Chapter 10: Building the regression model II: Diagnostics

Refined diagnostics for checking the adequacy of a regression model

- Improper functional form for a predictor variable
- Outliers
- Influential observations
- Multicollinearity

10.1 Model adequacy for a predictor variable – Added-variable plots

Review:

1. Extra sum of squares: Measurement of additional or extra reduction of the error sums of squares when an independent variable is added to the model given a set of independent variables are already in the model.
   For example, $SSR(X_2|X_1) = SSE(X_1) – SSE(X_1,X_2)$.

2. Residual plot of e_i vs. X_{ik} (residuals vs. kth independent variable): Determine the appropriateness of the specified relationship between X_k and Y. For example, if there is a random scattering of points, the relationship between X_k and Y is specified correctly. If there is a relationship between the points of e_i vs. X_{ik} (for example, a quadratic relationship), this suggests X_{ik} is not specified correctly in the model.

The problem with #2 is that it does not necessarily give information about the marginal relationship between X_k and Y given all of the independent variables in the model.

> Solution: Use **added-variable plots** (also called **partial regression plots** and **adjusted variable plots)**. SAS calls these partial regression leverage plots.

Suppose there are only two independent variables X_1 and X_2. The steps to create a partial regression plot for X_1 are:

1. Fit the model using Y as the dependent variable and X_2 as the independent variable. Obtain the residuals. Symbolically, $\hat{Y}_i(X_2) = \hat{\beta}_0 + \hat{\beta}_1 X_{i2}$ and $e_i(Y\,|\,X_2) = Y_i - \hat{Y}_i(X_2)$.
2. Fit the model with X_1 as the dependent variable and X_2 as the independent variable. Obtain the residuals. Symbolically, $\hat{X}_{i1}(X_2) = \hat{\beta}_0^* + \hat{\beta}_1^* X_{i2}$ and $e_i(X_1\,|\,X_2) = X_{i1} - \hat{X}_{i1}(X_2)$.
3. Plot $e_i(Y\,|\,X_2)$ vs. $e_i(X_1\,|\,X_2)$.

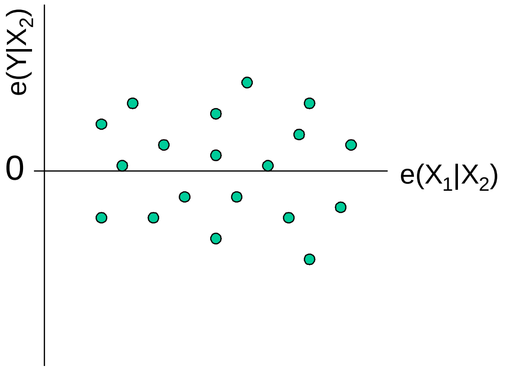
If there are more than two independent variables in the model, then plot $e_i(Y\,|\,X_2, X_3,\ldots,X_{p-1})$ vs. $e_i(X_1\,|\,X_2, X_3,\ldots,X_{p-1})$. In addition, make the appropriate changes to do added-variable plots for X_2, X_3,…, X_{p-1}.

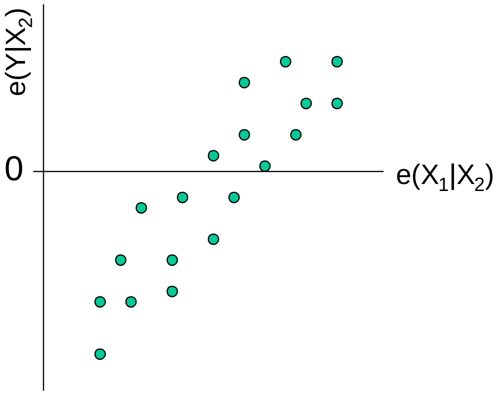
Interpretation:

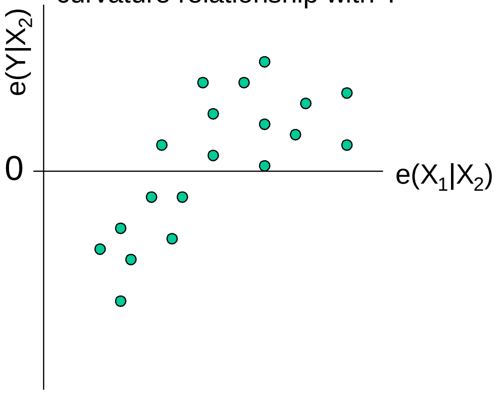
The added-variable plots
(1) show the marginal importance of this variable in reducing the residual variability;
(2) help to find the correct functional form of an independent variable in a multiple regression model;

(3) provide information about the strength of this relationship (If the scatter of the points around the line through the origin is much less than the scatter around the horizontal line, inclusion of the variable in the regression model will provide a substantial further reduction in the error sum of squares).

