Chapter 11

Outliers and Influential Observations

Overview

- Outliers and Influential observations
  - Ancombe example
- Leverage
- Identifying an outlier
  - Studentized residual, RSTUDENT
- Identifying an influential observation
  - DFFIT and DFFITS
  - DFBETA and DFBETAS

11.1 Introduction

- Sometimes, some observations do not fit the proposed model.

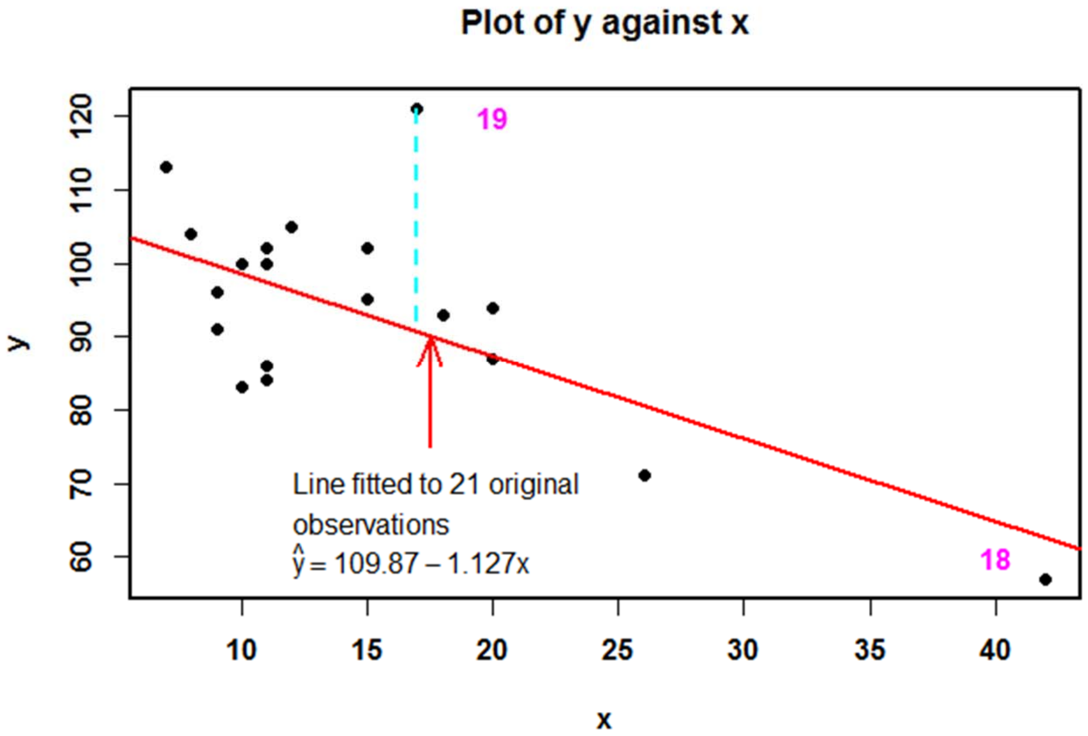
- Observations that do not belong to the model often exhibit numerically large residuals. They are called **outliers**.

Introduction (Continued)

- Two main reasons for outliers:
  - mistakes in inputting or recording the data, (i.e., they don't belong to the model);
  - the algebraic form of the model is incorrect
- Therefore, instead of discarding the outliers, we should study them carefully.
- These outliers may tell us something about the model that we do not know.
- This information can lead to substantial improvements in the model.

