3. Diagnostics I: Residuals and Influence

The methods of estimating and testing of chapters 1 – 2 & 5 – 6 are only half the analysis required for regression. This part of regression analysis is called the **aggregate analysis**. All these models make **assumptions** about the **errors** e_i: $e_i \sim \text{NID}(0, \sigma^2)$. The second half of a regression analysis consists of determining unusual cases and checking assumptions – collectively known as **diagnostics**.

In Chapter 3, we study methods for determining if a particular case is "**unusual**" in any way. Once identified, these unusual cases could be omitted from the model to obtain a better fit. This is known as **case analysis** and the methods are based on the residuals, $\hat{e}_i$.

The residuals, $\hat{e}_i$, provide information about the true errors and we study **regression diagnostic** plots of these residuals to test the assumptions about the errors, i.e. the assumptions of normality and constant variance. This is covered in Chapter 4.

Outliers:

In the simple linear model $y_i = \beta_0 + \beta_1 x_i + e_i$, the ith case is said to be an **outlier** if the corresponding **residual** , $\hat{e}_i = y_i - \hat{y}_i$, is **large**.

Cases of high leverage:

The ith case is said to be a **case of high leverage** if the corresponding ***x* – value**, x_i, is far from the mean of the x–values, $\bar{x}$, i.e. if the x-value is unusually **large** or unusually **small**.

Cases of high influence:

The ith case is said to be a **case of high influence** if **omitting** this case from the data set would cause **large changes** in the values of the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

Video 3.1

Example: The usefulness of plots (Anscombe data set)

This example uses artificial data to illustrate an important point: always draw the scatter-plot before fitting a regression line.

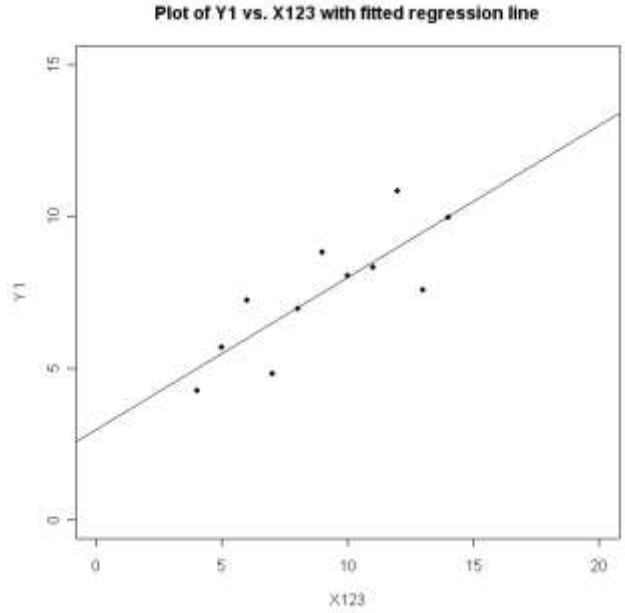
The example consists of four data sets of (x,y)-values, each with 11 cases. A simple linear regression model $y_i = \beta_0 + \beta_1 x_i + e_i$ is fitted to each data set. Each data set leads to an **identical aggregate analysis**:

$$\hat{\beta}_0 = 3.0, \hat{\beta}_1 = 0.5, \hat{\sigma}^2 = 1.53, \mathrm{R}^2 = 0.667$$

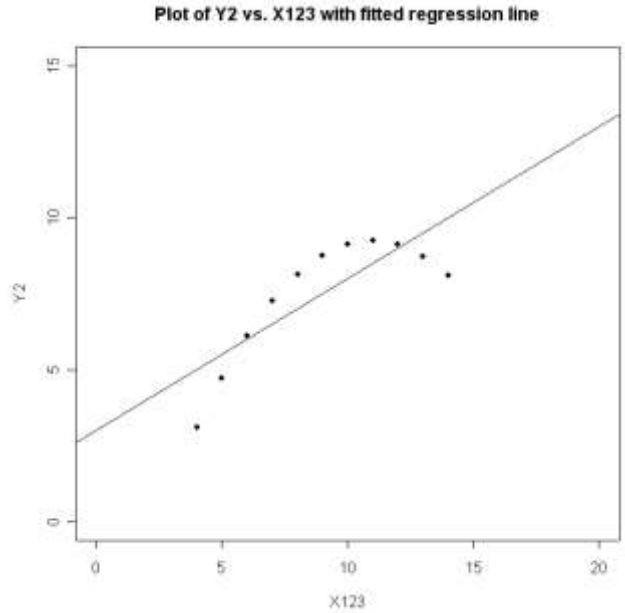
X1	Y1	X2	Y2	X3	Y3	X4	Y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

However, the scatter-plots show that it is **not appropriate** to fit a simple linear regression model for some of the four data sets.

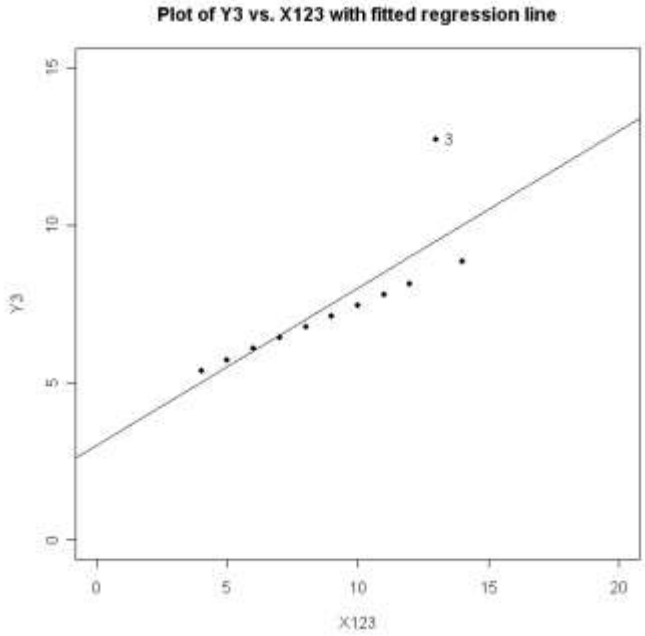
The scatter-plot for the first data set indicates that a simple linear regression model **is appropriate**.



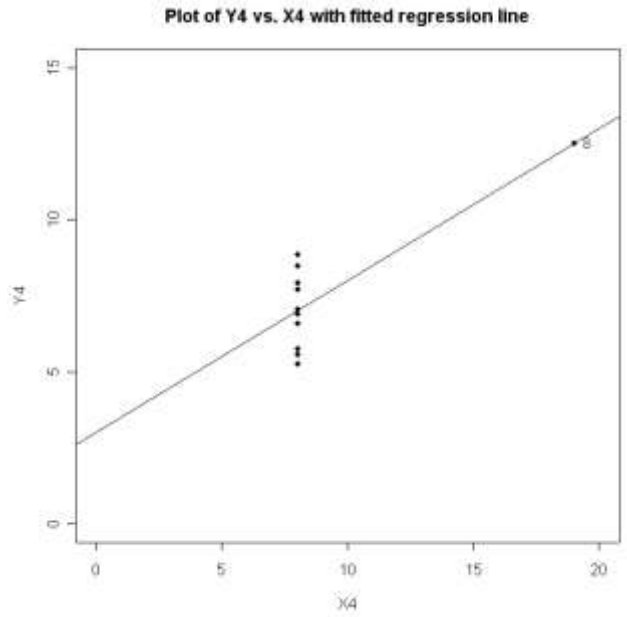
The scatter-plot for the second data set indicates that a simple linear regression model is **inappropriate** and that a smooth curve, perhaps a **quadratic** model in X instead of a linear model, could be fitted to the data set.



The scatter plot for the third data set indicates that a simple linear regression model may be correct for most of the data**,** but that there is one **outlier**, case 3. If this case is omitted, then the remaining 10 points are perfectly fitted by the regression line $\hat{y}$ = 4.0 + 0.346*x*, which is **quite different** from that obtained from all 11 cases. Thus case 3 is a **case of high influence**.



In the scatter plot for the fourth data set, the *x*-value for case 8 is *x* = 19.0, but the *x*-values for all the remaining cases are all the **same**: *x* = 8.0. Thus case 8 is a **point of high leverage**, since its *x*-value is unusual. The slope parameter, $\hat{\beta}_1$, is largely determined by case 8. If case 8 is omitted, we could not even estimate β_1.



Residuals

Recall that a multiple regression model can be expressed in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}$$

It may be shown that

$$\mathrm{E}(\hat{e}_i) = 0$$
$$\mathrm{var}(\hat{e}_i) = \sigma^2(1 - h_{ii})$$

where h_{ii} = ith diagonal element of the **hat matrix**

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

The hat matrix is so called because

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

where $\hat{\mathbf{Y}}$ is the vector of fitted values $\hat{y}_i$. Thus the hat matrix transforms the vector of observed responses $\mathbf{Y}$ into $\hat{\mathbf{Y}}$. The quantity h_{ii} is called the **leverage** of the ith case. It may be shown that

$$h_{ii} \leq 1$$

As $\mathrm{var}(\hat{e}_i) = \sigma^2(1 - h_{ii})$, it follows that

as $h_{ii} \to 1$, $\mathrm{var}(\hat{e}_i) \to 0$, i.e. $\hat{e}_i = y_i - \hat{y}_i \to 0$

For a case with h_{ii} close to 1, the fitted value $\hat{y}_i$ will be close to y_i, so for a case with high leverage, the fitted line will pass close to the corresponding point. It may be shown that for regression models with an intercept,

$1/n \leq h_{ii}$, where n = number of cases.

