Stat 151a Lecture Notes 17

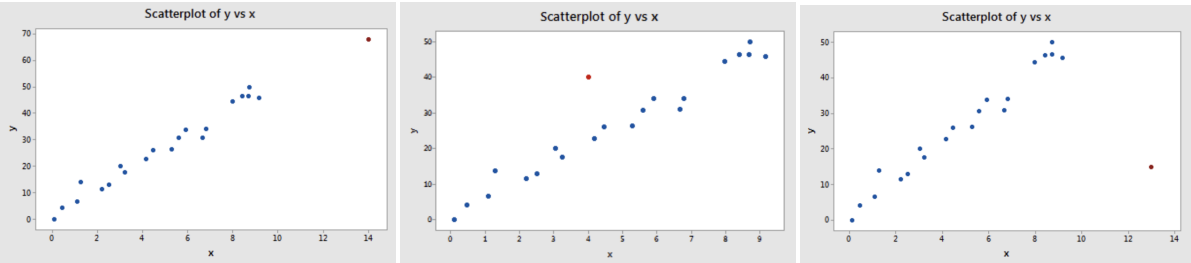
Chae Yeon Lee

May 1, 2023

1 Overview

Today:

- Outlier
- Leverage
- Influence



The first image shows high **leverage** point and the second image shows an **outlier**.The last image shows both an outlier and a high leverage. The first two image does not have influential point because when we create a model without the red point and construct a model, the model does not change much from when we include the red point and constructing the model. However, the last image includes a influential point because R^2 value, standard error, and p-value change significantly.

2 Outlier

Outlier is a data point whose response y does not follow the general trend of the rest of the data.

For multivariate case, the outlier is identified as the response (y_i) that is conditionally unusual given X_i.
Residuals are defined as $\sigma_\epsilon^2 Q$. However, the diagonal values of this Q matrix are not equal, which indicates that standardizing residuals are necessary.
So, one attempt would be to standardize the residuals.

$$\tilde{e_i} = \frac{e_i}{\sigma_\epsilon^2 \sqrt{I - H_i}}$$

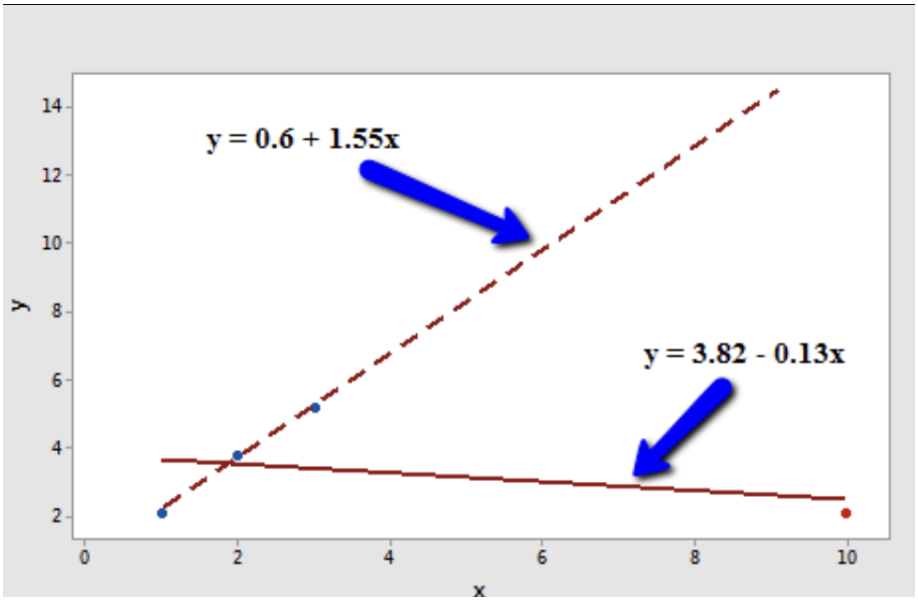
However, the problem with this equation is that e_i is related to σ_ϵ^2.
Thus, this brings **studentized residuals** into picture.
When trying to identifying outliers, one problem is that the potential outliers are so influential that the regression model is pulled towards the potential outlier, and thus the outlier is not flagged as an outlier. Studentized residuals address this issue.
Let

- y_i : observed response of the i^{th} observation.
- $\hat{y}_i$: predicted response for the i^{th} observation based on the estimated model with the i^{th} observation deleted.

Then, the **deleted residuals** are defined as

$$d_i = y_i - \hat{y}_i$$

The **basic idea** is to delete the observations one at a time, each time refitting the regression model on the remaining n-1 observations. Then, we compare the observed response value to their fitted value based on the models with the i^{th} model deleted. This produces **deleted residuals**, and standardizing the deleted residuals produces **studentized residuals.**



The estimated regression equation only with the three blue points are $y = 0.6 + 1.55x$. Based on this equation, when we predict the fourth (red) point, we have $y = 0.6 + 1.55(10) = 16.1$
The deleted residual for the red data point is then

$$d_4 = 2.1 - 16.1 = -14$$

.
Studentized residual is then

$$t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{MSE_{(i)}(1 - h_{ii})}$$

.
Another definition of studentized residual is

$$t_i = r_i\left(\frac{n-k-2}{n-k-1-{r_i}^2}\right)^{1/2}$$

where r_i is the ith standarized residual, n = number of observations, and k = number of predictors.
After calculating the studentized residual, we have to decide if this value is large enough by relying on the fact that studentized residuals follow a t-distribution with $(n-k-2)$ degrees of freedom.

3 Leverage

A data point has a high leverage is it has extreme predictor x values.
As we discusses previously, the predicted responses can be written as $\hat{y} = Hy$. We can also write the predicted responses as

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + h_{i3}y_3 + \cdot + h_{in}y_n$$

The leverage, h_{ii} quantifies the influence that the observed response y_i has on the predicted value $\hat{y}_i$. Therefore, if h_{ii} is small, then the observed response y_i plays a small role in the predicted response $\hat{y}_i$. If h_{ii} is large, then the obseved response plays a large role in the predicted response $\hat{y}_i$.
Properties of leverages:

- The leverage h_{ii} measures the distance between the x value for the i^{th} data point and the mean of the x values of all n data points.
- The leverage h_{ii} is the number $\in [0,1]$.
- The sum of h_{ii} equal $k+1$, which is the number of parameters.

4 INFLUENCE

A data point is influential if it has influence on other regression analysis, such as the predicted response, estimated slope coefficients, or the hypothesis test results. Outliers and high leverage data points have the potential to be influential but we need to further investigate in order to determine whether they are actually influential or not.
We use the cook's distance to identify the influential data points. The basic idea is to delete the observations one at each time, and each time refit the regression model on the remaining $n-1$ observations. then, we compare the results from all n observations to the results with the ith observation deleted to see how much influence that the observation has on the analysis.
Question: Why not we just use influence of ith data point =

$$\|\hat{\beta_{(-i)}} - \beta\|_2^2$$

Answer: This is treating that each β are in the same scale. However, each β may have different units, and therefore, just subtracting each beta and calculating the norm makes an assumption that each β are on the same scale.
Cook's distance

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(k+1) * MSE}\Big[\frac{h_{ii}}{(1-h_i i)^2}\Big]$$

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(-i)})^T X^T X(\hat{\beta} - \hat{\beta}_{(-i)})}{(p+1)\hat{\sigma_\epsilon^2}}$$

Note here that Cook's distance D_i depends on both the residual, e_i, and the leverage h_{ii}.