question of the inclusion or exclusion of the new variable in the regression equation can be resolved.

It is apparent from this discussion that the effect a variable has on the regression equation determines its suitability for being included in the fitted equation. The results presented in this chapter influence the formulation of different strategies devised for variable selection. Variable selection procedures are presented in Chapter 11.

4.14 ROBUST REGRESSION

Another approach (not discussed here), useful for the identification of outliers and influential observations, is *robust regression*, a method of fitting that gives less weight to points with high leverage. There is a vast amount of literature on robust regression. The interested reader is referred, for example, to the books by Huber (1981), Hampel et al. (1986), Rousseeuw and Leroy (1987), Staudte and Sheather (1990), and Birkes and Dodge (1993). We must also mention the papers by Krasker and Welsch (1982), Coakley and Hettmansperger (1993), Chatterjee and Mächler (1997), and Billor, Chatterjee, and Hadi (2006), which incorporate ideas of bounding influence and leverage in fitting. In Section 13.5 we give a brief discussion of robust regression and present a numerical algorithm for robust fitting. Two examples are given as illustration.

EXERCISES

4.1 Check to see whether or not the standard regression assumptions are valid for each of the following data sets: **This exercise can be ignored.**

(a) The Milk Production data described in Section 1.3.1.

(b) The Right-To-Work Laws data described in Section 1.3.2 and given in Table 1.3.

(c) The Egyptian Skulls data described in Section 1.3.4.

(d) The Domestic Immigration data described in Section 1.3.3.

(e) The New York Rivers data described in Section 1.3.5 and given in Table 1.9.

4.2 Find a data set where regression analysis can be used to answer a question of interest. Then:

(a) Check to see whether or not the usual multiple regression assumptions are valid.

(b) Analyze the data using the regression methods presented thus far, and answer the question of interest.

4.3 Consider the computer repair problem discussed in Section 2.3. In a second sampling period, 10 more observations on the variables Minutes and Units

Table 4.6 Expanded Computer Repair Times Data: Length of Service Calls (Minutes) and Number of Units Repaired (Units)

Row	Units	Minutes	Row	Units	Minutes
1	1	23	13	10	154
2	2	29	14	10	166
3	3	49	15	11	162
4	4	64	16	11	174
5	4	74	17	12	180
6	5	87	18	12	176
7	6	96	19	14	179
8	6	97	20	16	193
9	7	109	21	17	193
10	8	119	22	18	195
11	9	149	23	18	198
12	9	145	24	20	205

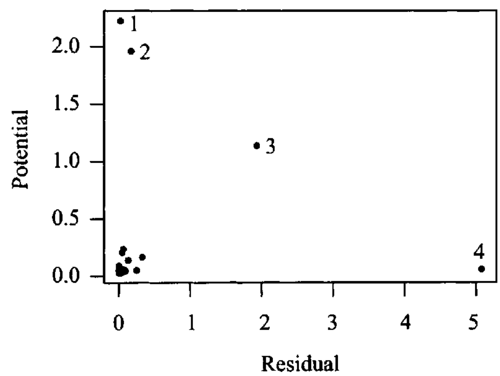

Figure 4.14 P-R plot used in Exercise 4.4.

were obtained. Since all observations were collected by the same method from a fixed environment, all 24 observations were pooled to form one data set. The data appear in Table 4.6.

(a) Fit a linear regression model relating Minutes to Units.

(b) Check each of the standard regression assumptions and indicate which assumption(s) seems to be violated.

4.4 In an attempt to find unusual points in a regression data set, a data analyst examines the P-R plot (shown in Figure 4.14). Classify each of the unusual points on this plot according to type.

4.5 Name one or more graphs that can be used to validate each of the following assumptions. For each graph, sketch an example where the corresponding assumption is valid and an example where the assumption is clearly invalid.

(a) There is a linear relationship between the response and predictor variables.

(b) The observations are independent of each other.

(c) The error terms have constant variance.

(d) The error terms are uncorrelated.

