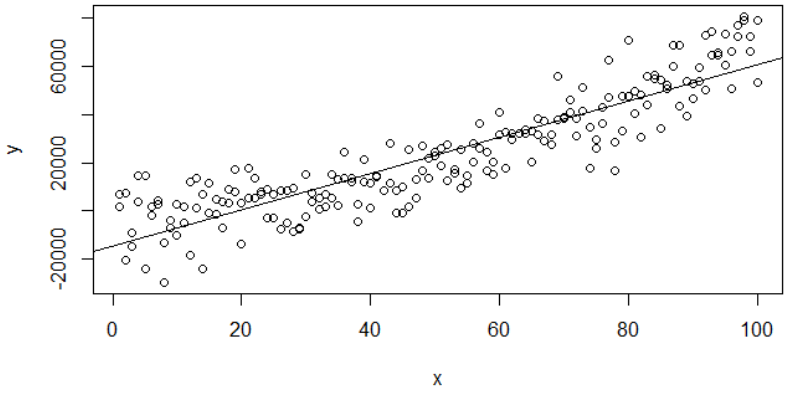
Kitu Komya (UID: 404-491-375)
Statistics 101A (Discussion 3A)
Homework 5 (due: 10/28/16)

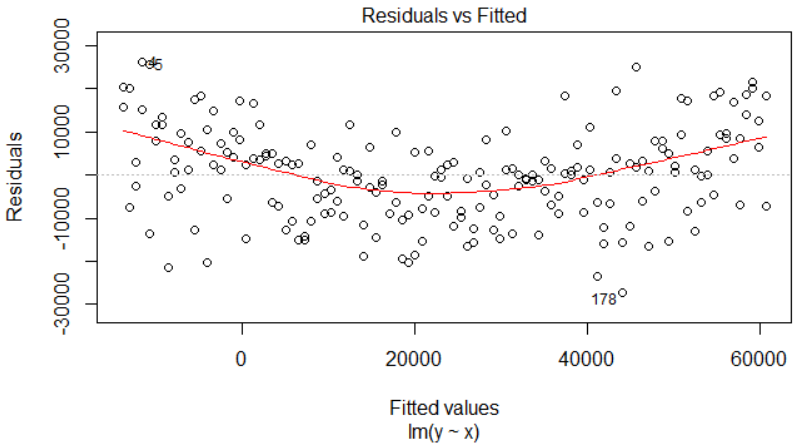
Please note: I'm concurrently taking Stats 20, so I'm still familiarizing myself with R.

1a. No, the linear model is not valid on this data. The equation has the quadratic term $7x^2$ in it, which is clearly evident of a polynomial (specifically, quadratic) model, not a linear model.

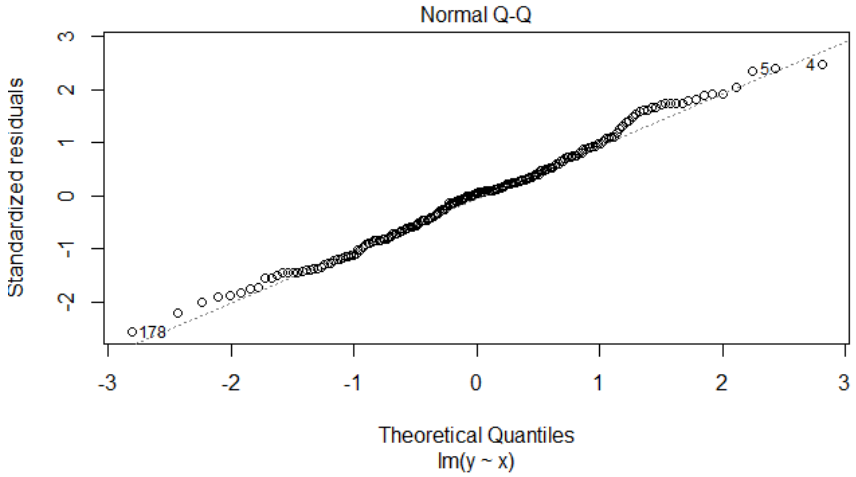
1b.



