 (x_2)	% Weight 3 (x_3)
26.4	4.10	6.62	8.00
25.9	6.32	8.72	9.10
32.0	4.08	4.42	8.70
25.2	4.15	7.60	9.20
39.7	10.15	4.83	9.40
35.9	1.72	3.12	7.60
26.5	1.70	5.30	8.20

Assume the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$.

a. Compute $X'X$, $(X'X)^{-1}$, and $X'\boldsymbol{y}$.

b. Plot the response y versus each predictor variable. Comment on these plots.

c. Obtain the least squares estimates of $\boldsymbol{\beta}$ and give the fitted equation.

d. Construct a 90% confidence interval for

i. the predicted mean value of y when $x_1 = 3$, $x_2 = 8$, and $x_3 = 9$;

ii. the predicted individual value of y when $x_1 = 3$, $x_2 = 8$, and $x_3 = 9$.

e. Construct the ANOVA table and test for a significant linear relationship between y and the three predictor variables.

4.15. An experiment was conducted to study the toxic action of a certain chemical on silkworm larvae. The relationship of $\log_{10}$ (survival time) to $\log_{10}$(dose) and $\log_{10}$(larvae weight) was investigated. The data, obtained by feeding each larvae a precisely measured dose of the chemical in an aqueous solution and recording the survival time until death, are given in the following table. The data are stored in the file **silkw**.

$\log_{10}$ Survival Time (y)	$\log_{10}$ Dose (x_1)	$\log_{10}$ Weight (x_2)
2.836	0.150	0.425
2.966	0.214	0.439
2.687	0.487	0.301
2.679	0.509	0.325
2.827	0.570	0.371
2.442	0.590	0.093
2.421	0.640	0.140
2.602	0.781	0.406
2.556	0.739	0.364
2.441	0.832	0.156
2.420	0.865	0.247
2.439	0.904	0.278
2.385	0.942	0.141
2.452	1.090	0.289
2.351	1.194	0.193

Assume the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

a. Plot the response y versus each predictor variable. Comment on these plots.

b. Obtain the least squares estimates for $\boldsymbol{\beta}$ and give the fitted equation.

c. Construct the ANOVA table and test for a significant linear relationship between y and the two predictor variables.

d. Which independent variable do you consider to be the better predictor of log(survival time)? What are your reasons?

e. Of the models involving one or both of the independent variables, which do you prefer, and why?

4.16. You are given the following matrices computed for a regression analysis:

$$X'X = \begin{bmatrix} 9 & 136 & 269 & 260 \\ 136 & 2{,}114 & 4{,}176 & 3{,}583 \\ 269 & 4{,}176 & 8{,}257 & 7{,}104 \\ 260 & 3{,}583 & 7{,}104 & 12{,}276 \end{bmatrix}$$

$$X'\boldsymbol{y} = \begin{bmatrix} 45 \\ 648 \\ 1{,}283 \\ 1{,}821 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 9.610 & 0.008 & -0.279 & -0.044 \\ 0.008 & 0.509 & -0.258 & 0.001 \\ -0.279 & -0.258 & 0.139 & 0.001 \\ -0.044 & 0.001 & 0.001 & 0.0003 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}(X'\boldsymbol{y}) = \begin{bmatrix} -1.163461 \\ 0.135270 \\ 0.019950 \\ 0.121954 \end{bmatrix}$$

$\boldsymbol{y}'\boldsymbol{y} = 285$

a. Use these results to construct the analysis of variance table.

b. Give the computed regression equation and the standard errors of the regression coefficients.

c. Compare each estimated regression coefficient to its standard error and use the t test to test the simple hypotheses that each individual regression coefficient is equal to zero. State your conclusions about β_1, β_2, and β_3.