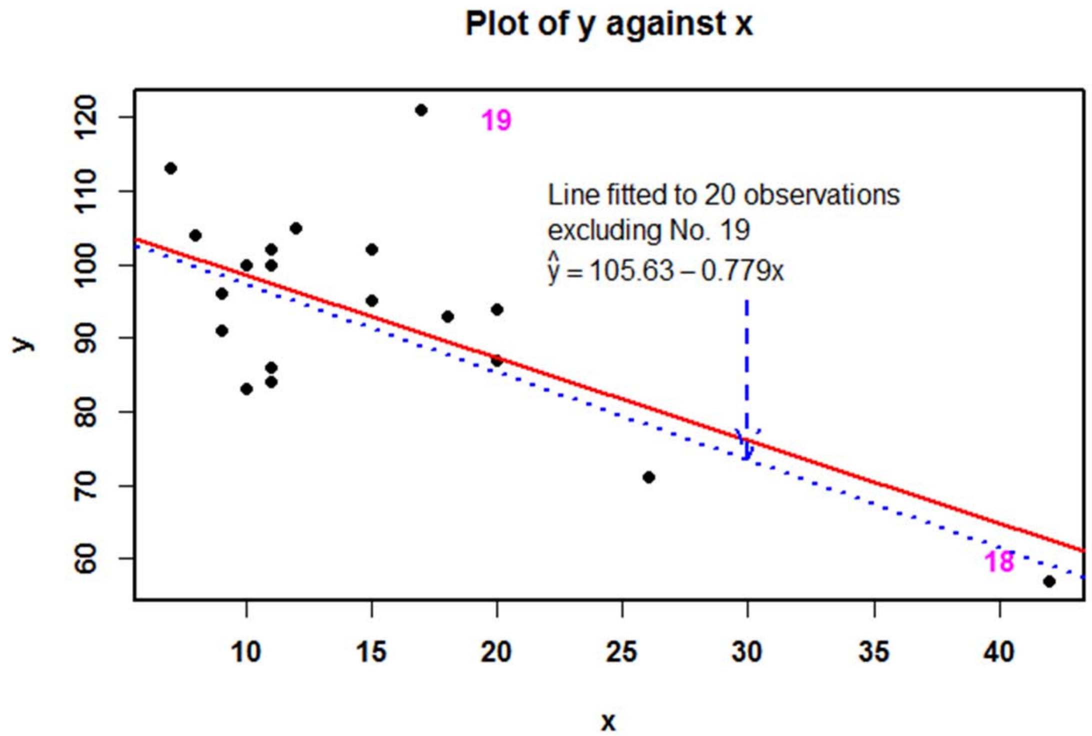
Introduction (Continued)

- A point has undue influence when it has
  - a large residual or
  - is located far away from other points in the space of the predictor variables.
- Observation 19 is considered as an outlier
- It has a large residual



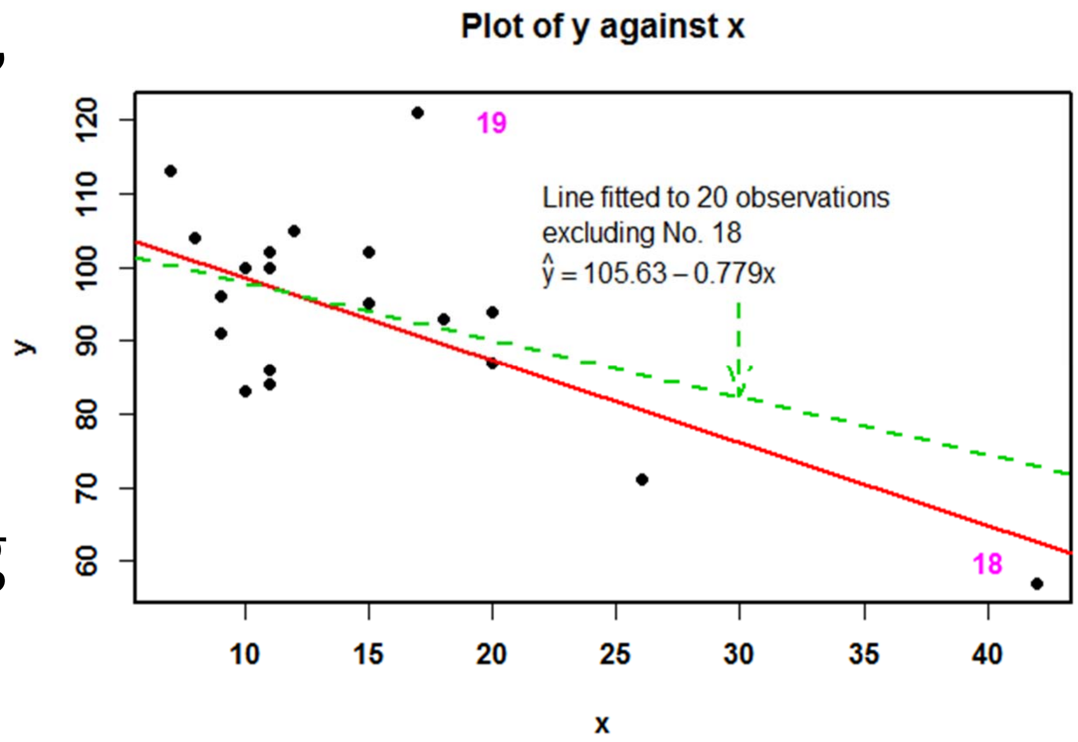
Introduction (Continued)

