Advanced Diagnostic Measurement in MLR

Why do we perform diagnostics?

Looking for,

- Outliers
- Evidence of a non-normal error distribution
- Evidence of non-independence in the errors
- Evidence of a disproportionate influence by one or more individual data points
- Evidence of multicollinearity

Previously we have used:

- Plots of residuals against predicted values for multiple symptoms
- Plots of residuals against independent variables for functional relationship
- Plots of residuals against time, collection order for independency checking
- Normal quantile plots of residuals for Normality
- Histograms of residuals for Normality, outliers
- Plots of independent variables against each other for multicollinearity

However, these plots have a limitation in that they do not reveal the marginal effect of a predictor variable when other variables are present.

To overcome this limitation, we require more advanced models and specialized tools.

- Added-variable plots to observe the marginal effect of a predictor variable
- Studentized deleted residuals, hat matrix diagonals to observe the outliers
- Cook's D, DFFITS, and DFBETAS to detect influential points
- Variance inflation factor (VIF) and tolerance to diagnose multicollinearity

These tools can provide more insights into the relationships between variables and improve the accuracy of our model.

Added-variable plot

Added-variable plots, also called *partial regression plots* or *adjusted variable plots*, are refined residual plots that provide graphic information about the marginal importance of a predictor variable X_k, given the other predictors are already in.

Regress Y on X2 $\hat{Y}_{\mathrm{i}}(\mathrm{X2}) = \mathrm{bo} + \mathrm{b2X_{i2}}$

$$e_i(\hat{Y}_{\mathrm{i}}|\mathrm{X}_2) = \mathrm{Y_i} \ -\hat{Y}_i(\mathrm{X}_2)$$

Regress X1 on X2 $\hat{X}_{\mathrm{ii}}(\mathrm{X2}) = \mathrm{bo} + \mathrm{b2X_{i2}}$

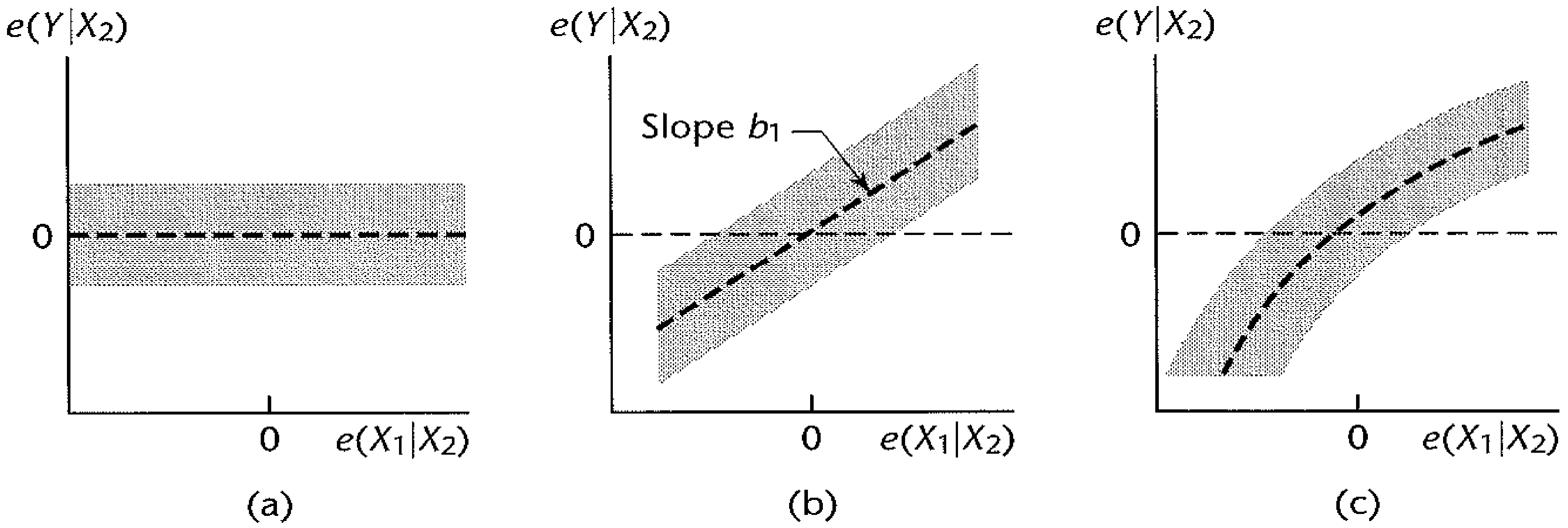
$$\mathrm{e_i}(\mathrm{X}_1|\mathrm{X}_2) = \mathrm{X_{i1}} - \hat{X}_{i1}(\mathrm{X}_2)$$

- The added-variable plot for X1 is a plot of the $e_i(\hat{Y}_{\mathrm{i}}|\mathrm{X}_2)$ aganist $\mathrm{e_i}(\mathrm{X}_1|\mathrm{X}_2)$ which shows the relationship between the residual error of the response variable Y and the residual error of the explanatory variable X1, while holding all other explanatory variables $(X_2, X_3, \ldots)$constant.

- The plot can help identify any patterns or trends in the relationship between the ***marginal effect*** of X1 and Y, which may not be apparent from simple scatter plots or other graphical tools.

- It can also help evaluate the validity of the linear regression model assumptions, such as linearity, independence, and homoscedasticity.

Added-variable plot prototypes

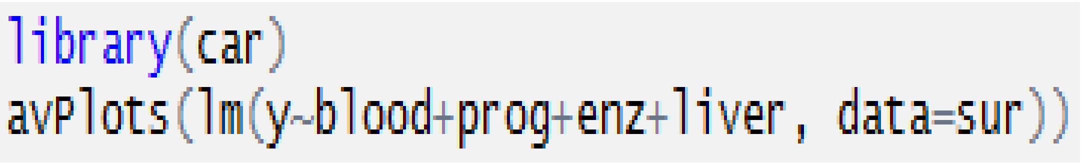


Plot (a) indicates that X1 contain no additional information useful for predicting Y beyond that contained in X2

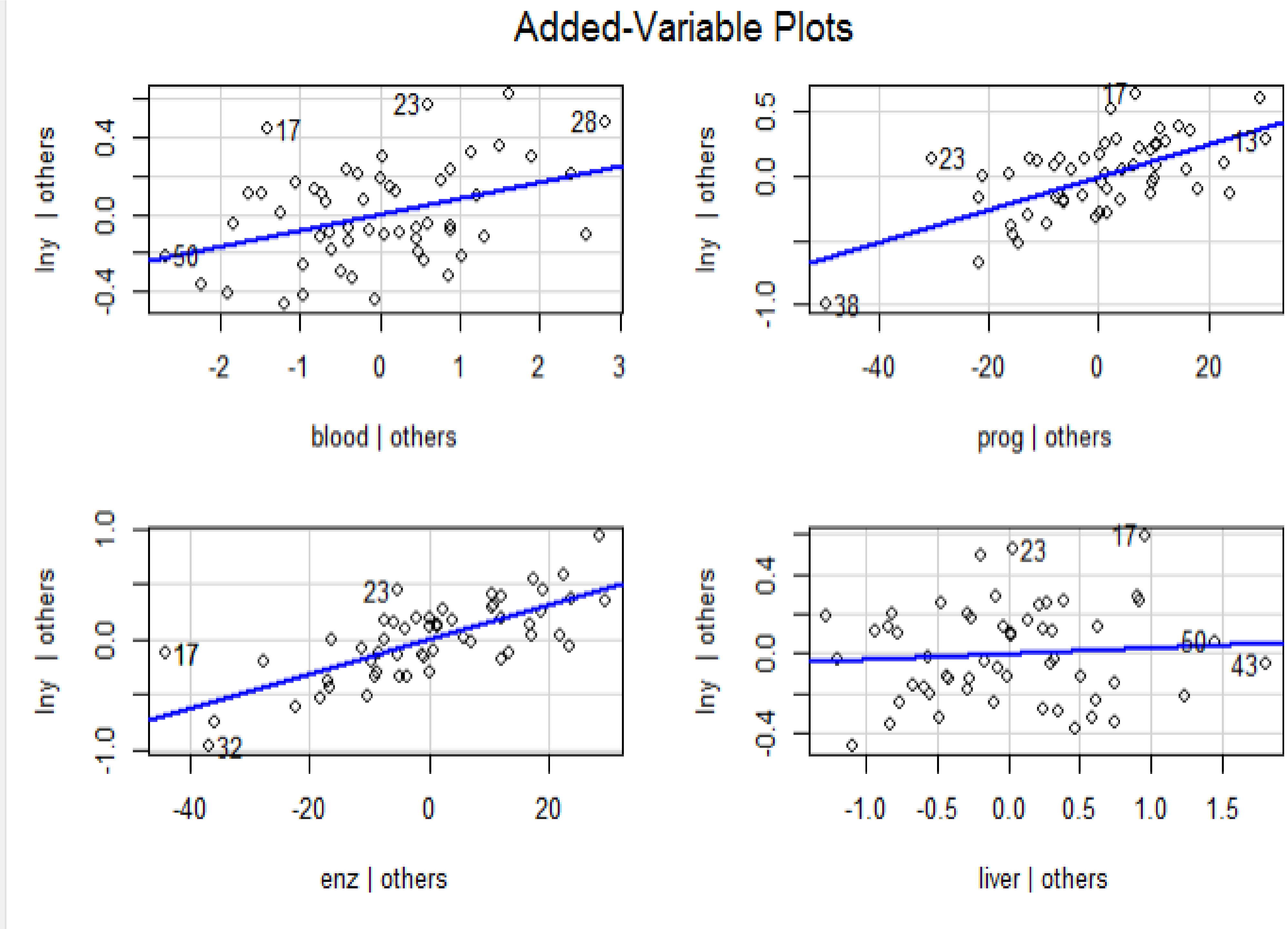
Plot (b) indicates that a linear term in X1 may be a helpful addition to the regression model already containing X2.

