5

Diagnostic Procedures

5.1 Overview

In Chapters 3 and 4 we discussed procedures for point and confidence interval estimation of parameters of interest in a linear regression model. While most of these procedures are valid under assumptions (A), some procedures relied on assumptions (B). In actual applications of these procedures, the investigator needs to decide whether or not the assumptions are reasonable for the given situation. To assist the investigator in making this judgment, we suggested that a **residual analysis** be carried out. Several graphical methods of examining the residuals to assess the validity of the model assumptions were suggested. These checks of model assumptions are by no means the only ones, and there are many other graphical and numerical *diagnostic procedures* that can be of assistance in the understanding of the sample data. Some of these procedures are discussed in this chapter.

Section 5.2 introduces **studentized deleted residuals** that are useful in identifying *outliers*, i.e., sample observations that have unusual values for the *response* variable Y. Section 5.3 contains a discussion of **hat values** or **leverages** that are useful in identifying sample observations that have unusual values for the *predictor* variables X_i. Section 5.4 deals with the identification of **influential observations**, i.e., observations whose inclusion or exclusion from the analyses may give very different results. Problems associated with **ill-conditioned** data matrices, such as *numerical instability* of regression computations, and **multicollinearity** are discussed in Section 5.5. Section 5.6 contains chapter exercises. In the laboratory manuals we discuss some MINITAB and SAS commands that can be used to carry out the computations required for the procedures in this chapter.

5.2 Outliers

Often we find that there are a few (perhaps one or two) sample values that do not seem to be consistent with the regression model being fitted, but the remaining values do agree with the model. If we can identify these few and exclude them, the remaining sample may satisfy assumptions (A) or (B). Such situations are often revealed by residual plots discussed in Chapters 3 and 4. If the response y_i corresponding to sample item i deviates considerably more from its fitted value $\hat{\mu}_Y(x_{i,1}, \ldots, x_{i,k})$ than other sample values deviate from their fitted values, we say that sample observation i (case i) is an **outlier**. Such sample values need to be examined further to find out why they do not agree more closely with the fitted model. Following are some of the possible reasons.

1 Even when the model holds exactly, there are occasionally some observations that fall far from the regression curve, giving rise to an apparent outlier. (Recall that when sampling from a Gaussian population with mean zero and standard deviation one we do expect values outside the interval $(-3, 3)$ about 1% of the time). Such data values are typically included in the analysis.

2 An examination of the circumstances under which an observed value was obtained reveals that an error of some sort was made and that the sample value is incorrect. Such instances occur due to instrument malfunction during an investigation, human errors, or even errors occurring during data transcription or entry of data into a computer. If such an explanation is available, then the sample value in question should be excluded from the analysis.

3 The model under consideration does describe most of the observations adequately, but there are some combinations of the predictor factors for which the model does not give an adequate approximation, resulting in apparent outliers. Here the model should be reexamined to determine the range of values of predictor factors for which it is adequate. Observed values outside this range should be analyzed separately, or a better model should be constructed.

Occasionally we can find no explanation for an apparent outlier, and this poses a dilemma. Should this sample value be included in the analysis or excluded from the analysis? An often-used tactic is to analyze the data with the value in question included and then analyze it again with it excluded. If the results of the two analyses do not differ much, or if decisions reached are the same with either of the two sets of results, then at least our conclusions will be unaffected by the presence or absence of the value in question. If the two analyses result in different conclusions, we are forced to present both situations. Until an explanation is found for the outlying point, we have no way of verifying which conclusion is correct. In experimental situations, the factor values corresponding to the outlier may perhaps be repeated, one or more additional observations may be obtained, and the data reanalyzed. The advice of a professional statistician might be useful here.

While the standardized residuals r_i defined in (4.5.1) may be used to identify unusual observations (called *outliers*), keep in mind that the ith fitted value

$\hat{\mu}_Y(x_{i,1}, \ldots, x_{i,k})$, and hence the ith residual $\hat{e}_i$, will already be affected by including the outlying observations. For this reason, many authors suggest the use of **studentized deleted residuals**, which we discuss next.

Studentized Deleted Residuals

Studentized deleted residuals are defined as follows. Delete case i (the ith sample value) from the data and carry out the regression analysis with the remaining data. Then use the estimated regression coefficients to predict the response for sample item i. More specifically, omit the ith sample point from the data and use the remaining data to compute a regression of Y on the k predictor variables $X_1, \ldots, X_k$. Let

$$\hat{\beta}_{(-i)0},\ \hat{\beta}_{(-i)1},\ \hat{\beta}_{(-i)2}, \ldots,\ \hat{\beta}_{(-i)k}$$

denote the estimates of the regression coefficients and

$$\hat{\sigma}_{(-i)Y|X_1,\ldots,X_k}$$

which we abbreviate as

$$\hat{\sigma}_{(-i)}$$

denote the estimate of the subpopulation standard deviation σ with the ith sample point omitted from the data. Likewise, let

$$\hat{Y}_{(-i)}(x_{i,1}, \ldots, x_{i,k})$$

be the estimate of $Y(x_{i,1}, \ldots, x_{i,k})$ when the ith sample value is not used. Thus

$$\hat{Y}_{(-i)}(x_{i,1}, \ldots, x_{i,k}) = \hat{\beta}_{(-i)0} + \hat{\beta}_{(-i)1}x_{i,1} + \cdots + \hat{\beta}_{(-i)k}x_{i,k}$$

is the predicted value corresponding to the observed y_i when the ith sample value is excluded from the anaylsis. The estimates $\hat{\beta}_{(-i)0}, \ldots, \hat{\beta}_{(-i)k}$ are not influenced by sample item i because it is not included in the analysis. Thus

$$y_i - \hat{Y}_{(-i)}(x_{i,1}, \ldots, x_{i,k}) \tag{5.2.1}$$

is the difference between the observed value y_i and predicted value of y_i when sample item i is not used to estimate $\beta_0, \ldots, \beta_k$. This difference between the actual value and the predicted value is then divided by its standard error to obtain the *studentized deleted residual* for sample item i. It can be shown that

$$SE(y_i - \hat{Y}_{(-i)}(x_{i,1}, \ldots, x_{i,k})) = \frac{\hat{\sigma}_{(-i)}}{\sqrt{1 - h_{i,i}}} \tag{5.2.2}$$

where $h_{i,i}$ is the ith diagonal element of the hat-matrix $H = X(X^TX)^{-1}X^T$ defined in (4.5.2) where all observations are included in the X matrix.

D E F I N I T I O N

We define the *studentized deleted residual* for case i, denoted by T_i, as

$$T_i = \frac{y_i - \hat{Y}_{(-i)}(x_{i,1}, \ldots, x_{i,k})}{\hat{\sigma}_{(-i)} / \sqrt{1 - h_{i,i}}} \quad \blacksquare \tag{5.2.3}$$

It can be shown that T_i has a student's-t distribution with $(n-1)-k-1 = n-k-2$ degrees of freedom. For this reason studentized deleted residuals are sometimes called t-residuals.

Thus T_i is computed and examined for each $i = 1, \ldots, n$. Roughly speaking, if the absolute value of any T_i is greater than 2 (some investigators use 3), this indicates that the ith sample observation should be carefully scrutinized as a possible outlier. Outliers stand out more clearly when we examine studentized deleted residuals T_i than when we examine the standardized residuals r_i. We illustrate the use of studentized deleted residuals in Example 5.2.1.

E X A M P L E 5.2.1

In a study to investigate how automobile insurance premiums Y (in dollars) for collision coverage are related to the age X_1 (in months) and the purchase price X_2 (in dollars) of a car, a simple random sample of 36 car owners was selected from a study population, and each was asked to provide the above information. The data are displayed in Table 5.2.1 and are also stored in the file **premiums.dat** on the data disk.

T A B L E 5.2.1

Premiums Data

Observation Number	Premium Y (dollars)	Age of Car X_1 (months)	Price of Car X_2 (dollars)
1	221	57	11804
2	448	8	12926
3	515	6	14054
4	632	12	17486
5	48	47	8700
6	189	30	8570
7	581	34	18982
8	102	39	9198
9	404	33	14986
10	83	59	8473
11	280	56	13891
12	565	13	16127
13	1105	10	29480
14	388	46	15868
15	435	2	10782

TABLE 5.2.1
(Continued)

Observation Number	Premium Y (dollars)	Age of Car X_1 (months)	Price of Car X_2 (dollars)
16	309	11	8645
17	322	17	9086
18	741	32	22559
19	500	34	14969
20	626	1	14861
21	1051	34	29733
22	845	4	22893
23	278	59	15198
24	333	56	16696
25	650	34	20411
26	772	27	23128
27	477	19	16507
28	443	37	13704
29	692	3	16472
30	618	36	18422
31	1050	7	27110
32	643	45	22968
33	116	46	9177
34	269	9	8977
35	259	38	10514
36	491	16	13739

Exhibit 5.2.1 contains a MINITAB output for the regression of Y on X_1, X_2. The following are included: observation number, age X_1, price X_2, premium Y, the standardized residuals r_i (labeled `stdresid`), the fitted values $\hat{\mu}_Y(x_{i,1}, x_{i,2})$ (labeled `fits`), the residuals $\hat{e}_i$ (labeled `residual`), and the studentized deleted residuals T_i (labeled `tresid`). ■

EXHIBIT 5.2.1
MINITAB Output for Example 5.2.1

```
The regression equation is
premium  = 6.9 - 5.10 age + 0.0395 price

Predictor       Coef      Stdev    t-ratio       p
Constant        6.90      24.52       0.28   0.780
Age          -5.0996     0.3894     -13.10   0.000
Price       0.039533   0.001209      32.69   0.000

s = 41.94       R-sq = 97.7%     R-sq(adj) = 97.6%
```

EXHIBIT 5.2.1
(Continued)

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	2492087	1246043	708.49	0.000
Error	33	58038	1759		
Total	35	2550125			

Unusual Observations

Obs.	age	premium	Fit	Stdev.Fit	Residual	St.Resid	
27	19.0	477.00	562.58	7.85	-85.58	-2.08R	(5.2.4)
28	37.0	443.00	359.97	7.99	83.03	2.02R	(5.2.5)

R denotes an obs. with a large st. resid.

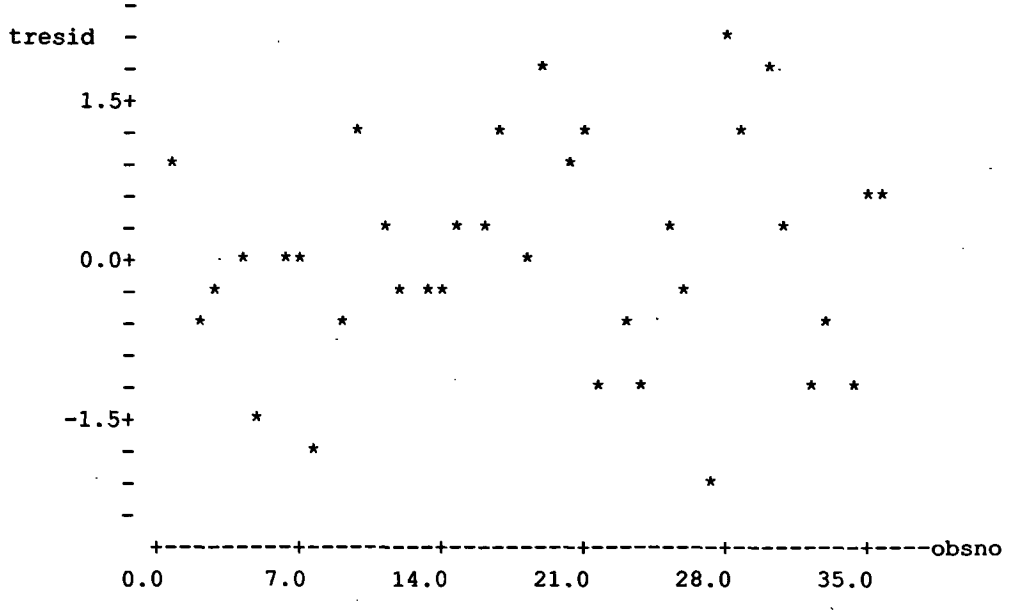
obsno	premium	age	price	fits	residual	stdresid	tresid
1	221	57	11804	182.87	38.1338	0.95962	0.95844
2	448	8	12926	477.10	-29.1040	-0.72149	-0.71614
3	515	6	14054	531.90	-16.8965	-0.41918	-0.41388
4	632	12	17486	636.98	-4.9763	-0.12178	-0.11995
5	48	47	8700	111.15	-63.1519	-1.57678	-1.61473
6	189	30	8570	192.71	-3.7062	-0.09163	-0.09025
7	581	34	18982	583.93	-2.9259	-0.07124	-0.07016
8	102	39	9198	171.64	-69.6364	-1.71939	-1.77448
9	404	33	14986	431.05	-27.0514	-0.65491	-0.64914
10	83	59	8473	40.98	42.0176	1.07656	1.07924
11	280	56	13891	270.47	9.5287	0.23852	0.23508
12	565	13	16127	578.15	-13.1512	-0.32131	-0.31690
13	1105	10	29480	1121.33	-16.3350	-0.43316	-0.42776
14	388	46	15868	399.62	-11.6244	-0.28516	-0.28115
15	435	2	10782	422.94	12.0571	0.30633	0.30208
16	309	11	8645	292.56	16.4359	0.41454	0.40927
17	322	17	9086	279.40	42.5995	1.06031	1.06237
18	741	32	22559	735.53	5.4651	0.13511	0.13309
19	500	34	14969	425.28	74.7202	1.80976	1.87775
20	626	1	14861	589.30	36.7021	0.91975	0.91754
21	1051	34	29733	1008.95	42.0543	1.12130	1.12584
22	845	4	22893	891.53	-46.5284	-1.17327	-1.18023
23	278	59	15198	306.84	-28.8421	-0.72822	-0.72294
24	333	56	16696	381.36	-48.3615	-1.21368	-1.22275
25	650	34	20411	640.42	9.5814	0.23453	0.23114
26	772	27	23128	783.53	-11.5273	-0.28539	-0.28138
27	477	19	16507	562.58	-85.5760	-2.07728	-2.19403
28	443	37	13704	359.97	83.0284	2.01678	2.12100

EXHIBIT 5.2.1
(Continued)

29	692	3	16472	642.79	49.2136	1.22453	1.23420
30	618	36	18422	551.59	66.4119	1.61684	1.65923
31	1050	7	27110	1042.94	7.0596	0.18290	0.18020
32	643	45	22968	685.41	-42.4088	-1.06941	-1.07182
33	116	46	9177	135.11	-19.1088	-0.47524	-0.46959
34	269	9	8977	315.89	-46.8884	-1.18467	-1.19221
35	259	38	10514	228.76	30.2385	0.74147	0.73631
36	491	16	13739	468.45	22.5526	0.55066	0.54476

In many regression computer packages, the output will indicate when a sample value should be examined as a possible outlier by tagging a standardized (or studentized deleted) residual if its absolute value is larger than a certain number, generally larger than 2 (some packages use 3). As you can see in (5.2.4) and (5.2.5), sample observations 27 and 28 have been tagged with an R to indicate that these observations have standardized residuals greater than 2 in magnitude and should be examined.

