Review for Midterm Exam 2

Exam Details

Date: Thursday, March 31st, 2011
Time: 8:00-9:00 PM
Location: LWSN B151
Content: Chapters 8-12 & 15-18.1 (KNNL) - Lectures 15-24

The exam will test conceptual ideas. You will need to be able to do simple calculations and you will need to be able to read and interpret SAS output. You will not need to prove anything, so theory is important only with regard to understanding the concepts.

Nothing specific from exam 1 material will be covered in detail on the second exam. However, general concepts related to regression may be tested, since the material from Chapters 8-12 covers regression. Note that for the Lecture 18 material (chapters 8, 11, and 12), you may need to know generally when you would apply certain methods, but do not need to be able do any detailed calculations for anything that I did not give a specific example for in class. For example, you might be asked to interpret output from incorporating a qualitative variable into a regression analysis (since we did this in class and on the homework), but you would not need to interpret output from a weighted least squares or robust regression analysis (just know when it is appropriate to use them).

Materials permitted

You are allowed a one page (8.5 x 11 in) sheet (both sides) of handwritten notes. Calculators are allowed. BUT, laptop computers, phones, or any devices capable of wireless communication are not permitted.

Chapter 8: Polynomial Models, Interaction Models, and Qualitative Predictor Variables

- Have a general understanding of polynomial models, noting that intercorrelation can be an issue which can sometimes be remedied by centering the variables. Know advantages and disadvantages of using a polynomial model. (Do not need to be able to interpret SAS output)
- Have a general understanding of interaction models and how to incorporate interaction terms. Know the difference between additive, reinforcing, and interference effects and be able identify from a graph. (Do not need to be able to interpret SAS output)
- Understand the use of an indicator (0-1) variable in a regression model to describe a qualitative predictor (such as male/female). Be able to give the meaning of parameter estimates in such a model, as well as test for differences between the overall line, intercepts, and slopes. Be able to give the estimated regression line for both groups. (You should able to interpret SAS output for qualitative predictors)
Chapter 9: Building the Regression Model: Selection of Predictors and Model Validation

- Know and understand how to use the various criteria for model selection ($R^2$, Adjusted $R^2$, Mallow’s $C_p$, AIC, BIC, PRESS). Be able to compare and contrast the various methods. Know how to read the SAS output in an effort to determine the “best” (or a couple of good candidate) model(s).
- Understand the “Best” subsets algorithm approach to model selection.
- Understand the concepts and advantages/disadvantages of forward selection, backward elimination, and stepwise regression as automated selection procedures. Be able to interpret SAS output for forward stepwise regression.
- Understand when to use different model selection procedures.
- Understand the basic concept of model validation.

Chapter 10: Diagnostics

- Have a general understanding of added variable (partial regression) plots and how they may be used to consider the marginal role of an independent variable given that other variables are already in the model.
- Understand how to identify outliers in the response (Y) using studentized or studentized deleted residuals. Understand why using studentized residuals is advantageous for outlier detection over using the regular residuals.
- Understand how to identify outliers in the predictors (X) using leverage values from the hat-matrix.
- Understand the other measures of influence including DFFITS, Cook’s D, and DFBetas.
- Be able to interpret SAS output for identifying outliers and influential observations using the above measures.

Chapters 11/12: Remedial Measures/ Autocorrelation

- Know what weighted least squares regression is and why it would be used.
- Be somewhat familiar with what other remedial measures are available (e.g. Ridge regression, Bootstrap, Time Series) and their uses.
- You do not need to be able to interpret any SAS output or know a lot of details for these methods.
Chapter 15: Introduction to Design of Experiments

- Know and understand the following terminology:
  - Categorical/qualitative variable
  - Factor/ factor levels
  - Experimental vs. observational factor
  - Qualitative vs. quantitative factor
  - Treatment/Treatment combination
  - Experimental unit
  - Balanced design
  - Crossed vs. nested factors
  - Control group
  - Fixed vs. random effect
  - Randomization
- Be able to read the description of an experiment and identify any of the above terms.

Chapters 16-18.1: One-way Analysis of Variance (ANOVA)

Note: It will be important to know how to interpret SAS output for many concepts in these chapters.

- Understand the difference between a regression analysis and an ANOVA in terms of
  - The difference between a regression model and an ANOVA model in terms of the assumed relationship between the dependent and independent variable(s).
  - The difference in the variable types used (qualitative vs quantitative)
- Be able to understand the connection between regression and ANOVA by interpreting SAS output from a regression analysis using indicator variables for r-1 factor levels of the ANOVA.
- Be able to interpret plots (scatterplots/boxplots) to identify visual differences between groups.
- Understand the cell means model and the factor effects model (with constraints).
- Be able to compute parameter estimates for either model given a table of means.
- Understand the partitioning of SS and the ANOVA table in general. Recall that the pooled variance estimate \( \hat{\sigma}^2 = \text{MSE} \).
- Be able to conduct the overall F-test for testing the equality of all factor level means and give details associated with the test.
- Be able to compute CI’s and conduct hypothesis tests for means and differences between means (which may require calculation of the standard errors)
- Be able to determine the minimum significant difference under LSD, Tukey, Bonferroni, or Scheffe methods and test for differences among the means. Understand the differences between these methods! Know how to interpret SAS output for these methods.
- Be able to set up and use linear combinations/contrasts to construct CI’s or test hypotheses about different groups of means.
- Know the three basic assumptions (on the error terms) of ANOVA. Be able to suggest diagnostics for determining whether these are satisfied. Note: We will cover other items related to diagnostics/remedies, but you will not be tested over these in Exam 2.
- Understand the differences between one vs. two-sided tests. Understand the different types of errors that can be made in testing and discuss the importance of sample size.
- Be able to determine the appropriate sample size needed given other required information either from table B.12 (which would be provided, if needed) or output from Proc Power.
Review Problems for Exam II

