Chris Cole

Project 3

**Descriptive Statistics**

A picture containing text, indoor

Description automatically generated

The graphs of blood clotting and liver function are both skewed left. Both graphs have 1 or 2 outliers that stretch the data out to the left. Prognostic index and enzyme functions are right skewed and also have 1-2 outliers to the right. For enzyme function and liver function the mean is slightly lower and higher respectively relative to the median. The age graph is relatively symmetrical and is the only one that has no outliers in the data.

The y graph of the distribution is skewed heavily left due to the mean being higher than the median and appears to have 5 outliers to the left of the boxplot. The data has a median of about 605 survival days but has a mean of 702 survival days. The addition of the log transformed graph helps with the skewness of the graph. The log transformed graph is more symmetrical around the mean. Since the graph is log transformed, the data on this graph has been exponentiated. The median of the natural log graph is simply the natural log of the median of the y graph. The median and mean of the data are much closer together in the log transformed graph than the y graph, but there is still a slightly larger mean than median. The log transformed graph has 5 total outliers on both sides of the boxplot. Because log transformed graph is more symmetric, I would prefer to do the multiple linear regression on it.

A picture containing table

Description automatically generated

I used multivariate correlations and the scatterplot matrix to get a better view of how each of the variables correlated with one another on the model. Blood clotting, the prognostic index, enzyme function, and liver function all show a positive correlation with log transformed survival time in the liver operation. The survival time is likely to increase as each of these variables increase. Based on the scatterplots, age does not seem to have a strong correlation with survival time, but it is slightly negative as seen by the correlation matrix.

The correlation matrix above the scatterplot matrix shows the correlation between the coefficients and log transformed survival time. Based on the values of .653 and .640, enzyme function and liver function have some of the strongest correlations with survival time. Age has the weakest correlations with -.1451. Because it is a negative value, this means that as age increases survival time decreases. We are also able to see the linearity between the variables. For the most part there is very little correlation between the coefficients with the exception of liver function score. Liver function score has the highest correlation values with the other variables in the model.

Chart, box and whisker chart

Description automatically generated

Since gender is nominal in this model, it cannot be analyzed using the scatterplot matrix. I used a one-way analysis of the log-transformed survival time by gender to display the data. According to the graph, males have a slightly above average survival time and a skewed right distribution. Females have a median below the average and a skewed left distribution. There is one outlier outside of the female boxplot that is significantly higher than the rest of the data. In general, this data seems to suggest that males have a greater mean survival time when undergoing the liver operation. The variability in the survival times for females is higher than that of males.

**Regression analysis**

A picture containing table

Description automatically generated

A.

ln(y) = β0 + β1 Blood Clotting + β2 PI + β3 Enzyme Function Test + β4 Liver Function test + β5Age + β6 Gender + ε.

Due to the model being log transformed, the coefficient can give you the percent increase or decrease in the response for every one-unit increase only if the other values remain constant. For each of these we will assume that the other variables are held constant. For every 1 unit increase in blood-clotting score survival time increases by 9.95%. For unit increase in prognostic index, survival time increases by 1.31%. For every unit increase in enzyme function test, survival time increases by 1.64%. For every unit increase in the liver function test, survival time decreases by about .312%. Every unit increase in age shows a .485% decrease in survival time. Finally, holding all other variables constant, the estimated survival time for females is approximately 3.32% lower than that for males.

Chart, box and whisker chart

Description automatically generated

B.

The box plot of residual helps to ensure that the normality assumption is not being violated in our model. The box plot of residuals is fairly symmetrical around the mean eliminating any fears of issues with normality. There are no outliers in the data that are easily visible in the data, however some of the z-scores falls outside of 2 standard deviations (95%) of the observations on the positive end. On the negative end, the minimum z-score falls well within the 2 standard deviations. This could mean that there are outliers in the data.

C.

The residual plots help to affirm the assumptions that are being made of our model. The residual by predicted plot at the beginning of our regression analysis checks the homogeneous assumption. This checks if the residuals have equal variance. The random scatter of points means that there is no issues with this assumption.