(e) The error terms are normally distributed.

(f) The observations are equally influential on least squares results.

4.6 The following graphs are used to verify some of the assumptions of the ordinary least squares regression of Y on $X_1, X_2, \cdots, X_p$:

1. The scatter plot of Y versus each predictor X_j.
2. The scatter plot matrix of the variables $X_1, X_2, \cdots, X_p$.
3. The normal probability plot of the internally standardized residuals.
4. The residuals versus fitted values.
5. The potential-residual plot.
6. Index plot of Cook's distance.
7. Index plot of Hadi's influence measure.

For each of these graphs:

(a) What assumption can be verified by the graph?

(b) Draw an example of the graph where the assumption does not seem to be violated.

(c) Draw an example of the graph which indicates the violation of the assumption.

4.7 Consider again the Cigarette Consumption data described in Exercise 3.15 and given in Table 3.17. **(check Chapter 3 Practice Exercise)**

(a) What would you expect the relationship between Sales and each of the other explanatory variables to be (i.e., positive, negative)? Explain.

(b) Compute the pairwise correlation coefficients matrix and construct the corresponding scatter plot matrix.

(c) Are there any disagreements between the pairwise correlation coefficients and the corresponding scatter plot matrix?

(d) Is there any difference between your expectations in part (a) and what you see in the pairwise correlation coefficients matrix and the corresponding scatter plot matrix?

(e) Regress Sales on the six predictor variables. Is there any difference between your expectations in part (a) and what you see in the regression coefficients of the predictor variables? Explain inconsistencies if any.

(f) How would you explain the difference in the regression coefficients and the pairwise correlation coefficients between Sales and each of the six predictor variables?

(g) Is there anything wrong with the tests you made and the conclusions you reached in Exercise 3.15?

4.8 Consider again the Examination Data used in Exercise 3.3 and given in Table 3.10: **(check Chapter 3 Practice Exercise)**

(a) For each of the three models, draw the P-R plot. Identify all unusual observations (by number) and classify as outlier, high-leverage point, and/or influential observation.

(b) What model would you use to predict the final score F?

4.9 Either prove each of the following statements mathematically or demonstrate its correctness numerically using the Cigarette Consumption data described in Exercise 3.15 and given in Table 3.17: **(check Chapter 3 Practice Exercise)**

(a) The sum of the ordinary least squares residuals is zero.

(b) The relationship between $\hat{\sigma}^2$ and $\hat{\sigma}^2_{(i)}$ is

$$\hat{\sigma}^2_{(i)} = \hat{\sigma}^2 \left(\frac{n - p - 1 - r_i^2}{n - p - 2} \right). \tag{4.26}$$

4.10 Identify unusual observations for the data set in Table 4.7

Table 4.7 Data for Exercise 4.10

Row	Y	X	Row	Y	X
1	8.11	0	7	9.60	19
2	11.00	5	8	10.30	20
3	8.20	15	9	11.30	21
4	8.30	16	10	11.40	22
5	9.40	17	11	12.20	23
6	9.30	18	12	12.90	24

4.11 Consider the Scottish hills races data in Table 4.5. Choose an observation index i (e.g., $i = 33$, which corresponds to the outlying observation number 33) and create an indicator (dummy) variable U_i, where all the values of U_i are zero except for its ith value which is one. Now consider comparing the following models:

$$\begin{aligned} \text{H}_0 : \text{Time} &= \beta_0 + \beta_1 \text{ Distance} + \beta_2 \text{ Climb} + \varepsilon, && (4.27) \\ \text{H}_1 : \text{Time} &= \beta_0 + \beta_1 \text{ Distance} + \beta_2 \text{ Climb} + \beta_3 \, U_i + \varepsilon. && (4.28) \end{aligned}$$

Let r_i^* be the ith externally standardized residual obtained from fitting model (4.27). Show (or verify using an example) that

(a) The t-Test for testing $\beta_3 = 0$ in Model (4.28) is the same as the ith externally standardized residual obtained from Model (4.27), that is, $t_3 = r_i^*$.