Plot (c) indicates that a nonlinear term in X1 may be a helpful addition to the regression model already containing X2.

Added-variable plots (surgical example)



```
library(car)
avPlots(lm(y~blood+prog+enz+liver, data=sur))
```



- The added-on effect can be seen on Prog, Enz, Blood
- The first order seems sufficient
- Liver doesn’t show added-on effect

Comments

- Added-variable plots need to be used with caution. They may not show the proper form of the marginal effect if the functional relations for some of the predictor variables are miss specified.

E.g., if X1 and X2 are related in a curvilinear fashion to Y but the regression model is linear term only

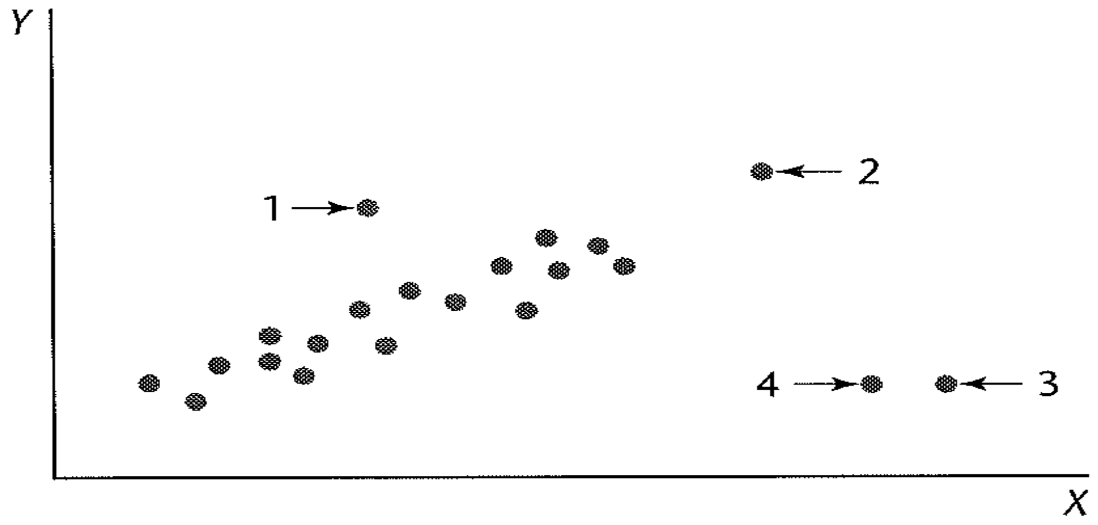
$\mathrm{e_i}(\mathrm{X_1}|\mathrm{X_2}) \neq \mathrm{X_{i1}} - \hat{X}_{i1}(\mathrm{X_2})$

- High multicollinearity among the predictor variables may cause the added-variable plots to show an improper functional relation for the marginal effect of a predictor.

$e_i(\hat{Y}_\mathrm{i}|\mathrm{X_2}) \neq \mathrm{Y_i} - \hat{Y}_i(\mathrm{X_2})$

Identifying outliers in MLR

- In SLR, we can identify outliers by means of boxplots, scatter plots, residual plots etc.
- In MLR, it is difficult to identify outliers by simple graphic means because it might not be extreme in a multiple regression model anymore.
- We now discuss the use of some refined measures for identifying outliers
  - ➢ Residuals and semi-studentized residuals (outlying Y observation)
  - ➢ Studentized deleted residuals (outlying Y observation)
  - ➢ Hat Matrix (outlying X observation)
  - ➢ Identifying influential cases (DFFITS, Cook’s distance and DFBETAS measures)



A simple review on the Hat matrix and residuals

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

$$\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$$

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

$$\boldsymbol{\sigma}^2\{\boldsymbol{e}\} = \boldsymbol{\sigma}^2\ (\boldsymbol{I} - \boldsymbol{H})$$

- The hat matrix (H) is a matrix that is used to compute the predicted values (y-hat) for the dependent variable based on the observed values of the independent variables.
- It can be interpreted as a matrix that projects the observed values of the dependent variable (y) onto the predicted values (y-hat) based on the observed values of the independent variables (x)

The variance of residual e_i is:

$$\boldsymbol{\sigma}^2\{\boldsymbol{e}_i\} = \boldsymbol{\sigma}^2\ (\boldsymbol{1} - \boldsymbol{h}_{ii}) \xrightarrow{\text{Estimated by}} \boldsymbol{s}^2\{\boldsymbol{e}_i\} = \boldsymbol{MSE}\ (\boldsymbol{1} - \boldsymbol{h}_{ii})$$

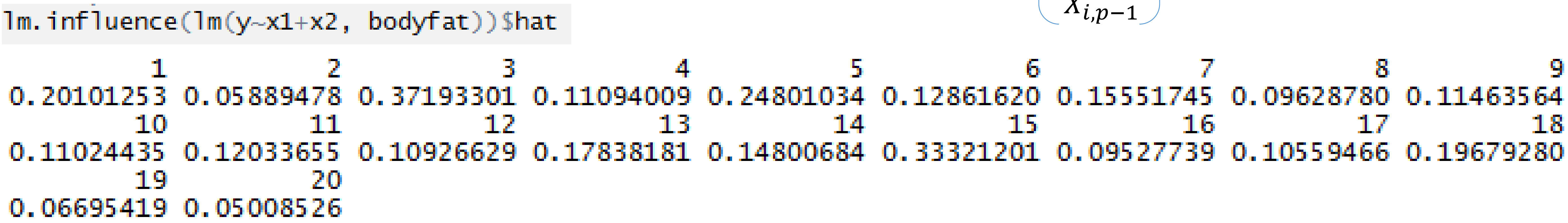
The covariance between e_i and e_j is:

$$\boldsymbol{\sigma}^2\{\boldsymbol{e}_i, \boldsymbol{e}_j\} = \boldsymbol{\sigma}^2\ \left(\boldsymbol{0} - \boldsymbol{h}_{ij}\right) = -\boldsymbol{h}_{ij}\boldsymbol{\sigma}^2 \longrightarrow \boldsymbol{s}^2\{\boldsymbol{e}_i, \boldsymbol{e}_j\} = -\boldsymbol{h}_{ij}(\boldsymbol{MSE})$$

