Homework 3

IE 7275 Data Mining in Engineering

Readings and Practice:

1. Chapter 4: Multiple Linear Regression
2. Read the book chapters "Simple Linear Regression" and "Multiple Linear Regression" (attached to the assignment). Practice example problems given in the book chapters.
3. Read R Tutorial on "Regression". The data sets (women.R, mtcars.R and states.R) referenced in the tutorial are included in the assignment folder to allow you practice regression model building.

Problem 1 (25points): Gradient Descent Algorithm for Multiple Linear Regression

The file concrete.csv includes 1,030 types of concrete with numerical features indicating characteristics of the concrete. The variable "strength" is treated as the response variable.

- Standardize all variables (including the response variable "strength"). Split the data set into a training set (60%) and a validation set (40%).
- Implement the gradient descent algorithm in R with the ordinary least square cost function.
- Fit the multiple linear regression model using the gradient descent algorithm and the training set. Try out different learning rates: $\alpha = 0.01, 0.1, 0.3, 0.5$ and compare the speed of convergence by plotting the cost function. Determine the number of iterations needed for each α value.
- Apply the fitted regression model to the validation set and evaluate the model performance (ME, RMSE, MAE, MPE, MPAE). Calculate the correlation between the predicted strength and the actual strength. Create a lift chart to show model performance.

Problem 2 (25points): Multiple Linear Regression Model for Concrete Slump Test Data

- Read the included research article "Modeling Slump Flow Concrete". It is sufficient to consider "Slump Flow" as the response variable in this problem just as in the included article.
- Create a scatterplot matrix of "Concrete Slump Test Data" and select an initial set of predictor variables.
- Build a few potential regression models using "Concrete Slump Test Data"
- Perform regression diagnostics using both typical approach and enhanced approach
- Identify unusual observations and take corrective measures
- Select the best regression model
- Fine tune the selection of predictor variables

- Interpret the prediction results

Problem 3 (25points): The file insurance.csv includes 1,338 examples of beneficiaries currently enrolled in the insurance plan, with features indicating characteristics of the patient as well as the total medical expenses charged to the plan for the calendar year. The features are given as follows:

• *age*: An integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government).

• *sex*: The policy holder's gender, either male or female.

• *BMI*: The body mass index (BMI), which provides a sense of how over- or under-weight a person is relative to their height. BMI is equal to weight (in kilograms) divided by height (in meters) squared. An ideal BMI is within the range of 18.5 to 24.9.

• *children*: An integer indicating the number of children/dependents covered by the insurance plan.

• *smoker*: A yes or no categorical variable that indicates whether the insured regularly smokes tobacco.

• *region*: The beneficiary's place of residence in the US, divided into four geographic regions: northeast, southeast, southwest, or northwest.

Here, we want to predict medical "charges" from the variables given above by applying Multiple Linear Regression. (Hint: When loading the data, use stringsAsFactors = TRUE because it is appropriate to convert the three nominal variables to factors.)

a. Prior to building a regression model, it is often helpful to check for normality. Although linear regression does not strictly require a normally distributed dependent variable, the model often fits better when this is true. Look at the summary statistics and draw the histogram of the dependent variable. Comment on the results.
b. Create a correlation matrix and a scatterplot matrix for the four numeric variables in the insurance data frame. Do you notice any patterns in these plots in the scatterplot matrix?
c. Build a regression model using the independent variables, then evaluate the model performance.
d. Perform regression diagnostics using both typical approach and enhanced approach for the regression model you build in part (c).
e. Improve the regression model by adding a non-linear term for age and creating an indicator for obesity. Assume that BMI is strongly related to higher costs for the obese (that is, BMI of 30 or above), but has zero impact on medical expenditures for individuals in the normal weight range. Compare the results with the part (c).

Problem 4 (25points). Multiple Linear Regression Model for Forest Fire Data

- Create a scatterplot matrix of "Forest Fire Data" and select an initial set of predictor variables
- Build a few potential regression models using "Forest Fire Data"

- Perform regression diagnostics using both typical approach and enhanced approach
- Identify unusual observations and take corrective measures
- Select the best regression model
- Fine tune the selection of predictor variables
- Interpret the prediction results

Files Included in the Assignment:

Homework 3.docx
R Tutorial on Regression.pdf
Simple Linear Regression.pdf
Multiple Linear Regression.pdf
concrete.csv
Modeling Slump Flow Concrete.pdf
Concrete Slump Test Data Description.pdf
Concrete Slump Test Data.xlsx
insurance.csv
Data mining Approach to Predict Forest Fires.pdf
Forest Fires Data Description.pdf
Forest Fires Data.xlsx
mtcars.R
states.R
women.R