SESUG 2020 Paper 154

Outlier Detection and Treatment

Ross Bettinger, Silver Spring, MD

Abstract

Outliers in a set of data represent observations that are distinguished from the expected patterns in the observed data. We are attracted to them because they are somehow atypical of what we expect to see in the distribution of the data that we have collected or generated. We may be troubled by them because they may contain important information about the process which we are attempting to model, or they may simply be spurious noise or even data points that have been corrupted by noise. In any case, we must be careful in our treatment of them because we cannot easily decide whether they represent some valuable aspect of the process under study or are simply aberrations.

Keywords

Data distribution, binary indicator variable, Cook's D statistic, local outlier factor, local outlier probability, Mahalanobis distance, outliers, robust regression, studentized deleted residual, Winsorizing

Introduction

In the final week of the semester in a statistics class, the professor was reviewing the topics that were covered. He talked about exploratory data analysis, analysis of variance, regression, variable selection, sampling bias, hypothesis testing, data collection and data cleaning, and many more of the useful topics that students regularly encounter in their second statistics course. A student raised his hand and said, "Professor, can you give us more problems to work so we can practice our skills prior to the final exam?" And the professor turned to face him, pointed at him, smiled, and proclaimed "Outlier!"

This little anecdote highlights an important statistical topic that is often encountered in statistical and machine learning applications. What do we mean when we call an observation an outlier? What makes an outlying observation different from other observations? What, if anything, must we do when we have outlying observations in our data? Are outliers bad? Can we simply include them with the rest of our data and build models with them? Can we exclude them and conduct analyses without accounting for them?

Detecting Outliers in Data

An outlier is an observation that is so different from other observations that it attracts our attention.[1] It may be extreme in some sense, or simply "doesn't look right" to our trained eye. In the anecdote above, the professor labelled the student as an outlier perhaps because, in the professor's experience, no student had ever asked for more problems to work. He may have been an exception to the professor's experience-based characterization of the distribution of students' willingness to learn statistics.

[1] Typically, we are concerned with continuous data, but we may also consider the case for categorical data. For example, a rose growing in a vegetable garden may be considered to be an outlier. Or, a categorical value with unusually low frequency may be due to a data labelling error. Domain knowledge must be used in such cases so that appropriate remedial measures may be applied. Perhaps the observation is so atypical that it must be omitted from the analysis, or it can be relabeled and merged with another category based on additional characteristics.

A more formal definition is that an outlier is “an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism” [1]. Also, “an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs” [2].

Outlying observations cannot be ignored because they may affect the parameter estimation process. High-*leverage* points are extreme or outlying values of $\boldsymbol{x_i}$ that distort fitted regression model estimates by minimizing the error criterion, $\sum_i e_i^2$. If an observation has high *influence*, it will significantly change the parameter estimates of a fitted regression model if it is omitted from the calculation [3].

Detecting Outliers Using Univariate Regression

A first step in detecting outliers in data is to perform exploratory data analysis of individual variables. We will use Series III of Anscombe’s quartet[2] [4] to represent the importance of visual analysis before any statistical algorithms are applied. Table 1 indicates the Series III data points.

Table 1 Anscombe’s Quartet Series III Data

x	10.0	8.0	13.0	9.0	11.0	14.0	6.0	4.0	12.0	7.0	5.0
III_y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

Figure 1 shows the scatterplot of the data overlaid on the 45° line of exact fit between actual III_y and predicted III_y.

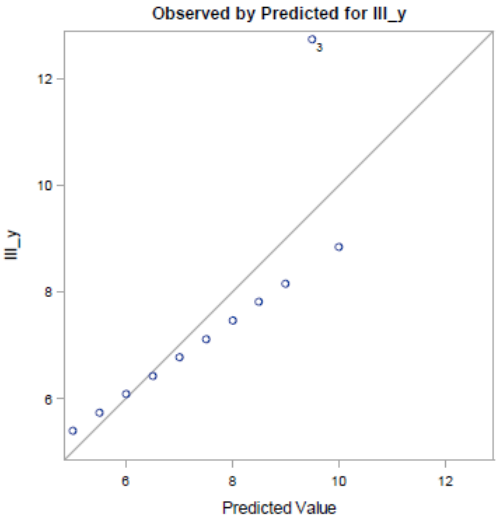


Figure 1 Anscombe’s Quartet Series III Scatterplot and Trend Line

We see that observation 3 (note the “3” subscript below the data point in the scatterplot) is an exception to the trend and biases the trend line upward. The coefficient of determination, R^2, is 0.6663, indicating moderate correlation between the dependent and independent variable. The equation of the trend line is

$$\widehat{III\_y} = 3.00245 + 0.49973 \cdot x \tag{1}$$

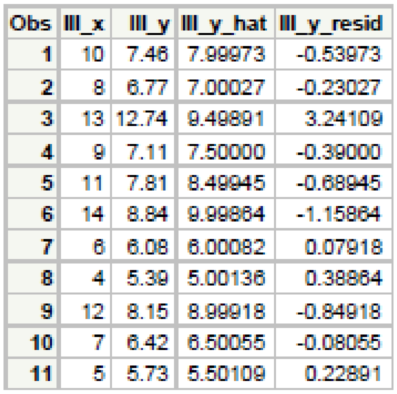
which is close to what Anscombe originally specified to be

$$III\_y = 3 + 0.5 \cdot x \tag{2}$$

[2] Anscombe’s quartet is a collection of four datasets each of which has the same mean, variance, and correlation as the others. However, their distributions are strikingly different. Anscombe created the quartet to emphasize the need for visual analysis of data in addition to computing summary statistics.

The regression results are given in Table 2. Note that the residuals vary widely in magnitude, and that the residual of the outlier at observation 3 is much larger than those of any of the data points since its predicted value, $\widehat{III\_y}$, is too low.

Table 2 Series III Regression Results



Obs	III_x	III_y	III_y_hat	III_y_resid
1	10	7.46	7.99973	-0.53973
2	8	6.77	7.00027	-0.23027
3	13	12.74	9.49891	3.24109
4	9	7.11	7.50000	-0.39000
5	11	7.81	8.49945	-0.68945
6	14	8.84	9.99864	-1.15864
7	6	6.08	6.00082	0.07918
8	4	5.39	5.00136	0.38864
9	12	8.15	8.99918	-0.84918
10	7	6.42	6.50055	-0.08055
11	5	5.73	5.50109	0.22891

If we include a binary indicator variable to signify the presence of the exceptional data point, we have new data shown below (Table 3).

Table 3 Series III Data and Binary Indicator Variable

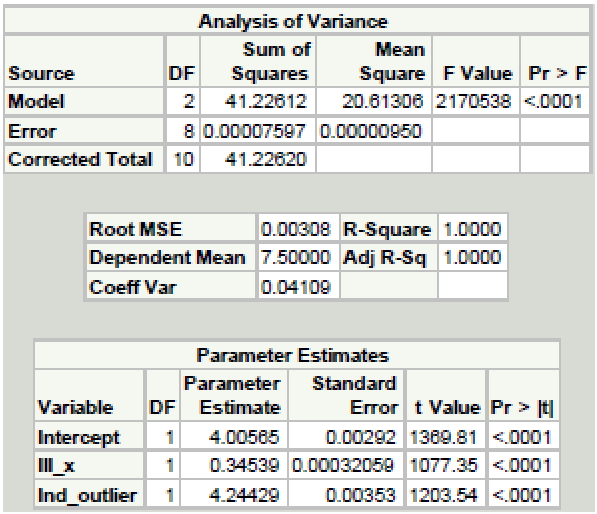
x	10.0	8.0	13.0	9.0	11.0	14.0	6.0	4.0	12.0	7.0	5.0
III_y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
Binary	0	0	**1**	0	0	0	0	0	0	0	0

We use OLS regression to compute the least-squares coefficients of the model

$$III\_y = b_0 + b_1 x + b_2 BinaryIndicator \qquad (3)$$

The results from PROC REG are shown in Table 4.

Table 4 Regression of III_y on x Using Binary Indicator



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	41.22612	20.61306	2170538	<.0001
Error	8	0.00007597	0.00000950		
Corrected Total	10	41.22620			

Root MSE	0.00308	R-Square	1.0000
Dependent Mean	7.50000	Adj R-Sq	1.0000
Coeff Var	0.04109		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.00565	0.00292	1369.81	<.0001
III_x	1	0.34539	0.00032059	1077.35	<.0001
Ind_outlier	1	4.24429	0.00353	1203.54	<.0001

The coefficient of determination, R^2, is 1, indicating exact correlation between the dependent and independent variable (to four significant digits of accuracy). The equation of the trend line for the outlier model is

$$\widehat{III\_y} = 4.00565 + 0.34539 \cdot x + 4.24429 \cdot BinaryIndicator \qquad (4)$$

Figure 2 shows how the binary indicator variable improves the prediction of III_y.

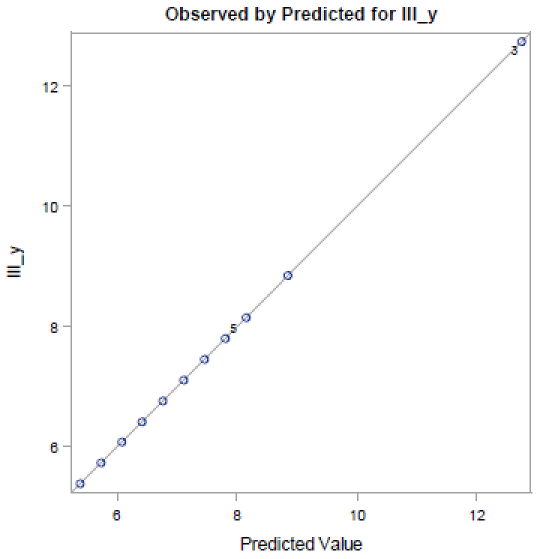
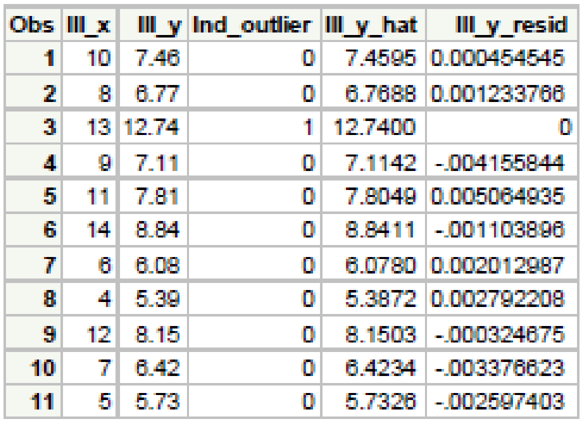


Figure 2 Scatterplot of III_y vs $\widehat{III\_y}$ for Outlier Model

We see that the use of the binary indicator variable to distinguish the outlier for observation 3 removes the bias caused by the outlier and results in a near-perfect fit. The regression results are given in Table 5. Note that the residuals are quite small.

Table 5 Series III Regression Results for Binary Indicator Model



Obs	III_x	III_y	Ind_outlier	III_y_hat	III_y_resid
1	10	7.46	0	7.4595	0.000454545
2	8	6.77	0	6.7688	0.001233766
3	13	12.74	1	12.7400	0
4	9	7.11	0	7.1142	-.004155844
5	11	7.81	0	7.8049	0.005064935
6	14	8.84	0	8.8411	-.001103896
7	6	6.08	0	6.0780	0.002012987
8	4	5.39	0	5.3872	0.002792208
9	12	8.15	0	8.1503	-.000324675
10	7	6.42	0	6.4234	-.003376623
11	5	5.73	0	5.7326	-.002597403

The effect of adding the binary indicator variable is illustrated by the following computations as shown in Table 6.

Table 6 Effect of Adding Binary Indicator Variable

Obs #3	$Regression\ Equation$	Pred III_y	Resid III_y
No Indicator	$Eqn\ 1\colon\ \widehat{III\_y} = 3.00245 + 0.49973 \cdot 13$	9.4989	3.24109
With Indicator	$Eqn\ 4\colon \widehat{III\_y} = 4.00565 + 0.34539 \cdot 13 + 4.24429 \cdot 1$	12.7400	0

Since outliers produce large residuals, several measures have been developed to detect outlying observations based on their associated residuals. PROC REG produces many measures of outlyingness. Among them are studentized deleted residuals [5] and Cook's distance (Cook's D) statistic [6].

- Studentized deleted residuals are residuals that have been computed from the original III_y value and the predicted III_y value where the *i*th observation has been deleted from the regression so as to eliminate its influence as an outlier. The residual of the *i*th observation is then normalized by the adjusted mean square error of the regression excluding the deleted observation. Each studentized deleted residual has a Student's t distribution with $n - p - 1$ degrees of freedom.
- Cook's D statistic is "[A]n overall measure of the combined impact of the *i*th case on all of the estimated regression coefficients" [7, p. 403]. Similar to the studentized deleted residual, Cook's D is calculated by removing the *i*th observation and recalculating the regression estimates. The computed value of Cook's D may be referred to an F distribution with p and $n - p$ degrees of freedom at the median, e.g., $F_{.50}(p, n - p)$. If D_i > 1 then observation *i* may be considered to be influential. Another rule-of-thumb is to consider any observation for which $Cook's\ D\ \geq 4/n$, where n is the number of data points, to be an outlier.

