CHAPTER 6

Diagnostics for Leverage and Influence

Importance of Detecting Influential Observations

- **Leverage Point:**
  - unusual x-value;
  - very little effect on regression coefficients.

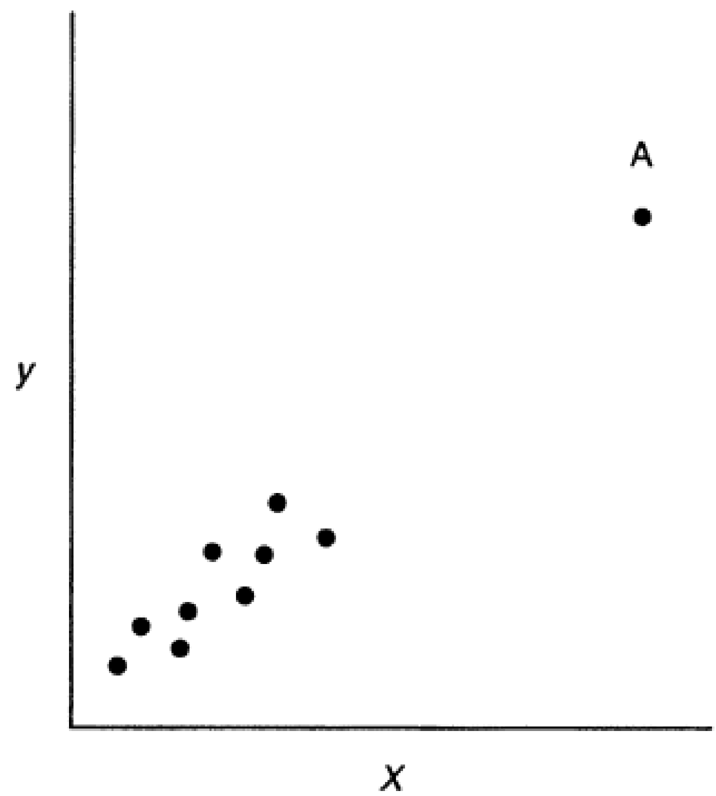

Figure 6.1 An example of a leverage point.

Importance of Detecting Influential Observations

- **Influence Point**: **unusual in y and x;**

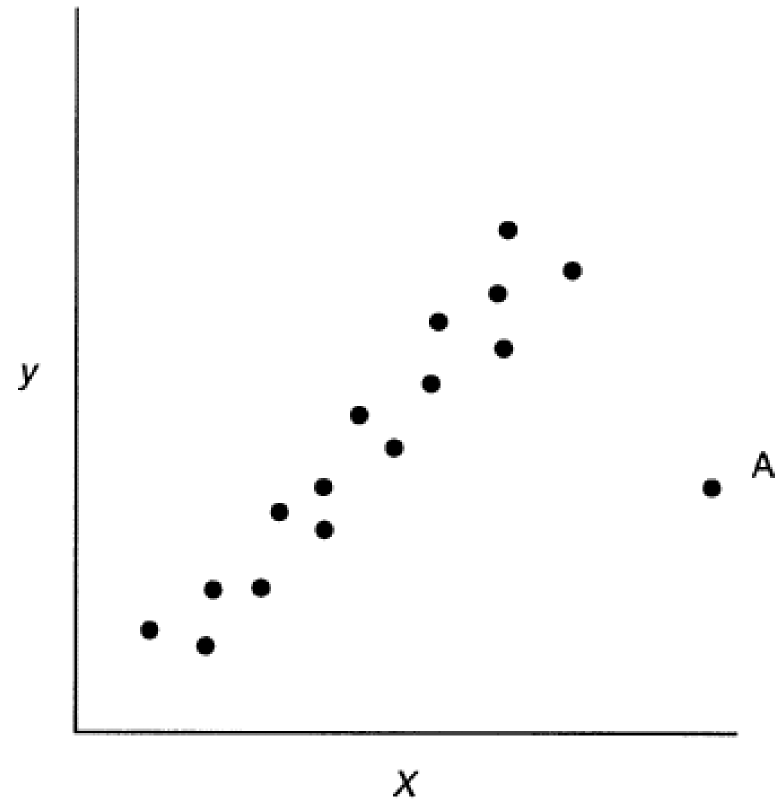

Figure 6.2 An example of an influential observation.

