Things people have learned so far in 617 – responses to Quiz 2 question

Fall 2022

From the course I understand when to transform the X and Y, and the reason why they need to be transformed.

review how to use the R to compute the linear model

* use normal Q-Q plot to check whether variables are linearly distributed;
* get a brief understanding about the IMRaD writing style and learn what does each part should contain and check whether the article fits this writing style;
* use diagnostic plots to check whether the assumption model fits the original data;
* SS decompositions and F statistics;
* hat matrix and the proof for SST, SSreg, RSS;
* use R to summarize the statistical data for variables and draw diagnostic plot

I didn't know about the usage of a log transform to interpret percentage changes until today. I could have used this in my last job!

I have developed a much better understanding of case diagnostic plots. Before this class I was really only comfortable interpreting the residuals vs fitted and the QQ plot. However, now I feel like I understand what to look for in the other two plots as well and the impact they can have on the strength of the fit of the model. I also think my interpretations for interactions are much stronger now.

I've learned to interpret residual plots and to prove R^2 = correlation^2 in terms of matrix. The lectures enhanced my understanding and interpretation of multiple linear regression.

I have learned some of the basic theory for the matrix form of multiple regression. Before this course I was unaware of how much linear algebra was involved in regression. In hindsight it makes a lot of sense that matrices play a heavy role.

Cooks distance and the residual-leverage plot, it was always one of those plots I saw but never understood, however now I know its purpose.

I've worked with regression tools in R for a fair bit but never really got around to fully understanding the information that diagnostic plots offered. Once I actually understood that through 36-617, I was quite surprised by how powerful and informative those four plots are -- they provide a plethora of information about the data and model fit at a single glace. Moreover, they serve as more reliable sources to assess the model fit/performance than some other indicators.

I've learned that the mean of residuals from a linear regression model is zero and this does not depend on the normality assumption. I always understood that we want our residuals to have a center of zero but I did not realize that the mean error (sum of all the errors) will always equal zero.

Tools that help check if a model if good: R-square, F-statistic, diagnostic plots (Residuals vs Fitted values, Residuals vs Leverage, Normal QQ plot, etc.)

It's the first time for me to systematically learn the three reasons for transformation and the interpretations for them.

So far in this course I have learned a lot more about leverage and what it means. I previously took an Experimental Design stats course that mentioned leverage briefly but I had no idea what Cook's distance really meant or what having a high leverage and high residual signified about the data until completing the first homework in this course. I found it interesting that the variance across values in a data set can vary so much and have such an impact on how the data should be treated in order to accommodate it. I now know that leverage is measure of distance between xi and the mean of x, and that to have an effect, the residual e-hat must also be large. This means we only look for points in the upper and lower right side of the residuals vs leverage plot. (In my previous course I would look for points all over the graph that were past Cook's distance because I didn't understand where they were expected to be located).

I've learnt to interpret the Residuals vs Fitted, Normal Q-Q, Scale-Location and Residuals vs Leverage plot. And I have not learnt about leverage before during my undergraduate years.

There are many new things that I've learned in this course, such as matrix form of regression. and how to use R to draw diagnostic plots.

(1) Use diagnostic plot to find whether the model fits well. Use plots to find whether there are outliers with unusual large y or high leverages with unusual large x.
(2) Judge What variable need to be transformed and how to transform variable to make it more normal-like by analyzing the Normal QQ plot and find whether there is a long(short) left(right) tail.

(3) Use SS decompositions and F statistics to select suitable model.
(4) Write paper in IMRD method

Something new I've learned in this course so far is the nuances of the relationship between residuals and leverage. I previously thought that high leverage values were almost always concerning, but I didn't think about the cases where they could help the model given that the residual magnitude isn't too large.

In this course, I have learned about the leverage and residual plot, and the meaning of low/high leverage and low/high residual in comparison to Cooks distance. In prior classes, I had not covered leverage in my diagnostic plots, so learning another element to be aware of for observations that may have high influence on the regression and data is important.

Implementing Box-Cox for X in R and the method of computing optimal lambda using the Box-Cox likelihood.

First, I learned how to use the matrix to transform mathematical formulas and borrow matrix properties to derive expression formulas; second, I learned how to use diagnostic plots to determine the linear regression fit and to determine who has a better fit by looking at R^2; finally, I learned how to perform basic linear regression fitting with R and output summary and plot to make judgments about the fit.

I have learned a lot about matrix identities and operations, especially with hat matrix and derivation of MLR. Besides that, I learned to use the heuristic approach to find variables needed for transformations in a dataset, rather than brute-forcing the problem.

I've learned how to determine the fit of a linear regression model using summary statistics and diagnostic plots. I've also learned how to use R to conduct data analysis.

Something new I learned so far in this course is the hierarchy principle. Previous to this course I did not know that if you include a k-way interaction term in your model, then all lower order interactions should be included in the model. I assumed you would only include them if their significant. It was interesting to see in class that this principle holds true through examples and plots.

I learned simple and multiple linear regression models, and how to express them in their matrix forms. Also, I learned how to interpret the diagnostic plots of these models and how to transform variables.

I have never heard about hat matrix before, and I learnt the definition and the property of hat matrix in this course.

I learned why to do the transformation for data and the reason for log transformation explained in the change of E(y) and x.

* hat matrix and SS decomposition.
* Methods to make transformation: like box-cox, inverse response plot.
* Do not delete any data before I figure out the reason. I thought those data with high leverage and pointed as outliers should be deleted, however, now I understand it's not good to delete any data easily.
* In Monday's class, I learned how to interpret qq-plot's tails, and observe it with a histogram.

Something new that I have learned new in this course has to be the Box-Cox transformations and how they can be used to improve the distribution to maintain linearity.

I have actually never seen a scale-location plot before! It is completely new to me.

Box-cox transformation for finding the exact but not intuitive power for variable transformation!

I would say the transformation of the matrix. Before this course, I only know to use the definition but not very clear about the whole transformation process and the property of the hat matrix.

Taking this course, I realized how important the diagnostic plot, scatter plot, and histogram are in regression. Looking at the test statistic and r^2 is not enough to judge the model.

Do different transformations on x or y so that they are more normally distributed. We can use boxCox() or powerTransform() to compute estimated transformation parameter.

Through the lecture and homework, I was able to understand how statistical equations are actually applied to the real-world problem analysis. Furthermore, I learned new statistical tools to assess the validity of the statistical models. For instance, I did not know how to utilize or interpret such diagnostic graphs such as Residual vs. Fitted, Normal Q-Q, Scale-Location, and Residuals vs Leverage plot. I am excited to learn more of these tools later in the semester to enrich my ability to interpret the data set.

Something that I have learned in this course is that I need to be more efficient when studying technical/mathematical topics. Working away till the last minute is not helpful and I am inefficient in time managing.

The hat matrix H_1 is new to me, so expressing SS decomposing in matrix form is also new to me. Furthermore, the Variance-stabilizing Transformations is also new to me.

I did not know what high leverage points are. I have seen the plot of Leverage vs. Standardized Residuals in other statistics classes before when inspecting the fit for linear models, but I never knew what Cook's Distance meant.