Figure 3 contains a PROC REG-produced side-by-side chart of studentized deleted residuals and Cook's D for Anscombe's series III data without a binary indicator variable to distinguish outliers. The legend at the bottom of the chart indicates that the *t* statistic for observation 3 is significantly outlying, and the value of Cook's D statistic is similarly greater than the cutoff value of 1.

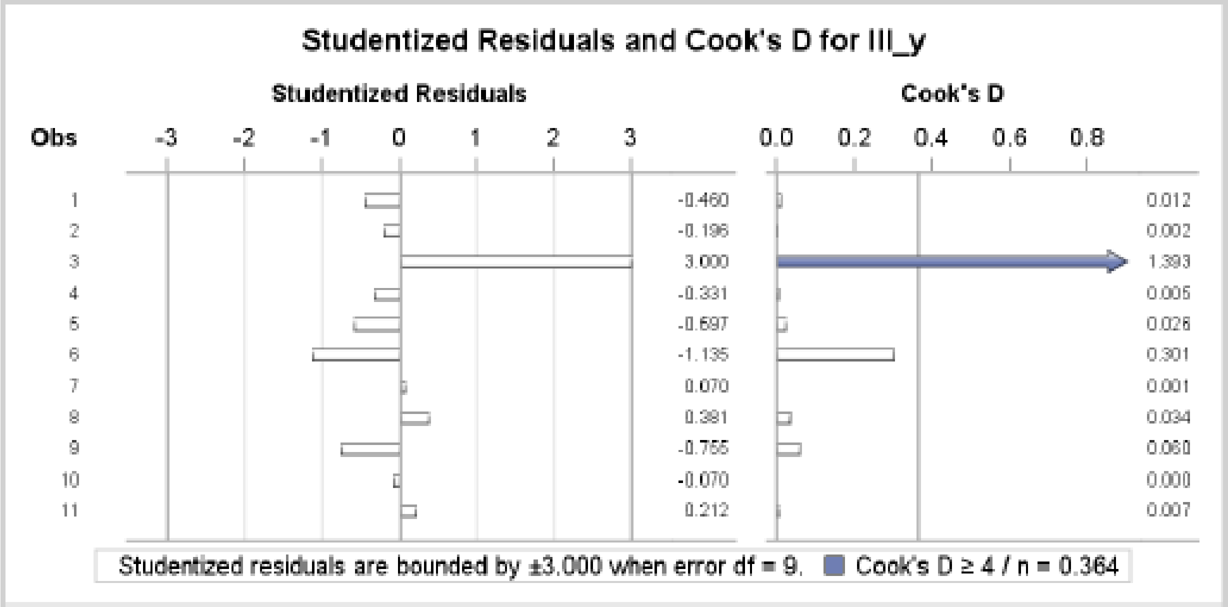


Figure 3 Series III_y Studentized Deleted Residuals and Cook's D

Figure 4 indicates the moderating effect of distinguishing outliers by using a binary indicator variable. None of the observations produces an extremely large residual that biases the regression estimates.

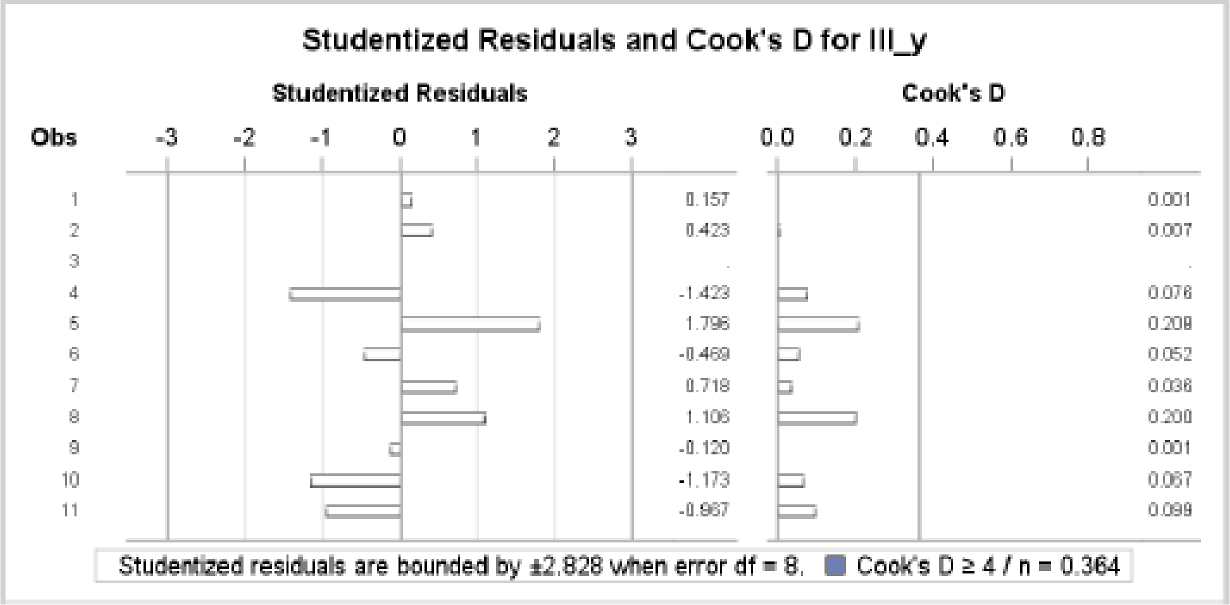


Figure 4 Series III_y Studentized Deleted Residuals and Cook's D for Binary Indicator Model

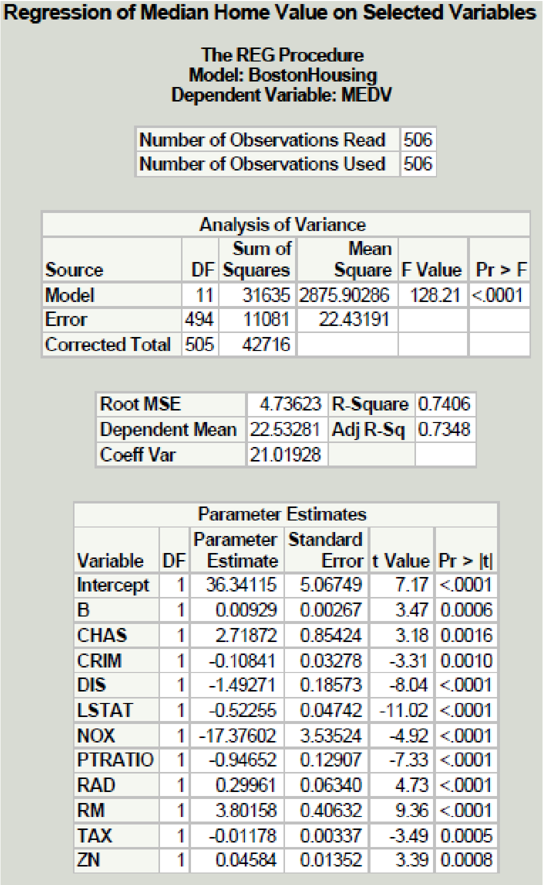
Detecting Outliers Using Multivariate Regression

Treatment of outliers in the univariate case may be straightforwardly applied to the more complex case of multivariate regression. We will use the Boston Housing dataset [8] to investigate treatment of outliers in the multivariate case.

The Boston housing dataset represents housing values in suburbs of Boston. It consists of 506 observations that were collected in 1978. There are 13 interval-scaled attributes and one binary variable. The dependent variable in this exercise is median home value in $1,000's. Appendix A contains attribute information.

We used PROC REG to perform variable selection using the adjusted R^2 model selection option[3]. The predictors AGE and INDUS were omitted from the final model. Table 8 shows the results.

Table 7 Regression of Median Home Value on Selected Variables



Regression of Median Home Value on Selected Variables

The REG Procedure
Model: BostonHousing
Dependent Variable: MEDV

Number of Observations Read	506
Number of Observations Used	506

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	31635	2875.90286	128.21	<.0001
Error	494	11081	22.43191		
Corrected Total	505	42716			

Root MSE	4.73623	R-Square	0.7406
Dependent Mean	22.53281	Adj R-Sq	0.7348
Coeff Var	21.01928		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	36.34115	5.06749	7.17	<.0001
B	1	0.00929	0.00267	3.47	0.0006
CHAS	1	2.71872	0.85424	3.18	0.0016
CRIM	1	-0.10841	0.03278	-3.31	0.0010
DIS	1	-1.49271	0.18573	-8.04	<.0001
LSTAT	1	-0.52255	0.04742	-11.02	<.0001
NOX	1	-17.37602	3.53524	-4.92	<.0001
PTRATIO	1	-0.94652	0.12907	-7.33	<.0001
RAD	1	0.29961	0.06340	4.73	<.0001
RM	1	3.80158	0.40632	9.36	<.0001
TAX	1	-0.01178	0.00337	-3.49	0.0005
ZN	1	0.04584	0.01352	3.39	0.0008

[3] The adjusted R^2 method finds subsets of independent variables that best predict a dependent variable by linear regression in the given sample. The method finds the models with the highest adjusted R^2 for a given combination of variables in a subset of the predictors. See, e.g., https://documentation.sas.com describing PROC REG model selection methods.

Figure 5 is a scatterplot of observed versus predicted median home values. It shows the outlying observations that are not within the 95% prediction ellipse[4]. The observations at the upper right corner of the plot all have the same median home value, so it is possible that they were capped at $50,000.

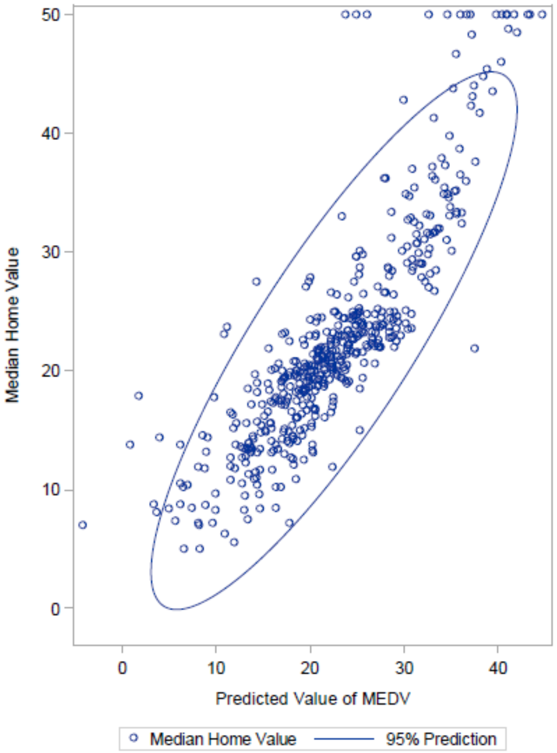
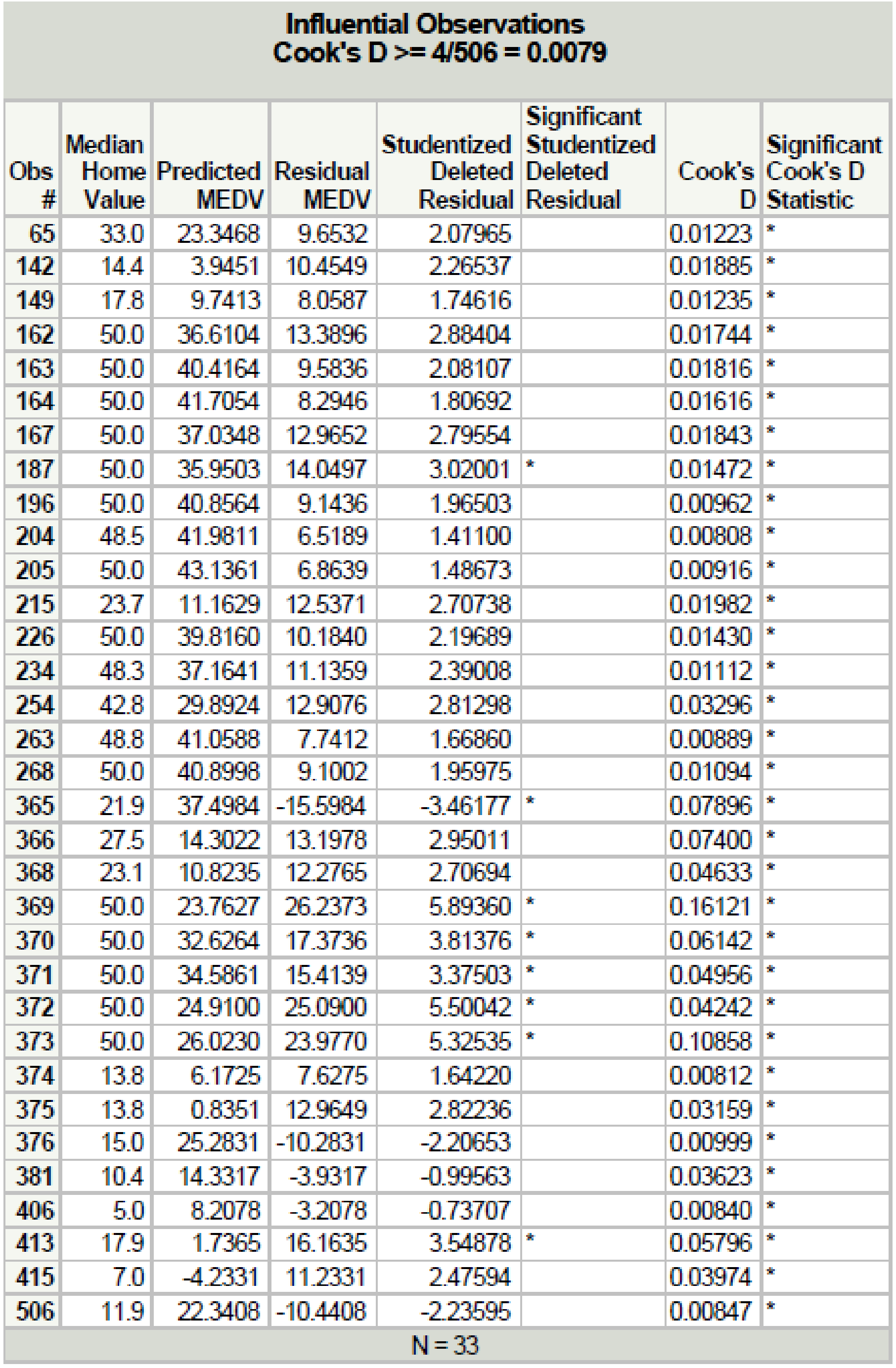