Note that these questions do not provide an extensive review. For example, you will also need to know how to identify necessary values from SAS output and no such questions are included here. This is just meant to get you thinking about possible topics and questions you may see. Solutions will not be provided, however I will be glad to discuss any of the questions during office hours or in class review.

1. Review your homework/lecture notes for regression with qualitative predictors (Lecture 18) and how to interpret SAS output from that model.

2. Explain the conceptual difference between $R^2$ and adjusted-$R^2$. Also indicate in what situations each should be used.

3. If all $p - 1$ available predictor variables are included in the model, the value of the Mallow’s $C_p$ statistic will be $p$. Explain (conceptually, in terms of what Mallow’s $C_p$ actually measures) why this is the case.

4. Discuss in what situations it would be preferable to use an automated selection procedure rather than using model selection criteria. Also, discuss the advantages/disadvantages of forward/backward selection and why stepwise is more preferable.

5. What quantity would you use to identify outliers in the response? In the predictor variables? What values would you compare these to?

6. Which influence statistics would you consider if you were interested in the influence that a single data point was having (a) on the estimated regression line, (b) on the predicted value at that point, (c) on the parameter estimate for $\beta_1$?

7. Give the three basic assumptions on the error terms for ANOVA analysis. Indicate how you would check they are satisfied. What assumption is not made in an ANOVA model that IS made in a regression model?

8. Suppose that in a one-way ANOVA you have 5 treatments for which the respective means are 7, 8, 11, 8, and 15. Write the cell means and factor effects models (with assumptions and constraint(s)). Give the parameter estimates for both models. For the factor effects model, show that the tau’s sum to zero.

9. How many indicator variables would be necessary if we wanted to analyze the situation in Problem #8 as a regression model (using the approach discussed in class)?

10. (15 points) Suppose that factor A has 5 levels and there are 4 observations per level. Complete the following ANOVA table and conduct the overall F-test for the equality of factor level means. Note that you can find the F-critical value to compare your F-statistic to using table B.4 (but you may need to round down for the denominator degrees of freedom). Based on your conclusion, would it be appropriate to conduct any further analysis to identify differences in factor level means? What is the estimate for $\sigma$?

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>___</td>
<td>60</td>
<td>___</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>___</td>
<td>___</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>___</td>
<td>___</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
11. Compare and contrast the methods of multiple comparison (LSD, Tukey, Bonferroni, and Scheffe).
   a. Which would be most appropriate in a one-way ANOVA having 3 treatment levels, but you plan to compare the first treatment to the average of the other two and in separate comparison look for a difference between the last two treatments?
   b. Which would be most appropriate if the factor has 8 levels with the same number of observations per level and you plan to check all pairwise differences of means?
   c. Which would be most appropriate if you have looked at your data before deciding what to test?

12. Suppose that in a one-way ANOVA we have four treatments and the estimates for the treatment means are 25, 32, 27 and 31, respectively. Assume that MSE = 9, n = 5, and you want to do all pairwise comparisons. Provide a Bonferroni adjusted confidence interval for the pairwise difference $\mu_4 - \mu_1$. You can use table B.2 to find an appropriate critical t (it may not be exact). What is the minimum significant difference? Be able to calculate such a CI or minimum significant difference for LSD, Tukey, or Scheffe, if given the appropriate critical values.

13. In a one-way ANOVA having 6 treatments, write a contrast that could be used to test whether the combination of levels 1 and 2 has the same mean as both the combination of levels 3-6.

14. In a one-way ANOVA the four treatment levels are (men, smoker), (women, smoker), (men, nonsmoker), and (women, non-smoker), with 16 subjects per group. These have means of 3, 5, 9, 11 respectively. The MSE is 4.
   a. Find a 95% CI for the contrast that would test whether there is a significant difference between smokers and non-smokers. Note that $t_{0.975}(60) = 2$. What is your conclusion?
   b. If we were interested in testing whether the smokers had a lower average than the non-smokers, state the null and alternate hypotheses. What critical value would you compare the test statistic to?

15. Explain what the power of a test is and why it is important. What is one thing we can do to both increase power and decrease the probability of a type I error simultaneously?

16. What is the total sample size you would need to detect a difference of 5 or greater between the three treatment means at a significance level of 0.05 and with power of 0.80 (assume $\sigma=4$)? You can use table B.12.