Comment:

$$\boldsymbol{h}_{ii} = \boldsymbol{X}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_i \qquad \boldsymbol{Where}\ \boldsymbol{X}_i = \begin{pmatrix} 1 \\ X_{i,1} \\ X_{i,2} \\ . \\ . \\ . \\ X_{i,p-1} \end{pmatrix}$$

$i.e., p = 3$



```
lm.influence(lm(y~x1+x2, bodyfat))$hat
```

```
         1          2          3          4          5          6          7          8          9 
0.20101253 0.05889478 0.37193301 0.11094009 0.24801034 0.12861620 0.15551745 0.09628780 0.11463564 
        10         11         12         13         14         15         16         17         18 
0.11024435 0.12033655 0.10926629 0.17838181 0.14800684 0.33321201 0.09527739 0.10559466 0.19679280 
        19         20 
0.06695419 0.05008526 
```

Studentized residuals (r_i)

Studentized residuals: $r_i = \frac{e_i}{s\{e_i\}}$, where s$\{e_i\} = \sqrt{MSE\ (1 - h_{ii})}$

The body fat example

To compute r_3 for case 3

$e_3 = -3.176$ $h_{33} = 0.372$

$\mathrm{MSE} = 6.4677$

$$r_i = \frac{e_i}{s\{e_i\}} = -\frac{3.176}{\sqrt{MSE\ (1 - h_{ii})}} = -1.576$$

- This means that the residual for case 3 is 1.576 standard deviations smaller than what would be expected based on the overall distribution of residuals.
- Studentized residuals mainly measures the outliers in the Y scale.
- When the Normal assumption is met and a significant value of 0.05 is used, a case with $|r_i| > 2$ could be considered an outlier for a reasonably large sample.

Studentized deleted residual (t_i)

$$d_i = Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1 - h_{ii}}$$

$$s^2\{d_i\} = \frac{MSE_{(i)}}{1 - h_{ii}}$$

$$t_i = \frac{d_i}{s\{d_i\}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} = e_i\left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}\right]^{1/2}$$

- $\hat{Y}_{i(i)}$ *and* $MSE_{(i)}$are computed from data set after the ith case is deleted in fitting the LM.
- The larger is the value h_{ii}, the larger will be the deleted residual as compared to the ordinary residual of e_i
- The studentized deleted residual mainly measure the outliers in the Y scale. It is also considered a balanced measurement for other outlying situation such as the impact on the estimation of regression coefficients.
- Like the studentized residual (r_i), a value of 2 is often used as a threshold to determine outliers with the studentized deleted residual (t_i) when the sample size is large. A more precise method is the Bonferroni procedure, especially for small sample size or when the Normality assumption is violated.

The Bonferroni Procedure to determine Y Outliers with the Studentized deleted residual $(\alpha = 0.1, \mathrm{n} = 20, \mathrm{p} = 3)$

Ho: Case i is not outlying in Y-scale

Ha: Case i is outlying in Y-scale

$$t_i = \frac{d_i}{s\{d_i\}} \sim t(n-p-1)$$

With Bonferroni procedure: $g = n$

Bonferroni critical value = $t(1 - \frac{\alpha}{2n}; n-1-p)$

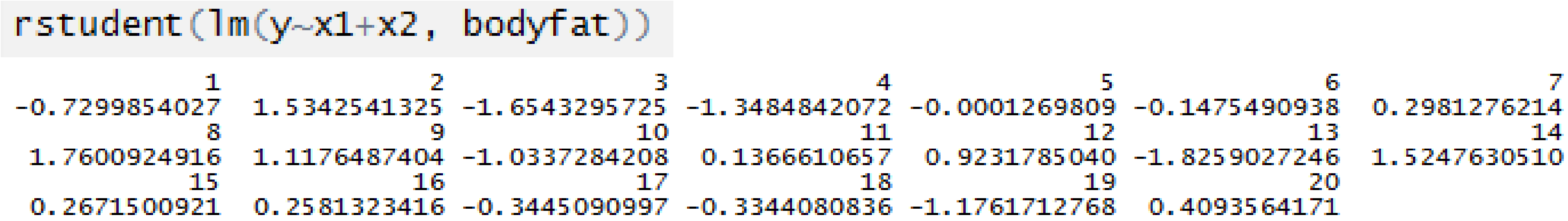
= $t(0.9975; 16) = 3.252$

The body fat example $Y = -17.174 + 0.2224X1 + 0.6594X2 + \epsilon$

To test if case 1 ($\mathrm{X}_{11} = 19.5, X_{12} = 43.1$) is an outlier: $e_1 = -1.683$ $h_{11} = 0.201$ $SSE = 195.9508$

$$t_1 = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}\right]^{1/2} = -1.683 \left[\frac{20-3-1}{195.9508(1-0.201) - (-1.683)^2}\right]^{\frac{1}{2}} = -0.73$$

Since $|t_1| = 0.73 < 3.252$, we conclude that case1 is not an outlier.



```
rstudent(lm(y~x1+x2, bodyfat))
```

```
           1            2            3            4            5            6            7 
-0.7299854027  1.5342541325 -1.6543295725 -1.3484842072 -0.0001269809 -0.1475490938  0.2981276214 
           8            9           10           11           12           13           14 
 1.7600924916  1.1176487404 -1.0337284208  0.1366610657  0.9231785040 -1.8259027246  1.5247630510 
          15           16           17           18           19           20 
 0.2671500921  0.2581323416 -0.3445090997 -0.3344080836 -1.1761712768  0.4093564171 
```

Comment

In addition to outliers, large Studentized deleted residuals can be caused by a non-normal error distribution or non-constant variance.

If the data for a potential outlier does not have any obvious problems, consider transforming Y.

Identifying X Outliers with Hat Matrix leverage values

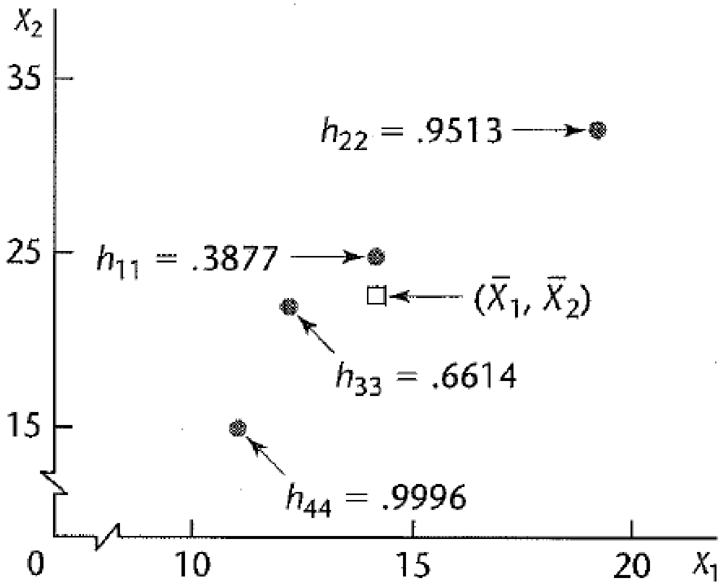
- The hat matrix projects Y to Y-hat based on the value of X: $\hat{Y} = HY$

- The i^{th} diagonal element, h_{ii} is called the leverage of the i^{th} x value has on the predicted values of Y.