We plot the studentized deleted residuals against the corresponding observation numbers to see if any stand out as being unduly large. We could examine the magnitudes of studentized deleted residuals in the preceding exhibit, but with a plot of them against observation numbers we can view all of them together.



This plot does not indicate the presence of any outliers because none of the studentized deleted residuals are exceptionally large in magnitude. For instance, none of them exceeds 3 units in magnitude, and the studentized deleted residuals for observations 27 and 28 are just slightly greater than 2 in magnitude.

Suppose, for illustration, that the Y value for sample number 36 in Table 5.2.1 had been incorrectly entered as 1491 instead of 491. We carry out the calculation of studentized deleted residuals along with other diagnostic statistics for this modified data set. The results are given in Exhibit 5.2.2.

EXHIBIT 5.2.2
MINITAB Output for Example 5.2.1—Modified Data

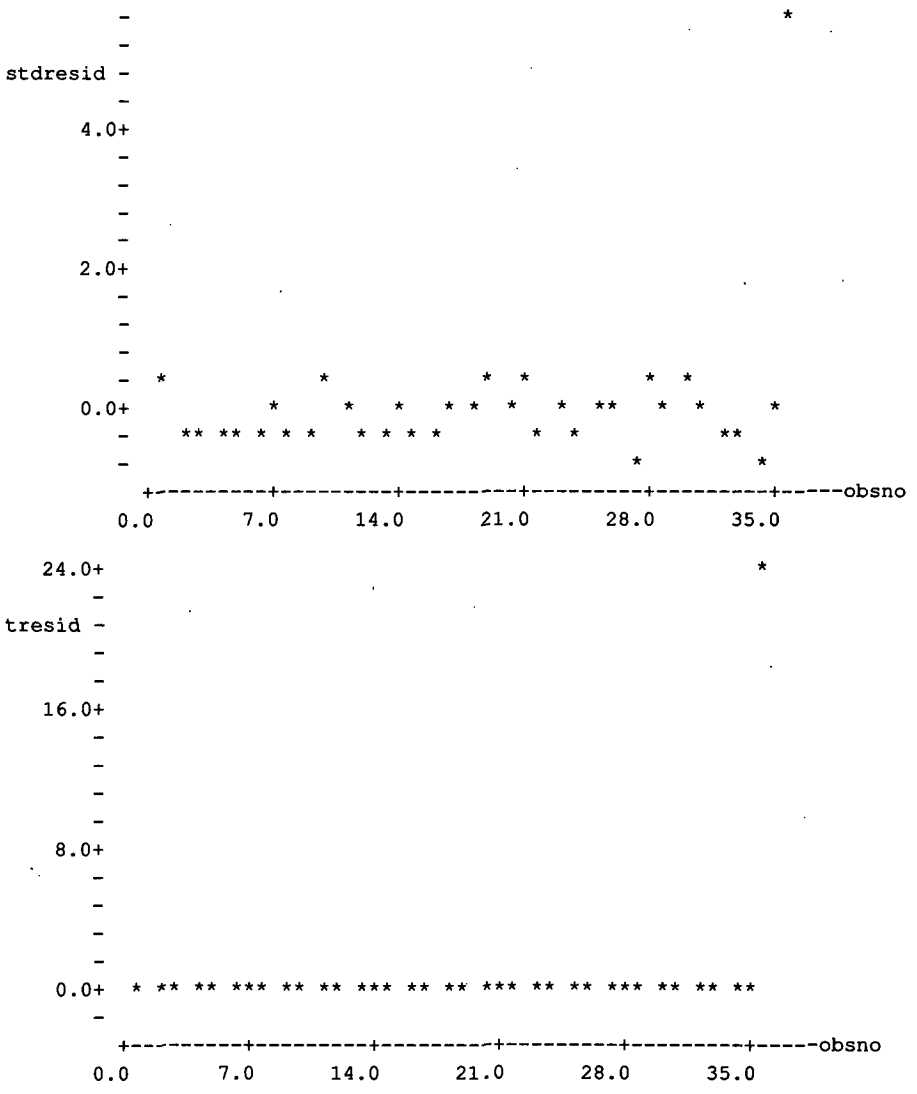
```
The regression equation is
premium = 102 - 6.24 age + 0.0373 price


Predictor       Coef        Stdev    t-ratio        p
Constant       101.8        104.6       0.97    0.338
Age           -6.244        1.662      -3.76    0.001
Price       0.037324     0.005161       7.23    0.000


s = 179.0       R-sq = 70.1%     R-sq(adj) = 68.3%


Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    2     2476278     1238139     38.66    0.000
Error        33     1056902       32027
Total        35     3533181

  Unusual Observations

Obs.     Age     Premium       Fit Stdev.Fit  Residual   St.Resid

36      16.0      1491.0     514.7      38.5     976.3       5.59R

R denotes an obs. with a large st. resid.
```

(5.2.6)

EXHIBIT 5.2.2
(Continued)

obsno	premium	age	price	fits	residual	stdresid	tresid
1	221	57	11804	186.48	34.523	0.20358	0.2006
2	448	8	12926	534.29	-86.293	-0.50129	-0.4955
3	515	6	14054	588.88	-73.881	-0.42951	-0.4241
4	632	12	17486	679.51	-47.515	-0.27248	-0.2686
5	48	47	8700	133.06	-85.061	-0.49768	-0.4919
6	189	30	8570	234.35	-45.350	-0.26275	-0.2590
7	581	34	18982	597.99	-16.991	-0.09695	-0.0955
8	102	39	9198	201.60	-99.597	-0.57626	-0.5703
9	404	33	14986	455.09	-51.089	-0.28984	-0.2858
10	83	59	8473	49.66	33.336	0.20015	0.1972
11	280	56	13891	270.62	9.384	0.05505	0.0542
12	565	13	16127	622.55	-57.548	-0.32948	-0.3250
13	1105	10	29480	1139.66	-34.664	-0.21540	-0.2123
14	388	46	15868	406.84	-18.841	-0.10831	-0.1067
15	435	2	10782	491.73	-56.733	-0.33776	-0.3332
16	309	11	8645	355.78	-46.779	-0.27648	-0.2726
17	322	17	9086	334.78	-12.777	-0.07452	-0.0734
18	741	32	22559	743.99	-2.986	-0.01730	-0.0170
19	500	34	14969	448.21	51.789	0.29394	0.2898
20	626	1	14861	650.22	-24.220	-0.14223	-0.1401
21	1051	34	29733	999.26	51.741	0.32328	0.3189
22	845	4	22893	931.27	-86.274	-0.50979	-0.5040
23	278	59	15198	300.67	-22.667	-0.13411	-0.1321
24	333	56	16696	375.31	-42.309	-0.24881	-0.2452
25	650	34	20411	651.33	-1.327	-0.00761	-0.0075
26	772	27	23128	796.44	-24.441	-0.14180	-0.1397
27	477	19	16507	599.27	-122.269	-0.69550	-0.6900
28	443	37	13704	382.27	60.735	0.34571	0.3410
29	692	3	16472	697.86	-5.861	-0.03418	-0.0337
30	618	36	18422	564.60	53.397	0.30464	0.3004
31	1050	7	27110	1069.94	-19.937	-0.12105	-0.1192
32	643	45	22968	678.08	-35.084	-0.20732	-0.2043
33	116	46	9177	157.11	-41.108	-0.23957	-0.2361
34	269	9	8977	380.66	-111.658	-0.66109	-0.6553
35	259	38	10514	256.96	2.041	0.01173	0.0116
36	1491	16	13739	514.69	976.312	5.58610	23.5827

In (5.2.6) we note that observation number 36 is tagged as having a standardized residual greater than 2. We now plot the standardized residuals r_i against observation numbers, and also plot the studentized deleted residuals T_i against the observation numbers, to see if any of these residuals stand out as being unduly large.



Both plots clearly indicate that observation 36 is an outlier, although the studentized deleted residual for case 36 is considerably bigger than its standardized residual. This demonstrates the fact that it is often easier to identify outlying observations based on their studentized deleted residuals T_i than on their standardized residuals r_i.

In a real problem we would have to look into why observation 36 is an outlier and make a decision as to whether or not it should be included in any further analyses. As

mentioned earlier, if no explanation can be found, we should carry out the analysis both ways, once with the questionable point included and once with it excluded.

Problems 5.2

5.2.1 Consider Task 3.4.1 where crystalline forms of certain chemical compounds are used in various electronic devices and where it is often more desirable to have large crystals than small ones. Crystals of one particular compound are to be produced by a commercial process, and an investigator wants to examine the relationship between Y, the weight of a crystal in grams and X, the time taken (in hours) for the crystal to grow to its final size. The following data are from a laboratory study in which 14 crystals of various sizes were obtained by allowing the crystals to grow for different preselected amounts of time. The data are reproduced in Table 5.2.2, and are also stored in the file **crystal.dat** on the data disk. A SAS output from a regression analysis of these data is in Exhibit 5.2.3.

TABLE 5.2.2

Crystal Number	Weight Y (in grams)	Time X (in hours)
1	0.08	2
2	1.12	4
3	4.43	6
4	4.98	8
5	4.92	10
6	7.18	12
7	5.57	14
8	8.40	16
9	8.81	18
10	10.81	20
11	11.16	22
12	10.12	24
13	13.12	26
14	15.04	28

EXHIBIT 5.2.3
SAS Output for Problem 5.2.1

The SAS System

0:00 Saturday, Jan 1, 1994

Model: MODEL1
Dependent Variable: WEIGHT

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	230.63070	230.63070	204.578	0.0001
Error	12	13.52819	1.12735		
C Total	13	244.15889			

Root MSE	1.06177	R-square	0.9446
Dep Mean	7.55286	Adj R-sq	0.9400
C.V.	14.05782		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	0.001429	0.59938725	0.002	0.9981
TIME	1	0.503429	0.03519723	14.303	0.0001

OBS	WEIGHT	TIME	FITS	RESIDUAL	STDRESID	TRESID
1	0.08	2	1.0083	-0.92829	-1.01438	-1.01572
2	1.12	4	2.0151	-0.89514	-0.94518	-0.94063
3	4.43	6	3.0220	1.40800	1.44726	1.52513
4	4.98	8	4.0289	0.95114	0.95781	0.95424
5	4.92	10	5.0357	-0.11571	-0.11481	-0.10998
6	7.18	12	6.0426	1.13743	1.11767	1.13055
7	5.57	14	7.0494	-1.47943	-1.44682	-1.52456
8	8.40	16	8.0563	0.34371	0.33614	0.32335
9	8.81	18	9.0631	-0.25314	-0.24874	-0.23877
10	10.81	20	10.0700	0.74000	0.73420	0.71929
11	11.16	22	11.0769	0.08314	0.08373	0.08018
12	10.12	24	12.0837	-1.96371	-2.01847	-2.37793
13	13.12	26	13.0906	0.02943	0.03107	0.02975
14	15.04	28	14.0974	0.94257	1.02999	1.03285

a Examine a plot of Y versus X and decide if there appear to be any outliers in this data set. If so, state which observations you regard as outliers. Why?

b Examine the standardized residuals and studentized deleted residuals given in the computer output in Exhibit 5.2.3 and decide whether there appear to be any outliers in this data set. If so, state which observations you regard as outliers. Why?

5.2.2 For this problem the data are the same as in Table 5.2.1 except one Y value has been changed (suppose it was incorrectly recorded). A SAS output containing various diagnostic statistics for the incorrect data is given in Exhibit 5.2.4. Identify the sample item that has been recorded incorrectly by first looking at the studentized deleted residuals and then by checking the standardized residuals.