In addition, it may be shown that

$\sum_{i=1}^{n} h_{ii} = p'$, where $p' = p + 1$ is the number of predictors (or columns in the **X** matrix.

Thus the **average leverage** is

$$\frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p'}{n}$$

By convention, if

$$h_{ii} > \frac{2p'}{n},$$

the i[th] case is said to **high leverage**.

In the simple linear model, $y_i = \beta_0 + \beta_1 x_i + e_i$,

it may be shown that

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

Hence we can confirm by direct summation that

$$\sum_{i=1}^{n} h_{ii} = 2 = p'$$

The leverage h_{ii} will achieve its minimum value $\frac{1}{n}$ when $x_i = \bar{x}$ and will increase as x_i deviates from $\bar{x}$. Thus the leverage h_{ii} of the i[th] case measures how **unusual** x_i is.

In multiple regression, the leverage h_{ii} of the i[th] case measures how **unusual** are the *x*- values for the ith case, i.e. h_{ii} essentially measures the **distance** between the vector of *x*-values for the i[th] case

$$(x_{i1}, x_{i2}, ..., x_{ip})$$

and the vector of mean values of the *X*- variables

$$(\bar{x}_1, \bar{x}_2, ..., \bar{x}_p)$$

Thus the leverage h_{ii} can be large if individual deviations $x_{ij} - \bar{x}_i$ are large.

In the fuel consumption example, the number of cases is n = 48. If the following model is fitted to the fuel consumption data in Chapter 2,

$$\text{FUEL} = \beta_0 + \beta_1 \text{ TAX } + \beta_2 \text{ DLIC},$$

then $p' = p + 1 = 3$ and so the i[th] case has **high leverage** if

$$h_{ii} > \frac{2p'}{n} = \frac{2(3)}{48} = 0.125$$

For the fuel consumption data set, it may be shown that the case with highest leverage is Texas (TX) (case number 37) with $h_{ii} = 0.2067 > 0.125$, so this case does have high leverage. We compare the values of TAX and DLIC for Texas with the summary statistics for the data set:

	TAX	DLIC
37 Texas	5.00	56.6
Minimum	5.00	45.100
Average	7.663	57.033
Maximum	10.00	72.400

This shows that the TAX rate of 5c per gallon in Texas is the **smallest** of all the 48 states. The percentage of the population in Texas with driver's licences (DLIC = 56.6%) is close to the average value over all 48 states (57.033%). Thus Texas is a case of high leverage because of the low TAX rate.

Video 3.2

Outliers and Studentized residuals

From earlier, $\mathrm{E}(\hat{e}_i) = 0$ and $\mathrm{var}(\hat{e}_i) = \sigma^2(1-h_{ii})$.

This implies that the residuals $\hat{e}_i$ do **not have the same variance**, in general.

For cases of high leverage, h_{ii} is **large** and so $\mathrm{var}(\hat{e}_i)$ is **small**.
For cases of low leverage, h_{ii} is **small** and so $\mathrm{var}(\hat{e}_i)$ is **large**.

We **scale** the residuals so that the scaled residuals all have the same variance 1.
The i[th] (internally) **Studentized residual** r_i is defined as

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{(1-h_{ii})}}$$

If the model is correct, then

$\mathrm{E}(r_i) = 0$,
$\mathrm{Var}(r_i) = 1$

By convention, a case is said to be an **outlier** if

$$|\, r_i \,| > 2$$

i.e. if the absolute value of the i[th] Studentized residual exceeds 2.

For the fuel consumption data, it may be shown the cases with highest Studentized residuals are North Dakota (ND) and Wyoming (WY)

State	r_i
37 ND	2.0659
40 WY	3.3453

Both of these states are outliers.

Video 3.3

Influence of cases

For the fuel consumption data set, with the model

$$\text{FUEL} = \beta_0 + \beta_1 \text{ TAX } + \beta_2 \text{ DLIC},$$

the i^{th} case is said to be a **case of high influence** if **omitting** that case from the data set would cause **large changes** in the values of the least squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$. The model when fitted to all 48 states is:

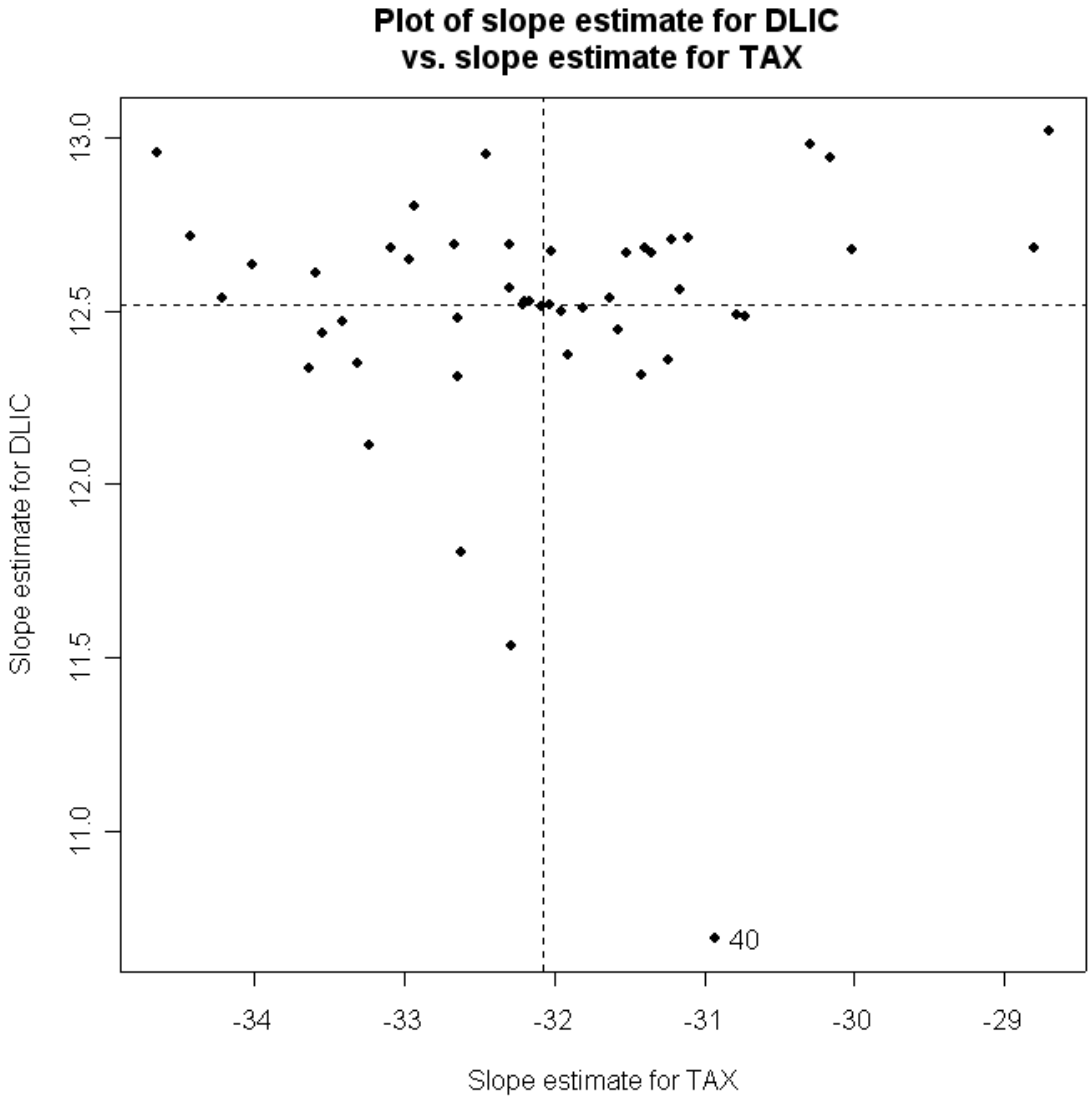
$$\text{FUEL} = 108.971 - 32.075 \text{ TAX} + 12.515 \text{ DLIC}$$

If the States of Texas and Wyoming are omitted in turn and these estimates are recalculated, we get:

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
All states	108.971	–32.075	12.515
Omit Texas	125.231	–33.549	12.436
Omit Wyoming	198.652	–30.933	10.691

Thus Texas would appear to have **low influence** and Wyoming to have **high influence**.

Each of the 48 states is omitted in turn and the estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$ are recorded. In the scatter-plot below, these estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$ are plotted and compared to the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ obtained using all 48 states.



From this, it appears that the point for Wyoming (observation number 40) is the farthest from the point for the original estimates for all 48 states, suggesting that Wyoming may have **high influence**.

To objectively assess how far a point is from the original estimates, we need a numerical measure.

The **Cook's distance**, D_i, for case i is a measure of the distance from the point $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ with case i omitted to the point $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ with all cases included.
The statistic D_i is calculated as follows:

$$D_i = \frac{1}{p'} r_i^2 \left(\frac{h_{ii}}{1 - h_{ii}} \right)$$

where r_i is the ith Studentized residual and h_{ii} is the leverage of the ith case.

If D_i is large, then case i has **high influence**.

