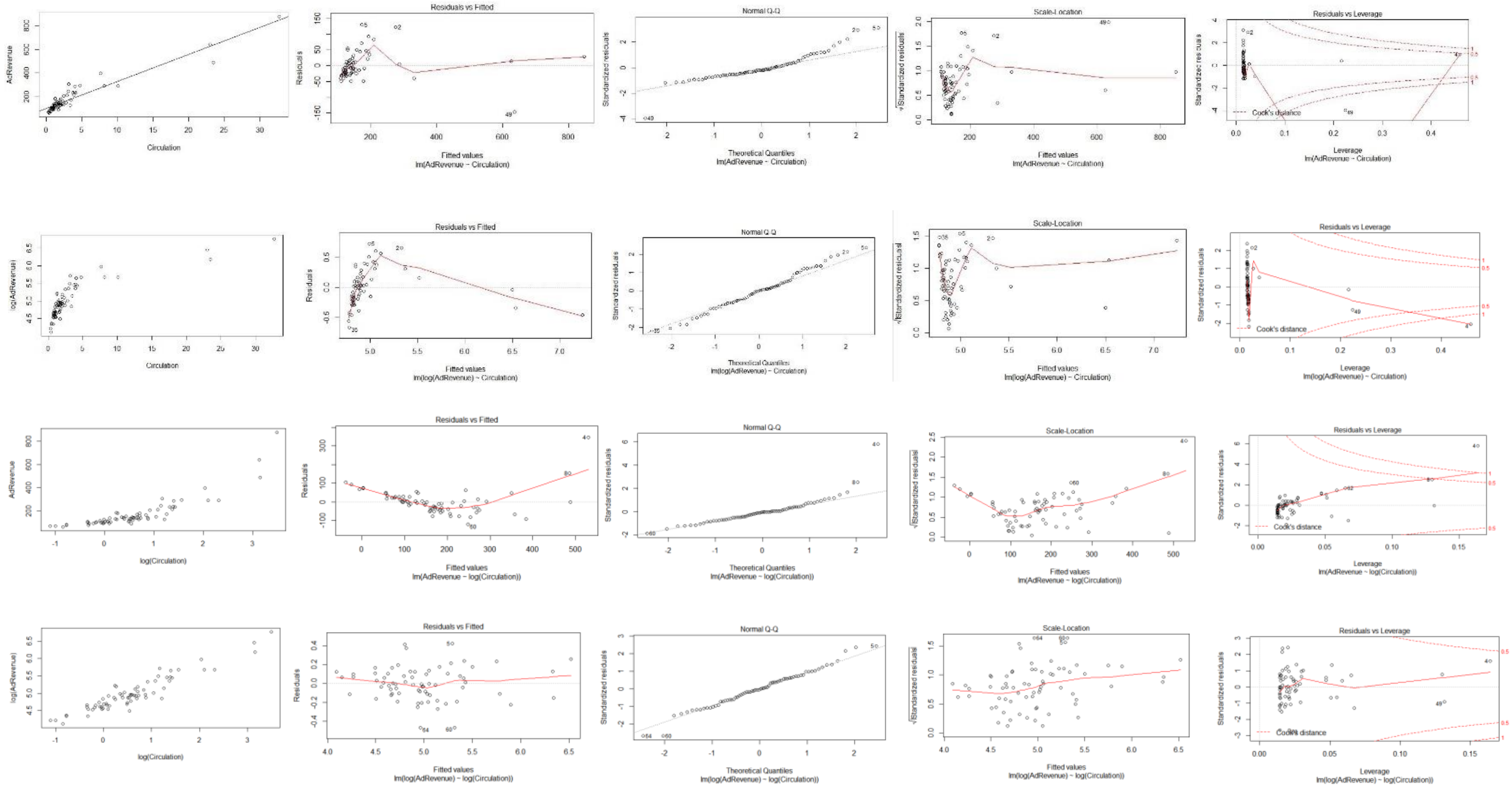
Kitu Komya (UID: 404-491-375)
Statistics 101A (Discussion 3A)
Homework 6 (due: 11/04/16)

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

1. Part A

scatterplot residuals qq plot scale-location leverage



(a) From top to bottom, the models are non-transformed linear (1), transformed log for response variable (2), transformed log for predictor variable (3), and transformed both variables (4). From here on out, I will be referring to these models by their assigned (#).