Suppose the only independent variables are X_1 and X_2. Below are example partial regression plots:


Given X_2 in the model, X_1 does not give any additional information about Y


Given X_2 in the model, X_1 has a linear relationship with Y


Given X_2 in the model, X_1 has a curvature relationship with Y

Notes:

1. Remember what a t-test does – it tests the linear relationship between Y and X_k given all of the other variables in the model. Therefore, the partial regression plots and t-tests partially give the same information. With the added-variable plots, there is also information about the type of relationship between Y and X_k.
2. The added-variable plots are dependent on which independent variables are already included in a model.
3. See Figure 10.2 of KNN for an additional interpretation of added-variable plots. This further helps to relate added-variable plots to extra sum of squares.
4. Added-variable plots also help to identify outliers and "leverage" or influential points (more on this later).

Example 1: Table 10.1 shows a portion of the data on average annual income (X_1) of managers during the past two years, a score measuring each manager's risk aversion (X_2), and the amount of life insurance carried (Y) for a sample of 18 managers in the 30-39 age group.

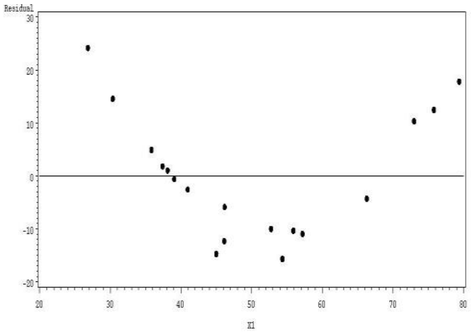
See graphs below.

The residual plot suggests that a linear relationship for X_1 is not appropriate.

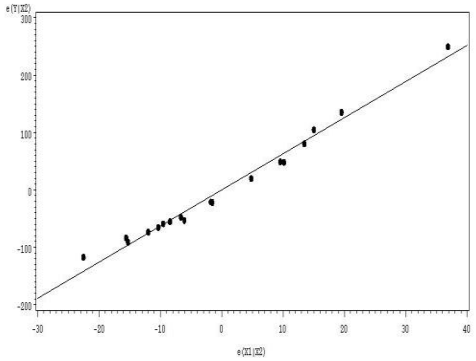
The added-variable plot suggests that the curvilinear relationship between Y and X_1 when X_2 is already in the regression model is strongly positive.

Note that the scatter of the points around the least squares line through the origin with slope 6.2880 is much smaller than is the scatter around the horizontal axis, indicating that adding X_1 to the model with a linear relation will substantially reduce the error of sum squares. In fact, $R^2_{Y1|2}=0.984$, and the slope 6.2880 equals to the regression coefficient for X_1 when is X_1 added to the fitted model.

Residual plot against X1 when Y is regressed on both X1 and X2



Added-variable plot for X1



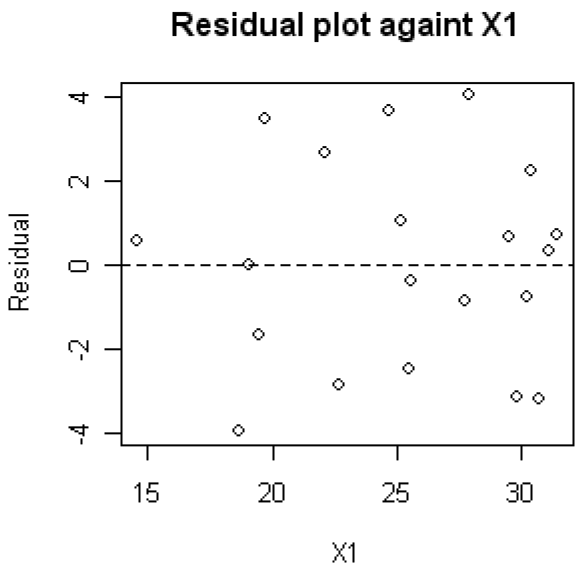
Example 2 Consider the body fat example (Table 7.1) again. For demonstration purpose, we only consider the regression of body fat (Y) on triceps thickness (X_1) and thigh circumference(X_2).