EXHIBIT 5.2.4
SAS Output for Problem 5.2.2

The SAS System

0:00 Saturday, Jan 1, 1994

Model: MODEL1
Dependent Variable: PREMIUM

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	2214896.6332	1107448.3166	40.958	0.0001
Error	33	892283.67239	27038.89916		
C Total	35	3107180.3056			

Root MSE	164.43509	R-square	0.7128
Dep Mean	513.36111	Adj R-sq	0.6954
C.V.	32.03108		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-53.402539	96.13561819	-0.555	0.5823
AGE	1	-2.472629	1.52683741	-1.619	0.1149
PRICE	1	0.040413	0.00474230	8.522	0.0001

EXHIBIT 5.2.4
(Continued)

OBS	PREMIUM	AGE	PRICE	FITS	RESIDUAL	STDRESID	TRESID
1	221	57	11804	282.69	-61.695	-0.39595	-0.3908
2	448	8	12926	449.20	-1.197	-0.00757	-0.0075
3	515	6	14054	499.73	15.271	0.09662	0.0952
4	632	12	17486	623.59	8.409	0.05248	0.0517
5	48	47	8700	181.98	-133.979	-0.85315	-0.8495
6	189	30	8570	218.76	-29.760	-0.18765	-0.1849
7	581	34	18982	629.65	-48.651	-0.30211	-0.2979
8	102	39	9198	221.89	-119.885	-0.75493	-0.7499
9	404	33	14986	470.63	-66.633	-0.41142	-0.4062
10	83	59	8473	143.13	-60.133	-0.39294	-0.3878
11	280	56	13891	369.51	-89.510	-0.57143	-0.5655
12	565	13	16127	566.20	-1.197	-0.00746	-0.0073
13	1105	10	29480	1113.25	-8.252	-0.05581	-0.0550
14	388	46	15868	474.13	-86.133	-0.53888	-0.5330
15	435	2	10782	377.39	57.613	0.37331	0.3684
16	309	11	8645	268.77	40.229	0.25877	0.2551
17	322	17	9086	271.76	50.243	0.31894	0.3146
18	741	32	22559	779.15	-38.154	-0.24057	-0.2371
19	500	34	14969	467.47	32.527	0.20092	0.1980
20	626	1	14861	544.71	81.295	0.51957	0.5137
21	1051	34	29733	1064.13	-13.133	-0.08931	-0.0880
22	845	4	22893	861.89	-16.886	-0.10859	-0.1070
23	1278	59	15198	414.91	863.088	5.55770	21.6335
24	333	56	16696	482.87	-149.869	-0.95922	-0.9580
25	650	34	20411	687.40	-37.402	-0.23349	-0.2301
26	772	27	23128	814.51	-42.513	-0.26843	-0.2646
27	477	19	16507	566.72	-89.718	-0.55543	-0.5495
28	443	37	13704	408.93	34.068	0.21105	0.2080
29	692	3	16472	604.87	87.134	0.55294	0.5470
30	618	36	18422	602.07	15.926	0.09888	0.0974
31	1050	7	27110	1024.89	25.110	0.16592	0.1635
32	643	45	22968	763.54	-120.539	-0.77521	-0.7704
33	116	46	9177	203.73	-87.728	-0.55644	-0.5505
34	269	9	8977	287.13	-18.133	-0.11684	-0.1151
35	259	38	10514	277.54	-18.542	-0.11596	-0.1142
36	491	16	13739	462.27	28.728	0.17889	0.1762

5.3 Leverages or Hat Values

The **hat matrix** H was defined in (4.5.2), and the diagonal elements $h_{i,i}$ of the hat matrix were used in the calculation of standardized residuals (see (4.5.1)) and also the studentized deleted residuals (see (5.2.2) and (5.2.3)). The numbers $h_{i,i}$ are called **leverages** or **hat values.**

Hat values (leverages) are determined entirely by the sample values of the predictor variables, $X_1, \ldots, X_k$, and are not affected by the values of the response variable Y.

Hat values have a practical interpretation. Loosely speaking, the value $h_{i,i}$ is a measure of how *typical* or *atypical* the values of the predictor variables are for observation i. Recall that hat values are used in calculating $SE(\hat{\mu}_Y(x_{i,1}, \ldots, x_{i,k}))$ and $SE(\hat{Y}(x_{i,1}, \ldots, x_{i,k}))$. In fact, from (4.6.5), using $x^T = (x_{i,1}, \ldots, x_{i,k})$ we have

$$SE(\hat{\mu}_Y(x_{i,1}, \ldots, x_{i,k})) = \hat{\sigma}\sqrt{h_{i,i}}$$

and from (4.6.6) we have

$$SE(\hat{Y}(x_{i,1}, \ldots, x_{i,k})) = \hat{\sigma}\sqrt{1 + h_{i,i}}$$

To better understand the practical meaning of the leverages, it is useful to explicitly examine what they are in straight line regression. Recall that for straight line regression ($k = 1$), the observations satisfy the model

$$y = X\beta + e$$

in matrix form where

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad X\beta + e = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \tag{5.3.1}$$

The hat matrix is

$$H = X(X^TX)^{-1}X^T \tag{5.3.2}$$

$$= \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix}$$

From this we find that

$$h_{i,i} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SSX} \qquad \text{for} \qquad i = 1, \ldots, n \tag{5.3.3}$$

From (5.3.3) we see that the values of x_i that are far away from $\bar{x}$ will result in large values of $h_{i,i}$ relative to the others. It can be shown mathematically that $\sum_{i=1}^n h_{i,i} = 2$

for the straight line regression model and, more generally, that

$$\sum_{i=1}^{n} h_{i,i} = p \tag{5.3.4}$$

for the multiple regression model, where p is the number of β parameters in the regression function. The *average* of $h_{1,1}, \ldots, h_{n,n}$ is thus equal to p/n. In a set of data, if any $h_{i,i}$ value is large relative to p/n, say greater than $2p/n$ or $3p/n$, then the corresponding sample item may be considered to have an unusual set of values for the predictor variables, and the corresponding data point is called a *high leverage point*. Some of the possible reasons for such unusual values are

- The values of one or more of the predictor variables are recorded incorrectly.
- The corresponding sample item may not actually belong to the population under investigation.
- The investigator may have designed the study to include population items that have extreme values for the predictor variables. If the investigator is certain that the regression function being used is correct, then such a design may result in more precise estimates of the parameters. In this case, however, the investigator is already aware of such extreme cases and the hat values (leverages) provide a confirmation of this.

Roughly speaking, if $h_{i,i}$ is large, then for multiple regression with k predictors, this implies that the ith observation $x_{i,1}, \ldots, x_{i,k}$ is far removed from the means $\bar{x}_1, \ldots, \bar{x}_k$ of all observations. This unusual set of hat values might have an influence on point estimates and confidence intervals for the population regression quantities. This influence may be useful or it may be detrimental. An examination of the hat values is thus advisable, particularly in observational studies where it is not uncommon for the sample to be contaminated with observations that really do not belong to the study population. They are especially useful in conjunction with the procedures discussed in the next section for examining *influential observations*. Example 5.3.1 illustrates the use of hat values to find unusual predictor (X) values.

EXAMPLE 5.3.1

A wildlife biologist interested in a certain species of mammals has collected the data in Table 5.3.1 on 25 offspring of that species under the age of 12 months. The data were obtained by simple random sampling from a target population. The quantities observed are

Y = weight of the offspring (in pounds)
X_1 = age of the offspring (in months)
X_2 = length of the offspring (in inches)

The investigator is interested in the relationship among Y, X_1, and X_2. It is supposed that assumptions (A) hold and the regression function of Y on X_1 and X_2 is given by

$$\mu_Y(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

The data are also stored in the file **mammalwt.dat** on the data disks.

Using MINITAB we have computed several relevant quantities that are given in Exhibit 5.3.1. This exhibit also contains a plot of the hat values (`hatvals`) against observation numbers and a plot of length X_2 against age X_1.

Observe that the hat value corresponding to sample observation 18 is high relative to the other hat values. The value of p in this problem is 3 because there are three β parameters in the regression function. Hence $p/n = 3/25 = 0.12$ and $3p/n = 0.36$. The value $h_{18,18}$ is 0.618561, which is much greater than $3p/n = 0.36$. We therefore conclude that the values of X_1 and X_2 for offspring 18 are unusual relative to the X_1 and X_2 values for the other offspring in the sample. From Table 5.3.1

TABLE 5.3.1
Mammal Offspring Data

Observation Number	Weight Y	Age X_1	Length X_2
1	16.4	10	23.1
2	16.3	7	22.8
3	18.2	12	24.0
4	14.8	6	22.1
5	14.7	5	21.5
6	16.4	11	23.2
7	17.3	10	22.7
8	13.0	1	20.2
9	18.8	12	24.2
10	16.3	11	23.5
11	15.0	8	22.5
12	16.8	9	22.6
13	16.0	8	23.2
14	18.4	11	23.4
15	13.5	5	21.2
16	12.4	2	20.6
17	15.6	9	22.2
18	16.6	4	23.4
19	16.4	10	23.1
20	14.1	3	21.0
21	17.9	10	23.1
22	17.7	11	22.8
23	16.3	10	23.5
24	16.6	8	22.3
25	15.2	6	22.5

EXHIBIT 5.3.1

MINITAB Output for Example 5.3.1

```
The regression equation is
weight = - 5.94 + 0.200 age + 0.902 length

Predictor       Coef      Stdev    t-ratio       p
Constant      -5.939      5.636      -1.05   0.303
Age           0.19969    0.08829       2.26   0.034
Length         0.9021     0.2753       3.28   0.003

s = 0.7261      R-sq = 82.3%     R-sq(adj) = 80.7%


Unusual Observations
Obs.     age     length       Fit  Stdev.Fit  Residual   St.Resid
 18      4.0     16.600    15.970      0.571     0.630       1.41 X          (5.3.5)

hatvals

  0.057404   0.076032   0.121121   0.056115   0.087465   0.084756   0.085440
  0.274175   0.144781   0.079048   0.041413   0.055036   0.091957   0.078075
  0.124700   0.205291   0.109248   0.618561   0.057404   0.149319   0.057404
  0.132624   0.075377   0.052852   0.084405
```

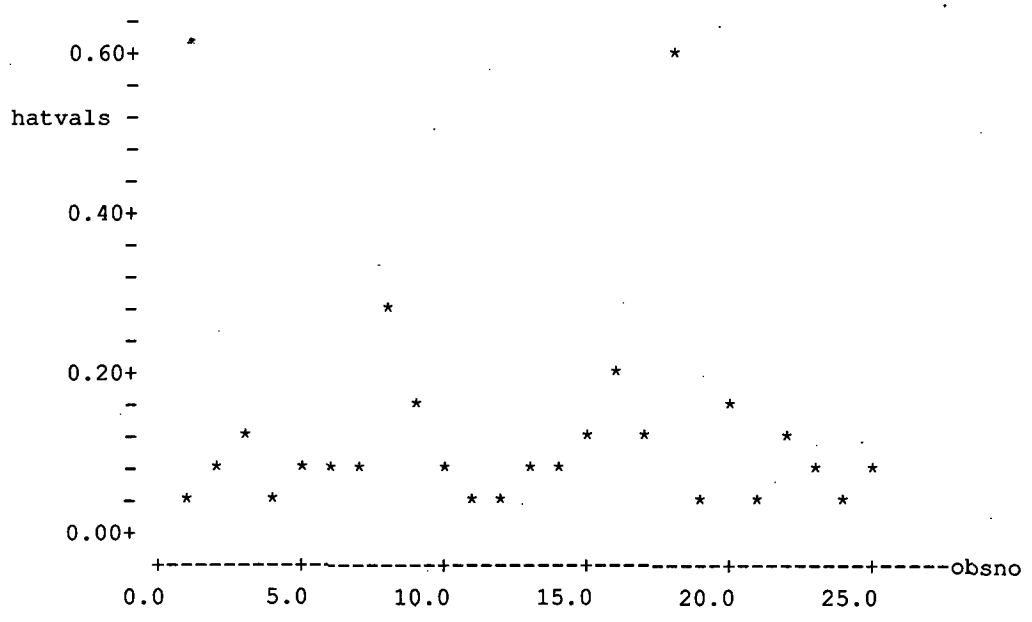
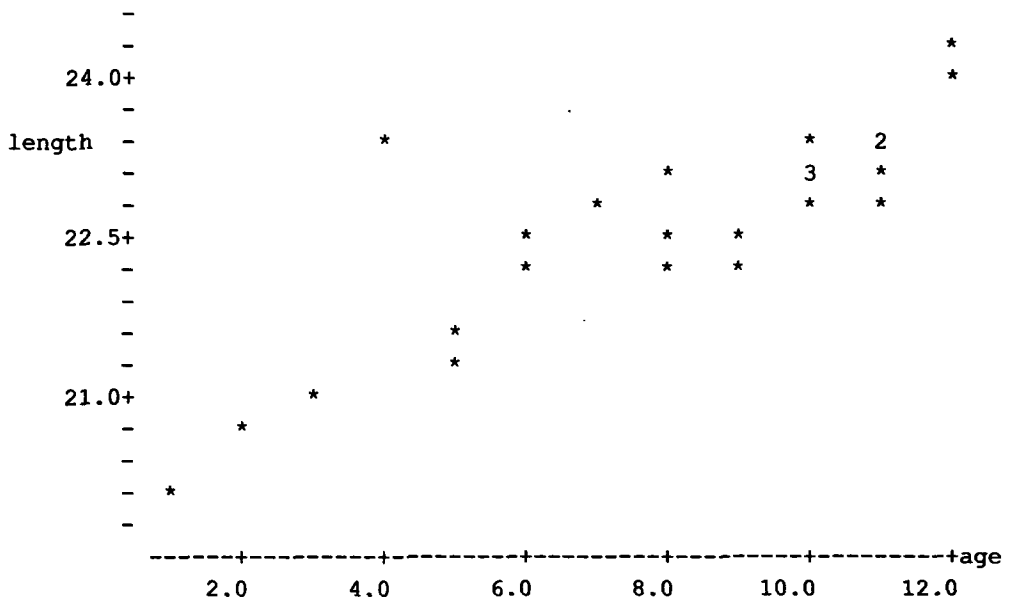
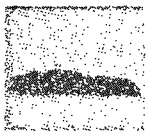


EXHIBIT 5.3.1
(Continued)



we find that the age and length for offspring 18 are 4 months and 23.4 inches, respectively. Neither of these values in itself is unusual; what is unusual is the combination $X_1 = 4, X_2 = 23.4$. Offspring 18 is rather long for its age. An examination of the plot of X_2 against X_1 given in Exhibit 5.3.1 makes this clearer.