- The possible influence of Observation 19 is moderated by the fact that there are observations at neighbouring X-space



Introduction (Continued)

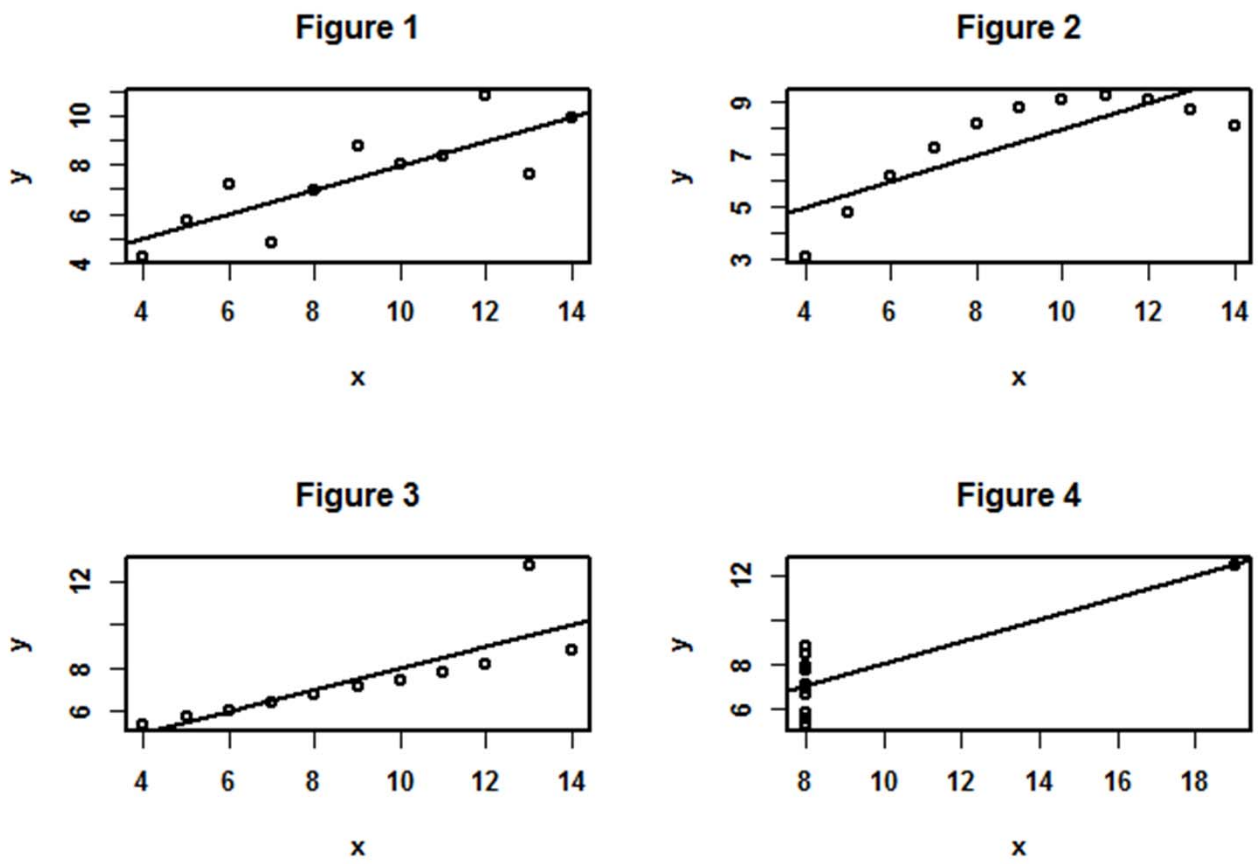
- Observation 18 is considered as an **influential observation**.
- Being alone in the region, it may have a major influence on the position of the model there
- It may or may not have a large residual, depending on the model fitted and the rest of the data