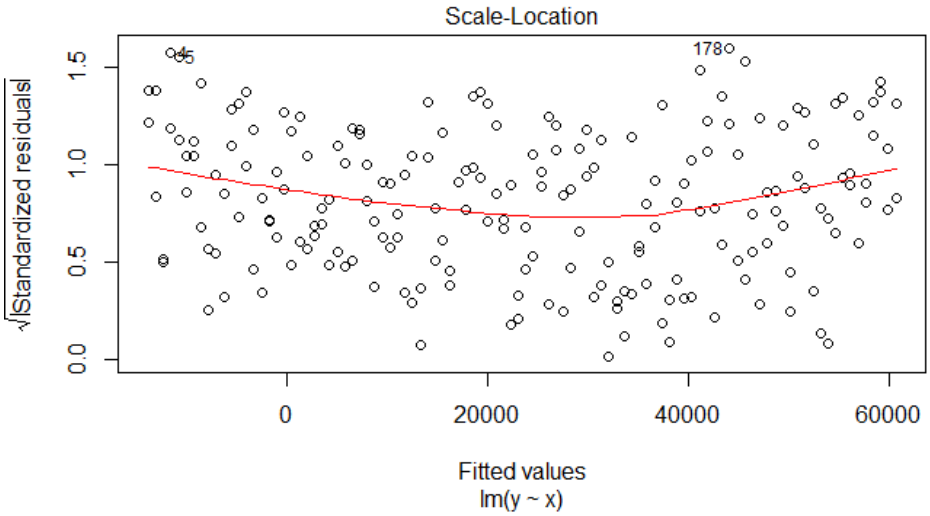
Scatterplot: on first glance, although it seems like the least squares linear regression model trendline fits the data well, upon a closer well we see that the line overestimates at the extrema and underestimates near the middle, clearly indicative of a quadratic function, and thus in violation of the linearity assumption in using a linear model.



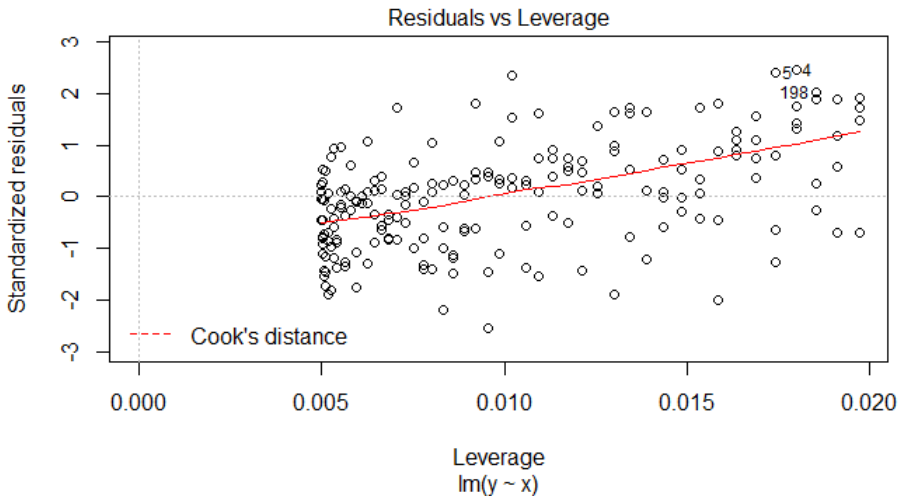
Residual plot: there is a clear parabolic trend in this plot. This is clearly indication of violating the linearity assumption. The red smoother further shows that at the extrema the line overestimates the data and underestimates it near the middle, which also violates the constant variance of the errors assumption since the variance follows a parabolic trend as well. We should have seen a randomly scattered residual plot with no pattern and a flat red smoother to conclude validity of the linear model.



QQ plot: this is actually a good QQ plot since for the most part the data points lie on the line. There is some deviation at the extrema, but they are not too drastic. This means that the errors are normally distributed. This makes sense because upon seeing the R code, we see that the standard error, e, was calculated using a random, normal distribution centered at 0. Therefore, the linear model does not violate the assumption for normality of the errors.