Figure 5 Observed vs Predicted Median Home Values

[4] A prediction ellipse is a graphical representation of the $100(1-\alpha)\%$ confidence region of prediction for the location of a new observation assuming bivariate normality of observed and predicted points. For a Type 1 error of 5%, the 95% prediction ellipse represents the bivariate region of the plane in which 95% of predicted points would be contained. Thus, points outside of the prediction ellipse boundary may be potential outliers.

Table 8 represents 33 observations that equaled or exceeded the Cook’s D value of 0.0079, which indicates the threshold of an influential observation for this dataset. Note that there are studentized deleted residual t-values that do not exceed the critical value of $t \geq 3$, so we conclude that Cook’s D statistic may produce more “false alarms”[5] than the studentized deleted residual t-statistic.

Table 8 Multivariate Regression Influential Observations



Influential Observations
Cook's D >= 4/506 = 0.0079

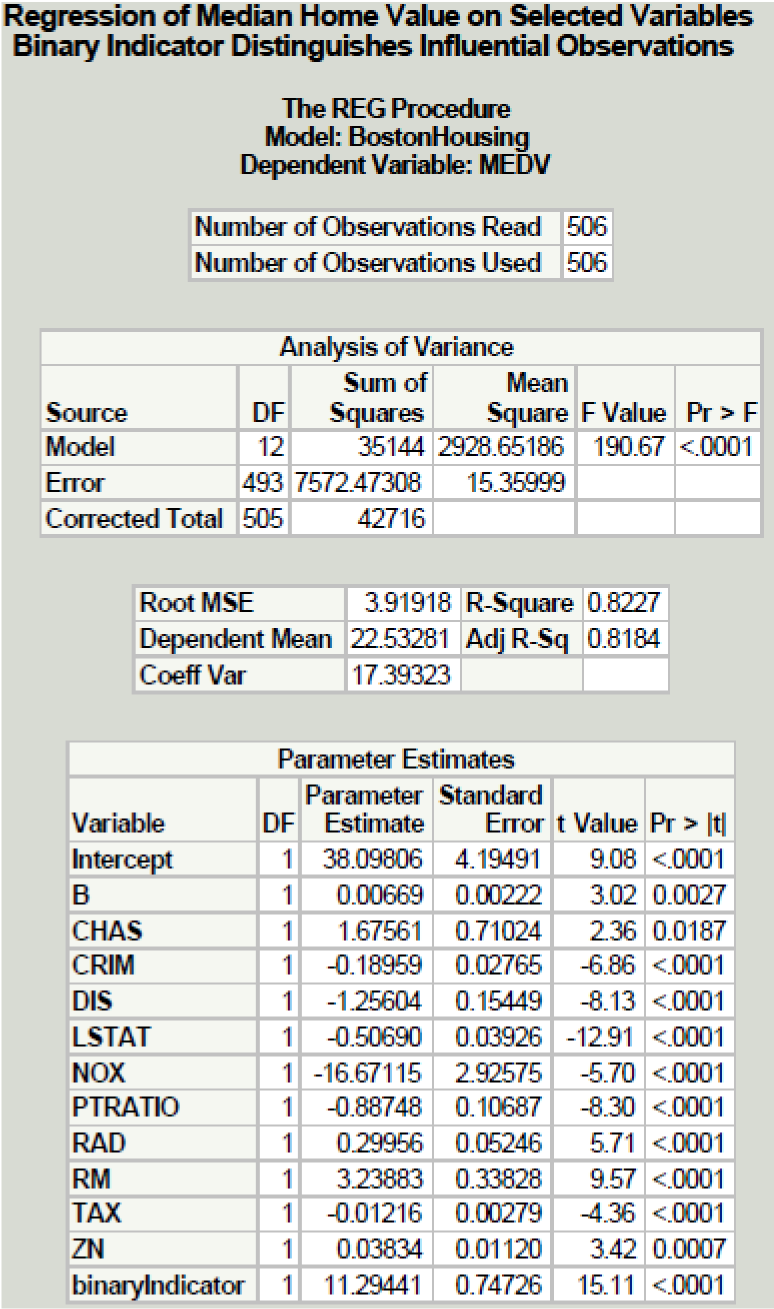
					Significant		
	Median			Studentized	Studentized		Significant
Obs	Home	Predicted	Residual	Deleted	Deleted	Cook's	Cook's D
#	Value	MEDV	MEDV	Residual	Residual	D	Statistic
65	33.0	23.3468	9.6532	2.07965		0.01223	*
142	14.4	3.9451	10.4549	2.26537		0.01885	*
149	17.8	9.7413	8.0587	1.74616		0.01235	*
162	50.0	36.6104	13.3896	2.88404		0.01744	*
163	50.0	40.4164	9.5836	2.08107		0.01816	*
164	50.0	41.7054	8.2946	1.80692		0.01616	*
167	50.0	37.0348	12.9652	2.79554		0.01843	*
187	50.0	35.9503	14.0497	3.02001	*	0.01472	*
196	50.0	40.8564	9.1436	1.96503		0.00962	*
204	48.5	41.9811	6.5189	1.41100		0.00808	*
205	50.0	43.1361	6.8639	1.48673		0.00916	*
215	23.7	11.1629	12.5371	2.70738		0.01982	*
226	50.0	39.8160	10.1840	2.19689		0.01430	*
234	48.3	37.1641	11.1359	2.39008		0.01112	*
254	42.8	29.8924	12.9076	2.81298		0.03296	*
263	48.8	41.0588	7.7412	1.66860		0.00889	*
268	50.0	40.8998	9.1002	1.95975		0.01094	*
365	21.9	37.4984	-15.5984	-3.46177	*	0.07896	*
366	27.5	14.3022	13.1978	2.95011		0.07400	*
368	23.1	10.8235	12.2765	2.70694		0.04633	*
369	50.0	23.7627	26.2373	5.89360	*	0.16121	*
370	50.0	32.6264	17.3736	3.81376	*	0.06142	*
371	50.0	34.5861	15.4139	3.37503	*	0.04956	*
372	50.0	24.9100	25.0900	5.50042	*	0.04242	*
373	50.0	26.0230	23.9770	5.32535	*	0.10858	*
374	13.8	6.1725	7.6275	1.64220		0.00812	*
375	13.8	0.8351	12.9649	2.82236		0.03159	*
376	15.0	25.2831	-10.2831	-2.20653		0.00999	*
381	10.4	14.3317	-3.9317	-0.99563		0.03623	*
406	5.0	8.2078	-3.2078	-0.73707		0.00840	*
413	17.9	1.7365	16.1635	3.54878	*	0.05796	*
415	7.0	-4.2331	11.2331	2.47594		0.03974	*
506	11.9	22.3408	-10.4408	-2.23595		0.00847	*

N = 33

[5] The concept of a false alarm is relative in this discussion since there is no definitive knowledge that an observation is sufficiently distinct from the other points in the sample to be considered an outlier. Hence, we observe that Cook’s D statistic may be a more sensitive indicator than the studentized deleted residual t-statistic.

If we add to the model a binary indicator variable for which $BinaryIndicator = 1$ when $Cook's\ D \geq 0.0079$ and 0 otherwise, we get the following improved result as seen in Table 9:

Table 9 Multivariate Regression Model Using Binary Indicator



Regression of Median Home Value on Selected Variables
Binary Indicator Distinguishes Influential Observations

The REG Procedure
Model: BostonHousing
Dependent Variable: MEDV

Number of Observations Read	506
Number of Observations Used	506

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	35144	2928.65186	190.67	<.0001
Error	493	7572.47308	15.35999		
Corrected Total	505	42716			

Root MSE	3.91918	R-Square	0.8227
Dependent Mean	22.53281	Adj R-Sq	0.8184
Coeff Var	17.39323		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	38.09806	4.19491	9.08	<.0001
B	1	0.00669	0.00222	3.02	0.0027
CHAS	1	1.67561	0.71024	2.36	0.0187
CRIM	1	-0.18959	0.02765	-6.86	<.0001
DIS	1	-1.25604	0.15449	-8.13	<.0001
LSTAT	1	-0.50690	0.03926	-12.91	<.0001
NOX	1	-16.67115	2.92575	-5.70	<.0001
PTRATIO	1	-0.88748	0.10687	-8.30	<.0001
RAD	1	0.29956	0.05246	5.71	<.0001
RM	1	3.23883	0.33828	9.57	<.0001
TAX	1	-0.01216	0.00279	-4.36	<.0001
ZN	1	0.03834	0.01120	3.42	0.0007
binaryIndicator	1	11.29441	0.74726	15.11	<.0001

We compare the prediction performance of the two models below. Figure 5, reproduced from above, shows the 95% prediction ellipse of the regression model *sans* binary indicator variable. Figure 6 is a scatterplot of observed versus predicted median home values for the regression model that includes a binary indicator to distinguish outlying observations.

We observe that the observations are grouped more closely within the 95% prediction ellipse for the binary indicator model, and that there appear to be fewer outliers in the neighborhood of the centroid of median home value. We conclude that the accuracy of the model has been improved by the inclusion of the binary indicator variable.