- The larger is h_{ii}, the smaller is the variance of the residual e_i: $\sigma^2\{e_i\} = \sigma^2\ (1 - h_{ii})$

- Observations with extreme values on one or more X values tend to have large h_{ii}

$$h_{ii} = X_i'(X'X)^{-1}X_i$$

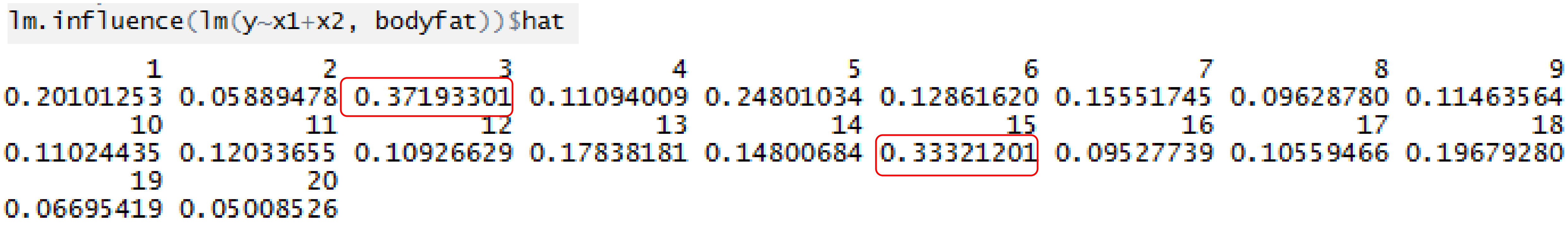


- h_{ii} is considered *large if* it is more than twice as large as the mean leverage value $\bar{h} = \text{p}/\text{n}$,

$$h_{ii} > 2p/n$$

In the body fat example, $\bar{h} = \dfrac{\Sigma h_{ii}}{n} = \dfrac{p}{n} = \dfrac{3}{20} = 0.15$

Hence, any case with a $h_{ii} > 2(0.15) = 0.3$ is considered outlying in term of their X values.



```
lm.influence(lm(y~x1+x2, bodyfat))$hat
```

```
         1          2          3          4          5          6          7          8          9 
0.20101253 0.05889478 0.37193301 0.11094009 0.24801034 0.12861620 0.15551745 0.09628780 0.11463564 
        10         11         12         13         14         15         16         17         18 
0.11024435 0.12033655 0.10926629 0.17838181 0.14800684 0.33321201 0.09527739 0.10559466 0.19679280 
        19         20 
0.06695419 0.05008526 
```

Comments

- High leverage does not necessarily mean that an observation is an outlier in the y-scale (the space of the dependent variable).

- If the dataset has small n or large p, a lower cutoff value may be appropriate; if the dataset has a large n or small p, a higher cutoff value may be appropriate. (0.2-0.5)

- For observations with very high leverage, examine the pattern of leverage values across the independent variables to determine which independent variable(s) may be driving the high leverage values.
  - ➢ If multiple observations have high leverage values that are spread out across multiple independent variables, it may be an indication of collinearity in the dataset.

Identifying influential cases—DFFITS, Cook's Distance, and DFBETAS Measures

A case is ***influential*** if its exclusion causes major changes in the fitted regression function: either on the coefficients or the fitted values.

We take up three measures of influence that are widely used in practice, each based on the omission of a single case to measure its influence.

Influence on Single Fitted Value--DFFITS

$$(DFFITS)_i = \frac{\left(\hat{Y}_i - \hat{Y}_{i(i)}\right)}{\sqrt{MSE_{(i)}h_{ii}}}$$

It represents the estimated standard deviation of the fitted value increases or decrease with the inclusion ($\hat{Y}_i$) or exclusion ($\hat{Y}_{i(i)}$)of the ith case.

I.e., the influence on prediction

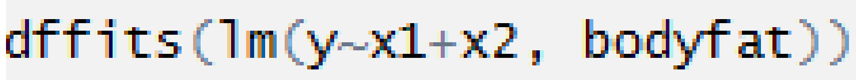
$$(DFFITS)_i = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}\right]^{\frac{1}{2}} \left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}} = t_i \left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}}$$

A guideline for identifying influential cases is that

$|(DFFITS)_i| > 1$ for small to medium data set, and $> 2\sqrt{\frac{p}{n}}$ for large data sets.

In the body fat example, $t_3 = -1.656, h_{33} = 0.372$

$$(DFFITS)_i = t_i \left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}} = -1.656\left(\frac{0.372}{1-0.372}\right)^{\frac{1}{2}} = -1.27$$



```
dffits(lm(y~x1+x2, bodyfat))
```

Since 1.27 is somewhat large than 1, but since it is not too far greater than 1, the case may not be influential enough to require remedial action

Influence on All Fitted Value—Cook's distance

$$D_i = \frac{\Sigma\left(\hat{Y}_i - \hat{Y}_{i(i)}\right)^2}{pMSE} \qquad = \frac{e_i^2}{pMSE}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right)$$

- An aggregate measure represents the influence of the *i*th case on all n fitted values

- For interpreting Cook's distance measure, relate D_i to the $F(p, n-p)$ distribution, and ascertain the corresponding percentile value.

If the percentile < 20%, the ith case has no influence on the fitted values

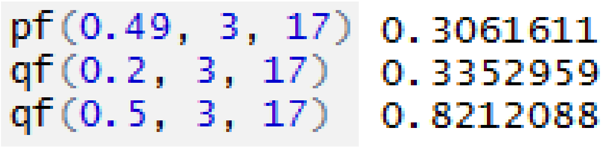
If the percentile is between 20% and 50%, the ith case has minor influence on the fitted values

If the percentile > 50%, the ith case has major influence on the fitted values

- Large influence depending on e_i and h_{ii}

In the body fat example, $e_3 = -3.176, h_{33} = 0.372$

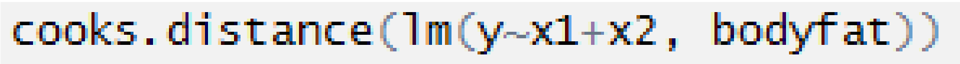
$$D_3 = \frac{e_i^2}{pMSE}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right) = \frac{(-3.176)^2}{3(6.47)}\left(\frac{0.372}{(1-0.372)^2}\right) = 0.49$$



```
pf(0.49, 3, 17)  0.3061611
qf(0.2, 3, 17)   0.3352959
qf(0.5, 3, 17)   0.8212088
```

- Case 3 with a D=0.49, is the 30.6% percentile of the distribution, hence has a minor influence on the fitted values.
- In general, case with $D_i < 0.33$ has no influence on the fitted values and,

a case with $D_i > 0.82$ has major influence on the fitted values.



```
cooks.distance(lm(y~x1+x2, bodyfat))
```

Influence on the Regression Coefficients—DFBETAS

"S" means "standardized".

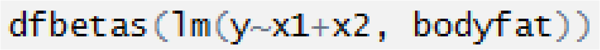
$$(DFBETAS)_{k(i)} = \frac{(b_k - b_{k(i)})}{\sqrt{MSE_{(i)}c_{kk}}}\,, \qquad k = 0,1, \ldots p-1$$

- A measure of the influence of the *i*th case on all each regression coefficients.

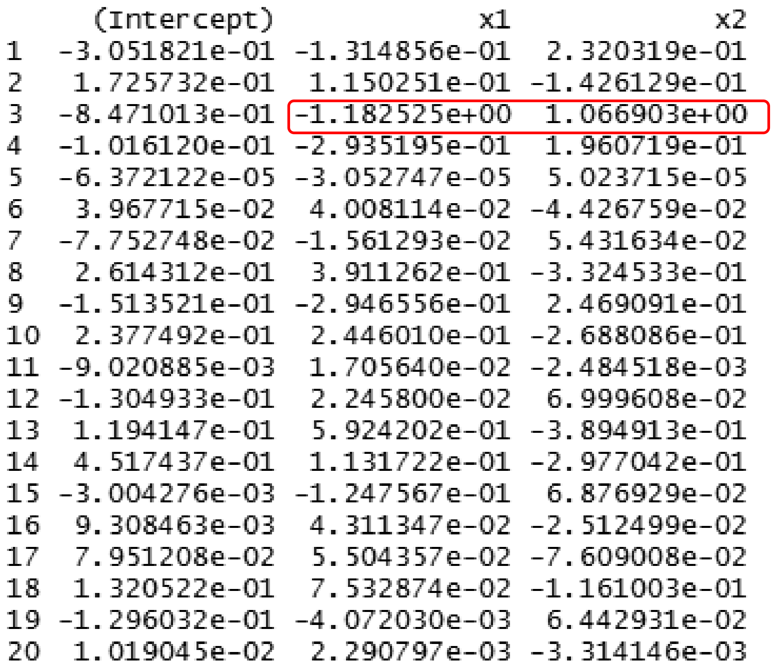
- c_{kk} is the kth diagonal element of $(X'X)^{-1}$.
- $\sigma^2\{b\} = \sigma^2(X'X)^{-1}$, hence $\sigma^2\{b_k\} = \sigma^2 c_{kk}$ $\quad (estimated\ by\ MSE_{(i)}c_{kk})$

- Large value of $(DFBETAS)_{k(i)}$ is indicative of a large impact of the ith case on the kth regression coefficient. A large value means

  $(DFBETAS)_{k(i)} > 1$ for small to medium data set and $> 2/\sqrt{n}$ for large data set.

