Understanding Regression Statistics – Part 2

This is the second publication that covers statistics that are sometimes generated by software when running a multiple linear regression. Part 1 examined regression statistics such as R^2, PRESS, adjusted R^2, VIF, standardized coefficients, etc. Part 2 focuses on residuals and how they can be used to determine how adequate the model is.

Multiple linear regression is used to build a model where one or more predictor variables (the X's) can be used to predict the response variable (Y). The model is an equation that describes the response variable in terms of the statistically significant predictor variables.

Software packages have the capability to generate a lot of other regression statistics beyond the model – all designed to help decide how "good" the model is. This publication focuses on residuals that are often included in the multiple linear regression analysis

In this publication:

Part 1 can be accessed at this link.

Example Data and the Model

We will use the same data that was presented in Part 1, which is from the book "Introduction to Linear Regression Analysis" by Douglas Montgomery, Elizabeth Peck, and Geoffrey Vining. A soft drink distributor wants to predict the amount of time (y) a delivery driver will take to service vending machines. There are two predictor variables that he is interested in exploring: the number of cases of product stocked (x_1) and the distance walked in feet by the delivery driver (x_2).

The data are given in Table 1. There are a total of 25 observations (delivery trips). For each trip, the number of cases stocked, the distance walked in feet, and the time taken in minutes were recorded.

A multiple linear regression was performed using these data and the SPC for Excel software. The purpose of regression analysis, of course, is to generate a model that predicts the response variable (delivery time) from the values of the predictor variables (number of cases and the distance walked).

© 2022 BPI Consulting, LLC
www.spcforexcel.com

Table 1: Delivery Time Data

Obs. Number	Number of Cases	Distance	Delivery Time	Obs. Number	Number of Cases	Distance	Delivery Time
1	7	560	16.68	14	6	462	19.75
2	3	220	11.50	15	9	448	24.00
3	3	340	12.03	16	10	776	29.00
4	4	80	14.88	17	6	200	15.35
5	6	150	13.75	18	7	132	19.00
6	7	330	18.11	19	3	36	9.50
7	2	110	8.00	20	17	770	35.10
8	7	210	17.83	21	10	140	17.90
9	30	1460	79.24	22	26	810	52.32
10	5	605	21.50	23	9	450	18.75
11	16	688	40.33	24	8	635	19.83
12	10	215	21.00	25	4	150	10.75
13	4	255	13.50				

The form of the model is:

$$y = b_0 + b_1x_1 + b_2x_2$$

where y is the response variable (delivery time), b_0 is the intercept, b_1 is the coefficient for x_1 (number of cases) and b_2 is the coefficient for x_2 (distance). The model from the SPC for Excel analysis is given below.

Delivery Time = 2.341231 + 1.615907(Number of Cases) +0.014385(Distance)

This model can now be used to predict the delivery time based on the number of cases and distance walked.

Predicted versus Observed Values

One method of determining how good the model predicts the observations is to plot the predicted values versus the observed values. If the model is good, you would expect the points to lie close to a straight line. This is done in Figure 1.

The SPC for Excel software was used to make the charts for the delivery analysis in this publication. The closer the points are to the line, the better the fit. This looks like a pretty good fit based on Figure 1.

Residuals

The difference between the predicted value and the observed value is called the residual. It is "how far" the point is off the line in Figure 1. A residual is defined as:

$$e_i = y_i - \hat{y}_i$$

© 2022 BPI Consulting, LLC
www.spcforexcel.com

where e_i is the i^{th} residual, y_i is the i^{th} observed value and $\hat{y}_i$ is the i^{th} predicted value. Table 2 shows the residuals for the data.