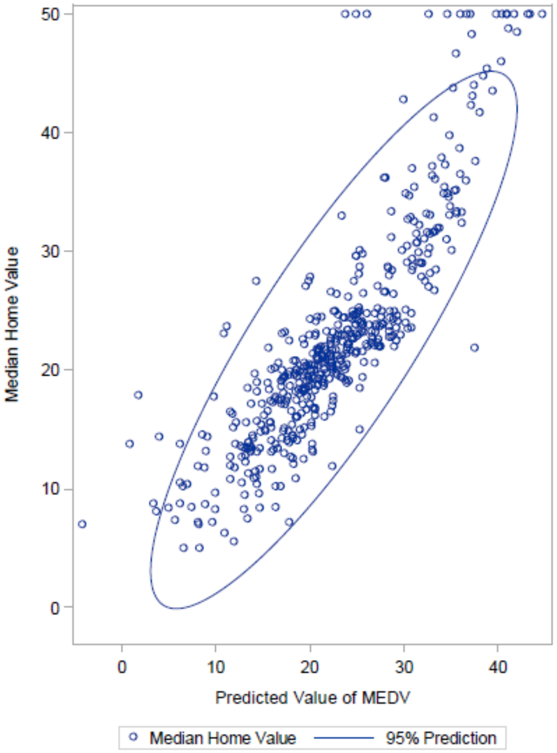


Figure 5 Observed vs Predicted Med Home Values

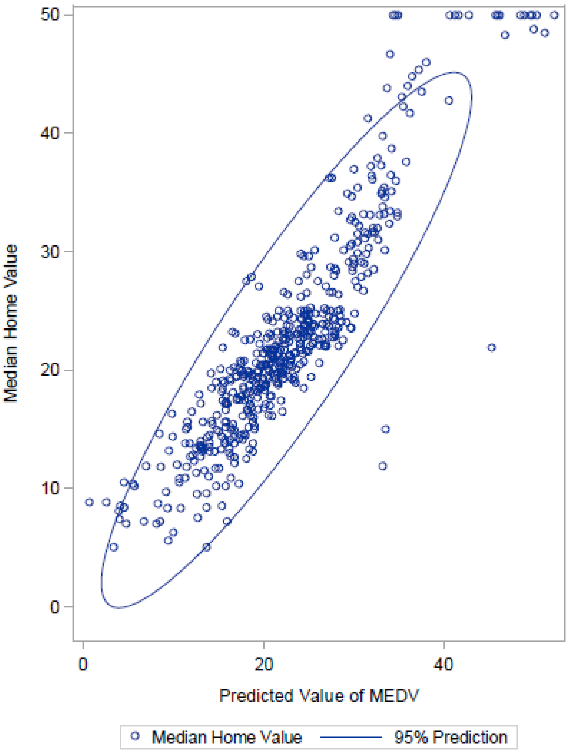
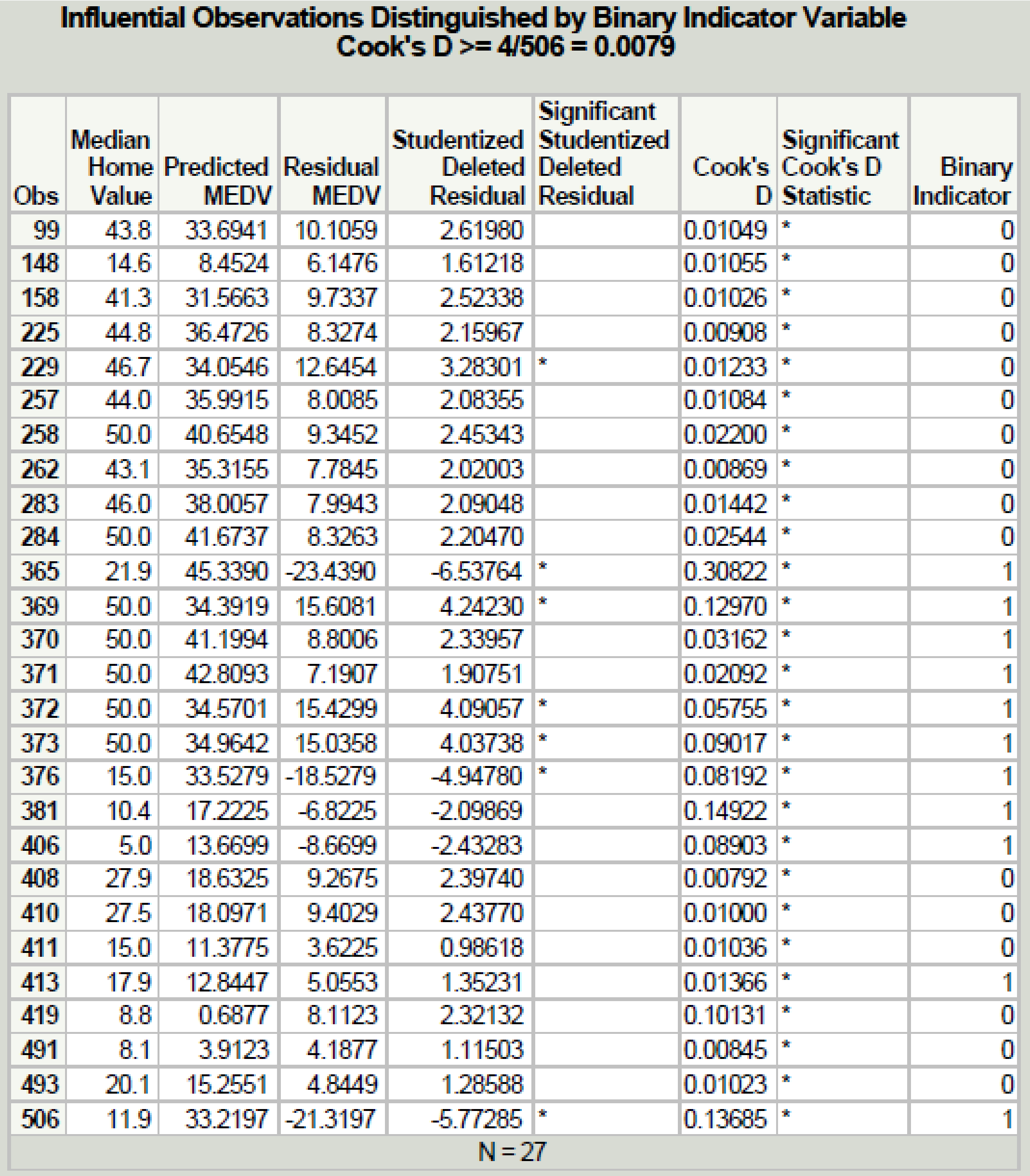


Figure 6 Observed vs Predicted Median Home Values, Binary Indicator Model

Table 10 contains the 27 observations detected by Cook's D statistic. Every observation in Table 8 was assigned $BinaryIndicator = 1$. In Table 10, only 11 of the original 33 distinguished observations surpass the Cook's D threshold of 0.0079.[6] The inclusion of the binary indicator variable has resulted in reducing the total number of suspect outliers from 33 to 27.

Table 10 Influential Observations Distinguished by Binary Indicator Variable



Influential Observations Distinguished by Binary Indicator Variable
Cook's D >= 4/506 = 0.0079

Obs	Median Home Value	Predicted MEDV	Residual MEDV	Studentized Deleted Residual	Significant Studentized Deleted Residual	Cook's D	Significant Cook's D Statistic	Binary Indicator
99	43.8	33.6941	10.1059	2.61980		0.01049	*	0
148	14.6	8.4524	6.1476	1.61218		0.01055	*	0
158	41.3	31.5663	9.7337	2.52338		0.01026	*	0
225	44.8	36.4726	8.3274	2.15967		0.00908	*	0
229	46.7	34.0546	12.6454	3.28301	*	0.01233	*	0
257	44.0	35.9915	8.0085	2.08355		0.01084	*	0
258	50.0	40.6548	9.3452	2.45343		0.02200	*	0
262	43.1	35.3155	7.7845	2.02003		0.00869	*	0
283	46.0	38.0057	7.9943	2.09048		0.01442	*	0
284	50.0	41.6737	8.3263	2.20470		0.02544	*	0
365	21.9	45.3390	-23.4390	-6.53764	*	0.30822	*	1
369	50.0	34.3919	15.6081	4.24230	*	0.12970	*	1
370	50.0	41.1994	8.8006	2.33957		0.03162	*	1
371	50.0	42.8093	7.1907	1.90751		0.02092	*	1
372	50.0	34.5701	15.4299	4.09057	*	0.05755	*	1
373	50.0	34.9642	15.0358	4.03738	*	0.09017	*	1
376	15.0	33.5279	-18.5279	-4.94780	*	0.08192	*	1
381	10.4	17.2225	-6.8225	-2.09869		0.14922	*	1
406	5.0	13.6699	-8.6699	-2.43283		0.08903	*	1
408	27.9	18.6325	9.2675	2.39740		0.00792	*	0
410	27.5	18.0971	9.4029	2.43770		0.01000	*	0
411	15.0	11.3775	3.6225	0.98618		0.01036	*	0
413	17.9	12.8447	5.0553	1.35231		0.01366	*	1
419	8.8	0.6877	8.1123	2.32132		0.10131	*	0
491	8.1	3.9123	4.1877	1.11503		0.00845	*	0
493	20.1	15.2551	4.8449	1.28588		0.01023	*	0
506	11.9	33.2197	-21.3197	-5.77285	*	0.13685	*	1

N = 27

[6] Observations 365, 369, 370, 371, 372, 373, 376, 381, 406, 413, and 506 were identified as outliers in the non-binary indicator regression, as reported in Table 8.

We can gauge the effect of adding the binary indicator variable by comparing regression performance statistics in the two models, as shown in Table 11.

Table 11 Comparison of Models

Performance Metric	No Binary Indicator	Includes Binary Indicator
Root MSE	4.73623	3.91918
Coefficient of Variation	21.01928	17.39323
R^2	0.7406	0.8227
Adjusted R^2	0.7348	0.8184

We see that adding the binary indicator to distinguish observations detected by Cook's D statistic has the effect of improving the model's accuracy by 17%[7], with a corresponding increase in adjusted R^2 of 8.36%.

Detecting Outliers Using Mahalanobis Distance

One or more outlying values in the univariate case, e.g., Anscombe's Series III data, will be easy to spot because a histogram of the data will show them at a significant distance from the body of the data. In the multivariate case, however, creating histograms for each variable can quickly become tedious to analyze, and may not readily reveal a pattern that can be used for labelling certain observations as outliers since there is no direct way to visualize the interactions between multiple variables.

The Mahalanobis distance is a commonly-used metric that converts an observation consisting of several continuous values into a single scalar value. The histogram of the distribution of these distances can be used to highlight observations whose distance from the centroid of the data is considered to be excessive.

The formula for the Mahalanobis distance is

$$D^2(\boldsymbol{x_i}) = (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^{\boldsymbol{T}} \boldsymbol{S}^{-\boldsymbol{1}} (\boldsymbol{x_i} - \overline{\boldsymbol{x}}) \qquad (5)$$

where $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}), i = 1, \ldots, N$ is a row vector of *p* variables, $\overline{\boldsymbol{x}}$ is the row vector of means of each variable, i.e., the centroid of the data, and $S^{-\boldsymbol{1}}$ is the inverse of the sample covariance matrix.[8] We may specify a boundary point, $\chi^2_{p,0.975}$, that determines an ellipsoid in p dimensions. If $D^2(\boldsymbol{x_i}) \leq \chi^2_{p,0.975}$, then the point $\boldsymbol{x_i}$ lies within the ellipsoid and is not an outlier. Otherwise $\boldsymbol{x_i}$ is an outlier. One difficulty with using the Mahalanobis distance to identify outliers is that "[D]ata sets with multiple outliers or clusters of outliers are subject to *masking* and *swamping* effects" [9, p. 7]. An observation that would be considered to be an outlier by itself is masked if there is another observation close to it which skews the mean and covariance estimates significantly to reduce the distance of the outlying points to the centroid of the data. Swamping occurs when an observation that would normally be a non-outlier is grouped with outlying instances that skew the mean and the covariance estimates toward the centroid of the group and not the centroid of the main body of the data. Distances in this case are large so that the normally non-outlying observation is classified as an outlier.

Since outliers significantly influence the mean and covariance estimates of a dataset, estimates of distance ought to be computed using robust procedures such as PROC ROBUSTREG. Further discussion of outliers using robust algorithms is a substantial topic that is outside the scope of this paper.

[7] We compute (RootMSE_NoIndicator-RootMSE_IncludingIndicator)/ROOTMSE_NoIndicator to get -0.17238, or about 17% decrease in Root MSE due to inclusion of a binary indicator variable for suspected outliers.

[8] The Mahalanobis distance is the multivariate equivalent of the z-score where $z = (x - \bar{x})/s$ in the univariate case.

Using PROC ROBUSTREG, we computed Mahalanobis and robust distances and compared them in Figure 7. We see that there is significant masking of Mahalanobis distances which the robust distance calculation technique uncovers. The Mahalanobis distance-observation number plot and the corresponding box-and-whisker plot show only moderate distances from observations to the centroid of the data, while the equivalent robust results from PROC ROBUSTREG reveal significant deviations.