(b) The F-Test for testing Model (4.27) versus (4.28) reduces to the square of the ith externally standardized residual, that is, $F = r_i^{*2}$.

(c) Fit Model (4.27) to the Scottish hills races data without the ith observation.

(d) Show that the estimates of β_0, β_1, and β_2 in Model (4.28) are the same as those obtained in (c). Hence adding an indicator variable for the ith observation is equivalent to deleting the corresponding observation!

4.12 Consider the data in Table 4.8, which consist of a response variable Y and six predictor variables. The data can be obtained from the book's Website.[2] Consider fitting a linear model relating Y to all six X-variables.

(a) What least squares assumptions (if any) seem to be violated?

(b) Compute r_i, C_i, DFITS_i, and H_i.

(c) Construct the index plots of r_i, C_i, DFITS_i, and H_i as well as the Potential-Residual plot.

(d) Identify all unusual observations in the data and classify each according to type (i.e., outliers, leverage points, etc.).

4.13 Consider again the data set in Table 4.8. Suppose now that we fit a linear model relating Y to the first three X-variables. Justify your answer to each of the following questions with the appropriate added-variable plot:

(a) Should we add X_4 to the above model? If yes, keep X_4 in the model.

(b) Should we add X_5 to the above model? If yes, keep X_5 in the model.

(c) Should we add X_6 to the above model?

(d) Which model(s) would you recommend as the best possible description of Y? Use the above results and/or perform additional analysis if needed.

4.14 Consider fitting the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ to the data set in Table 4.8. Now let u be the residuals obtained from regressing Y on X_1. Also, let X_2 and v be the residuals obtained from regressing X_3 on X_1. Show (or verify using the data set in Table 4.8 as an example) that:

(a) $\hat{\beta}_3 = \sum_{i=1}^{n} u_i v_i / \sum_{i=1}^{n} v_i^2$

(b) The standard error of $\hat{\beta}_3$ is $\hat{\sigma} / \sqrt{\sum_{i=1}^{n} v_i^2}$.

[2] http://www.aucegypt.edu/faculty/hadi/RABE5

Table 4.8 Data for Exercises 4.12–4.14

Row	Y	X_1	X_2	X_3	X_4	X_5	X_6
1	443	49	79	76	8	15	205
2	290	27	70	31	6	6	129
3	676	115	92	130	0	9	339
4	536	92	62	92	5	8	247
5	481	67	42	94	16	3	202
6	296	31	54	34	14	11	119
7	453	105	60	47	5	10	212
8	617	114	85	84	17	20	285
9	514	98	72	71	12	−1	242
10	400	15	59	99	15	11	174
11	473	62	62	81	9	1	207
12	157	25	11	7	9	9	45
13	440	45	65	84	19	13	195
14	480	92	75	63	9	20	232
15	316	27	26	82	4	17	134
16	530	111	52	93	11	13	256
17	610	78	102	84	5	7	266
18	617	106	87	82	18	7	276
19	600	97	98	71	12	8	266
20	480	67	65	62	13	12	196
21	279	38	26	44	10	8	110
22	446	56	32	99	16	8	188
23	450	54	100	50	11	15	205
24	335	53	55	60	8	0	170
25	459	61	53	79	6	5	193
26	630	60	108	104	17	8	273
27	483	83	78	71	11	8	233
28	617	74	125	66	16	4	265
29	605	89	121	71	8	8	283
30	388	64	30	81	10	10	176
31	351	34	44	65	7	9	143
32	366	71	34	56	8	9	162
33	493	88	30	87	13	0	207
34	648	112	105	123	5	12	340
35	449	57	69	72	5	4	200
36	340	61	35	55	13	0	152
37	292	29	45	47	13	13	123
38	688	82	105	81	20	9	268
39	408	80	55	61	11	1	197
40	461	82	88	54	14	7	225

Source: Chatterjee and Hadi (1988)