Introduction (Continued)

Anscombe's Example

Ref: Anscombe FJ (1973), "Graphs in Statistical Analysis," The American Statistician, 27, 17-21)



11.2 The Leverage

How to identify influential/outlier observations

- Let $H = X(X'X)^{-1}X'$.
- H is called the hat matrix.
- The leverages h_{ii}'s are the diagonal elements of the hat matrix H. That is,

$$h_{ii} = \underline{x}_i'(X'X)^{-1}\underline{x}_i$$

  where $\underline{x}_i'$ is the row of X

- The leverage h_{ii} describes how far away the i^{th} individual data point is from the centre of all data points, $\underline{\bar{x}} = \sum_{i=1}^{n} \underline{x}_i / n$

The Leverage (Continued)

- Greater influence can be generated at a point far away from the centre, than a point closer to it.

- It can be shown that $\sum_{i=1}^{n} h_{ii} = p + 1$, where p is the number of predictor variables.

- Hence, if all h_{ii}'s are **close** to $(p + 1)/n$ and if all the residuals turn out to be acceptably small, no point will have an undue influence.

The Leverage (Continued)

Drawback

- This method for finding influential observations treats all predictor variables the same regardless of how each one affects the response variable.

Remark:

- Leverage does not make use of the information about the i-th observation y_i, or the i-th residuals e_i

11.3 Studentized Residuals

- The Studentized residuals, often called RSTUDENT, is defined by

$$e_i^* = \frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}} \qquad (11.1)$$

where e_i is the i-th residual and $s_{(i)}$ is similar to s (i.e. $\sqrt{MSE}$) but the least squares method is run after deleting the i-th observation.

Studentized Residuals (Continued)

- Let $\underline{y}_{(i)}$ denote $\underline{y}$ without the i-th entry and $X_{(i)}$ denote X without the i-th row
- Let $\hat{\underline{\beta}}_{(i)}$ be the least squares estimate of $\underline{\beta}$ based on $\underline{y}_{(i)}$ and $X_{(i)}$, i.e.

$$\hat{\underline{\beta}}_{(i)} = \left(X_{(i)}{}'X_{(i)}\right)^{-1} X_{(i)}{}'\underline{y}_{(i)}$$

- Clearly

$$(n-p-2)s_{(i)}^2 = \sum_{\substack{k=1 \\ k\neq i}}^{n} \left[y_k - \underline{x}_k'\hat{\underline{\beta}}_{(i)}\right]^2$$

Studentized Residuals (Continued)

Note:

- $s^2_{(i)}$ is an unbiased and consistent estimate of σ^2

- It can be shown that

$$(n-p-2)s_{(i)}^2 = (n-p-1)s^2 - e_i^2(1-h_{ii})^{-1}$$

- Hence $s^2_{(i)}$ depends on
  - The s, (i.e. $\sqrt{MSE}$), for the full data set
  - The i-th residual, e_i and
  - The i-th leverage value, h_{ii}

Studentized Residuals (Continued)

- Consider adding to the list of predictor variables an indicator variable w which is 1 for the i-th case but is zero otherwise.
- It can be shown that the t-value associated with this indictor variable w is exactly e_i^*.
- Hence e_i^* has a t-distribution when the underlying distribution is normal.
- With the presence of w in the model, the estimates of the coefficients of the other predictor variables and the intercept are not affected by the i-th observation.

Studentized Residuals (Continued)

- Therefore, e_i^* is a standardized measure of the distance between the *i*-th case and the model estimated on the remaining cases.

- Hence it can be served as a test statistic to decide if the *i*-th data point belongs to the model.

- The *i*-th point is considered as a potential outlier or influential observation if $|e_i^*| > 2$ and should be tagged for further investigation.

11.4 DFFIT and DFFITS

- Some other possible measures for influential observations or outliers are to consider how much $\underline{\hat{\beta}}$ and $\underline{\hat{y}}$ would change if a given data point were deleted.