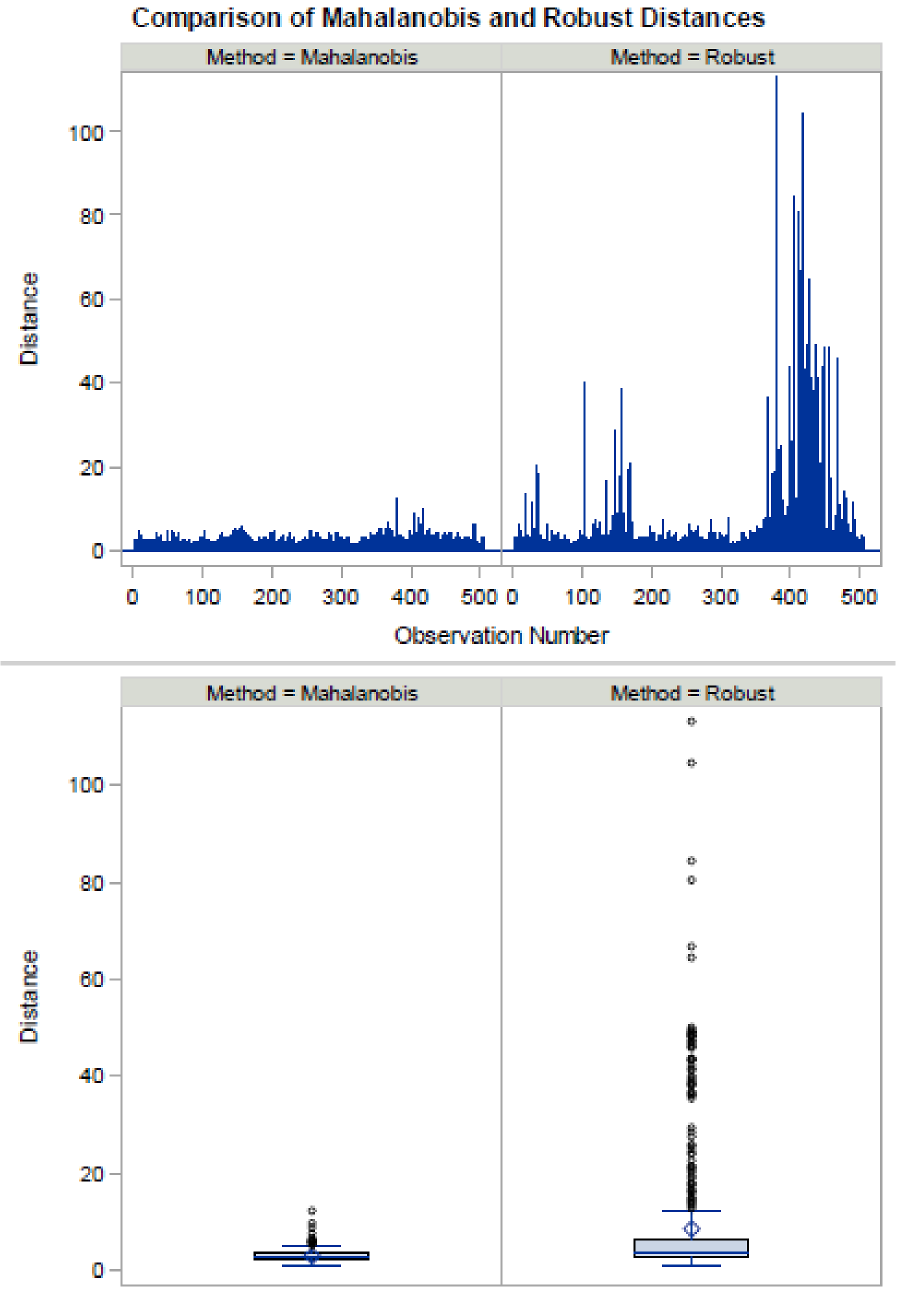


Figure 7 Comparison of Mahalanobis and Robust Distances

In Figure 8, we compare the predictions generated by the OLS binary indicator model using Cook's D statistic (Figure 6) and the OLS regression using the binary outlier indicator created by PROC ROBUSTREG as a predictor of outlyingness.[9] We observe that the two results are closely matched.

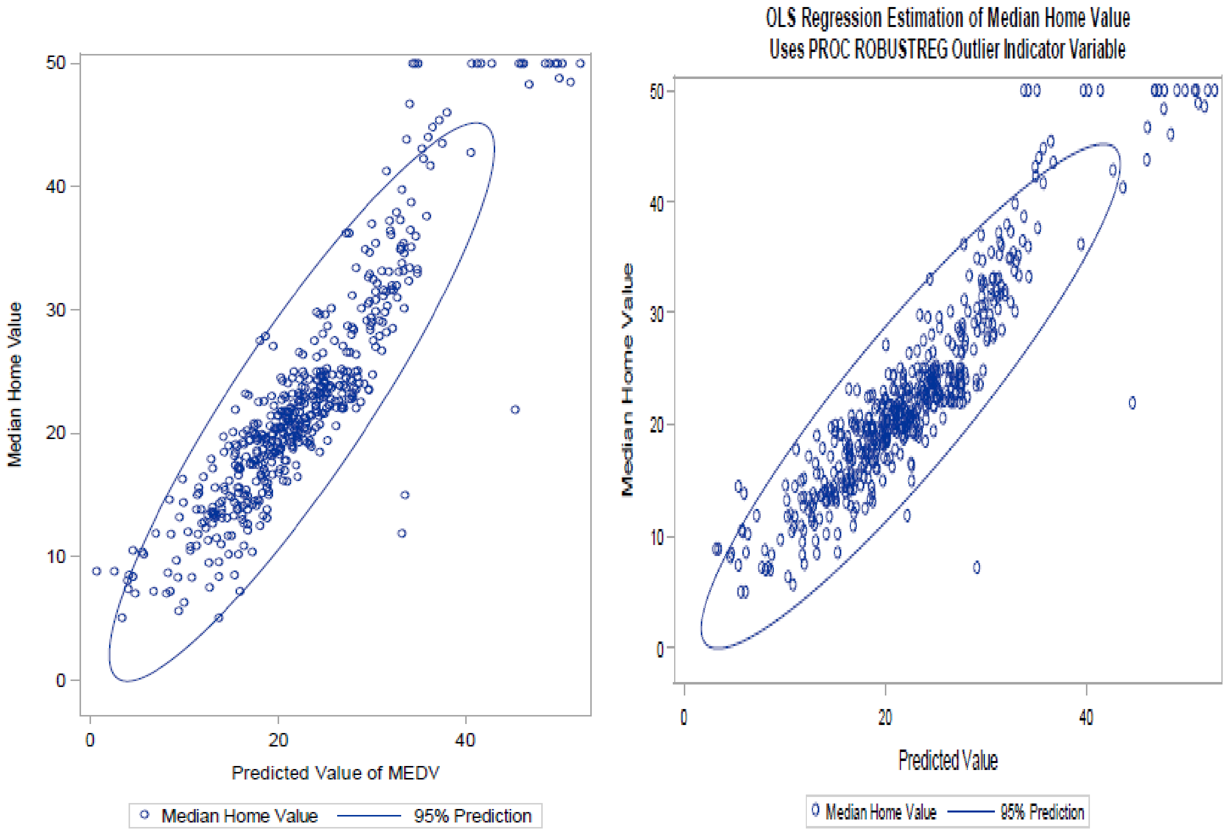


Figure 8 Comparison of Robust and OLS Regression Predictions

Table 12 contains selected performance statistics of the two regression models. We see that the robust regression detection of outliers is slightly more powerful than using Cook's D as a measure of outlyingness.

Table 12 Comparison of OLS Regression and Robust Regression

Performance Metric	OLS Regression Using Cook's D Binary Indicator	Robust Regression Using ROBUSTREG Binary Indicator
Root MSE	3.91918	3.62416
Coefficient of Variation	17.39323	16.08394
R^2	0.8227	0.8481
Adjusted R^2	0.8184	0.8447

[9] PROC ROBUSTREG creates a binary variable, eponymously called "OUTLIER", to distinguish observations that exceed a cutoff value for outlyingness. The feature is documented in the PROC ROBUSTREG MODEL statement and described more fully in the Leverage-point and Outlier Detection portion of the Details section of the ROBUSTREG Procedure documentation.

Detecting Outliers Using Local Outlier Factor and Local Outlier Probability

An outlier detection algorithm may compare an observation to all other observations in a set of data, in which case it is deemed to be global in scope, or it may compare the observation to its neighbors in a cluster, in which case it is local in scope. The detection techniques described above are global in scope, and here we discuss two techniques that are local in scope, pertaining to outliers relative to observations in a particular cluster.

Local Outlier Factor

If a point is relatively isolated from its nearest neighbors in a cluster, it may be an outlier with a degree of outlyingness related to its distance from its nearest local neighbors in the cluster. Using this concept, Breunig et al. [11] described an algorithm for finding locally-outlying observations relative to their nearest neighbors. A point is "*local* in that the degree of outlyingness depends on how isolated the object is with respect to the surrounding neighborhood" [11, p. 93]. A tutorial example is presented in [12] in detail.

The Local Outlier Factor, a positive real number that represents the density[10] of observations around a point compared to the density of its nearest k neighbors, can be used to quantify the isolation of a given point and thus measure its outlyingness. The LOF algorithm requires the specification of an integer, k, which represents the locality, i.e., the number of nearest neighbors to a point. Interpretation of the LOF is not straightforward since it is the average ratio of the density of the k nearest neighbors of a point A to the density of point A and is not necessarily comparable from one cluster to another. For a given point, a $LOF \sim 1$ indicates similar density as its neighbors, $LOF < 1$ indicates higher density than its neighbors, and $LOF > 1$ indicates lower density, i.e., fewer neighbors so higher likelihood of outlyingness.

We have written a SAS® macro, %LOF_LoOP, to compute the LOF and the LoOP (Local Outlier Probability, discussed below). Several graphical representations of the data are also produced to facilitate interpretation and definition of outlyingness based on the LOF and the LoOP. We applied the %LOF_LoOP macro to Fisher's iris data [13] and present selected results.[11]

[10] Density is measured in number of points per unit of distance.

[11] The %LOF_LoOP macro invocation code used to produce the LOF and LoOP graphics is given in Appendix B.

A histogram of the LOF may be used to suggest a cutoff value above which to declare an observation to be an outlier. Values ≤ 1 represent inliers, while values slightly > 1 may be “near outliers” but not conclusively so. In Figure 9, values >> 1, e.g., > 1.5, are likely outliers. Since the LOF is not a uniform metric between clusters, it is difficult to make conclusive statements such as “an observation with a LOF value above 1.5 ought to be treated as an outlier.”

Figure 9 shows the frequency histogram of LOF factors for Fisher’s iris data using $k = 20$ nearest neighbors. The choice of k is important since the LOF may change markedly according to k. Xu et al. [15] have developed a technique for finding an optimal value of k.

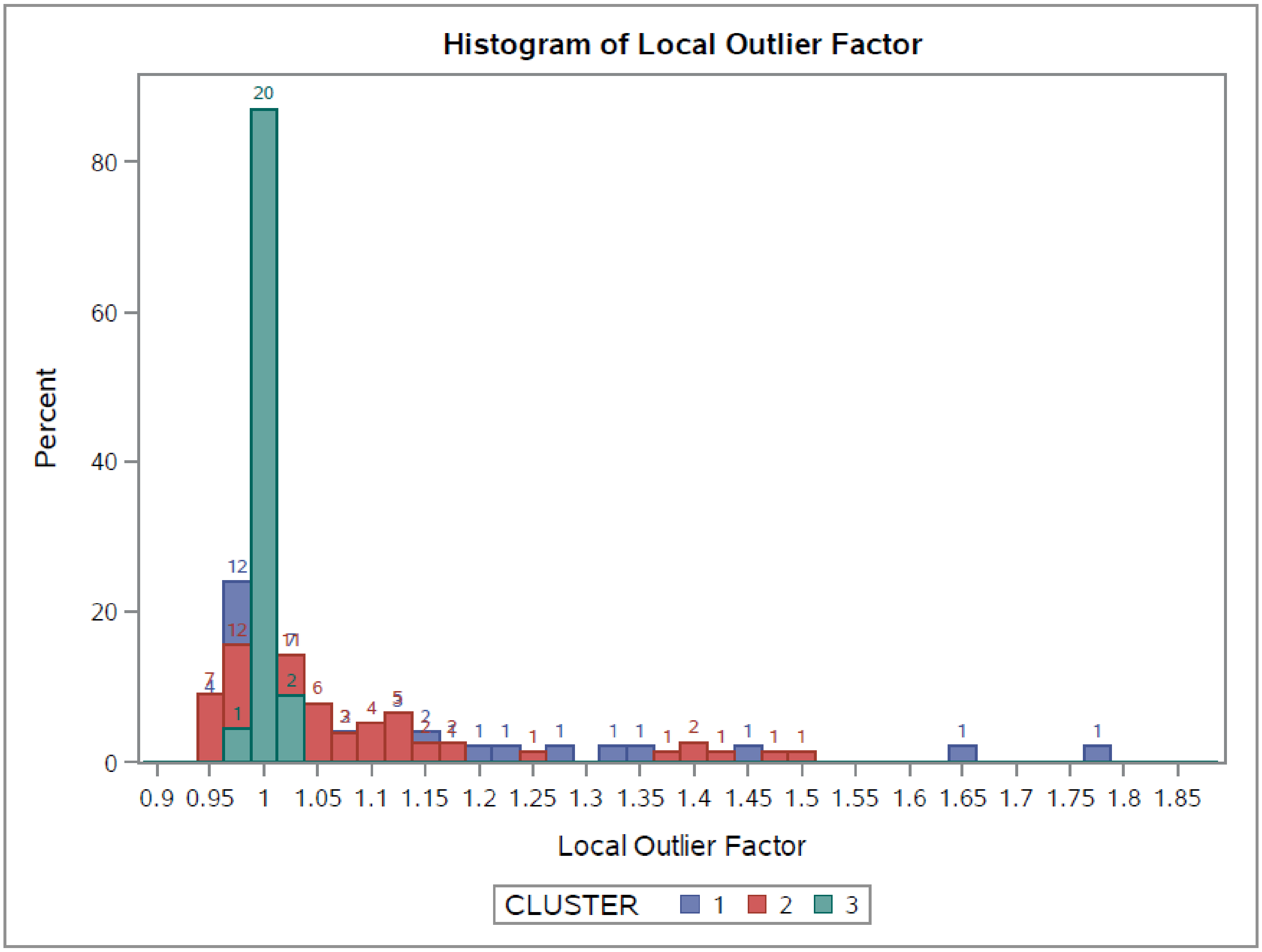


Figure 9 Frequency Histogram of LOF

The empirical cumulative distribution function may be used to detect the LOF value at which the increase in the frequency of distinct values levels off, indicating minor changes in the number of outlying observations. It is another way to depict the distribution of LOF values in addition to the frequency histogram.

Figure 10 represents the distribution of LOF values as a cumulative proportion of the entire set of scores.