Scatterplots: only (1) and (4) show a sign of linearity via a linear least squares regression line. (2) and (3) clearly do not fit the pattern, so they seem like invalid models.

Residual plots: (1) has a slight increasing red smoother line at the end of its range after a large dip, (2) has a strong negative linear trend, (3) has a parabolic trend, and (4) is generally very flat. Since we want to see no pattern or trend in the residual plot in order to prove linearity and constant variance, only (4) seems like a valid model.

QQ plots: (1) and (3) have large deviation at their extrema. A QQ plot should follow a straight line closely in order to demonstrate normality of the errors, which only (2) and (4) satisfy.

Scale-location plots: similar to residual plots, we should see no trend in order to satisfy the linearity and constant variance assumptions. However, in (1) we see a negative linear trend, in (2) a large dip then rise then dip again and then finally increasing slowly, in (3) a parabolic like trend, and in (4) a generally flat line with small linear increases. Thus, (4) seems like the only valid model here.

Leverage points: although having leverage points don't affect the four assumptions for a model, it's useful to check for any bad leverage points, which are outliers, to see how many outliers a model has to determine how badly a model fits its data. In (1) and (2) and (3) we see 1 influential points since it is outside the red dashed contour lines. There are also 3 leverage points, 2 of which are likely to be good and 1 bad because that is outside the [-2, 2] standardized residual interval. In (4) there are no influential points because there are no points outside of the red dashed contour lines. However, there are likely 2 leverage points, both of which are good since they lie in the [-2, 2] standardized residual interval. Overall, we see that the model of (4) best represents its data points with the least influential and bad leverage points.

In conclusion, we see that model 4 passes all 4 assumptions of linearity: it's linear, its errors are constant, its errors are normally distributed, and the errors are independent, as seen from the diagnostics. Thus, the best simple linear regression model is when both of the variables are transformed.

Using summary(), the equation of this model is log(Y) = 4.67473 + 0.52876*log(x).

(b)

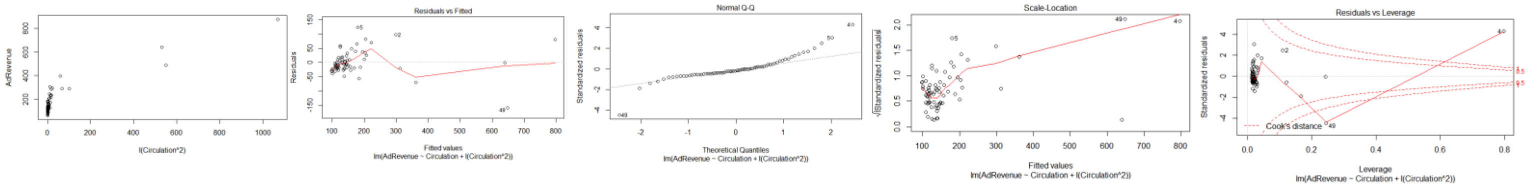
(i) Using the predict function, [3.947855, 4.6686] is the 95% prediction interval of advertising revenue per page for 0.5 million magazines in circulation.

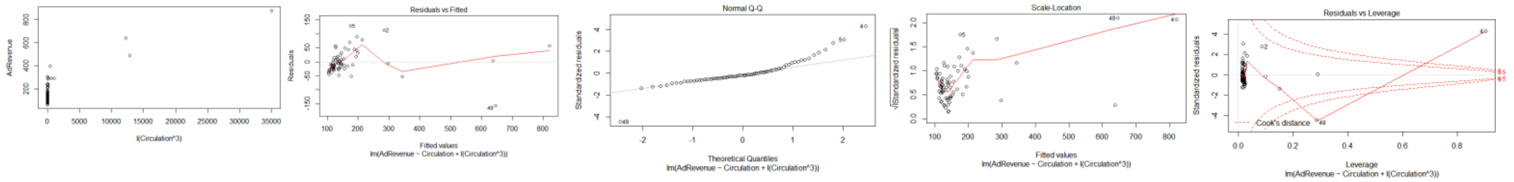
(ii) Using the predict function, [5.885815, 6.631689] is the 95% prediction interval of advertising revenue per page for 20 million magazines in circulation.