A large value of D_i may be due to large r_i, large h_{ii}, or both.

Thus a case may have high influence if it is an **outlier** (large r_i), or if it has **high leverage** (large h_{ii}), or both.

However, even if a case is an outlier, it may not have high influence if its leverage h_{ii} is small.

Similarly, even if a case has high leverage, it may not have high influence if its Studentized residual r_i is small.

Summary for Case Analysis of 3 States for the Fuel Consumption data:

State	r_i	h_{ii}	D_i
18 ND	2.0659	0.0444	0.0661
37 TX	–0.2497	0.2067	0.0054
40WY	3.3453	0.0930	0.3826

North Dakota (ND) is an **outlier** (r_i = 2.0659 >2), has low leverage (h_{ii} = 0.0444) and low influence (D_i = 0.0661).

Texas (TX) is not an outlier (r_i = –0.2497), has **high leverage** (h_{ii} = 0.2067 > 0.125) and low influence (D_i = 0.0054).

Wyoming (WY) is an **outlier** (r_i = 3.3453 > 2), has low leverage (h_{ii} = 0.0930) and **high influence** (D_i = 0.3826).

Video 3.4

Rat Data Set

An experiment was conducted to investigate the amount of a particular drug **retained** in the liver of a rat. Large livers would absorb more of a given dose than smaller livers, so ideally the dose administered should be **proportional to liver weight**. However, liver weight will not be known until **after** the rat is killed, so assuming that body weight is proportional to liver weight, the dose administered is made **proportional to body weight** as follows

	BodyWeight	LiverWeight	Dose	Y
1	176	6.5	0.88	0.42
2	176	9.5	0.88	0.25
3	190	9.0	1.00	0.56
4	176	8.9	0.88	0.23
5	200	7.2	1.00	0.23
6	167	8.9	0.83	0.32
7	188	8.0	0.94	0.37
8	195	10.0	0.98	0.41
9	176	8.0	0.88	0.33
10	165	7.9	0.84	0.38
11	158	6.9	0.80	0.27
12	148	7.3	0.74	0.36
13	149	5.2	0.75	0.21
14	163	8.4	0.81	0.28
15	170	7.2	0.85	0.34
16	186	6.8	0.94	0.28
17	146	7.3	0.73	0.30
18	181	9.0	0.90	0.37
19	149	6.4	0.75	0.46

The heaviest rat is rat 5, with a body weight of 200g. The **relative dose** for this rat is 200/200 = 1.00, indicating that this rat is given the **full dose**. Rat 19 has a body weight of 149g. The relative dose for this rat is 149/200 = 0.75, indicating that this rat is given 75% of the full dose.

After a fixed length of time, each rat was sacrificed, the liver weighed and the percentage of the dose in the liver determined. Thus rat 1 had a body weight of 176 grams and a liver weight of 6.5 grams. The relative dose administered to this rat was 0.88 (i.e. 88% of the full dose) and the percentage of the dose administered that was retained in the liver was 0.42 = 42%.

Note that rat 3 has a bodyweight of 190g and so the relative dose for this rat should be 190/200 = 0.95. However, the relative dose for this rat was 1.00, indicating that the full dose was mistakenly administered to this rat. This makes rat 3 an "**unusual**" case and it is of interest to see whether the regression diagnostics detect it as such.

The **experimental hypothesis** is that, for the method of determining the dose, there is **no relationship** between the percentage of the dose retained in the liver (Y) and the body weight, the liver weight and the relative dose.

The correlation coefficients are:

	BodyWeight	LiverWeight	Dose	Y
BodyWeight	1.0000000	0.5000101	0.9902126	0.1510855
LiverWeight	0.5000101	1.0000000	0.4900711	0.2033302
Dose	0.9902126	0.4900711	1.0000000	0.2275436
Y	0.1510855	0.2033302	0.2275436	1.0000000

Note the high correlation (0.990) between relative dose and body weight, because the relative dose was chosen to be proportional to body weight, albeit with the error for rat 3. However, the correlation coefficient between body weight and liver weight is only 0.500.

To test the experimental hypothesis, we fit each of the following models in turn:

$$Y = \beta_0 + \beta_1 X_1,\ Y = \beta_0 + \beta_2 X_2,\ Y = \beta_0 + \beta_3 X_3$$

and $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

X_1: Body weight X_2: Liver weight X_3: relative Dose

Note that in the last model, the variables X_1 and X_3 are **highly collinear** (r = 0.990) and so the variances of the estimates of the parameters in this model are **greatly inflated**.

The parameter estimates (and t-values) in these models are:

	Model Including			
Coefficient	X_1	X_2	X_3	X_1, X_2, X_3
β_0	0.196	0.220	0.133	0.266
	(0.89)	(1.64)	(0.63)	(1.37)
β_1 (Body Weight)	0.0008			-0.0212
	(0.63)			(-2.66)
β_2 (Liver Weight)		0.0147		0.0143
		(0.86)		(0.83)
β_3 (Dose)			0.235	4.178
			(0.96)	(2.74)

None of the simple regression models of dose retained (Y) on any of the independent variables (X_1, X_2 or X_3) is significant, all having t-values less than 1.

However, the regression of Y on X_1, X_2 and X_3 gives different and contradictory results.

In this model, X_1 and X_3 have significant t-values (t = –2.66 and 2.74, respectively), indicating that X_1 and X_3 **combined** are a useful indicator of Y.

If X_2 is dropped from the model, the phenomenon remains, i.e. X_1 and X_3 have significant t-values. The **aggregate analysis** suggests that a combination of body weight (X_1) and relative dose (X_3) **is associated** with the response (Y), the percentage of the dose retained in the liver.

We use **case analysis** to explore this paradox.
The residuals, Studentized residuals, leverages and Cook's distances for each of the n =19 cases are shown below:

	Yi	$\hat{e}_i$	r_i	h_{ii}	D_i
1	0.42	0.123758021	1.76604714	0.17798270	1.688268e-01
2	0.25	-0.089136157	-1.27303970	0.17934099	8.854024e-02
3	0.56	0.024088214	0.80715401	**0.85091457**	**9.296160e-01**
4	0.23	-0.100557321	-1.37723226	0.10761585	5.718456e-02
5	0.23	-0.067711915	-1.12309856	0.39153825	2.029162e-01
6	0.32	0.007131221	0.10073807	0.16115958	4.874208e-04
7	0.37	0.056580285	0.78795102	0.13688107	2.461564e-02
8	0.41	0.049584065	0.74258710	0.25367448	4.685795e-02
9	0.33	0.012310932	0.16490096	0.06701578	4.883028e-04
10	0.38	-0.002844507	-0.03922453	0.11968672	5.229549e-05
11	0.27	-0.080146346	-1.10507028	0.11950583	4.143644e-02
12	0.36	0.042357750	0.60240838	0.17239599	1.889847e-02
13	0.21	-0.098151103	-1.53565893	0.31618336	2.726019e-01
14	0.28	-0.027142867	-0.37680527	0.13140699	5.370022e-03
15	0.34	0.031614703	0.42556276	0.07617481	3.733265e-03
16	0.28	-0.058754717	-0.85886404	0.21661460	5.099189e-02
17	0.30	-0.018353809	-0.26470261	0.19522441	4.249284e-03
18	0.37	0.060682327	0.85093422	0.14872221	3.162543e-02
19	0.46	0.134691226	1.92204135	0.17796183	1.999403e-01

- There are **no outliers**, since all the Studentized residuals r_i satisfy $|r_i| < 2$.
- The Cook's distance for case 3 is D_3 = 0.93. The next largest value of Cook's distance is 0.27, so case 3 has very **high influence**.
- Case 3 also has the **highest leverage** h_{33} = 0.85, indicating that case 3 has "unusual" x-values.

When case 3 is deleted from the regression of Y on X_1, X_2 and X_3, none of the predictor variables have significant t- values and the paradox disappears. Thus the apparent relationship can be ascribed to case 3 alone.

Recall that case 3 should have X_3 = 0.95, rather than X_3 = 1.00.

Video 3.5

Summer 2006 Question 3

Data were collected to study the relationship between degree of brand liking, and the moisture content and sweetness of 16 similar products. A model of the following form was fitted to these data:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i, \quad e_i \sim \text{NID}(0, \sigma^2),$$

where

Y = Liking = index of brand liking,
X_1 = Moisture = moisture content of product,
X_2 = Sweetness = sweetness of product.

Excerpts from the R output for this model are shown on the next page.
This output includes a table showing the residuals (e), the studentized residuals (r), leverages (h) and Cook's distances (d) for selected cases.

(a) What is meant by a residual? By how much does the above model under/over-estimate the Liking in case 3?

(b) What is meant by an outlier? Which of the cases shown are outliers?

(c) What is meant by a case of high leverage? Which of the cases shown have high leverage? Can you explain why these cases have high leverage?
Can you explain why all 3 cases shown in the output have the same leverage value?

(d) What is meant by a case of high influence? Which of the cases shown has the highest influence? Can you explain why this case has high influence? What are the regression coefficients for the model with the case of highest influence omitted?