Figure 1: Predicted versus Observed Values for the Delivery Data

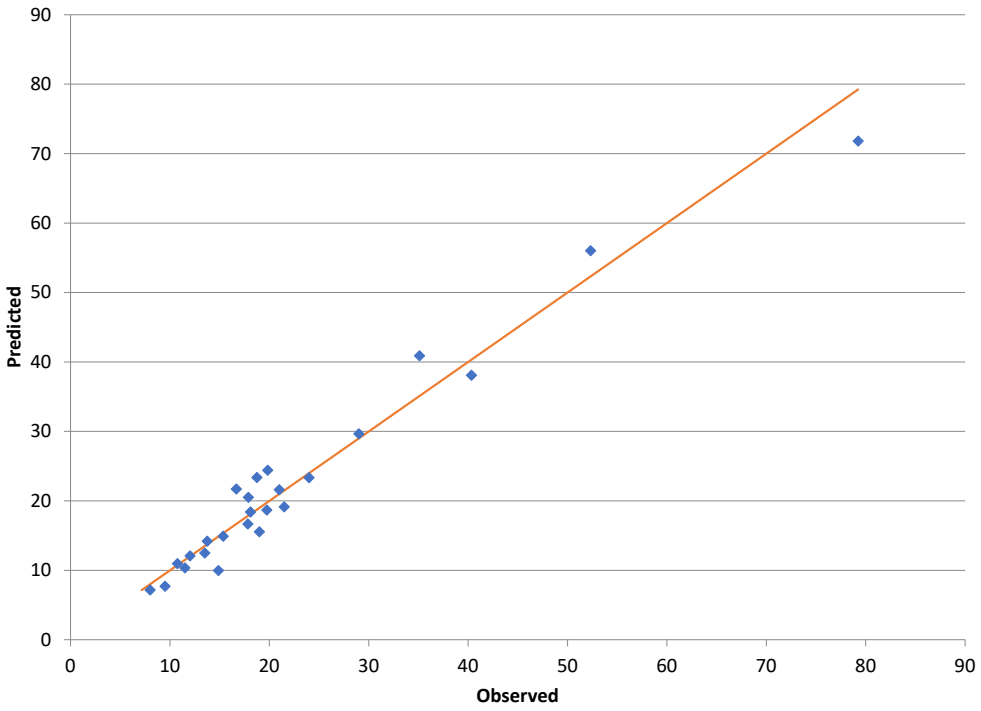


Table 2: Residuals

Obs. Number	Observed Value	Predicted Value	Residuals	Obs. Number	Observed Value	Predicted Value	Residuals
1	16.68	21.708	-5.028	14	19.75	18.682	1.068
2	11.5	10.354	1.146	15	24	23.329	0.671
3	12.03	12.080	-0.050	16	29	29.663	-0.663
4	14.88	9.956	4.924	17	15.35	14.914	0.436
5	13.75	14.194	-0.444	18	19	15.551	3.449
6	18.11	18.400	-0.290	19	9.5	7.707	1.793
7	8	7.155	0.845	20	35.1	40.888	-5.788
8	17.83	16.673	1.157	21	17.9	20.514	-2.614
9	79.24	71.820	7.420	22	52.32	56.007	-3.687
10	21.5	19.124	2.376	23	18.75	23.358	-4.608
11	40.33	38.093	2.237	24	19.83	24.403	-4.573
12	21	21.593	-0.593	25	10.75	10.963	-0.213
13	13.5	12.473	1.027				

© 2022 BPI Consulting, LLC
www.spcforexcel.com

The first observation was with 7 cases walking 560 feet. The predicted value is then:

$$\text{Delivery Time} = 2.341231 + 1.615907(\text{Number of Cases}) + 0.014385(\text{Distance})$$

$$\text{Delivery Time} = 2.341231 + 1.615907(7) + 0.014385(560)$$

$$\text{Delivery Time} = 21.70818$$

The observed value for observation 1 is 16.68, so the residuals for observation 1 is:

$$e_1 = y_1 - \hat{y}_1 = 16.68 - 21.70818 = -5.028$$

The rest of the residuals are calculated the same way.

