DSO 529
Professor Dawn Porter
Midterm Examination
Due by Tuesday, March 7, 2017 at 11:59pm on Blackboard.

Please answer all three 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).

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.

The exam should take you no more than 5 hours if you're prepared! ☺

Good luck!!

I. The first question is based on the data found in *Wages.xls.*

1. Fit a standard regression model to predict ***Wage*** (hourly rate) based on ***Union*** (1 if in a union, 0 if not), ***Educ*** (years of schooling), ***Experience*** (years of work experience), ***Female*** (1 for female, 0 for male), and ***Married*** (1 if married, 0 otherwise).
   a) What is your fitted equation?

   b) Does there seem to be a significantly different wage for women over men based on this regression? Prove your answer.

   c) Create a scatterplot using ***Wage*** as the dependent variable and ***Educ*** as the independent variable, but split the results by the ***South*** variable (1 means from the south). [JMP: Analyze – Fit Y by X, and enter the variables. Once the graph appears, go to the red triangle and select the Group By option. Highlight the *South* variable. Then choose the Fit Line option from the red triangle.] What does this plot indicate would be a potentially valuable inclusion to the model?

   d) Re-run the regression model including the suggestion from above. What is your new fitted equation? Does this improve your model?

2. Using the original model from part 1 (using the 5 suggested independent variables) and create the residual plot for this model.
   a) Is there any indication that there might be an issue with heteroscedasticity?

   b) Prove your answer to the above question statistically (not just visually).

   c) What variable(s) do you think might be the source of heteroscedasticity? Prove this.

   d) How might you attempt to alleviate this? Do whatever you think is reasonable to correct for the heteroscedasticity (don't bother dealing with any other model issues) and include your final output here.

3. This one is open-ended… do whatever you think might help you create a "good" regression model; include your final output and residual plot.

II. **(*Dataset: KidCreative.xls*)** A magazine reseller is trying to decide what magazines to include in emails to customers as part of a marketing campaign. All the e-mails that will be sent will go to customers that have previously bought a magazine subscription with them. Currently, they are focusing on who would be most likely to purchase a subscription to a children's magazine called "Kid Creative," whose target audience is children between the ages of 9 and 12. An experimental e-mail with an ad for "Kid Creative" was sent to a test group of 673 customers and the purchase behavior recorded.

Variables are:

Buy	(1 if purchased "Kid Creative", 0 otherwise) (Y variable)
Income	(in thousands)
IsFemale	(1 if the person is female, 0 otherwise)
IsMarried	(1 if married, 0 otherwise)
HasCollege	(1 if has one or more years of college education, 0 otherwise)
IsProfessional	(1 if employed in a profession, 0 otherwise)
IsRetired	(1 if retired, 0 otherwise)
Unemployed	(1 if not employed, 0 otherwise)
ResLength	(years living in same city)
Dual Income	(1 if dual income, 0 otherwise)
Minors	(1 if children under 18 are in the household, 0 otherwise)
Own	(1 if own residence, 0 otherwise)
English	(1 if the primary language in the household is English, 0 otherwise)
PrevParent	(1 if previously purchased a parenting magazine, 0 otherwise).
PrevChild	(1 if previously purchased a children's magazine)

a) Run an appropriate regression model to predict the *probability* a particular person will choose to purchase "Kid Creative" magazine. Don't adjust the data or remove any variables yet. What is your regression equation?

b) Using a significance level of 0.05, remove variables one at a time that don't seem to be statistically significant. Include here a table outlining your decisions and reasons (brief). For the last variable you wanted to remove in your table, prove statistically that it should or should not be included.

c) What is your final model choice? Include your output here.

d) Is this model statistically significant overall? Justify your answer.

e) If you removed the variable *IsFemale,* put it back in now. How does being a female customer change the *odds* of her buying "Kid Creative"?

f) Based on your final model from part d, what percent of the data was *correctly* re-classified?

g) Go back to the first row of the dataset. According to your final model, what is the *probability* that person will purchase "Kid Creative?"

III. This question will use the dataset *Unemployment.xls.* The data represent civilian unemployment in thousands of persons aged 16 years and over for the years 1964 – 2010.

a) Create a standard linear regression model to predict *unemployment* based on *year* (don't do anything fancy here!). What is your fitted equation?

b) Do you notice any issues with autocorrelation? How do you know, statistically?

c) Create a valid and good model using techniques we discussed in class to alleviate any model violations. Document the steps you take to get to this model (just list them out… I don't need to see all the results.). Make sure to include your final output and residual plot.

d) Predict *unemployment* for the years 2011 and 2012, as well as the 95% intervals for each of the years. (you can use JMP to do this for you if you want.)