R output for Question 3

```
> brands1.lm <- lm(Liking ~ Moisture + Sweetness,
+ data=brands.df)

> coef(brands1.lm)
(Intercept)    Moisture   Sweetness
     37.650       4.425       4.375

> cbind(e,r,h,d)[c(3,14,15),]
       e         r      h         d
3  -3.10 -1.318128 0.2375 0.1803922
14 -4.40 -1.870891 0.2375 0.3634123
15  3.35  1.424429 0.2375 0.2106609


> brands.df[c(3,14,15),]
   Liking Moisture Sweetness
3      61        4         2
14     95       10         4
15     94       10         2

> summary(Moisture)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    4.0     5.5     7.0     7.0     8.5    10.0

> summary(Sweetness)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2       2       3       3       4       4

> lm.influence(brands1.lm)$coefficients[c(3,14,15),]
   (Intercept)   Moisture  Sweetness
3  -2.08360656  0.1524590  0.2540984
14  2.23606557 -0.2163934 -0.3606557
15 -0.05491803  0.1647541 -0.2745902
```

Practical (Assignment) 3

Instructions for this practical

- Open the template "Surname Forename Chpt x" from Canvas (in "Practicals").
- Complete the grid on the first page.
- Save this file (as a Word document) using your own surname, forename and the appropriate chapter number.

- Practice Question:
- Type the commands one by one into R.
- Compare the results in the R text output and graphics with the corresponding results and figures in your notes.
- Use appropriate R output to answer the questions, adapting the R code if necessary.

- Exam Question:
- Adapt the relevant R code you used for the practice question to answer the questions.
- Copy and paste the relevant R text output and graphics into your Word document to support your answers. Change the text font to "Courier New" to align columns.

- Restrict your Word document to a **maximum of 2 pages** (re-sizing graphics and deleting irrelevant R output will help).
- Submit this Word document **via Canvas** by **5.00pm 13th November 2020** (**STRICT** deadline)
- Note that submitting the practical is a declaration that the practical is your own work. Plagiarism/copying will not be tolerated.

Practice Question (not to be submitted)

For the data in the fuel.txt dataset for 48 contiguous US states, fit a model of the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e,\ e \sim \text{IN}(0, \sigma^2),$$

where X_1 = TAX, X_2 = DLIC and Y = FUEL (see Practical 2 for a description of this dataset).

(a) By how much does the above model under/over-estimate the fuel consumption per capita in the state of Maine (ME)?
(b) Plot the studentized residuals vs. observation number and identify any outliers.
(c) Plot the leverages vs. observation number and identify any cases of high leverage. Which of the states has the highest leverage? Can you explain why this case has high leverage?
(d) Plot the Cook's distances vs. observation number and identify any cases of high influence. Which of the states has the highest influence? Can you explain why this case has high influence?
(e) Give a table showing the residuals, studentized residuals, leverages and Cook's distances for the outliers, cases of high leverage and cases of high influence.

Exam Question (Winter 2019-20, Question 3) (to be submitted)

A recruitment consulting company has taken a random sample of executives running private companies to investigate the variables that potentially influence the salaries of those executives. The data are stored in **Executives.txt (on Canvas)**.

Fit a model of the following form to these data:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + e_i , \quad e_i \sim \text{NID}(0, \sigma^2),$$

where

Y = Salary = salary (€) of the executive,
X_1 = Experience = experience (years) of the executive,
X_2 = Education = length of education (years) of the executive,
X_3 = Profits = profits (€) of the executive's company in last year,
X_4 = Sales = sales (€) of the executive's company in last year.

Calculate the following for each case:
the residuals (e), the studentized residuals (r), leverages (h) and Cook's distances (d).

Note: As there is a large number of cases, it may be easier to answer the questions below using plots rather than lists of the values.

(a) What is meant by a residual? By how much does the model under/over-estimate salary for case 1? (8 marks)

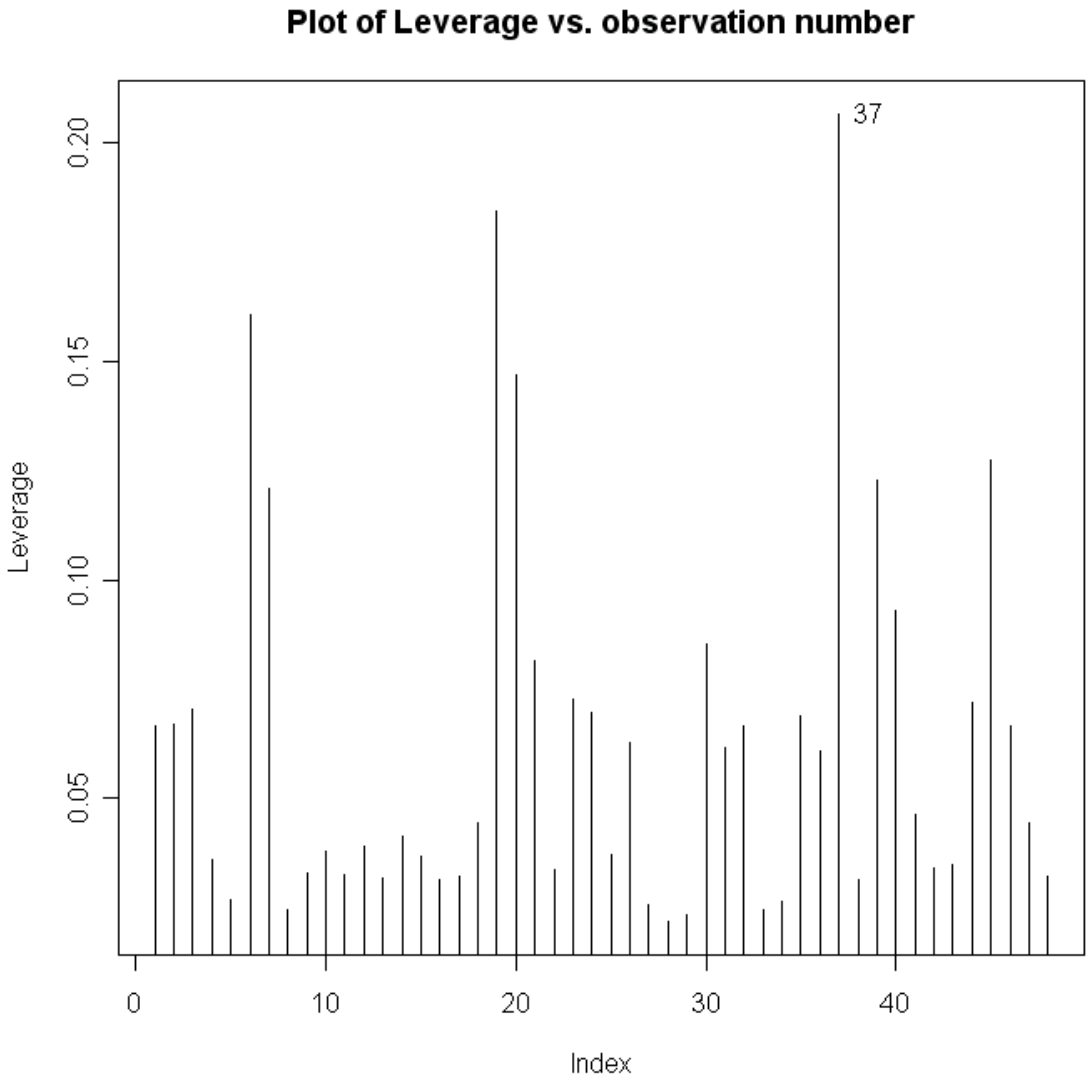
(b) How is an outlier detected? Explain why this is a reasonable criterion. Using this criterion, how many outliers would you expect in this set of data? How many cases are identified as outliers? (15 marks)

(c) What is meant by a case of high leverage in *multiple* regression? Which case has the highest leverage? Explain why the most extreme case has high leverage. (10 marks)

(d) What is meant by a case of high influence? Which case has the highest influence? Quote the value of Cook's distance. Explain why this case has high influence. (9 marks)

