**Data Analysis And Regression**

**Assignment-2** | **Total points: 15 pts for DSC 323; 20 pts for DSC 423**

Note:

* All assignments should be submitted in a **single MS WORD format**, no PDFs or any other file types will be accepted. If you submit any other file type, it will not be graded.
* No extensions will be given unless for a documented reason specified in the syllabus, no late assignments past the due date even a couple of minutes late will be accepted as you have an extra day (7-days) to submit your assignments.
* Submitting work that is not yours is grounds for an automatic ‘F’ for the entire course – this includes taking content and ideas from others or consulting others to complete your deliverables other than your instructor.
* SAS software and virtual server stalls, gets slow and crashes; so start early and keep multiple backups in multiple places/mediums. Late submission or inability to do the assignment due to server and/or software issues will not be accepted. Any issues relating with SAS, contact IS using the phone number provided in the syllabus, I won’t be able to help you with DePaul software related issues.

***Note: For all questions, immaterial if whether the relevant output is asked to be attached or not, make sure to include it. Also, it is important to include the sign (negative/positive or increase/decrease, and units of measurements e.g. $ or $ 99 million,%, etc.) otherwise points will be deducted.***

**PROBLEM 1 [15 pts] – to be answered by everyone**

The file unemployment.txt attached to this assignment provides data acquired from census records for different zip codes. The data show

* median age of the population (Age)
* median income (Income) in $
* average bank balance (Balance) in $
* median years of education (Education) in years
* percent unemployment (Unemployment) in %

In this exercise you are asked to apply regression analysis techniques to describe the effect of age education, income and average account balance on unemployment.

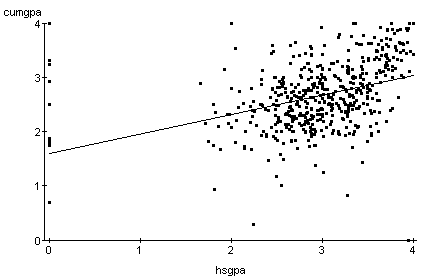
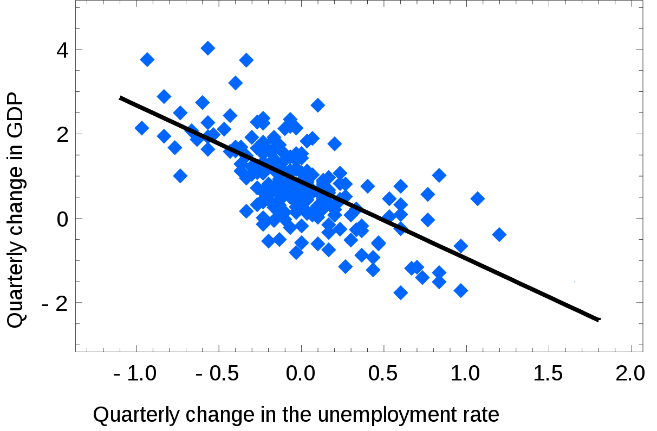
1. Analyze the distribution of unemployment using histogram, and compute appropriate descriptive statistics. Write a paragraph describing distribution of unemployment and use appropriate descriptive statistics to describe center and spread of the distribution. Discuss your findings. Also, explain if you see any outliers, if so, why do you think they are outliers? Include the histogram.
2. Create scatterplots to visualize the associations between unemployment and the other variables. Include the scatterplot. Discuss the patterns and association displayed by the scatterplot.
3. Compute correlation values of unemployment vs the other variables. Include the relevant output that shows the correlation values. Interpret the correlation values, and discuss which pairs of variables appear to be strongly/not strongly associated.
4. What is the dependent variable and what are the independent variables in this regression analysis?
5. Use SAS to run the full regression model to predict unemployment from age, education, income and balance. Include the relevant regression output. Analyze the model. Which predictors have a significant effect on unemployment, explain why you came to the conclusion that they are significant or insignificant?

1. If one of the predictors is not significant, remove it from the model and refit the new regression model. Include the relevant regression output. Make sure you include all the steps and outputs at arriving at the final model. Write the expression of the newly fitted regression model.
2. Interpret the value of the parameter estimates for the variables in the model.
3. Include the portion of the output that includes the R2and Adj-R2 coefficient values. Report the value for the R2and Adj-R2 coefficient and describe what it indicates. What can you conclude?
4. According to census data, the population for a certain zip code area has median age equal to 44.2 years, median education equal to 11.5 years, median income equal to $51,324 and balance is $34,200.
5. Use the final model computed in step (f) above to compute the predicted unemployment for the zip code area.
6. If the observed unemployment for the zip code area is 13.5%, what’s the model prediction error?
7. Copy and paste your FULL SAS code into the word document along with your answers.

**PROBLEM 2 [5 pts] – For Graduate Students ONLY**

* + - 1. Based on the two plots you see in figures 1 and 2, which one will produce a more accurate prediction? Need to provide explanation as to why you think it produce a better model.

**Figure-1 Figure-2**

* + - 1. Explain the type of correlation you see on both figures.