4.17. Consider the following two models:

$$\text{Model A}: \quad y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\text{Model B}: \quad y_i = \beta_1 x_i + \epsilon_i$$

Suppose that model A is fitted to 22 data points (x_i, y_i) with the following results:

$$\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1) = (4.0, -4.5), \quad V(\hat{\beta}_0) = 4.0,$$
$$V(\hat{\beta}_1) = 9.0, \quad \text{and} \quad \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0.0$$

a. Construct individual 95% confidence intervals for β_0 and for β_1. What conclusions can you draw?

b. Construct a joint 95% confidence region for (β_0, β_1). Draw this confidence region on the plane of possible values for (β_0, β_1). On the basis of this region, what conclusions can you draw about the relative merits of models A and B?

c. Do the results of (a) and (b) conflict? Carefully explain your reasoning.

4.18. Consider the model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$$

Let $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$, $\hat{\boldsymbol{\mu}} = H\boldsymbol{y}$, and $\boldsymbol{e} = (I - H)\boldsymbol{y}$, where $H = X(X'X)^{-1}X'$. Show that $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{e}$ are statistically independent.

4.19. Consider a regression through the origin,

$$y_i = \beta x_i + \epsilon_i, \quad \text{with } E(\epsilon_i) = 0,$$
$$V(\epsilon_i) = \sigma^2 x_i^2, \quad i = 1, 2, \ldots, 12$$

a. Derive the generalized least squares estmate of β in Eq. (4.58) and obtain its variance. Note that the covariance matrix V and its inverse V^{-1} are diagonal matrices. The generalized least squares estimate minimizes a weighted sum of squares with weights given by the diagonal elements in V^{-1}. Hence, one refers to it as the **weighted least squares** estimate.

b. Suppose that $z_i = y_i/x_i$ and $\sum_{i=1}^{12} z_i = 30$. Find the numerical value for the weighted least squares estimate in (a) and express its variance as a function of σ^2.

4.20. Consider a regression through the origin,

$$y_i = \beta x_i + \epsilon_i, \quad \text{with } E(\epsilon_i) = 0,$$
$$V(\epsilon_i) = \sigma^2 x_i, \quad x_i > 0, \quad i = 1, 2, \ldots, 10$$

a. Derive the generalized (weighted) least squares estmator of β and obtain its variance.

b. Assume that the experimenter recorded only the sample means $\bar{x} = 15$ and $\bar{y} = 30$. If possible, obtain a numerical value for the weighted least squares estimate in (a) and express its variance as a function of σ^2.

4.21. The data are taken from Davies, O. L., and Goldsmith, P. L. (Eds.). *Statistical Methods in Research and Production* (4th ed.). Edinburgh, UK: Oliver & Boyd, 1972. The data are given in the file **abrasion**.

The hardness and the tensile strength of rubber affect its resistance to abrasion. Thirty samples of rubber are tested for hardness (in degrees Shore; the larger the number, the harder the rubber) and tensile strength (in kilograms per square centimeter). Each sample was subjected to steady abrasion for a certain fixed period of time, and the loss of rubber (in grams per hour of testing) was measured.

Develop a model that relates the abrasion loss to hardness and tensile strength.

Construct scatter plots of abrasion loss against hardness and tensile strength. Fit appropriate regression models, obtain and interpret the estimates of the coefficients, calculate the ANOVA table, and discuss the adequacy of the model fit. Use your model(s) to obtain a 95% confidence interval for the mean abrasion loss for rubber with hardness 70 and tensile strength 200.

$y =$ Abrasion Loss (g/hr)	$x_1 =$ Hardness (degree Shore)	$x_2 =$ Tensile Strength (kg/cm^2)
372	45	162
206	55	233
175	61	232
154	66	231
136	71	231
112	71	237
55	81	224
45	86	219
221	53	203
166	60	189
164	64	210
113	68	210
82	79	196
32	81	180
228	56	200
196	68	173
128	75	188
97	83	161
64	88	119
249	59	161
219	71	151
186	80	165
155	82	151
114	89	128
341	51	161
340	59	146
283	65	148
267	74	144
215	81	134
148	86	127

4.22. The data are taken from Joglekar, G., Schuenemeyer, J. H., and LaRiccia, V. Lack-of-fit testing when replicates are not available. *American Statistician,* 43, 135–143, 1989. The data are given in the file **woodstrength**.