- It can be shown that

$$\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(i)} = \frac{(X'X)^{-1}\underline{x}_i e_i}{1 - h_{ii}}$$

DFFIT

- Hence

$$\text{DFFIT} = \hat{y}_i - \hat{y}_{i(i)} = \underline{x}_i'\hat{\underline{\beta}} - \underline{x}_i'\hat{\underline{\beta}}_{(i)} = \frac{h_{ii}e_i}{1 - h_{ii}}$$

which tells us how much the predicted value $\hat{y}_i$, at the point $\underline{x}_i$, would be affected if the i-th case were deleted.

DFFITS

- In order to eliminate the effect of units of measurement, standardized version of the statistic DFFITS_i is used.

- It can be shown that the variance of $\hat{y}_i$ can be estimated by $s_{(i)}^2 h_{ii}$

- Hence

$$\text{DFFITS}_i = \frac{\sqrt{h_{ii}}e_i}{s_{(i)}(1 - h_{ii})}$$

DFFITS (Continued)

- The i-th case is considered as an influential observation and tagged for further investigation if

$$|\mathrm{DFFITS}_i| > 2\sqrt{\frac{p+1}{n-p-1}}$$

11.5 DFBETA and DFBETAS

- Let

$$(X'X)^{-1}\underline{x}_i = \begin{pmatrix} a_{0,i} \\ \vdots \\ a_{p,i} \end{pmatrix}$$

- The effect of the i-th observation on the estimate of β_j, $j = 0, \ldots, p$ is given by

$$\text{DFBETA}_{ij} = \underline{\hat{\beta}}_j - \underline{\hat{\beta}}_{j(i)} = \frac{a_{ji}e_i}{1 - h_{ii}}$$

DFBETAS (Continued)

- Since variance of $\hat{\beta}_j$ is $q_{jj}\sigma^2$, where q_{jj} is the j-th diagonal element of $(X'X)^{-1}$, hence an estimate of the variance of $\hat{\beta}_j$ is given by $s_{(i)}^2 q_{jj}$.

- Therefore the standardized version of DFBETA_{ij} is given by

$$\mathrm{DFBETAS}_{ij} = \frac{a_{ji} e_i}{s_{(i)}(1 - h_{ii})\sqrt{q_{jj}}}$$

DFBETAS (Continued)

- The i-th case is considered as having large influence on the estimate of β_j and tagged for further investigation if

$$\left|\mathrm{DFBETAS}_{ij}\right| > \frac{2}{\sqrt{n}}$$

DFBETAS (Continued)

- Both DFFITS and DFBETAS are functions of the leverage and the RSTUDENT

$$\text{DFFITS}_i = \sqrt{\frac{h_{ii}}{1 - h_{ii}}}\, e_i^*$$

$$\text{DFBETAS}_{ij} = \frac{a_{ji}}{\sqrt{q_{jj}(1 - h_{ii})}} e_i^*$$

- Therefore, if either the leverage increases or the Studentized residual increases, both measures of influence will increase.

11.6 Other Measures of Influence

- Covariance Ratio

$$\text{Covariance Ratio} = \frac{\det\left(s_{(i)}^2\left(X_{(i)}'X_{(i)}\right)^{-1}\right)}{\det(s^2(X'X)^{-1})}$$

  – A value of this ratio close to 1 would indicate lack of influence of the i-th data point

- Cook's Statistics

$$\frac{\left(\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(i)}\right)' X'X \left(\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(i)}\right)}{(p+1)s^2}$$

  – It can be shown that it is essentially the same as the square of the DFFITS_i

11.7 Programs

SAS program

```
proc reg data = ch11ex1;
  model y = x / influence;
run;
```

R program

```
library(car)
model1=lm(y~x)
influence.measures(model1)
rstudent(model1)
```

11.8 Examples

Example 1

The following data set gives the plots on pages 11.5-7

Case	1	2	3	4	5	6	7	8	9	10	11
y	95	71	93	91	102	87	93	100	104	94	113
x	15	26	10	9	15	20	18	11	8	20	7

Case	12	13	14	15	16	17	18	19	20	21
y	96	83	84	102	100	105	57	121	86	100
x	9	10	11	11	10	11	42	17	11	10

