CPH 931 — Fall 2008 — Dr. Charnigo

Midterm Examination

Complete two exercises of your choice. This Midterm Examination is due on Tuesday 21 October at the end of lecture and is a strictly individual activity.

[50] 1. The file {BMIData2.xls} contains a modified version of the data in {BMIData.xls}. The modifications are that I have introduced many missing values and that I have deleted WEIGHT1, which you will not use in this exercise.

[10] a. Using records with no missing values, apply ordinary least squares to fit a linear regression model with BMI as the response variable and explanatory variables chosen from among ENERGY, TOTFAT, CHOL, FIBER, and ALCOHOL via backward elimination. Report the parameter estimates, standard errors, and p-values for the model chosen via backward elimination.

[10] b. Refit the chosen model from part a using not just the records with no missing values but also the records with missing values only on variables not included in the chosen model. So, a line of SAS code like

```
MODEL Y = X1 X2 X3 X4 X5 / SELECTION=backward;
```

would be replaced by a line of SAS code like

```
MODEL Y = X1 X2 X3;
```

if the chosen model included X1, X2, X3 but not X4, X5. Compare the results to those obtained in part a.

[10] c. Use multiple imputation with SEED = 12345 to generate five complete data sets. Apply ordinary least squares to fit a linear regression model for each complete data set with BMI as the response variable and explanatory variables ENERGY, TOTFAT, CHOL, FIBER, and ALCOHOL. Combine the five sets of results to obtain overall estimates of parameters, standard errors, and p-values. Report the overall estimates of parameters, standard errors, and p-values.

[10] d. Continuing from the previous item, perform backward elimination "manually" by: (1) identifying which explanatory variable has the largest p-value; (2) refitting the linear regression model for each complete data set without this variable; (3) updating the overall estimates of parameters, standard errors, and p-values; and, (4) repeating steps (1) to (3) until all remaining explanatory variables have p-values less than 0.05. Compare the results to those obtained in part a.

Note: Do not regenerate the five complete data sets by removing variables in PROC MI. The variables used in PROC MI need not be the same as those used in subsequent analyses.

[10] e. Suppose you are presenting your results from part d and someone asks you whether the assumption of normally distributed errors is valid. Can you think of some meaningful way to address this question in a linear regression analysis entailing multiple imputation?

[50] 2. The file {BloodPressure.xls} contains data from a small study comparing two treatments for hypertension (Treatment = 1 or 2). Each subject's systolic blood pressure (CurrentBP) is recorded at one week (Time = 1), one month (Time = 2), and three months (Time = 3) after the subject begins treatment. Also recorded for each subject are a baseline systolic blood pressure before treatment (BaselineBP) and whether the subject is a nonsmoker (Smoker = 1) or a smoker (Smoker = 2).

[10] a. Fit a linear mixed model for CurrentBP with explanatory variables as indicated by

```
CLASS SUBJECT TREATMENT TIME SMOKER;
MODEL CURRENTBP = TREATMENT TIME TREATMENT*TIME SMOKER BASELINEBP / SOLUTION;
```

and random effects for subjects. If the linear mixed model is written out in the form $Y_{ij} = \beta_0 + \sum_{k=1}^{m} \beta_k x_{k;ij} + \alpha_i + \epsilon_{ij}$, what is m and how are $X_1, \ldots, X_m$ defined?

Note: Unless X_k is continuous, defining X_k requires you to state explicitly when $X_k = 1$.

[10] b. Let "Subject a" be a nonsmoker with baseline blood pressure 150 who receives the first treatment, and let "Subject b" be a nonsmoker with baseline blood pressure 150 who receives the second treatment. Estimate the expected value of blood pressure for "Subject a" at one week ("a1"), at one month ("a2"), and at three months ("a3"). Do the same for "Subject b" ("b1", "b2", "b3").

[10] c. Test the null hypothesis that the two treatments are equally effective at one week. How can this null hypothesis be expressed in terms of $\beta_1, \ldots, \beta_m$?

Hint: Using notation from the previous item, you are looking for whether "b1 - a1" = 0.

[10] d. Test the null hypothesis that the two treatments are equally effective at all three time points. How can this null hypothesis be expressed in terms of $\beta_1, \ldots, \beta_m$?

[10] e. If "b3 - b1" is different from "a3 - a1", what does this mean scientifically? Estimate "b3 + a1 - b1 - a3", state whether the estimate is significantly different from 0, and interpret scientifically.

[50] 3. The file {CigaretteConsumption.xls} contains historical data on state-level variables thought to be related to annual cigarette sales (Sales, expressed in millions of dollars), including median age (Age), percent of residents with a high school education (HS), per capita income (Income, expressed in dollars), percent of African-American residents (Black), percent of female residents (Female), and average price per pack of cigarettes (Price, expressed in cents — yes, this is a very old data set!).

This exercise does not have parts a, b, c, d, and e. Rather, your task is to write a "Case Study" illustrating the application of linear regression to a real data set. You may assume that the reader is familiar with Lectures 1 through 3 of CPH 930 and Lectures 1 through 2 of CPH 931.

Be selective in what you present. Do not try to demonstrate every idea that you learned in CPH 930 and CPH 931. Rather, synthesize what you learned in CPH 930 and CPH 931 to address as best you can the question of what drives cigarette sales at the state level.

Note: You are limited to 1000 words and to three tables or figures.