MAT2017: Assignment 3 – 50 points

Due: By midnight on Wednesday, June 14, 2017

The 3rd project for MAT 2017 this semester is to conduct a linear regression analysis on a given dataset. This project is ***optional*** and the work should be done ***individually***. The successful completion of the project will contribute extra 5% toward to your final grad.

1. Overview

Regression is a mathematical form that represents the type of relation among variables. Regression models relate a response variable to one or several explanatory variables. The model is used to predict unknown values of the response variable. The performance of the model is measured by multiple metrics and one of the common metrics is the root mean squared error (RMSE). The objective is to find a linear regression model that minimizes the root mean squared error (RMSE).

2. Questions Studied

Each data in a given dataset consists of multiple explanatory variables, $X_1, X_2, \dots, X_n$, and one response variable Y. Students are free to choose one dataset among the following three datasets for their study.

(1) Franchise stores' annual net sales data

The data (X1, X2, X3, X4, X5, Y) are for each franchise store.
X1 = number sq. ft./1000
X2 = inventory/$1000
X3 = amount spent on advertising/$1000
X4 = size of sales district/1000 families
X5 = number of competing stores in district
Y = annual net sales/$1000

(2) Residents' health data

The data (X1, X2, X3, X4, Y) are by city.
X1 = doctor availability per 100,000 residents
X2 = hospital availability per 100,000 residents
X3 = annual per capita income in thousands of dollars
X4 = population density people per square mile
Y = death rate per 1000 residents

(Reference: *Life In America's Small Cities*, by G.S. Thomas)

(3) Basketball players data

The following data (X1, X2, X3, X4, Y) are for each player.
X1 = height in feet
X2 = weight in pounds
X3 = percent of successful field goals (out of 100 attempted)
X4 = percent of successful free throws (out of 100 attempted)
Y = average points scored per game

(Reference: *The official NBA basketball Encyclopedia*, Villard Books)

Once students decide the dataset for their analysis, they will define the followings before the analysis:

- The population
- Explanatory variables and a response variable
- The question: what you want to find from a regression analysis

3. Regression Analysis

First, students will need to conduct correlation analysis on each pair (Xi, Y) of one explanatory variable and the response variable. This includes draw a scatterplot of (Xi, Y) and finding correlation coefficient, and develop a linear regression model.

Second, students will perform a regression analysis using all explanatory variables and the response variable according to the following steps:

Step 1: Preparing X and y

Step 2: Splitting X and y into training and testing sets

Step 3: Develop a linear regression model

Step 4: Making predictions

Step 5: Computing the RMSE for the predictions

Step 6: Feature selection consideration

The following notebook example will guide you to conduct the analysis. This Python notebook program and a dataset file can be find in the project description file on the course Portal.

http://nbviewer.jupyter.org/github/sada-narayanappa/Notebooks/blob/STATS/linearRegression/regressionAnalysis.ipynb

Students are free to use and modify this program for their analysis. Students can also use any other statistical software, such as R, IPSS, MathLab for their study.

3 Final Report

After you perform the analysis, report the results in the final report. The report should include the following:

- Question studied - 5 points
  - The population, explanatory variables, response variables, and the question
- Correlation analysis on all pairs of the explanatory variables and the response variable - 15 pts
- A regression analysis - 30 points