(c) We are using log transforms on both variables, and although this model is a better fit than the other three models (as seen in part a), we don't know that reality actually follows this. The majority of the weaknesses that were identified in the non-transformed linear model were addressed by using the log transform on both of the variables. There are some high leverage points that may need examination and a possibility of non-constant variance seen in scale-location. Furthermore, our model can't be used for extrapolation beyond the range of its values because we have not determined that it follows the same trend below and above these ranges. Hence, extrapolation is a weakness. Moreover, causality cannot be determined from our model. These data points come from an observational study, not an experiment. We cannot say that as the circulation increases, the ad revenue changes. No, we simply see an association between the two variables, and there may be confounding variables such as budget or location. We only see what happens when we compare two magazines with different circulation values. Hence, making causal claims is another weakness of our model.

Part B

(a) The equation of the quadratic model is E(Y|x) = 88.1390 + 29.5006x – 0.2394x^2. The equation of the cubic model is E(Y|x) = 59.17037 + 51.23582x – 2.50538x^2 + 0.05223x^3. The multiple r-squared value of the quadratic model is 0.901, while for the cubic model it's 0.9333, so clearly the cubic model is a better fit since the model describes more of the variability in the diamond ring price than the quadratic model. Following are the diagnostic outputs for both the quadratic (top) model and the cubic (bottom model). From left to right, we see the scatterplot, residual plot, QQ plot, scale-location plot, and the residuals-leverage plot.





The scatterplots are transformed so that is why we cannot deduce any trend. The residual plots for both models don't look too good or too bad. There is some kind of a trend of a dip and then slight increase, so our assumptions for linearity and for constant variance must be cautioned. Moreover, the qq plots are okay. The majority of the points follow the straight line, but there is heavy deviation at both ends, not more notably in either. The scale-location plot proves that there is non-constant variance in both of the models, which is indicative of a bad fit. However, the cubic model's constant variance assumption is less violated. In both of the graphs, there is one influential point outside of the red, dashed contour lines, and two bad high leverage points outside of the standardized residuals interval of [-2, 2]. Overall, though by little, the cubic model is a better fit because its deviations from the assumptions are less severe than the quadratic model.

(b) (i) 95% prediction interval for a circulation of 0.5 million copies on the cubic model is [14.92314, 153.4138].

(ii) 95% prediction interval for a circulation of 20 million copies on the cubic model is [418.179, 580.8878].

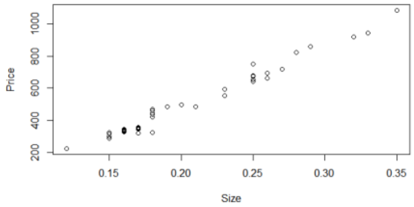
(c)

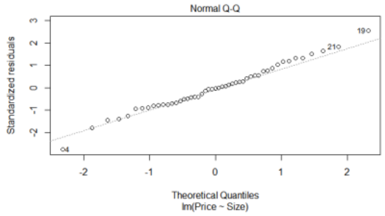
Part C

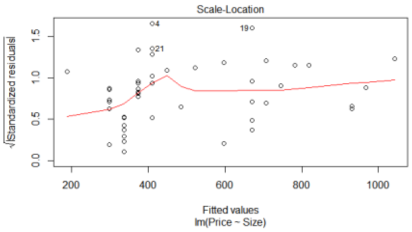
As mentioned, the assumptions of constant variance and linearity are questionable in the cubic model. This is not the best model, but it is also not the worst. Furthermore, our model can't be used for extrapolation beyond the range of its values because we have not determined that it follows the same trend below and above these ranges. Hence, extrapolation is a weakness. Moreover, causality cannot be determined from our model. These data points come from an observational study, not an experiment. We cannot say that as the circulation increases, the ad revenue changes. No, we simply see an association between the two variables, and there may be confounding variables such as budget or location. We only see what happens when we compare two magazines with different circulation values. Hence, making causal claims is another weakness of our model

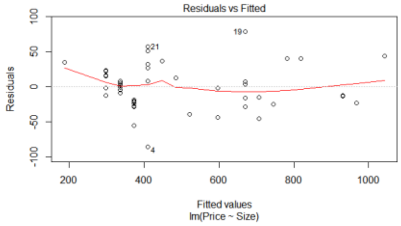
2. Part 1

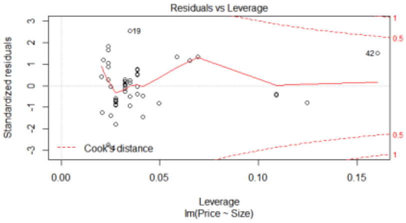
(a) I have used a linear model to describe the data. Below are the R outputs. From left to right, top to bottom: scatterplot, residual plot, scale-location plot, QQ plot, and residuals-leverage plot. The equation of the model, as seen from the summary finction in R, is E(Y|x) = -258.05 + 3715.02x. An interpretation of the slope is as follows: Diamond stones that are 1 carat heavier, on average, cost the diamond ring $3715.02 more. Note that this is an association conclusion, not a causality one.