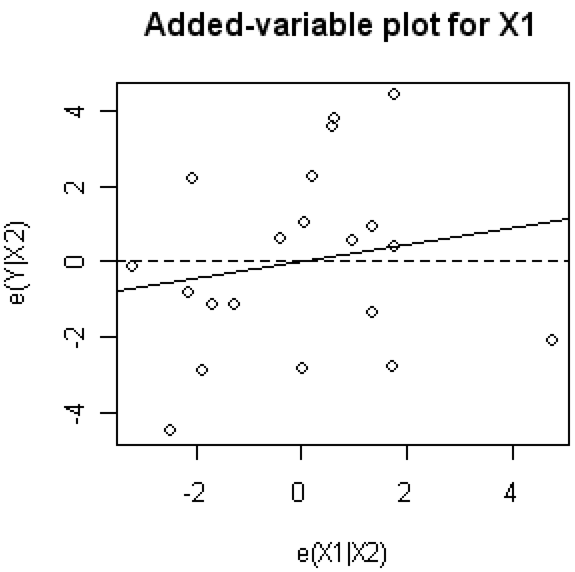
See the graphs below.

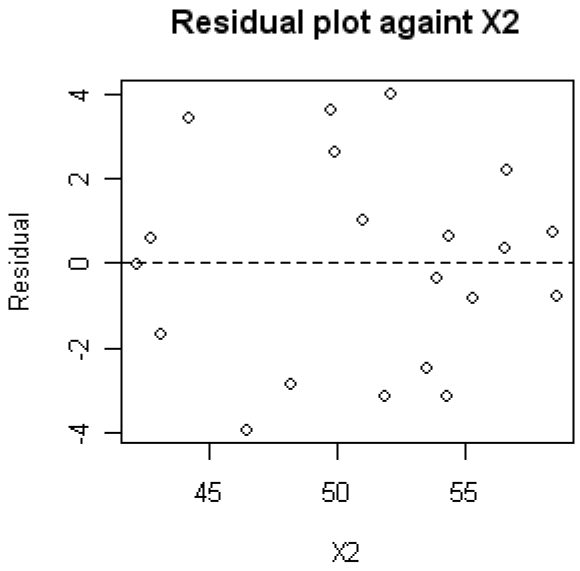
The two residual plots do not indicate any lack of fit for the linear terms or the existence of unequal variances of the error terms when the first-order model is used.

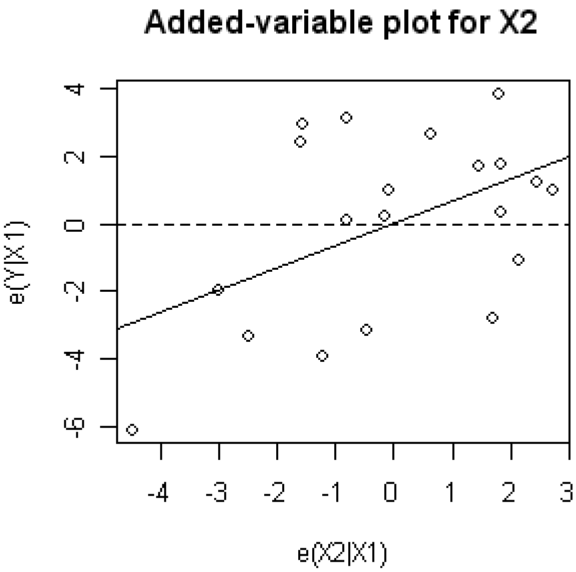
The added-variable plot for X_1 suggests that X_1 is of little help when X_2 is already in the regression model. In fact $R^2_{Y1|2}=0.031$.

The added-variable plot for X_2 suggests that a linear term in X_2 may be helpful even when X_1 is already in the regression model. In fact $R^2_{Y2|1}=0.232$. This plot also suggests the presence of one potentially influential case in the lower left corner.


Residual plot againt X1


Added-variable plot for X1


Residual plot againt X2

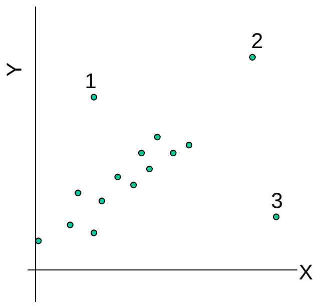

Added-variable plot for X2

10.2 Identifying outlying Y observations – studentized deleted residuals

Outlying cases

When there is only 1 independent variable, identifying outliers is not too difficult. Below is a partial reproduction of Figure 10.5 in KNN:

Scatter plot showing outliers



1. Outlying point with respect to its Y value. May not be very influential to the regression model fit since there are similar X values.
2. Outlying point with respect to its X and Y value. May not be very influential to the regression model fit since the Y value is consistent with the others.
3. Outlying point with respect to its X value. May be influential since the X value is outlying and not consistent with respect to the other X values.

In multiple regression, we generally can not look at plots as shown above (too many dimensions). Therefore, we need to look at numerical measures that give information about a particular observation being outlying or not.