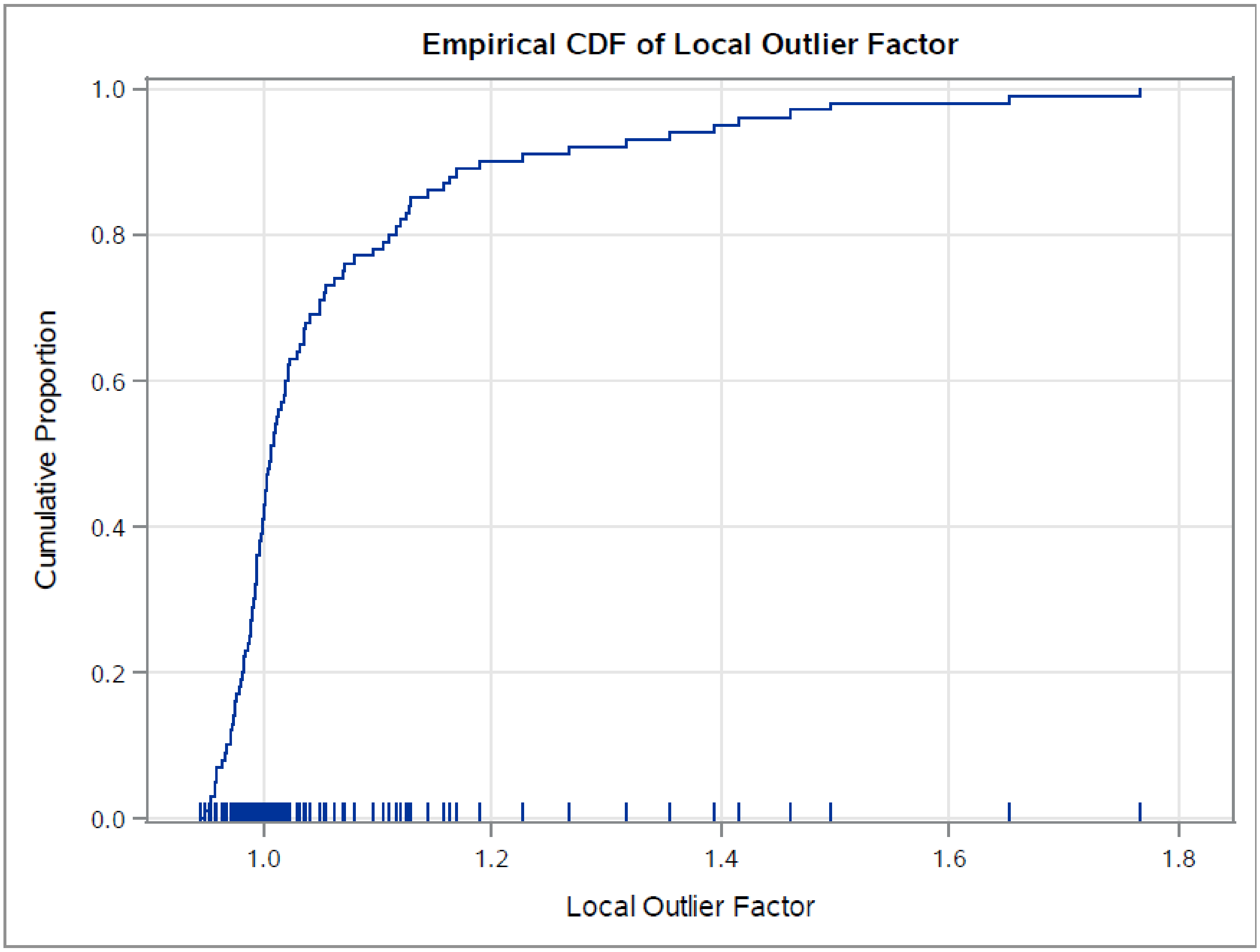


Figure 10 Empirical CDF of Local Outlier Factor

Figure 11 is a scatterplot of the clustered iris data using the $k = 20$ nearest neighbors. In the interests of legibility, we set the minimum LOF value to 1.17 so that circles of LOF > 1.17 would be drawn, thus suppressing circles for smaller values of the LOF.

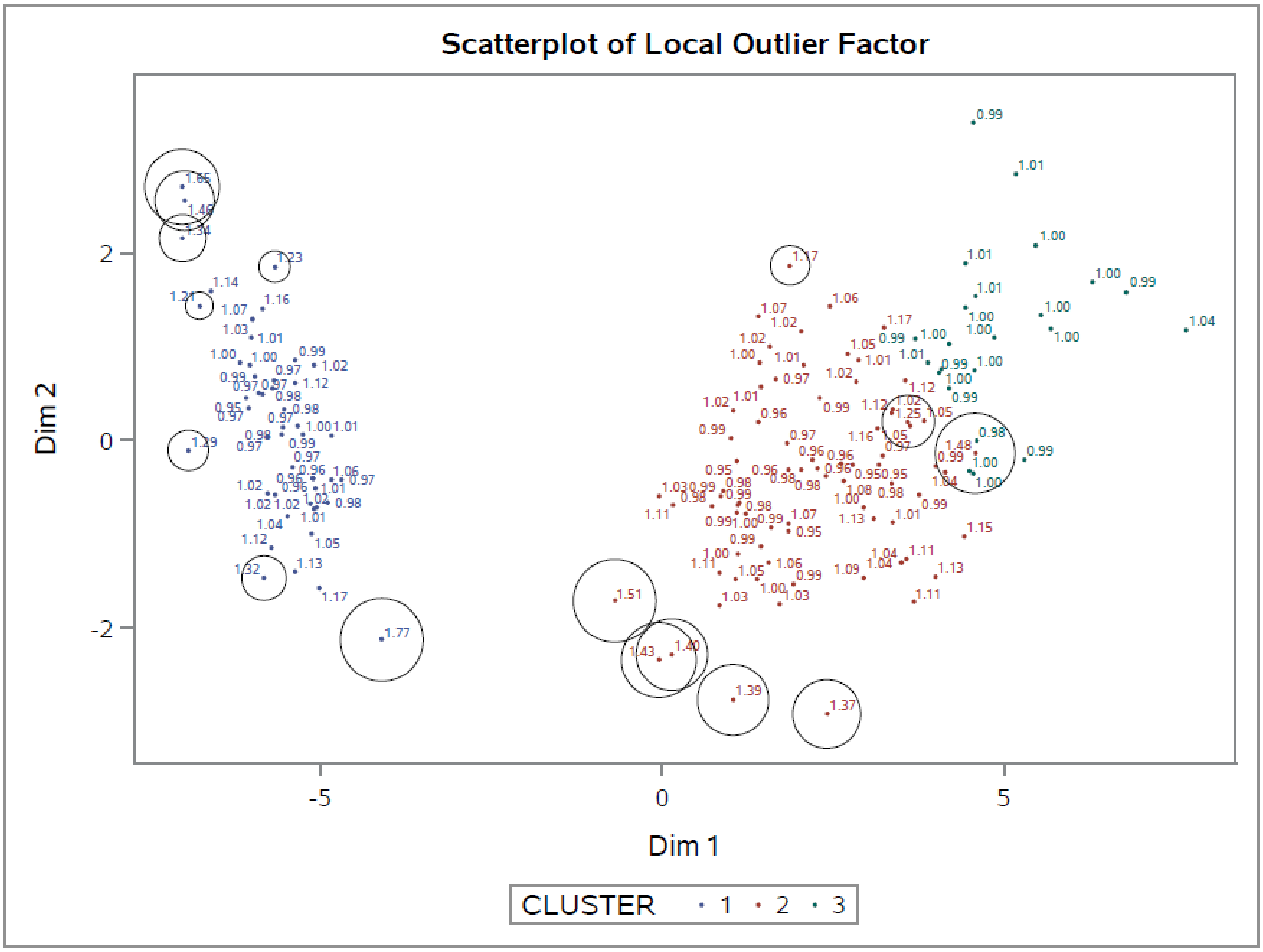


Figure 11 Scatterplot of Local Outlier Factor, Minimum Radius=1.17

Local Outlier Probability

Kriegel et al. [14] have extended the local outlier paradigm to produce a probability of outlyingness (Local Outlier Probability) which is easier to interpret than the LOF itself. The LoOP is a normalized probability that can be compared between clusters and is thus more applicable to all of the points in a set of data.

The distribution of the Local Outlier Probability for the iris data is shown in Figure 12. Similar reasoning as for LOF suggests that LoOP values above 0.6 may indicate that the algorithm has detected an outlying observation.

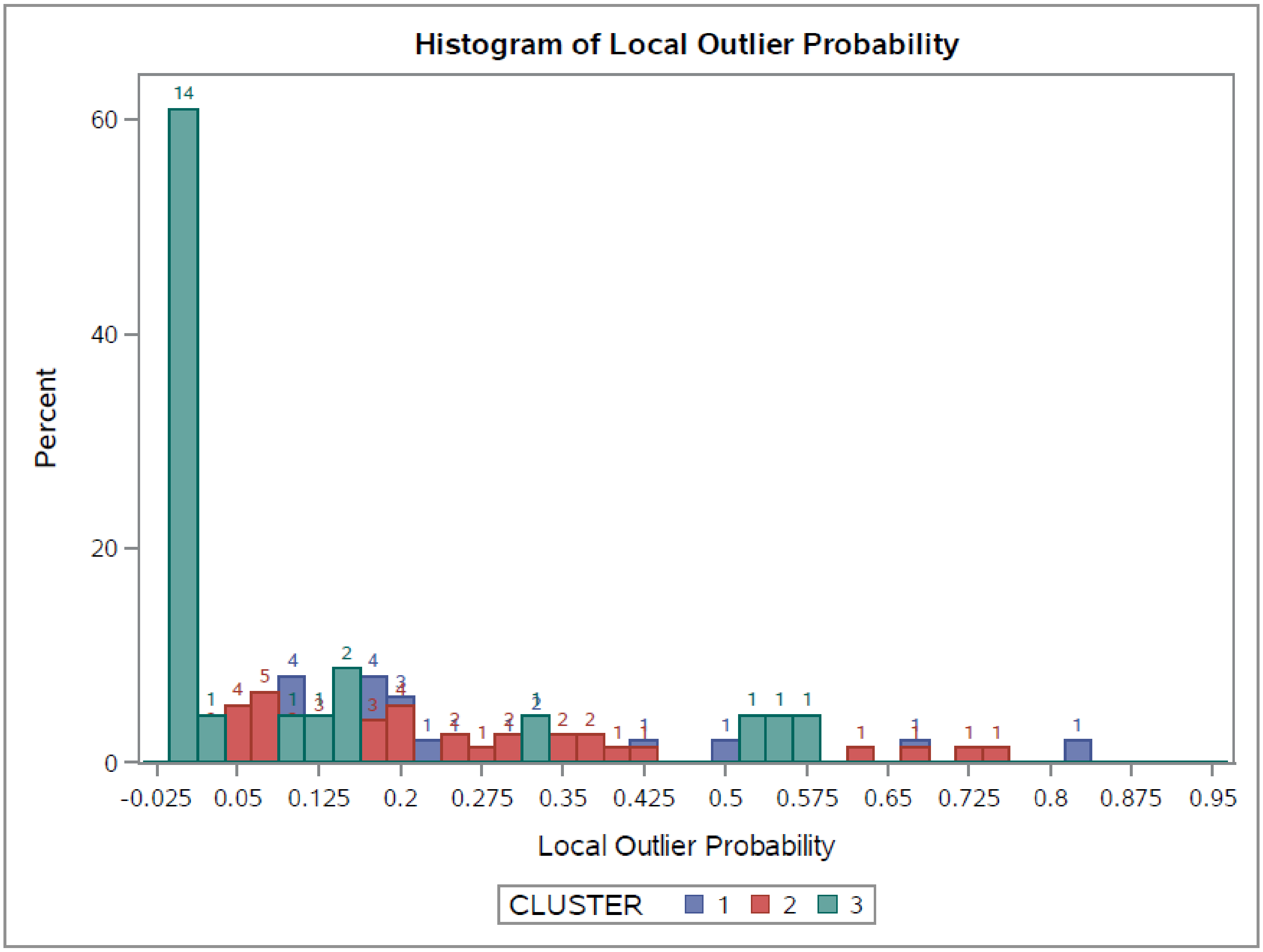


Figure 12 Histogram of Local Outlier Probability

The empirical CDF of the LoOP in Figure 13 serves as a graphical description of the cumulative LoOP values and may be used similarly as the ECDF of the LOF.

