Computational Statistics - Seminar 3

Andrew Struthers

February 2023

Honor code: I pledge that I have neither given nor received help from anyone other than the instructor or the TAs for all work components included here. – Andrew

Introduction

The purpose of this seminar is to discuss some influences on regression coefficients. We use regression analysis to understand the relationships between variables, as we have seen in previous assignments. These coefficients can, however, be heavily influenced by certain observations and data points. This report will discuss two measures, Cook's distance, and leverage, that can help us to identify these influences and the math behind both measures.

Cook's Distance Resampling Method

Cook's distance is a way for us to measure the influential outliers from a set of predictor variables when doing regression analysis. The method works by running the regression to calculate various coefficients with and without the i^{th} observation. Cook's distance tells us the change in the regression coefficients when the i^{th} observation is removed from the analysis regarding the number of predictors and the sample size. This way, just having many predictor variables doesn't necessarily mean that each variable is super influential.

Cook's distance for the i^{th} observation is given by the following formula:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \cdot \text{MSE}}$$

where $\hat{Y}_j$ and $\hat{Y}_{j(i)}$ are the predicted values of the dependent variable for the j^{th} observation with and without the i^{th} observation respectively. The number of predictor variables in the model is p, the sample size is n, and MSE is the mean squared error of the regression model.

A general rule of thumb is if the value of Cook's distance for a specific i is more than three times as large as the mean of all Cook's distances from the model, the i^{th} observation might be an outlier. In general, larger values indicate a greater influence on the regression coefficients. The Cook's distance resampling method involves repeatedly estimating the regression model after deleting each observation one at a time and then computing the Cook's distance for each observation in each resampled model. This provides a more accurate estimate of the influence of each observation on the regression coefficients.

Leverage

Similar to Cook's Distance, "hat value" is a measure that we can use to test the influence of individual observations in a regression analysis. It is based on the diagonal elements of the "hat matrix", which takes observed values from the independent variables and maps them to predicted values of a dependent variable. The hat value for the i^{th} observation is the i^{th} element along the diagonal of the hat matrix. We see the hat value concept appear in a few different places.

Average Hat Value

The average hat value represents the average influence of the observations on the regression coefficients. The average hat value, $\bar{h}$, can be calculated by the following equation:

$$\bar{h} = \frac{(k+1)}{n}$$

where k is the number of predictors in the model and n is the number of datapoints. A larger average hat value indicate that the observations have a greater influence on the regression model.

Simple Regression

In simple linear regression, we can measure the distance between the i^{th} observation and the mean of the data. The equation to calculate the hat value for the i^{th} observation is:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

where x_i is the value of the predictor variable for the i^{th} observation and $\bar{x}$ is the mean of the predictor variable. The hat value is a function of the distance between the i^{th} predictor value and the mean predictor value. If this hat value is much larger than the average hat value, we can say that this predictor value has a larger influence on the regression coefficients.

Multiple Regression

Since the basic linear regression model with p predictor variables is $y = \mathbf{X}\beta + \epsilon$ where $\mathbf{X}$ is an $n \times p$ matrix, we can calculate the hat value for the i^{th} observation by:

$$h_{ii} = \mathbf{x}_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i^T$$

where $\mathbf{x}_i$ is the vector of predictor values for the i^{th} observation and $\mathbf{X}$ is the matrix of predictor values for all observations. Just like in simple reindicates, the hat value is a function of the distances between the i^{th} predictor values and the predictor values of all other observations.

Conclusion

Both Cook's distance and hat values are useful tools we can use to identify influential observations in our data and estimate their impact on the coefficients of the regression model we fit to the data. We can use Cook's distance or the hat value concepts to potentially highlight outliers or errors in measurement, both of which could be affecting our regression model. Overall, these are tools that we can apply when doing regression analysis to judge the impact of outliers and judge the overall influence each observation in our data has on the resulting regression model.

References

[1] Stephanie Glen. Cook's distanc/cook's d: Definition, interpretation. *Statistics How To: Elementary Statistics for the rest of us!*, Jan 2022.

[2] R. Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977.

[3] Cook's distance. *Cook's Distance - MATLAB & Simulink*, 2022.

[4] Michael H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li. *Applied Linear Statistical Models*. McGraw-Hill Education Private Limited, 2013.

[5] Till Bergmann. *Identifying outliers and influential cases*, Oct 2015.

[6] Using leverages to help identify extreme x values. *PennState: Statistics Online Courses*.