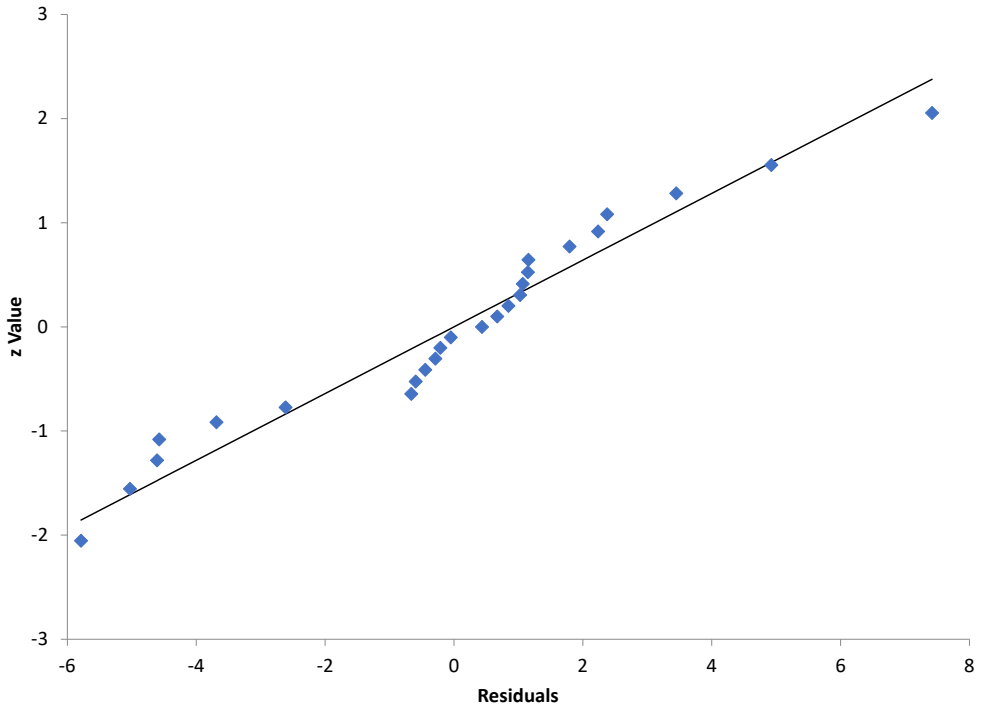
Note that observation 9 has the highest residual. This observation also had the largest number of cases and distance walked.

Now we have the residuals. What do we do with them? One use of them is to check some of the assumptions used in multiple linear regression. Another is to give insights in the how good the model is.

Residuals and Normality

One assumption is that the residuals are normally distributed. This should be checked. If they are not normally distributed, you may have to look at a different type of regression analysis. One way to check if the residuals are normally distributed is to create a normal probability plot as shown in Figure 2.

Figure 2: Residuals Normal Probability Plot for the Delivery Data



© 2022 BPI Consulting, LLC
www.spcforexcel.com

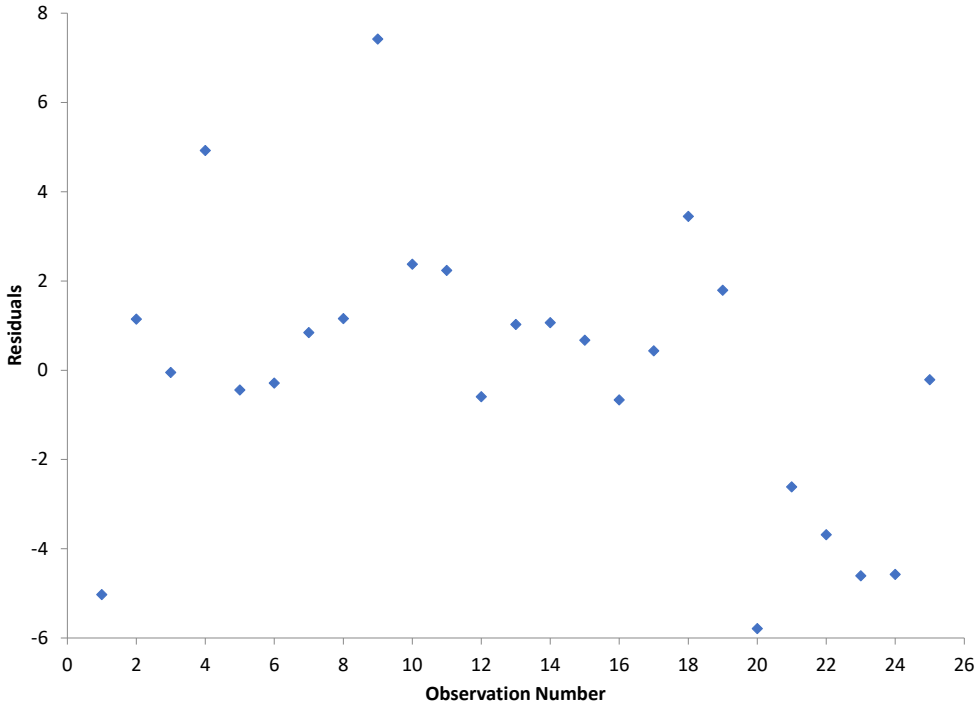
If the residuals are normally distributed, the points should fall along a straight line. Do the points lie along a straight line? Maybe, but maybe not. The p value for this plot is 0.11. Since it is greater than 0.05, you assume the residuals are normally distributed. But with a p value of 0.11, it is probably a borderline conclusion. There might be one or more outliers influencing the results, such as observation 9.