The tensile strength of Kraft paper (in pounds per square inch) is measured against the percentage of hardwood in the batch of pulp from which the paper was produced. Data for 19 observations are given here.

Develop a model that relates tensile strength to the percentage of hardwood in the paper. Construct scatter plots of tensile strength against the percentage of hardwood.

a. Fit a linear model and comment on your findings.

b. Consider a model that also includes the square of the percentage of hardwood. Fit the quadratic model, obtain and interpret the estimates of the coefficients, calculate the ANOVA table, and discuss the adequacy of the model fit. Add the fitted line to your scatter plot. Discuss whether the quadratic component is needed. Use your model to obtain a 95% confidence interval for the mean tensile strength of paper with 6% hardwood content. How is this interval different from a corresponding prediction interval? Discuss whether it is reasonable to obtain a confidence interval for the mean tensile strength of paper with 20% hardwood content.

$x =$ Hardwood Concentration	$y =$ Tensile Strength
1.0	6.3
1.5	11.1
2.0	20.0
3.0	24.0
4.0	26.1
4.5	30.0
5.0	33.8
5.5	34.0
6.0	38.1
6.5	39.9
7.0	42.0
8.0	46.1
9.0	53.1
10.0	52.0

x = Hardwood Concentration	y = Tensile Strength
11.0	52.5
12.0	48.0
13.0	42.8
14.0	27.8
15.0	21.9

4.23. The data are taken from Humphreys, R. M. Studies of luminous stars in nearby galaxies. I. Supergiants and O stars in the Milky Way. *Astrophysics Journal, Supplementary Series,* 38, 309–350, 1978. The data are given in the file **lightintensity**.

Light intensity and surface temperature were determined for 47 stars taken from the Hertzsprung–Russel diagram of Star Cluster CYG OB1. The objective is to find a relationship between light intensity and surface temperature.

Construct a scatter plot of light intensity against surface temperature. Fit a quadratic regression model, obtain and interpret the estimates of the coefficients, calculate the ANOVA table, and discuss the adequacy of the model fit. Add the fitted line to your scatter plot.

What other interpretations of the scatter plot are possible? For example, could it be that four stars are different in the sense that they do not follow the linear pattern established by the other stars? What questions would you ask the astrophysicist?

Index	x = Log Surface Temp	y = Log Light Intensity
1	4.37	5.23
2	4.56	5.74
3	4.26	4.93
4	4.56	5.74
5	4.30	5.19
6	4.46	5.46
7	3.84	4.65
8	4.57	5.27
9	4.26	5.57
10	4.37	5.12
11	3.49	5.73
12	4.43	5.45
13	4.48	5.42
14	4.01	4.05
15	4.29	4.26
16	4.42	4.58
17	4.23	3.94
18	4.42	4.18
19	4.23	4.18
20	3.49	5.89
21	4.29	4.38
22	4.29	4.22
23	4.42	4.42
24	4.49	4.85
25	4.38	5.02
26	4.42	4.66
27	4.29	4.66
28	4.38	4.90
29	4.22	4.39
30	3.48	6.05
31	4.38	4.42
32	4.56	5.10
33	4.45	5.22
34	3.49	6.29
35	4.23	4.34
36	4.62	5.62
37	4.53	5.10
38	4.45	5.22
39	4.53	5.18
40	4.43	5.57
41	4.38	4.62
42	4.45	5.06
43	4.50	5.34
44	4.45	5.34
45	4.55	5.54
46	4.45	4.98
47	4.42	4.50

4.24. Consider the UFFI data set in Table 1.2 ($n = 24$ observations). Estimate the model with three regression coefficients, $y = \beta_0 + \beta_1 x_1(\text{UFFI}) + \beta_2 x_2(\text{TIGHT}) + \varepsilon$.