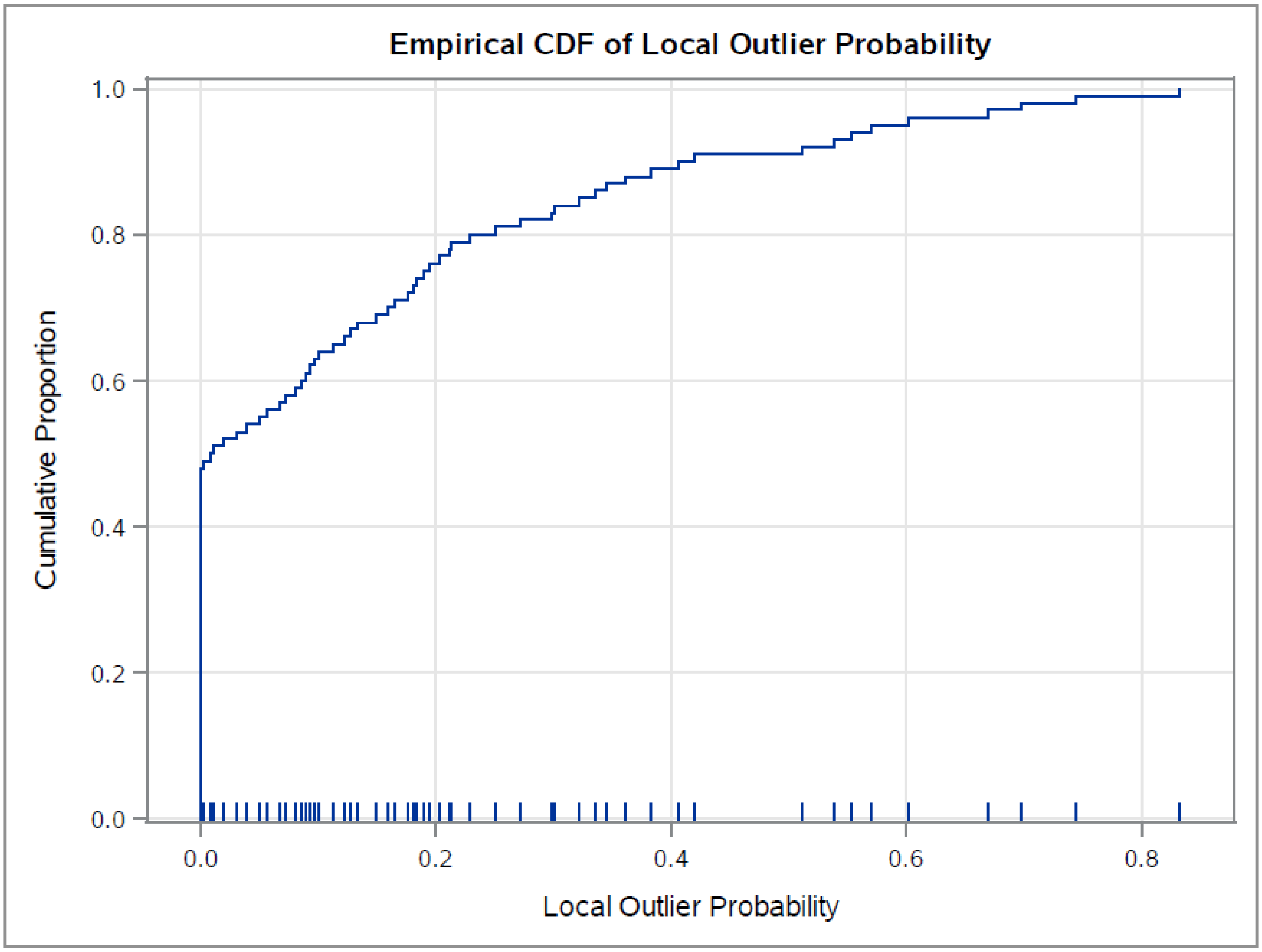


Figure 13 Empirical CDF of Local Outlier Probability

Figure 14 is the scatterplot of the LoOP of Fisher's iris data, with circles drawn around points for which LoOP ≥ 0.2. The %LOF_LoOP macro has the provision to create an outlier flag for observations whose probability of outlyingness exceeds a threshold value. For example, if the macro parameter $OUTLIER\_PROB\ =\ 0.6$ and the LoOP score is > 0.6, the variable OUTLIER would be set to 1 in the SAS output dataset produced by the macro.

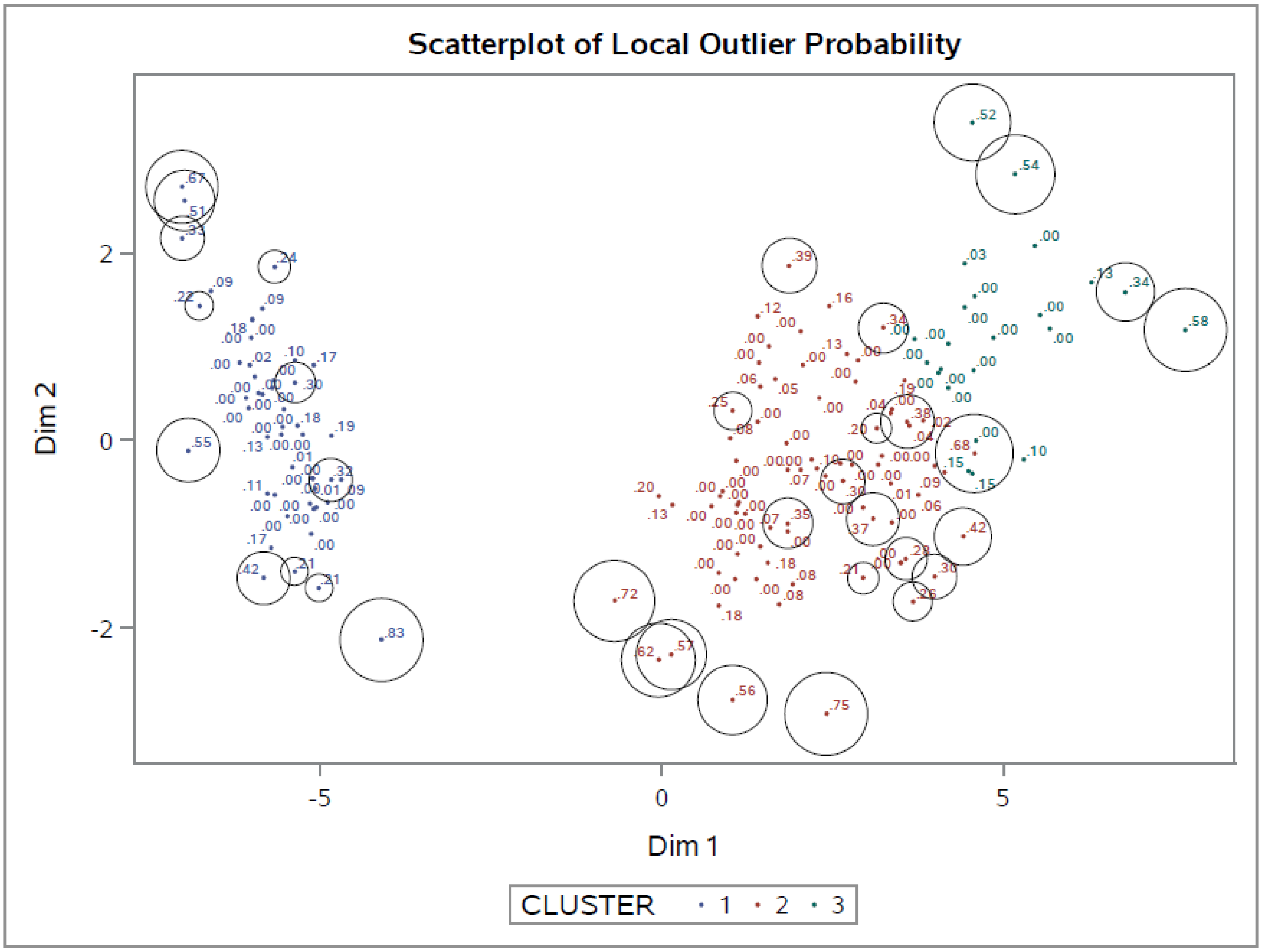


Figure 14 Scatterplot of Local Outlier Probability, Minimum Radius=0.2

Treatment of Outliers

Detecting outliers is the first step in remediating their tendency to influence the predictions of a regression model. They may be identified visually or through analytical means, and a binary indicator variable can be used to mark their presence in a regression analysis.

Treatment of Outliers by Excision

Our first impulse regarding outliers, once we believe that we have confidently detected them, may be to excise them from the sample data. We may tell ourselves that, since these outlying observations are not characteristic of the data, we may remove them from the analysis without consequence, and all will be well. But we may have fooled ourselves with the false assumption that these data points are truly outliers or noise-contaminated data. What if they represent the beginning of an emerging trend?

For example, in ATM fraud detection, a one-time spike in the amount of cash withdrawn from an individual's account may represent money withdrawn for a vacation, or it may signify that a fraudster has hacked the account. If the issuing bank denies the transaction and it is legitimate, the accountholder will be angry. If the transaction is permitted and the accountholder denies making the transaction, the bank may be required to absorb the cost of a false positive decision and incur the expenses associated with cancelling the account and issuing a new debit card. A sequence of unusually large withdrawals in a short amount of time may signify an emerging trend of fraudulent activity, which is very expensive to the bank. Merely to excise such transactions as "contaminated by noise" would be to make a costly and easily-avoided error.

Hopefully, we understand that simply throwing out data points ought not to be a reflex action, but must be considered in light of the consequences of ignoring potentially valuable information. At the very least, the properties of the sample data have changed, and data selection bias has been introduced into what was previously a random sample.

Treatment of Outliers Using Binary Indicator Variables

We have discussed the use of binary indicator variables to distinguish and remediate the effects of outlying values and observe that their use is substantiated in practice as a valid technique. We note that the use of binary indicators does not affect the original distribution of the sample data.

Treatment of Outliers Using Winsorization

Winsorization [10] is the process of replacing a specified set of extreme values of a given variable in a set of sample data with specified values computed from the data. Small extreme values are replaced by larger ones and large extreme values are replaced by smaller ones. While this substitution may reduce the effects of outlying values, it changes the distribution characteristics of the variable that is Winsorized.

For example, in the Boston Housing data, the minimum and maximum values of CRIM are 0.00632 and 88.97620, and the 1st and 99th percentiles are 0.01360 and 41.52920. So a 98% Winsorization of CRIM would substitute the values 0.01360 in place of any value lower than 0.01360 and 41.52920 in place of any value higher than 41.52920. We see from the box-and-whisker plots in Figure 15, which contains the z-score standardization of the sample data, that the range of the CRIM variable is attenuated in the Winsorized sample.

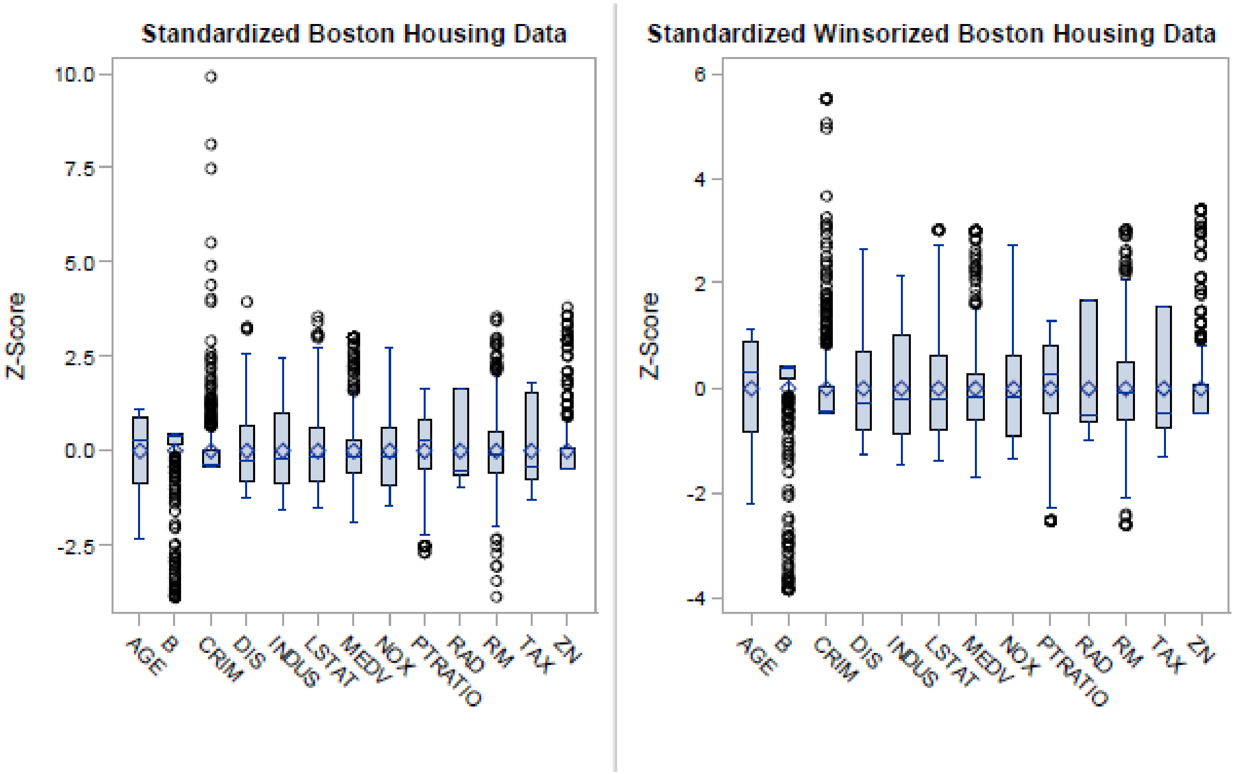
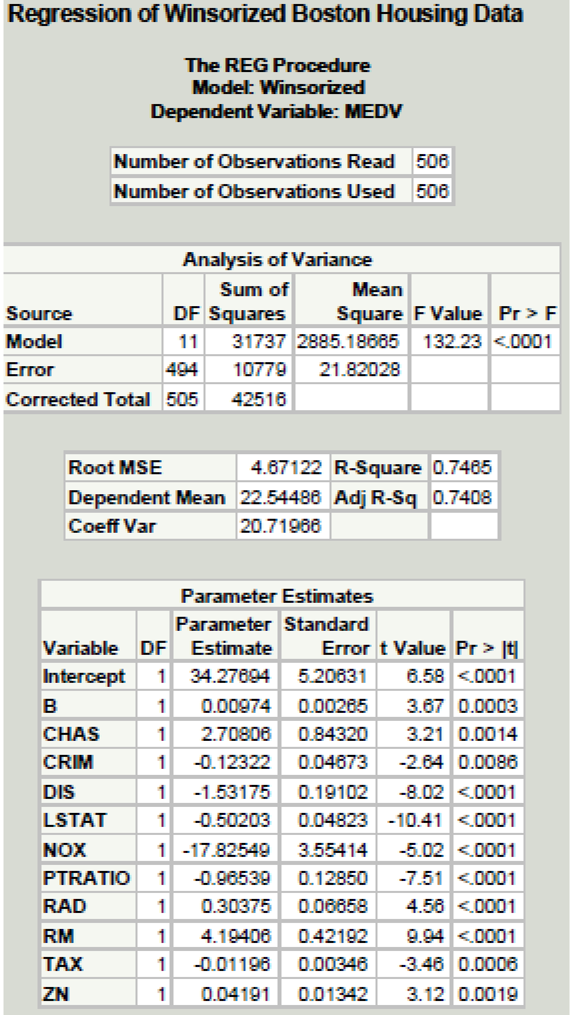


