**Data Analysis And Regression**

**Assignment-5** | **Total Points: 25 pts for DSC 323; 30 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 [25 pts] – to be answered by everyone**

This problem asks you to build a model for the college dataset (college.csv) that contains the following variables:

*School School name*

*Private public/private indicator. YES if university is private, NO if university is public.*

*Accept.pct percentage of applicants accepted*

*Elite10 Elite schools with majority of students from the top 10% of their high school class*

*(0- Not Elite, 1-Elite)*

*F.Undergrad number of full-time undergraduate students*

*P.Undergrad number of part-time undergraduate students*

*Outstate Out-of-state tuition*

*Room.Board room and board costs*

*Books estimated book costs*

*Personal Estimated personal spending*

*PhD Percent of faculty with PhD*

*Terminal Faculty with terminal degrees (terminal degree is a university degree that is either*

*highest on the academic track or highest on the professional track in a given field*

*of study)*

*S.F.Ratio Student/faculty ratio*

*perc.alumni Percent of alumni who donate*

*Expend Instructional expenditure per student*

*Grad.Rate Graduation rate in 4 years*

Apply regression analysis techniques to analyze the relationship among the observed variables and build a model to predict Graduation Rates (Grad.Rate). **Note: Depending on how you import you data (INFILE or IMPORT) the SAS may relabel the column names. Make sure to use the variable names that appear when you use a proc print.**

***Note: Before you start, open the college.csv file, and examine the data.***

Answer the following questions.

1. Analyze the distribution of Grad.Rate and discuss if the distribution is symmetric, or if you need to apply any transformation (This is the data exploration stage, therefore use the appropriate statics to explore your data).
2. Create scatterplots for Grad.Rate vs each of the independent variables. What conclusions can you draw about the relationships between Grad.Rate and the independent variables? (No need to include the scatterplots in your submission).
3. Build boxplots to evaluate if graduation rates vary by university type (private vs public) and by status (elite vs not elite). Include the boxplots and discuss your findings. (See SAS Procedures section on D2L if you need the code to generate a boxplot).
4. Fit a full model (with all independent variables) to predict Grad.Rate. Discuss the parameter estimates, significance, goodness-of-fit and AdjR2 values. Include the relevant output.
5. Does multi-collinearity seem to be a problem here? What is your evidence? Compute and analyze the VIF statistics. Include the relevant output and discuss your answer.
6. Apply TWO variable selection procedures to find an optimal subset of independent variables to predict Grad.Rate*.* You can choose any two procedures among the ones we learned in class: backward selection, forward selection, adj-R2, Cp, stepwise. Make sure to include the o/p of the 2 selection methods. No need to discuss the models, include the outputs.
7. Fit a final regression model **M1** for Grad.Rate based on the results in f) – i.e. optimal model. Explain your choice. Write down the expression of the estimated model **M1**.
8. Draw a plot of the studentized residuals against the predicted values. Does the plot show any striking pattern indicating problems in the regression analysis? Include the outputs and explain.
9. Analyze normal probability plot of residuals. Is there any evidence that the assumption of normality is not satisfied? Include the outputs and explain.
10. Are there any outliers or Influential Points? Compute appropriate statistics. Include the outputs. Take any action you think is necessary and explain why/why not you took these actions?
11. Analyze the AdjR2 value for the final model and discuss how well the model explains the variation in graduation rates among the universities.
12. Draw conclusions on graduation rates based on your regression analysis. What are the most important predictors in your model? Does your model show a significant difference in graduation rates between private and public universities? Do “elite” universities have higher graduation rates? Explain.
13. Use the final regression model to predict the graduation rate for the following values. Using SAS, compute the predicted graduation rate, 95% confidence interval and prediction interval for your estimate. Make sure to use SAS coding to determine the values. Include all relevant outputs. Discuss your findings.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Private | Yes |  | Books | 250 |
| Accept.pct | 0.87 |  | Personal | 1350 |
| Elite10 | Not Elite |  | PhD | 40 |
| F.Undergrad | 3000 |  | Terminal | 34 |
| P.Undergrad | 524 |  | S.F.Ratio | 30.2 |
| Outstate | 6500 |  | perc.alumni | 13 |
| Room.Board | 3300 |  | Expend | 5201 |

1. Copy and paste your FULL SAS code into the word document along with your answers.

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

* 1. What selection methods did you use in Problem-1 (f)?
  2. Explain why you chose these two selection methods in Problem-1 (f) as opposed to the other methods. The reason(s) should have a methodological approach. If you say I wanted to try it, or it is the easiest to use or something to that effect you will receive zero. Provide reason(s) for both.