Mini-Project duo group 7

Venkata Bapanapalli

Karthik Meyyappan

Contributions: Both members equally contributed to analytically solve and implement the code of the given two questions.

Chart, box and whisker chart

Description automatically generatedA picture containing diagram

Description automatically generated Original Natural Log

The natural log of the PSA values will be used as they have less outliers than the normal distribution of the PSA.

Cancervol, Weight, Benign Prostatic Hyperplasia, and Capsular penetration are quantitative values and the Seminal vesicle invasion, and Gleason score are qualitative values. The models will be generated using the quantitative variables.

Diagram

Description automatically generated

The biggest factors as can be seen from the plots above are Cancer Volume, Benign Prostatic Hyperplasia, and Capsular Penetration.

first\_fit = lm(psa.log ~ cancervol + weight + age + benpros + capspen)

second\_fit = lm(psa.log ~ cancervol + benpros + capspen)

third\_fit = lm(psa.log ~ cancervol + benpros)

Weight and age are small, so they are set to zero since we accept the null hypothesis. Since the p-value is large at 0.2561 for third\_fit we accept the null hypothesis for capspen.

Chart, scatter chart

Description automatically generatedAfter removing the parameters that have the least influence, we can choose model the model with cancervol and benpros as the best model as the other quantitative variables are either very small or have high p-values. Then we factor in the qualitative variables vesinv and gleason.

The addition of vesinv and gleason prove to be important to the model.

The linear model using the remaining parameters is:

Chart, line chart, scatter chart

Description automatically generated

From this we can conclude a strong linear relationship with the other predictors. A transformation of response was necessary for cancer volume, weigh, and seminal vesicle invasion due to a lack of fit.

The predicted PSA level for a patient is 10.176.