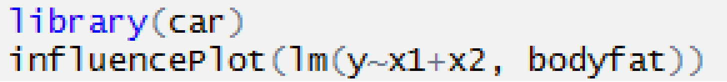


```
dfbetas(lm(y~x1+x2, bodyfat))
```

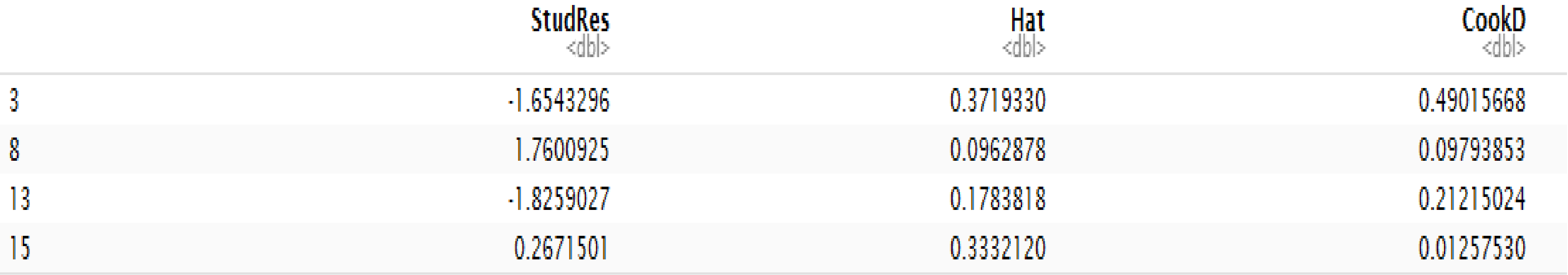


```
     (Intercept)            x1            x2
1  -3.051821e-01 -1.314856e-01  2.320319e-01
2   1.725732e-01  1.150251e-01 -1.426129e-01
3  -8.471013e-01 -1.182525e+00  1.066903e+00
4  -1.016120e-01 -2.935195e-01  1.960719e-01
5  -6.372122e-05 -3.052747e-05  5.023715e-05
6   3.967715e-02  4.008114e-02 -4.426759e-02
7  -7.752748e-02 -1.561293e-02  5.431634e-02
8   2.614312e-01  3.911262e-01 -3.324533e-01
9  -1.513521e-01 -2.946556e-01  2.469091e-01
10  2.377492e-01  2.446010e-01 -2.688086e-01
11 -9.020885e-03  1.705640e-02 -2.484518e-03
12 -1.304933e-01  2.245800e-02  6.999608e-02
13  1.194147e-01  5.924202e-01 -3.894913e-01
14  4.517437e-01  1.131722e-01 -2.977042e-01
15 -3.004276e-03 -1.247567e-01  6.876929e-02
16  9.308463e-03  4.311347e-02 -2.512499e-02
17  7.951208e-02  5.504357e-02 -7.609008e-02
18  1.320522e-01  7.532874e-02 -1.161003e-01
19 -1.296032e-01 -4.072030e-03  6.442931e-02
20  1.019045e-02  2.290797e-03 -3.314146e-03
```

Diagnostic with the influencePlot() Output



```
library(car)
influencePlot(lm(y~x1+x2, bodyfat))
```



	StudRes <dbl>	Hat <dbl>	CookD <dbl>
3	-1.6543296	0.3719330	0.49015668
8	1.7600925	0.0962878	0.09793853
13	-1.8259027	0.1783818	0.21215024
15	0.2671501	0.3332120	0.01257530

List all possible cases that are identified as outlying regarding its Y value.

None, since the studentized deleted residuals do not exceed the Bonferroni critical value.

With the Bonferroni procedure, a case is possibly an outlier if $|t_i| >$ $t\left(1-\frac{\alpha}{2n};n-1-p\right) = t(1-0.05/2(20);20-1-3)$

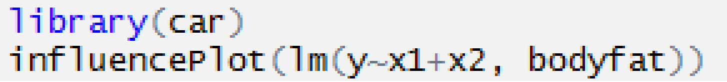
$= t(0.99875;16) = 3.58$

List all possible cases that are identified as outlying with regard to its X value.

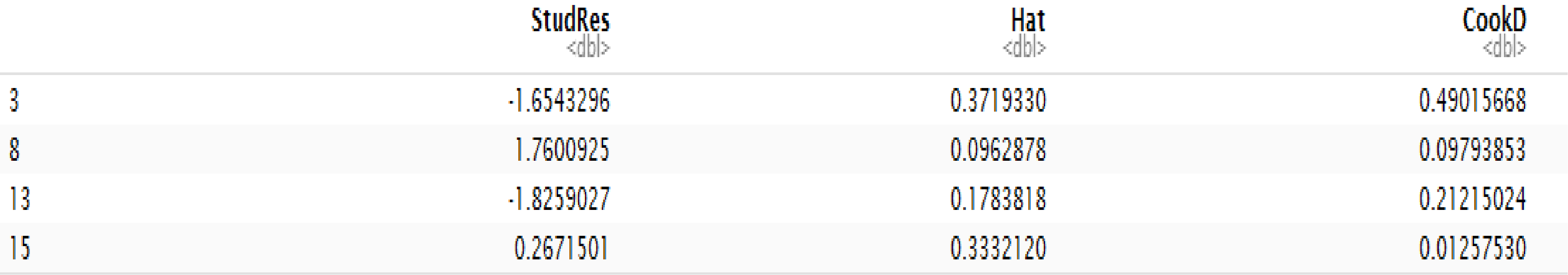
Case 3 and 15

A case is considered an outlier if $h_{ii} >$ $2\,\bar{h} =$ 2p/n=2(3)/20=0.3

Diagnostic with the influencePlot() Output



```
library(car)
influencePlot(lm(y~x1+x2, bodyfat))
```

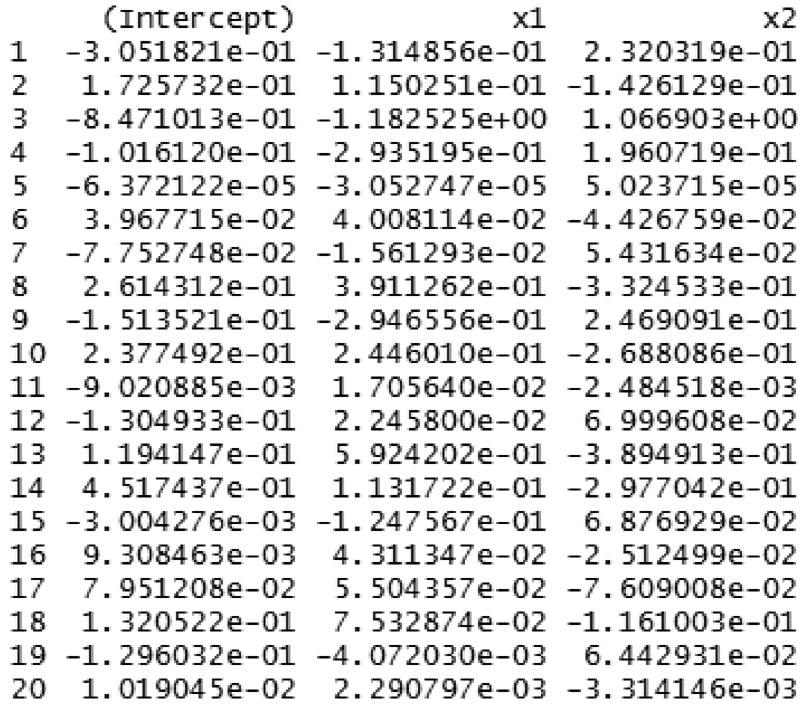


	StudRes <dbl>	Hat <dbl>	CookD <dbl>
3	-1.6543296	0.3719330	0.49015668
8	1.7600925	0.0962878	0.09793853
13	-1.8259027	0.1783818	0.21215024
15	0.2671501	0.3332120	0.01257530

A case has major influence if D $> \mathrm{F}(0.5; p, n-p) = F(0.5; 3, 20-3) = 0.821$;
Moderate influence if D is less than 0.821 but greater than $\mathrm{F}(0.2; p, n-p) = \mathrm{F}(0.2; 3, 20-3)$ =0.335

Are there any potential influential points? Case 3 has moderate influence.