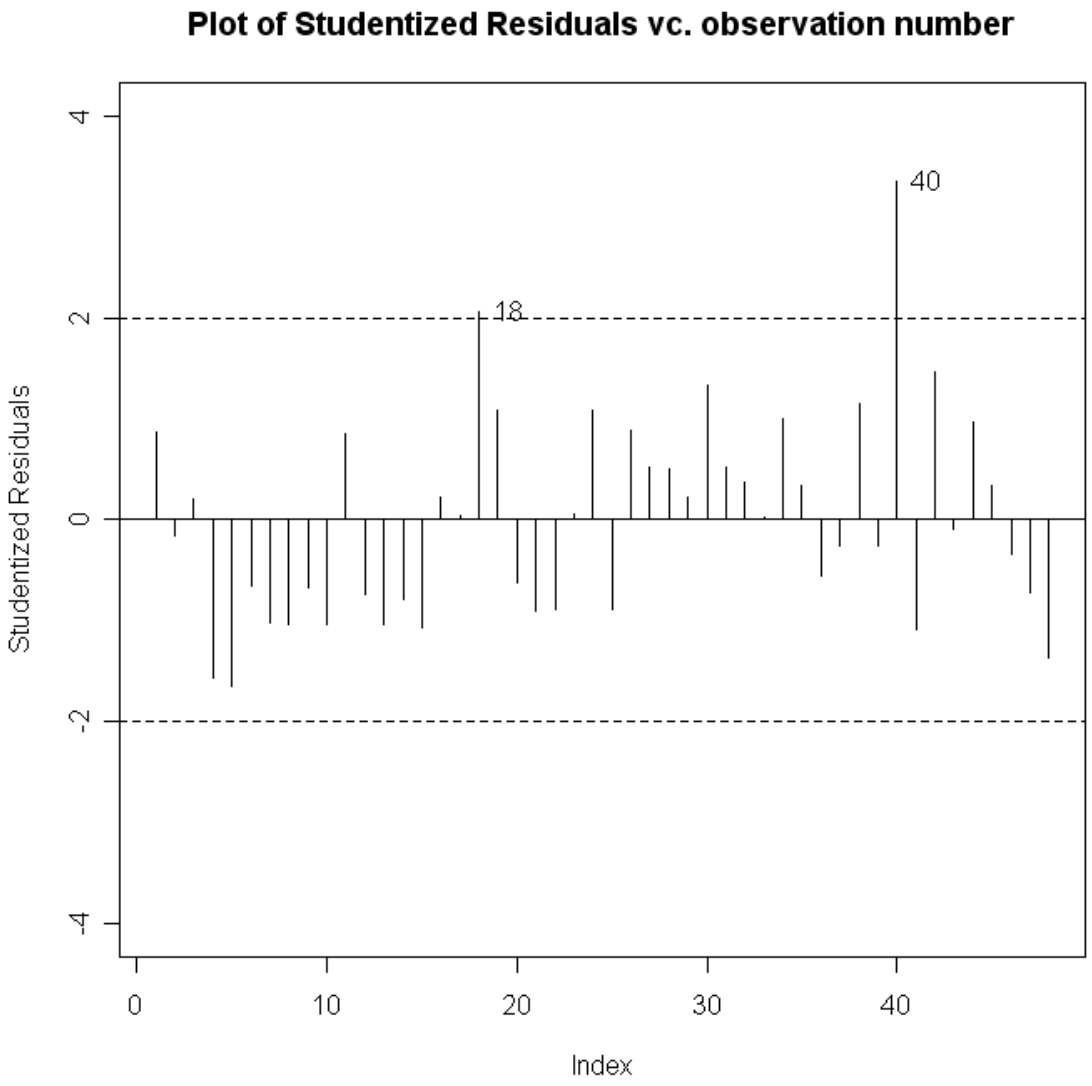
(e) Based on the above analyses, what recommendation(s) would you make for this model? Explain. (8 marks)

```
> # R code and output for Chapter 3

> # read fuel consumption data
> fuel.cons.df <-
+ read.table("P:\\ST2053\\fuel.txt",header=T)

> # regression model of FUEL on TAX and DLIC
> fuel.cons5.lm <- lm(FUEL~TAX+DLIC,data=fuel.cons.df)

> formula(fuel.cons5.lm)
FUEL ~ TAX + DLIC

> # calculate the leverages h
> h <- lm.influence(fuel.cons5.lm)$hat
>
> # plot of leverage vs. observation number
> plot(h,type="h",
+ main="Plot of Leverage vs. observation number",
+ ylab ="Leverage")
> identify(h,n=1)
[1] 37
```



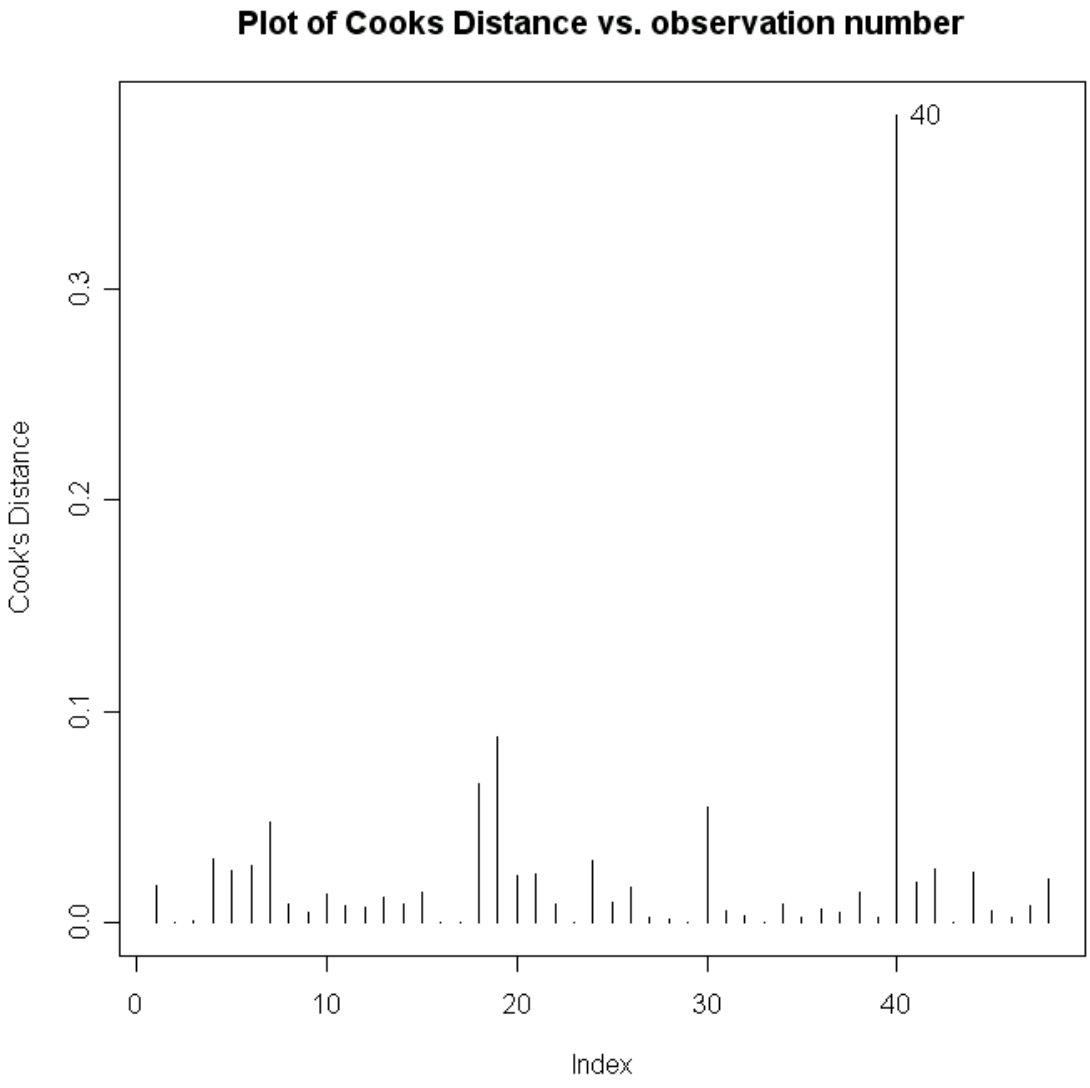

```
> # Texas(37) has high leverage
> # why does Texas(37) have high leverage in this model?

> fuel.cons.df[37,c("TAX","DLIC")]
   TAX DLIC
TX   5 56.6
> # in Texas, TAX = 5c/gallons, DLIC = 56.6%
> summary(fuel.cons.df$TAX)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.000   7.000   7.500   7.668   8.125  10.000
> summary(fuel.cons.df$DLIC)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  45.10   52.98   56.45   57.03   59.52   72.40
> # Texas has the lowest TAX (5c/gallon) of the 48 states;
> # The value of DLIC for Texas (56.6%) is about average
```

```
> # calculate the studentized residuals
> e <- resid(fuel.cons5.lm)
> s <- summary(fuel.cons5.lm)$sigma
> r <- e / (s*(1-h)^0.5)
> # plot of studentized residuals vs. observation number
> plot(r,type="h",
+ main="Plot of Studentized Residuals vs. observation
number",ylab ="Studentized Residuals",ylim=c(-4,4))
> abline(h=0,lty=1)
> abline(h=c(-2,2),lty=2)
> identify(r,n=2)
[1] 18 40
```




```
> # North Dakota(18) and Wyoming(40) are outliers,
> # i.e. have large studentized residuals
```

