Dr.K.S.Bhanu

Department of Statistics

Institute of Science, Nagpur.

Semester III - Unit I

Outliers , leverage points and influential observations

Contents of the topic: Divided into five parts

A- What are outliers and influential observations?
B- Reasons for the presence of outliers/Influential observations:
C- Detection of leverage and outliers (Diagnostic plots):
D- Deletion Diagnostics (for influential observations
E- Treatment – Remedial measures.

- This topic deals with situations where one or a few observations are different-in some way-from the rest of the data.
- We distinguish between two ways a few points may differ from the remaining points.
- A data point might lie far from the general trend in the rest of the data: such a point is called an outlier.
- Sometimes, a statistical analysis is very sensitive to a single (or a few) data point(s), in the sense that if the value of this point is changed even slightly, the outcome of the analysis alters greatly. Such points are called leverage points ..

[A] Outliers & Influential observations: (Discordant observations):

Discordant observations are those observation the presence of which causes deviations from the four assumptions of the regression model.

(a)Outliers:

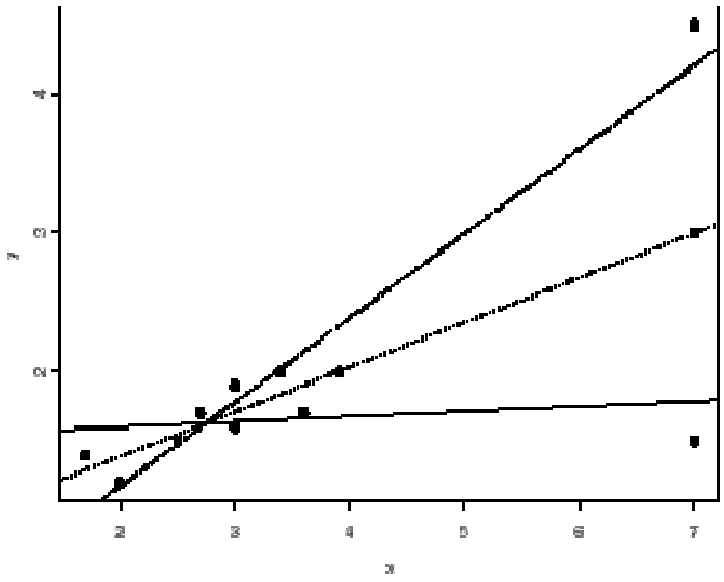
An outlier in a regression model is a data point that produces a large residual. In the space of the explanatory variables(X) if an observation is far removed (away) from the centre of the data it is referred to as an X-outlier. If the observation is such that the corresponding value of the response is far removed from the other values of the response variable that is called Y-outlier.

b) Leverage points:

Leverage :

In the regression setup, the values of the regressors (X) can be very high or very low. A single large or low value of X can pull the reg. line to one side .Such observations of X which are far away from the bulk of the data are called **leverage points**. These points can be treated as outliers in the space of the explanatory variables(x)

A leverage point is a point for which the observed value of this particular point has a great influence on the analysis. Suppose we have a cluster of data with x-values not too far apart; also, we have one observation corresponding to an x-value further away. The value of this isolated point is disproportionately influential on the least squares line: one might say that it works as a **lever-**if the value of this observation is changed, the least squares line changes considerably (as illustrated in the figure). In contrast, if the value of one of the points within the cluster is changed, the least squares line will not be affected to the same extent.



- The leverage of the ith observation is defined as the ith diagonal element of the hat matrix. $H = X(X' X)^{-1} X'$ (i.e hii)
- The ith leverage is a no. between 0 and 1 and it indicates the amount of leverage , or influence , the ith observation has on the least-squares line.
- The larger the value of hii, the more influence the observation has on the least squares line.
- A rule of thumb says that an observation is a leverage point if it has a hat-diagonal hii is greater than $2p/n$ where $p=k+1$.
- Note that the hat-matrix H only depends on the design matrix (X) and not on the response variables Yi.. That is, the observed value of the response variable is irrelevant with regard to whether or not a point (xi,yi) is a leverage point.

- The leverage point may be remote in the x-space yet may lie almost on the regression line passing through the rest of the sample points.
- Note that leverage points do not necessarily constitute a problem. If the observation y_i corresponding to a leverage point lies close to the general trend in the data, the point is called a good leverage point, and there is no reason to do anything about the data point.
- (Such points do not affect the estimates of the regression coefficients but may affect the value of R^2)

- However, if yi differs from the main trend-in particular, if yi corresponds to an outlier-the point is called a bad leverage point, and should be removed from the dataset.

(c) Influential observations:

Certain observations can considerably influence the analysis of the data due to their presence. They need not necessarily be the outliers or leverage points .These observations which individually or together with several other observations significantly influence the reg. analysis are called influential observations. An observation is influential if removing it would significantly change the position of the reg. line.