Leverage

- The **hat matrix** is:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

- The diagonal elements of the hat matrix h_{ii} – standardized measure of the distance of the *i*th observation from the center of the x.

Leverage

- The average size of the hat diagonal is p/n.
- Traditionally, any $h_{ii} > 2p/n$ indicates a **leverage** point.
- Appropriate for large *n;* otherwise consider large as compared to other values
- An observation with large h_{ii} and a large residual is likely to be **influential**

Treatment of Influential Observations

- Discard if:
  - there is an error in recording a measured value;
  - the sample point is invalid; or,
  - the observation is not part of the population that was intended to be sampled
- Do not discard if:
  - the influential point is a valid observation

Treatment of Influential Observations

- Robust estimation techniques
  - These techniques offer an alternative to deleting an influential observation
  - Observations are retained but ***downweighted*** in proportion to residual magnitude or influence.

Measure of Influence

- Reading materials
- Optional

Example 6-1. The Delivery Time Data

The model of interest is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

TABLE 3.2 Delivery Time Data for Example 3.1

Observation Number	Delivery Time (Minutes) y	Number of Cases x_1	Distance (Feet) x_2
1	16.68	7	560
2	11.50	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330
7	8.00	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.50	5	605
11	40.33	16	688
12	21.00	10	215
13	13.50	4	255
14	19.75	6	462
15	24.00	9	448
16	29.00	10	776
17	15.35	6	200
18	19.00	7	132
19	9.50	3	36
20	35.10	17	770
21	17.90	10	140
22	52.32	26	810
23	18.75	9	450
24	19.83	8	635
25	10.75	4	150

Example 6-1 Excel Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.980
R Square	0.960
Adjusted R Square	0.956
Standard Error	3.259
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	5550.811	2775.405	261.235	4.68742E-16
Residual	22	233.732	10.624		
Total	24	5784.543			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	2.341	1.097	2.135	0.044	0.067	4.616	-0.750	5.433
Number of Cases, x_1	1.616	0.171	9.464	3.25E-09	1.262	1.970	1.135	2.097
Distance, x_2 (ft)	0.014	0.004	3.981	0.001	0.007	0.022	0.004	0.025

TABLE 3.3 Observations, Fitted Values, and Residuals for Example 3.1

Observation Number	y_i	$\hat{y}_i$	$e_i = y_i - \bar{y}_i$
1	16.68	21.7081	−5.0281
2	11.50	10.3536	1.1464
3	12.03	12.0798	−0.0498
4	14.88	9.9556	4.9244
5	13.75	14.1944	−0.4444
6	18.11	18.3996	−0.2896
7	8.00	7.1554	0.8446
8	17.83	16.6734	1.1566
9	79.24	71.8203	7.4197
10	21.50	19.1236	2.3764
11	40.33	38.0925	2.2375
12	21.00	21.5930	−0.5930
13	13.50	12.4730	1.0270
14	19.75	18.6825	1.0675
15	24.00	23.3288	0.6712
16	29.00	29.6629	−0.6629
17	15.35	14.9136	0.4364
18	19.00	15.5514	3.4486
19	9.50	7.7068	1.7932
20	35.10	40.8880	−5.7880
21	17.90	20.5142	−2.6142
22	52.32	56.0065	−3.6865
23	18.75	23.3576	−4.6076
24	19.83	24.4028	−4.5728
25	10.75	10.9626	−0.2126

