Exercises on Chapter 2: Linear Regression with one independent variable:

Summary:

Simple Linear Regression Model: (distribution of error terms unspecified)

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i} \qquad \boldsymbol{, i = 1, 2, \dots, n} \qquad \textbf{(2.1)}$$

where, $\boldsymbol{Y_i}$ **…** value of the response variable in the *i*th trial
$\boldsymbol{X_i}$ **…** value of the independent (predictor) variable in the *i*th trial; a known constant.
$\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$ **…** are model parameters; $\boldsymbol{\beta_0}$, *Y*-intercept of the regression line – (mean of the probability distribution of *Y* at *X*=0; $\boldsymbol{\beta_1}$, the slope of the regression line – (change in the mean of the probability distribution of *Y* per unit increase in *X*.
$\boldsymbol{\varepsilon_i}$ **…** random error term with mean $\boldsymbol{E(\varepsilon_i) = 0}$ and constant variance $\boldsymbol{\sigma^2(\varepsilon_i) = \sigma^2}$**;**
$\boldsymbol{\varepsilon_i}$ ***and*** $\boldsymbol{\varepsilon_j}$ are uncorrelated so that $\boldsymbol{cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall i, j;\ i \neq j}$

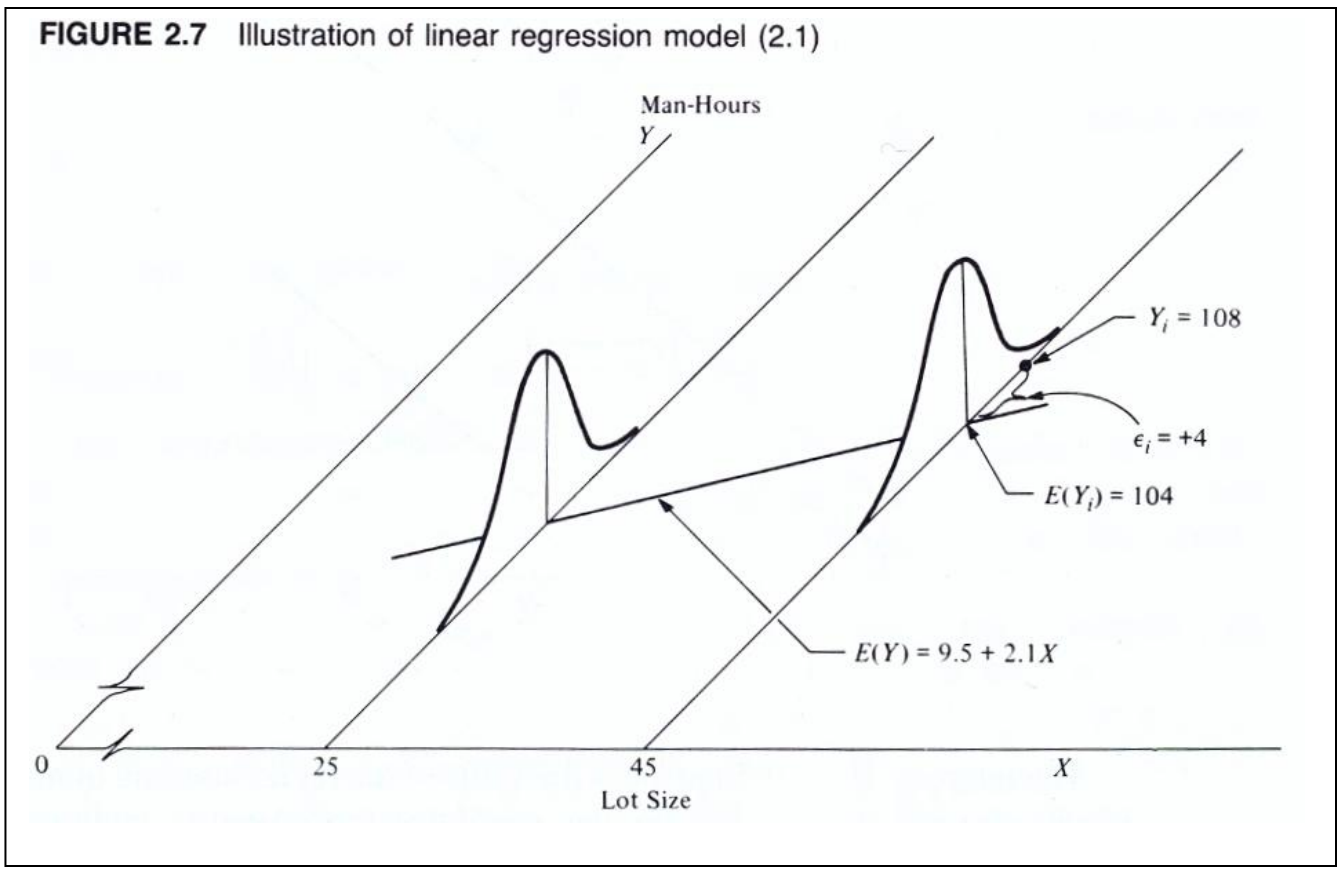
1- The observed value of Y consists of: a constant term $\beta_0 + \beta_1 X_i$ and a random term ε_i

