REGRESSION OUTLIERS AND INFLUENTIAL OBSERVATIONS USING FATHOM

Lindsey Bell
lbell2@coastal.edu

Keshav Jagannathan
kjaganna@coastal.edu

Department of Mathematics and Statistics
Coastal Carolina University
P.O. Box 261954 Conway, SC 29528

1 Introduction

Simple linear regression is a concept that is taught in every introductory statistics course. Students are taught how to compute the equation of the regression line, interpret the slope and intercept, and sometimes how to judge the fit of the line using outliers and r^2. What is seldom taught in the introductory statistics classroom is the distinction between outliers and outliers of influence, hereafter referred to as influential observations.

Many statisticians (Cook [2], Hoaglin and Iglewicz [4], Liu et al. [6], and Belsley et al. [1] among others) have developed tools and discussed the importance of detecting outliers and influential observations. There are many reasons to detect outliers; removing observations that are from populations different from the population being considered, removing observations that have been designated as typographical errors, and distinguishing observations that are peculiar to name a few.

Jaykumar and Thomas [8] mention important applications in which detecting outliers plays an important role including fraud detection [9], computer network intrusions [11], and criminal activities in e-commerce [10].

Detecting influential observations is equally important because these observations may have a disproportionate effect on the slope of the regression line as well as any predictions made using the regression line.

For example, consider the data found in Mosteller and Tukey [7] regarding the fertility rates of 47 (french speaking) Swiss cantons in 1888. From Figure (1), we can see that the inclusion of a single point (Geneva; in red) appreciably changes the slope of the regression line. This data point therefore can be considered to be "influential" in the data set. One strategy, and a bad one at that, is to simply remove any of these types of observations. However, it is our opinion that each of these observations must be investigated in depth and considered on a case by case basis for potential exclusion.

Diagnostic measures associated with regression models is a topic that is covered in depth in most (if not every) upper division undergraduate course on regression models. In these courses, instructors typically first cover the validation of the regression

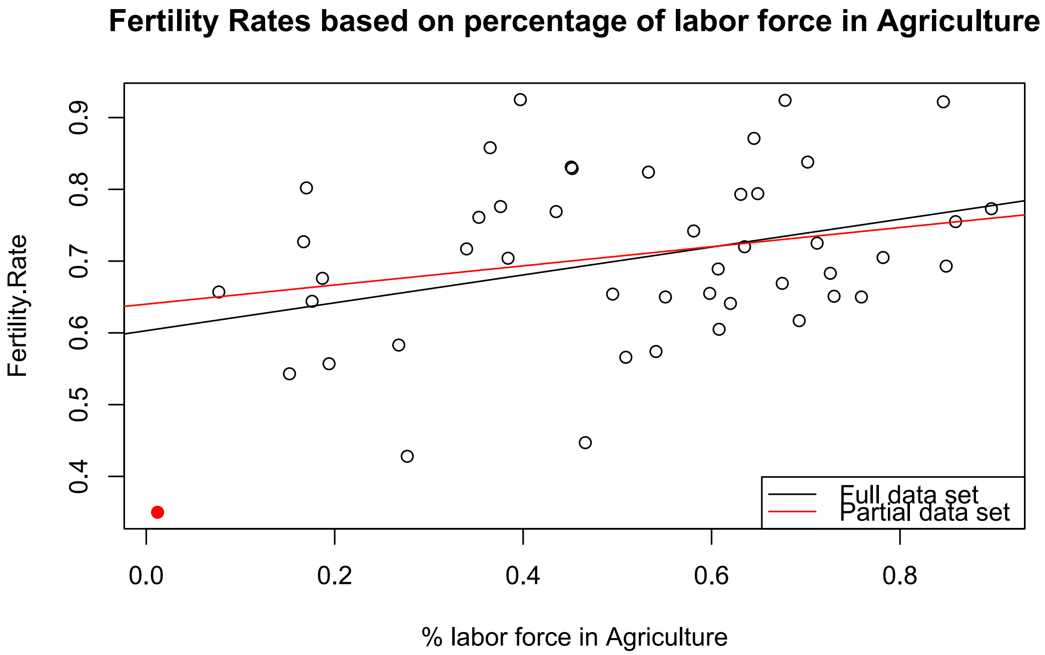

Figure 1: Fertility rates and percentage of labor force in agriculture in 1888

assumptions. Then, detection of outliers and influential observations using a variety of methods (residual plots, hat values, leverage, and Cook's distance etc.) is discussed. The mathematical level of these methods, while perfectly acceptable in an upper division course, is unsuitable for an introductory course in statistics.