Diagnostic with the influencePlot() Output

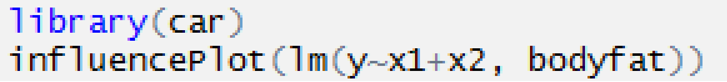


```
     (Intercept)            x1            x2
1  -3.051821e-01 -1.314856e-01  2.320319e-01
2   1.725732e-01  1.150251e-01 -1.426129e-01
3  -8.471013e-01 -1.182525e+00  1.066903e+00
4  -1.016120e-01 -2.935195e-01  1.960719e-01
5  -6.372122e-05 -3.052747e-05  5.023715e-05
6   3.967715e-02  4.008114e-02 -4.426759e-02
7  -7.752748e-02 -1.561293e-02  5.431634e-02
8   2.614312e-01  3.911262e-01 -3.324533e-01
9  -1.513521e-01 -2.946556e-01  2.469091e-01
10  2.377492e-01  2.446010e-01 -2.688086e-01
11 -9.020885e-03  1.705640e-02 -2.484518e-03
12 -1.304933e-01  2.245800e-02  6.999608e-02
13  1.194147e-01  5.924202e-01 -3.894913e-01
14  4.517437e-01  1.131722e-01 -2.977042e-01
15 -3.004276e-03 -1.247567e-01  6.876929e-02
16  9.308463e-03  4.311347e-02 -2.512499e-02
17  7.951208e-02  5.504357e-02 -7.609008e-02
18  1.320522e-01  7.532874e-02 -1.161003e-01
19 -1.296032e-01 -4.072030e-03  6.442931e-02
20  1.019045e-02  2.290797e-03 -3.314146e-03
```

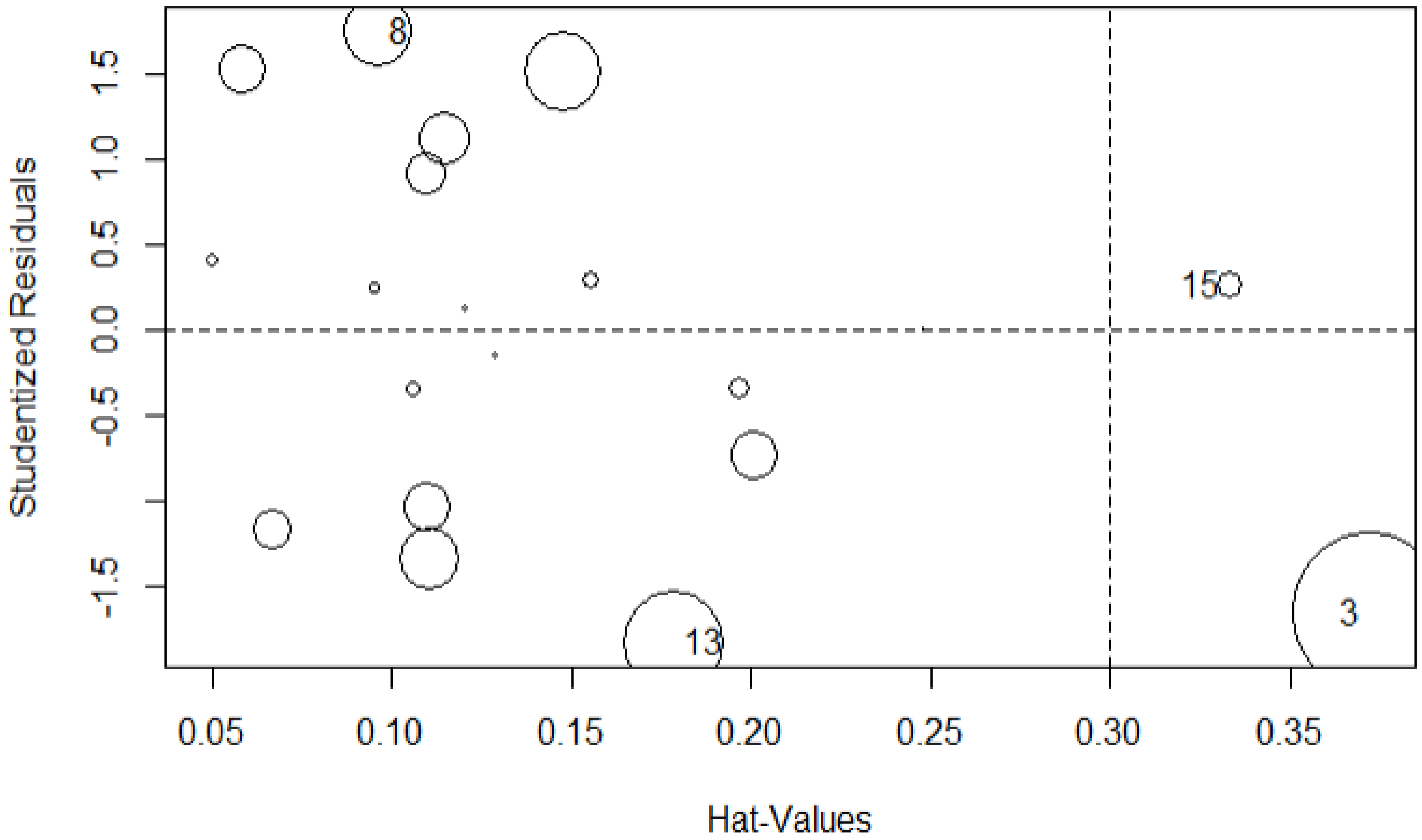
A case i is considered has a large impact on the regression coefficient if the |DEBETAS| >1, or > $\frac{2}{\sqrt{n}}$ (=0.45) for large data set.

Any case has significant impact? Case 3 since |DEBETAS| for beta1 and beta2 are bigger than 1 but not by much.

Diagnostic with the influencePlot() Output



```
library(car)
influencePlot(lm(y~x1+x2, bodyfat))
```

