Homework 1

Due on March 7, 11:59pm

We will use the cleaned Toyota corolla pricing data set from the "Data cleaning" assignment (the one due on Feb 7) in this homework.

a) Conduct a principal component analysis (correlation-based PCA) on all binary variables that survive after cleaning. Please note that those binary variables also include dummy variables created from categorical ones. Don't forget to remove the reference variable from the set of dummy variables (usually the most common category).

b) How many principal components should we keep if we want to capture at least 80% of the total variance?

c) What is the difference between the principal component and principal component scores?

d) Which three variables contribute most to the first principal component? Try to give some interpretation of the first principal component if possible.

e) Consider the principal components retained in (b) as new variables. Construct a multiple linear regression model (potential predictors include those new variables from PCA analysis as well as all the continuous variables). The dependent variable is *Price*. More specifically, you need to do the following.

i) Partition the data into training (60%) and validation (40%) first.

ii) Try all four variable selection methods we have learned (forward, backward, stepwise, best subset).

f) Write down the linear relationship for each of the four selected models in (e). For example, $Price = beta_0 + beta_1 * x_1 + beta_2 * x_2 + \ldots + e$. Replace beta's by actual values.

g) What are RMSE's for each of the four models on the validation data? Which variable selection method gives you the smallest RMSE?

h) Check the three linear model assumptions for the model with the smallest RMSE on the validation data (linearity, normality, independency).

i) Are those assumptions satisfied? If not, what should you do?