Prediction Intervals in the Presence of Outliers

David J. Olive*

Southern Illinois University

July 21, 2003

Abstract

This paper presents a simple procedure for computing prediction intervals when the data comes from a population that produces a small percentage of easily detected randomly occurring outliers. The multiple linear regression model with normal errors is used to illustrate the procedure.

KEY WORDS: Regression, Robust Statistics

*David J. Olive is Assistant Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. E-mail address: dolive@math.siu.edu. This research was supported by NSF grant DMS 0202922.

1 INTRODUCTION

Outliers are observations that are far from the bulk of the data and can cause classical estimators to perform very poorly. Typing and recording errors may create outliers, and a data set can have a large proportion of outliers if there is an omitted categorical variable (e.g. gender, species, or geographical location) where the data behaves differently for each category. Outliers should always be examined to see if they follow a pattern, are recording errors, or if they could be explained adequately by a more complicated model. Recording errors can sometimes be corrected and omitted variables can be included, but often there is no simple explanation for a group of data which differs from the bulk of the data.

Assume that the population that generates the data is such that a certain proportion γ of the cases will be easily identified but randomly occurring unexplained outliers where $\gamma < \alpha < 0.2$, and assume that remaining proportion $1 - \gamma$ of the cases will be well approximated by the statistical model.

A common suggestion for examining a data set that has unexplained outliers is to run the analysis on the full data set and to run the analysis on the "cleaned" data set with the outliers deleted. Then the statistician may consult with the collectors of the data in order to decide which analysis is "more appropriate." Although the analysis of the cleaned data may be useful for describing the bulk of the data, the analysis is not very useful if prediction or description of the entire population is of interest.

Similarly, the analysis of the full data set will likely be unsatisfactory for prediction since numerical statistical methods tend to be inadequate when outliers are present.

Classical estimators will frequently fit neither the bulk of the data nor the outliers well, while an analysis from a good practical robust estimator (if available) should be similar to the analysis of the cleaned data set.

Hence neither of the two analyses alone is appropriate for prediction or description of the actual population. Instead, information from both analyses should be used. The cleaned data will be used to show that the bulk of the data is well approximated by the statistical model, but the full data set will be used along with the cleaned data for prediction and for description of the entire population.

To illustrate the above discussion, consider the multiple linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{1.1}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of errors. The ith case $(Y_i, \boldsymbol{x}_i^T)$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element Y_i of $\boldsymbol{Y}$. Since prediction intervals are desired, also assume that the errors e_i are independent identically distributed (iid) random variables from a normal distribution with zero mean and variance σ^2.

Finding prediction intervals for future observations is a standard problem in multiple linear regression. Let $\hat{\boldsymbol{\beta}}$ denote the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$ and let

$$MSE = \frac{\sum_{i=1}^{n} r_i^2}{n-p}$$

where $r_i = Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ is the ith residual. Following Neter, Wasserman and Kutner (1983, p. 246), if the errors are iid normal, then a $(1-\alpha)100\%$ prediction interval (PI) for a new observation $Y_{h(new)}$ corresponding to a vector of predictors $\boldsymbol{x}_h$ is given by

$$\hat{Y}_h \pm t_{1-\alpha/2,n-p} se(pred) \tag{1.2}$$

where $\hat{Y}_h = \boldsymbol{x}_h^T \hat{\boldsymbol{\beta}}$, $P(t \leq t_{1-\alpha/2,n-p}) = 1 - \alpha/2$ where t has a t distribution with $n - p$ degrees of freedom, and

$$se(pred) = \sqrt{MSE(1 + \boldsymbol{x}_h^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_h)}.$$

For discussion, suppose that $1 - \gamma = 0.92$ so that 8% of the cases are outliers. If interest is in a 95% PI, then using the full data set will fail because outliers are present, and using the "clean" data set with the outliers deleted will fail since only 92% of future observations will behave like the "clean" data.

