Jared Yu
Module 7 Discussion

- Discuss the utilities of various residuals covered in Chapter 4 for what purposes.

Response:

Thus far in the textbook there have been a few core assumptions of the linear regression model. They are described as follows:

1. Relationship between y and regressor $\mathbf{x}$ is linear
2. The error term ε has zero mean
3. ε has constant variance σ^2
4. The errors are uncorrelated
5. The errors are normally distributed

The ordinary residual derived from the linear regression model is defined as the following,

$$e_i = y_i - \hat{y}_i, \qquad i = 1, \cdots, n.$$

By doing an analysis of these residuals, they help in guiding those using linear regression to try to check and see if assumptions of the linear regression model are being met. This can be done by plotting and through statistical tests, but the textbook tends to favor the former over the latter.

The ordinary residual is not always the best at trying to understand the assumptions, however. There are a variety of issues that arise and so having different types of residuals to try and get a better understanding of the model relative to the characteristics of the data is important.

A possibility is to perform scaling on the residuals, which is useful in identifying outliers that are perhaps impacting the ability of the model to perform as desired. Of these there are the *standardized residuals* and *studentized residuals*.

The standardized residual is as follows,

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}}, \qquad i = 1, \cdots, n,$$

which have approximately zero mean and $Var(e_i) \approx 1$. Therefore, in the case where d_i is large, e.g., $d_i > 3$, then that is indicative that a subject from the dataset is an outlier.

In the previous approach, the methodology is to divide by the square root of MS_{Res}, but this can be improved by dividing instead by the standard deviation of that i'th residual instead. This is possible by finding $Var(\boldsymbol{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$, where $Var(e_i) = \sigma^2(1 - h_{ii})$. The formula then for the studentized residuals is as follows,

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1 - h_{ii})}}, \qquad i = 1, \cdots, n.$$

A note about them is that they have constant variance, $Var(r_i) = 1$. The way to think about them is to see if both r_i and h_{ii} are large. This is indicative that it is an *influential point*.

An alternative is to look at *PRESS residuals*. These are based on looking at the *prediction error*, which is defined as follows,

$$e_{(i)} = y_i - \hat{y}_{(i)},$$

where $\hat{y}_{(i)}$ is the prediction for that subject when it is removed from the dataset before fitting a model. There is an alternative formulation that doesn't require fitting n different models and so this is more convenient. It depends on e_i and h_{ii}, so when both of these are large, it again is indicative of highly influential points. In the case where there's a large difference between e_i and $e_{(i)}$, that indicates that the model that lacks that subject will have poor prediction.

A note about some of the previous methods is that they are referred to as *internal scaling*, where they can be calculated based on the n points already modeled. Therefore there is the idea about using a calculation based on a dataset that excludes the i'th subject. It is called the $R\text{-}student$ and is defined as,

$$t_i = \frac{e_i}{\sqrt{S_{(i)}^2(1-h_{ii})}}, \qquad i = 1, \cdots, n,$$

where

$$S_{(i)}^2 = \frac{(n-p)MS_{Res} - \frac{e_i^2}{1-h_{ii}}}{n-p-1}.$$

Given that the i'th subject is highly influential, then $S_{(i)}^2$ will differ significantly from MS_{Res}. The test statistic for t_i follows a t_{n-p-1} distribution.

- Describe and discuss the difference between random errors and residuals.

Response:

The linear regression model can be expressed in matrix form as the following,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

In this formulation, it can be seen that the random error vector, $\boldsymbol{\varepsilon}$, is a term that helps to balance out the true response variable from the linear model. In the simple linear regression model, it can be seen as how far the response variable departs from the fitted line. However, in multiple linear regression the interpretation is not as simple.

It can be seen that it is in sense a random error, since it is unlikely in a set of real-world data that the response variable exactly fits the linear model of predictors. There is inevitably some variance from this linear formulation. We can make assumptions about this random variable, which have been stated above. For example, having zero mean, constant variance, and being normally distributed. This above linear regression model can be thought of as expressing the true population parameters of the real-world data, which generally are unknown and can only be estimated. Therefore, this random error remains some random variable.

The residuals however are calculated after fitting a model and calculating the difference between the fitted and true values. We can think of the residuals as being the deviation between the fitted model and the sampled data. So, analyzing them can help us to better understand how well the fitted model's assumptions conform to the sampled data itself. In the case of serious violations of the assumptions, that would indicate that the fitted model is inappropriate to use.

The residuals themselves are not independent, despite having zero mean and approximately average variance. This can be seen in how from their sum of squares, the degree of freedom is

$n - p$ rather than n. We can see that in the MS_{Res} calculation, we are utilizing the estimated values of the coefficients rather than their true values, hence the degree of freedom being the way it is.

- Discuss the impact of violation of the statistical assumptions on the validity of regression analyses.

Response:

Depending on the methodology used, for example using various types of residuals, plotting, or statistical tests, there can be different levels of insight into the data itself. These insights may allow the researcher to see that there is in fact a violation of the initial statistical assumptions of the linear regression model.

There is also the possibility for transformations on either the regressor and/or the response variable to help change the resulting residual plots. This can help when the response is between 0 and 1 or when the residual plots seem to indicate nonlinearity.

Before it has been stated that there are possible influential points in the dataset and so these can be damaging to the model. It can be dragging the fitted model towards one direction, and so this could be indicative that the current model is insufficient. A possibility is to fit polynomial terms to try and help lessen the impact of such outliers. This is useful when the residual plots indicate that the data is nonlinear.

Something else is the possibility of timeseries data showing that there is autocorrelation involved. In such an event, the normal linear regression model is insufficient and there is an entire area of time-series analysis that would be more appropriate for modeling and understanding such data.

These fixes are important when it becomes clear that the regression assumptions are not being met. If instead researchers proceeded with utilizing the model, despite violations of the assumptions, then the same approach on a new sample of data could give a wildly different model with an opposite conclusion about the data.

In certain cases, the outliers can be removed from the dataset and the model can be refitted. This could yield different results, but it could be important in the case that the outliers are having a negative impact on the resulting estimate. However, given that they are removed, it would be important then for this to be noted in the conclusion and results. For example, it should be mentioned that the given model is based on outlier removal having taken place.