Residuals and semistudentized residuals (Chapters 1 and 3)

$$e_i = Y_i - \hat{Y}_i \text{ and } e_i^* = \frac{e_i}{\sqrt{MSE}}$$

Remember that $\sqrt{MSE}$ is not quite the estimated variance of e_i. Thus, e_i^* is not quite a random variable with variance of 1.

Two refinements are made to the analysis of residuals more effective for identifying outlying Y observations. These refinements require the use of hat matrix.

Hat matrix (Chapters 5 and 6)

Remember that:

$\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$,

The variance-covariance matrix for e is Cov(**e**) = σ^2(**I**-**H**), and its estimate is

$$\overset{\wedge}{Cov}(\mathbf{e}) = MSE(\mathbf{I} - \mathbf{H})$$

Note that **H** is a n×n matrix with elements of

$$\begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix}$$

And, Cov(**e**) is

$$\sigma^2\left(\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} - \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix}\right) = \sigma^2\begin{bmatrix} 1-h_{11} & -h_{12} & \cdots & -h_{1n} \\ -h_{21} & 1-h_{22} & \cdots & -h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -h_{n1} & -h_{n2} & \cdots & 1-h_{nn} \end{bmatrix}$$

Thus, Cov(e_1, e_1) = Var(e_1) = σ^2(1-h_{11}), Var(e_2)=σ^2(1-h_{22}), … Var(e_n)=σ^2(1-h_{nn}). In general, Var(e_i)=σ^2(1-h_{ii})

The estimated variance of the ith residual is

$$\hat{Var}(e_i) = MSE(1-h_{ii})$$

The **studentized residual** is:

$$r_i = \frac{e_i}{\sqrt{\hat{Var}(e_i)}} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

r_i~t(n-p)

Deleted residuals

From Chapter 9, PRESS prediction error = Y_i - $\hat{Y}_{i(i)}$

Example: Let i=3. Then the PRESS prediction error = Y_3 - $\hat{Y}_{3(3)}$. To obtain $\hat{Y}_{3(3)}$, a regression model is fit to the data where observation 3 is removed from the data set. For observation 3's X_1,…,X_{p-1}, the predicted Y value is obtained - $\hat{Y}_{3(3)}$.

Deleted residual for the ith observation: d_i=Y_i - $\hat{Y}_{i(i)}$

An algebraically equivalent expression without the recomputation of the fitted regression function omitting the ith case is:

$$d_i = \frac{e_i}{(1-h_{ii})}$$

Notes:

1. Suppose X_i is very influential on the regression model fit (for example, point #3 on Figure 10.5). Then the regression line will be "pulled" toward (X_i,Y_i) resulting in a $\hat{Y}_i$ "close" to Y_i. If this observation is deleted from the data set and a new

regression model is fit, $\hat{Y}_{i(i)}$ will not be as “close” to Y_i. This will result in a d_i that is larger (in absolute value) than e_i.

2. Suppose X_i is not very influential on the regression model fit (for example, a point within the main cluster of points). The regression line will not be heavily influenced by (X_i,Y_i). Thus, d_i should be about the same as e_i.

From Section 6.7, the estimated variance of a “new” observation is

$$s^2\{pred\} = MSE(1 + \mathbf{X}_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$$

for $\mathbf{X}_h=(1, X_{h1}, X_{h2}, \ldots, X_{h,p-1})'$. This variance is used in calculating prediction intervals.

Since the i^{th} observation is removed from the data set in calculating d_i, the i^{th} observation can be thought of as a “new” observation and the estimated variance from Section 6.7 can be used for its variance. Thus,

$$s^2\{d_i\} = MSE_{(i)}(1 + X'_i(X'_{(i)}X_{(i)})^{-1}X_i)$$

An algebraically equivalent expression is $s^2\{d_i\} = \dfrac{MSE_{(i)}}{1-h_{ii}}$

The **studentized deleted residual** is:

$$t_i = \frac{d_i}{s\{d_i\}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

$t_i \sim t(n\text{-}1\text{-}p)$,

where $MSE_{(i)}$ is the MSE from the model where the i^{th} observation is deleted. Also, $\mathbf{X}_{(i)}$ has the row of $\mathbf{X}$ corresponding to the i^{th} observation deleted. Note that $t_i \sim t(n\text{-}p\text{-}1)$ where the degrees of freedom comes from n-1 observations and p parameters.

One more expression can be found for the studentized deleted residuals! Note that

$$(n-p)MSE = (n-p-1)MSE_{(i)} + \frac{e_i^2}{(1-h_{ii})}$$

Thus,

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

Therefore, n different regression models DO NOT need to be calculated!