The importance of influential observations in linear regression, as evidenced in the above example, necessitates the development of effective methods for teaching these topics in introductory statistics courses.

2 Methods

Technology in the classroom has made it easier for instructors to illustrate concepts; there is no longer a reason for instructors to not cover the topic of influential observations in the introductory statistics classroom.

Our purpose is twofold. First, we wish to provide a working "definition" for an influential observation that can be used in introductory level courses. Once we define an influential observation, we will compare and contrast it to outliers using several examples. We will then build some intuition into detecting these observations. Second, we wish to provide a hands-on activity that will further illustrate these concepts. The activity contains several portions that can either be used or discarded based on the instructor's needs and time constraints.

We will use Fathom [3] as our primary software for the activity. It is worth noting that while there are several software packages in the market that can be used, we chose Fathom for its ease of use and simplicity. If the instructor does not have access to Fathom, the activity can be easily modified for use with any statistical software.

2.1 Defining an Influential Observation

There are two key factors in determining whether an observation has influence: (1) the observation must be an outlier and (2) the observation must have "leverage".

2.1.1 Outliers

Many textbooks define an outlier as an observation that does not follow the general pattern of the data set. This definition, however, is vague and subjective, leaving it to the student to classify observations based on how he or she feels about that observation following the pattern of the data.

Further, the scale of a scatterplot could erroneously guide a student into classifying an observation as an outlier or miss classifying an outlier using this definition.

Other textbooks classify an observation as an outlier in the following manner.

Definiton 1. *An outlier is any point that falls more than two standard deviations away from its corresponding univariate mean.*

This definition works well for symmetric data but not so well when the data are skewed. This is due to the fact that neither the mean nor the standard deviation are resistant measures.

In upper level courses, students are taught to use studentized residuals or standardized residuals to identify outliers (see for example Kutner et. al. [5]). We now define those two terms.

Definiton 2. *The standardized residual of an observation y_i with predicted value $\hat{y}_i$ has a residual of $e_i = y_i - \hat{y}_i$. The standardized residual, $\tilde{e}_i$ can be calculated as*

$$\tilde{e}_i = \frac{e_i}{\sqrt{\widehat{Var}(e_i)}}$$

Definiton 3. *The studentized residual of an observation y_i, $\tilde{\hat{e}}_i$ can be calculated as*

$$\tilde{\hat{e}}_i = \frac{e_i}{\sqrt{\widehat{Var_i}(e_i)}}$$

where $\widehat{Var_i}(e_i)$ is the variance of the data set with the i^{th} observation removed.

The above definition, clearly, is not appropriate for an introductory statistics course. One might argue that the definition of the standardized residual is also rather advanced for introductory statistics students. However, it is our opinion that it makes more sense to teach students about scaled or standardized residuals. We will therefore use the following as our working definition of a regular outlier.

Definiton 4. *Any observation that has standardized residual greater than two is classified as an outlier.*

2.1.2 Leverage

The concept of leverage is almost never covered in the introductory course. Reasons for the exclusion of this topic could include the mathematical level and the lack of time. The formal definition of leverage is given in terms of "hat values" as follows.

Definiton 5. *Consider the matrix model for linear regression $\mathbf{Y} = \mathbf{X}\beta + \epsilon$. We can then express the fitted values of $\mathbf{Y}$, $\hat{\mathbf{Y}}$ as*

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

The hat matrix is then defined as $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and values in the hat matrix are called hat values. The diagonal elements h_{ii} of $\mathbf{H}$ are the leverages of the observations.

A large value of h_{ii} indicates that the i^{th} observation is distant from its univariate mean. But what constitutes "large"?

One guideline is that if $h_{ii} > \dfrac{2p}{n}$, where there are $p-1$ independent variables, the i^{th} observation is said to have large leverage.

A simpler guideline for identifying observations with "large" leverages is to classify observations with $h_{ii} > 0.5$ as having high leverage.

The above definition of leverage drives home the point that defining and computing leverages in an introductory course is time consuming and of an inappropriate level. There exists, however, some simple intuition for understanding what leverage is.

