Chapter 6

Diagnostics for Leverage and Influence

In Chapter 5, we focused considerable attention on studying residuals in order to shed light on possible violation of assumptions. Among the techniques presented was the so-called outlier analysis, which is designed to highlight suspect data points. Outliers are detected for further checking by the analyst, the lab technician, recorder, or whomever is involved in the data-taking process. A natural concern is that an erroneous observation will exert an undue amount of influence on the regression results – influence that is counterproductive. The influence on regression statistics from such an observation is the "fallout" that results from the regression being "pulled" toward the errant measured response. The analyst needs to be able to identify these observations and assess the extent of the influence.

6.1 Source of the influence

First it may be an outlier as described previously. Recall that the Studentized residual, given by

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1 - h_{ii})}}$$

is a natural diagnostic used to detect data points that exert influence created by errors in **the y-direction**. However, all high influence observations are not due to errors in the y-direction. An influential observation may be a legitimate and very important part of the data set. Influence can occur when a single observation is extreme in **the x-direction**; i.e., it is a proper observation and does not necessarily represent evidence of a model fallacy. There is a clear danger involved with allowing one piece of information to totally dictate.

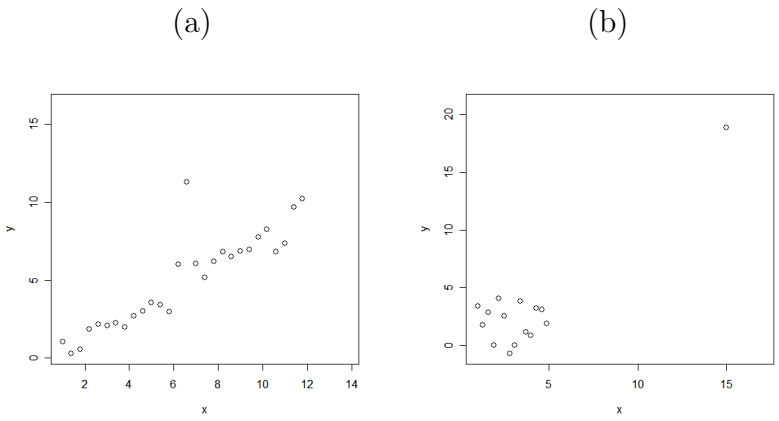

Figure 6.1: Examples of the influence. (a) Single observation with error in y-direction; (b) Single influential observation remote from center.

From now, we focus on diagnostic of individual data points which exert disproportionate influence on the regression.

6.2 Diagnostics: Residuals and Hat Matrix

We begin by recalling the $\mathbf{X}$ matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

and the vector of responses

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

where the ith data points is given by $[\mathbf{x}_i^T, y_i]$. Now, recall that the Studentized residual, given by

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1 - h_{ii})}}$$

is a natural diagnostic used to detect data points that exert influence created by errors in the y-direction.

As observed above, remote points potentially have disproportionate impact on the parameter estimates, standard errors, predicted values, and model summary statistics. The hat matrix

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

plays an important role in identifying influential observations. We usually focus attention on the diagonal elements h_{ii} of the hat matrix $\mathbf{H}$ which can be written as

$$h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i, \quad i = 1, \ldots, n,$$

where $\mathbf{x}_i^T$ is the ith row of the $\mathbf{X}$ matrix. For a simple linear regression, it can be shown that

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

and, in general, it can be shown that

$$h_{ii} = \frac{1}{n} + \mathbf{x}_{c,i}^T(\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{x}_{c,i}$$

where $\mathbf{X}_c$ is the centered matrix of $\mathbf{X}$ and $\mathbf{x}_{c,i}^T = (x_{i1} - \bar{x}_1, \ldots, x_{ik} - \bar{x}_k)^T$ is the ith row of the $\mathbf{X}_c$ matrix. The hat matrix diagonal is a standardized measure of the distance of the ith observation from the center of the x space. Thus, large hat diagonals reveal observations that are potentially influential because they are remote in x space from the rest of the sample.

The hat diagonal h_{ii} also shows how y_i affects $\hat{y}_i$. For example, if $h_i \approx 1$ then $\hat{y}_i \approx y_i$ since $\text{Var}(e_i) = (1 - h_{ii})\sigma^2 \approx 0$ and $y_i = \hat{y}_i + e_i$. That is, the predicted value, $\hat{y}_i$, will be close to the actual value, y_i, no matter what values of the rest of the data take. Notice that $\frac{1}{n} \leq h_{ii} \leq 1$. Also, note that h_{ii} depends only on the x's.