If the points are significantly off the line or the p value is less than 0.05, other regression methods beside linear should be considered.

Residuals and Observation Number

Another assumption is that the residuals are not correlated. One method of determining if the residuals are correlated is to plot the residuals over time (versus observation number assuming the observation numbers are in time-order). The residuals should bounce randomly around 0. Figure 3 is a plot of the residuals versus observation number.

Figure 3: Residuals versus Observation Number for the Delivery Data



The points appear to fluctuate randomly around 0, implying that the residuals are not correlated. If a positive correlation exists between the residuals, you get a chart like Figure 4.

In this example, a large residual tends to be followed by another large residual, while a small residual tends to be followed by another small residual.

If a negative correlation exists, then there is more of a sawtooth pattern as shown in Figure 5.

© 2022 BPI Consulting, LLC
www.spcforexcel.com

Figure 4: Residuals with a Positive Correlation Example

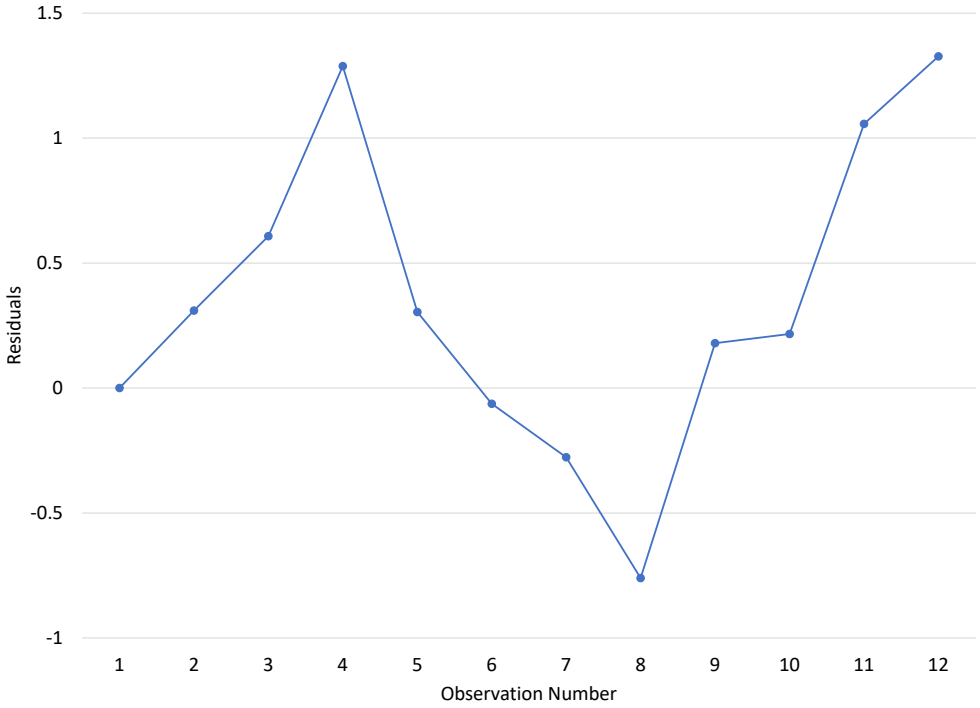
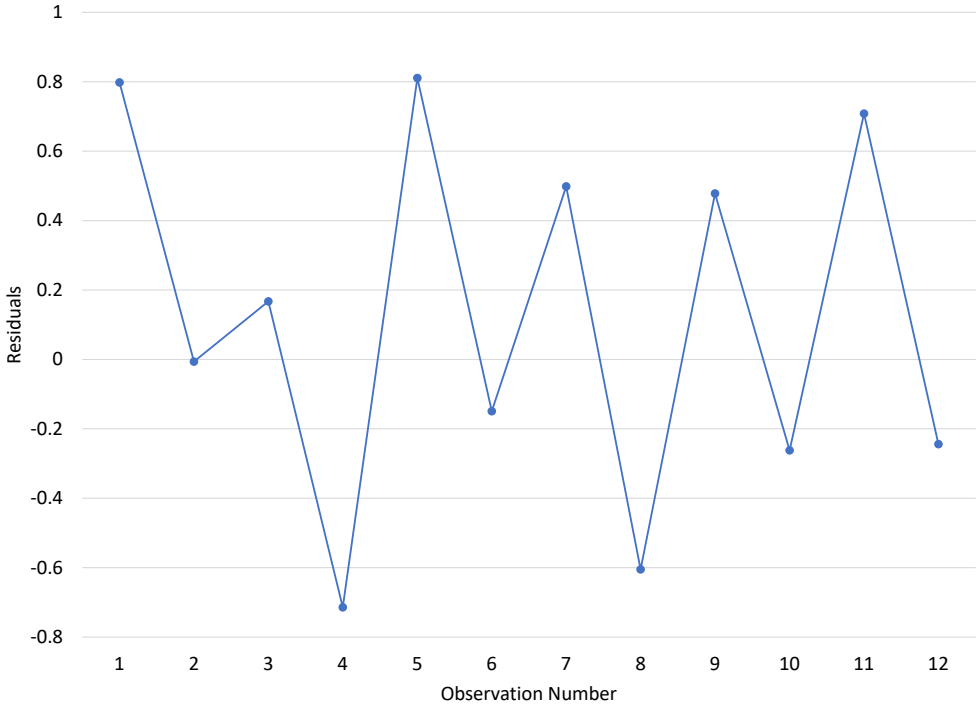