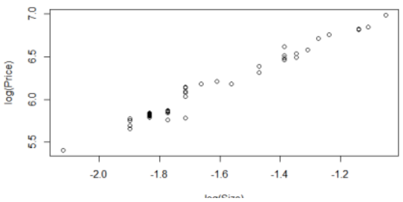


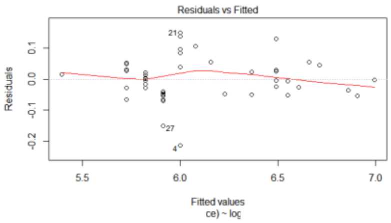
The scatterplot looks approximately linear, which is a good indication of following the linear assumption. The residuals plot has a slight parabolic trend in its plot, but the trend is small, yet we should still be cautious about our linear assumption. The scale location plot shows a slight increase in constant variance over time, yet since the slope of the red smoother line is small, we can still continuo. The QQ plot looks great as the data points lie on the straight line, indicating that the assumption of normality of the errors have been met. And the final graph shows that there are no influential points since there aren't any points outside of the red dashed contour lines and also no bad leverage points. 42 is a good high leverage point because it is within the standardized residuals' interval of [-2, 2] so we can keep this point.

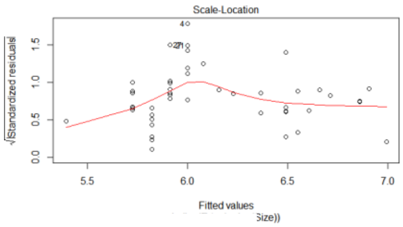
(b) As mentioned in part a, there is a slight parabolic trend in the residual plot which leads us to believe that possibly the linear model is the wrong model for the data. The scale-location plot also shows a non-flat line, indicative of a non-constant variance. Moreover, our model can't be used for extrapolation beyond the range of its values because we have not determined that it follows the same trend below and above these ranges. Hence, extrapolation is a weakness. Furthermore, causality cannot be determined from our model. These data points come from an observational study, not an experiment. We cannot say that as the carat size/weight of a diamond stone increases, the price of the diamond ring increases. No, we simply see an association between the two variables, and there may be confounding variables such as vendor or cost to produce. We only see what happens when we compare two diamonds stones of different sizes/weights. Hence, making causal claims is another weakness of our model.

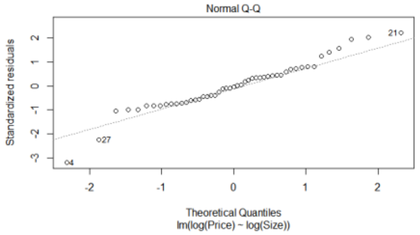
Part 2

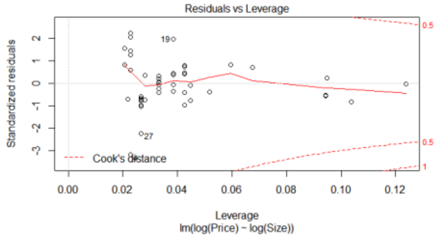
(a) I have used a log transform on both of the variables to describe the data. Below are the R outputs. From left to right, top to bottom: scatterplot, residual plot, scale-location plot, QQ plot, and residuals-leverage plot. The equation of the model, as seen from the summary finction in R, is $E(\log(Y)|\log(x)) = 8.56317 + 1.49566x$. An interpretation of the slope (in terms of elasticity, or percent change) is as follows: A 1% increase in the carat weight of a diamond stone is associated with a 1.49566% increase in the mean of the cost the diamond ring. Note that this is an association conclusion, not a causality one.