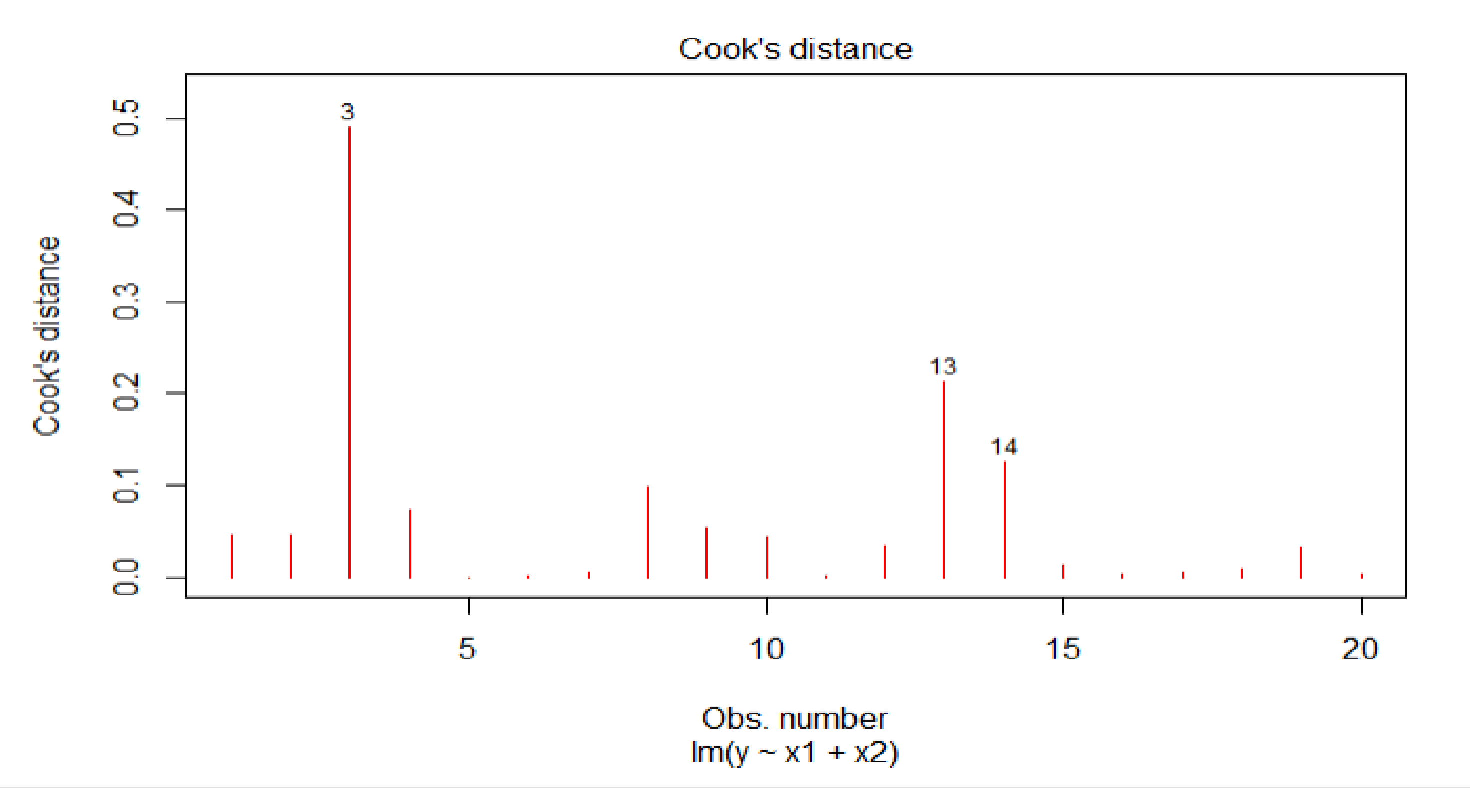


- The areas of the circles are proportional to DFFIT and Cooks distance.
- Vertical reference are drawn at twice and three times the average hat value
- Horizontal reference lines at -2, 0 and 2 on the Studentized-residual scale (for reference).

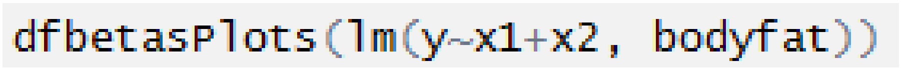
Outlying Y observations deviate in the Y axis (Studentized residuals)
Outlying X observations deviate in the X axis (Hat values)
The areas of the circles imply potential influential point, the bigger it is, the more likely that the point is influential.

Diagnostic with Cook’s distance Plot

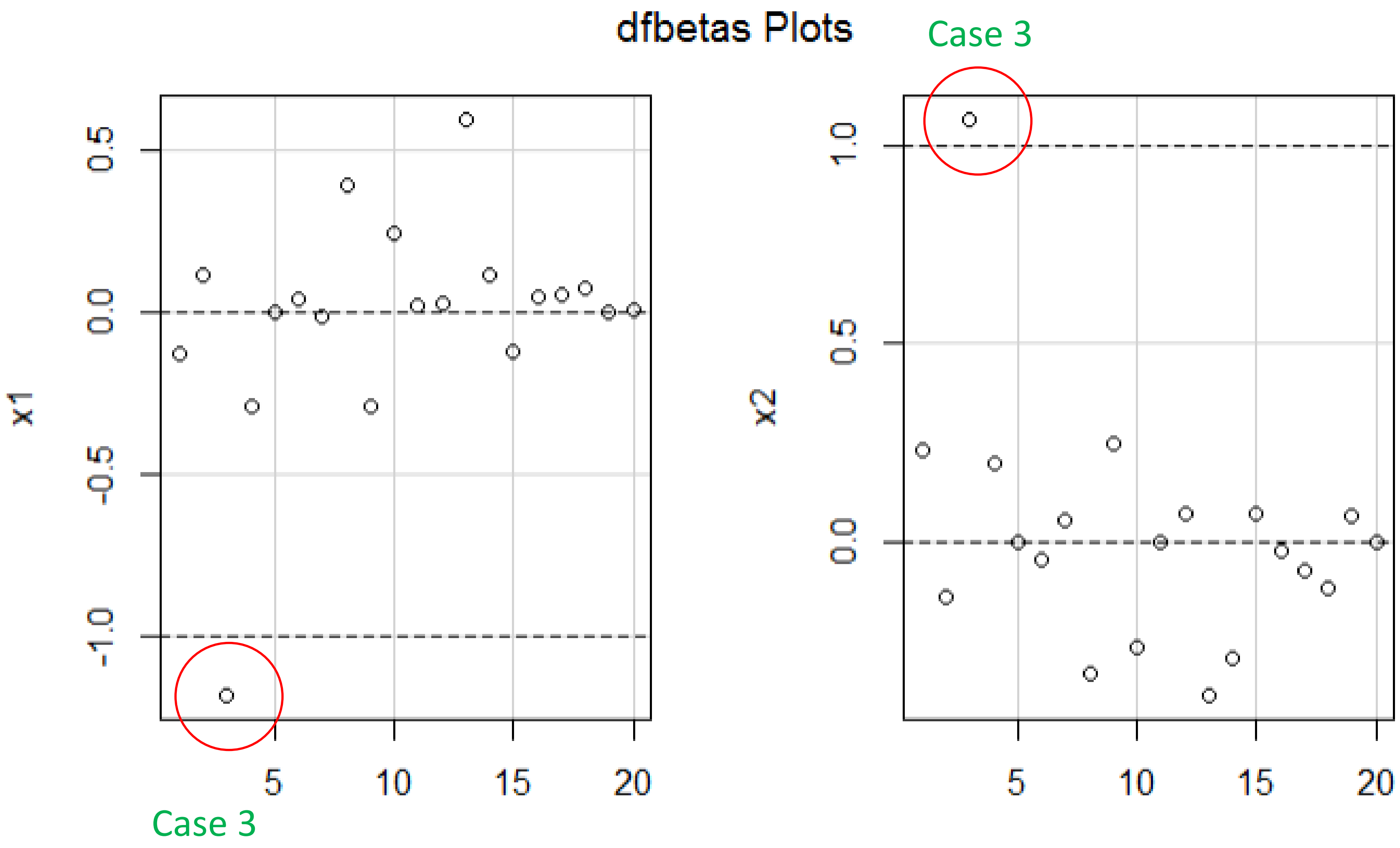
```
plot(lm(y~x1+x2, bodyfat), pch=18, col="red", which=c(4))
```



Diagnostic with dfbetasPlots() Plot



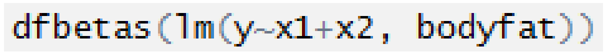
```
dfbetasPlots(lm(y~x1+x2, bodyfat))
```