Leverage points

A leverage point is a point whose x-value is distant from the other x-values.

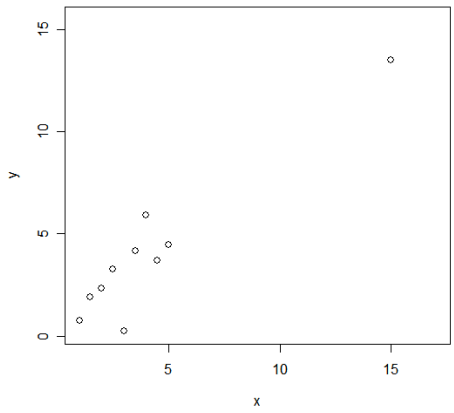

Figure 6.2: An example of a leverage point.

Rule for identifying high leverage points

A popular numerical rule is to identify x_i as a leverage point in a multiple linear regression model with k predictors if

$$h_{ii} > 2\bar{h} = \frac{2p}{n},$$

where $\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{1}{n}\text{tr}(\mathbf{H})$ and $p = k + 1$.

6.3 Influence points

Data points whose deletion causes substantial changes in the fitted model.

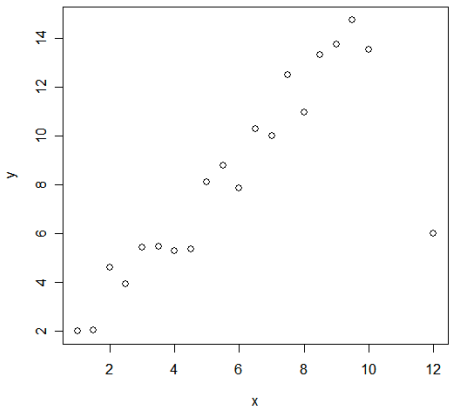

Figure 6.3: An example of an influence point.

How to detect influence points?

Measure for detecting influence points

Cook (1977) has suggested a way to measure the influence points given by

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{pMS_{Res}} = \frac{\sum_{j=1}^{n} (\hat{y}_{j(i)} - \hat{y}_j)^2}{pMS_{Res}},$$

where $\hat{\boldsymbol{\beta}}_{(i)}$ is a vector of estimates obtained by deleting the ith point and $\hat{y}_{j(i)}$ is the jth fitted value based on the fit obtained when the ith case has been deleted from the fit. It can be shown that

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}},$$

where r_i is the studentized residual and h_{ii} is the ith leverage value. Thus, Cook's distance is the product of the square of the ith stuentized residual and $h_{ii}/(1-h_{ii})$. The first quantity measures the extent to which the ith case is outlying and the second quantity measures the leverage of the ith case. Hence a large value of D_i may be due to a large value of r_i, a large value of h_{ii} or both.

We usually consider points whose $D_i > F_{.5,p,n-p}$ to be influential. One of rules of thumb for cut-off is that the points whose $D_i > 1$ is considered to be influential since $F_{.5,p,n-p} \approx 1$. There are various computational cut-off for Cook's distance.

Other measures for detecting influence points

Read sections 6.4 and 6.5 on pp.195–198 in textbook.

- DFBETAS

  DFBETAS measures how much the regression coefficient $\hat{\beta}_j$ changes if the ith observation were deleted.

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S^2_{(i)} C_{jj}}}, \quad j = 1, \ldots, p,$$

  where $S^2_{(i)}$ is another unbiased estimate of σ^2, C_{jj} is the jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$ and $\hat{\beta}_{j(i)}$ is the jth regression coefficient computed without use of the ith observation. A large value of $DFBETAS_{j,i}$ indicates that observation i has considerable influence on the jth regression coefficient. Belsley, Kuh and Welsch (1980) suggest a cut-off of $2/\sqrt{n}$ for $DFBETAS_{j,i}$; that is, if $|DFBETAS_{j,i}| > 2/\sqrt{n}$, then the ith observation warrants examination.

- DFFITS

  In similar sense to DFBETAS, DFFITS measures the effect on the fitted value when the ith observation were deleted.

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

  where $\hat{y}_{(i)}$ is the fitted value of y_i obtained without the use of the ith observation. Belsley, Kuh and Welsch (1980) suggest a cut-off of $2\sqrt{p/n}$ for $DFFITS_i$; that is, if $|DFFITS_i| > 2\sqrt{p/n}$, then the ith observation warrants attention.

