DSO 529
Professor Dawn Porter
Final Examination
Due by midnight Thursday, May 4, 2017 on Blackboard

Please answer your questions thoroughly and, when requested provide calculations and/or justification for your responses. Turn in only RELEVANT output, and please try to answer the specific questions on the test sheet. If you have any clarification questions, you can reach me by e-mail (dawn.porter@marshall.usc.edu).

Please sign the following statement when you have completed the exam.

Please print (or type) your name here: _____

I, _____, certify that I have neither given nor received assistance of any kind from anyone regarding the content or specific subject matter of this examination.

**If you know what you're doing,** the exam should take you **no more than** 5 hours!!! ☺

Good luck!!

I. (*Dataset: Rankings.xlsx*) This dataset contains US News full-time Business School rankings from 2002 – 2009. The variables are:

| | |
|---|---|
| *Rank:* | Rank of school |
| *School:* | School name |
| *Year:* | 2002 through 2009 |
| *GPA:* | Average undergraduate GPA of incoming class |
| *GMAT:* | Average GMAT score of incoming class |
| *AcceptRate:* | Percent of applicants accepted to school |
| *Salary:* | Average starting salary and bonus of graduates |
| *EmpGrad:* | Percent of students employed at time of graduation |
| *Tuition*[1]*:* | Annual tuition and fees for MBA program |
| *Enrollment:* | Overall student program enrollment |

A. For this first part, the goal is to predict the *Salary* variable.

a. Run a standard OLS regression using the variables just as they are (without *School* name), with no modifications at all, to predict the Salary variable. What is your fitted equation?

b. Are there any issues or violations with your output results from part a?

c. Attempt to run an FEM (Fixed Effects Model) with this data. (at this point, still don't do anything else and do NOT bother eliminating insignificant variables yet) Is the *group* of dummy variables statistically significant? Prove your answer statistically.

d. Are there any other issues remaining in the output of the model from part c?

e. Work with the data to correct for the problem you noted in part d. (You shouldn't need to spend TOO much time doing this!) Again, predict the *Salary* variable. Include a brief report of any models you tried and why you discarded them. (maybe a short table) What is your final model, including the output and residual plot?

B. For this second part, you are to use your knowledge of Clustering Analysis. ☺

a. Using a combination of *Hierarchical* and *K-Means* clustering methods, what number of clusters would you suggest using all the data except the *School* names? What is your reason for choosing that number?

b. Regardless of your answer to the previous question, form 6 clusters of the data and save the cluster numbers to your worksheet (under the red triangle, choose *Save Clusters*). Create a summary table of the clusters' statistics and copy it to your answer sheet. (*Analyze – Tabulate*, drag summary statistics of *N, Mean, Std Dev, Min,* and *Max* to the top to be the column headings. Then drag the *Cluster Number* variable to the left side to be the row headings, and finally select the 8 numerical variables and drag them just to the right of the *Cluster Number* variable. You can save this to its own table by using the red triangle next to *Tabulate*.)

[1] Some modifications were made to estimate annual costs for schools that reported values as either total program cost or cost per academic credit.

c. How would you characterize each cluster? Give just a brief description of each of the 6 in a summary table.

d. Which, if any, of the quantitative variables are significantly predictable using only the *Cluster Number?* Justify your answer statistically. What method did you use to answer this?

II. (*Dataset: Movies.xlsx*) A set of movies from various years has been compiled to see if there is a significant difference in *US Domestic Revenues* between different movie *Genres* or *MPAA Ratings*. There are a total of 11 different *Genres* and 4 different *MPAA Ratings* listed in the dataset.

a. What type of analysis is appropriate to analyze this? What are the null hypotheses for these two goals?

b. Determine whether there is a statistically significant difference overall in *US Domestic Revenues* between different *Genres*. Which *Genres*, if any, are statistically different from each other in terms of *US Domestic Revenues*?

c. Determine whether there is a statistically significant difference overall in *US Domestic Revenues* between different *MPAA Ratings*. Which *Ratings*, if any, are statistically different from each other in terms of *US Domestic Revenues*?

d. Give a *general* interpretation of your final analysis.

III. *(Dataset: CarSales.xlsx)* Monthly car sales were recorded in Quebec during the years 1960-1968. The variable definitions are self-explanatory. ☺

a. Is the *Car Sales* variable stationary? How do you know (statistically)? If it is *not* stationary, what needs to be done to make it stationary? Justify your answer.

b. Does there seem to be a seasonal, or cyclical, pattern to the data? Justify your answer.

c. Using your vast knowledge of *Time Series* analysis, create a solid model to predict *Car Sales* for the next 4 months (I actually have the true values and can compare them to your estimates!), including 95% intervals.