Figure 5: Residuals with a Negative Correlation Example



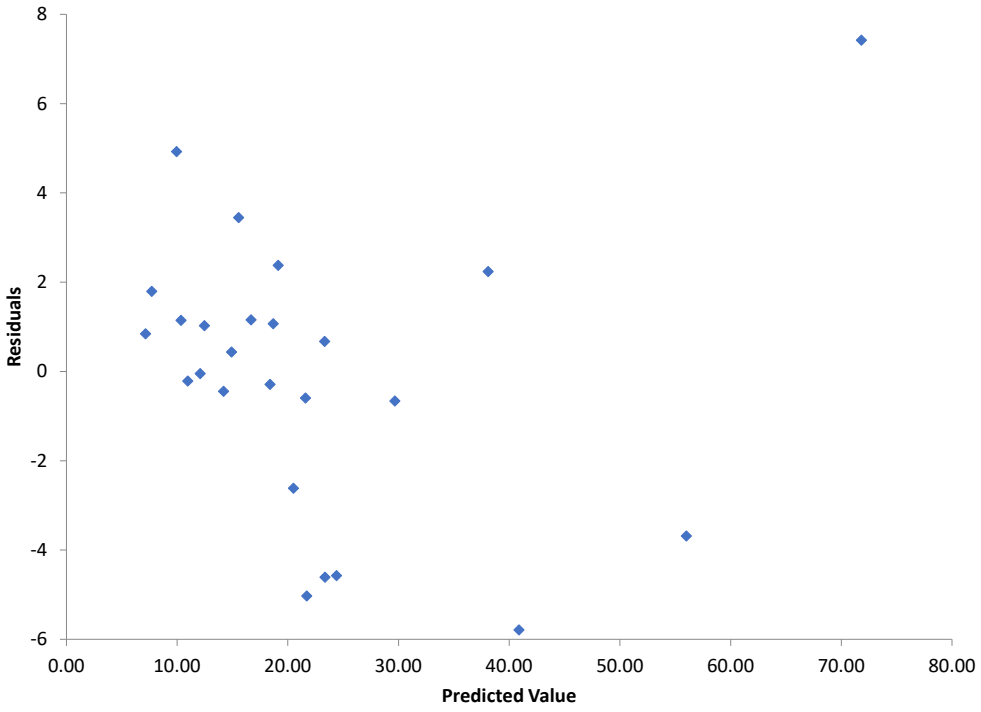
© 2022 BPI Consulting, LLC
www.spcforexcel.com

With a negative correlation, a “up” value is usually followed by a “down” value and vice versa.

Residuals and Predicted Values

Another assumption is the variance of the residuals is consistent. One method of checking this is to plot the residuals versus the predicted values for delivery time. The pattern on the chart will give you insights into the variance. Figure 6 shows the plot of the residuals versus the fitted values.

