Predicted revenue and prediction interval

Science, Statistics

From the Scatterplot of Revenue vs. Circulation, it can be seen that the variance of the dependent variable, Revenue, is increasing. This is a violation of the Gauss-Markov condition of constant variance in the error terms. Also, since a linear relationship seems appropriate, transformation of both the dependent and independent variables are necessary. 2) Fitting polynomial models to the data may be better than fitting a straight line model to the untransformed data because this allows for curvature and can fit the data more closely.

However, this might not be sufficient because it does not account for nonconstant variance. 3) The natural log transformation of both variables provides the best model of the three. From the plot of the Regression Line for lnRevenue vs. lnCirculation, it can be seen that the points are relatively equally scattered around the regression line. Also, the nonconstant variance seems to be fixed. This is evident in the plot of the residuals vs. predicted values, as the points are randomly scattered about the center line.

The square root transformation of both variables improves linearity, as indicated in the plot of the Regression Line for sqrtRevenue vs. sqrtCirculation, but does not fix the problem of non-constant variance. This can be clearly seen in the plot of the residuals vs. predicted values. The points are not randomly scattered around the center line, but seem to be bunched up on the left side and spread outwards, indicating increasing variance. The inverse transformation of both variables does not improve linearity, as curvature can be seen in the plot of the Regression Line for invRevenue vs. invCirculation.

Although non-constant variance is slightly improved over the square root transformation, as can be seen in the plot of the residuals vs. predicted values, it is still insufficient. Therefore, both variables natural log transformed seems to be the best model of the three choices. 4) The model used is . This implies that . From this result, it can be seen that a k-fold change in the circulation in millions results in a change in revenue in thousands of dollars. From the regression, = 0. 5334. This means that if circulation changes by a factor of k, its revenue will also change by a factor of k0. 334. 5) From SAS, a 95% prediction interval with a circulation of 1 million for the natural log of the revenue is (4. 3005, 5. 0202) with a predicted value of 4. 6604.

This translates to a prediction interval of ($73 736. 65, $151 441. 59) with a predicted revenue of $105 678. 35. 6) Since the threshold for Cook’s D is 4/(n-2), where n= 70, the threshold is 0. 059. There are five values with Cook’s D greater than 0. 059, which indicates that they are influential points. From the normal Q-Q plot of the residuals, these 5 points can be seen to be utliers at the ends of the graph. Therefore, they can greatly affect the fit of the model. Also from the normal Q-Q plot, it can be seen that the residuals are not exactly normally distributed. The curvature at the ends of the plot indicates heavy tails in the distribution. By the Central Limit Theorem confidence intervals, and the values for , , and E(Y) are valid. However, since a prediction interval deals only with a single point, it is not valid. Due to the heavy tails in the distribution of the error terms, the prediction interval calculated in 5) may not be accurate.