TABLE 6.1 Statistics for Detecting Influential Observations for the Soft Drink Delivery Time Data

Observation i	(a) h_{ii}	(b) D_i	(c) $DFFITS_i$	(d) Intercept $DFBETAS_{0,i}$	(e) Cases $DFBETAS_{1,i}$	(f) Distance $DFBETAS_{2,i}$	(g) $COVRATIO_i$
1	0.10180	0.10009	−0.5709	−0.1873	0.4113	−0.4349	0.8711
2	0.07070	0.00338	0.0986	0.0898	−0.0478	0.0144	1.2149
3	0.09874	0.00001	−0.0052	−0.0035	0.0039	−0.0028	1.2757
4	0.08538	0.07766	0.5008	0.4520	0.0883	−0.2734	0.8760
5	0.07501	0.00054	−0.0395	−0.0317	−0.0133	0.0242	1.2396
6	0.04287	0.00012	−0.0188	−0.0147	0.0018	0.0011	1.1999
7	0.08180	0.00217	0.0790	0.0781	−0.0223	−0.0110	1.2398
8	0.06373	0.00305	0.0938	0.0712	0.0334	−0.0538	1.2056
9	0.49829	3.41835	4.2961	−2.5757	0.9287	1.5076	0.3422
10	0.19630	0.05385	0.3987	0.1079	−0.3382	0.3413	1.3054
11	0.08613	0.01620	0.2180	−0.0343	0.0925	−0.0027	1.1717
12	0.11366	0.00160	−0.0677	−0.0303	−0.0487	0.0540	1.2906
13	0.06113	0.00229	0.0813	0.0724	−0.0356	0.0113	1.2070
14	0.07824	0.00329	0.0974	0.0495	−0.0671	0.0618	1.2277
15	0.04111	0.00063	0.0426	0.0223	−0.0048	0.0068	1.1918
16	0.16594	0.00329	−0.0972	−0.0027	0.0644	−0.0842	1.3692
17	0.05943	0.00040	0.0339	0.0289	0.0065	−0.0157	1.2192
18	0.09626	0.04398	0.3653	0.2486	0.1897	−0.2724	1.0692
19	0.09645	0.01192	0.1862	0.1726	0.0236	−0.0990	1.2153
20	0.10169	0.13246	−0.6718	0.1680	−0.2150	−0.0929	0.7598
21	0.16528	0.05086	−0.3885	−0.1619	−0.2972	0.3364	1.2377
22	0.39158	0.45106	−1.1950	0.3986	−1.0254	0.5731	1.3981
23	0.04126	0.02990	−0.3075	−0.1599	0.0373	−0.0527	0.8897
24	0.12061	0.10232	−0.5711	−0.1197	0.4046	−0.4654	0.9476
25	0.06664	0.00011	−0.0176	−0.0168	0.0008	0.0056	1.2311

Example 6.1 The Delivery Time Data

- Examine Table 6.1. If some possibly influential points are removed here is what happens to the coefficient estimates and model statistics:

Run	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	MS_{Res}	R^2
9 and 22 in	2.341	1.616	0.014	10.624	0.9596
9 out	4.447	1.498	0.010	5.905	0.9487
22 out	1.916	1.786	0.012	10.066	0.9564
9 and 22 out	4.643	1.456	0.011	6.163	0.9072

Measures of Influence

- The influence measures discussed here are those that measure the effect of deleting the *i*th observation.
  1. Cook's D_i, which measures the effect on $\hat{\beta}$
  2. $DFBETAS_{j(i)}$, which measures the effect on $\hat{\beta}_j$
  3. $DFFITS_i$, which measures the effect on $\hat{Y}_i$
  4. $COVRATIO_i$, which measures the effect on the variance-covariance matrix of the parameter estimates.