Aside from the possibility that this information may be of interest to the investigator in its own right, it is important to be aware of sample observations with high hat values. Such observations may have a strong influence on the estimates of the regression coefficients and, unless the assumed form of the regression function is known to be valid at all of the (x_1, x_2) values occurring in the sample, the estimates and confidence intervals for various quantities in regression may strongly depend on whether or not the observation under question is included in the analysis. Some computer programs warn about this possibility by tagging cases with high hat values. In (5.3.5) of Exhibit 5.3.1 note that observation 18 has been tagged by MINITAB using the symbol "X", indicating that the hat value for observation 18 is greater than $3p/n$. Further examination of such cases is warranted, and we discuss this in the next section. ■



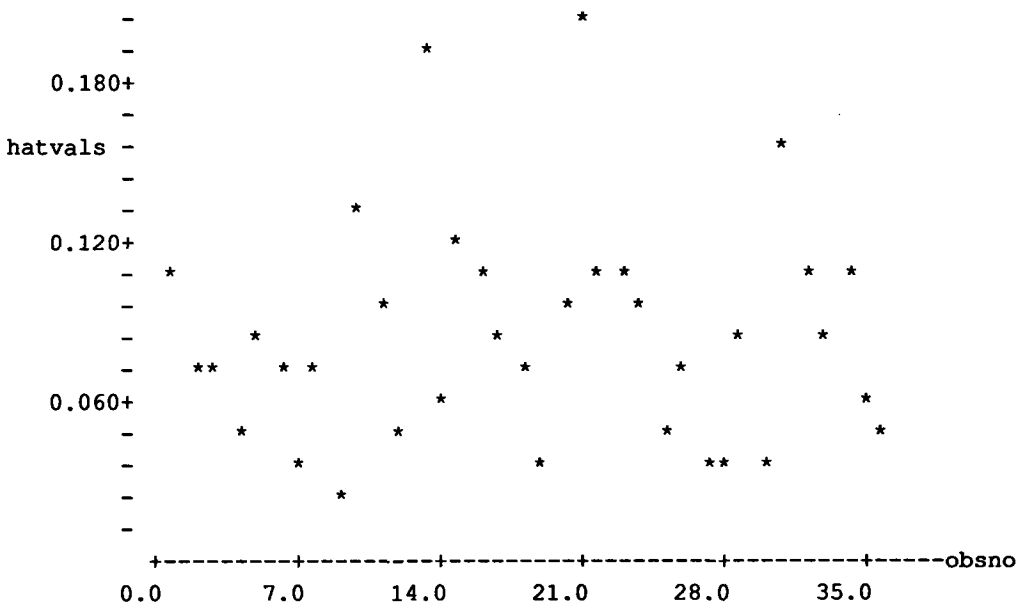
Problems 5.3

Problems 5.3.1–5.3.3 refer to the insurance premium data in Table 5.2.1. The MINITAB output from a regression analysis of these data is given in Exhibit 5.3.2. The exhibit also contains the hat values and a plot of the hat values against observation numbers.

EXHIBIT 5.3.2
MINITAB Output for Problems 5.3.1–5.3.3

```
The regression equation is
premium = 6.9 - 5.10 age + 0.0395 price

Predictor       Coef       Stdev    t-ratio        p
Constant        6.90       24.52       0.28    0.780
Age          -5.0996      0.3894     -13.10    0.000
Price       0.039533    0.001209      32.69    0.000

s = 41.94       R-sq = 97.7%     R-sq(adj) = 97.6%


Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    2     2492087     1246043    708.49    0.000
Error        33       58038        1759
Total        35     2550125


  hatvals

  0.102106   0.074772   0.076159   0.050586   0.087922   0.069855   0.040922
  0.067328   0.029894   0.133848   0.092546   0.047447   0.191368   0.055131
  0.119122   0.106149   0.082200   0.069749   0.030746   0.094581   0.200206
  0.105777   0.108070   0.097195   0.051014   0.072351   0.035019   0.036306
  0.081593   0.040688   0.152939   0.105813   0.080722   0.109292   0.054343
  0.046241
```

EXHIBIT 5.3.2
(Continued)



5.3.1 **a** What is the value of n? What is the value of p? What is the value of $2p/n$?

b Are there any hat values that are larger than $2p/n$? If so, what are they?

5.3.2 Carry out an analysis of hat values for these data and explain your findings in a short report.

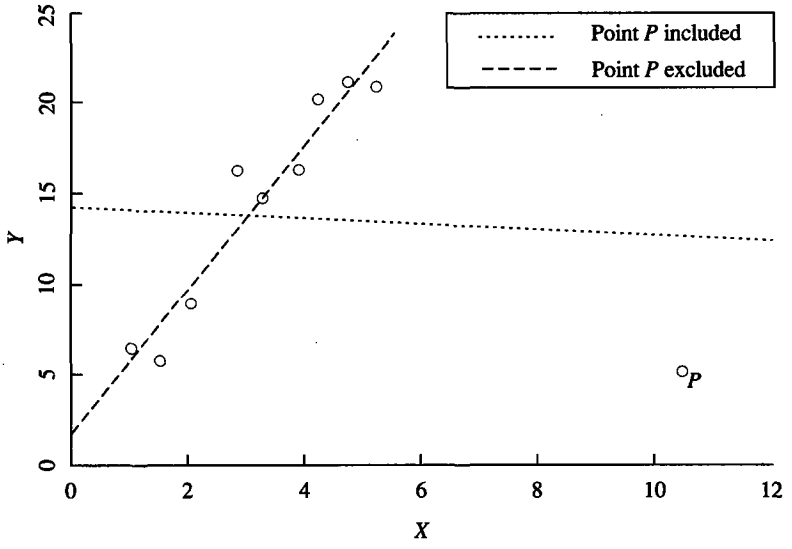
5.3.3 If the Y value (premium) for car 8 in the sample is changed from its value 102 to an incorrect value 302, how will this change the hat value $h_{8,8}$? How will it change the hat values $h_{i,i}$ for $i \neq 8$?

5.4
Influential Observations—Cook's Distance and DFFITS

There are instances when we find that the conclusions derived from a regression analysis are highly influenced by one or more specific sample observations. A sample observation is said to be an **influential observation** if the exclusion of this observation from the analysis results in conclusions that are very different from the conclusions reached when this observation is included in the analysis. Figure 5.4.1 illustrates one such situation.

Observe that while most of the points in Figure 5.4.1 suggest a straight line regression function with a positive slope, the point with the largest X value (this point is labeled as P in the figure) will unduly influence the estimation of the slope

FIGURE 5.4.1



and the intercept, and its inclusion in the analysis will actually result in a negative estimate for the slope. In this section we discuss two commonly used diagnostic measures for identifying such influential observations. They are:

1 Cook's distance
2 DFFITS

Cook's Distance

The *influence* of any data point may be assessed by examining the amount by which the estimates of the regression coefficients (i.e., β parameters) change if this data point is deleted from the analysis. This approach leads to the measure known as *Cook's distance* [1], [2].

Cook's distance for sample observation i is defined by

$$c_i = \frac{1}{p}\left(\frac{h_{i,i}}{1-h_{i,i}}\right) r_i^2 \tag{5.4.1}$$

where $h_{i,i}$ is the hat value for sample item i, r_i is the ith standardized residual defined in (4.5.1), and p is the number of β's in the regression function. If $\hat{\beta}_j$ is the estimate of β_j using all of the data points and $\hat{\beta}_{j(-i)}$ is the estimate of β_j using all of the data points *but excluding observation i*, then c_i *is a measure of the difference between*

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$$

and

$$(\hat{\beta}_{0(-i)}, \hat{\beta}_{1(-i)}, \ldots, \hat{\beta}_{k(-i)})$$

A large value of c_i indicates that the estimates of some or all of the β's will change substantially if observation i is omitted from the analysis. Some authors recommend examining those cases for which c_i is greater than the tabled F-value $F_{0.50;p,n-p}$ in Table T-5 of Appendix T. The interpretation of Cook's distance is illustrated in Example 5.4.1.

DFFITS

The *influence* of an individual sample observation may also be assessed by examining, for each i, the amount by which the predicted Y value for sample item i changes when this item is excluded from the analysis; i.e., by examining the quantity

$$\hat{Y}(x_{i,1}, \ldots, x_{i,k}) - \hat{Y}_{(-i)}(x_{i,1}, \ldots, x_{i,k})$$

We sometimes write this quantity as $\hat{Y}_i - \hat{Y}_{(-i)i}$ for ease of notation. We first standardize this quantity by dividing by its standard error based on data with the ith sample value removed. This standard error is $\hat{\sigma}_{(-i)}\sqrt{h_{i,i}}$, where $h_{i,i}$ is the hat value for the ith observation using all the data and the standardized value is called DFFITS (*dif*ference in the *fit*ted value–standardized) for the ith observation. Thus we define

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{(-i)i}}{\hat{\sigma}_{(-i)}\sqrt{h_{i,i}}} \tag{5.4.2}$$

If the magnitude of $\hat{Y}_i - \hat{Y}_{(-i)i}$ is large relative to its standard error, this tells us that it makes a difference whether or not we include the ith observation in predicting $Y(x_{i,1}, \ldots, x_{i,k})$. In other words, if the absolute value of DFFITS_i is large, the ith sample observation is said to influence the estimate of $Y(x_{i,1}, \ldots, x_{i,k})$ and hence is called an *influential observation*. An alternate formula for computing DFFITS_i is

$$\text{DFFITS}_i = T_i\sqrt{\frac{h_{i,i}}{1 - h_{i,i}}} \tag{5.4.3}$$

where T_i is the studentized deleted residual for sample item i defined in (5.2.3). Some authors have suggested that observations for which the absolute value of DFFITS_i is greater than $2\sqrt{p/n}$ may be considered *influential* and should be examined.

Example 5.4.1 illustrates the use and interpretation of Cook's distance and DFFITS.

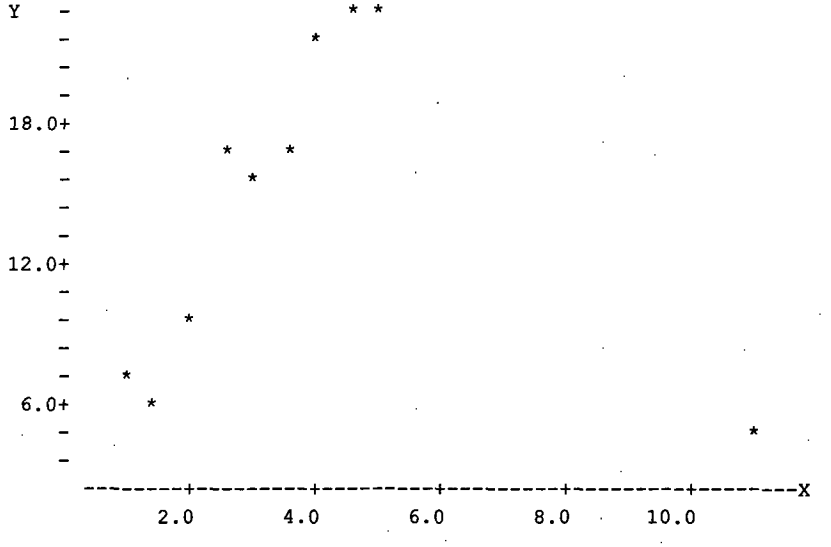
EXAMPLE 5.4.1

Consider the artificial data in Table 5.4.1, which consists of 10 observations of a predictor variable X and a response variable Y. The data are also stored in the file **table 541.dat** on the data disk.

TABLE 5.4.1

Observation Number	*Y*	*X*
1	6.8	1.0
2	6.1	1.5
3	9.5	2.0
4	16.2	2.5
5	15.2	3.0
6	16.4	3.5
7	21.5	4.0
8	22.5	4.5
9	22.3	5.0
10	4.7	11.0

The plot of these data is shown in Figure 5.4.1. Note that if we were to carry out a straight line regression analysis for these data, the results using all the data values would be very different from the results obtained when sample observation 10 is excluded. The fitted straight lines with and without sample observation 10 are also shown in Figure 5.4.1. Thus it is clear that observation 10 is an *influential* data point. An examination of the Cook's distances or DFFITS should also reveal this. First we plot *Y* against *X*.



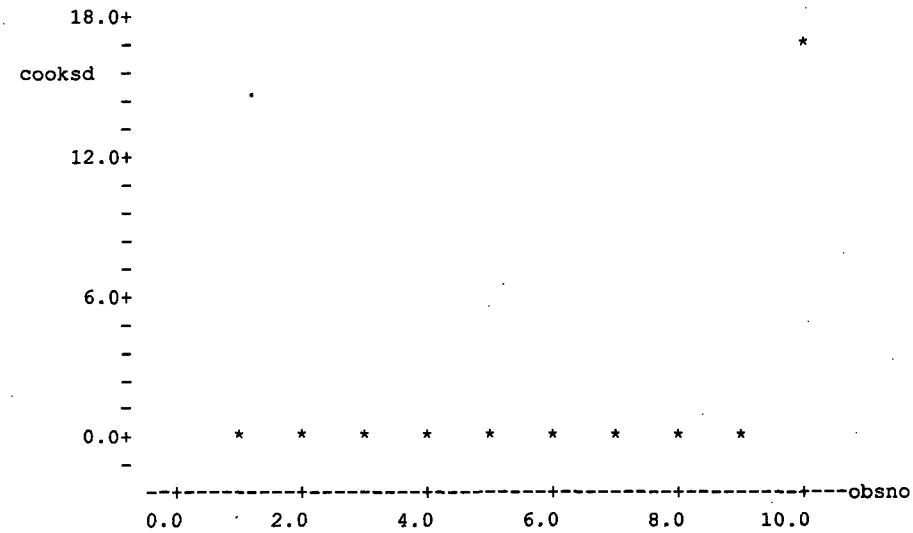
Next we regress *Y* on *X* and compute the standardized residuals (`stdresid`), Cook's distances (`cooksd`), DFFITS (`dffits`), and hat values (`hatvals`). The results are in Exhibit 5.4.1.

