J. Jedediah Smith

BIFX-551

Assignment 2 & 3

4/12/2022

1. Using the data provided in Bronchitis.csv generate a Linear Model, evaluate the model, and provide the evidence of the appropriateness of the model. The aim of the study was to define if the probability of having Bronchitis is influenced by smoking and/or pollution. The data are stored contains information on 212 participants.

The data showed that 46 of the 144 participants had Bronchitis. A binomial generalized linear model was created to compare the “cigs” and “poll” variables against “bron.” Model summary showed that both variables seemed to have a significant impact, but that Cigs had a slightly larger one. The cigs p-value was 2.83e-08, while the poll one was 0.007113. Further analysis reveals that the odds of getting Bronchitis are increased by a factor of 1.235761 per unit increase in cigs, and 1.140834 in poll. This supports what we saw from the model summary. Due to both variables being significant, we can conclude that the model is appropriate.

Text

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence



1. 25 students aspiring to enter to Computer Science have been evaluated on 5 different subjects: Differential Geometry (DifGeo), Complex Analysis (CompAn), Algebra (Alg), Real Analysis (RealAn) and Statistics (Stat). The objective of this study is to obtain a global ranking of students for entry into the Faculty of Mathematics, through a global score, extracted as a certain linear combination of grades in the five subjects tested. Use PCA to obtain that global score, sort the candidates and generate a formula to evaluate future candidates.

After data was imported, a scree plot was done to see how many components would be needed. It said that 1 component would be ideal. After running the PCA, we can see this was indeed the right conclusion, as all of the subjects have high correlation with PC1 and are all approximately 80% explained by the single component. Global scores were also calculated with the PCA, then displayed using “SubPC$scores.” With the PCA results, the following formula was calculated:

PC1 = 0.90\*DifGeo+0.88\*ComAn+0.93\*Alg+0.90\*RealAn+0.91\*Stat

Chart, line chart

Description automatically generated

Text

Description automatically generated with low confidence

Text

Description automatically generatedText

Description automatically generated

1. The following sample of a random vector (X, Y) has been obtained: (-3,7), (1,4), (0,6), (3,5), (1,9), (7,7), (6,9), (8,10), (7,12), (10,11). Perform a principal component analysis and a Regression on this sample. Represent a scatter plot, together with the first principal component and the regression line.

After the data was created, PCA was performed. The Scree Plot produced strange results and said that 0 components was ideal. This was the number of components used. It created RC1 and RC2, but the h2 values indicated that only RC1 effectively represented the data. Finally, a regression model was created. It yielded an insignificant p-values. This was subsequently represented on a scatter plot, with the model line going right through the middle of the randomly scattered set of points.

Chart, line chart

Description automatically generated

Text

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generatedText, letter

Description automatically generated