Measures of Influence: Cook's D

$$D_i(X'X, pMS_{\text{Res}}) = D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{pMS_{\text{Res}}}$$

$$= \frac{r_i^2}{p} \frac{Var(\hat{y}_i)}{Var(e_i)} = \frac{r_i^2}{p} \frac{h_{ii}}{(1-h_{ii})}$$

- What contributes to D_i:
  - How well the model fits the *i*th observation, y_i
  - How far that point is from the remaining dataset
- Large values of D_i indicate an influential point, usually if $D_i > 1$.

Measures of Influence: Cook's D

- To interpret Cook's distance measure:
  - Relate D_i to the F(p, n-p) distribution and compute the percentile value
  - If percentile less than 20 percent i^{th} case has little influence
  - If percentile near 50 percent than i^{th} case has a major influence

TABLE 6.1 Statistics for Detecting Influential Observations for the Soft Drink Delivery Time Data

Observation i	(a) h_{ii}	(b) D_i	(c) $DFFITS_i$	(d) Intercept $DFBETAS_{0,i}$	(e) Cases $DFBETAS_{1,i}$	(f) Distance $DFBETAS_{2,i}$	(g) $COVRATIO_i$
1	0.10180	0.10009	−0.5709	−0.1873	0.4113	−0.4349	0.8711
2	0.07070	0.00338	0.0986	0.0898	−0.0478	0.0144	1.2149
3	0.09874	0.00001	−0.0052	−0.0035	0.0039	−0.0028	1.2757
4	0.08538	0.07766	0.5008	0.4520	0.0883	−0.2734	0.8760
5	0.07501	0.00054	−0.0395	−0.0317	−0.0133	0.0242	1.2396
6	0.04287	0.00012	−0.0188	−0.0147	0.0018	0.0011	1.1999
7	0.08180	0.00217	0.0790	0.0781	−0.0223	−0.0110	1.2398
8	0.06373	0.00305	0.0938	0.0712	0.0334	−0.0538	1.2056
9	0.49829	3.41835	4.2961	−2.5757	0.9287	1.5076	0.3422
10	0.19630	0.05385	0.3987	0.1079	−0.3382	0.3413	1.3054
11	0.08613	0.01620	0.2180	−0.0343	0.0925	−0.0027	1.1717
12	0.11366	0.00160	−0.0677	−0.0303	−0.0487	0.0540	1.2906
13	0.06113	0.00229	0.0813	0.0724	−0.0356	0.0113	1.2070
14	0.07824	0.00329	0.0974	0.0495	−0.0671	0.0618	1.2277
15	0.04111	0.00063	0.0426	0.0223	−0.0048	0.0068	1.1918
16	0.16594	0.00329	−0.0972	−0.0027	0.0644	−0.0842	1.3692
17	0.05943	0.00040	0.0339	0.0289	0.0065	−0.0157	1.2192
18	0.09626	0.04398	0.3653	0.2486	0.1897	−0.2724	1.0692
19	0.09645	0.01192	0.1862	0.1726	0.0236	−0.0990	1.2153
20	0.10169	0.13246	−0.6718	0.1680	−0.2150	−0.0929	0.7598
21	0.16528	0.05086	−0.3885	−0.1619	−0.2972	0.3364	1.2377
22	0.39158	0.45106	−1.1950	0.3986	−1.0254	0.5731	1.3981
23	0.04126	0.02990	−0.3075	−0.1599	0.0373	−0.0527	0.8897
24	0.12061	0.10232	−0.5711	−0.1197	0.4046	−0.4654	0.9476
25	0.06664	0.00011	−0.0176	−0.0168	0.0008	0.0056	1.2311

Measures of Influence: DFFITS and DFBETAS

DFBETAS – measures how much the regression coefficient changes in standard deviation units if the *i*th observation is removed

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S_{(i)}^2 C_{jj}}}$$

where $\hat{\beta}_{j(i)}$ is an estimate of the *j*th coefficient when the *i*th observation is removed