Partial Printout for Example 1 Using SAS

Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	DFBETAS x
1	2.0310	0.1840	0.0479	1.1659	0.0413	0.0166	0.0033
2	-9.5721	-0.9416	0.1545	1.1970	-0.4025	0.1886	-0.3348
3	-15.6040	-1.5108	0.0628	0.9363	-0.3911	-0.3310	0.1924
4	-8.7309	-0.8143	0.0705	1.1151	-0.2243	-0.2000	0.1279
5	9.0310	0.8329	0.0479	1.0850	0.1869	0.0753	0.0149
6	-0.3341	-0.0306	0.0726	1.2013	-0.0086	0.0011	-0.0050
7	3.4120	0.3112	0.0580	1.1702	0.0772	0.0045	0.0327
8	2.5230	0.2297	0.0567	1.1742	0.0563	0.0443	-0.0225
9	3.1421	0.2899	0.0799	1.1997	0.0854	0.0791	-0.0543
10	6.6659	0.6177	0.0726	1.1521	0.1728	-0.0228	0.1014
11	11.0151	1.0508	0.0908	1.0878	0.3320	0.3156	-0.2289
12	-3.7309	-0.3428	0.0705	1.1833	-0.0944	-0.0842	0.0538
13	-15.6040	-1.5108	0.0628	0.9363	-0.3911	-0.3310	0.1924
14	-13.4770	-1.2798	0.0567	0.9923	-0.3137	-0.2468	0.1254
15	4.5230	0.4132	0.0567	1.1590	0.1013	0.0797	-0.0405
16	1.3960	0.1274	0.0628	1.1867	0.0330	0.0279	-0.0162
17	8.6500	0.7983	0.0521	1.0964	0.1872	0.1333	-0.0549
18	-5.5403	-0.8451	0.6516	2.9587	-1.1558	0.8311	-1.1127
19	30.2850	3.6070	0.0531	0.3964	0.8537	0.1435	0.2732
20	-11.4770	-1.0765	0.0567	1.0426	-0.2638	-0.2076	0.1054
21	1.3960	0.1274	0.0628	1.1867	0.0330	0.0279	-0.0162

Partial Printout for Example 1 Using R

Influence measures of
lm(formula = y ~ x) :

	dfb.1_	dfb.x	dffit	cov.r	cook.d	hat	inf
1	0.01664	0.00328	0.04127	1.166	8.97e-04	0.0479	
2	0.18862	-0.33480	-0.40252	1.197	8.15e-02	0.1545	
3	-0.33098	0.19239	-0.39114	0.936	7.17e-02	0.0628	
4	-0.20004	0.12788	-0.22433	1.115	2.56e-02	0.0705	
5	0.07532	0.01487	0.18686	1.085	1.77e-02	0.0479	
6	0.00113	-0.00503	-0.00857	1.201	3.88e-05	0.0726	
7	0.00447	0.03266	0.07722	1.170	3.13e-03	0.0580	
8	0.04430	-0.02250	0.05630	1.174	1.67e-03	0.0567	
9	0.07907	-0.05427	0.08541	1.200	3.83e-03	0.0799	
10	-0.02283	0.10141	0.17284	1.152	1.54e-02	0.0726	
11	0.31560	-0.22889	0.33200	1.088	5.48e-02	0.0908	
12	-0.08422	0.05384	-0.09445	1.183	4.68e-03	0.0705	
13	-0.33098	0.19239	-0.39114	0.936	7.17e-02	0.0628	
14	-0.24681	0.12536	-0.31367	0.992	4.76e-02	0.0567	
15	0.07968	-0.04047	0.10126	1.159	5.36e-03	0.0567	
16	0.02791	-0.01622	0.03298	1.187	5.74e-04	0.0628	
17	0.13328	-0.05493	0.18717	1.096	1.79e-02	0.0521	
18	0.83112	-1.11275	-1.15578	2.959	6.78e-01	0.6516	*
19	0.14348	0.27317	0.85374	0.396	2.23e-01	0.0531	*
20	-0.20761	0.10544	-0.26385	1.043	3.45e-02	0.0567	
21	0.02791	-0.01622	0.03298	1.187	5.74e-04	0.0628	

Partial Printout for Example 1 Using R (Continued)

```
> rstudent(model1)
          1           2           3           4           5           6 
 0.18396849 -0.94158335 -1.51081192 -0.81426336  0.83286292 -0.03063183 
          7           8           9          10          11          12 
 0.31124676  0.22971575  0.28991014  0.61766026  1.05084716 -0.34283148 
         13          14          15          16          17          18 
-1.51081192 -1.27977575  0.41315320  0.12739342  0.79828114 -0.84511086 
         19          20          21 
 3.60697972 -1.07648108  0.12739342 
```

Example 1 (Continued)

From the printout, we have the following.