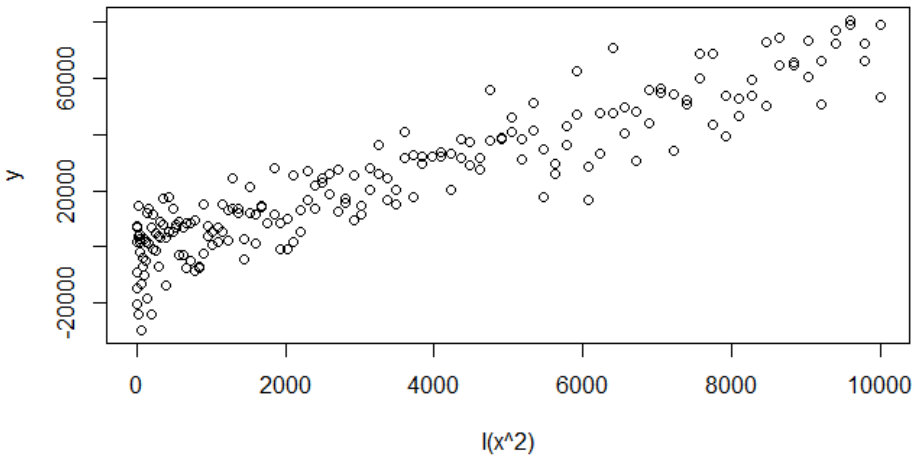


Scale-Location plot: similar to the residual plot, this plot too should not show any trend. However, again, similarly to the residual plot, there is a clear parabolic trend. This means that there is non-constant variance in the errors, which further confirms from the residual plot that the linear model violates the constant variance in the errors assumption.



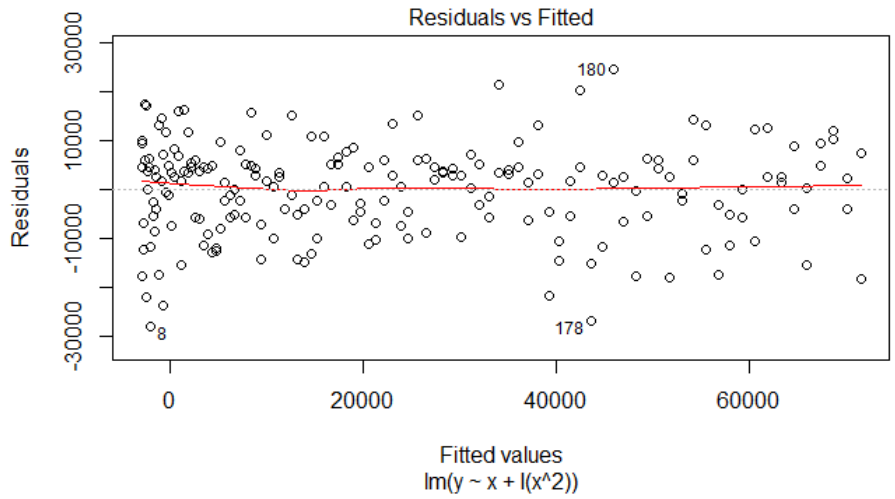
Residuals vs Leverage plot: this plot numbers two points that are most likely high leverage points. They are on the border of falling within the [-2, 2] interval of standardized residuals, so perhaps 198 is a good leverage point and 504 a bad leverage point. Regardless, two high leverage points exist, which is not desirable, yet still manageable to deal with. Moreover, since no dashed contour lines are visible using Cook's distance, none of the points are influential, which is good.

1c.

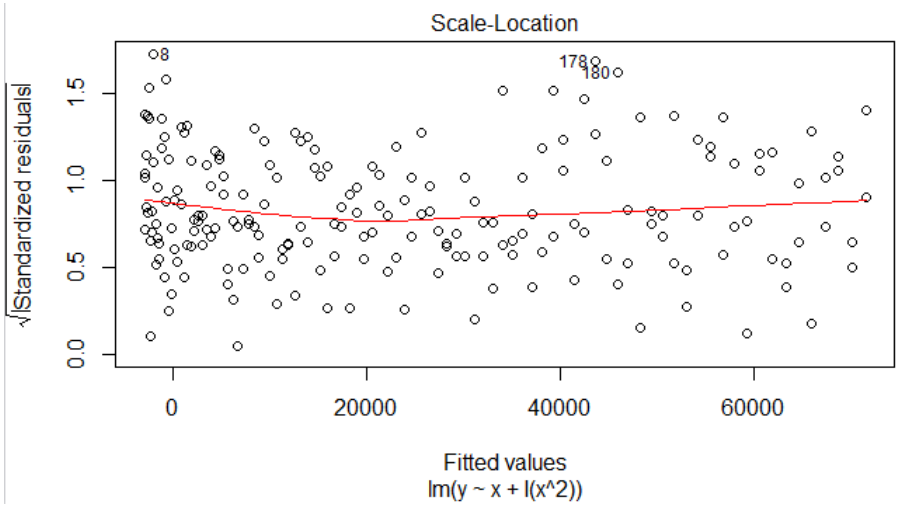


I have fit the model in R. Since I do not know how to add a quadratic transformation model's trendline, above I have only reproduced the scatterplot without any model fitting the data. Just by visual inspection it is clear that a transformed linear model will now better fit the data than a non-transformed linear model fitting the data without the quadratic transformation (as seen in 1b).