- COVRAIO

  COVRATIO measures overall precision of estimation.

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

  where $\mathbf{X}_{(i)}$ is the original $\mathbf{X}$ matrix with the ith observation removed.

  If $COVRATIO_i > 1$ the ith observation improves the precision of estimation, while if $COVARTIO_i < 1$, inclusion of the ith point degrades precision.

  Belsley, Kuh and Welsch (1980) suggest for large samples that if $|COVARTIO_i - 1| > 3p/n$, then the ith point should be considered influential.

  Note that the generalized variance of $\hat{\boldsymbol{\beta}}$ is defined as

$$|\mathrm{Var}(\hat{\boldsymbol{\beta}})| = |\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}|.$$

Remark that the suggested cut-off values for all measures for detecting the influence points are various and it is difficult to determine the cut-off values.

Notes:

1. An influence point may not be an outlier.
2. An outlier may not be an influence point.
3. Leverage point may not be influence point.

Examples

1. Rocket propellant data.

   SAS Program:

```
proc reg data=rocket;
model Strength = Age / influence;
plot cookd.*obs.;    /* plot of Cook's D against observation number */
output out=newout p=pred r=resid student=student cookd=cookd;
run; quit;
proc print data=newout; var cookd; run;
```

   Output:

Output Statistics

Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	DFBETAS Age
1	106.7583	1.1526	0.0541	1.0197	0.2757	0.0623	0.0761
2	-67.2746	-0.7488	0.1475	1.2324	-0.3115	0.1331	-0.2533
3	-14.5936	-0.1536	0.0760	1.2099	-0.0441	-0.0399	0.0258
4	65.0887	0.6890	0.0620	1.1311	0.1771	0.0094	0.0778
5	-215.9776	-2.7882	0.1059	0.5904	-0.9594	-0.9296	0.6969
6	-213.6041	-2.6856	0.0787	0.5999	-0.7851	0.1100	-0.4742
7	48.5638	0.5379	0.1523	1.2786	0.2280	-0.0997	0.1868
8	40.0616	0.4437	0.1566	1.2991	0.1912	0.1904	-0.1578
9	8.7296	0.0921	0.0811	1.2188	0.0274	0.0252	-0.0169
10	37.5671	0.3926	0.0550	1.1652	0.0947	0.0690	-0.0287
11	20.3743	0.2117	0.0501	1.1741	0.0486	0.0257	-0.0024
12	-88.9464	-0.9939	0.1335	1.1556	-0.3901	-0.3857	0.3085
13	80.8174	0.9204	0.1724	1.2291	0.4201	-0.1992	0.3540
14	71.1752	0.7554	0.0618	1.1186	0.1939	0.1588	-0.0847
15	-45.1434	-0.4893	0.1174	1.2352	-0.1785	0.0615	-0.1352
16	94.4423	1.0198	0.0694	1.0699	0.2786	-0.0138	0.1474

```
17    9.4992    0.1012    0.0990    1.2428    0.0336    0.0322   -0.0236
18   37.0975    0.3867    0.0507    1.1604    0.0893    0.0522   -0.0103
19  100.6848    1.1585    0.1667    1.1557    0.5181    0.5168   -0.4335
20  -75.3202   -0.8232    0.1098    1.1647   -0.2892    0.0916   -0.2134
```

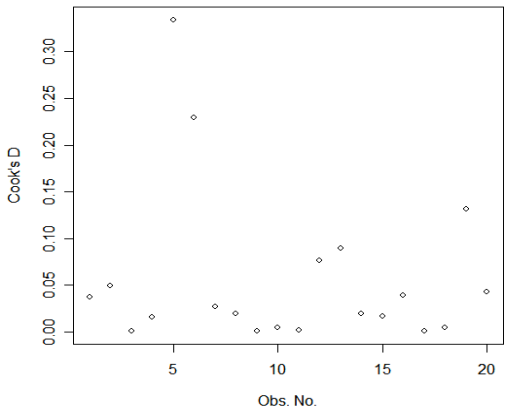

Figure 6.4: A plot of Cook's distance against observation number.

No leverage point is detected because all $h_{ii} < \frac{4}{20} = .2$. No serious influential point is detected since $D_i < F_{.5,2,18} = .72$.

