STAT 4360 Mini Project 2

Name: Jaemin Lee

Section 1: Answers to the specific question asked

1. Given in the code.
2. After applying the log transformation to PSA, the graph looked a lot better compared to the graph of the original PSA against other variables. I plotted PSA against Cancer Volume to see if transformation was needed. Below represents a side-by-side plot of psa vs cancervol and psa.log vs cancervol.



I wanted to check if the graph could look better after some kind of transformation on cancer volume. I decided to apply sqrt on cancer volume. Below is the side-by-side graph of PSA.log vs Cancer Volume PSA.log vs CancerVolume.sqrt. The correlation between the two on the left is clearer on the left plot.



1. My fit1 is the linear model for psa.log against cancervol.sqrt. Its slope is 0.6366 which indicates that there is a correlation. The p-value for cancervol.sqrt is 2.2e^-16 (less than 0.05) which is statistically significant. The t-value for it is 10 (greater than 2) which is significant. Therefore, cancervol.sqrt is a very useful variable. Fit2 is the linear model for psa.log against weight. I applied sqrt to weight to see if the result is better. The slope for psa.log vs weight is 0.1395, meaning there is a slight correlation. The p-value for weight.sqrt is 0.01(less than 0.05), which is statistically significant. The t-value is 2.49 (greater than 2) which is significant. Thus, weight can be a useful predictor. Fit3 is a linear model for psa.log vs age. The slope is 0.026 which represents a weak correlation. The p-value for age is 0.096 which is greater than 0.05, so it is not significant. The t-value is 1.68 which is less than 2, so it is not significant. Thus, age is not a significant predictor. Fit4 is a linear model for psa.log vs benpros. The slope is 0.0599 which represents a weak correlation. The p-value for benpros is 0.124 which is greater than 0.05, so it’s not significant. The t-value is 1.55 which is less than 2, so it is not significant. Fit5 is a boxplot of psa.log vs vesinv.factor. The p-value for vesinv is 1.48e^-9 which is statistically significant. The t-value is 6.698 which is significant. Thus, vesinv is a significant predictor. Fit6 is a linear model for psa.log vs capspen.sqrt. I applied sqrt to capspen for the better result. The slope is 0.55, meaning there is some correlation. The p-value for capspen is 5.29e^-9 which is statistically significant. The t-value is 6.4 which is significant. Thus, capspen is a significant predictor. Fit7 is a linear model for psa.log vs gleason. Its slope is 0.84 which represents a high correlation. The p-value for gleason is 1.23^-18 which is statistically significant. The t-value is 6.2 which is significant. Thus, gleason is a significant predictor. In conclusion, cancervol, weight, vesinv, capspen, and gleason turned out to be the significant predictors in PSA level using linear regression.
2. Fit8 is a multiple regression model using all of the predictors. Cancervol, benpros, vesinv, gleason have p-values less than 0.05 in this model. When p-values are less than 0.05, then we can reject the null hypothesis. Thus, we can reject the null hypothesis for those 4 predictors above. The only difference between simple linear model and multiple regression model is that weight is not significant anymore for multiple regression. The R^2 for this model is 0.557. We can use this to compare with my final model.
3. We want to look at R^2 to check if the model is reasonably good or not. If R^2 is close to 1, then the model is reasonable. If the model is close to 0, then the model is not reasonable. I decided to apply interactions between predictors to fit a reasonably good model. Potential candidates for those predictors are cancervol, vesinv, capspen, and gleason. First, I applied interactions between cancervol and other variables. They are fit9, fit10, fit11, fit12, and fit13. However, all of them have R^2 around 0.5. So, I decided to pick the model that had the highest R^2 for the interactions and add other predictors to get better R^2. For this model (fit14), I applied interactions between cancervol and benpros, because it had the highest adjusted R^2(= 0.54). Fit14’s adjusted R^2 = 0.604 which is better than my previous models. Compared to the multiple regression model with all predictors, R^2 went up from 0.557 to 0.604. Thus, fit14 is my final model as it has the highest reliability out of all my models.
4. psa.log = -1.23129 + (0.52993 \* cancervol.sqrt) + (0.16209 \* benpros) + (0.69893 \* vesinv.factor1)

+ (-0.09462 \* capspen.sqrt) + (0.32221 \* gleason) + (-0.03576 \* cancervol.sqrt:benpros) with vesinv = 1, if seminal vesicle is present

vesinv = 0, if seminal vesicle is absent

1. There are two different ways to check if the model is doing a good job on predicting.

The first one is to find the mean of necessary variables and plug them into the equation in (f).

-1.23129 + (0.52993 \* 2.307854) + (0.16209 \* 2.534725) + (0.69893 \* 0) + (-0.09462 \* 0.9627874) + (0.32221 \* 7) + (-0.03576 \* 5.584535) = 2.367. Thus, PSA level = 2.367233.

The second one is to use the predict() function and let R do the job for us. After using the predict function, the result turned out to be 2.357729.

The difference between the first method and the second method is 0.01. My assumption on this discrepancy is due to the interaction terms in fit14. Either my way of finding the mean of interaction (mean(cancervol.sqrt\*prostate$benpros) is not accurate or predict() function doesn't include interactions as parameters. However, my model is doing a good job on predicting. Also, psa level around below 2.5 is safe. Thus, this patient shouldn’t have to worry about getting prostate cancer.

Section 2: R code











