LINEAR REGRESSION ANALYSIS

MODULE – VI

Lecture - 23

Tests for Leverage and Influential Points

Dr. Shalabh

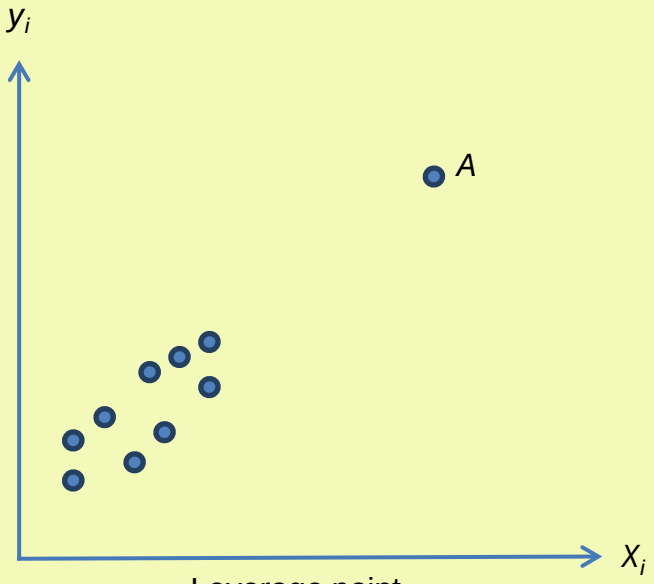
Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

The location of observations in *x*-space can play an important role in determining the regression coefficients.

Consider a situation like in the following

The point *A* in this figure is remote in *x*-space from the rest of the sample but it lies almost on the regression line passing through the rest of the sample points. This is a **leverage point.**



Leverage point

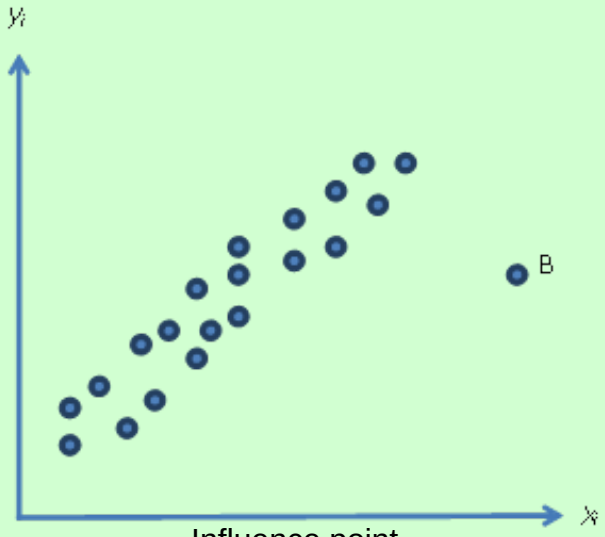
It is an unusual *x*-value and may control certain model properties.

- This point does not affect the estimates of the regression coefficients.
- It affects the model summary statistics e.g., R^2, standard errors of regression coefficients etc.

Now consider the point *B* following figure:

This point has a moderately unusual *x*-coordinate and the *y*-value is also unusual. This is an **influence point**

- It has a noticeable impact on the model coefficients.
- It pulls the regression model in its direction.



Influence point

Sometimes a small subset of data exerts a disproportionate influence on the model coefficients and properties.

In an extreme case, the parameter estimates may depend more on the influential subset of points than on the majority of the data. This is an undesirable situation.

A regression model has to be a representative of all the sample observations and not only of a few. So we would like to find these influential points and asses their impact on the model.

- If these influential points are “bad” values, they should be eliminated from the sample .
- If nothing is wrong with these points but if they control the model properties then it is to be found that how do they effect the regression model in use.

Leverage

The location of points in *x*-space affects the model properties like parameter estimates, standard errors, predicted values, summary statistics etc.

The hat matrix $H = X(X'X)^{-1}X'$ plays an important role is identifying the influential observations. Since

$$V(\hat{y}) = \sigma^2 H$$

$$V(e) = \sigma^2 (I - H),$$

($\hat{y}$ is fitted value and e is residual) so the elements h_{ij} of H may be interpreted as the amount of leverage excreted by the i^{th} observation y_i on the i^{th} fitted value $\hat{y}_i$.

The i^{th} diagonal element of H is

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

where x_i' is the i^{th} row of X-matrix. The i^{th} diagonal element of this hat matrix is a standardized measure of the distance of i^{th} observation from the center (or centroid) of the x-space. Thus large diagonal elements reveal observations that are potentially influential because they are remote in X-space from the rest of the sample.

