Chapiter 8. Outliers and Influential Observations

Sometimes, while most of the observations fit the model and meet G-M conditions at least approximately, some of the observations do not. This occurs when there is Something wrong with the observations or if the model is faulty.

As far as the observations are concerned, there could have been a mistake in inputting or recording data. A few observations might reflect conditions or situations different from those under which other observations were obtained.

The leverage.

The leverages hii are the diagonal elements of

$H = X(X'X)^{-1}X'$. $h_{ii} = x_i'(X'X)^{-1}x_i$

The key features of a leverage h_{ii} is that it describes how far away the individual data point is from the centroid of all data points in the space of independent variables, i.e., how far removed x_i is from $\bar{x} = \frac{\sum x_i}{n}$.

$$\hat{y} = Hy$$

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i}^{n} h_{ij} y_j.$$

when h_{ii} is large, y_i influence $\hat{y}_i$

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

$$= \frac{1}{n} + (x_i - \bar{x})' \left[\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' \right]^{-1} (x_i - \bar{x})$$

$\sum_{i=1}^{n} h_{ii} = k+1$, if we had the option of choosing

(k: # of variables)

independent variable values we would choose them

So as to make each $h_{ii} = \frac{k+1}{n}$.

In the practice, a point i is a leverage point if

(1) $h_{ii} > \frac{2(k+1)}{n}$ criterion

(2) $h_{ii} > \frac{3(k+1)}{n}$ for $k+1 > 6$ and $n-k-1 > 12$.

(3) $h_{ii} > 0.5$

The Residuals

For the purpose of detecting observations that do not belong to the model, more valuable than the residuals are the Studentized residuals, often called RSTUDENT and defined by

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

$$e_i = y_i - x_i'\hat{\beta}$$

$$\text{var}(e_i) = \sigma^2(1-h_{ii})$$

$$e_i^{(\sigma)} = \frac{e_i}{\sigma\sqrt{1-h_{ii}}}$$

where e_i is a residual.

$S_{(i)}$ is equivalent to s if least squares is

Run after deleting the i^{th} case.

$y_{(i)}$ and $X_{(i)}$ the results of subtracting the i^{th} row from y and X

dim: $y_{(i)} : (n-1) \times 1$

$X_{(i)} : (n-1) \times (k+1)$

$$\hat{\beta}_{(i)} = (X'_{(i)} X_{(i)})^{-1} X'_{(i)} y_{(i)}$$

$$(n-K-2)\, s^2_{(i)} = \sum_{\substack{l=1 \\ l \neq i}}^{n} [\, y_l - x'_l \hat{\beta}_{(i)} \,]^2$$

Theorem:

(1) Relations between $\hat{\beta}$ et $\hat{\beta}_{(i)}$:

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

If $\hat{\beta} - \hat{\beta}_{(i)}$ is large, i is an outliar.

(2) under G-M conditions,

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

(3) Relations between $S^2_{(i)}$ and S^2:

$$(n-K-2)\,S^2_{(i)} = (n-K-1)\,S^2 - \frac{e_i^2}{1-h_{ii}}$$

Detecting Outliers and Points That do not belong to the model.

An outlier is an observation that is poorly explained by the model and has a high residual $|e_i^*| > 2$.

For a search for points that might not fit the model, the Studentized residuals are much more useful. One could examine a listing or a plot of these against case numbers. One can easily identify from these which e_i^* are significant at some given level. e.g 5%.

(1) An outlier if $|e_i^*| > t_{n-K-2;\frac{\alpha}{2}}$

(2) Normal probability plot.

(3) box-plot

to detect outliers.

Influential Observations.

An influential point is an observation that has a major impact on the parameter estimation process. It should be noted that not all the influential points have high values of studentized residuals.

Measures of influence:

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

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

$a_{ji} = [(X'X)^{-1} x_i]_j \quad (X'X)^{-1} x_i = (a_{0i}, \cdots a_{ki})'$

(3) $DFFIT_j = \hat{y}_i - \hat{y}_{i(i)} = x_i' \hat{\beta} - x_i' \hat{\beta}_{(i)} = \frac{h_{ii} e_i}{1 - h_{ii}}$

(14) DFBETAS$_{ij}$ (standard DFBETA$_{ij}$) =

$$\frac{\hat{\beta}_j - \beta_{j(i)}}{S_{(i)}\sqrt{g_{jj}}} = \frac{a_{ji}\, e_i}{S_{(i)}(1-h_{ii})\sqrt{g_{jj}}} = \frac{a_{ji}\, e_i^*}{\sqrt{g_{jj}(1-h_{ii})}}$$

g_{jj} : j^{th} diagonal element of $(X'X)^{-1}$.

DFBETASij to examine the i^{th} observation's influence on the estimation.

Criterion : $\frac{2}{\sqrt{n}}$ if $> \frac{2}{\sqrt{n}}$, requires investigation

(15) DFFITS$_i$ (Standard DFFIT$_i$)

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

$\text{var}(\hat{y}_{i(i)}) = S^2_{(i)}\, h_{ii}$ $\quad h_{ii} = \frac{(k+1)}{n}$ $\quad |e_i^*| = 2$.

to inform the change in the predicted value with and without the observation.

Criterion level: if $> 2\sqrt{\frac{k+1}{n-k-1}} \Rightarrow$ requires investigation.

(6) A very frequently used measure of influence can be defined as the distance between $\hat{\beta}$ and $\hat{\beta}_{(i)}$ standardized by the estimated covariance matrix $s^2(X'X)^{-1}$ of $\hat{\beta}$.

Cook's distance: $$\frac{(\hat{\beta}-\hat{\beta}_{(i)})'(X'X)(\hat{\beta}-\hat{\beta}_{(i)})}{(k+1)s^2}$$

$$\sim F_{k+1;\, n-k-1}.$$

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

$$\Rightarrow \text{Cook's distance} = \frac{h_{ii}\,(e_i^{(s)})^2}{(k+1)(1-h_{ii})}$$

influential observation when Cook's distance is large.

$F_{k+1,\, n-k-1;\, \alpha=0.5} \simeq 1.$ or $\frac{4}{n}$ (criterion)

(7) Covariance Ratio measures the effect on the estimated covariance matrix of $\hat{\beta}$.

$$\frac{\det\left[s^2_{(i)} \left(X'_{(i)} X_{(i)} \right)^{-1} \right]}{\det\left[s^2 \left(X'X \right)^{-1} \right]}$$

A value of this ratio close to 1 would indicate lack of influence of the i^{th} data point.

$$\det\left[X'_{(i)} X_{(i)} \right] = (1 - h_{ii}) \det(X'X)$$

$$\Rightarrow \text{Covariance Ratio} = \frac{(n-k-1)^{k-1}}{(1-h_{ii})(n-k-2+e_i^{*2})^{k+1}}$$

when $|\text{Covariance Ratio} - 1| > \frac{3(K+1)}{n}$ $\Rightarrow$ investigation

- Outliers that deviate from the model are detected by analysis of the studentized residuals e_i^*. They do not necessarily influence the plot of regression line.

- Leverage points can have an impact on the estimation of the model. They have high value of hii. Their distance from the center of gravity is large.
- Influential observations have high DFBETAS and/or DFFITS. and/or Cook's distance. They play a large role in the regression line.