Data Analysis Final Exam

out Thursday 12/03/09
due **Friday 12/11/09 5pm - HARD DEADLINE, NO EXTENSIONS**

This exam is a week-long take-home data analysis exam. You are allowed to use your textbook as well as other reference books you feel you might need. You should use the statistical software package R to perform your analysis. **You are under no circumstances allowed to consult with any person other than your professor and your teaching assistants.** You are expected to comply with the CMU policy on academic integrity. Unauthorized help will result in failing the exam.

Due Date: The exam is due to me by Friday December 11, 2009 5pm. Bring a printed copy to my office, mailbox, etc. and put an electronic copy in the Blackboard Digital Dropbox. You may email your exam if you will be out of town (rnugent@stat.cmu.edu).

Format: Your answers should be written in a report-style format with the following sections: *Introduction, Exploratory Data Analysis, Modeling the Data & Diagnostics, Results, Subgroup Analyses, Conclusions/Discussion*.

You have only eight pages (not double-sided) for all text and graphs. Anything over eight pages will not be read. Font size should be no smaller than 10 point. Graphs should be either imbedded in the report in a seamless manner or at the end of the report. Margins should be acceptable.
At the end of your eight page report, please attach an Appendix of your R code used in your analysis. I should be able to run your code and get the results in your report.

Language: Your language should be very clear and precise. Do not make claims for which you have no evidence. Do not say "will" or "would" when you really mean "may" or "might". Do not use language that implies causation. You are studying associations between variables only.
Grammar will be marked; writing style and syntax will weigh more heavily in your grade. Move away from wordy phrases *(e.g: This is because, this is due to, the reason that this is, this means that, I believe that this, I think the reason is that).* Do not start any sentence with This, These, etc without a noun following it. (*These results show vs. This shows*). Write strong sentences using careful scientific language. Use your space wisely.

Exam Instructions: This exam gives far fewer prompts than the previous exams. The primary goal is to build a good, theoretically valid model. While there are multiple approaches, any decision you make must be completely justified. Leaving out details will result in a loss of points.

Practice thinking about how to analyze data. While there are wrong answers, there are many possible right answers. Again, any data analysis decisions or conclusions that you make should be justified and explained. Your job is to correctly analyze the data, not force the analysis to match a pre-conceived idea. Discuss whether or not your results match your hypotheses. Brainstorm ways to improve the research study.

Again, you are more than welcome to discuss the research problem or any questions you have with the professor or the teaching assistant. **Help from any other individual is prohibited.**

Research Problem

Background info from P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physiochemical properties. In Decision Support Systems, Elsevier, 47(4): 547-553. ISSN: 0167-9236

The wine industry is a multi-billion dollar industry worldwide. Wines are made up of chemical compounds which are similar or identical to those in fruits, vegetables, and spices. The composition of these chemical compounds determines the taste and quality of the wine. For example, the sweetness of wine is determined by the amount of residual sugar in the wine after fermentation, relative to the acidity present in the wine. Some of the chemical compounds are more easily manipulated than others. The alcohol concentration can be changed by monitoring the grape sugar concentration prior to harvesting the grapes. The residual sugar is increased by suspending the sugar fermentation carried out by yeasts. The volatile acidity depends on the lactic bacteria control activity.

One of the top ten exporters of wine is Portugal (3.17% of the market share in 2005). Some of Portugal's most famous wines come from the Minho region of Portugal (in the northwest). These wines are referred by the term *vinho verde*. They are medium in alcohol and account for 15% of the total wine produced in Portugal. Exports of this wine have increased by 36% from 1997 to 2007, and Portugal has invested in new technological development to aid the growth of this industry. In particular, wine certification and quality assessment are two new fields of interest. The certification process protects the consumer and assures quality; the assessment, however, helps determine which chemical variables are useful in determining wine quality. The relationships between the objective physiochemical variables and the subjective sensory variable are not well understood. The wine industry is heavily invested in answering the question: what goes into a wine that people like?

In addition, having an accurate model predicting the quality of the wine from the chemical concentrations will greatly reduce costs at the wineries. When wine is in the certification phase, by Portuguese law, human tasters have to be used for the sensory analysis (i.e. determining the final quality score). This process is costly and time-consuming and subject to evaluator bias. However, under the assumption that the evaluators will (or should) choose a quality score near the predicted quality, we can decrease the number of necessary sensory evaluations. If a wine is predicted to be of low quality, we may not include it in the batch to be evaluated or re-evaluated by the tasters.

