INFERENTIAL MODELLING – HDS 5320-02

ASSIGNMENT – 01

Vaishnavi Gannavaram

Instructor: Dr. Srikanth Mudigonda

Introduction

The dataset used in this assignment is "California Hospital Performance Ratings for Coronary Artery Bypass Graft (CABG) Surgery." It is present in the Comma-Separated Value file format and the source of this data set is https://data.ca.gov/dataset/california-hospital-performance-ratings-for-coronary-artery-bypass-graft-cabg-surgery .

This data has been named as "hsptldata" in the R file. "Readr" is the library that I have used to read the comma-separated value file. There are 10 variables in the data set. These Variables have been cleaned by dropping NA values using "na.omit" function and also have been assigned to new data set name called "hsptldata.cleaned." Out of those the variables that I have observed to be converted into Factor form using "as.factor" function are Performance Measure, Adverse Events, and Performance Rating.

Plotting Linear Regression models is the next step that I have taken to determine whether some of the variables are significant predictors for the outcome variable "*Risk-adjusted Rate"* and to show the amount of variance accounted by each variable.

Model 1:

This is the single-predictor model where "*# of Cases"* is the predictor and the '*Risk-adjusted Rate"* is the outcome variable. This step is done to determine whether "*# of Cases"* is a significant predictor of "*Risk-adjusted Rate"* or not.

In this model, the probability of observing an F-statistic value as 1.282 for the 1 predictor and the 422 observations by chance is 0.0002, Therefore, we can infer that model 1 is reliable. It means "*# of Cases"* can be a significant predictor for "*Risk-adjusted Rate".*

The adjusted R square value is 0.03679. The amount of variance shown by the predictor is shared by the combination is 3.69% variability. Here we can also infer that the other variables can also be better predictors than this one.

The p-value obtained for this model is 0.0002506 which is <0.05. Therefore, "# of cases" can be statistically significant predictor variable for 'Risk-adjusted rate.'

Model 2:

In addition to *# of Cases*, *County* is added to the model to test whether it is a significant predictor or not.

In this model, the probability of observing an F-statistic value as 1.341 for the predictor and the 451 observations by chance is 1.13e-05. Therefore, we can infer that model 2 is also reliable. It means "county" can be a significant predictor for "*Risk-adjusted Rate."*

The adjusted R square value is 0.047. The amount of variance shown by the predictor is shared by the combination is 4.7% variability. Here we can infer that this model is slightly more significant than the previous one.

The p-value obtained for this model is 1.13e-05 i.e.,<0.05. Therefore, "county" can be statistically significant predictor variable for 'Risk-adjusted rate.'

Model 3:

This is the three-predictor model where "# of Cases", "County", "Performance Measure" are the predictors, and the 'Risk-adjusted Rate" is the outcome variable. This step is done to determine whether "Performance Measure" is a significant predictor of "Risk-adjusted Rate" or not.

In this model, the probability of observing an F-statistic value as 5.084 for the 3 predictor and the 457 observations by fluke is <0.001, Therefore, we can infer that model 3 is reliable. It means "Performance Measure" can be a significant predictor for "Risk-adjusted Rate".

The adjusted R square value is 0.3742. The amount of variance change is 32.72 compared to the previous model. Here this model gives 37.42% variability. Here we can also infer that the other variables can also be better predictors than this one because there is 62.58% of variability left to predict.

The p-value obtained for this model is <2.2e-16 which is <0.05. Therefore, "Performance Measure" can be statistically significant predictor variable for 'Risk-adjusted rate.'

Model 4:

This is the four-predictor model where “# of Cases,” "County", "Performance Measure", "# of Adverse Events" are the predictors and the ‘Risk-adjusted Rate” is the outcome variable. This step is done to determine whether "# of Adverse Events" is a significant predictor of “Risk-adjusted Rate” or not.

In this model, the probability of observing an F-statistic value as 6.391 for the 4 predictor and the 458 observations by fluke is <0.001, Therefore, we can infer that model 4 is reliable. It means "# of Adverse Events" can be a significant predictor for “Risk-adjusted Rate”.

The adjusted R square value is 0.4417. The amount of variance change is 6.75 compared to the previous model. Here this model gives 44.17% variability. Here we can also infer that the other variables can also be better predictors than this one because there is 55.83% of variability left to predict.

The p-value obtained for this model is <2.2e-16 which is <0.05. Therefore, "# of Adverse Events" can be statistically significant predictor variable for ‘Risk-adjusted rate.

Anova test:

The Anova test is done to identify the best model in terms of statistical and practical significance among the four models.

Here, we can clearly see that the fourth model is the best fit among all the other models as it has 66337 - Residual sum of squares, and it also has the least p-value which is nearer to zero.

The second most significant model is the third model with 74377 – RSS and p value nearer to zero. The third and fourth best fits are model 2 and model 1 considering the RSS values.

Model 5:

The most significant and best fit model with F-statistic: 7.851 and p- value being less than < 2.2e-16 and Adjusted R-square value: 0.5046. with a variance of 50.46. Therefore, concluding this model as best fit among all the above four models.