26 BE 7023 & 26 PH 7023: Advanced Biostatistics

Autumn 2018

MB Rao

Homework Sheet No. 4 Due Date: October 03, 2018 Maximum Points: 30

Model Selection

Download and activate the ‘car’ package. Download the data ‘Highway1.’ This data looks at highway fatalities in various segments of highways in Minnesota. The goal of the study is to identify risk factors for fatalities. This data has been analyzed in a Master’s thesis on Transportation Studies.

1. Determine the dimension of the data. Show the top ten rows of the data. Obtain summary statistics of the data. 3 points
2. Describe the data. 3 points
3. The response variable is ‘rate.’ In an analysis done at the University of Minnesota, log2(rate) is taken to be the response variable. The predictors are taken to be: log2(len); log2(ADT); log2(trks); log2(sigs1); slim; shld; lane; acpt; itg; lwid; hwy. The last predictor is categorical with four levels. You respect the transformations recommended. Fit a full regression model. Comment on the output including R2. Write the prediction model. Identify the significant predictors. Explain how the categorical variable ‘hwy’ is handled. Estimate the standard deviation σ of the error. 7 points
4. The model in Question 3 is too elaborate. We need a tight model. Employ the ‘backward elimination procedure’ to obtain a tight model according to the Akaike Information Criterion. Write the final prediction model. Compare its R2 with the R2 of the full model. Compare the estimates of the standard deviation of the error terms. 6 points
5. What is the total number of all possible regressions for the data on hand? Use the R function ‘regsubsets’ (package = “leaps”) on the highway data. Explain the output. 5 points
6. Apply the ‘forward selection procedure’ on the ‘highway1’ data. Write the final prediction model. Compare and contrast this model with the one in Question 4. 6 points