- Large DFBETAS indicates *i*th observation has considerable influence
- In general, $|DFBETAS_{j,i}| > 2/\sqrt{n}$

Measures of Influence: DFFITS and DFBETAS

DFFITS – measures the influence of the *i*th observation on the fitted value, again in standard deviation units.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S_{(i)}^2 h_{ii}}}$$

- Cutoff: If $|DFFITS_i| > 2\sqrt{p/n}$, the point is most likely influential
- For small and medium size data sets consider a case influential if *DFFITS* greater than 1

TABLE 6.1 Statistics for Detecting Influential Observations for the Soft Drink Delivery Time Data

Observation i	(a) h_{ii}	(b) D_i	(c) $DFFITS_i$	(d) Intercept $DFBETAS_{0,i}$	(e) Cases $DFBETAS_{1,i}$	(f) Distance $DFBETAS_{2,i}$	(g) $COVRATIO_i$
1	0.10180	0.10009	−0.5709	−0.1873	0.4113	−0.4349	0.8711
2	0.07070	0.00338	0.0986	0.0898	−0.0478	0.0144	1.2149
3	0.09874	0.00001	−0.0052	−0.0035	0.0039	−0.0028	1.2757
4	0.08538	0.07766	0.5008	0.4520	0.0883	−0.2734	0.8760
5	0.07501	0.00054	−0.0395	−0.0317	−0.0133	0.0242	1.2396
6	0.04287	0.00012	−0.0188	−0.0147	0.0018	0.0011	1.1999
7	0.08180	0.00217	0.0790	0.0781	−0.0223	−0.0110	1.2398
8	0.06373	0.00305	0.0938	0.0712	0.0334	−0.0538	1.2056
9	0.49829	3.41835	4.2961	−2.5757	0.9287	1.5076	0.3422
10	0.19630	0.05385	0.3987	0.1079	−0.3382	0.3413	1.3054
11	0.08613	0.01620	0.2180	−0.0343	0.0925	−0.0027	1.1717
12	0.11366	0.00160	−0.0677	−0.0303	−0.0487	0.0540	1.2906
13	0.06113	0.00229	0.0813	0.0724	−0.0356	0.0113	1.2070
14	0.07824	0.00329	0.0974	0.0495	−0.0671	0.0618	1.2277
15	0.04111	0.00063	0.0426	0.0223	−0.0048	0.0068	1.1918
16	0.16594	0.00329	−0.0972	−0.0027	0.0644	−0.0842	1.3692
17	0.05943	0.00040	0.0339	0.0289	0.0065	−0.0157	1.2192
18	0.09626	0.04398	0.3653	0.2486	0.1897	−0.2724	1.0692
19	0.09645	0.01192	0.1862	0.1726	0.0236	−0.0990	1.2153
20	0.10169	0.13246	−0.6718	0.1680	−0.2150	−0.0929	0.7598
21	0.16528	0.05086	−0.3885	−0.1619	−0.2972	0.3364	1.2377
22	0.39158	0.45106	−1.1950	0.3986	−1.0254	0.5731	1.3981
23	0.04126	0.02990	−0.3075	−0.1599	0.0373	−0.0527	0.8897
24	0.12061	0.10232	−0.5711	−0.1197	0.4046	−0.4654	0.9476
25	0.06664	0.00011	−0.0176	−0.0168	0.0008	0.0056	1.2311

A Measure of Model Performance

- Information about the overall precision of estimation can be obtained through another statistic, COVRATIO_i

$$COVRATIO_i = \frac{|(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1} S^2_{(i)}|}{|(\mathbf{X}'\mathbf{X})^{-1} MS_{\text{Res}}|}$$

$$= \frac{(S^2_{(i)})^p}{MS^p_{\text{Res}}}\left(\frac{1}{1-h_{ii}}\right)$$

A Measure of Model Performance

Cutoffs and Interpretation

- If $COVRATIO_i > 1$, the *i*th observation improves the precision.
- If $COVRATIO_i < 1$, *i*th observation can degrade the precision.
- Cutoffs: $COVRATIO_i > 1 + 3p/n$

  or $COVRATIO_i < 1 - 3p/n$; (the lower limit is really only good if $n > 3p$).