Consider the regression line. We know that the regression line always passes through the point $(\bar{x}, \bar{y})$. Let us think of the regression line as a teeter-totter that balances at that point. Now, if a child (observation) sits close to the center of the teeter-totter $(\bar{x})$, then the balance of the teeter-totter is not going to be severely affected. However, the further the child sits away from the balance point $(\bar{x})$, then more drastic the oscillation of the teeter-totter.

There are several methods used in upper division courses for identifying influential observations. Those methods include DFFITS, Cook's Distance and DFBETAS. Those methods are not practical for an introductory statistics course. We therefore present our "working" definition of an influential observation.

Definiton 6. *An observation is deemed to be influential if it is an outlier and is in a position of high leverage.*

For example, consider the data presented on DASL [12] on the per capita income of 20 OECD countries in 1960 based on percentage of labor force employed in industry. A scatterplot of the data is presented below in Figure (2).

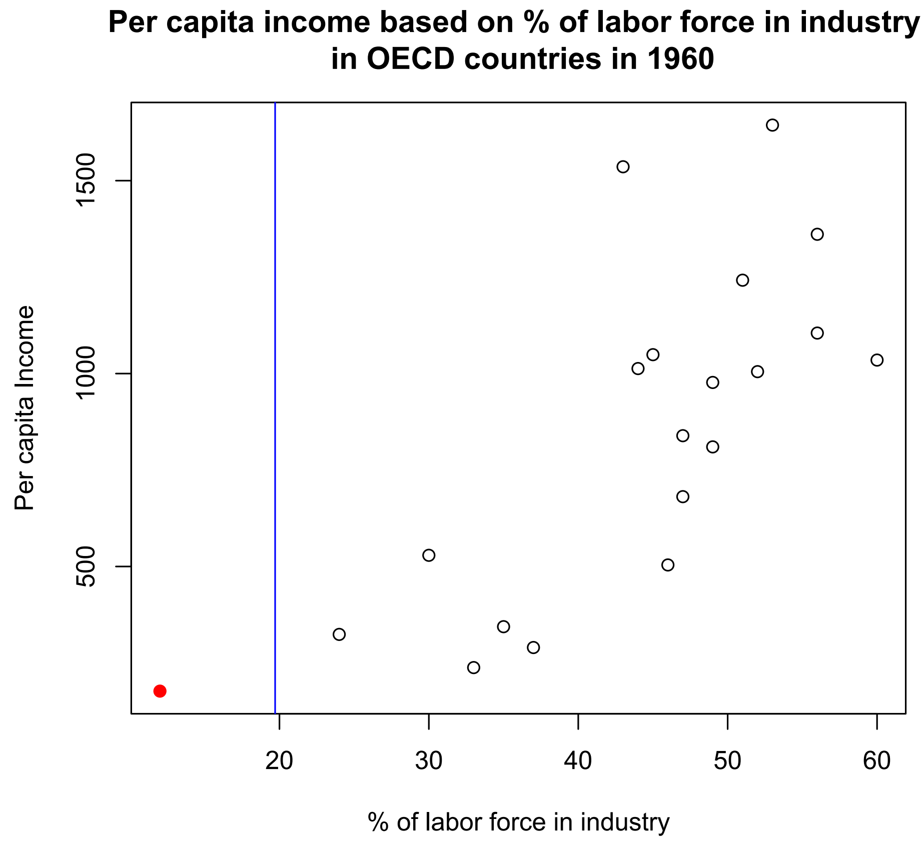

Figure 2: Per capita income and percentage of labor force in industry in 1960 OECD countries

The line in blue represents two standard deviations away from the mean of the independent variable. As we can see, the one observation (Turkey) in red has leverage on the regression line as it falls far away from the mean of the independent variable. Furthermore, it is the only observation in the data set that is in a position of leverage. Based on our definition, as it is an observation in a location of leverage and an outlier (falling more than 2 standard deviations from the mean), it is deemed to be an influential observation.

In Figure (3), we can see that excluding Turkey from consideration drastically changes the slope of the linear regression line.