Test for outlying Y observations

KNN recommend doing the following (p.396) for studentized deleted residuals:

> Use a Bonferroni critical value of $t[1-\alpha/(2n); n\text{-}p\text{-}1]$ to determine if the observation is an outlier. In other words, H_o: Not outlier vs. H_a: outlier. If $|t_i| < t(1-\alpha/(2n); n\text{-}p\text{-}1)$, then there is not sufficient evidence that the i^{th} observation is an outlier. Otherwise, the i^{th} observation is an outlier.

Use the same method for studentized residuals, but use n-p degrees of freedom.

Problem: t[1-α/(2n); n-p-1] typically is large since the number of tests being done is n. Therefore, this procedure will be VERY conservative.

Example: Consider the body fat example (Table 7.1) again. For demonstration purpose, we only consider the regression of body fat (Y) on triceps thickness (X_1) and thigh circumference(X_2).

Question: Is case 13 which has the largest absolute studentized deleted residual an outlier? When α =0.10, the appropriate Bonferroni critical value is:

$$t(1-\alpha/(2n); n-p-1)=t(0.9975;16)=3.252$$

```
> Data <- read.table(file="C://Hongmei/Teaching/stat350/DataSet/CH07TA01.txt", header=FALSE)
> colnames(Data) <- c("X1", "X2", "X3", "Y")
> fit <- lm(Y~X1+X2, data=Data)
> rstudent(fit)# studentized deleted residuals
            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 
> influence.measures(fit)# DFFITS, Cook’s distance, DFBETAS
Influence measures of
	 lm(formula = Y ~ X1 + X2, data = Data) :

      dfb.1_    dfb.X1    dfb.X2     dffit cov.r   cook.d    hat inf
1  -3.05e-01 -1.31e-01  2.32e-01 -3.66e-01 1.361 4.60e-02 0.2010    
2   1.73e-01  1.15e-01 -1.43e-01  3.84e-01 0.844 4.55e-02 0.0589    
3  -8.47e-01 -1.18e+00  1.07e+00 -1.27e+00 1.189 4.90e-01 0.3719   *
4  -1.02e-01 -2.94e-01  1.96e-01 -4.76e-01 0.977 7.22e-02 0.1109    
5  -6.37e-05 -3.05e-05  5.02e-05 -7.29e-05 1.595 1.88e-09 0.2480   *
6   3.97e-02  4.01e-02 -4.43e-02 -5.67e-02 1.371 1.14e-03 0.1286    
7  -7.75e-02 -1.56e-02  5.43e-02  1.28e-01 1.397 5.76e-03 0.1555    
8   2.61e-01  3.91e-01 -3.32e-01  5.75e-01 0.780 9.79e-02 0.0963    
9  -1.51e-01 -2.95e-01  2.47e-01  4.02e-01 1.081 5.31e-02 0.1146    
10  2.38e-01  2.45e-01 -2.69e-01 -3.64e-01 1.110 4.40e-02 0.1102    
11 -9.02e-03  1.71e-02 -2.48e-03  5.05e-02 1.359 9.04e-04 0.1203    
12 -1.30e-01  2.25e-02  7.00e-02  3.23e-01 1.152 3.52e-02 0.1093    
13  1.19e-01  5.92e-01 -3.89e-01 -8.51e-01 0.827 2.12e-01 0.1784    
14  4.52e-01  1.13e-01 -2.98e-01  6.36e-01 0.937 1.25e-01 0.1480    
15 -3.00e-03 -1.25e-01  6.88e-02  1.89e-01 1.775 1.26e-02 0.3332   *
16  9.31e-03  4.31e-02 -2.51e-02  8.38e-02 1.309 2.47e-03 0.0953    
17  7.95e-02  5.50e-02 -7.61e-02 -1.18e-01 1.312 4.93e-03 0.1056    
18  1.32e-01  7.53e-02 -1.16e-01 -1.66e-01 1.462 9.64e-03 0.1968    
19 -1.30e-01 -4.07e-03  6.44e-02 -3.15e-01 1.002 3.24e-02 0.0670    
20  1.02e-02  2.29e-03 -3.31e-03  9.40e-02 1.224 3.10e-03 0.0501    
```

10.3 Identifying outlying X observations – hat matrix leverage values

The diagonal elements of the hat matrix, h_{ii}, are useful in detecting outlying X observations.

h_{ii} measures distance of the i^th observation to the mean (center) of all observations. The farther away from the mean, the larger the h_{ii}. See Figure 10.6 of KNN.