EXHIBIT 5.4.1
MINITAB Output for Example 5.4.1

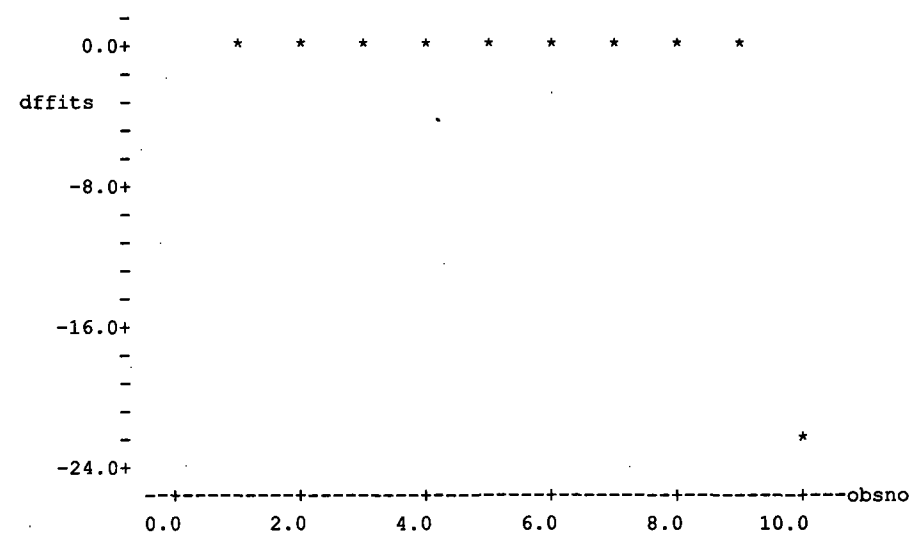
```
The regression equation is
Y = 14.5 - 0.106 X

Predictor       Coef      Stdev    t-ratio       p
Constant      14.521      4.006       3.63   0.007
X            -0.1055     0.8599      -0.12   0.905

s = 7.327       R-sq = 0.2%      R-sq(adj) = 0.0%

Analysis of Variance

SOURCE       DF          SS          MS         F       p
Regression    1        0.81        0.81      0.02   0.905
Error         8      429.47       53.68
Total         9      430.28

Unusual Observations
Obs.       X         Y       Fit Stdev.Fit  Residual   St.Resid
 10     11.0      4.70     13.36      6.61     -8.66    -2.74RX

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.

obsno       Y      X    stdresid    hatvals     cooksd    dffits

    1     6.8    1.0    -1.16791   0.207989     0.1791   -0.6147
    2     6.1    1.5    -1.23997   0.172865     0.1607   -0.5900
    3     9.5    2.0    -0.70981   0.144628     0.0426   -0.2820
    4    16.2    2.5     0.28319   0.123278     0.0056    0.0998
    5    15.2    3.0     0.14394   0.108815     0.0013    0.0471
    6    16.4    3.5     0.32368   0.101240     0.0059    0.1023
    7    21.5    4.0     1.06509   0.100551     0.0634    0.3596
    8    22.5    4.5     1.22081   0.106749     0.0891    0.4376
    9    22.3    5.0     1.20843   0.119835     0.0994    0.4613
   10     4.7   11.0    -2.74104   0.814050    16.4458  -21.7503
```

An examination of the diagnostic statistics in Exhibit 5.4.1 reveals that observation 10 is an influential observation. Various plots may also be examined.

First we plot Cook's distances against the corresponding observation numbers.

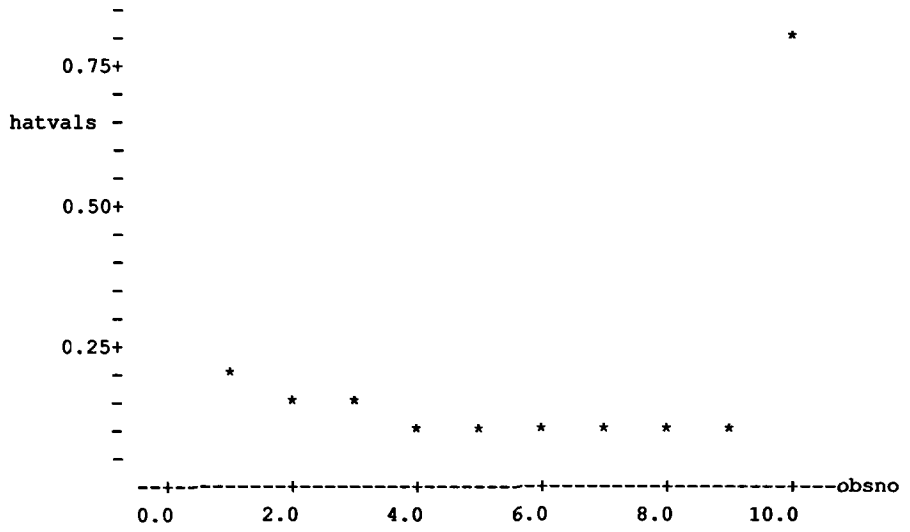


The value of $F_{0.5:2,8}$ is 0.76, as given in Table T-5 in Appendix T and so sample observations with Cook's distances that are greater than 0.76 are candidates for further scrutiny. In this problem the only such case is observation 10. Next we plot DFFITS against sample item numbers.



We observe that the value of DFFITS for case 10 is very large in absolute value relative to the remaining cases. The value of $2\sqrt{p/n}$ is $2\sqrt{2/10} = 0.894427$, and the value of DFFITS for case 10 is -21.7503, which is larger than 0.894427 in absolute value. Thus case 10 is identified as an influential observation.

Next we examine a plot of the hat values against observation numbers.



This plot indicates that case 10 is a *high leverage point.* It has a hat value of 0.81405, which is larger than $2p/n = 0.4$. Its X value is unusual relative to other X values. We thus have a possible explanation of why case 10 is an *influential* observation. ■

Although in this example an examination of a plot of Y against X would have revealed the fact that case 10 has an unusual X value, and that its inclusion or exclusion from the regression analysis may unduly affect the values of the estimated parameters, this may not be possible in general in problems involving several predictor variables. In such cases, diagnostic statistics such as hat values (leverages), Cook's distances, and DFFITS are of great assistance in identifying and interpreting influential observations.

To see what effect observation 10 has on the conclusions, we now reanalyze the data with case 10 removed. A computer output from a regression analysis for the data in Table 5.4.1, with sample item number 10 removed, is given in Exhibit 5.4.2. Note that the parameter estimates have changed substantially as a result of deleting case 10. When all the data are included in the analysis we get (see Exhibit 5.4.1)

$$\hat{\beta}_0 = 14.521 \qquad \hat{\beta}_1 = -0.1055 \qquad \hat{\sigma} = 7.327 \tag{5.4.4}$$

When case 10 is omitted from the analysis we get (see Exhibit 5.4.2)

$$\hat{\beta}_{(-10),0} = 1.627 \qquad \hat{\beta}_{(-10),1} = 4.5133 \qquad \hat{\sigma}_{(-10)} = 1.932 \tag{5.4.5}$$

The predicted Y value for case 10 using $\hat{\beta}_0$ and $\hat{\beta}_1$ in (5.4.4) is

$$\hat{Y}(11) = 14.521 + 11(-0.1055) = 13.3605$$

but if the estimates in (5.4.5) are used (i.e., if sample number 10 is omitted) we get

$$\hat{Y}_{(-10)}(11) = 1.627 + 11(4.5133) = 51.2733$$

EXHIBIT 5.4.2
MINITAB Output for Example 5.4.1—Observation 10 Removed

Observation Number	Y	X
1	6.8	1.0
2	6.1	1.5
3	9.5	2.0
4	16.2	2.5
5	15.2	3.0
6	16.4	3.5
7	21.5	4.0
8	22.5	4.5
9	22.3	5.0

The regression equation is
Y = 1.63 + 4.51 X

Predictor	Coef	Stdev	t-ratio	p
Constant	1.627	1.629	1.00	0.351
X	4.5133	0.4988	9.05	0.000

s = 1.932 R-sq = 92.1% R-sq(adj) = 91.0%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	305.55	305.55	81.86	0.000
Error	7	26.13	3.73		
Total	8	331.68			

Clearly we obtain substantially different values for $\hat{\beta}_i$ and for $\hat{Y}(11)$ depending on whether or not we include case 10 in the analysis. In a real study the investigator would now look for possible explanations for observation 10 being unusual. It may be that the values for this item are incorrectly recorded, or it may be the case that a straight line regression model is not appropriate over the entire range of data values. The true population regression function may be a quadratic function or some other type of function over the range of the X values in the sample data. The investigator has to make a decision, using all available information and having considered various possible explanations, about how to treat influential observations. When no explanations are available concerning influential observations, it may be

best to present results from both analyses, one with the sample item in question included and one with it excluded.

EXAMPLE 5.4.2

The data for this example were obtained by a slight modification of the data for Example 5.4.1. Specifically, suppose that the Y value for item 10 had been wrongly recorded as 4.7 instead of the correct value, which equals 52.2. Table 5.4.2 gives the corrected data set. These data are also stored in the file **table542.dat** on the data disk. The results of a regression analysis of these data, performed using MINITAB, appear in Exhibit 5.4.3.

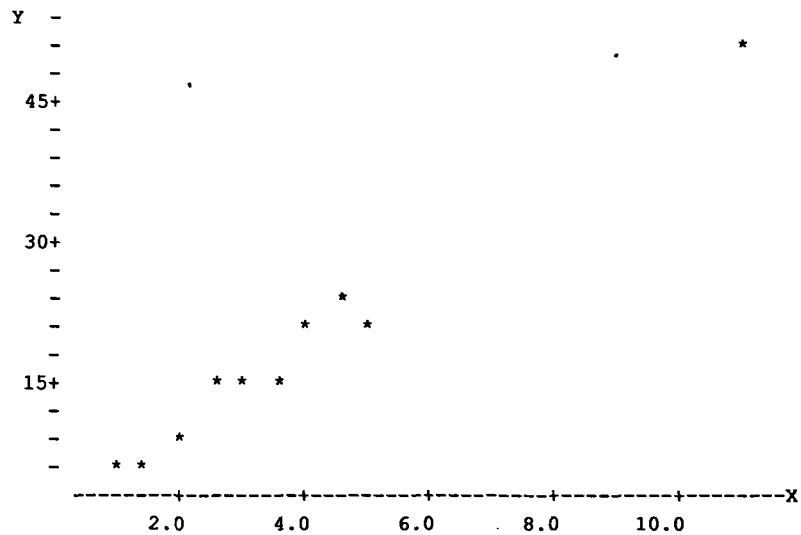
TABLE 5.4.2

Observation Number	Y	X
1	6.8	1.0
2	6.1	1.5
3	9.5	2.0
4	16.2	2.5
5	15.2	3.0
6	16.4	3.5
7	21.5	4.0
8	22.5	4.5
9	22.3	5.0
10	52.2	11.0

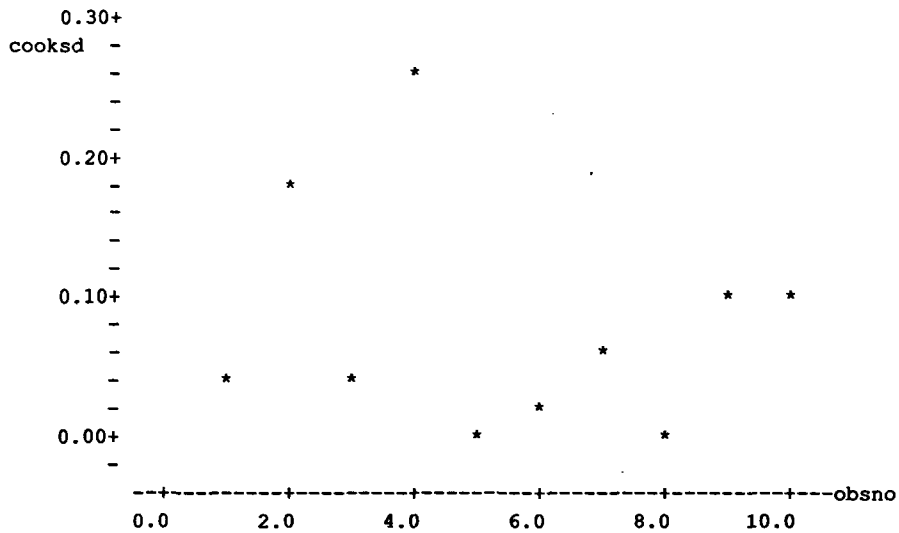
EXHIBIT 5.4.3
MINITAB Output for Example 5.4.2

```
The regression equation is
Y = 1.37 + 4.61 X

Predictor       Coef      Stdev    t-ratio       p
Constant      1.3701     0.9910       1.38   0.204
X             4.6052     0.2127      21.65   0.000

s = 1.813       R-sq = 98.3%     R-sq(adj) = 98.1%

Analysis of Variance

SOURCE       DF          SS          MS         F       p
Regression    1      1539.7      1539.7    468.59   0.000
Error         8        26.3         3.3
Total         9      1566.0

Unusual Observations

Obs.       X         Y       Fit Stdev.Fit  Residual   St.Resid
 10     11.0    52.200    52.028     1.636     0.172     0.22 X          (5.4.6)
X denotes an obs. whose X value gives it large influence.

obsno      Y      X  stdresid      hatvals     cooksd      dffits

    1    6.8    1.0   0.51119     0.207989   0.034312   0.249145
    2    6.1    1.5  -1.32110     0.172865   0.182378  -0.638922
    3    9.5    2.0  -0.64455     0.144628   0.035122  -0.254616
    4   16.2    2.5   1.95417     0.123278   0.268486   0.948144
    5   15.2    3.0   0.00829     0.108815   0.000004   0.002710
    6   16.4    3.5  -0.63336     0.101240   0.022593  -0.204024
    7   21.5    4.0   0.99407     0.100551   0.055235   0.332089
    8   22.5    4.5   0.23718     0.106749   0.003361   0.076967
    9   22.3    5.0  -1.23266     0.119835   0.103436  -0.472709
   10   52.2   11.0   0.22044     0.814050   0.106369   0.432762
```

We plot Y against X. This plot shows that case 10 has an unusual X value relative to the X values of the other cases, but it appears that the estimated regression line will not change substantially if this point is omitted from the analysis.