[B] Reasons for the presence of outliers/Influential observations:

1. Gross errors in measuring either the response or explanatory variables.

2. Need for additional variables in the model.

3. Need of an alternate model like a non linear model or a mixture model.

[C] Detection of leverage and outliers (Diagnostic plots):

1) using leverage (for X outliers):

In the case of simple linear regression we can plot the data point (X,Y) of the fitted model & identify the outlying observation. If the regression model is of the form

The hat matrix plays an important role in identifying influential observations. hii is taken as a measure of the distance of the ith observation from the center of the x-space. Large hat diagonals reveal observations that are potentially influential.

Any observation for which the hat diagonal exceeds 2p/n (where p =k+1 is considered to be a leverage point.

Note that not all leverage points will be influential.

Observations with large hat diagonals(for which hii is large) and large residuals are likely to be influential.

2) Residual Analysis (to identify Y-outliers):

Y-outliers can be identified by looking for an unusual pattern in the residual. The ith residual is given by,

$$\varepsilon i = yi - \hat{y}i; \quad i = 1,2, \ldots . . n$$

Plotting the residuals is a very effective way of investigating how well the regression model fits the data. If the model is correct, then the residuals ei should behave as zero mean and uncorrelated random variables with constant variance.

The residual ei measures the degree of misfit of the ith observation. Thus, residuals contain information which can be used to identify the outliers.

1. Standardized Residual:

$$d_i = \frac{\varepsilon_i}{\sqrt{MSRes}}$$

$$where, MSRes = \frac{SSRes}{n-p} = \frac{\sum {\varepsilon_i}^2}{n-p}; \quad p = k + 1$$

$$SSRes = \sum (Yi - \hat{y}i)^2$$

If di>3 ; ith data point is an outlier.

2. Studentised Residual:
   a) Internally studentised residual:

$$r_i = \frac{\varepsilon i}{\sqrt{MSE(1 - hii)}}$$

A very large value of r_i implies that ith observation can be an outlier.
A plot of the above two residuals against the fitted value yi (hat)can given an indication about the presence of outliers in the model.

   b) Externally Studentised residual (R-student statistcs):

$$t_i = \frac{e_i}{{\hat{\sigma}_i}^2 (1 - hii)}$$

Where, ${\hat{\sigma}_i}^2$is the estimate of σ^2 when the ith observation is left out.

3. R^2adjusted=$1 - \frac{\frac{SSRes}{n-p}}{\frac{SST}{n-1}}$

4. $Var(\hat{\beta}_j) = \sigma^2 Cjj \quad S.E(\hat{\beta}_j) = \sqrt{\sigma^2 Cjj}$

[D] DELETION DIAGNOSTICS (for influential observations): The influential observations can be identified using deletion diagnostics whereby the regression analysis is carried out both with and without the observation and the effect of deletion is investigated. The deletion of such an observation may change

1) The estimates of the regression coefficient$\hat{\beta}$.
2) The fit of that observation or other observations
3) The precision of the estimate i.e. V($\hat{\beta}$).
   These can be identified using deletion diagnostics.

1. DIAGNOSTICS FOR PARAMETER ESTIMATES:

They measure the effect of the influential observation on the estimates of the regression coefficient $\hat{\beta}$. The influence of deleting the ith observation on the jth regression coefficient $\hat{\beta}_j$ can be detected using the following measure:

DFBETAS= $S_{ij}(\beta) = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\hat{\sigma}_{(i)}{}^2 Cjj}}$

Where C = (Cij) = $(X'X)^{-1}$

And $\hat{\beta}_{j(i)}$ and $\hat{\sigma}_{(i)}{}^2$ are the estimates of$\hat{\beta}_j$ and $\hat{\sigma}^2$ after deleting the ith observation. A large value of Sij (β) indicates that the ith observation has considerable influence on the jth regression coefficient.

The suggested cutoff point for Sij (β) is 2/√n.

i.e if |Sij (β)|>2/√n it implies that the ith observation could be an influential observation.

2. DIAGNOSTICS FOR FIT:

This diagnostic method is used to investigate the effect of deleting the ith observation on the predicted or fitted value. The change in the fit of the ith observation when it is deleted is obtained using.

$$\text{DFFITSi} = \text{Si(FIT)} = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{\hat{\sigma}_{(i)}{}^2 hii}}$$

Where, $\hat{y}_{(i)}$ is the fitted value of yi obtained by deleting the ith observation. It can be shown that

$$\text{Si(fit)} = \text{ti}\sqrt{\frac{\text{hii}}{1-\text{hii}}}$$

Where ti=R-student statistic

hii=leverage of the ith observation.