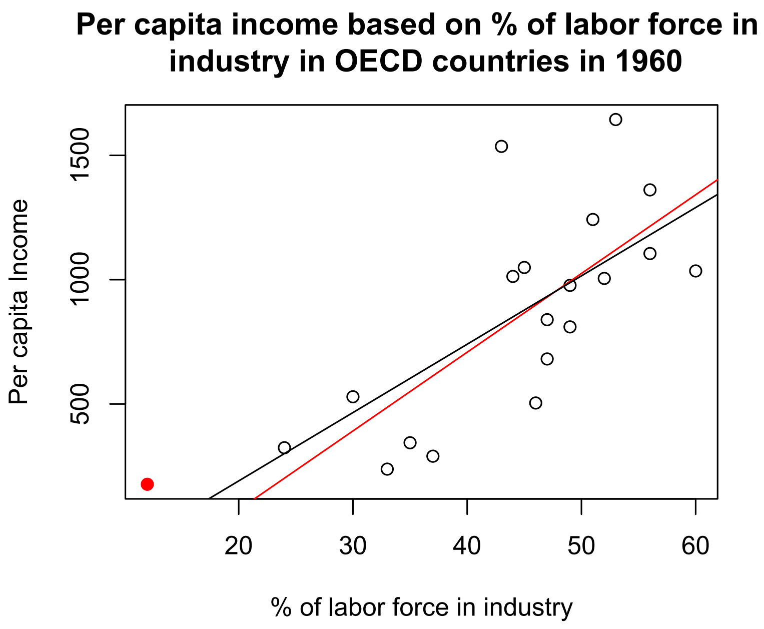

Figure 3: Per capita income and percentage of labor force in industry in 1960 OECD countries

Often times, in the classroom, we define influential observations as outliers in the x direction. It would be beneficial to do more than just talk about influential observations. It would be nice to have the students understand influential observations on a conceptual level. A hands on activity (included in the appendix) using Fathom [3] will hopefully help the students understand influential observations.

3 Conclusion

Not all higher level topics in statistics lend themselves nicely to the introductory statistics classroom. However, influential observations and outliers do just that. Given the importance of linear regression and predictions using regression lines, it is critical that students are able to do more than blindly compute regression lines and predicted values.

From a standpoint of making accurate predictions and being aware of observations that play an important role in the regression relationship, it is critical that students understand in detail the concepts of outliers and influential observations and the differences between the two.

The activity provided in the appendix is designed for a standard 50 minute lecture and in our opinion, adequately fulfills the goal of imparting a conceptual understanding of two critical concepts.

References

[1] Belsley, D.A., Kuh, E., Welsch, R.E., (1980). Regression diagnostics: Identifying influential data and sources of collinearity, *John Wiley*, New York, NY.

[2] Cook, R.D., (1977). Detection of influential observations in linear regression, *Technometrics*, **19**, 15-18.

[3] KCP Technologies, (2001). Fathom Dynamic Statistics Software, *Key Curriculum Press*, Emeryville, CA.

[4] Hoaglin, D. and Iglewicz, B., (1993). How to detect and handle outliers, *The ASQC Basic References in Quality Control: Statistical Techniques*, **16**.

[5] Kutner, M., Natchtsheim, C., Neter, J. (2004). *Applied Linear Regression Models*, 4th edition, *McGraw-Hill Irwin*, New York, NY.

[6] Liu, H., Shah, S., Jiang, W., (2004). On-line outlier detection and data cleaning, *Computers and Chemical Engineering*, **28**, 1635-1647.

[7] Mosteller, F., Tukey, J.W., (1977). Data Analysis and Regression: A Second Course in Statistics, *Addison-Wesley Series in Behavioral Science: Quantitative Methods*.

[8] Jayakumar, G.S.D.S., Thomas, B.J., (2013). A new procedure of clustering based on multivariate outlier detection, *Journal of Data Science*, **11**, 69-84.

[9] Bolton, R.J., Hand, D.J., (2002). Statistical fraud detection: a review, *Statistical Science*, **17**, 235-255.

[10] Chiu, A.L., Fu, A.W., (2003). Enhancement on local outlier detection, *Proceedings of the Seventh International Database Engineering and Applications Symposium (IDEAS03)*, 298-307.

[11] Lane, T., Brodley, C.E., (1999). Temporal sequence learning and data reduction for anomaly detection, *ACM Transactions on Information and System Security*, **2**, 295-331.

[12] Data and Story Library, `http://lib.stat.cmu.edu/DASL/`.