Leverage

- The value of h_{ii} for Observation 18 is 0.6516.
- It is much higher than the expected value $(p+1)/n = 0.0952$.
- Hence Observation 18 is a potential influential observation.

Studentized residuals RSTUDENT

- The value of e_i^* for Observation 19 is 3.6070.
- It is much higher than 2.
- Hence Observation 19 is a potential influential observation.

Example 1 (Continued)

DFFITS

- The value of DFFITS for Observations 18 and 19 are -1.1558 and 0.8537.
- They are much higher than $2\sqrt{2/19} = 0.6489$.
- Hence Observations 18 and 19 are potential influential observations.

Example 1 (Continued)

DFBETAS

For β_0

- The value of DFBETAS for β_0 for Observation 18 is 0.8311.

- It is bigger than $2/\sqrt{21} = 0.4364$.

- Hence Observation 18 is a potential influential observation.

Example 1 (Continued)

DFBETAS

For β_1

- The value of DFBETAS for β_1 for Observation 18 is -1.1127.
- It is much higher than 0.4364.
- Hence Observation 18 is a potential influential observation.
- To summarize, Observations 18 and 19 are potential influential observations or outliers and should be tagged for further study.

Example 2

The data for Example 2 are given in the file "ch11ex2.txt" in the IVLE.

Partial Printout for Example 2 using SAS

			Hat Diag	Cov		-----------	DFBETAS	-----------
Obs	Residual	RStudent	H	Ratio	DFFITS	Intercept	x1	x2
1	-0.8092	-0.3780	0.2291	1.5679	-0.2061	0.0482	-0.1776	-0.0454
2	-1.5768	-0.6812	0.0766	1.2176	-0.1963	-0.0973	-0.0536	0.0599
3	-1.0650	-0.4715	0.1364	1.3746	-0.1874	-0.1714	0.1085	0.1173
4	7.7691	9.9314	0.1256	0.0023	3.7646	2.5511	0.8506	-2.2690
5	-0.6770	-0.2909	0.0931	1.3506	-0.0932	-0.0716	0.0518	0.0362
6	0.2861	0.1329	0.2276	1.6104	0.0721	-0.0358	0.0026	0.0603
7	0.5104	0.2437	0.2669	1.6805	0.1471	-0.0815	0.0138	0.1278
8	0.3437	0.1601	0.2318	1.6162	0.0880	-0.0379	-0.0082	0.0702
9	0.3860	0.1729	0.1691	1.4929	0.0780	-0.0235	-0.0139	0.0551
10	-0.2317	-0.0989	0.0852	1.3622	-0.0302	-0.0138	0.0161	-0.0001
11	-0.3165	-0.1353	0.0884	1.3644	-0.0421	-0.0343	0.0191	0.0199
12	0.2649	0.1150	0.1152	1.4073	0.0415	0.0244	0.0132	-0.0211
13	0.9924	0.4382	0.1339	1.3800	0.1723	0.1612	-0.0845	-0.1172
14	-1.8408	-1.3005	0.6233	2.2972	-1.6729	0.3139	-1.5494	-0.0812
15	-0.1413	-0.0598	0.0699	1.3417	-0.0164	-0.0017	0.0018	-0.0055
16	-1.6099	-0.7447	0.1891	1.3589	-0.3597	-0.3485	0.1999	0.2687
17	-2.2845	-1.0454	0.1386	1.1381	-0.4194	-0.3769	0.2593	0.2495