Figure 6: Delivery Time Residuals vs Fitted Values for the Delivery Data



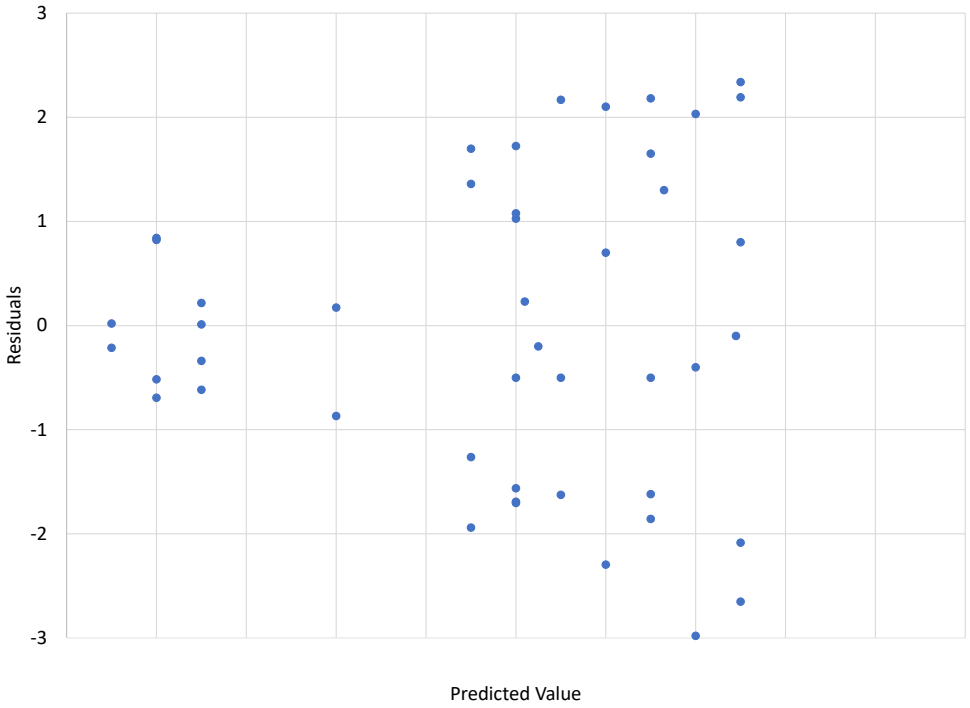
If the residuals can be contained within a band, then there is no issue with the variance of the residuals. You can see from Figure 6 that most of the results are within the band from -6 to 6. The numerical values of the band are not important; it is the pattern on the chart. There is one point outside that band (observation 9), which may be an outlier as pointed out earlier. Overall, it appears that there is no issue with the variance changing.

A somewhat common occurrence is for the variation in the residuals to increase as the observed values get larger. If this occurs, the points on the chart will form a cone – increasing in width as the values increase. An example of this is shown in Figure 7.

The variance of the residuals was determined in Part 1 in the ANOVA table. The variance is given by the mean square of the residuals (MS_{Res}), which is the sum of squares due to the residuals divided by n – p, where n is the number of observations and p is the number of parameters in the model. See Part 1 for more information.

© 2022 BPI Consulting, LLC
www.spcforexcel.com

Figure 7: Residuals with Increasing Variance



Scaled Residuals

Scaled residuals are very helpful in finding outliers. You can plot the scaled residuals to see if any are beyond certain limits. We begin with the standardized residuals.

Standardized Residuals

The standardized residuals are calculated by dividing the residuals by the square root of MS_{Res}.

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}}$$

MS_{Res} is 10.62 from Part 1 of this series. The first residual from Table 2 is -5.028, so the first standardized residual is given by:

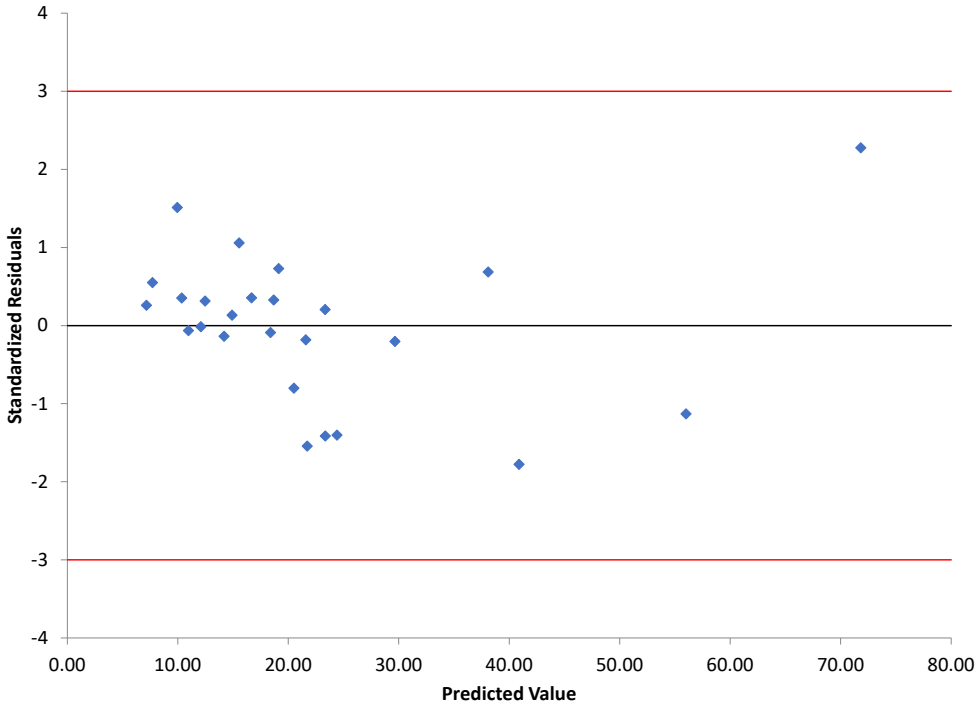
$$d_1 = \frac{e_1}{\sqrt{MS_{Res}}} = \frac{-5.028}{\sqrt{10.62}} = -1.543$$