A simple remedy is to create a nominal $100(1 - \alpha)\%$ PI for future cases from this population by making a classical $100(1 - \alpha^*)$ PI from the clean cases where

$$1 - \alpha^* = (1 - \alpha)/(1 - \gamma). \tag{1.3}$$

Since this PI is valid when $Y_{h(new)}$ is clean, if no outliers will fall in the PI then

$$\mathrm{P}(Y_{h(new)} \text{ is in the PI}) \approx \mathrm{P}(Y_{h(new)} \text{ is in the PI and clean}) =$$

$$\mathrm{P}(Y_{h(new)} \text{ is in the PI} \mid Y_{h(new)} \text{ is clean})\ \mathrm{P}(Y_{h(new)} \text{ is clean}) \approx (1 - \alpha^*)(1 - \gamma) = (1 - \alpha).$$

Assume that there are n_c clean cases and n_o outlying cases where $n_c + n_o = n$. Then the formula for this PI is

$$\hat{Y}_h \pm t_{1-\alpha^*/2,n_c-p} se(pred) \tag{1.4}$$

where $\hat{Y}_h$ and $se(pred)$ are obtained after performing OLS on the n_c clean cases. For example, if $\alpha = 0.1$ and $\gamma = 0.08$, then $1 - \alpha^* \approx 0.98$. Since γ will be estimated from the data, the coverage will be approximately valid.

2 An Example

The following example illustrates the procedure. STATLIB provides a data set that is available from the website (http://lib.stat.cmu.edu/datasets/bodyfat). The data set (contributed by Roger W. Johnson) includes 252 cases, 14 predictor variables, and a response variable Y = *bodyfat*. The correlation between Y and the first predictor x_1 = *density* is extremely high, and the plot of x_1 versus Y looks like a straight line except for four points. If simple linear regression is used, the residual plot of the fitted values versus the residuals is curved and five outliers are apparent. The curvature suggests that x_1^2 should be added to the model, but the least squares fit does not resist outliers well. If the five outlying cases are deleted, four more outliers show up in the plot. The residual plot for the quadratic fit looks reasonable after deleting cases 6, 48, 71, 76, 96, 139, 169, 182 and 200. Cases 71 and 139 were much less discrepant than the other seven outliers.

These nine cases appear to be *outlying at random*: if the purpose of the analysis was description, we could say that a quadratic fits 96% of the cases well, but 4% of the cases are not fit especially well. If the purpose of the analysis was prediction, deleting the outliers and then using the clean data to find a 99% PI would not make sense if 4% of future cases are outliers. To create a nominal 90% PI for future cases from this population, make a classical $100(1-\alpha^*)$ PI from the clean cases where $1-\alpha^* = 0.9/(1-\gamma)$. For the bodyfat data, we can take $1 - \gamma \approx 1 - 9/252 \approx 0.964$ and $1 - \alpha^* \approx 0.94$. Notice that $(0.94)(0.96) \approx 0.9$.

Figure 1 is useful for presenting the analysis. The top two plots have the nine outliers deleted. Figure 1a is a forward response plot of the fitted values $\hat{Y}_i$ versus the response

Y_i while Figure 1b is a residual plot of the fitted values $\hat{Y}_i$ versus the residuals r_i. These two plots suggest that the multiple linear regression model fits the bulk of the data well. Next consider using weighted least squares where cases 6, 48, 71, 76, 96, 139, 169, 182 and 200 are given weight zero and the remaining cases weight one. Figure 1c and 1d give the forward response plot and residual plot for the entire data set. Notice that seven of the nine outlying cases can be seen in these plots.

ARC (Cook and Weisberg 1999) was used to make the residual plots to find outliers and to find prediction intervals. After making the residual plot, the outliers can be highlighted and deleted from the data set. It took less that five minutes to detect the outliers graphically. The classical 90% PI using $\boldsymbol{x} = (1, 1, 1)^T$ and all 252 cases was $\hat{Y}_h \pm t_{0.95,249} se(pred) = 46.3152 \pm 1.651(1.3295) = (44.12, 48.51)$. When the 9 outliers are deleted, $n_c = 243$ cases remain. Hence the 90% PI using equation (1.4) with 9 cases deleted was $\hat{Y}_h \pm t_{0.97,240} se(pred) = 44.961 \pm 1.88972(0.0371) = (44.89, 45.03)$. Notice that the classical PI is about 31 times longer than the new PI.