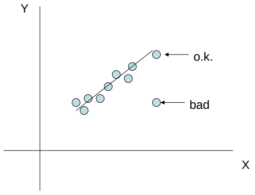
Notes:

1) Properties of h_{ii}'s: $0 \le h_{ii} \le 1$ and $\Sigma h_{ii} = p$
2) h_{ii} is often called the "leverage" (in terms of the X values) of the i^{th} observation.
3) If the i^{th} observation is outlying in terms of the X values, it has substantial leverage on determining the fitted value for $\hat{Y}_i$.
4) The larger is h_{ii}, the more important Y_i is to determining $\hat{Y}_i$. Remember that $\hat{\mathbf{Y}} = \mathbf{HY}$; i.e., $\hat{Y}_i$ is a linear combination of h_{ii}'s.
5) The larger is h_{ii}, the smaller are the denominators of r_j and t_i (i.e., the estimated variances) since

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}} \text{ and } t_i = e_i\left[\frac{n-p-1}{SSE(1-h_{ii})-e_i^2}\right]^{1/2}.$$

6) Since the estimated regression line is pulled toward observations with high leverage, e_i may not be large relative to the rest of the residuals.
7) Rule of thumb for determining if h_{ii} is "large":
   a) $h_{ii} > 2p/n$ where p/n can be shown to be the mean of h_{ii}
   b) $h_{ii} > 0.5$ indicates very high leverage, $0.2 < h_{ii} \le 0.5$ indicate moderate leverage.

   These rules apply only when the sample size is "large" relative to the number of parameters in the model.
8) Large h_{ii} are values are not necessarily bad!



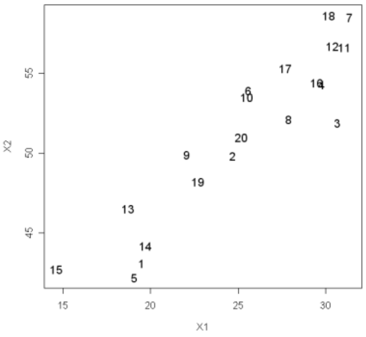
9) Use of hat matrix to identify hidden extrapolation

   Let X_{new} be the vector containing the X values for which an inference about a mean response or a new observation is to be made. Compute

$$h_{new,new} = X'_{new}(X'X)^{-1}X_{new}$$

   IF $h_{new,new}$ is within the range of leverage values h_{ii} for the cases in the data set, no extrapolation is involved.

   Example: Body fat example: Are there any outlying cases in terms of their X values? Look at previous SAS output, and notice $2p/n=2(3)/20=0.3$. Check out the graph below



10.4 Identifying influential cases – DFFITS, Cook's distance, and DFBETAS measures

Once outlying observations are identified, it needs to be determined if they are influential to the estimated regression function. An observation is considered to be **influential** if its exclusion causes major changes in the fitted regression function.

Influence on single fitted value – DFFITS

The influence of observation i on $\hat{Y}_i$ is measured by:

$$(\text{DFFITS})_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)}h_{ii}}}$$

The letters "DF" stands for the difference between the fitted values $\hat{Y}_i$ for the ith case when all n cases are used in fitting the regression function and the predicted value $\hat{Y}_{i(i)}$ for the ith case obtained when the ith case is omitted in fitting the regression function.

(DFFITS)$_i$ ≈ the number of standard deviations by which $\hat{Y}_i$ changes when observation i is removed from the data set.

It can shown that DFFITS can be calculated in the following manner:

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

Therefore, only one regression model needs to be fit.

Guideline for determining influential observations:

- |(DFFITS)$_i$|>1 for "small to medium" sized data sets
- |(DFFITS)$_i$|>$2\sqrt{p/n}$ for large data sets

Influence on all fitted values – Cook's Distance

Measures the influence of the i$^{\text{th}}$ observation on ALL n fitted values.

Cook's Distance is:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\text{MSE}}$$

Notes:

1. The numerator is similar to (DFFITS)$_i$. For Cook's Distance, ALL of the fitted values are compared.
2. The denominator serves as a standardizing measure.

It can be shown that Cook's Distance can be calculated in the following manner:

$$D_i = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

Therefore, only one regression model needs to be fit. From examining the above formula, note how D_i can be large (larger e_i or h_{ii} or both).

Guideline for determining influential observations:
Relate D_i to the F(p, n-p) distribution and ascertain the corresponding percentile value.

