STA3010 Regression Analysis

Feng YIN

The Chinese University of Hong Kong (Shenzhen)

yinfeng@cuhk.edu.cn

April 8, 2020

Overview

1. Leverage Point
2. Influential Point
3. Measures of Influence
4. Treatment of Influential Points

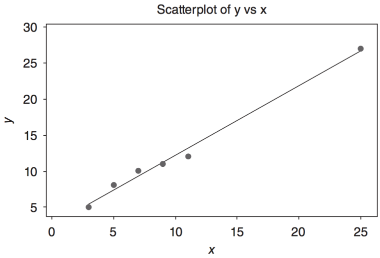
Leverage Point

- We have learned that $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ and h_{ii} is the i-th diagonal element of the hat matrix $\mathbf{H}$. The metric h_{ii} is also known as "leverage score" for the i-th data point $(y_i, \mathbf{x}_i)$. Every data point is a leverage point.
- Alternatively, the leverage score can be computed as $h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}$, in other words, it quantifies the weight of the output y_i in predicting $\hat{y}_i$. Larger h_{ii} value means higher weight, and vice versa.
- If the distance $||\mathbf{x}_i - \bar{\mathbf{x}}||$ is large, h_{ii} is in general also large.

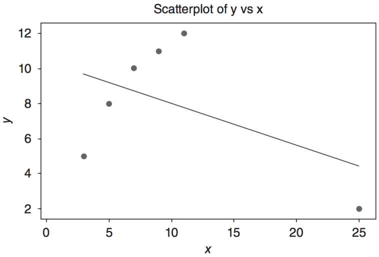
High-Leverage Point

Rule-of-Thumb: Treat any data point whose leverage score, h_{ii} is 2 times larger than the mean leverage value, i.e., $\bar{h} = \frac{p}{n}$, where n is the total number of data points, as a high-leverage point.

Two illustrating examples of high-leverage point:



NO influence on the LS fit!



Significant influence on the LS fit!

High-leverage point can be angel or devil!

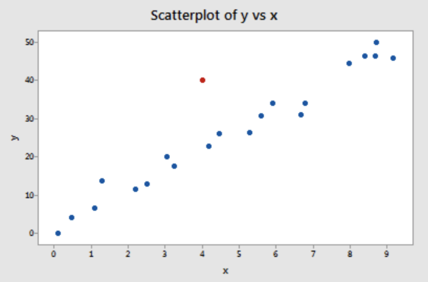
source:depositphotos.com

Influential Point

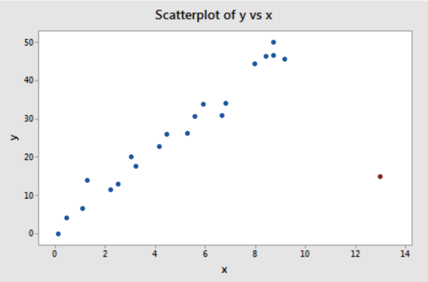
- In this lecture, influential points are defined to be the data points that considerably influence the parameter fit.
- More broadly, influential points are defined to be the data points that considerably influence the parameter fit in the training phase, inference/prediction in the test phase, and analysis.

Influential Point

Another view: Influential points could be ordinary-leverage points!



(a) Influential ordinary-leverage point



(b) Influential high-leverage point

Important observation: outputs with large h_{ii} and "large deviation from the expected profile" are likely to be high influential points.

But a high influential point is sly and good at hiding itself by attracting the fitting line...

Influential Point versus Outlier

- An influential point, in general, is a point that influences the parameter fit significantly.
- An outlier, often refers to a data point that demonstrates a distinct "pattern" as compared to the majority.

Relationship:

- An outlier is not necessarily an influential point.
- An influential point can be an outlier. But outliers are not always bad points that need to be removed.

Detection of Influential Points

Aim

There may exist influential points in the dataset and our aim is to identify them (especially the high influential ones).

Detection of influential points can help:

- dig out suspicious data subject to “man-made” errors;
- identify multi-modal data.

Quantitative Measures of Influence

So far, we haven't provided any measure to quantify the impact of an influential point.

In the following, we mainly consider two measures (out of three) that are more widely used to quantify the influence.

1. Cook's measure (proposed by Cook in 1977)
2. DFFITS meaure [1] (proposed by Belsley, Kuh, and Welsch in 1980)
3. DFBETAS [2] measure (proposed by Belsley, Kuh, and Welsch in 1980)

[1] DFFITS: difference in fitted output values

[2] DFBETAS: difference in fitted model parameters

Cook's Measure

- Consider multiple linear regression model first!
- Focus on single influence point first, then extend it to multiple points.
- Cook proposed a measure of the squared distance between the LS estimate based on all data points, i.e., $\hat{\beta}$ and the LS estimate obtained with the i-th data point deleted, i.e., $\hat{\beta}_{(i)}$.
- The squared distance measure is expressed in general form as:

$$D_i(\mathbf{M}, c) = \frac{\left(\hat{\beta} - \hat{\beta}_{(i)}\right)^T \mathbf{M} \left(\hat{\beta} - \hat{\beta}_{(i)}\right)}{c}. \tag{1}$$

Cook's Measure

Concretely, let $\mathbf{M} = \mathbf{X}^T\mathbf{X}$ and $c = p \cdot MS_{Res}$, we have

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)^T \mathbf{X}^T\mathbf{X} \left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)}{p \cdot MS_{Res}} < ? \tag{2}$$

- The magnitude of D_i is usually compared upon $F_{\alpha,p,n-p}$.
- For instance, if $D_i = F_{0.5,p,n-p}$, then deleting point i would move $\hat{\boldsymbol{\beta}}_{(i)}$ to the boundary of an approximate 50% confidence region (CR) of $\hat{\boldsymbol{\beta}}$.
- Since $F_{0.5,p,n-p} \approx 1$ for moderate n and not too small p, we usually consider points for which $D_i > 1$ to be influential.