Residuals vs Predicted values

Chart, scatter chart

Description automatically generated

To determine whether the model has been specified correctly, we would expect to see a random scatter of points on each of these plots. The normal probability plot below also validates that the residuals are normally distributed. All of the points fall within the confidence bands. There is no evidence of non-linearity in the model.

Chart, histogram

Description automatically generated

D.

The variance inflation factors test if there is multiple collinearities between any of the coefficients. Blood clotting had a value of 1.8951967, Prognostic Index had a value of 1.3008361, enzyme function was 1.7315693, 2.9992787 for liver function test, 1.0966944 for age, and 1.1267306 for gender. Since these are below the threshold value of 10, the independent variables are not linearly related with one another. Multicollinearity is not an issue for the model.

**Hypothesis Testing, Confidence Intervals, and Model Performance**

H0: x1=x2=x3=x4=x5=x6=0

H1: At least one is not equal to 0

The null hypothesis will test that the model is not significant, and the alternative hypothesis will test that it is significant. We have a f-value of 26.6915 and a significant p-value that is less than .0001. This p-value falls well within the error rate. This means we reject the null hypothesis and conclude that our model relating survival time to blood clotting, the prognostic index, enzyme function, liver function, age and gender is highly significant.

Text, table

Description automatically generated

Individual Tests

Blood Clotting Score

The null hypothesis is x1=0 vs alternative hypothesis of x1 not 0. If the error rate is 5% and our t-value is 3.23, we are given a p-value of .0022. This is less than the error rate of 5%. As a result we would reject the null hypothesis and determine that there is linearity between the blood clotting score and survival time.

Prognostic Index

The null hypothesis is x2=0 vs alternative hypothesis of x2 not 0. With an error rate of 5% we are given a t-value of 5.65 and a p-value that is highly significant (<.001). Because of this, we will reject the null hypothesis and determine that there is evidence of linearity between the Prognostic Index and survival time.

Enzyme Function Test Score

The null hypothesis is x3=0 vs alternative hypothesis of x3 not 0. With an error rate of 5% we are given a t-value of 7.68 and a p-value that is highly significant (<.001). Because of this, we will reject the null hypothesis and conclude with sufficient evidence that enzyme function has linearity with survival time.

Liver Function Test Score

The null hypothesis is x1=4 vs alternative hypothesis of x4 not 0. This coefficient has a t-statistic of -0.06. The p value of .9550 falls outside of our 5% error rate. So, we would fail to reject the null hypothesis and say we have insufficient evidence to conclude that there is linearity between Liver function test scores and survival time in days.

Age

The null hypothesis is x5=0 vs alternative hypothesis of x5 not 0. Age has a t-statistic of -1.51 and a p-value of .1371. Because the p-value falls outside of our 5% error rate, we fail to have sufficient evidence to reject the null hypothesis. We are unable to say with certainty that age is linearly correlated to survival time.

Indicator variable for gender

The null hypothesis is x6=0 vs alternative hypothesis of x6 not 0. Gender produces a t statistic of -.92 and a p value of .3631. Given an error rate of 5%, the p-value is to large to conclude linearity. Therefore we would fail to reject the null hypothesis and we are unable to conclude that the absence or presence of the gender variable (or whether the patient is a male or female) has any linearity with survival time.

The 90% confidence intervals mean that we are 90% confident that the regression coefficient is between the upper and lower value. For blood clotting we are 90% certain that the coefficient is between .0456 and .144. For Prognostic Index we are 90% certain that the regression coefficient is between .009 and .016. We are 90% certain that the coefficient for enzyme function is between .012 and .019. We have the same level of certainty that liver function is between -.0958 and .0895. We are also 90% certain that the age coefficient falls between -.010 and .0005.

**Mean Intervals, and Prediction Intervals**

The estimate from the model estimates that if you have a 30-year-old female with a blood clotting score of 3, a PI of 40, an enzyme function score of 50, and a liver function score of 30 we are 90% confident that the patient would live between 202.33 days and 336.72 days. This is found by plugging the data into the model and exponentiating the original responses of the confidence intervals.

The estimate from the model predicts that if you have an individual that is a 30-year-old female with a blood clotting score of 3, a PI of 40, an enzyme function score of 50, and a liver function score of 30 we are 90% confident that the patient would survive between 160.102 and 425.53 days.