Partial Printout for Example 2 Using R

```
> influence.measures(model2)
Influence measures of
         lm(formula = y ~ x1 + x2) :

     dfb.1_   dfb.x1    dfb.x2    dffit   cov.r   cook.d    hat inf
1   0.04820 -0.17760 -0.045398 -0.2061 1.56788 1.51e-02 0.2291
2  -0.09726 -0.05358  0.059948 -0.1963 1.21756 1.33e-02 0.0766
3  -0.17139  0.10847  0.117314 -0.1874 1.37460 1.24e-02 0.1364
4   2.55105  0.85060 -2.269011  3.7646 0.00226 5.92e-01 0.1256   *
5  -0.07157  0.05180  0.036238 -0.0932 1.35062 3.10e-03 0.0931
6  -0.03582  0.00261  0.060327  0.0721 1.61043 1.87e-03 0.2276
7  -0.08147  0.01377  0.127822  0.1471 1.68054 7.73e-03 0.2669   *
8  -0.03790 -0.00824  0.070233  0.0880 1.61625 2.77e-03 0.2318
9  -0.02348 -0.01388  0.055145  0.0780 1.49288 2.18e-03 0.1691
10 -0.01379  0.01607 -0.000133 -0.0302 1.36219 3.26e-04 0.0852
11 -0.03430  0.01909  0.019900 -0.0421 1.36436 6.37e-04 0.0884
12  0.02436  0.01321 -0.021125  0.0415 1.40734 6.17e-04 0.1152
13  0.16119 -0.08454 -0.117175  0.1723 1.38002 1.05e-02 0.1339
14  0.31395 -1.54939 -0.081173 -1.6729 2.29721 8.89e-01 0.6233   *
15 -0.00169  0.00185 -0.005484 -0.0164 1.34171 9.63e-05 0.0699
16 -0.34855  0.19991  0.268732 -0.3597 1.35890 4.45e-02 0.1891
17 -0.37688  0.25928  0.249471 -0.4194 1.13812 5.82e-02 0.1386
```

Partial Printout for Example 2 Using R (Continued)

```
> rstudent(model2)
          1           2           3           4           5           6 
-0.37801050 -0.68119342 -0.47146876  9.93141754 -0.29091422  0.13289608 
          7           8           9          10          11          12 
 0.24372410  0.16013046  0.17293852 -0.09885476 -0.13532374  0.11495241 
         13          14          15          16          17 
 0.43816486 -1.30053194 -0.05975554 -0.74468118 -1.04541119 
```

Example 2 (Continued)

From the printout, we have the following.

Leverage

- The value of h_{ii} for Observation 14 is 0.6233.
- It is much higher than the expected value $(p + 1)/n = 0.1764$.
- Hence Observation 14 is a potential influential observation.

Example 2 (Continued)

From the printout, we have the following.

Studentized residuals RSTUDENT

- The value of e_i^* for Observation 4 is 9.9314.
- It is much higher than 2.
- Hence Observation 4 is a potential influential observation.

Example 2 (Continued)

DFFITS

- The value of DFFITS for Observations 4 and 14 are 3.7646 and -1.6729.

- They are much higher than $2/\sqrt{3/14} = 0.9258$.

- Hence Observations 4 and 14 are potential influential observations.

Example 2 (Continued)

DFBETAS

For β_0

- The value of DFBETAS for β_0 for Observation 4 is 2.5511.
- It is bigger than $2/\sqrt{17} = 0.4851$.
- Hence Observation 4 is a potential influential observation.

Example 2 (Continued)

DFBETAS

For β_1

- The values of DFBETAS for β_1 for Observations 4 and 14 are 0.8506 and −1.5494 respectively
- It is much higher than 0.4851.
- Hence Observations 4 and 14 are potential influential observations.

Example 2 (Continued)

DFBETAS

For β_2

- The value of DFBETAS for β_2 for Observation 4 is -2.2690.
- It is much higher than 0.4851.
- Hence Observation 4 is a potential influential observation.
- To summarize, Observations 4 and 14 are potential influential observations or outliers and should be tagged for further study.

Recap

- Leverage:
  - Check if h_{ii} is **far away** from to $(p + 1)/n$
- Studentized residual (RSTUDENT), e_i^*
  - Check if $|e_i^*| > 2$
- DFFIT and DFFITS
  - Check if $|\text{DFFITS}_i| > 2\sqrt{(p + 1)/(n - p - 1)}$
- DFBETA and DFBETAS
  - Check if $\left|\text{DFBETAS}_{ij}\right| > 2/\sqrt{n}$