Joint CR of Model Parameters (Optional)

We can prove for Gaussian i.i.d. random errors that

$$\frac{\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^T \mathbf{X}^T\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)}{p\cdot MS_{Res}} \sim F_{p,n-p}. \tag{3}$$

Consequently, a $100(1-\alpha)$ percent joint CR for $\boldsymbol{\beta}$ is

$$\frac{\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^T \mathbf{X}^T\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)}{p\cdot MS_{Res}} \leq F_{\alpha,p,n-p}. \tag{4}$$

However, this joint confidence region is impractical for $p > 2$ as it is difficult to visualize.

Joint CR of Model Parameters (Optional)

The equation:

$$\frac{\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{T}\mathbf{X}^{T}\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)}{p\cdot MS_{Res}}=F_{\alpha,p,n-p} \tag{5}$$

indicates an elliptical boundary, see the following figure with $p=2$:

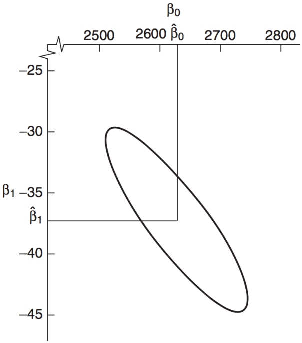


Figure 3.8 Joint 95% confidence region for β_0 and β_1 for the rocket propellant data.

Cook's Measure

Practically, D_i is calculated according to

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}, \quad i = 1, 2, ..., n \tag{6}$$

where D_i is the product of (1) a constant $1/p$ and (2) the square of the i-th studentized residual and (3) a ratio $\frac{h_{ii}}{1-h_{ii}}$.

DFFITS Measure

Belsley, Kuh, and Welsch proposed to use instead

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S_{(i)}^2 h_{ii}}}, \quad i = 1, 2, ..., n, \tag{7}$$

where $\hat{y}_{(i)}$ is the fitted value of y_i obtained without using the i-th data point. The denominator is a standardization factor due to $var(\hat{y}_i) = \sigma^2 h_{ii}$.

Remark: $DFFITS_i$ is the number of estimated standard deviations that the fitted value $\hat{y}_i$ changes if the i-th data point has been removed.

DFFITS Measure

Practically, DFFITS measure is computed according to

$$DFFITS_i = \sqrt{\frac{h_{ii}}{1-h_{ii}}}t_i, \quad i = 1, 2, ..., n, \tag{8}$$

where t_i is the R-student residual.

Belsley *et.al.* suggest that any data point with $|DFFITS_i| > 2\sqrt{p/n}$ warrants attention.

Remarks on the Influence Measures

1. Both the Cook's measure and DFFITS measure can be regarded as functions of the leverage score and residual, namely $g(h_{ii}, e_i)$ differ only in their specific form.
2. The cut-off values hold in general for large data records.
3. Selection of a proper cut-off value is application/case dependent for moderate data records.

Detecting Groups of Influential Points

Obviously, there could be situations where a group of points together exert undue influence on the regression mode.

We extend Cook's measure to assess the simultaneous influence of a group of m observations:

$$D_{\mathbf{i}} = \frac{\left(\hat{\boldsymbol{\beta}}_{(\mathbf{i})} - \hat{\boldsymbol{\beta}}\right)^T \mathbf{X}^T \mathbf{X} \left(\hat{\boldsymbol{\beta}}_{(\mathbf{i})} - \hat{\boldsymbol{\beta}}\right)}{p \cdot MS_{Res}} \tag{9}$$

where **i** denotes a vector of indices specifying which points are to be deleted.

Large values of $D_{\mathbf{i}}$ indicate that the set of m points in **i** are influential. But how to select the subset??

Influence Analysis for Nonlinear Regression Models

Representative works include:

- Cook and Weisberg, 1982, “Residual and inference in regression”
- Laurent and Cook, 1993, “Leverage, local influence and curvature in nonlinear regression”

Provided as additional readings for interested students.

Cook's Measure for Nonlinear Regression

Cook's measure for nonlinear regression is given as follows:

$$D_i = \frac{\left(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)}\right)^T \mathbf{Z}^T(\hat{\boldsymbol{\theta}})\mathbf{Z}(\hat{\boldsymbol{\theta}}) \left(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)}\right)}{p\hat{\sigma}^2}, \quad i = 1, 2, ..., n \qquad (10)$$

where

- p is the number of unknown nonlinear regression model parameters;
- $\hat{\boldsymbol{\theta}}$ is the parameter estimate trained with all n data points;
- $\hat{\boldsymbol{\theta}}_{(i)}$ is the parameter estimate trained with the i-th data point excluded from the whole data;
- $\mathbf{Z}(\hat{\boldsymbol{\theta}})$ is the Jacobian matrix of $\mathbf{f}(\mathbf{X}; \boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$;
- $\hat{\sigma}^2$ is a parameter estimate of the independent noise variance, σ^2.

Treatment of Influential Points

- Diagnostics for influential points offer the analyst insight about the data and alarm which observations may deserve more scrutiny.
- Should influential points/outliers ever be discarded?
- A compromise between deleting a data point and retaining it leads to robust regression that is less impacted than the ordinary least squares.

Summary

- Leverage point
- Influential point
- Influential point vs. Outlier
- Quantitative measure of influence
- Cook's measure, DFFITS measure for linear regression model
- Cook's measure for nonlinear regression model
- Discard an influential point or retain it? $\Longrightarrow$ Robust regression