2- Mean of the probability distribution of $\boldsymbol{Y_i}$: $\boldsymbol{E(Y_i) = \beta_0 + \beta_1 X_i}$, variance $\boldsymbol{\sigma^2(Y_i) = \sigma^2}$

Thus the regression function for model (2.1) is:

$$\boldsymbol{E(Y) = \beta_0 + \beta_1 X}$$

FIGURE 2.7 Illustration of linear regression model (2.1)



Estimation of regression function by method of Least Squares:

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{i}} = \boldsymbol{b_0} + \boldsymbol{b_1 X_i} \quad , \boldsymbol{i} = \mathbf{1}, \mathbf{2}, \dots, \boldsymbol{n} \qquad \textbf{(2.2)}$$

Least squares estimators:

$\hat{\beta}_0 = b_0$, $\hat{\beta}_1 = b_1$, are unbiased and have minimum variance among all unbiased linear estimators.

$$\boldsymbol{b_1} = \frac{\sum \boldsymbol{X_i Y_i} - \frac{\mathbf{1}}{\boldsymbol{n}} \sum \boldsymbol{X_i} \sum \boldsymbol{Y_i}}{\sum \boldsymbol{X_i^2} - \frac{\mathbf{1}}{\boldsymbol{n}} (\sum \boldsymbol{X_i})^2} = \frac{\sum (\boldsymbol{X_i} - \bar{\boldsymbol{X}})(\boldsymbol{Y_i} - \bar{\boldsymbol{Y}})}{\sum (\boldsymbol{X_i} - \bar{\boldsymbol{X}})^2} \qquad \text{(prove??)}$$

$$\boldsymbol{b_0} = \frac{\mathbf{1}}{\boldsymbol{n}} (\sum \boldsymbol{Y_i} - \boldsymbol{b_1} \sum \boldsymbol{X_i}) \qquad \text{(prove??)}$$

The point estimators b_0 and b_1 are linear functions of the observations Y_i: (prove??)

$$\boldsymbol{b_1} = \frac{\sum (\boldsymbol{X_i} - \bar{\boldsymbol{X}})(\boldsymbol{Y_i} - \bar{\boldsymbol{Y}})}{\sum (\boldsymbol{X_i} - \bar{\boldsymbol{X}})^2} = \frac{\sum (\boldsymbol{X_i} - \bar{\boldsymbol{X}}) \boldsymbol{Y_i}}{\sum (\boldsymbol{X_i} - \bar{\boldsymbol{X}})^2} = \sum \boldsymbol{k_i Y_i}$$