a. Use the statistical software of your choice and confirm the regression results in Table 4.1.

b. Determine the 3×3 matrix $X'X$ and its inverse $(X'X)^{-1}$. Determine the standard errors of the three estimates and the pairwise correlations among the estimates (there are three correlations).

c. Determine a 95% confidence region (ellipse) for the two slopes $\boldsymbol{\beta} = (\beta_1, \beta_2)'$. We know that the marginal distribution of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)'$ is a bivariate normal distribution with covariance matrix $\sigma^2 A^{-1}$, where A^{-1} is the appropriate 2×2 submatrix of $(X'X)^{-1}$ found in (b). Hence, the contours of the confidence ellipse can be traced out by solving $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' A (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = 2s^2 F(0.95; 2, n-3)$. Here, $F(0.95; 2, n-3 = 21)$ is the 95th percentile of the F distribution, and s^2 is the mean square error.

4.25. Confidence intervals for regression coefficients and the mean response and prediction intervals for future observations in Section 4.3 make use of the t distribution. The t distribution as the resulting sampling distribution of the coefficient estimates in Eq. (4.24) depends critically on the model assumptions, in particular the assumption that the independent errors are normally distributed. The distribution in Eq. (4.24) is not a t distribution and it is no longer known if the distribution of the errors is nonnormal.

Bootstrapping (or resampling) methods are commonly used to overcome problems of unknown sampling distributions. The bootstrap, originally proposed by Efron (1979), approximates the unknown theoretical sampling distribution of the coefficient estimates by an empirical distribution that is obtained through a resampling process.

Several versions of the bootstrap are proposed for the regression situation, and the references listed at the end of this exercise will give you more details. Here, we discuss the "bootstrap in pairs" method, which resamples directly from the original data $(y_i, \boldsymbol{x}_i), i = 1, 2, \ldots, n$. This method repeats the following steps B times. Sample with replacement n pairs from the original n observations $(y_i, \boldsymbol{x_i})$. From these n sampled pairs, calculate the least squares estimates and denote the jth coefficient estimate by $\hat{\beta}_j^{*(b)}$. The superscript asterisk denotes the fact that the estimate is obtained from data generated by the bootstrap procedure, the superscript b denotes the bth replication, and the subscript j refers to a particular scalar coefficient. The B independent replications supply the empirical bootstrap distribution function.

Percentile bootstrap intervals are proposed as confidence intervals for the regression coefficients. One approach determines the $100(\alpha/2)$ and $100(1-(\alpha/2))$ percentiles of the empirical bootstrap distribution function, $\hat{\beta}_j^*(\alpha/2)$ and $\hat{\beta}_j^*(1-(\alpha/2))$, and computes a $100(1-\alpha)\%$ bootstrap confidence interval for the parameter β_j as

$$\hat{\beta}_j^*(\alpha/2), \quad \hat{\beta}_j^*(1-(\alpha/2))$$

Here, we have given the very simplest bootstrap method for the regression situation. Modifications that improve on this simple procedure have been proposed and are discussed in the references. The modifications involve sampling residuals (compared to the resampling of cases discussed here) and refinements for improving the coverage properties of percentile bootstrap intervals [one modification calculates the lower and upper limits as $\hat{\beta}_j - [\hat{\beta}_j^*(1-(\alpha/2)) - \hat{\beta}_j]$ and $\hat{\beta}_j - [\hat{\beta}_j^*(\alpha/2) - \hat{\beta}_j]$, where $\hat{\beta}_j$ is the estimate from the original sample].

a. Select one or more of the listed references and write a brief summary that explains the bootstrap methods in regression and discusses their importance.

b. Consider the simple linear regression model. Use the fuel efficiency data in Table 1.3 and regress fuel efficiency (gallons per 100 traveled miles) on the weight of the car. Obtain a 95% bootstrap confidence interval for the slope. Use $B =$ 1,000 and 2,000 replications. Relate the results to the standard confidence interval based on the t distribution.