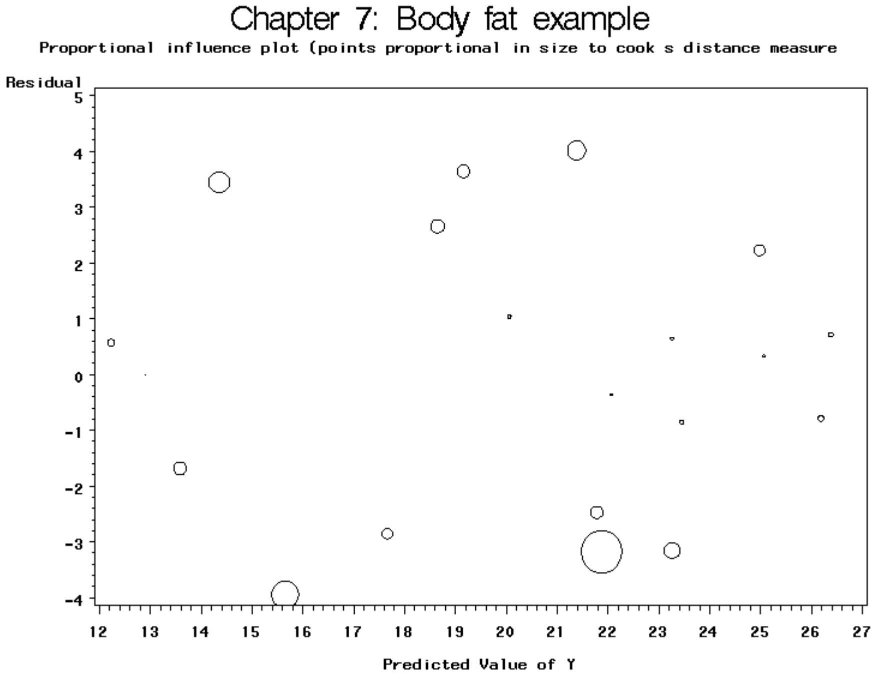
- If the percentile value is less than about 10% or 20%, little influence on the fitted values.
- If the percentile values is near 50% or more, the ith observation has a major influence on the fit of the regression function.

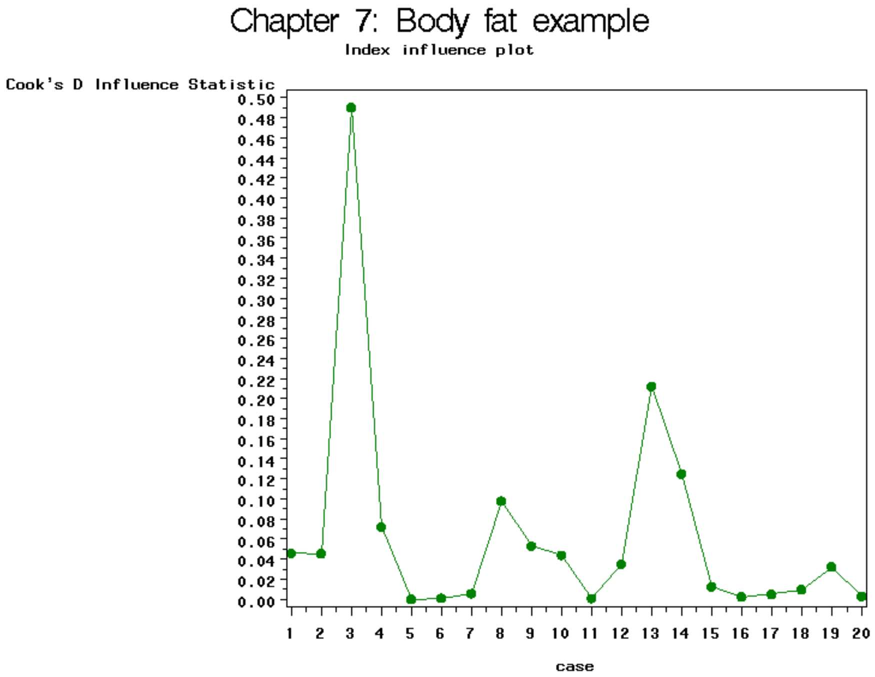
Question: For body fat example, is case an influential point?

For case 3, D_i=0.49 is the 30.6th percentile using F(3,17) distribution. It does have influence on the regression fit. However it may not be large enough to call for consideration of remedial measures.

(If you specify model Y=X/r in the SAS proc reg procedure, Cook's distance will be printed out automatically.)

Look at **Figure 10.8**.





Influence on the regression coefficients - DFBETAS

Measures the influence of the ith observation on each estimated regression coefficient, b_k.

Let $b_{k(i)}$ be the estimate of β_k with the ith observation removed from the data set, and c_{kk} be the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ (remember that $\mathbf{X}$ is a n×p matrix)

Then

$$(\text{DFBETAS})_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{\text{MSE}_{(i)} c_{kk}}}$$

for k=1,…,p-1

<u>Notes</u>:

1. Notice that DFBETAS is calculated for each β_k and each observation.
2. The variance of b_k is $\sigma^2 c_{kk}$. In this case, σ^2 is estimated by $\text{MSE}_{(i)}$. Therefore, the denominator serves as a standardizing measure.