An alternative regression model:

$$\boldsymbol{Y_i} = \boldsymbol{\beta_0^*} + \boldsymbol{\beta_1} (\boldsymbol{X_i} - \bar{\boldsymbol{X}}) + \boldsymbol{\varepsilon_i} \quad , \boldsymbol{i} = \mathbf{1}, \mathbf{2}, \dots, \boldsymbol{n} \qquad \textbf{(2.6)}$$

where, $\beta_0^* = \beta_0 + \beta_1 \bar{X}$. Then the least squares estimator for β_0^* is:

$$\boldsymbol{b_0^*} = \boldsymbol{b_0} + \boldsymbol{b_1} \bar{\boldsymbol{X}} = (\bar{\boldsymbol{Y}} - \boldsymbol{b_1} \bar{\boldsymbol{X}}) + \boldsymbol{b_1} \bar{\boldsymbol{X}} = \bar{\boldsymbol{Y}} \qquad \textbf{(2.14)}$$

Hence, the estimated regression equation for alternative model (2.6) is:

$$\widehat{\boldsymbol{Y}} = \bar{\boldsymbol{Y}} + \boldsymbol{b_1} (\boldsymbol{X} - \bar{\boldsymbol{X}}) \qquad \textbf{(2.15)}$$

Residuals

The ith residual is the difference between the observed response $\boldsymbol{Y_i}$ value and the corresponding fitted value $\widehat{\boldsymbol{Y}}_{\boldsymbol{i}}$

$$\boldsymbol{e_i} = \boldsymbol{Y_i} - \widehat{\boldsymbol{Y}}_{\boldsymbol{i}} = \boldsymbol{Y_i} - \boldsymbol{b_0} - \boldsymbol{b_1 X_i} \qquad \textbf{(2.16)}$$

Properties of fitted regression line:

1- The sum of the residuals is zero: $\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{e_i} = \mathbf{0}$ (2.17) prove??
2- The sum of the squared residuals, $\sum \boldsymbol{e_i^2}$,is a minimum
3- The sum of the observed values $\boldsymbol{Y_i}$ equals the sum of the fitted values $\widehat{\boldsymbol{Y}}_{\boldsymbol{i}}$
   $\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{Y_i} = \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \widehat{\boldsymbol{Y}}_{\boldsymbol{i}}$ prove??
4- The sum of the weighted residuals: $\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{X_i e_i} = \mathbf{0}$, $\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \widehat{\boldsymbol{Y}}_{\boldsymbol{i}} \boldsymbol{e_i} = \mathbf{0}$ prove??
5- The regression line always goes through the point $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{Y}})$ prove??

Point estimator of σ^2 for the regression model:

The deviation of an observation $\boldsymbol{Y_i}$ from its estimated mean $\widehat{\boldsymbol{Y}}_{\boldsymbol{i}}$; i.e. the residuals $\boldsymbol{e_i}$

Error (residual) sum of squares *SSE*,

$$\boldsymbol{SSE} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y_i} - \boldsymbol{\hat{Y}_i})^2 = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y_i} - \boldsymbol{b_0} - \boldsymbol{b_1 X_i})^2 = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{e_i^2}$$

Error (residual) mean square *MSE*,

$$\boldsymbol{MSE} = \frac{\boldsymbol{SSE}}{\boldsymbol{n-2}} = \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y_i} - \boldsymbol{\hat{Y}_i})^2}{\boldsymbol{n-2}} = \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y_i} - \boldsymbol{b_0} - \boldsymbol{b_1 X_i})^2}{\boldsymbol{n-2}} = \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{e_i^2}}{\boldsymbol{n-2}}$$

Normal Error Model

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i} \qquad \boldsymbol{, i = 1, 2, \dots, n} \qquad \textbf{(2.25)}$$

where, $\boldsymbol{Y_i}$ **…** the observed response in the *i*th trial
$\boldsymbol{X_i}$ **…** the level of the independent variable in the *i*th trial; a known constant.
$\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$ **…** are model parameters
$\boldsymbol{\varepsilon_i}$ **…** are independent $\boldsymbol{N(0, \sigma^2)}$

The likelihood function for the Normal error model:

$$\boldsymbol{L(\beta_0, \beta_1, \sigma^2)} = \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} \frac{\boldsymbol{1}}{\boldsymbol{(2\pi\sigma^2)^{\frac{1}{2}}}} \boldsymbol{\exp\left[-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right]}$$

$$= \frac{\boldsymbol{1}}{\boldsymbol{(2\pi\sigma^2)^{\frac{n}{2}}}} \boldsymbol{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right]}$$