2. Delivery time data.

SAS Program:

```
proc reg data=delivery;
model Time = Cases Distance / influence;
plot cookd.*obs.;
output out=newout p=pred r=resid student=student cookd=cookd;
run; quit;
proc print data=newout; var cookd; run;
```

Output:

```
                                Output Statistics

                           Hat Diag      Cov          ----------DFBETAS----------
 Obs  Residual  RStudent        H    Ratio   DFFITS Intercept    Cases Distance
```

1	-5.0281	-1.6956	0.1018	0.8711	-0.5709	-0.1873	0.4113	-0.4349
2	1.1464	0.3575	0.0707	1.2149	0.0986	0.0898	-0.0478	0.0144
3	-0.0498	-0.0157	0.0987	1.2757	-0.0052	-0.0035	0.0039	-0.0028
4	4.9244	1.6392	0.0854	0.8760	0.5008	0.4520	0.0883	-0.2734
5	-0.4444	-0.1386	0.0750	1.2396	-0.0395	-0.0317	-0.0133	0.0242
6	-0.2896	-0.0887	0.0429	1.1999	-0.0188	-0.0147	0.0018	0.0011
7	0.8446	0.2646	0.0818	1.2398	0.0790	0.0781	-0.0223	-0.0110
8	1.1566	0.3594	0.0637	1.2056	0.0938	0.0712	0.0334	-0.0538
9	7.4197	4.3108	0.4983	0.3422	4.2961	-2.5757	0.9287	1.5076
10	2.3764	0.8068	0.1963	1.3054	0.3987	0.1079	-0.3382	0.3413
11	2.2375	0.7099	0.0861	1.1717	0.2180	-0.0343	0.0925	-0.0027
12	-0.5930	-0.1890	0.1137	1.2906	-0.0677	-0.0303	-0.0487	0.0540
13	1.0270	0.3185	0.0611	1.2070	0.0813	0.0724	-0.0356	0.0113
14	1.0675	0.3342	0.0782	1.2277	0.0974	0.0495	-0.0671	0.0618
15	0.6712	0.2057	0.0411	1.1918	0.0426	0.0223	-0.0048	0.0068
16	-0.6629	-0.2178	0.1659	1.3692	-0.0972	-0.0027	0.0644	-0.0842
17	0.4364	0.1349	0.0594	1.2192	0.0339	0.0289	0.0065	-0.0157
18	3.4486	1.1193	0.0963	1.0692	0.3653	0.2486	0.1897	-0.2724
19	1.7932	0.5698	0.0964	1.2153	0.1862	0.1726	0.0236	-0.0990
20	-5.7880	-1.9967	0.1017	0.7598	-0.6718	0.1680	-0.2150	-0.0929
21	-2.6142	-0.8731	0.1653	1.2377	-0.3885	-0.1619	-0.2972	0.3364
22	-3.6865	-1.4896	0.3916	1.3981	-1.1950	0.3986	-1.0254	0.5731
23	-4.6076	-1.4825	0.0413	0.8897	-0.3075	-0.1599	0.0373	-0.0527
24	-4.5729	-1.5422	0.1206	0.9476	-0.5711	-0.1197	0.4046	-0.4654
25	-0.2126	-0.0660	0.0666	1.2311	-0.0176	-0.0168	0.0008	0.0056

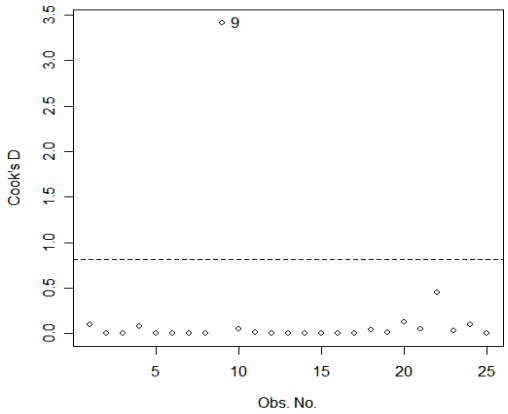

Figure 6.5: A plot of Cook's distance against observation number.

There are two leverage points (observations 9 and 22) since $h_{ii} > \frac{6}{25} = .24$. One influential point is detected since $D_9 > F_{.5,3,22} = .81$.

Treatment of influential points

Read Page 199 in textbook.

Diagnostics for leverage points or influence points are an important part of the regression model-builder's arsenal of tools. They are intended to offer the analyst insight about the data and to signal which observations may deserve more scrutiny.

Influential points, when detected, should be thoroughly investigated before any action is taken. When a point is deleted from an analysis, it should be justified and noted.