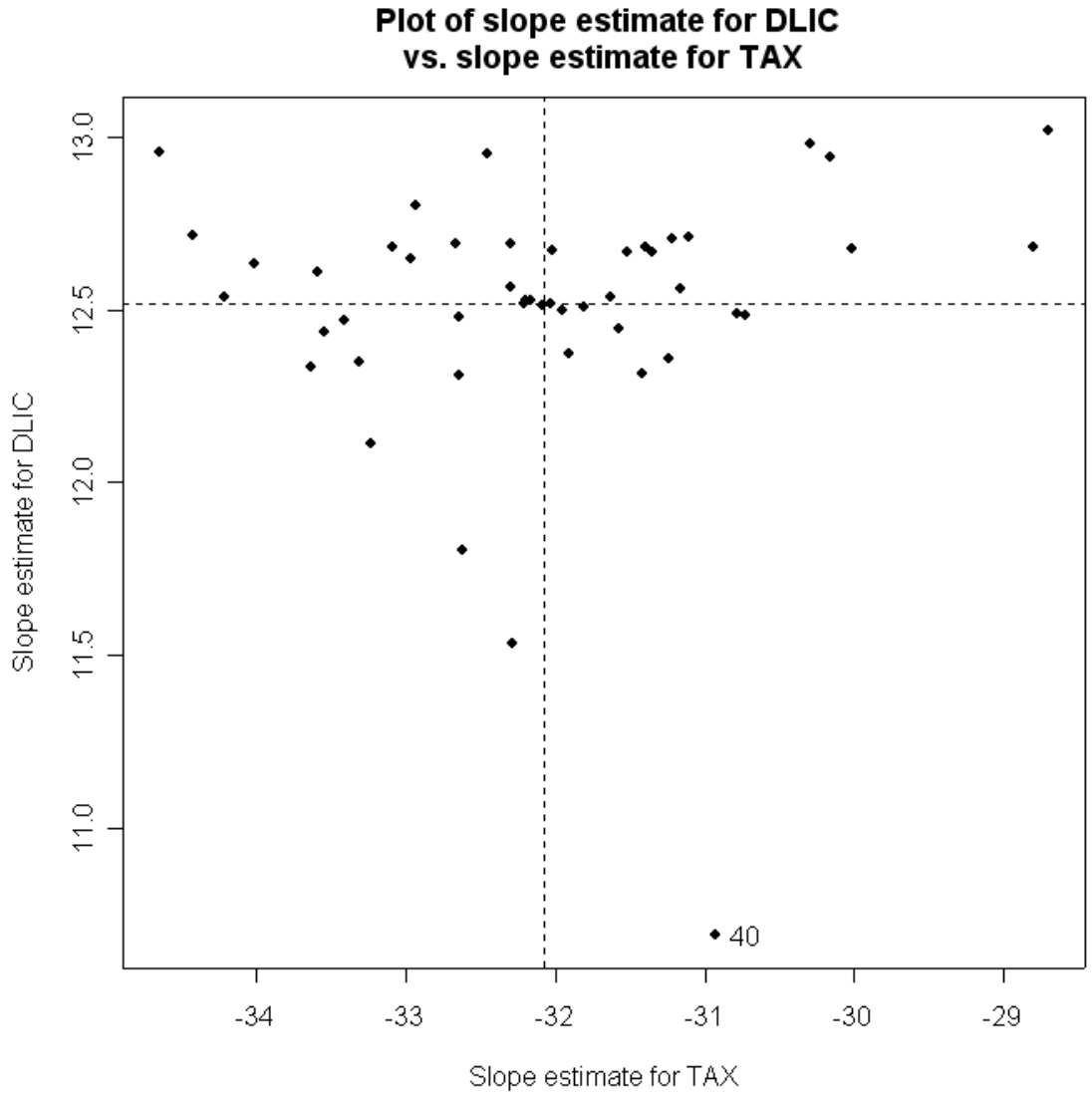
```
> # calculate the Cook's distance
> p <- length(coef(fuel.cons5.lm))
> d <- (1/p)*(h/(1 - h))*r^2
> # plot of Cook's distance vs. observation number
> plot(d,type="h",
+ main="Plot of Cook's Distance vs. observation number",
+ ylab="Cook's distance")
> identify(d,n=1)
[1] 40
```




```
> # Wyoming(40) has high influence;
```

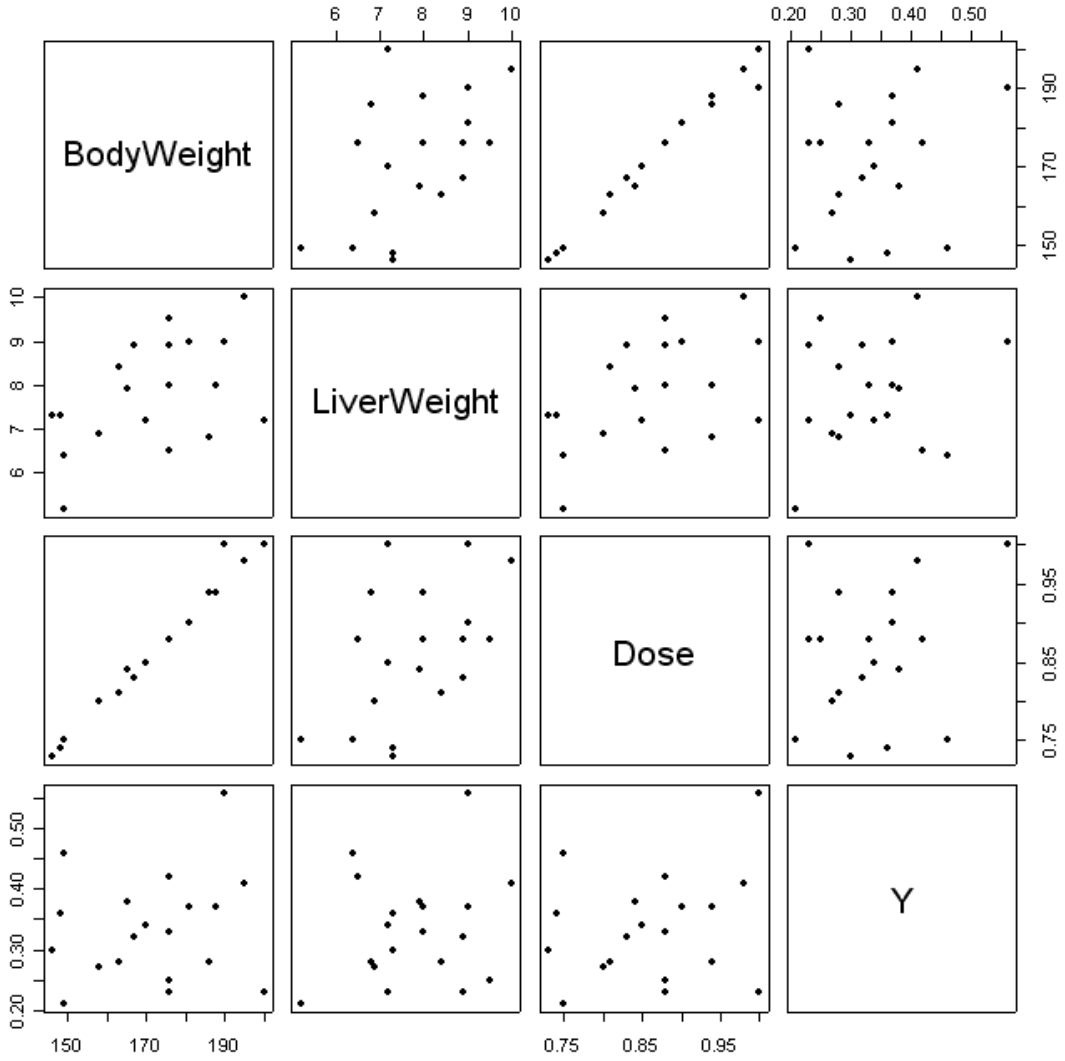
```
> # lm.influence(fuel.cons5.lm)$coefficients
> # is a (48x3) matrix containing the CHANGES in
> # the regression coefficients for each of the
> # 48 models with one case dropped
> coeffs.changes <-
lm.influence(fuel.cons5.lm)$coefficients
>
> # CHANGES in regression coefficients for
> # North Dakota(18),Texas(37),Wyoming(40)
> coeffs.changes[c(18,37,40),]
   (Intercept)       TAX        DLIC
ND    58.01803 -3.378348 -0.50426280
TX   -16.25973  1.473683  0.07915221
WY   -89.68060 -1.142211  1.82368759
> # Wyoming(40) seems to have high influence
>
> # regression coefficients for model with all 48 cases
> coef(fuel.cons5.lm)
(Intercept)         TAX        DLIC
  108.97087   -32.07532    12.51486

> coefficients <- t(coef(fuel.cons5.lm)-t(coeffs.changes))
> # coefficients is a (48x3) matrix containing
> # the regression coefficients for each of the
> # 48 models with one case dropped

> # regression coefficients for
> # North Dakota(18),Texas(37),Wyoming(40)
> coefficients[c(18,37,40),]
   (Intercept)       TAX     DLIC
ND    50.95284 -28.69697 13.01912
TX   125.23061 -33.54900 12.43571
WY   198.65147 -30.93311 10.69117
> # Wyoming(40) seems to have high influence
```

```
> # Plot of slope estimate for DLIC
> # vs. slope estimate for TAX
> # for 48 one-case-deleted models
> plot(coefficients[,2],coefficients[,3],
+ main=" Plot of slope estimate for DLIC
+ vs. slope estimate for TAX",
+ xlab="Slope estimate for TAX",
+ ylab="Slope estimate for DLIC",
+ pch=16)
> abline(v=coef(fuel.cons5.lm)[2],lty=2)
> abline(h=coef(fuel.cons5.lm)[3],lty=2)
> identify(coefficients[,2],coefficients[,3],n=1)
[1] 40
> # Wyoming(40) seems to have high influence
```



```
> # Rat data
>
> rats.df <- read.table("P:\\ST2053\\rats.txt",header=T)
> rats.df
   BodyWeight LiverWeight Dose    Y
1         176         6.5 0.88 0.42
2         176         9.5 0.88 0.25
3         190         9.0 1.00 0.56
4         176         8.9 0.88 0.23
5         200         7.2 1.00 0.23
6         167         8.9 0.83 0.32
7         188         8.0 0.94 0.37
8         195        10.0 0.98 0.41
9         176         8.0 0.88 0.33
10        165         7.9 0.84 0.38
11        158         6.9 0.80 0.27
12        148         7.3 0.74 0.36
13        149         5.2 0.75 0.21
14        163         8.4 0.81 0.28
15        170         7.2 0.85 0.34
16        186         6.8 0.94 0.28
17        146         7.3 0.73 0.30
18        181         9.0 0.90 0.37
19        149         6.4 0.75 0.46
>
> # attach(rats.df)
> attach(rats.df)
```

```
> # scatterplot matrix for rat data
> pairs(rats.df,pch=16)
```