Maximum Likelihood estimators:

$\hat{\beta}_0 = b_0$, $\hat{\beta}_1 = b_1$ same as least squares estimation

$$\hat{\sigma}^2 = \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y_i} - \boldsymbol{\hat{Y}_i})^2}{n}$$

$$MSE = \frac{n}{n-2}\hat{\sigma}^2$$

Exercises

Group (A)

2.1. Refer to the sales volume example on page 24. Suppose that the number of units sold is measured accurately but clerical errors are frequently made in determining the dollar sales. Would the relation between the number of units sold and dollar sales still be a functional one? Discuss.

2.2. The members of a health spa pay annual membership dues of $300 plus a charge of $2 for each visit to the spa. Let *Y* denote the total dollar cost for the year for a member and X the number of visits by the member during the year. Express the relation between X and *Y* mathematically. Is it a functional or a statistical relation?

2.3. Experience with a certain type of plastic indicates that a relation exists between the hardness (measured in Brinell units) of items molded from the plastic *(Y)* and the elapsed time since termination of the molding process *(X).* It is proposed to study this relation by means of regression analysis. A participant in the discussion objects, pointing out that the hardening of the plastic "is the result of a natural chemical process that doesn't leave anything to chance, so the relation must be mathematical and regression analysis is not appropriate." Evaluate this objection.

2.4. In Table 2.1, the lot size X is the same in production runs I and 8 but the man-hours *Y* differ. What feature of regression model (2.1) is illustrated by this?

2.5. When asked to state the simple linear regression model, a student wrote it as follows: $E(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$ Do you agree?

2.6. Consider the normal error regression model (2.25). Suppose that the parameter values are β_0 = 200, β_1 = 5.0, and σ = 4.

a- Plot this normal error regression model in the fashion of Figure 2.7. Show the distributions of *Y* for *X* = 10, 20, and 40.

b- Explain the meaning of the parameters β_0 and β_1. Assume that the scope of the model includes *X* = 0.

2.7. In a simulation exercise, regression model (2.1) applies with β_0 = 100, β_1 = 20, and σ^2 = 25. An observation on *Y* will be made for X = 5.

a- Can you state the exact probability that *Y* will fall between 195 and 205? Explain.

b- If the normal error regression model (2.25) is applicable, can you now state the exact probability that *Y* will fall between 195 and 205? If so, state it.

2.8. In Figure 2.7, suppose another observation were obtained at X = 45. Would *E(Y)* for this new observation still be 104? Would the *Y* value for this new observation again be 108?

2.9. A student in accounting enthusiastically declared: "Regression is a very powerful tool. We can isolate fixed and variable costs by fitting a linear regression model, even when we have no data for small lots." Discuss.

2.10. An analyst in a large corporation studied the relation between current annual salary (Y) and age *(X)* for the 46 computer programmers presently employed in the company. She concluded that the relation is curvilinear, reaching a maximum at 47 years. Does this imply that the salary for a programmer increases until age 47 and then decreases? Explain.

2.11. The regression function relating production output by an employee after taking a training program *(Y)* to the production output before the training program *(X)* is *E(Y)* = 20 + *.95X,* where X ranges from 40 to 100. An observer concludes that the training program does not raise production output on the average because β_1 is not greater than 1.0. Comment.

2.12. Evaluate the following statement: "For the least squares method to be fully valid, it is required that the distribution of *Y* is normal."

2.13. A person states that b_o and *b,* in the fitted regression equation (2.12) can be estimated by the method of least squares. Comment.

2.14. According to (2.17), $\sum e_i = 0$, when model (2.1) is fitted to a set of *n* observations by the method of least squares. Is it also true that $\sum e_i = 0$? Comment.