The focus of this article is on prediction intervals for multiple linear regression, but similar ideas hold for prediction intervals in the location model. Whitmore (1986) gives a useful introduction to prediction intervals in the location model, and Horn (1988) gives a partial solution for obtaining prediction intervals when the underlying distribution has heavy tails. Fisher and Horn (1994) discuss robust prediction intervals in the regression setting.

3 References

Cook, R.D., Weisberg, S. 1999. Applied Regression Including Computing and Graphics. Wiley, New York.

Horn, P.S., 1988. A biweight prediction interval for random samples. J. Amer. Statist. Assoc. 83, 249-256.

Fisher, A., Horn, P.S., 1994. Robust prediction intervals in a regression setting. Comput. Statist. Data Anal. 17, 129-140.

Neter, J., Wasserman, W., Kutner, M.H., 1983. Applied Linear Regression Models. Irwin, Homewood, IL.

Whitmore, G.A., 1986. Prediction limits for a univariate normal observation. Amer. Statist. 40, 141-143.

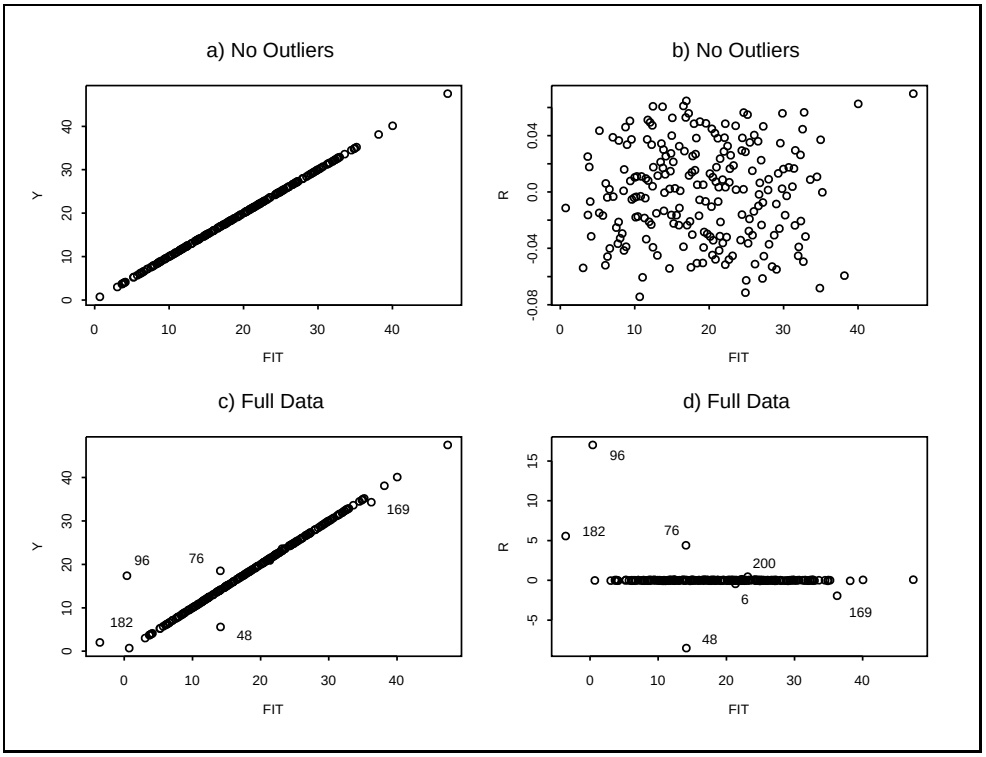

Figure 1: Plots for Summarizing the Entire Population