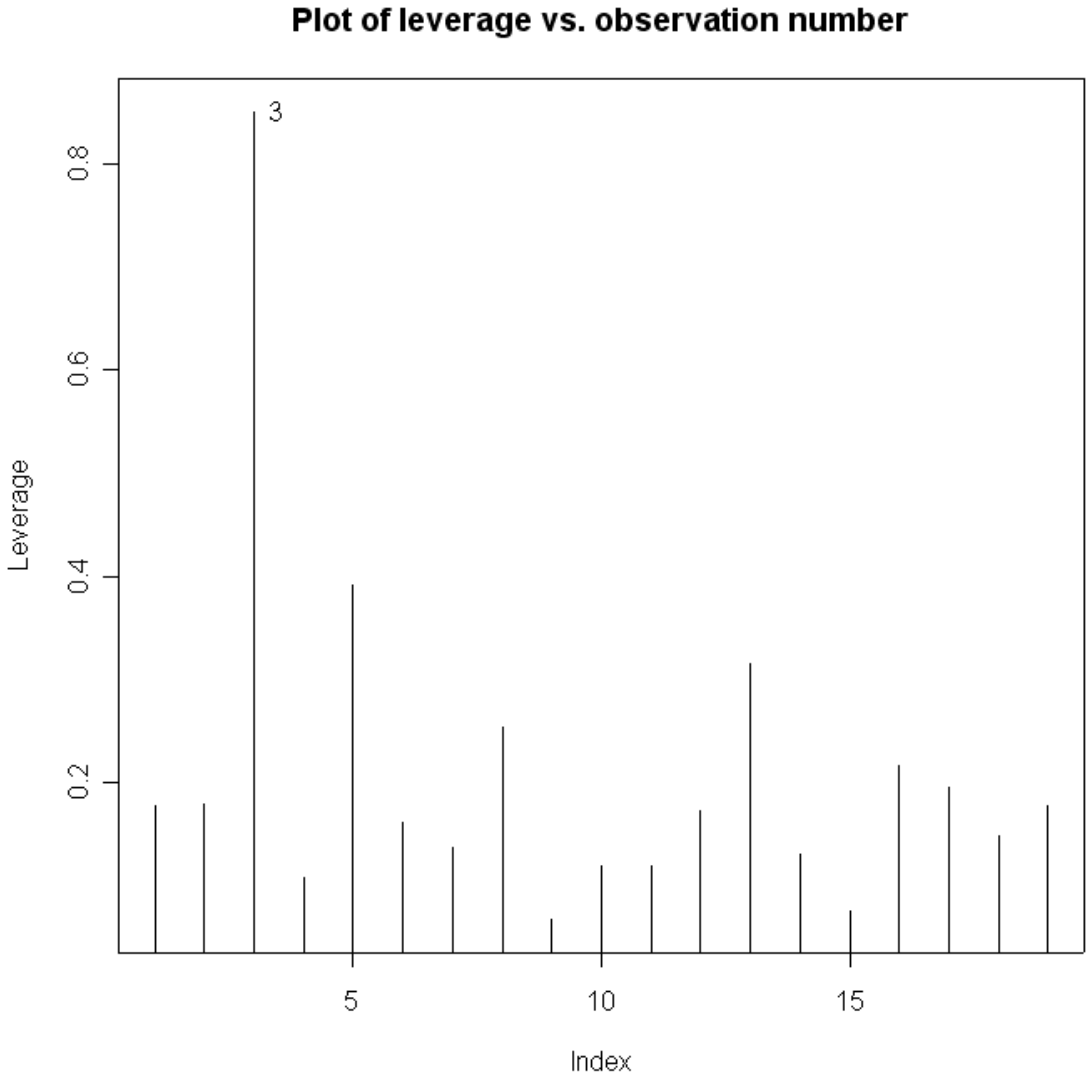



```
> # correlations for rat data
> cor(rats.df)
            BodyWeight LiverWeight      Dose         Y
BodyWeight   1.0000000   0.5000101 0.9902126 0.1510855
LiverWeight  0.5000101   1.0000000 0.4900711 0.2033302
Dose         0.9902126   0.4900711 1.0000000 0.2275436
Y            0.1510855   0.2033302 0.2275436 1.0000000
```

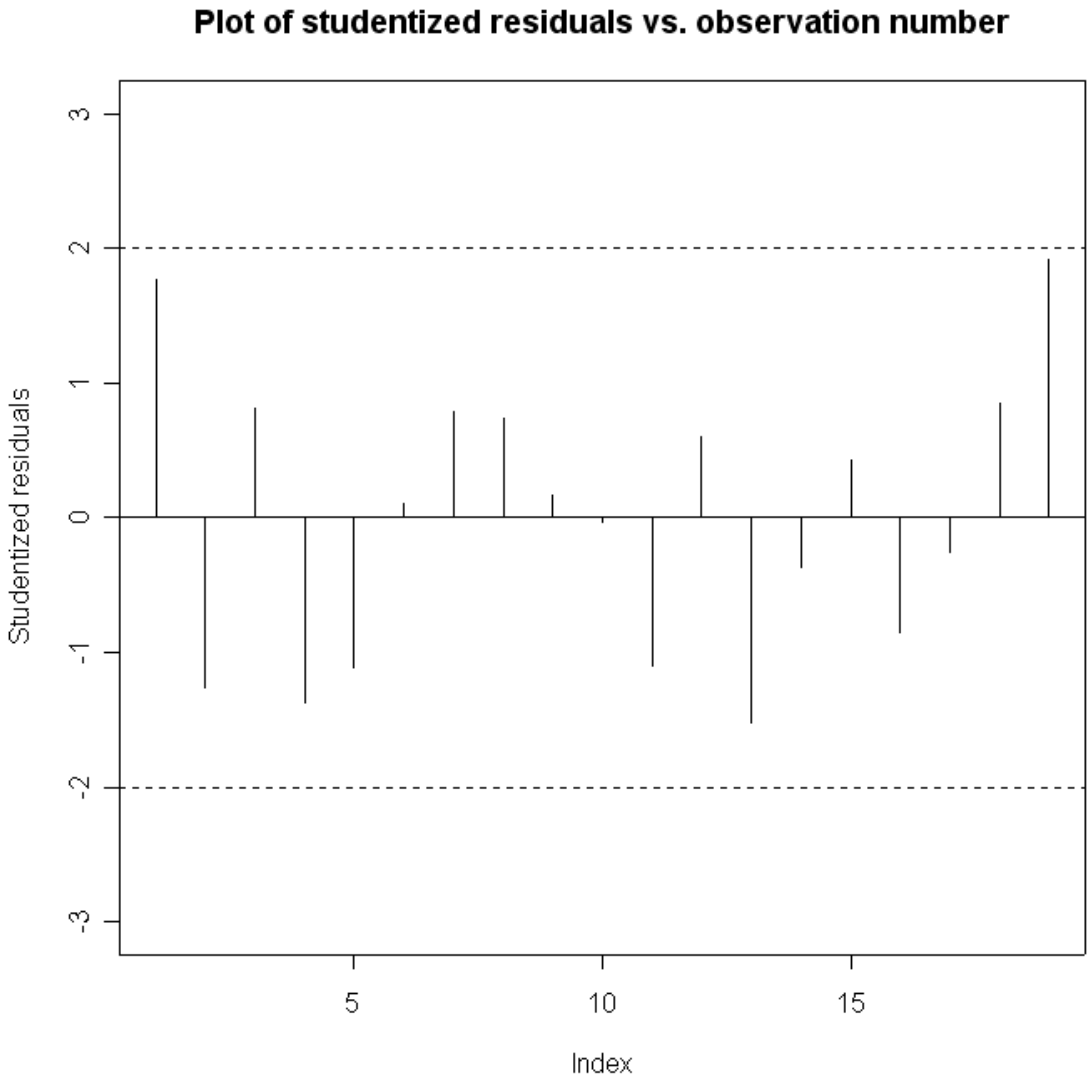
```
> # Regression models for rat data
>
> # regression model for Y on BodyWeight
> rats1.lm <-
+ lm( Y ~ BodyWeight,data = rats.df)
> summary(rats1.lm)

Call:
lm(formula = Y ~ BodyWeight, data = rats.df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.12834 -0.06065 -0.00889  0.04692  0.20976

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1962346  0.2215825   0.886    0.388
BodyWeight  0.0008105  0.0012862   0.630    0.537

Residual standard error: 0.08999 on 17 degrees of freedom
Multiple R-Squared: 0.02283,    Adjusted R-squared: -
0.03465
F-statistic: 0.3971 on 1 and 17 DF,  p-value: 0.537

>
> # regression model for Y on LiverWeight
> rats2.lm <-
+ lm( Y ~ LiverWeight ,data = rats.df)
> summary(rats2.lm)

Call:
lm(formula = Y ~ LiverWeight, data = rats.df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.12129 -0.05790 -0.00805  0.03739  0.20724

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.22037    0.13573   1.624    0.123
LiverWeight  0.01471    0.01718   0.856    0.404

Residual standard error: 0.08913 on 17 degrees of freedom
Multiple R-Squared: 0.04134,    Adjusted R-squared: -
0.01505
F-statistic: 0.7331 on 1 and 17 DF,  p-value: 0.4038
```

```
> # regression model for Y on Dose
> rats3.lm <-
+ lm( Y ~ Dose ,data = rats.df)
> summary(rats3.lm)

Call:
lm(formula = Y ~ Dose, data = rats.df)
Residuals:
     Min       1Q   Median       3Q      Max
-0.13761 -0.06212 -0.00427  0.04850  0.19239

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1330     0.2109   0.631    0.537
Dose          0.2346     0.2435   0.963    0.349

Residual standard error: 0.08864 on 17 degrees of freedom
Multiple R-Squared: 0.05178,    Adjusted R-squared: -
0.004002
F-statistic: 0.9283 on 1 and 17 DF,  p-value: 0.3488

> # regression model for Y on BodyWeight,
> # LiverWeight and Dose
> rats4.lm <-
+ lm( Y ~ BodyWeight + LiverWeight + Dose,
+ data = rats.df)
> summary(rats4.lm)
Call:
lm(formula = Y ~ BodyWeight + LiverWeight + Dose, data =
rats.df)
Residuals:
      Min        1Q    Median        3Q       Max
-0.100557 -0.063233  0.007131  0.045971  0.134691

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.265922   0.194585   1.367   0.1919
BodyWeight  -0.021246   0.007974  -2.664   0.0177 *
LiverWeight  0.014298   0.017217   0.830   0.4193
Dose         4.178111   1.522625   2.744   0.0151 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 0.07729 on 15 degrees of freedom
Multiple R-Squared: 0.3639,     Adjusted R-squared: 0.2367
F-statistic:  2.86 on 3 and 15 DF,  p-value: 0.07197
```