Average size of hat diagonal is

$$\begin{aligned}\bar{h} &= \frac{\sum h_{ii}}{n} = \frac{\text{rank}(H)}{n} \\ &= \frac{\text{rank}(X)}{n} \\ &= \frac{trH}{n} \\ &= \frac{k}{n}.\end{aligned}$$

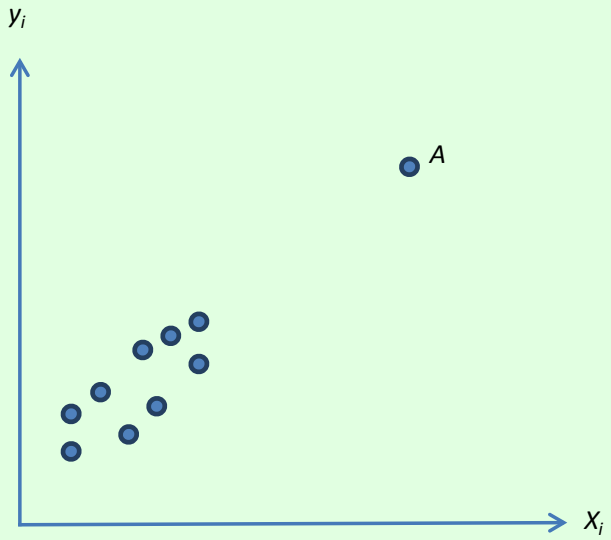
- If $h_{ii} > 2\bar{h} = \frac{2k}{n} \Rightarrow$ the point is remote enough from rest of the data to be considered a leverage point.

- Care is needed in using cutoff value $\frac{2k}{n}$ and magnitudes of k and n are to be assessed. There can be situations where $\frac{2k}{n} > 1$ and then this cut off does not apply.

All leverage points are not influential on the regression coefficients.

In the following figure, the point *A*

- will have a large hat diagonal and is surely a leverage point.
- have no effect of the regression coefficients as it lies on the same line passing through the remaining observations.



Hat diagonal examine only the location of observations in *x*-space, so we can look at the studentized residual or *R*-student in conjunction with the h_{ii}.

Observation with

- large hat diagonal and
- large residuals

are likely to be influential.

Measures of influence

(1) Cook's *D*-statistics

If data set is small, then the deletion of values greatly affects the fit and statistical conclusions.

In measuring influence, it is desirable to consider both

- the location of point is *x*-space and
- the response variable.

The Cook's distance statistics, denoted as Cook's *D*-statistic, is a measure of distance between the least-squares estimate based on all *n* observations in *b* and the estimate obtained by deleting the i^{th} point, say $b_{(i)}$. It is given by

$$D_i(M,C) = \frac{(b_{(i)} - b)' M (b_{(i)} - b)}{C}; \quad i = 1, 2, ..., n.$$

Usual choice of *M* and *C* are

$$M = X'X$$

$$C = kMS_{res}.$$

So

$$D_i(X'X, kMS_{res}) = \frac{(b_{(i)} - b)' X'X (b_{(i)} - b)}{kMS_{res}}; \quad i = 1, 2, ..., n$$

$$= \frac{(\hat{y} - \hat{y}_{(i)})'(\hat{y} - \hat{y}_{(i)})}{kMS_{res}}.$$

where $\hat{y} = Xb$

$\hat{y}_{(i)} = Xb_{(i)}$

$b = (X'X)^{-1}X'y.$

Points with large $D_i \Rightarrow$ the points have considerable influence of OLSE *b*.

Since

$$\frac{(b_{(i)} - b)'X'X(b_{(i)} - b)/k}{SS_{res}/(n-k)}$$

looks like a statistic having a $F(k, n-k)$ distribution. Note that this statistics is not having a $F(k, n-k)$ distribution.

Therefore the magnitude of D_i is assessed by comparing it with $F_\alpha(k, n-k)$. If $D_i = F_{0.5}(k, n-k)$, then deleting point *i* would move $b_{(i)}$ to the boundary of an approximate 50% confidence region for β based on the complete data set.

This displacement is large and indicates that the OLSE is sensitive to the i^{th} data point.

- Since $F_{0.5}(k, n-k) \approx 1$, we usually consider that points for which $D_i > 1$ to be influential.
- Ideally, each $\hat{\beta}_{(i)}$ is expected to stay within the boundary of a 10-20% confidence region.
- D_i is not an *F*-statistic but cut off of 1 works very well in practice.

Since $$D_i = \frac{(\hat{y}_{(i)} - \hat{y})'(\hat{y}_{(i)} - \hat{y})}{kMS_{res}},$$

so D_i can be interpreted as a squared Euclidian distance (apart from kMS_{res}) that the vector of fitted values moves when the i^{th} observation is deleted.

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

the expression of D_i can be written as

$$D_i = \frac{(b_i - b_{(i)})'X'X(b_i - b_{(i)})}{kMS_{res}}$$

$$= \frac{x_i'(X'X)^{-1}(X'X)(X'X)^{-1}x_i e_i^2}{(1-h_{ii})^2 kMS_{res}}$$

$$= \left(\frac{e_i}{1-h_{ii}}\right)^2 \left(\frac{h_{ii}}{kMS_{res}}\right)$$

$$= \frac{r_i^2}{k}\left(\frac{h_{ii}}{1-h_{ii}}\right)$$

where r_i is studentized residual.

D_i: product of squared i^{th} studentized residual and $\frac{h_{ii}}{1-h_{ii}}$.

: Reflects how well the model fits the i^{th} observation y_i and a component that measures how far that point is from the rest of the data.

: Either component or both may contribute to a large value of D_i.

: Thus D_i combines residual magnitude for i^{th} observation and location of points in x-space to assess influence.