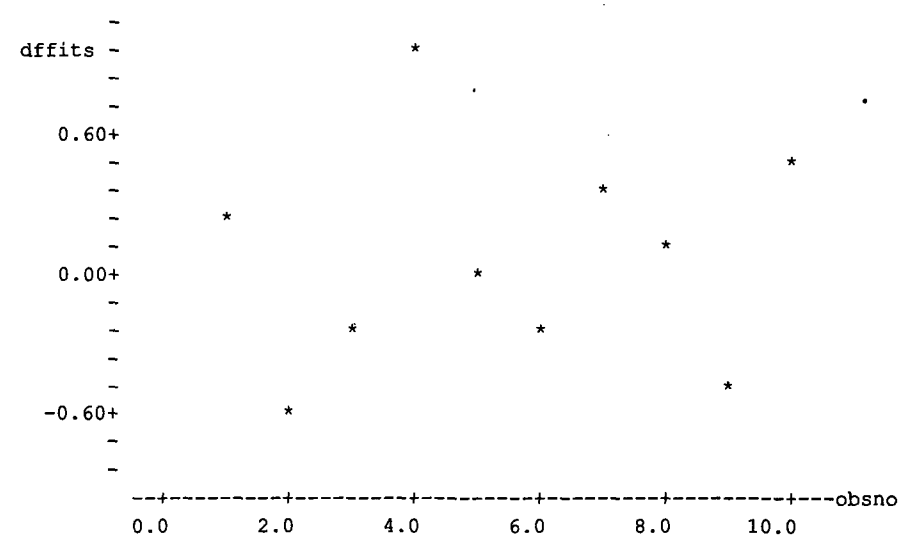


Next we plot Cook's distances against observation numbers.



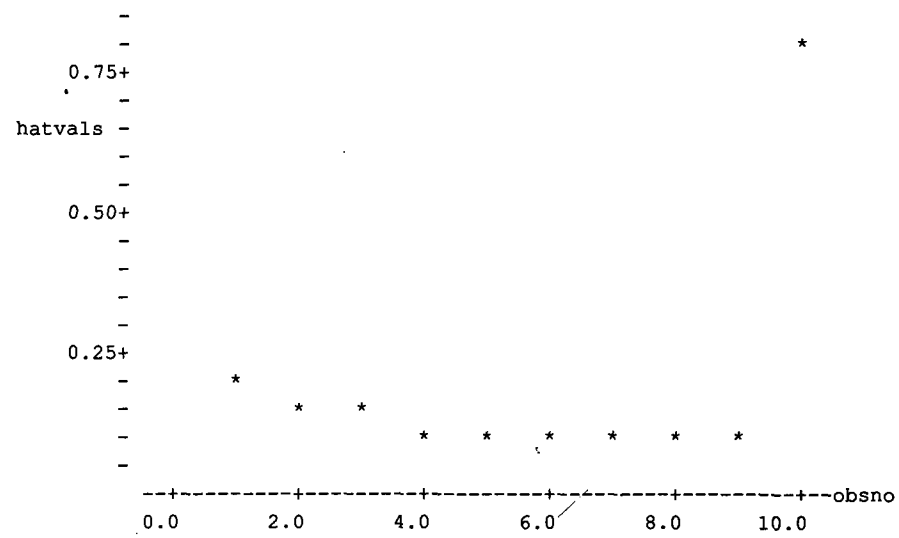
Based on this plot, we do not see any indication of influential points. To confirm this we calculate $F_{0.5;p,n-p} = F_{0.5;2,8} = 0.76$. Because all of the Cook's distances are less than this value, we can conclude that none of the observations is influential.

Next we plot the DFFITS against the observation numbers.



Again, based on the plot of DFFITS against observation numbers, one might conclude that none of the observations is influential since none is greater than $2\sqrt{p/n} = 0.8944$ in absolute value.

Finally, we plot the hat values against the observation numbers.



Note that, although case 10 has an unusual X value relative to the remaining data points, its inclusion or exclusion in the analysis does not substantially affect the estimated regression function. You should confirm this statement by verifying that the estimated regression function when case 10 is excluded is

$$\hat{Y}_{(-10)}(x) = 1.63 + 4.51x$$

which is not very different from the estimated regression function

$$\hat{Y}(x) = 1.37 + 4.61x$$

calculated with case 10 included.

However, you should note that although inclusion or exclusion of observations with large hat values may or may not substantially affect the estimated regression function, it *will* affect the standard errors of predicted values and the standard errors of the estimated regression coefficients, thus affecting the widths of confidence intervals. In the present example note that $SE(\hat{\beta}_i) = 0.4988$ when observation 10 is excluded, but $SE(\hat{\beta}_i) = 0.2127$ when observation 10 is included in the analysis. ■

In Box 5.4.1 we summarize the formulas for computing various diagnostic measures discussed in Sections 5.2–5.4.

BOX 5.4.1 Summary of Formulas for Diagnostic Statistics

Notation:

Y is the response variable.

$X_1, \ldots, X_k$ are predictor variables.

The regression function of Y on $X_1, \ldots, X_k$ is

$$\mu_Y(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

y_i is the observed Y value for sample item i for $i = 1, \ldots, n$.

$x_{i,1}, \ldots, x_{i,k}$ are values of the predictor variables $X_1, \ldots X_k$ for sample item i for $i = 1, \ldots, n$.

$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the least squares estimates of $\beta_0, \beta_1, \ldots, \beta_k$ using all n sample observations.

$\hat{\beta}_{(-i)0}, \hat{\beta}_{(-i)1}, \ldots, \hat{\beta}_{(-i)k}$ are the least squares estimates of $\beta_0, \beta_1, \ldots, \beta_k$ calculated after *leaving out the ith observation* but using the remaining $n - 1$ sample observations.

$\hat{\sigma}$ = the estimated subpopulation standard deviation using all n sample observations.

$\hat{\sigma}_{(-i)}$ = the estimated subpopulation standard deviation calculated after *leaving out the ith observation* but using the remaining $n - 1$ sample observations.

$\hat{Y}(x_{i,1}, \ldots, x_{i,k})$ is the estimated Y value of an item with $X_1 = x_{i,1}, \ldots, X_k = x_{i,k}$ calculated using all n sample observations and is given by

$$\hat{Y}(x_{i,1}, \ldots, x_{i,k}) = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_k x_{i,k}$$

$\hat{Y}_{(-i)}(x_{i,1}, \ldots, x_{i,k})$ is the estimated Y value of an item with $X_1 = x_{i,1}, \ldots, X_k = x_{i,k}$ calculated after *leaving out the ith observation* but

using the remaining $n-1$ sample observations and is given by

$$\hat{Y}_{(-i)}(x_{i,1},\ldots,x_{i,k}) = \hat{\beta}_{(-i)0} + \hat{\beta}_{(-i)1}x_{i,1} + \cdots + \hat{\beta}_{(-i)k}x_{i,k}$$

The **residual** corresponding to observation i is $\hat{e}_i$ and is given by

$$\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_k x_{i,k})$$

The **hat matrix** $\boldsymbol{H}$ is defined by

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

The **hat value** corresponding to observation i is $h_{i,i}$, which is given by

$$h_{i,i} = \text{the } i\text{th diagonal element of the hat matrix } \boldsymbol{H}$$

Observations with $h_{i,i} > 2p/n$ should be examined as potentially influential observations.

The **standardized residual** corresponding to observation i is r_i and is given by

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{i,i}}}$$

Observations with $|r_i| > 2$ should be examined as possible outliers.

Studentized deleted residual for observation i is given by

$$T_i = \frac{y_i - \hat{Y}_{(-i)}(x_{i,1},\ldots,x_{i,k})}{\hat{\sigma}_{(-i)}\Big/\sqrt{1-h_{i,i}}}.$$

Observations with $|T_i| > 2$ should be examined as possible outliers.

Cook's distance for observation i is denoted by c_i and is defined by

$$c_i = \frac{1}{p}\left(\frac{h_{i,i}}{1-h_{i,i}}\right)r_i^2$$

where p is the number of β parameters in the model. Thus $p = k+1$ for the model $\mu_Y(x_1,\ldots,x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$. Observations with $c_i > F_{0.5;p,n-p}$ should be examined as possible influential observations.

DFFITS for observation i is denoted by DFFITS_i and is given by

$$\text{DFFITS}_i = T_i\sqrt{\frac{h_{i,i}}{1-h_{i,i}}}$$

where T_i is the studentized deleted residual for observation i defined above. Observations with $|\text{DFFITS}_i| > 2\sqrt{p/n}$ should be examined as possible influential observations.

Problems 5.4

Problems 5.4.1 and 5.4.2 refer to the data in Table 5.4.2. In Exhibit 5.4.4 is a SAS output for regression analysis of these data. This exhibit also includes several diagnostic measures.

EXHIBIT 5.4.4
SAS Output for Problems 5.4.1–5.4.2

The SAS System

0:00 Saturday, Jan 1, 1994

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	1539.71399	1539.71399	468.585	0.0001
Error	8	26.28701	3.28588		
C Total	9	1566.00100			

Root MSE	1.81270	R-square	0.9832
Dep Mean	18.87000	Adj R-sq	0.9811
C.V.	9.60625		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	1.370110	0.99103083	1.383	0.2042
X	1	4.605234	0.21274399	21.647	0.0001

EXHIBIT 5.4.4
(Continued)

obs	Y	X	fits	residual	stdresid	tresid	hatvals	cooksd	dffits
1	6.8	1.0	5.97534	0.82466	0.51119	0.48618	0.20799	0.03431	0.24914
2	6.1	1.5	8.27796	-2.17796	-1.32110	-1.39760	0.17287	0.18238	-0.63892
3	9.5	2.0	10.58058	-1.08058	-0.64454	-0.61921	0.14463	0.03512	-0.25462
4	16.2	2.5	12.88320	3.31680	1.95417	2.52849	0.12328	0.26849	0.94814
5	15.2	3.0	15.18581	0.01419	0.00829	0.00776	0.10882	0.00000	0.00271
6	16.4	3.5	17.48843	-1.08843	-0.63336	-0.60789	0.10124	0.02259	-0.20402
7	21.5	4.0	19.79105	1.70895	0.99407	0.99323	0.10055	0.05523	0.33209
8	22.5	4.5	22.09366	0.40634	0.23718	0.22264	0.10675	0.00336	0.07697
9	22.3	5.0	24.39628	-2.09628	-1.23266	-1.28110	0.11983	0.10344	-0.47271
10	52.2	11.0	52.02769	0.17231	0.22044	0.20683	0.81405	0.10637	0.43276

5.4.1 From Exhibit 5.4.4 obtain the following:

a $h_{3,3}$ **c** Cook's distance c_4 **e** r_3

b DFFITS_4. **d** $\hat{Y}(3.5)$ **f** $\hat{e}_3$

g Studentized deleted residual T_6

5.4.2 **a** What is the largest (in absolute value) studentized deleted residual? Is this large enough to investigate the corresponding observation as a possible outlier?

b What is the largest hat value? Is this large enough to investigate the corresponding observation as a possible influential observation (high leverage point)?

c What is the largest Cook's distance? Is this large enough to conclude that the corresponding observation may be an influential observation?

d What is the largest (in absolute value) DFFITS value? Is this large enough to conclude that the corresponding observation may be an influential observation?

e Write a brief report summarizing your findings in parts (a)–(d).

Problems 5.4.3 through 5.4.6 refer to the data in Table 5.4.3. These data are also stored in the file **table543.dat** on the data disk. SAS output for regression of Y on X is given in Exhibit 5.4.5.

TABLE 5.4.3

Observation Number	Y	X
1	16.3	11.0
2	16.8	12.0
3	20.1	13.0

Observation Number	Y	X
4	27.3	14.0
5	28.3	15.0
6	50.3	32.0

EXHIBIT 5.4.5

SAS Output for Problems 5.4.3–5.4.6 (Regression of Y on X with Observation Number 6 Included)

The SAS System

0:00 Saturday, Jan 1, 1994

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	761.07908	761.07908	62.991	0.0014
Error	4	48.32925	12.08231		
C Total	5	809.40833			

Root MSE	3.47596	R-square	0.9403
Dep Mean	26.51667	Adj R-sq	0.9254
C.V.	13.10859		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	1.219517	3.48898438	0.350	0.7443
X	1	1.564772	0.19715657	7.937	0.0014

obs	Y	X	fits	residual	stdresid	tresid	hatvals	cooksd	dffits
1	16.3	11	18.43201	-2.13201	-0.70945	-0.65714	0.25255	0.08503	-0.38197
2	16.8	12	19.99678	-3.19678	-1.04302	-1.05865	0.22252	0.15568	-0.56636
3	20.1	13	21.56155	-1.46155	-0.46979	-0.41856	0.19893	0.02740	-0.20858
4	27.3	14	23.12633	4.17367	1.32741	1.53687	0.18177	0.19572	0.72437
5	28.3	15	24.69110	3.60890	1.14034	1.20211	0.17105	0.13416	0.54605
6	50.3	32	51.29222	-0.99223	-1.74337	-3.08082	0.97319	55.16410	-18.56178

EXHIBIT 5.4.5 (Continued)
(Regression of Y on X with Observation Number 6 Excluded)

The SAS System

0:00 Saturday, Jan 1, 1994

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	119.02500	119.02500	30.764	0.0116
Error	3	11.60700	3.86900		
C Total	4	130.63200			

Root MSE	1.96698	R-square	0.9111
Dep Mean	21.76000	Adj R-sq	0.8815
C.V.	9.03942		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-23.090000	8.13387362	-2.839	0.0657
X	1	3.450000	0.62201286	5.547	0.0116

obs	Y	X	fits	residual	stdresid	tresid	hatvals	cooksd	dffits
1	16.3	11	14.86	1.44	1.15753	1.27051	0.6	1.00491	1.55605
2	16.8	12	18.31	-1.51	-0.91755	-0.88330	0.3	0.18041	-0.57825
3	20.1	13	21.76	-1.66	-0.94355	-0.91868	0.2	0.11129	-0.45934
4	27.3	14	25.21	2.09	1.26998	1.52494	0.3	0.34561	0.99831
5	28.3	15	28.66	-0.36	-0.28938	-0.23965	0.6	0.06281	-0.29351

5.4.3 Exhibit the following:

a $\hat{Y}(13)$

b $\hat{Y}_{(-6)}(13)$

c $\hat{\sigma}$

d $\hat{\sigma}_{(-6)}$

5.4.4 Use the formula in (5.2.3) and compute T_6. Compare this value with the "tresidual" value for observation 6 in Exhibit 5.4.5.