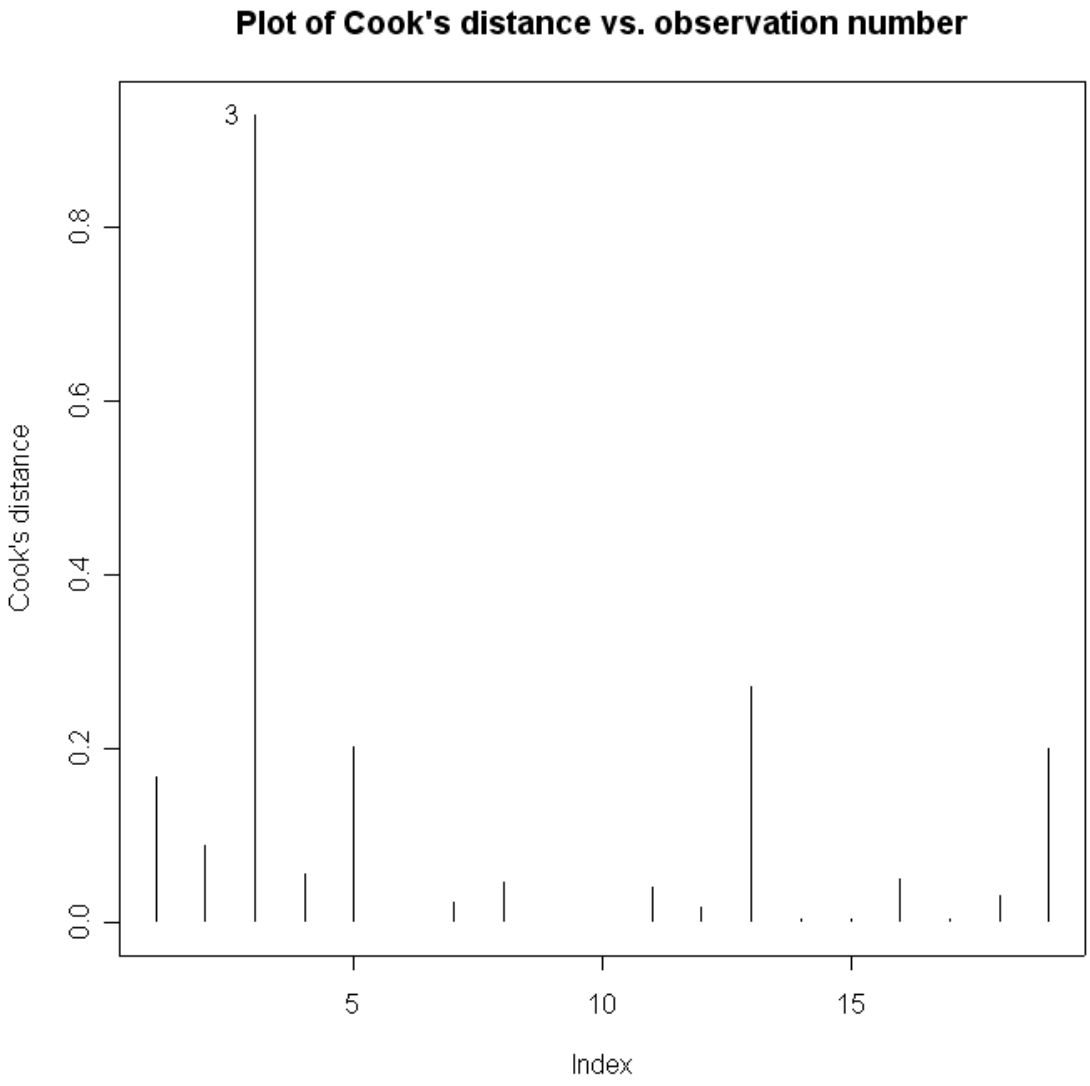
```
> # regression diagnostics for rat data
>
> # calculate the leverages h
> h <- lm.influence(rats4.lm)$hat
>
> # plot of leverage vs. observation number

> plot(h,type="h",
+ main="Plot of leverage vs. observation number",
+ ylab ="Leverage")
> identify(h,n=1)
[1] 3
> # case 3 has high leverage
```



```
# calculate the studentized residuals
e <- resid(rats4.lm)
s <- summary(rats4.lm)$sigma
r <- e / (s*(1-h)^0.5)

> # plot of studentized residuals vs. observation number
> plot(r,type="h",main="Plot of studentized residuals vs.
observation number",
> ylab ="Studentized residuals",ylim=c(-3,3))
> abline(h=0,lty=1)
> abline(h=c(-2,2),lty=2)
# there are no outliers
```



```
> # calculate the Cook's distances
> p <- length(coef(rats4.lm))
> d <- (1/p)*(h/(1-h))*r^2
> # plot of Cook's distance vs. observation number

> # see Table 5.6, p.123
> plot(d,type="h",
+ main="Plot of Cook's distance vs. observation number",
+ ylab="Cook's distance")
> identify(d,n=1)
[1] 3
> # case 3 has high influence
```



```
> # Table of diagnostic statistics
> tableDiag.df <- data.frame(Y,e,r,h,d)
> tableDiag.df
      Y            e           r          h            d
1  0.42  0.123758021  1.76604714 0.17798270 1.688268e-01
2  0.25 -0.089136157 -1.27303970 0.17934099 8.854024e-02
3  0.56  0.024088214  0.80715401 0.85091457 9.296160e-01
4  0.23 -0.100557321 -1.37723226 0.10761585 5.718456e-02
5  0.23 -0.067711915 -1.12309856 0.39153825 2.029162e-01
6  0.32  0.007131221  0.10073807 0.16115958 4.874208e-04
7  0.37  0.056580285  0.78795102 0.13688107 2.461564e-02
8  0.41  0.049584065  0.74258710 0.25367448 4.685795e-02
9  0.33  0.012310932  0.16490096 0.06701578 4.883028e-04
10 0.38 -0.002844507 -0.03922453 0.11968672 5.229549e-05
11 0.27 -0.080146346 -1.10507028 0.11950583 4.143644e-02
12 0.36  0.042357750  0.60240838 0.17239599 1.889847e-02
13 0.21 -0.098151103 -1.53565893 0.31618336 2.726019e-01
14 0.28 -0.027142867 -0.37680527 0.13140699 5.370022e-03
15 0.34  0.031614703  0.42556276 0.07617481 3.733265e-03
16 0.28 -0.058754717 -0.85886404 0.21661460 5.099189e-02
17 0.30 -0.018353809 -0.26470261 0.19522441 4.249284e-03
18 0.37  0.060682327  0.85093422 0.14872221 3.162543e-02
19 0.46  0.134691226  1.92204135 0.17796183 1.999403e-01
```

```
> # regression model for Y on BodyWeight,
> # LiverWeight and Dose, with case 3 deleted
>
> rats5.lm <-
+ lm( Y ~ BodyWeight + LiverWeight + Dose,
+ data = rats.df, subset = -3)
> summary(rats5.lm)

Call:
lm(formula = Y ~ BodyWeight + LiverWeight + Dose, data =
rats.df,
    subset = -3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.102154 -0.056486  0.002838  0.046519  0.137059

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.311427   0.205094   1.518    0.151
BodyWeight  -0.007783   0.018717  -0.416    0.684
LiverWeight  0.008989   0.018659   0.482    0.637
Dose         1.484877   3.713064   0.400    0.695

Residual standard error: 0.07825 on 14 degrees of freedom
Multiple R-Squared: 0.02106,    Adjusted R-squared: -0.1887
F-statistic: 0.1004 on 3 and 14 DF,  p-value: 0.9585


> # quit R
> q("yes")
```