Figure 15 Comparison of Standardized Boston Housing Data and Standardized Winsorized Boston Housing Data

Since we have seen the distorting effects of outliers in masking and swamping outlying and well-behaved observations, we may naturally expect to see an improvement in any model built with Winsorized data.

Table 13 presents the results of predicting median home value using 98% Winsorized Boston housing data. The variables AGE and INDUS were omitted by the adjusted R^2 selection algorithm.

Table 13 Regression Model Using 98% Winsorized Data



Regression of Winsorized Boston Housing Data

The REG Procedure
Model: Winsorized
Dependent Variable: MEDV

Number of Observations Read	506
Number of Observations Used	506

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	31737	2885.18665	132.23	<.0001
Error	494	10779	21.82028		
Corrected Total	505	42516			

Root MSE	4.67122	R-Square	0.7465
Dependent Mean	22.54486	Adj R-Sq	0.7408
Coeff Var	20.71966		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	34.27694	5.20631	6.58	<.0001
B	1	0.00974	0.00265	3.67	0.0003
CHAS	1	2.70806	0.84320	3.21	0.0014
CRIM	1	-0.12322	0.04673	-2.64	0.0086
DIS	1	-1.53175	0.19102	-8.02	<.0001
LSTAT	1	-0.50203	0.04823	-10.41	<.0001
NOX	1	-17.82549	3.55414	-5.02	<.0001
PTRATIO	1	-0.96539	0.12850	-7.51	<.0001
RAD	1	0.30375	0.06658	4.56	<.0001
RM	1	4.19406	0.42192	9.94	<.0001
TAX	1	-0.01196	0.00346	-3.46	0.0006
ZN	1	0.04191	0.01342	3.12	0.0019

Table 14 shows the comparison of OLS regression using a binary indicator variable and OLS regression using the Winsorized data.

Table 14 Comparison of OLS Regression Using Robust Binary Indicator and Winsorized Data

Performance Metric	**Robust Regression Using ROBUSTREG Binary Indicator**	**OLS Regression Using 98% Winsorized Data**
Root MSE	3.62416	4.67122
Coefficient of Variation	16.08394	20.71966
R^2	0.8481	0.7465
Adjusted R^2	0.8447	0.7408

The performance of the model built on the Winsorized sample data is clearly inferior to the model built on the original data using the binary indicator variable created by PROC ROBUSTREG. We conclude that there is significant information in the tails of the distributions of the Winsorized variables that is lost when a relatively indiscriminate Winsorization is performed on the Boston Housing sample data.

Treatment of Outliers Using Transformations

Sometimes, transformations may be used to mitigate the distorting effects of extreme values. For example, income is frequently distributed according to a log-normal distribution. Taking the logarithm of income then produces a predictor that is more normally distributed in the logarithm of the variable. This approach might be recommended to reduce the skewness of income, which if untreated would tend to produce disproportionate effects in the model's parameter estimates. However, introducing this transformed value into the regression equation changes the model from one where change in Y is proportional to change in $Income$ to one where change in Y is proportional to the change in the logarithm of $Income$ and hence the model is nonlinear in $Income$. For example, given the linear regression model

$$Y = \alpha + \sum_i \beta_i x_i + \gamma \cdot Income + \epsilon \,, \tag{6}$$

if we transform $Income$ to $\log(Income)$ then we must solve for the parameter estimates of the model

$$Y = \alpha + \sum_i \beta_i x_i + \gamma \cdot \log(Income) + \epsilon \,. \tag{7}$$

It can be rewritten into the regression model format

$$Y = \alpha + \sum\nolimits_i \beta_i x_i + \gamma \cdot x_L + \epsilon \tag{8}$$

where $x_L = \log(Income)$. This model is still a linear regression model because it is linear in the parameters, but is no longer straightforward. Explanation of the marginal change in Y with a unit change in $Income$ is now more complicated.

There are many transformations possible, but each may introduce complexity into the interpretation of the model's construction and complicate the insights into the data, which is the purpose of building a model. Domain knowledge may be useful in guiding the use of transformations and in explaining parameter estimates. Any factor that diminishes straightforward understanding of the model may also reduce the likelihood of acceptance of results.

Summary

We must approach outlier detection with respect for the data, since the observations in a set of data contain information relevant to the process being modeled. We must apply domain knowledge to help modelers decide which observations are typical and which are not characteristic of the generative process. Noisy data are especially subject to outliers, and will distort OLS regression model parameter estimates due to their extreme values. Robust regression algorithms such as PROC ROBUSTREG have been developed to mitigate the effect of extreme values. We may detect outliers by processing the data as a single group using global outlier detection algorithms, or in individual clusters using local detection algorithms.

Treating outliers using binary indicator variables to distinguish them from nonoutlying observations may produce more accurate results than Winsorizing them arbitrarily, which adds bias to the data and may introduce errors into the parameter estimation process. Modifying a variable's values via a mathematical or other transformation may improve its performance but at the cost of increasing the model's complexity and hence its interpretability. A model that is not straightforward to understand may not be accepted by the user community for which it has been developed.

Appendix A

The housing.names file contains descriptive information about the Boston housing data. Its URL is https://archive.ics.uci.edu/ml/machine-learning-databases/housing/ .

The attributes and metadata are given below.

Attribute Name	Description
CRIM	Per-capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 sq. ft.
INDUS	Proportion of non-retail business acres per town
CHAS	Charles River dummy variable (=1 if tract bounds river, 0 otherwise)
NOX	Nitric oxides concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centers
RAD	Index of accessibility to radial highways
TAX	Full-value property-tax rate per $10,000
PTRATIO	Pupil-teacher ratio by town
B	1000(Bk – 0.63)^2 where Bk is the proportion of African Americans by town
LSTAT	% lower status of the population
MED	Median value of owner-occupied homes in $1,000's

Appendix B

The %LOF_LoOP heading and parameter definitions are:

```
%macro LOF_LoOP( DSNIN, DSNOUT, K, VAR, DIST_FCN=euclid, LAMBDA=3, MIN_LOF=1,
                 MIN_LOOP=0.95, OUTLIER_PROB=.95, PLOT=NONE, PRINT=N
               ) / minoperator ;

/* PURPOSE: compute local outlier factor (LOF) and local outlier probability (LoOP)
 *          characterizing an observation as an outlier within a local neighborhood
 *
 * Create variable OUTLIER in &DSNOUT: if LoOP >= &OUTLIER_PROB then OUTLIER = 1 else 0
 *
 * Parameters:
 *    DSNIN        ::= name of input dataset containing interval-scaled observations
 *    DSNOUT       ::= name of output dataset containing k-NN info
 *    K            ::= k'th nearest neighbor within a local nbhd
 *    VAR          ::= list of interval-scaled variables to be treated for outliers
 *    DIST_FCN     ::= [optional] proximity measure, e.g., Euclidean or Manhattan distance
 *    LAMBDA       ::= [optional] multiplier of the standard distance value for outlier detection
 *    MIN_LOF      ::= [optional] min value to plot using LOF value as radius of plot bubble
 *    MIN_LOOP     ::= [optional] min value to plot using LOOP value as radius of plot bubble
 *    OUTLIER_PROB ::= [optional] if Local Outlier Prob >= &OUTLIER_PROB then OUTLIER = 1, else 0
 *    PLOT         ::= [optional] plot control flag: ALL, LOF, LoOP, NONE
 *    PRINT        ::= [optional] print control flag for PROC MODECLUS: Y or N
 */
```

We used the Output Display System to capture the graphics produced by the %LOF_LoOP macro.

```
ods listing close ;
ods pdf file=
"C:\Users\Username\Documents\My SAS Files\Outliers\SASCode\LOF_LoOP_Iris.pdf" ;
ods graphics on ;

%let IRIS_VARS = sepallen sepalwid petallen petalwid ;

%LOF_LoOP(LIBNAME.iris,iris_LOF_LOOP, 20, &IRIS_VARS,
          min_LOF=1.17, min_LOOP=0.2, outlier_prob=0.6, plot=all, print=n,
          dist_fcn=Euclid
         )

ods graphics off ;
ods pdf close ;
ods listing ;
```

The macro code for %LOF_LoOP is available upon request.

References

[1] Hawkins, D (1980). Identification of Outliers, Chapman and Hall, London.

[2] Barnett V., Lewis T. (1994). Outliers in Statistical Data, 3rd ed., John Wiley, New York.

[3] https://en.wikipedia.org/wiki/Leverage_(statistics)

[4] https://en.wikipedia.org/wiki/Anscombe%27s_quartet

[5] https://en.wikipedia.org/wiki/Studentized_residual

[6] https://en.wikipedia.org/wiki/Cook%27s_distance

[7] Neter, J, Wasserman, W, Kutner, M (1990). Applied Linear Statistical Models 3rd ed., Richard D. Irwin, Inc., Boston, MA.

[8] https://archive.ics.uci.edu/ml/machine-learning-databases/housing/

[9] Ben-Gal, Irad (2005). Outlier Detection, in: Maimon O. and Rockach L. (Eds) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer Academic Publishers, ISBN 0-387-24435-2.

[10] https://en.wikipedia.org/wiki/Winsorizing

[11] Breunig, Markus M.; Kriegel, Hans-Peter; Ng, Raymond T.; Sander, Jörg (2000). LOF: Identifying Density-Based Local Outliers (http://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf). SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp. 93–104.
[12] https://medium.com/@doedotdev/local-outlier-factor-example-by-hand-b57cedb10bd1
Local Outlier Factor | Simple Example By Hand, April 25, 2018.

[13] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems", Annual Eugenics, 7, Part II, pp. 179-188. Taken from the UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/iris]. Irvine, CA: University of California, School of Information and Computer Science.

[14] Kriegel, H; Kröger, P; Schubert, E; Zimek, A. (2009). LoOP: Local Outlier Probabilities (http://www.dbs.ifi.lmu.de/Publikationen/Papers/LoOP1649.pdf). *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. New York, NY, pp. 1649-1652.

[15] Xu, Zekun; Kakde, Deovrat; Chaudhur, Arin (2019). Automatic Hyperparameter Tuning Method for Local Outlier Factor, with Applications to Anomaly Detection. SAS Institute, Inc.

Bibliography

Wikipedia contributors. (2020, February 3). Mahalanobis distance. In *Wikipedia, The Free Encyclopedia*. Retrieved 00:07, February 20, 2020, from https://en.wikipedia.org/w/index.php?title=Mahalanobis_distance&oldid=938905763Acknowledgements

Acknowledgement

We thank Bruce Gilsen, JT Lehman, Mark Leventhal, Joseph Naraguma, and Donald Wedding for their generosity in taking the time to review this document.
Especial thanks to Jacob Madsen, who wrote the R package DDoutlier. This package, available from the CRAN.r-project.org and https://github.com/jhmadsen/DDoutlier, is an excellent collection of density-based outlier detection algorithms.

Contact Information

Your comments and questions are valued and encouraged. The code is available upon request. Contact the author at:

Name:	Ross Bettinger
Enterprise:	Consultant
E-mail:	rsbettinger@gmail.com