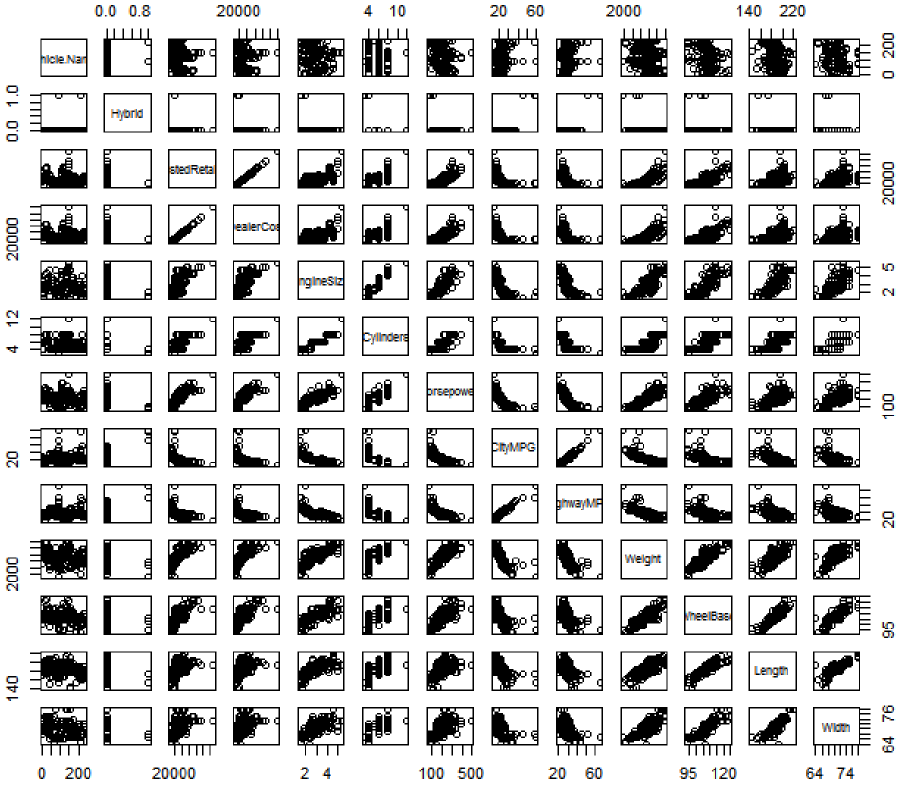
The scatterplot looks approximately linear, which is a good indication of following he linear assumption. The residual plot is mostly flat, except for a slight dip near the end of the plot. Still, it is largely flat and thus we can assume that the linearity and constant variance assumptions hold. The scale-location plot looks like a normal distribution, which is kind of troubling since we expect a flat line. Thus, we need to be cautious about the constant variance assumption. The QQ plot looks good enough to conclude normality of the errors, however there are deviations at the extrema that require attention. However, it is straight enough and enough on the line to verify the assumption. There are no influential points since there aren't any points outside of the dashed, red contour lines. There are also no bad high leverage points since all of the data fall within the standardized residuals' interval of [-2, 2].

(b) As mentioned in part a, the scale-location plot provides some worry over the constant variance assumption since it seems that it varies in a normal distribution like way. Moreover, the normality of the errors is not as normal as we would like, since the data points have deviation in the extrema in the QQ plot. Moreover, just like Part 1, our model can't be used for extrapolation beyond the range of its values because we have not determined that it follows the same trend below and above these ranges. Hence, extrapolation is a weakness. Furthermore, causality cannot be determined from our model. These data points come from an observational study, not an experiment. We cannot say that as the carat size/weight of a diamond stone increases, the price of the diamond ring increases. No, we simply see an association between the two variables, and there may be confounding variables such as vendor or cost to produce. We only see what happens when we compare two diamonds stones of different sizes/weights. Hence, making causal claims is another weakness of our model.

Part 3

In comparing the models from Part A and Part B, I find the model in Part A to be a more reliable and better model. The model from Part A clearly follows through the normality of the errors assumption as in its QQ plot, the data lie on a straight line, while the model in Part B doesn't seem to have as convincing of a verification of this assumption. If the normal condition fails, then we can only get good approximations of the slope, intercept, intervals, etc if the sample size is large. In our case, a sample size of 49 is only a moderate sample size, and thus the consequences of not following a normal distribution of errors may follow, which is pretty fatal, although unbiased estimates of intercept, slope, and predicted values. Moreover, the model in Part B is more worrisome in terms of the constant variance assumption, as seen through the scale-location plot, which should be flat, but instead follows a normal-distribution like. If the constant variance assumption fails, then everything fails, and there is no recovering. Although Part A also gives reason to be concerned about constant variance as seen through a slight dip in the residual plot and a slight increase in the scale-location plot, its concern is not nearly as grave as a very obvious trend in the scale-location plot in Part B. Thus, Model A is more reliable.

