**CSC423: Data Analysis And Regression / CSC 324: Data Analysis & Statistical Software II**

**Assignment-3** | **Total Points: 25pts for CSC 423 / 20pts for CSC 324**

**Due Date: 01/31/2017 by 11:59 pm**

**Problem 1 [5 pts] – ONLY for GRADUATES**

A university career center collects information on the job status and starting salary of graduating seniors. Data recently collected over a two-year period included over 900 seniors who had found employment at the time of graduation. The information was used to model starting salary Y as a function of two qualitative independent variables: COLLEGE at four levels {Business, Engineering, Liberal Arts, Nursing} and SEX (male and female).

1. Define the dummy variables to include college (use Business as your baseline) in a regression model for starting salary Y
2. Write down the general regression model relating starting salary Y to both college and sex.
3. How would your model change if students in Engineering have the same starting salary as students in Business? Show the final regression model.

**Problem 2 [4 pts] – to be answered by everyone**

You will continue the analysis of the banking.txt dataset that was analyzed in Assignment 2 – data file is attached.

1. Analyze the residuals of the regression model you found in your previous assignment. Include the residual plots.
2. Conduct a global F-test for overall model adequacy. Write down the test hypotheses and test statistic and discuss conclusions. Include the relevant output.

**Problem 3 [14pts] – to be answered by everyone**

A national homebuilder builds single-family homes and condominium style townhouses.

The file housesales.txt provides information on the selling price (PRICE), lot cost (COST), type of home (HOME) (SF=single family home or T=condominium style) and region of the country (REGION) (M=Midwest, S=south) for closings during one month.

1. Define the dummy variables for region and home (write them down here), and create them in SAS.
2. Analyze the association between selling price and each individual attribute (cost, home and region) using appropriate statistics and graphs. Discuss your findings. Include the relevant output.
3. Fit an adequate regression model for sales price as a function of lot cost, region of country, and type of home. Remove the terms that are not significant. The final model should only contain variables that are significantly associated with sale price. Write down the model equation. Include the relevant output.
4. Conduct goodness of fit test for the selected regression model. Include the relevant output.
5. Analyze model residuals to check if assumptions on data are satisfied. Discuss your findings. Include the relevant output.
6. Discuss what the regression model indicates for the relationship between price and home type (i.e. interpret the coefficient values).
7. Use the regression analysis to determine whether mean sale prices are different for the two regions? Explain.

**“Reflection” Problem [2 pts] – to be answered by everyone**

Post a message on the discussion board reflecting on the topics in week 4. Indicate the topic you found to be the easiest, the one you found to be the hardest, and why. I created a new Thread called “Assignment-3 Reflection Comments”.

**Submission instructions:**

* Submit the assignment using the dropbox “Assignment3” via D2L.
* Keep a copy of all your submissions.
* If you have questions about the homework, email me BEFORE the deadline.
* Assignments submitted 6 days after the due date will not be accepted. A 20% point penalty will be applied for late assignments.
* Assignments submitted four days after the due date will not be accepted.