Example 6.4 The Delivery Time Data

Column g of Table 6.1 contains the values of $COVRATIO_i$ for the soft drink delivery time data. The formal recommended cutoff for $COVRATIO_i$ is $1 \pm 3p/n = 1 \pm 3(3)/25$, or 0.64 and 1.36. Note that the values of $COVRATIO_9$ and $COVRATIO_{22}$ exceed these limits, indicating that these points are influential. Since $COVRATIO_9 < 1$, this observation degrades precision of estimation, while since $COVRATIO_{22} > 1$, this observation tends to improve the precision. However, point 22 barely exceeds its cutoff, so the influence of this observation, from a practical viewpoint, is fairly small. Point 9 is much more clearly influential.

TABLE 6.1 Statistics for Detecting Influential Observations for the Soft Drink Delivery Time Data

Observation i	(a) h_{ii}	(b) D_i	(c) $DFFITS_i$	(d) Intercept $DFBETAS_{0,i}$	(e) Cases $DFBETAS_{1,i}$	(f) Distance $DFBETAS_{2,i}$	(g) $COVRATIO_i$
1	0.10180	0.10009	−0.5709	−0.1873	0.4113	−0.4349	0.8711
2	0.07070	0.00338	0.0986	0.0898	−0.0478	0.0144	1.2149
3	0.09874	0.00001	−0.0052	−0.0035	0.0039	−0.0028	1.2757
4	0.08538	0.07766	0.5008	0.4520	0.0883	−0.2734	0.8760
5	0.07501	0.00054	−0.0395	−0.0317	−0.0133	0.0242	1.2396
6	0.04287	0.00012	−0.0188	−0.0147	0.0018	0.0011	1.1999
7	0.08180	0.00217	0.0790	0.0781	−0.0223	−0.0110	1.2398
8	0.06373	0.00305	0.0938	0.0712	0.0334	−0.0538	1.2056
9	0.49829	3.41835	4.2961	−2.5757	0.9287	1.5076	0.3422
10	0.19630	0.05385	0.3987	0.1079	−0.3382	0.3413	1.3054
11	0.08613	0.01620	0.2180	−0.0343	0.0925	−0.0027	1.1717
12	0.11366	0.00160	−0.0677	−0.0303	−0.0487	0.0540	1.2906
13	0.06113	0.00229	0.0813	0.0724	−0.0356	0.0113	1.2070
14	0.07824	0.00329	0.0974	0.0495	−0.0671	0.0618	1.2277
15	0.04111	0.00063	0.0426	0.0223	−0.0048	0.0068	1.1918
16	0.16594	0.00329	−0.0972	−0.0027	0.0644	−0.0842	1.3692
17	0.05943	0.00040	0.0339	0.0289	0.0065	−0.0157	1.2192
18	0.09626	0.04398	0.3653	0.2486	0.1897	−0.2724	1.0692
19	0.09645	0.01192	0.1862	0.1726	0.0236	−0.0990	1.2153
20	0.10169	0.13246	−0.6718	0.1680	−0.2150	−0.0929	0.7598
21	0.16528	0.05086	−0.3885	−0.1619	−0.2972	0.3364	1.2377
22	0.39158	0.45106	−1.1950	0.3986	−1.0254	0.5731	1.3981
23	0.04126	0.02990	−0.3075	−0.1599	0.0373	−0.0527	0.8897
24	0.12061	0.10232	−0.5711	−0.1197	0.4046	−0.4654	0.9476
25	0.06664	0.00011	−0.0176	−0.0168	0.0008	0.0056	1.2311

R code

- influence.measures(model1)