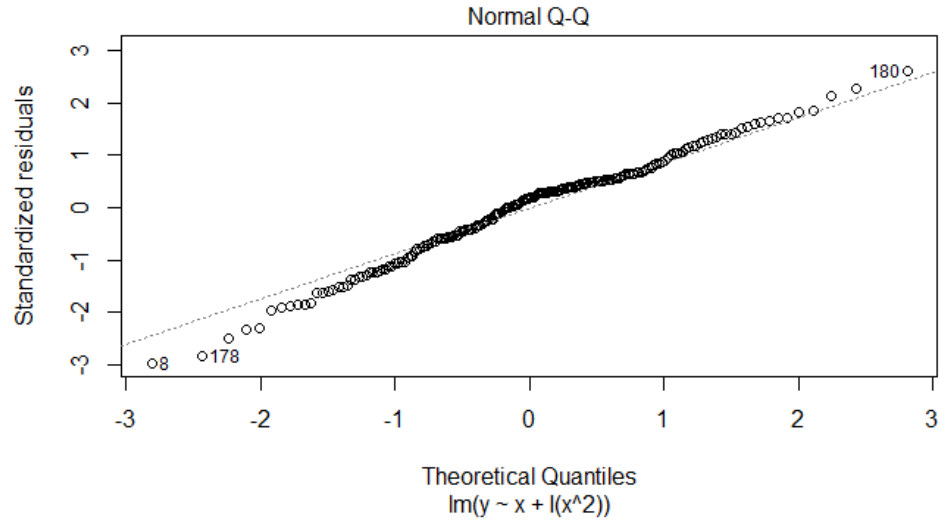
1d.



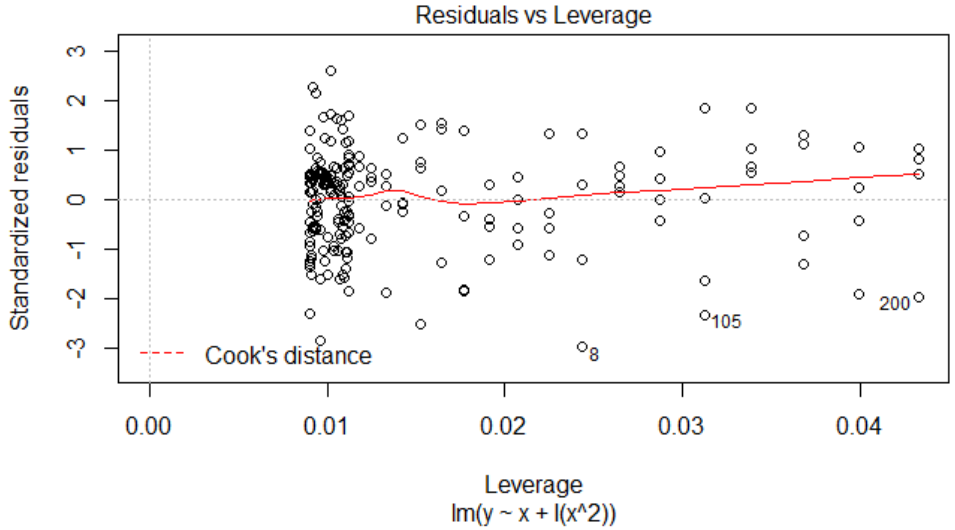
Residual plot: wow! The quadratic transformation sure did wonders to our residual plot! Now we see the red smoother basically flat, with the residuals showing no trend in its scatter. This transformation already seems like a better fit because the residual plot shows that the linearity and constant variance assumptions have been met.



Scale-Location plot: similar to the residual plot, this plot too should not show any trend. Furthermore, again, similarly to the residual plot, there is no major trend. although there is a slight movement of the red smoother, it is drastically better than the smoother from 1b because this line is nearly flat. This means that there is a constant variance in the errors, which further confirms from the residual plot that this transformed model follows the constant variance in the errors assumption.



QQ plot: this is a decent QQ plot. The majority of the data points lie on the straight line, with some deviation at the extrema. Interestingly, the deviation at the extrema is a bit larger than the deviation at the extrema of the non-transformed linear model. However, the difference is small enough to neglect any comment on the difference in normality of the errors between the two models. And thus, this plot can conclude normality of the errors, which is also evident by the R code since the standard error, e, was calculated using a random, normal distribution centered at 0, just like previously. Hence the transformation has no effect on the normality of the errors assumption, which was previously accepted as well.