5.4.5 Use the formula in (5.3.3) and compute $h_{6,6}$. Compare this value with the $h_{i,i}$ value for observation 6 in Exhibit 5.4.5.

5.4.6 Study the diagnostic statistics given in Exhibit 5.4.5 and write a short report summarizing your findings.

Conversation 5.4

Investigator: Good afternoon. I'm somewhat confused about all the diagnostic procedures. Do you have some time to help me sort these out?

Statistician: Certainly. I'll be glad to do what I can. What seems to be the difficulty?

Investigator: I've read about fits, residuals, standardized residuals, studentized deleted residuals, Gaussian scores, hat values, leverages, DFFITS, Cook's distances, and yet I haven't seen confidence intervals or tests for these quantities. It's been said that if some of these quantities are large, then this means something, etc. I really don't understand how to interpret them.

Statistician: All of these quantities have been developed to help determine whether certain assumptions are satisfied. And to check these assumptions, you must make other assumptions. First, let's discuss **residuals**. For simplicity we'll consider straight line regression even though the concepts apply to multiple regression as well.

Residuals are "estimates" of the random errors e_i in the model

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Assumptions (A) or (B) underlie the mathematical theory of regression as we have been describing it and imply the following about the residuals e_i.

The e_i for $i = 1, \ldots, n$ form a simple random sample from a population of errors E, and this population is Gaussian with mean zero and standard deviation σ (which is unknown). **(5.4.7)**

If these assumptions are not satisfied at least approximately, then many of the results (such as confidence intervals on β_i, etc.) may not be correct. So it's important to determine whether these assumptions are at least approximately satisfied. We'd like to examine the population of errors $\{E\}$, but they aren't available to us because we only have a sample of size n, viz., e_i, from this population. But even e_i aren't observable. However, we can estimate the e_i, and these estimates are the residuals. So we use the residuals to check assumptions. It may be helpful to view the residuals $\hat{e}_i$ in two ways.

1 As estimates of e_i for $i = 1, \ldots, n$ where e_i is a simple random sample from the population of errors $\{E\}$. From this point of view, residuals are used to help determine whether the population of errors $\{E\}$ is, at least approximately, Gaussian with mean zero.

2 The ith residual $\hat{e}_i = y_i - \hat{Y}(x_i)$ is the difference between the observed y_i and the estimated value, namely $\hat{Y}(x_i)$; i.e., the ith residual is a measure of how far y_i is from its estimated value. From this point of view, residuals can help determine whether any of the sample Y values is an outlier; i.e., if any of the Y values do not seem to agree with the estimated regression function. So you can see why residuals are important!

Investigator: I think I see what you're saying. But we not only use residuals, we also use *standardized residuals.*

Statistician: That's correct. The residuals $\hat{e}_i$ typically have different standard deviations that depend on the x_i values and are therefore not directly comparable with one another. Thus, to eliminate this dependence on x_i, we standardize them so that the standardized residuals approximate a simple random sample of size n (approximate because they are correlated) from a Gaussian population with mean zero and standard deviation *one* if the statement in (5.4.7) is correct. Thus we can use Gaussian *scores* to see whether the standardized residuals appear to be a simple random sample of size n from a Gaussian population with mean zero and standard deviation one. If they do not appear to be a simple random sample from a Gaussian population with mean zero and standard deviation one, then we suspect that (5.4.7) is not correct. To help evaluate whether the population of errors $\{E\}$ is Gaussian with mean zero, we plot the standardized residuals against Gaussian scores and examine how close the points on this plot are to a line through the origin with slope one. If they are close, we conclude that (5.4.7) is close enough to being correct so that our statistical procedures are approximately valid. This was discussed in Section 4.5.

Investigator: I think I see why standardized residuals are useful, but can you give me some additional insight about studentized *deleted* residuals? What is so important about deleted?

Statistician: First, let's discuss **fits**, viz., the fitted values, which are the estimates of the sample y_i values based on the estimated regression function. Of course we know the value of y_i, and so we are not interested in estimating it, but if the ith fit, namely $\hat{Y}(x_i)$, is close to y_i for each $i = 1, 2, \ldots, n$, then we feel that $\hat{Y}(x)$ will be close to $Y(x)$ for other x values. In fact, as you know, residuals $\hat{e}_i$ are measures of how good the fits are because

$$\hat{e}_i = y_i - \hat{Y}(x_i)$$

If one of the residuals, say the ith one, has a magnitude that is substantially different from the others, this means that something could be wrong with either y_i or $\hat{Y}(x_i)$. To decide which is wrong, we compute the regression of Y on X after omitting (y_i, x_i), the ith sample observation. Then $\hat{Y}_{(-i)}(x_i)$ is another estimate of y_i, but this estimate is not influenced by y_i because it is not used in computing $\hat{Y}_{(-i)}(x_i)$. So

$$y_i - \hat{Y}_{(-i)}(x_i)$$

is a measure of how far the ith sample value y_i is from its estimate. We standardize this quantity by dividing it by its standard error, which is

$$SE(y_i - \hat{Y}_{(-i)}(x_i)) = \hat{\sigma}_{(-i)}/\sqrt{1 - h_{i,i}}$$

and the standardized quantity is T_i, the studentized deleted residual given in (5.2.3). So the studentized deleted residual is similar to the standardized residual, except the studentized deleted residual is computed without using y_i.

Investigator: This helps me see why studentized deleted residuals are useful.

Statistician: In addition, it is true that if (5.4.7) obtains, then T_i is distributed as student's t. Hence if T_i is larger in absolute value than, say 3, then it seems unlikely that (5.4.7) is correct. There are other statistics that can be used to check these assumptions, and one of them is Cook's distance. This is similar to studentized deleted residuals. Cook's distance for sample observation i, denoted by c_i, is a measure of how much the two sets of estimated regression coefficients differ when they are obtained by the following two methods: (1) by using all the sample data or (2) by using all the sample data except the ith observation. Another useful diagnostic measure is DFFITS. DFFITS_i is a measure of how much the two estimates of $Y(x_{i1}, \ldots, x_{ik})$, the Y value of sample item i, differ when one estimate is obtained by using all the sample data and the other is obtained by using all the sample data except y_i.

Investigator: Can you summarize the usefulness of hat values for me?

Statistician: Hat values are useful because they identify sample observations whose X values are substantially further from the center (mean) of the data than the remainder of the X values. Observations with large hat values are called *high leverage points*, and each such sample value should be investigated to see if point estimates and confidence intervals are changed substantially when it is included or excluded from the analysis.

Investigator: You use terms like "close enough," "approximately," "similar to," "previous judgment," and "knowledge," etc. I thought that statistical inference was a very objective procedure and that personal judgment shouldn't enter into making decisions.

Statistician: Scientists would, perhaps, like to make decisions that do not depend on their personal judgments, but that is not possible. Investigators make decisions based on assumptions, and if they want to check to see whether these assumptions are correct, then they must make other assumptions, etc. So someplace in this chain of events, they must make some assumptions. In regression we make certain assumptions (A), or (B), but we ought to use all available information, including our judgments and previous experiences, to examine these assumptions. The evaluation of assumptions is necessarily a subjective exercise using descriptive statistics. Thus, even though we don't have exact procedures available to check assumptions, the approximate procedures are useful and important. If, based on these procedures, we believe that assumptions (A) or (B) are approximately satisfied, then point estimates, confidence intervals, etc. are valid enough to be useful.

Investigator: Thanks for your time. Perhaps I'll come again soon.

5.5
Ill-Conditioning and Multicollinearity

The computation of parameters in a multiple linear regression model often involves a large number of arithmetical operations, and the results may be affected by rounding errors. Typically the computations are done using a calculator or a computer. This usually means that only a finite number of significant digits are kept in the computer's memory. For most desktop computers, this is about seven digits when using single precision arithmetic and about twelve digits when using double precision. As a simple illustration that dramatically shows the effect of rounding, consider the problem of calculating the quantity c where

$$c = 10^{10}\left(\sqrt{2} - \frac{2 \times 29 \times 37 \times 659}{1000000}\right)$$

If we carry out the calculations, rounding to the nearest seven significant digits at each step, we get

$$\begin{aligned}
\sqrt{2} &= 1.414214 \\
2 \times 29 &= 58 \\
58 \times 37 &= 2146 \\
2146 \times 659 &= 1414214 \\
\frac{2 \times 29 \times 37 \times 659}{1000000} &= 1.414214 \\
c = 10^{10}(1.414214 - 1.414214) &= 0
\end{aligned}$$

If we carry out the calculations, rounding to the nearest twelve significant digits at each step, we get

$$\begin{aligned}
\sqrt{2} &= 1.41421356237 \\
2 \times 29 &= 58 \\
58 \times 37 &= 2146 \\
2146 \times 659 &= 1414214 \\
\frac{2 \times 29 \times 37 \times 659}{1000000} &= 1.414214 \\
c = 10^{10}(1.41421356237 - 1.414214) &= 10^{10}(-0.00000043763) = -4376.3
\end{aligned}$$

Certainly in this problem there is a considerable loss of accuracy in the result when only seven significant digits are kept. Admittedly this will not be the case in every problem, but regression calculations are particularly prone to rounding errors. Example 5.5.1 illustrates this point.

EXAMPLE 5.5.1

Recall that the estimate of β in a multiple linear regression model may be written as

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

where X is the matrix of predictors (with the first column being all 1's in the case of a regression model with an intercept), and y is the column vector consisting of the corresponding values of the response variable. Suppose, for illustration, that

$$X = \begin{bmatrix} 1 & 2.001 \\ 1 & 1.998 \\ 1 & 2.003 \\ 1 & 1.999 \end{bmatrix} \qquad y = \begin{bmatrix} 3.01 \\ 2.99 \\ 2.94 \\ 2.99 \end{bmatrix}$$

We calculate $\hat{\beta}$ in two ways. First, we use exact arithmetic with no loss of accuracy at any stage. In parallel to this we compute $\hat{\beta}$ again, but this time we *round* all intermediate calculations to seven significant digits. The actual calculations are given in Table 5.5.1. Most statistical packages compute $\hat{\beta}$ by using special numerical techniques; they do not apply the formula $\hat{\beta} = (X^T X)^{-1} X^T y$ directly. However, we use this formula to illustrate how rounding errors can be troublesome, and even the special numerical techniques are not immune to rounding error problems.

TABLE 5.5.1

Illustration of the Effect of Rounding in Regression Calculations

Exact Calculations	Rounding to Seven Significant Digits
$X^T X = \begin{bmatrix} 4 & 8001/1000 \\ 8001/1000 & 16004015/1000000 \end{bmatrix}$	$X^T X = \begin{bmatrix} 4.0000 & 8.0010 \\ 8.0010 & 16.00402 \end{bmatrix}$
$X^T y = \begin{bmatrix} 1193/100 \\ 2386286/100000 \end{bmatrix}$	$X^T y = \begin{bmatrix} 11.930 \\ 23.86286 \end{bmatrix}$
$(X^T X)^{-1} = \begin{bmatrix} 16004015/59 & -8001000/59 \\ -8001000/59 & 4000000/59 \end{bmatrix}$	$(X^T X)^{-1} = \begin{bmatrix} 202582.5 & -101278.5 \\ -101278.5 & 50632.91 \end{bmatrix}$
$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} 115609/5900 \\ -490/59 \end{bmatrix}$	$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} 14.00000 \\ -7.000000 \end{bmatrix}$
$= \begin{bmatrix} 19.59475 \\ -8.305085 \end{bmatrix}$ (final result rounded to seven significant digits)	

Thus rounding to seven significant digits has resulted in inaccuracy in the results, even in the first significant digit for β_1. ■

Sometimes the errors due to inexact arithmetic are not very serious and can be safely ignored. However, this is not always so, and it is possible for rounding errors in regression analysis to result in serious errors in the parameter estimates, which can lead to erroneous conclusions or decisions. The seriousness of the errors due to rounding usually depends on a property of the X matrix called the **condition of X**. Some X matrices are **well-conditioned**, while others, like the one in Example 5.5.1, are said to be **ill-conditioned**. We say that the X matrix is ill-conditioned if small errors in the sample data values translate into large errors in the final results. *This may be so even if all the calculations are carried out exactly.* When the X matrix in a regression problem is ill-conditioned, rounding is likely to lead to substantial errors in the results. Many of the software packages for regression use double precision arithmetic and also employ numerical techniques that minimize the effects of rounding. If the X matrix is **ill-conditioned**, then some computer packages inform the user that this is the case.

We illustrate these concepts in Example 5.5.2.