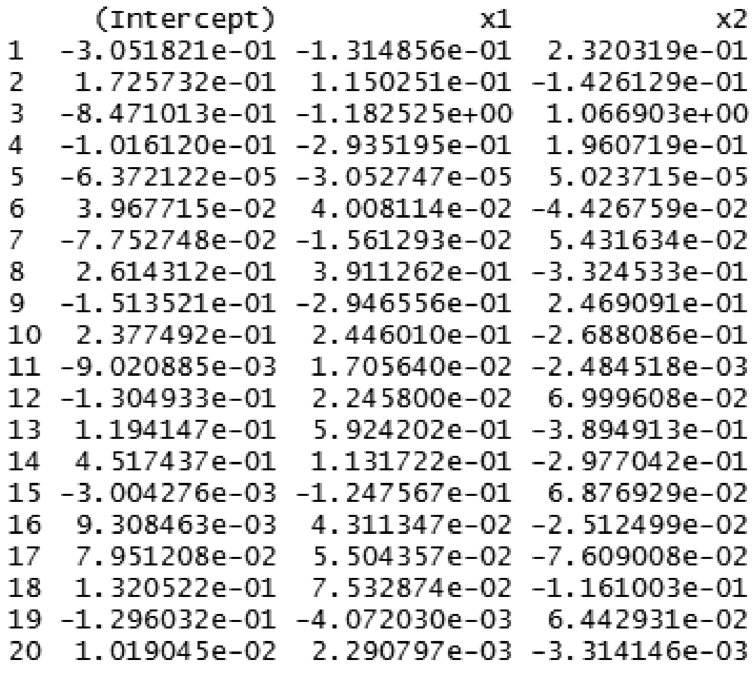
Case 3

Case 3

- Horizontal reference lines at 0, and +/- 1



```
dfbetas(lm(y~x1+x2, bodyfat))
```



```
     (Intercept)            x1            x2
1  -3.051821e-01 -1.314856e-01  2.320319e-01
2   1.725732e-01  1.150251e-01 -1.426129e-01
3  -8.471013e-01 -1.182525e+00  1.066903e+00
4  -1.016120e-01 -2.935195e-01  1.960719e-01
5  -6.372122e-05 -3.052747e-05  5.023715e-05
6   3.967715e-02  4.008114e-02 -4.426759e-02
7  -7.752748e-02 -1.561293e-02  5.431634e-02
8   2.614312e-01  3.911262e-01 -3.324533e-01
9  -1.513521e-01 -2.946556e-01  2.469091e-01
10  2.377492e-01  2.446010e-01 -2.688086e-01
11 -9.020885e-03  1.705640e-02 -2.484518e-03
12 -1.304933e-01  2.245800e-02  6.999608e-02
13  1.194147e-01  5.924202e-01 -3.894913e-01
14  4.517437e-01  1.131722e-01 -2.977042e-01
15 -3.004276e-03 -1.247567e-01  6.876929e-02
16  9.308463e-03  4.311347e-02 -2.512499e-02
17  7.951208e-02  5.504357e-02 -7.609008e-02
18  1.320522e-01  7.532874e-02 -1.161003e-01
19 -1.296032e-01 -4.072030e-03  6.442931e-02
20  1.019045e-02  2.290797e-03 -3.314146e-03
```

Measures of Multicollinearity

We can already diagnose multicollinearity by observing:

- A significant global F-test alongside non-significant t-tests for all individual βs
- Parameter estimates that change greatly when predictors are added to the model or removed
- Parameter estimates that “don’t make sense” and the standard errors become large.
- Large differences between Type I and Type II extra sums of squares

Variance Inflation Factor (Tolerance)

The VIF measures the extent to which the variance of the estimated regression coefficient of a predictor variable is increased due to multicollinearity.

$\Sigma\{b\} = \sigma^2(X'X)^{-1}$, *and* $Var(b_i) = \sigma^2(X'X)^{-1}[i,i]$ in MLR, and reduced to σ^2/SS_X in SLR.

Furthermore, $Var(b_i) = \frac{\sigma^2}{SS_{X_i}(1-R_i^2)} = \frac{\sigma^2}{SS_X} \times \frac{1}{1-R_i^2}$, where R_i^2 is the correlation determination of X_i and other predictors.

Example, regress X_1 on X_2 and X_3, then $R_{1|23}^2$ is the coefficient determination for $\mathrm{X}_1 \sim X_2 + X_3$

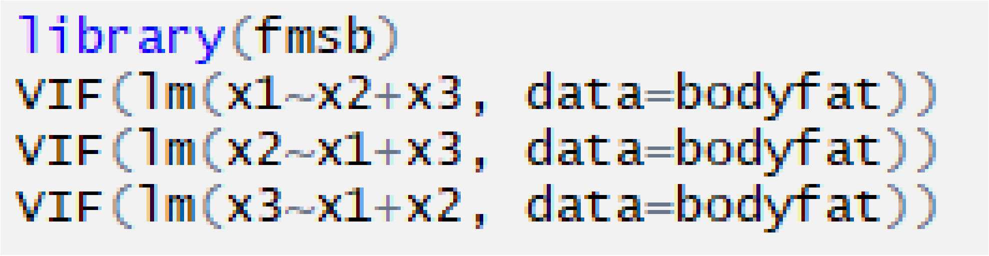
- *If none multicollinearity between* X_1 *and* (X_2, X_3), $R_{1|23}^2 = 0$, *then* $Var(b_1) = \frac{\sigma^2}{SS_{X_1}(1-R_1^2)} = \frac{\sigma^2}{SS_X}$.
- *If all multicollinearity between* X_1 *and* (X_2, X_3), $R_{1|23}^2 = 1$, *then* $Var(b_1) = \frac{\sigma^2}{SS_{X_1}(1-R_1^2)}$ *becomes indeterminate.*

$$\text{Define the Variance Inflation Factor}, VIF_i \ as \ \frac{1}{1-R_i^2} = \frac{1}{Tolenrance}$$

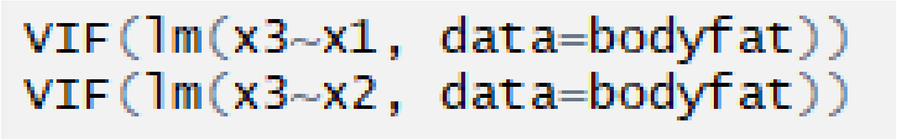
As a rule of thumb, a $VIF \geq 10$,or R_k^2 >0.9 indicate excessive multilinearity.

The Body-fat example

$$
\begin{aligned}
VIF_{1\,|2\,3} &= 708.84 \\
VIF_{2\,|1\,3} &= 564.34 \\
VIF_{3\,|1\,2} &= 104.61 \\
VIF_{3\,|1} &= 1.265 \\
VIF_{3\,|\,2} &= 1.007
\end{aligned}
$$

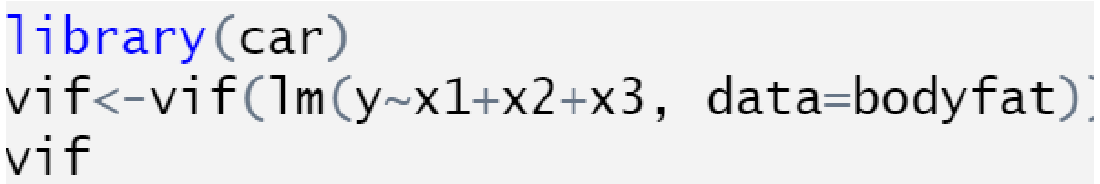


```
library(fmsb)
VIF(lm(x1~x2+x3, data=bodyfat))
VIF(lm(x2~x1+x3, data=bodyfat))
VIF(lm(x3~x1+x2, data=bodyfat))
```

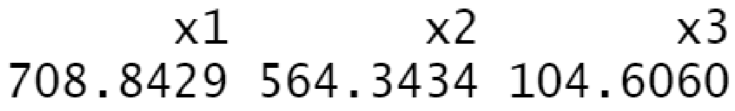


```
VIF(lm(x3~x1, data=bodyfat))
VIF(lm(x3~x2, data=bodyfat))
```

- The result proves that multicollinearity issue exists between X_3 and ($X_1 \, and \, X_2$) combined.
- It is a common practice to compute the VIF of a predictor with all other predictors considered in the full model.



```
library(car)
vif<-vif(lm(y~x1+x2+x3, data=bodyfat))
vif
```



```
      x1       x2       x3 
708.8429 564.3434 104.6060 
```