Kitu Komya (UID: 404-491-375)
Statistics 101A (Discussion 3A)
Homework 9 (due: 12/02/16)

Please note: I'm concurrently taking Stats 20, so I'm still familiarizing myself with R.

1a. Looking at the scatterplot matrix, we see non-linear relationships between the response variable of suggested retail price and each of the predictors. Moreover, the residual plot is not a random plot; it is heteroskedastic in nature in that the variance keeps increasing. This is further corroborated by the scale-location plot in which there is blatant evidence of non-constant variance since the plot is not flat but rather upwardly increasing. Because the predictors do not appear to be linearly related with the response variable, we need to consider transformations of the response and the predictor variables.

1b. Because there is a definite pattern in the residual plot, the linear model is a bad fit for the data. Really, however, when you look at the residual plot, you realize that there are a few outliers that make the trend of the residual plot curved. So perhaps removing these points will lead to a more flat line. Regardless of the outliers though, there is a heteroscedastic trend in which the variance is increasing, so this will be an invalid model still, since there should be a random scattering of the data points.

1c. Looking at the residuals vs leverage plot, we see a few high leverage points, notably points 223 and 67 since they are well outside the cut off line of [2*(p+1)]/n. From these two, only 223 is outside the standardized residuals interval of [-2, 2], so point 223 is an outlier as it is a bad high leverage point. It is noteworthy to note that since it falls outside of the red contour lines of Cook's distance, point 223 is also an influential point.

1d. Figure 6.55 shows a scatterplot matrix of the transformed response and predictor variables. The pairwise relationships are much more linear than those in Figure 6.53.

Moreover, in 6.56, the residual plot shows a random pattern in its scattering, despite a slight upward trend, thus proving that this is a valid model. Moreover, the normal qq plot shown makes it obvious that the errors are normally distributed since the data lie on a straight line. The scale-location plot, in conjunction with the residual plot, show that the non-constant variance issue has been fixed since now the trends are generally nonexistent and only a flat line exists, despite some slight dips. The last residuals-leverage graph shows that there are actually two influential points since they are outside the red contour lines of Cook's distance as well as a good high leverage point since it is inside the standardized residuals range of [-2, 2]. This difference of one more influential point is hardly deterring.

As seen from the MMP plots in Figure 6.57, after the transformation, all of the nonparametric estimates as solid blue line curves are almost indistinguishable from the smooth fitted values shown as dashed red lines. Thus, the transformation is valid.

1e. Using the relationship between SS_{reg} and RSS, the formula for calculating the F-statistic is as follows: $F = {\frac{RSS_{red} - RSS_{full}}{change\ in\ df}} \Big/ {\frac{RSS_{full}}{n-p-1}} = {\frac{7.232 - 6.717}{2}} \Big/ {\frac{6.717}{234-7-1}} = 8.664$

where RSS = df * $(RSE)^2$

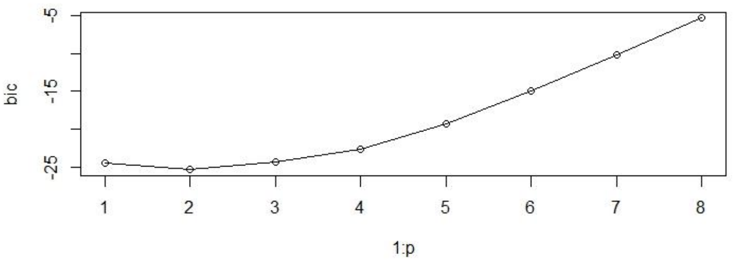
Using an online calculator, the critical F-value is 1.244. Since the F-statistic value of 8.664 is greater than the critical F-value of 1.244, we reject the null hypothesis. In essence, there is significant difference between the two models, so it is not sensible to remove the insignificant variables since that will give an entirely different model that is invalid.

1f. Had another categorical variable been added that displayed the vehicle manufacturer, the effect of manufacturer could also be measured. The variables would have to be factors, which are categorical variables that are not numerical but rather distinct categories. Like the hybrid variable (but in this case, with many more levels in the factors than just yes or no), the summary test would output all of the factors and their significance levels as if they are their own variables. So basically, all of the different levels of the factors of manufacturers would be treated as their own variables.

2a. The linear model for this dataset is (bear with me): Pounds = -3.855408 – 0.312481*sex + 0.011539*fage + 0.016641*mage + 0.023626*meduc + 0.267083*weeks + 0.006242*visits – 0.024871*cignum + 0.007445*gained – 0.160720*smoke

2b. At lambda = 0, the p-value is significant at a 5% significance level so we reject the null hypothesis that lambda = 0, and thus, it's best to not do a log transform. At lambda = 1, the p-value is not significant at a 5% significance level so we fail to reject the null hypothesis that lambda = 1 which means that no transformation is necessary. We can further confirm this as we see that the estimate from the R console is 0.969 which is nearly 1. Thus, the box-cox transformation did not improve the validity of the model.

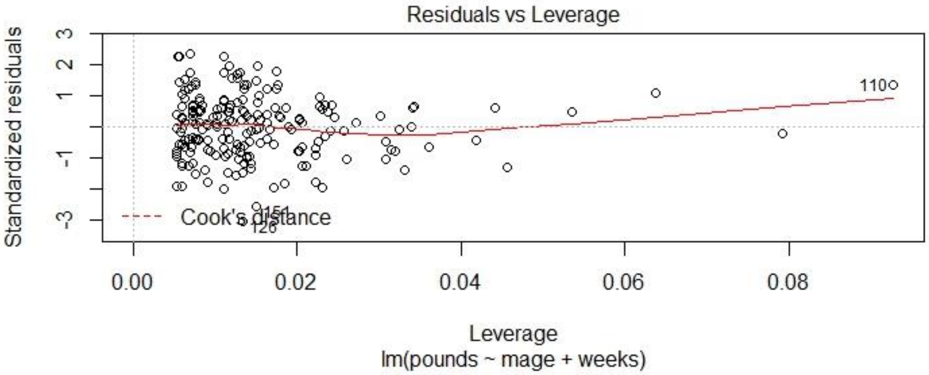
2c. Thus, let's use the inverse response plot to find a better model. However, upon looking at the lambda values, we see that there is not much of a difference in bettering the model just using the untransformed response variable. Hence, we will use the backwards stepwise procedure. Here is the bic plot:



Since the minimum value happens at 2, let's use that number once we get to the table. Using summary on the model and looking at row 2, mage and weeks are the significant variables. Thus, the model is: pounds = -4.05288 + 0.04063*mage + 0.26653* weeks. This is the smallest bic value so it is the best model that the backwards stepwise regression can achieve. I have used this method because it makes sense to start with all of the variables first and then rid of the useless ones. It's adjusted r square value is 0.2302, while the original model had an adjusted r square value of 0.2375. Although the value is smaller, perhaps we are avoiding overfitting.

2d. Using the variance inflation factor to understand the problem of collinearity, we receive values of 1.008118 for mage and 1.008118 for weeks as well. Since these numbers are less than 5, collinearity is not a problem.

2e.



There are no influential points since there are no points outside of the red contour lines as depicted by Cook's distance. There is, however, a high leverage point (point 110), but since it is within the standardized residuals interval of [-2, 2], this is a good high leverage point. We can further evaluate this point by seeing the effect on the adjusted r square before and after the point. But this is for a later discussion.

2f. Since our model does not have smoking as a variable, a mother's smoking habits does not affect the weight of her baby in terms of pounds. This is an interesting analysis of our data and we can conclude that sometimes our data does not match with what we expected.