EXAMPLE 5.5.2

Suppose we want to develop a function for predicting the weight Y using age X_1 and length X_2 for babies with ages ranging from 1 month to 12 months, and that a sample of size 12 was selected by first preselecting the ages and then randomly choosing one baby from each preselected age group. The length and weight of each chosen baby are recorded along with age. The data are displayed in Table 5.5.2. We

TABLE 5.5.2

Observation Number	Weight Y (pounds)	Age X_1 (months)	Length X_2 (inches)
1	9.2	1	20.4
2	9.8	2	20.9
3	9.1	3	22.1
4	9.6	4	21.7
5	11.7	5	22.9
6	10.7	6	24.2
7	12.7	7	24.9
8	13.0	8	26.1
9	13.4	9	26.9
10	14.7	10	27.6
11	14.4	11	28.1
12	15.2	12	29.2

suppose that assumptions (A) for regression are satisfied with

$$\mu_Y(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{5.5.1}$$

We estimate the parameters β_i using the formula (4.4.8) and carrying out all calculations exactly (by hand). The results are

$$X^T X = \begin{bmatrix} 12 & 78 & 295 \\ 78 & 650 & 2035.7 \\ 295 & 2035.7 & 7351.16 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 143.5 \\ 1018.5 \\ 3598.99 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} 63417951/236068 & 1357051/118034 & -824135/59017 \\ 1357051/118034 & 29723/59017 & -35460/59017 \\ -824135/59017 & -35460/59017 & 42900/59017 \end{bmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} 5743609/2360680 \\ 421987/1180340 \\ 34577/118034 \end{bmatrix} = \begin{bmatrix} 2.433031584 \\ 0.357513089 \\ 0.292941017 \end{bmatrix}$$

(final result rounded to nine decimals)

Now suppose that X_2 and Y values are measured slightly inaccurately, resulting in the data in Table 5.5.3. Note that the data values in Table 5.5.3 differ from the corresponding values in Table 5.5.2 by at most plus or minus 0.1.

TABLE 5.5.3

Observation Number	Weight Y (pounds)	Age X_1 (months)	Length X_2 (inches)
1	9.3	1	20.5
2	9.7	2	20.8
3	9.2	3	22.2
4	9.5	4	21.6
5	11.8	5	23.0
6	10.6	6	24.1
7	12.8	7	25.0
8	12.9	8	26.0
9	13.5	9	27.0
10	14.6	10	27.5
11	14.5	11	28.2
12	15.1	12	29.1

We again estimate the parameters β_i using formula (4.4.8) and carrying out all calculations exactly (by hand). The results are

$$X^TX = \begin{bmatrix} 12 & 78 & 295 \\ 78 & 650 & 2035.1 \\ 295 & 2035.1 & 7350.40 \end{bmatrix}$$

$$X^Ty = \begin{bmatrix} 143.5 \\ 1017.9 \\ 3598.42 \end{bmatrix}$$

$$(X^TX)^{-1} = \begin{bmatrix} 63612799/275428 & 1351165/137714 & -825305/68857 \\ 1351165/137714 & 29495/68857 & -35280/68857 \\ -825305/68857 & -35280/68857 & 42900/68857 \end{bmatrix}$$

$$\hat{\beta} = (X^TX)^{-1}X^Ty = \begin{bmatrix} -377089/2754280 \\ 335833/1377140 \\ 58877/137714 \end{bmatrix}$$

$$= \begin{bmatrix} -0.13691019 \\ 0.243862643 \\ 0.427530970 \end{bmatrix} \quad \text{(final result rounded to nine decimals)}$$

We see that small perturbations (changes or errors) in the sample values have resulted in substantial changes in the estimated parameter values, even though the calculations are exact. ■

The matrix X may be ill-conditioned because of one or both of the following reasons:

1 One or more columns of X consist of elements all of which are *very nearly* equal to zero.

2 One or more columns of X are *very nearly* obtainable as linear combinations of the remaining columns. In this case we say that **multicollinearity** exists among the columns of X. This is what happens in Example 5.5.2. You may verify that, in Example 5.5.2, X_1 and X_2 are nearly linearly related and that in fact

$$x_{i,2} \approx 19.2106 + 0.82657x_{i,1}$$

for each $i = 1, \ldots, 12$, causing the X matrix to be ill-conditioned.

Multicollinearity among the columns of X can occur due to one or more of the following reasons:

a In the population, one or more of the predictor variables $X_1, \ldots, X_k$ is nearly an exact linear combination of some or all of the remaining predictor variables. For instance, variable X_j may be very nearly an exact linear combination of the

remaining predictors so that we have

$$X_{I,j} \approx c_0 + c_1 X_{I,1} + \cdots + c_{j-1} X_{I,j-1} + c_{j+1} X_{I,j+1} + \cdots + c_k X_{I,k} \qquad (5.5.2)$$

for every $I = 1, \ldots, N$. In this situation we say that **multicollinearity** exists among the predictor variables in the population. If sample data are obtained by simple random sampling, then the sample values of the predictor variables also tend to exhibit a relation such as (5.5.2), resulting in multicollinearity among the columns of the X matrix, making it ill-conditioned. When the predictor variables exhibit multicollinearity in the population, then even if the data are obtained by sampling with preselected X values, we are unable to avoid an ill-conditioned X matrix because a relation such as (5.5.2) holds for every set of values $(X_{I,1}, \ldots, X_{I,k})$ that occurs in the population.

Consider, for instance, a study in which the predictor variables X_1, X_2, X_3, X_4, and X_5 are heights at ages 4, 5, 6, 7, and 8, respectively, of a population of children, and the response variable Y is height at age 9. In all likelihood, the height at age 8 can be predicted very well using the heights at ages 4, 5, 6, and 7 in a linear prediction function. This means that X_5 is very nearly a linear function of X_1, X_2, X_3, and X_4 in the population, and no matter how the sample is selected, the sample values of X_5 will also be very nearly a linear function of the sample values of X_1, X_2, X_3, and X_4.

b Data were obtained by sampling with preselected X values, but practical constraints such as cost, infeasibility of obtaining samples of the response variable at certain combinations of the predictors, etc., may have resulted in a choice of preselected values for the predictors leading to an ill-conditioned X matrix.

c The design of the study is bad. Here investigators could have selected values of the predictor variables in such a way that the X matrix would not be ill-conditioned, but they failed to take advantage of this opportunity.

Presence of *multicollinearity* among the columns of the X matrix has the following implications:

a Computations are very sensitive to rounding, and even if several significant digits are retained during various steps of the calculations, they often yield incorrect values for estimates of various parameters. This can perhaps be overcome by using double precision or multiple precision calculations.

b The results are highly sensitive to errors in the sample data. Even seemingly negligible errors in the measurements can lead to results that have no resemblence to the results that would be obtained if there were no errors in the data. Because practically all measurements are subject to errors, the resulting statistics cannot be taken seriously when the columns of the X matrix exhibit multicollinearity. The standard errors of the parameter estimates may reflect this situation by taking on values that are extremely large relative to the magnitude of the estimates.

c Based on the sample at hand, it is not possible to separate the influences of each of the predictors on the response. This is again related to the fact that the estimated regression coefficients tend to have large standard errors relative to their magnitudes. Whereas we may be able to find good prediction functions,

we have to choose arbitrarily from among several sets of nearly equally good prediction functions. Knowledge related to the field of application can often guide us in making a rational selection.

Variance Inflation Factors (VIF)

Several diagnostic procedures have been proposed in the literature for detecting the presence of approximate linear relationships among the columns of the X matrix. Associated with each predictor variable X_j is a number denoted by VIF_j, called the **variance inflation factor** for X_j, which is defined as

$$\text{VIF}_j = \frac{1}{1 - \hat{\rho}^2_{X_j(X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_k)}} \tag{5.5.3}$$

where $\hat{\rho}^2_{X_j(X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_k)}$ is the sample coefficient of determination of X_j on the remaining predictor variables (see (4.9.20)). If one or more of these variance inflation factors is large, we can conclude that *nearly linear relationships* exist among the columns of the X matrix. It has been suggested, as a rule of thumb, that values of VIF_j greater than 10.0 may be considered large enough for us to suspect serious multicollinearity. Note that $\hat{\rho}^2_{X_j(X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_k)}$ in (5.5.3) is not in general a valid estimate of $\rho^2_{X_j(X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_k)}$ (since no valid estimate is available if the X values are preselected), but it can be computed using the formula in (4.9.20).

What can or should an investigator do if the diagnostics reveal multicollinearity among the columns of X? Two cases need to be distinguished.

1 *The main objective of the study is to obtain a good prediction function for the response variable.* If the data were obtained by simple random sampling, then the existence of multicollinearity among the columns of X indicates that, even in the population, some of the predictor variables are nearly linear functions of other predictors and thus are redundant. Prediction functions based on only a subset of the predictors can be found that are nearly as good as the one based on *all* the predictors.

   If the data were obtained by sampling with preselected X values, then the preselected values of the predictor variables were not chosen appropriately; i.e., the sampling design was bad. In this case any prediction function based on the sample cannot be expected to predict the Y values very well for values of the predictor variables that are not similar to the ones in the sample. More data have to be collected using a better sampling design.

2 *The main objective is to estimate the parameters in the regression function or to assess the importance of each predictor in predicting the response variable Y.* There are two ways in which this situation can be handled:

   a If the multicollinearity among the columns of X is due to a bad sampling design (not preselecting the values of the predictors judiciously), then the investigator may be able to compensate for this mistake by obtaining an additional sample of suitable size by judiciously preselecting the values of the predictor variables at which to sample the response variable. The preselected values of the predictors should be chosen to make the X matrix

well-conditioned. A professional statistician's assistance can be invaluable here.

b If the multicollinearity among the columns of X is due to the existence of exact (or nearly exact) linear relationships among the predictor variables in the population, then realistically the only way to obtain useful unbiased or nearly unbiased estimates of the parameters is to collect additional data. In cases of serious multicollinearity, the sample size needed to obtain useful unbiased or nearly unbiased estimates of the parameters may be extremely large. Some authors have suggested the use of alternate approaches such as **ridge regression**, but we do not discuss them here. To find out more about these approaches, consult other texts [6].

When multicollinearity among the predictors is detected, it is often useful to understand the nature of the multicollinearity—i.e., understand which predictor variables are approximately linearly related. If you are interested in this topic, you may refer to other texts [1], [30].

5.6 Exercises

5.6.1 Consider Problem 4.12.4 where an investigator is studying a population of people who have lived in mountain isolation for several generations. She is interested in studying the relationship of Y, the height of males at age 18, to the following variables.

X_1 = length at birth
X_2 = mother's height at age 18
X_3 = father's height at age 18
X_4 = maternal grandmother's height at age 18
X_5 = maternal grandfather's height at age 18
X_6 = paternal grandmother's height at age 18
X_7 = paternal grandfather's height at age 18

All heights and lengths are in inches. A random sample of 20 males of age 18 or more was obtained from the study population, and the preceding information was recorded. The data for this problem are a modification (for illustrative purposes) of the data in Problem 4.12.4. The data appear in Table 5.6.1. For convenience, they are also stored in the file **table561.dat** on the data disk.

TABLE 5.6.1

Observation Number	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	67.2	19.7	60.5	70.3	65.7	69.3	65.7	67.3
2	69.1	19.6	64.9	70.4	62.6	69.6	64.6	66.4
3	67.0	19.4	65.4	65.8	66.2	68.8	64.0	69.4
4	72.4	19.4	63.4	71.9	60.7	68.0	64.9	67.1
5	63.6	19.7	65.1	65.1	65.5	65.5	61.8	70.9
6	72.7	19.6	65.2	71.1	63.5	66.2	67.3	68.6
7	68.5	19.8	64.3	67.9	62.4	71.4	63.4	69.4
8	69.7	19.7	65.3	68.8	61.5	66.0	62.4	67.7
9	68.4	19.7	64.5	68.7	63.9	68.8	62.3	68.8
10	70.4	19.9	63.4	70.3	65.9	69.0	63.7	65.1
11	67.5	18.9	63.3	70.4	63.7	68.2	66.2	68.5
12	73.3	18.3	63.1	65.2	65.4	66.6	61.7	64.0
13	70.0	20.3	64.9	68.8	65.2	70.2	62.4	67.0
14	69.8	19.7	63.5	70.3	63.1	64.4	65.1	67.0
15	63.6	19.9	62.0	65.5	64.1	67.7	62.1	66.5
16	64.3	19.6	63.5	65.2	63.9	70.0	64.2	64.5
17	68.5	21.3	66.1	65.4	64.8	68.4	66.4	70.8
18	70.5	20.1	64.8	70.2	65.3	65.5	63.7	66.9
19	68.1	20.2	62.6	68.6	63.7	69.8	66.7	68.0
20	66.1	19.2	62.2	67.3	63.6	70.9	63.6	66.7

Exhibit 5.6.1 contains a computer output from a regression analysis for this problem. Observe that variance inflation factors are given as part of the output. Other diagnostic statistics such as residuals, standardized residuals, fitted values, studentized deleted residuals, DFFITS, etc. are also given in the computer output.

a Are there any sample items that have high leverage values? If so, what are the sample item numbers?

b Are there any sample items that are outliers? If so, what are the sample item numbers?

c Are there any sample items that are influential observations? If so, what are the sample item numbers?

d Write a short summary discussing unusual sample items. Explain what an investigator should do about them.

e Is there any indication of multicollinearity among the predictor variables? If so, explain what an investigator should do about it.

EXHIBIT 5.6.1

MINITAB Output for Exercise 5.6.1

```
The regression equation is
Y = 5.2 - 0.77 X1 + 1.01 X2 + 0.635 X3 + 0.093 X4 - 0.134 X5
          + 0.210 X6 - 0.589 X7
```

Predictor	Coef	Stdev	t-ratio	p	VIF
Constant	5.15	56.80	0.09	0.929	
X1	-0.769	1.082	-0.71	0.491	1.4
X2	1.0113	0.4744	2.13	0.054	1.6
X3	0.6355	0.2934	2.17	0.051	1.6
X4	0.0926	0.3953	0.23	0.819	1.2
X5	-0.1343	0.2915	-0.46	0.653	1.2
X6	0.2104	0.3799	0.55	0.590	1.5
X7	-0.5891	0.3616	-1.63	0.129	1.6

s = 2.312 R-sq = 56.0% R-sq(adj) = 30.3%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	7	81.597	11.657	2.18	0.113
Error	12	64.148	5.346		
Total	19	145.746			

Unusual Observations

Obs.	x1	y	Fit	Stdev.Fit	Residual	St.Resid
12	18.3	73.300	68.711	1.786	4.589	3.12R
16	19.6	64.300	67.751	1.623	-3.451	-2.10R

R denotes an obs. with a large st. resid.