Guideline for determining influential observations:

- $|(\text{DFBETAS})_{k(i)}| > 1$ for "small to medium" sized data sets
- $|(\text{DFBETAS})_{k(i)}| > 2/\sqrt{n}$ for large data sets

Influence on inferences

Examine the inferences from the estimated regression model with and without the observation(s) of concerned.

Average absolute percent difference: $\dfrac{\sum_{j=1}^{n}\left|\dfrac{\hat{Y}_{j(i)}-\hat{Y}_j}{\hat{Y}_j}\right|}{n}$

- If inferences are unchanged, remedial action is not necessary.
- If inferences are changed, remedial action is necessary.

R codes:

```
Data <- read.table(file="C:// CH07TA01.txt", header=FALSE)
colnames(Data) <- c("X1", "X2", "X3", "Y")
fit <- lm(Y~X1+X2, data=Data)
rstudent(fit)                    # studentized deleted residuals
influence.measures(fit)          # DFFITS, Cook's distance, DFBETAS
```

10.5 Multicollinearity diagnostics – variance inflation factor

Chapter 7 discusses informal ways to detect multicollinearity and the results of multicollinearity. This section discusses a more formal measure of multicollinearity – the variance inflation factor (VIF).

Problems arise when predictor variables are highly correlated:

1) Adding or deleting a predictor variable changes the regression coefficients
2) The extra sum of squares associated with a predictor variable varies, depending upon which predictor variables are already included in the model
3) The estimated standard deviations of the regression coefficients become large
4) The estimated regression coefficients individually may not be statistically significant even though a definite statistical relation exists between the response variable and the set of predictor variables

Informal Diagnostics: indications of the presence of serious multicollinearity

1) Large changes in the estimated regression coefficients when a predictor variables is added or deleted
2) Nonsignificant results in individual tests on regression coefficients for important predictor variables
3) Estimated regression coefficients with an algebraic sign that is the opposite of that expected from theoretical considerations or prior experience
4) Large coefficients of simple correlation between pairs of predictor variables in the correlation matrix
5) Wide confidence intervals for the regression coefficients representing important predictor variables

Variance Inflation factor

$(\mathrm{VIF})_k = (1-R_k^2)^{-1}$ for k=1,…,p-1

Where R_k^2 is the coefficient of multiple determination when X_k is regressed on the p-2 other X variables in the model.

R_k^2 measures the relationship between X_k and the other independent variables.

If R_k^2 is small (weak relationship) then $(\mathrm{VIF})_k$ is small. For example, suppose R_k^2=0, then $(\mathrm{VIF})_k$=1. If R_k^2=0.5, then $(\mathrm{VIF})_k$=2.

If R_k^2 is large (strong relationship) then $(\mathrm{VIF})_k$ is large. For example, suppose R_k^2=0.9, then $(\mathrm{VIF})_k$=10. If R_k^2=0.99, then $(\mathrm{VIF})_k$=100.

A large $(\mathrm{VIF})_k$ indicates the existence of multicollinearity.

$(\mathrm{VIF})_k>10$

The mean of the VIF values: $(\overline{\mathrm{VIF}}) = \dfrac{\sum_{k=1}^{p-1}(\mathrm{VIF})_k}{p-1}$.

Mean VIF values considerably larger than 10 are indicative of serious multicollinearity problems.

<u>Example</u>: Body fat example: The maximum of the VIF values is 708.84 and the mean value is 459.26. Thus the expected sum of the squared errors in the least squares standardized regression coefficients is nearly 460 times as large as it would be if the X variables are uncorrelated. These indicate there serious multicollinearity problems exist.

```
Variable       VIF
X1           708.84291
X2           564.34339
X3           104.60601
```

10.6 Surgical unit example
A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random selection of 108 patients was available for analysis. From each patient record, the following information was extracted from the pre-operation evaluation:

X_1	blood clotting score
X_2	prognostic index
X_3	enzyme function test score
X_4	liver function test score
X_5	age in years
X_6	indicator variable for gender (0=male, 1=female)
X_7 and X_8	indicator variable for history of alcohol use

Alcohol Use	X_7	X_8
None	0	0
Moderate	1	0
Severe	0	1

Use the first-order model with X_1, X_2, X_3 and X_8 based on Section 9.6.

R outputs and Figure 10.10

Questions:

(1) Is there a multicollinearity among the four predictor variables?

(2) Are there outlying Y observations?

(3) Are there outlying X observations?

(4) Are there influential cases? Check case 17.