3A.

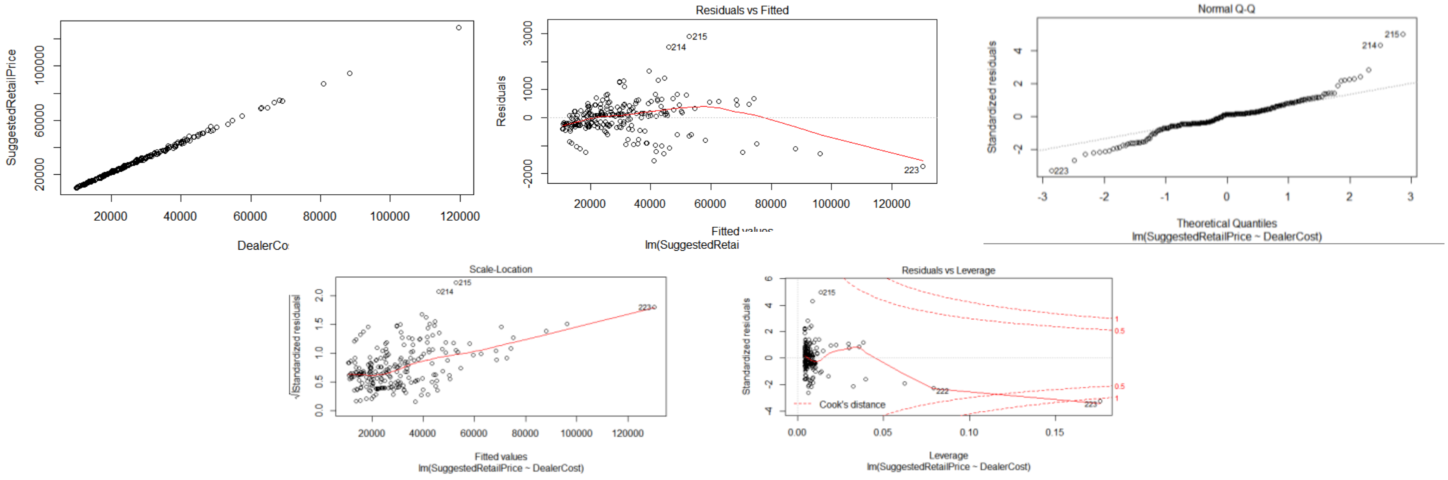


Looking at the scatterplot matrix, the linear association between dealer cost and suggested retail price are very strongly related. Using summary(), the equation of the model is

$$E(Y|x) = -61.904248 + 1.088841x.$$

The An interpretation of the slope is as follows: dealer costs that are $1 more, on average, suggest a retail price of $1.088841 more. Note that this is an association conclusion, not a causality one. Using confint() a 95% confidence interval of the slope is [1.083644, 1.094039].

3B. I have provided the outputs of the following diagnostics, from top to bottom, left to right: scatterplot, residual plot, qq plot, scale-location plot, and residuals-leverage plot.



This is unfortunately a bad model. Although it seems from the scatterplot that the linear model will be a great fit for the data, we see from the residuals plot that the constant variance and linearity assumptions are violated since the red smoother line is not flat, but rather has a negative trend, indicative of an invalid model fit. Moreover, although the majority of the points in the QQ plot do lie on the line, there is heavy deviation at the extrema, cautioning us to be aware of non-normality of the errors. Moreover, the scale-location plot is what breaks the model. There is very clear suggestion of non-constant variance as the red smoother line indicates that that the variance increases over x. Violating this assumption is horrible and usually not recoverable. From the leverages plot, we see one influential point since it is outside of the red, dashed contour lines, and two bad high leverage points (labelled 215 and 222) since they are outside the standardized residuals' interval of [-2, 2]. Hence, the weaknesses of this model include non-constant variance and possible non-normality of the errors, the first of which is a deadly error. Thus, the model does not fit the data well.

3C. Assuming all the necessary conditions held are valid, and using the summary function, at the 95% significance level, the variables that are good predictors of retail price are all those variables whose p-values are smaller than 0.05. This, in essence, means that the association between that particular variable and the retail price is nonzero, indicative of an association. The significant variables are: Dealer Cost, Cylinders, Weight, and Width.

3D. The co-efficient (estimate) of Weight is 0.7318. The interpretation of this slope is that for cars that are 1 kilogram heavier, on average have a retail price of $0.7318 more.