LINEAR REGRESSION ANALYSIS

MODULE – VI

Lecture - 24

Tests for Leverage and Influential Points

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

2. *DFFITS* and *DFBETAS*

Cook's distance measure is a deletion diagnostic, i.e., it measures the influence of i^{th} observation if it is removed from the sample.

There are two more statistics:

(i) *DFBETAS* which indicates that how much the regression coefficient changes if the i^{th} observation were deleted. Such change is measured in terms of standard deviation units. This statistic is

$$DFBETAS_{j,i} = \frac{b_j - b_{j(i)}}{\sqrt{S^2_{(i)} C_{jj}}}$$

where C_{jj} is the j^{th} diagonal element of $(X'X)^{-1}$ and $b_{j(i)}$ regression coefficient computed without use of i^{th} observation.

Large (in magnitude) value of $DFBETAS_{j,i}$, indicates that i^{th} observation has considerable influence on the j^{th} regression coefficient.

i. The values of $DFBETAS_{j,i}$ can be expressed in a n x k matrix that conveys similar information to the composite influence information in Cook's distance measure.

ii. The n elements in the j^{th} row of R produce the leverage that the n observations in the sample have on $\hat{\beta}_j$. $DFBETAS_{j,i}$ is the j^{th} element of $(b - b_{(i)})$ divided by a standardization factor

$$b_i - b_{(i)} = \frac{(X'X)^{-1} x_i' e_i}{1 - h_{ii}}.$$

The j^{th} element of $(b_i - b_{(i)})$ can be expressed as

$$b_i^j - b_{(i)}^j = \frac{r_{j,i} e_i}{1 - h_{ii}}.$$

r_{ij} = ((*R*)) denotes the $(i, j)^{th}$ elements of *R*

$$(RR')' = \left[(X'X)^{-1}X'X(X'X)^{-1}\right]'$$

$$= (X'X)^{-1} = C = R'R.$$

Since

$$C_{jj} = r_j' r_j,$$

so

$$\sqrt{S^2_{(i)}C_{jj}} = \sqrt{S^2_{(i)}r_j' r_j}$$

$$DFBETAS_{j,i} = \frac{b_j - b_{j(i)}}{\sqrt{S^2_{(i)}C_{jj}}}$$

$$= \left(\frac{r_{j,i}e_i}{1-h_{ii}}\right)\frac{1}{\sqrt{S^2_{(i)}r_j' r_j}}$$

$$= \frac{r_{j,i}}{\sqrt{r_j' r_j}} \qquad \frac{t_i}{\sqrt{1-h_{ii}}}$$

↓ ↓

Measures leverage (impact of i^{th} observation on b_i	Measures effect of large residuals

where t_i is the i^{th} *R*-student residual. Now if $\left|DFBETAS_{j,i}\right| > \frac{2}{\sqrt{n}}$, then it indicates that i^{th} observation warrants examination.

2. *DFFITS*

The deletion influence of i^{th} observation on the predicted or fitted value can be investigated by using diagnostic by Belsley, Kuh and Welsch as

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S^2_{(i)} h_{ii}}},\ i = 1, 2, ..., n$$

where $\hat{y}_{(i)}$ is the fitted value of y_i obtained without the use of the i^{th} observation. The denominator is just a standardization, since $Var(\hat{y}_i) = \sigma^2 h_{ii}$.

This $DFFITS_i$ is the number of standard deviations that the fitted value $\hat{y}_i$ changes if i^{th} observation is removed.

Computationally,

$$DFFITS_i = \sqrt{\frac{h_{ii}}{1 - h_{ii}}} \frac{e_i}{S_{(i)}\sqrt{1 - h_{ii}}}$$

$$= t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

$$= \text{Studentized } R \times \text{leverage of } i^{th} \text{ observation}$$

where t_i is *R*-student.

- If the data point is an outlier, then *R*-student will be large is magnitude.
- If the data point has high leverage, then h_{ii} will be close to unity.
- In either of these cases, $DFFITS_i$ can be large.
- If $h_{ii} \approx 0$, then the effect of *R*-student will be moderated.
- If *R*-student is near to zero, then combined with high leverage point, the value of $DFFITS_i$ can be small.
- Thus $DFFITS_i$ is affected by both leverage and prediction error. Belsley, Kuh and Welsch suggest that any observation for which

$$\left|DFFITS_{\mathrm{i}}\right| > 2\sqrt{\frac{k}{n}}$$

warrants attention.

Note: The cutoff values of This $DFFITS_{j,i}$ and This $DFFITS_i$ are only guidelines. It is very difficult to provide cutoffs that are correct for all cases. So analyst is recommended to utilize information about both what is diagnostic means and the application environment in selecting a cutoff.

For example, if $DFFITS_i = 1$, say, we could translate this into actual response units to determine just how much $\hat{y}_i$ is affected by removing the i^{th} observation.

Then use $DFFITS_{j,i}$ to see whether this observation is responsible for the significance (or perhaps nonsignificance) of particular coefficients or for changes is sign in a regression coefficient.

$DFFITS_{j,i}$ can be used to determine how much change in actual problem-specific units a data point has on the regression coefficient. Sometimes these changes will be of importance in a problem-specific context even though the diagnostic statistic do not exceed the formal cutoff.

The recommended cutoff are a function of sample size n. Certainly, any formal cutoff should be a function of sample size.

However, is practice, these cutoffs often identify more data points than an analyst may wish to analyze. This is particularly true in small samples. The cutoff values provided by Belsley, Kuh and Welsch make more sense for large samples. When n is small, then diagnostic views are preferred.

A measure of model performance generalized variance

The diagnostics D_i , $DFFITS_{j,i}$ and $DFFITS_i$ provide insight about the effect of observations on the estimated coefficient $\hat{\beta}_j$ and fitted values $\hat{y}_i$. They do not provide any information about overall precision of estimation.

The **generalized variance** is defined as the determinant of covariance matrix and is a convenient scalar measure of precision. The generalized variance of OLSE b is

$$GV(b) = |V(b| = |\sigma^2 (X'X)^{-1}|.$$

To express the role of i^{th} observation on the precision of estimation, define

$$COVRATIO_i = \frac{|(X'_{(i)}X_{(i)})^{-1} S^2_{(i)}|}{|(X'X)^{-1} MS_{res}|}, \quad i = 1, 2, ..., n.$$

If $COVRATIO_i > 1 \Rightarrow i^{th}$ observation improves the precision of estimation.

If $COVRATIO_i < 1 \Rightarrow i^{th}$ inclusion of i^{th} observation degrades the precision computationally,

$$COVRATIO_i = \frac{S^2_{(i)}}{MS^k_{res}} \left(\frac{1}{1 - h_{ii}} \right)$$

where

$$\frac{1}{1 - h_{ii}} = \frac{|(X'_{(i)}X_{(i)})^{-1}|}{|(X'X)^{-1}|}.$$

- So high leverage point will make $COVRATIO_i$ large. This is logical, since a high-leverage point will improve the precision unless the point is an outlier in *y*-space.

- If i^{th} observation is outlier, then $\frac{S^2_{(i)}}{MS^k_{res}}$ will be much less than unity.

- Cut-off values for *COVRATIO* are not easy to obtain. It is suggested that

$$\text{if} \quad COVRATIO_i > 1 + \frac{3k}{n}$$

$$\text{or if} \quad COVRATIO_i < 1 - \frac{3k}{n},$$

then i^{th} point should be considered influential. The lower bound is only appropriate when $n > 3k$.

These cut-offs are only recommended for large samples.