2.15. Grade point average. The director of admissions of a small college administered a newly designed entrance test to 20 students selected at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year *(Y)* can be predicted from the entrance test score *(X).* The results of the study follow. Assume that the first-order regression model (2.1) is appropriate.

i	*1*	2	3	4	5	6	7	8	9	10
X_i	5.5	4.8	4.7	3.9	4.5	6.2	6.0	5.2	4.7	4.3
Y_i	3.1	2.3	3.0	1.9	2.5	3.7	3.4	2.6	2.8	1.6

i	11	12	13	14	15	16	17	18	19	20
X_i	4.9	5.4	5.0	6.3	4.6	4.3	5.0	5.9	4.1	4.7
Y_i	2.0	2.9	2.3	3.2	1.8	1.4	2.0	3.8	2.2	1.5

Summary calculation results are: $\sum X_i$ = 100.0, $\sum Y_i$ = 50.0, $\sum X_i^2$= 509.12, $\sum Y_i^2$ =134.84, $\sum X_i Y_i$ = 257.66.

a. Obtain the least squares estimates of β_0 and β_1, and state the estimated regression function.

b. Plot the estimated regression function and the data. Does the estimated regression function appear to fit the data well?

c. Obtain a point estimate of the mean freshman GPA when the entrance test score is X = 5.0.

d. What is the point estimate of the change in the mean response when the entrance test score increases by one point?

2.16. Calculator maintenance. The Tri-City Office Equipment Corporation sells an imported desk calculator on a franchise basis and performs preventive maintenance and repair service on this calculator. The data below have been collected from 18 recent calls on users to perform routine preventive maintenance service; for each call, X is the number of machines serviced and *Y* is the total number of minutes spent by the service person. Assume that the first-order regression model (2.1) is appropriate.

i	1	2	3	4	5	6	7	8	9
X_i	7	6	5	I	5	4	7	3	4
Y_i	97	86	78	10	75	62	101	39	53

i	10	11	12	13	14	15	16	17	18
X_i	2	8	5	2	5	7	I	4	5
Y_i	33	118	65	25	71	105	17	49	68

Summary calculation results are: $\sum Y_i = 1{,}152$, $\sum X_i = 81$, $\sum(\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})^2 = 16{,}504$, $\sum(\boldsymbol{X_i} - \overline{\boldsymbol{X}})^2 = 74.5$, $\sum(\boldsymbol{X_i} - \overline{\boldsymbol{X}})(\boldsymbol{Y_i} - \overline{\boldsymbol{Y}}) = 1{,}098$.

a. Obtain the estimated regression function.

b. Plot the estimated regression function and the data. How well does the estimated regression function fit the data?

c. Interpret b_o in your estimated regression function. Does b_o provide any relevant information here? Explain.

d. Obtain a point estimate of the mean service time when X = 5 machines are serviced.

2.17. Airfreight breakage. A substance used in biological and medical research is shipped by airfreight to users in cartons of 1,000 ampules. The data below, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraft to another over the shipment route *(X)* and the number of ampules found to be broken upon arrival (Y). Assume that the first-order regression model (2.1) is appropriate.

i	1	2	3	4	5	6	7	8	9	10
X_i	1	0	2	0	3	1	0	1	2	0
Y_i	16	9	17	12	22	13	8	15	19	11

a. Obtain the estimated regression function. Plot the estimated regression function and the data. Does a linear regression function appear to give a good fit here?

b. Obtain a point estimate of the expected number of broken ampules when X = 1 transfer is made.

c. Estimate the increase in the expected number of ampules broken when there are 2 transfers as compared to 1 transfer.

d. Verify that your fitted regression line goes through the point $(\bar{X}, \bar{Y})$.

2.18. Plastic hardness. Refer to Problem 2.3. Twelve batches of the plastic were made, and from each batch one test item was molded and the hardness measured at some specific point in time. The results are shown below; X is elapsed time in hours, and *Y* is hardness in Brinell units. Assume that the first-order regression model (2.1) is appropriate.

i	1	2	3	4	5	6	7	8	9	10	11	12
X_i	32	48	72	64	48	16	40	48	48	24	80	56
Y_i	230	262	323	298	255	199	248	279	267	214	359	305

a. Obtain the estimated regression function. Plot the estimated regression function and the data. Does a linear regression function appear to give a good fit here?

b. Obtain a point estimate of the mean hardness when X = 48 hours.

c. Obtain a point estimate of the change in mean hardness when X increases by one hour.