Your research group has been given a sample of red and white *vinho verde* wines. Your data were collected from May 2004 to February 2007 and tested at an official certification entity (CRVV), an organization dedicated to improving the quality and marketing of *vinho verde*. The most common physiochemical tests were chosen for the analytical variables. The quality was measured via blind taste tests. Each wine was evaluated by a minimum of three people; the median score was recorded. As happens in many analyses, some of the physiochemical test results were not recorded correctly or are missing.

Your research group has been asked by the CRVV to analyze the data and develop a model to predict the quality of the wine. They believe that a multivariate linear regression model is appropriate and are particularly interested in several hypothesized relationships. Previous research has suggested that more acidic wines (as measured by pH) and wines with higher amounts of sulfates are associated with lower quality scores. In addition, it is thought that the effect of alcohol concentration on quality is different for red and white wines. Furthermore, as alcohol concentration is tied to grape sugar concentration, it is hypothesized that the alcohol concentration has an effect on the relationship between amount of residual sugar and wine quality.

You have been given the following variables:

quality: quality score of the wine (1 - poor; 10 - excellent)

fix.acid: fixed acidity (g(tartaric acid)/dm^3)

vol.acid: volatile acidity (g(acetic acid)/dm^3)

citric: citric acid (g/dm^3)

sugar: residual sugar (g/dm^3)

chlorides: chlorides (g(sodium chloride)/dm^3)

free.sd: free sulfur dioxide (mg/dm^3)

total.sd: total sulfur dioxide (mg/dm^3)

density: density of the wine; categorized as follows:
1: < 0.99 (g/dm^3); 2: [0.99, 1.00] (g/dm^3); 3: > 1.00 (g/dm^3)

pH: measure of the acidity of the wine (lower values are more acidic)

sulphates: (g(potassium sulphate)/dm^3)

alcohol: ($\%$ volume)

type: type of wine (red or white)

You have been asked to build a simple, interpretable, theoretically valid multivariate linear regression model to predict wine quality that addresses the research hypotheses.

Your report should have the following sections:

Introduction: Write a short introduction describing the research problem. Provide some background information for context for the audience. Clearly state all research hypotheses at the end. Cite all sources you use for background information.

Exploratory Data Analysis: **Stop and think before doing the EDA.**

Examine all variables individually. Summarize them (numerically/graphically as appropriate).
How many observations do you have? What do they look like? Do you have any missing data?

If you have missing data, describe it. How many observations are missing data?
Which variables are missing the most data?

Contrast the complete observations with the incomplete observations.
Are there any striking differences? If so, how will they affect your model?

Think about what type of multivariate EDA would be the most useful/efficient for the variables.
Do numerical/graphical multivariate EDA on your variables. How are they correlated?
Describe any trends or interesting features that you see.

NOTE: All graphs should be very clearly labeled/titled, etc.
Extra time spent on cleverly displaying data is not such a bad thing.

Modeling the Data & Diagnostics:

Build an effective model using the complete observations for predicting wine quality from the given predictors. You should explore all main effects as well as interactions that seem appropriate.

Everything you do should be justified, both statistically and with respect to the research problem.

Removing variables from the model should be statistically justified. Explain how and why you chose to model your variables. Clearly explain how and why your model was chosen.

When building your model, look for outliers; do they affect your analysis?
Are there any observations that you remove? Why and what do they look like?

Create residual plots to determine the appropriateness of your model. Are the assumptions met? If not, what steps do you take to transform the data in order to meet them? Should you transform your response variable? Why or why not? If you decide transformations are necessary, do them and then report/discuss your final estimated regression function.

NOTE: Again, all plots/graphs should be clearly labeled/titled, etc.

Results:

Decide on a final model.

Report results in a table with coefficient estimates, standard errors, confidence intervals, and p-values. Is your regression significant? Why or why not? Discuss your results in context.
Interpret all significant parameters.

Discuss your results with respect to the research hypotheses.

Imputation Analysis:

For the incomplete observations, impute appropriate values for each variable.
Explain why you chose those values to impute (include graphical evidence if appropriate).

Re-run the final model that you chose above on all of your observations (the original complete observations and the imputed incomplete observations). *Don't re-build another model.*

Describe any changes you see in the model results (if any).
Which of the two models would you choose to report and why?

Conclusions/Discussion:

Summarize your main findings in the analysis. Report interesting findings in your exploratory data analysis and in your modeling. Discuss possible reasons for these findings. What is the final conclusion with regards to the original research hypothesis? Make some recommendations for future work or studies. What can be done to improve the research study?