Residuals vs Leverage plot: this plot numbers three data points that are possibly high leverage points. Two data points seem to be at the border of the [2, 2] interval of the standardized residual, while 1 is clearly outside of the interval. Interestingly, that is one more leverage point than the non-transformed linear model. However, a difference of one leverage point hardly makes a difference. Moreover, no contour lines calculated from Cook's distance are present meaning that in this transformation there are also no influential points.

Overall, we see that the quadratic polynomial produces a valid model. As evident from the residual plot, the assumption of linearity is met. The residual plot along with the scale-location plot also prove constant variance of the errors. These two assumptions (linearity and constant variance of the errors) were previously violated in the non-transformed linear model. Hence, the transformed quadratic model is already a better fit. The quadratic model, similar to the non-transformed linear model, also had a few leverage points, but they can easily be handled. Moreover, both models also follow a normal distribution of the errors, as seen from the QQ plot. Taking the independence of errors assumption on faith, we can conclude that the transformed quadratic model has met all four assumptions for it to be a better and more valid model for its data.

1e. Using the summary function in R, we see the following outputs for p-values for

$\beta_0 = 0.152$

$\beta_1 = 0.438$

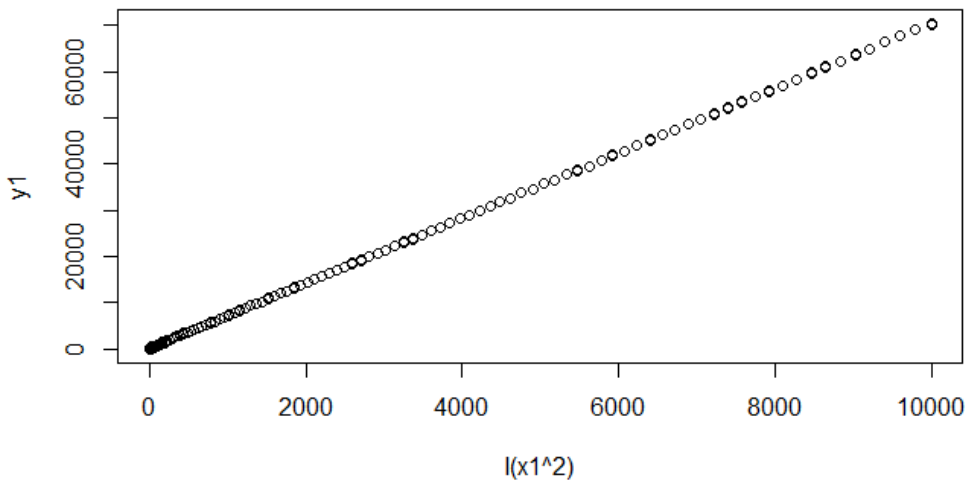
$\beta_2 = 2.87 * 10^{-12}$

At a significance level of 0.05, we see that we fail to reject the claim that the intercept (β_0) along with x (β_1) are 0. However, we can reject that x^2 (β_2) is 0, indicative of an association between x^2 and y, which matches our theory because we have fit a quadratic model. Thus, clearly the β_2 term belongs in the model, while the other two cannot be confirmed to belong.

Using the summary statistics, the equation of the estimated model is

$$\hat{E} = (Y|x) = -2964.1185 + 73.1069x + 6.7241x^2$$

1f.



I have fit the model in R. Just by visual inspection it is clear that a transformed linear model will fit the data very, very well! Because the standard error is much smaller than the model in 1c, the SYY (variability in y) is small, and thus the RSS will be small and the SS_{reg} large.

Using the summary function in R, we see the following outputs for p-values for

$\beta_0 = 0.166$

$\beta_1 = 1.35 * 10^{-6}$

$\beta_2 = < 2 * 10^{-16}$

At a significance level of 0.05, we see that we fail to reject the claim that the intercept (β_0) is 0. However, we can reject that x (β_1) along with x^2 (β_2) are 0, indicative of an association between x and y as well as x^2 and y. This matches our theory because for β_1 we see a very strong, positive linear association between x^2 and y because of the small standard error, and for β_2 we have fit a quadratic model. Had we been given an r-squared value, we would have seen a higher value in this model than the model in 1f, so that makes it fitting that the β_1 term has significance in terms of association with the model. Thus, clearly the β_1 and β_2 terms belongs in the model, while β_0 cannot be confirmed to belong.