2.19. Refer to **Grade point average** Problem 2.15.

a. Obtain the residuals e_i , Do they sum to zero in accord with (2.17)?

b. Estimate σ^2 and σ, **In** what units is σ expressed?

2.20. Refer to **Calculator maintenance** Problem 2.16.

a. Obtain the residuals e_i , and the sum of the squared residuals $\sum \boldsymbol{e_i^2}$.

b. Estimate σ^2 and σ, **In** what units is σ expressed?

2.21. Refer to **Airfreight breakage** Problem 2.17.

a. Obtain the residual for the first observation. What is its relation to ε_i*?*

b. Compute $\sum \boldsymbol{e_i^2}$ and *MSE.* What is estimated by *MSE?*

2.22. Refer to **Plastic hardness** Problem 2.18.

a. Obtain the residuals *e.*. Do they sum to zero in accord with (2.17)?

b. Estimate σ^2 and σ, In what units is σ expressed?

2.23. Muscle mass. A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected four women from each 10-year age group, beginning with age 40 and ending with age 79. The results follow; *X* is age, and *Y* is a measure of muscle mass. Assume that the first-order regression model (2.1) is appropriate.

i	1	2	3	4	5	6	7	8
X_i	71	64	43	67	56	73	68	56
Y_i	82	91	100	68	87	73	78	80

i	9	10	11	12	13	14	15	16
X_i	76	65	45	58	45	53	49	78
Y_i	65	84	116	76	97	100	105	77

a. Obtain the estimated regression function. Plot the estimated regression function and the data. Does a linear regression function appear to give a good fit here? Does your plot support the anticipation that muscle mass decreases with age?

b. Obtain the following:

(1) a point estimate of the difference in the mean muscle mass for women differing in age by one year,

(2) a point estimate of the mean muscle mass for women aged *X* = 60 years,

(3) the value of the residual for the eighth observation, (4) a point estimate of σ^2.

2.24. Robbery rate. A criminologist studying the relationship between population density and robbery rate in medium-sized U. S. cities collected the following data for a random sample of 16 cities; *X* is the population density of the city (number of people per unit area), and *Y* is the robbery rate last year (number of robberies per 100,000 people). Assume that the first-order regression model (2.1) is appropriate.

i	1	2	3	4	5	6	7	8
X_i	59	49	75	54	78	56	60	82
Y_i	209	180	195	192	215	197	208	189

i	9	10	11	12	13	14	15	16
X_i	69	83	88	94	47	65	89	70
Y_i	213	201	214	212	205	186	200	204

a. Obtain the estimated regression function. Plot the estimated regression function and the data. Does the linear regression function appear to give a good fit here? Discuss.

b. Obtain point estimates of the following: (1) the difference in the mean robbery rate in cities that differ by one unit in population density, (2) the mean robbery rate last year in cities with population density $X = 60$, (3) ε_{10}, (4) σ^2.

Group (B)

2.25. Refer to regression model (2.1). Assume that $X = 0$ is within the scope of the model. What is the implication for the regression function if $\beta_0 = 0$, so that the model is $Y_i = \beta_1 X_i + \varepsilon_i$? How would the regression function plot on a graph?

2.26. Refer to regression model (2.1). What is the implication for the regression function if $\beta_1 = 0$ so that the model is $Y_i = \beta_0 + \varepsilon_i$? How would the regression function plot on a graph?

2.27. Refer to **Plastic hardness** Problem 2.18. Suppose one test item was molded from a single batch of plastic and the hardness of this one item was measured at 12 different points in time. Would the error term in the model for this case still reflect the same effects as for the experiment initially described? Would you expect the error terms for the different points in time to be uncorrelated? Discuss.

2.28. Derive the expression for b_1 from the normal equations in (2.9).

2.29. Refer to the model $Y_i = \beta_0 + \varepsilon_i$ in Exercise 2.26. Using the method of least squares, derive the estimator of β_0 for this model.

2.30. Prove that the least squares estimator of β_0 obtained in Exercise 2.29 is unbiased.

2.31. Prove that the sum of the residuals weighted by the fitted values is zero.