If the data point is an outlier, then ti will be large in magnitude while if the data point has high leverage, then hii will be close to 1. In either case, Si(fit) will be large.

Thus Si(fit) depends both on leverage as well as prediction error.

The suggested cutoff point for Si(fit) is $2\sqrt{p/n}$.

i.e. if Si(fit) > $2\sqrt{p/n}$ it implies that the ith observation could be an influential observation.

3. COOK'S DISTANCE:

To find whether a point is influential or not, it is desirable to consider the location of the point in both the X-space as well as the Y-space. The cook's distance is based upon the squared distance between the least square estimates $\hat{\beta}$

based on all n points and the estimates obtained by deleting the ith point say $\hat{\beta}_{(i)}$. This distance measure is given by(in vector form)

$$Di = \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)' X' X \left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{(p)(MSRes)} \quad \ldots\ldots (1) \qquad ; i = 1,2, \ldots . n$$

Where, $MSRes = \frac{SSRes}{n-p} = \frac{\sum {e_i}^2}{n-p}$

Point with large values of Di will have considerable influence on the least square estimate β. Points for which Di ≥ 1 are considered to be influential.

Since $\hat{Y} = X\hat{\beta}$

$$\hat{Y}_{(i)} = X\hat{\beta}_{(i)}$$

Expression (1) can be rewritten as

$$Di = \frac{\left(\hat{y}_{(i)} - \hat{y}\right)' \left(\hat{y}_{(i)} - \hat{y}\right)}{(p)(MSRes)} \quad \ldots\ldots\ldots \quad (2)$$

The numerator of Di is the squared Euclidean distance that the vector of fitted values moves when the ith observation is deleted.

The Di statistic can also be written as

$$Di = \frac{{r_i}^2}{p}\left(\frac{hii}{1 - hii}\right) \quad = \frac{1}{p}$$

$$(square\ of\ the\ studentised\ residual)(\frac{hii}{1 - hii})$$

Thus Di combines the residual magnitude of the ith observation given by ri and the leverage of the ith point hii in the in the X-space to find the influence of the ith observation.

The magnitude of Di is compared with Fp,n-p,.

i.e if Di ≥ Fp,n-p, it implies that the ith point is highly influential.

4.PRECISION DIAGNOSTICS:

These measures study the influence of deleting the ith observation on the precision of an estimator through its variance and are called precision diagnostics.

We define the generalized variance of β as

$$\mathrm{GV}\ (\hat{\beta})\ = |\mathrm{var}\ (\hat{\beta})| = |\sigma^2\ (X'X)^{-1}|$$

To express the role of the ith observation on the precision of estimation, we define

$$\mathrm{COVRATIO(i)} = \frac{\left|\hat{\sigma}_{(i)}{}^2\left(\mathrm{X'}_{(i)}\mathrm{X}_{(i)}\right)^{-1}\right|}{|\mathrm{MSRES}(\mathrm{X'X})^{-1}|}$$

Where, MSRES = $\hat{\sigma}^2\ =$ U. E. of σ^2

$$= \frac{\left|\hat{\sigma}_{(i)}{}^2(\mathrm{X'}_{(i)}\mathrm{X}_{(i)})^{-1}\right|}{|\hat{\sigma}^2(\mathrm{X'X})^{-1}|}$$

$$= \left(\frac{\hat{\sigma}_{(i)}{}^2}{\hat{\sigma}^2}\right)^{p}\left(\frac{1}{1-(hii)}\right)$$

If COVRATIO (i)>1, the ith observation improves the precision of estimation, while if COVRATIO (i) <1, inclusion of the observation degrades the precision.

CUTOFF: The suggested cutoff point for COVRATIO is 1+3 p/n.

I.e. if COVRATIO>1+3 P/N. or if COVRATION<1-3 P/N., then the ith point should be considered influential.

NOTE:A high leverage point will make the COVRATIO(i)large. This is because a high leverage point will improve the precision unless the point is an outlier in the Y-space.

E] TREATMENT OF OUTLIERS / INFLUENTIAL OBSERVATIONS.

1. The presence of outliers / influential observations should be carefully investigated to see if a reason for their unusual behavior can be found. Care should be taken to see if the outlier is important to the model before it is discarded. The effect of outliers on the regression model may be checked by dropping these points and refitting the regression model. If the estimates of the parameters of the model do not change much, we can conclude that the points are not influential observation. Hence these points can be safely deleting these points, the precision of the estimates can be improved and the width of the C.I of β can be decreased.
2. If the influential observation is due to an error in measurement or if the sample point is really not needed for the model, then it can be safely discarded. However, if the analysis reveals that an influential point is a valid observation, then it should be retained.

These decisions are taken using robust estimation procedures