The mean of the standardized residuals is 0 and the standard deviation is 1. You can use this then to define that an outlier occurs if any standardized residuals is greater then 3 or less than -3. Figure 8 shows this.

You can see point 9 is within the +/- 3 limits for the standardized residuals implying it is not an outlier.

© 2022 BPI Consulting, LLC
www.spcforexcel.com

Figure 8: Standardized Residuals versus Predicted Values for Delivery Data



Internally Studentized Residuals

The internally studentized residuals improves on the standardized residuals. It takes into account the inequality of variances across the factors. It makes use of the H matrix (see Part 1). The internally studentized residuals are given by:

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1 - h_{ii})}}$$

where h_{ii} is the i^{th} diagonal element in the H matrix. The H matrix was shown in the Excel workbook that you can download from Part 1.

For the first observation, h_{11} is 0.1018. So, the internally studentized residual for the first observation is:

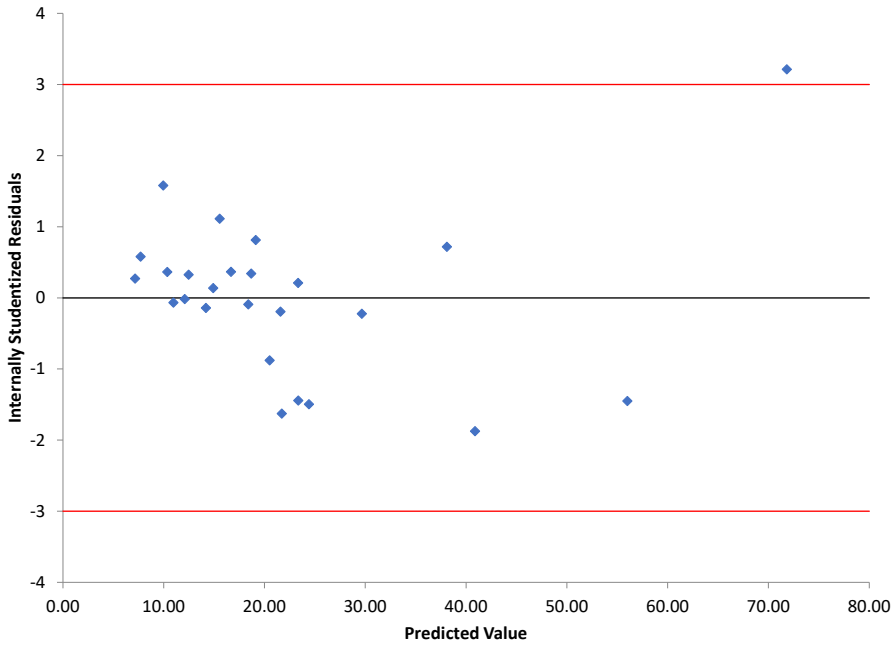
$$r_1 = \frac{e_1}{\sqrt{MS_{Res}(1 - h_{11})}} = \frac{-5.028}{\sqrt{10.62(1 - 0.1018)}} = -1.628$$

Internally studentized residuals greater than 3 or less than -3 are considered outliers. It should be noted some resources us 2 and – 2 as the limits for outliers. Figure 9 shows the internally studentized residuals versus the predicted value.

Note that point 9 is now an outlier.

