Linear Regression Analyses of HELP Data

Introduction

The Health Evaluation and Linkage to Primary Care (HELP) study investigated the effectiveness of intervention treatments that were designed to connect people undergoing detoxification to primary healthcare system, as those who are addicts tend to not seek regular medical care and are often hard to reach. Adult patients were recruited from a drug and alcohol detoxification clinic at Boston Medical Center and the clinical trial was funded by the National Institute of Health. Analysis of the data collected before intervention treatments was performed using RStudio to assess the association between the number of drinks and a participant's age.

Exploratory Analysis

Baseline information of 465 adult patients was collected, which included their sex, age, and reported average number of drinks per day over the past 30 days. Exploratory analysis showed no missing values in the dataset. 23.66% of the participants were female.

The ages ranged from 18-60, with a mean of 35.77, median of 35, first quartile of 30, third quartile of 41, and standard deviation of 7.81. Visual exploration of the age variable showed it to be approximately normally distributed.

The number of drinks ranged from 0-142, with a mean of 18.42 median of 13, first quartile of 3, third quartile of 26, and standard deviation of 20.17. 67 subjects reported no drinking over the past 30 days. When viewing a histogram of the number of reported average daily drinks, it was discovered that the data was right skewed, and therefore not normally distributed.

Linear Regression

A simple linear regression was performed on the number of drinks and a participant's age. The t-score was 4.341 with 463 degrees of freedom and the resulting p-value was very close to 0. With a p-value less than the $\alpha = 0.05$ significance level, the null hypothesis that there is no correlation/linear association between the two variables was rejected. There was evidence suggesting that there is a positive linear association between the average number of drinks consumed in a day and a participant's age with an estimated slope of 0.5111 and y-intercept of 0.1379.

$$drinks = 0.5111(age) + 0.1379$$

The coefficient of determination was 0.0391, meaning that a participant's age accounted for 3.91% of the variability in the number of drinks consumed in this dataset. Using the equation for the regression line, the predicted number of drinks for a subject who is 50 years old is 25.6906 in a day. Residual plots of the predicted number of drinks were generated to assess the linearity and homoscedasticity assumptions. The residuals appeared to be randomly scattered around the residuals=0 horizontal line, which confirmed the linearity assumption. However, the

spread in residuals increased as the predicted number of drinks increased, which violated the homoscedasticity assumption.

Log-Linear Regression

Because the reported average number of drinks per day was not normally distributed, the variable was logarithmically transformed into a new variable, logdrinks. A correction value of 1 was added to the number of drinks before the logarithmic transformation to avoid undefined values for those who reported no drinking, as the natural log of 0 is negative infinity.

$$logdrinks = log(drinks + 1)$$

The number of logdrinks ranged from 0 to 4.963, with a mean of 2.322, median of 2.639, first quartile of 1.386, third quartile of 3.296, and standard deviation of 1.3050. A subject who reported no drinking would also have a value of 0 logdrinks because the natural log of 1 is 0. Visual exploration of the logdrinks variable showed it to be more normally distributed than the drinks variable.

A simple linear regression was performed on the number of logdrinks and a participant's age. The t-score was 4.514 with 463 degrees of freedom and the resulting p-value was very close to 0. With a p-value less than the $\alpha = 0.05$ significance level, the null hypothesis that there is no correlation/linear association between the two variables was rejected. There was evidence suggesting that there is a positive linear association between the number of logdrinks consumed in a day and a participant's age with an estimated slope of 0.0343 and y-intercept of 1.0940.

$$logdrinks = 0.0343(age) + 1.0940$$

The coefficient of determination was 0.0422, meaning that a participant's age accounted for 4.22% of the variability in the number of logdrinks consumed in this dataset. Using the equation for the regression line, the predicted number of logdrinks for a subject who is 50 years old is 2.8102 in a day. Undoing the logarithmic transformation, the predicted number of drinks for a subject who is 50 years old is 15.6127 in a day using the log-linear regression model. Residual plots of the predicted number of logdrinks were generated to assess the linearity and homoscedasticity assumptions. The residuals appeared to be randomly scattered around the residuals=0 horizontal line, which confirmed the linearity assumption. The spread in the log-linear residuals was not as extreme as the variance seen in the linear residuals, thus the homoscedasticity assumption was more supported in this model.

Comparing Linear Regression Models

A final scatterplot of the data was produced to visually compare the both the linear and log-linear regression model. The linear regression model gave larger predicted number of drinks, as seen previously when a 50-year old subject was predicted to consume 25.6906 drinks per day by the linear regression model and only 15.6127 drinks per day by the log-linear regression model. The logarithmic transformation standardized the outliers that used to be very influential

in the linear regression model, such as subjects who reported having over 100 drinks in a day. Those outlier data points were shrunk down in the logarithmic transformation, thus not having as much impact on the linear regression. Reversing the logarithmic transformation produced a curved exponential line of best fit in the log-linear regression model.

Conclusion

Linear and log-linear regression analyses of the correlation between the reported average number of daily drinks consumed and a participant's age both showed evidence supporting a positive linear association between the two variables. Further analysis of the data after various intervention methods were implemented can continue with that knowledge.