© 2022 BPI Consulting, LLC
www.spcforexcel.com

Figure 9: Internally Studentized Residuals versus Predicted Values for Delivery Data



Externally Studentized Residuals

The externally studentized residuals are similar to the internally studentized residuals but use a different estimate of the variance. The estimate of the variance is based on a dataset with the i^{th} observation removed. It is given by S_i^2.

$$S_i^2 = \frac{(n-p)MS_{Res} - e_i^2/(1-h_{ii})}{n-p-1}$$

where n= number of observations and p = number of parameters (including the constant is present). This variance is used to determine the externally studentized residuals:

$$t_i = \frac{e_i}{\sqrt{S_i^2(1-h_{ii})}}$$

For the first observation:

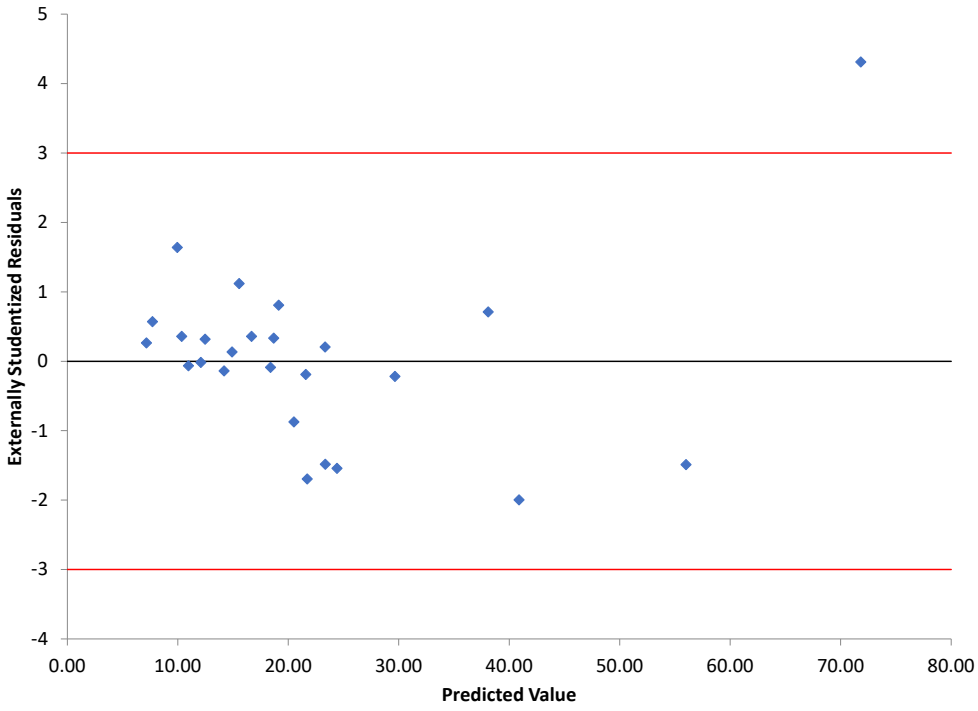
$$S_1^2 = \frac{(25-3)10.6239 - 7.41971^2/(1-0.4983)}{25-3-1} = 5.9$$

$$t_1 = \frac{7.41971}{\sqrt{5.9(1-0.4983)}} = 4.31$$

© 2022 BPI Consulting, LLC
www.spcforexcel.com

Figure 10 shows the externally studentized residuals versus the predicted value.

Figure 10: Externally Studentized Residuals versus Predicted Values for Delivery Data



Comparing Figures 9 and 10 show that observation 9 is even further out when looking at the externally studentized residual. There are indications that observation 9 may be an outlier. You could rerun the regression leaving out observation 9 to see what impact it has on the results.

Summary

This publication has examined the residuals that are often calculated as part of multiple linear regression. These residuals give you insights to how adequate the model is. Plotting residuals is a very effective way of visually seeing how adequate the model and your assumptions are. Scaled residuals give you additional insights.

Quick Links

Visit our home page

SPC for Excel Software

Our YouTube Channel for Videos

SPC Training

SPC Consulting

© 2022 BPI Consulting, LLC
www.spcforexcel.com

SPC Knowledge Base

Ordering Information

Thanks so much for reading our publication. We hope you find it informative and useful. Happy charting and may the data always support your position.

Sincerely,

Dr. Bill McNeese
BPI Consulting, LLC

© 2022 BPI Consulting, LLC
www.spcforexcel.com