STAT 203 PROBLEM SET 2

Due date: February 4, 2010

(1) In one-way ANOVA, the model is:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad \epsilon_{i,j} \sim N(0, \sigma^2).$$

The mean sum-of-treatment-squares term is

$$MSTR = \sum_{i=1}^{r} n_i \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2 / (r-1).$$

Assume that $n_i = n$, i.e. each level has n observations. Show the following:

(a) For each i,

$$\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot} = \alpha_i + (\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}),$$

where $\bar{\epsilon}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} \epsilon_{ij}$ and $\bar{\epsilon}_{\cdot\cdot} = \frac{1}{nr}\sum_{i=1}^{r}\sum_{j=1}^{n} \epsilon_{ij}$

(b) $\sum_{i=1}^{r} \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2 = \sum \alpha_i^2 + \sum (\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot})^2 + 2\sum \alpha_i (\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot})$,

(c) $E[2\sum \alpha_i(\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot})] = 0$,

(d) Use (a-c), and the fact

$$E\left[\sum (\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot})^2\right] = \frac{(r-1)\sigma^2}{n},$$

to show that

$$E(MSTR) = \sigma^2 + \frac{n\sum_{i=1}^{r} \alpha_i^2}{r-1}.$$

(2) In the surgery rehab data of lecture 8 (data file Rehab.txt), design and test the hypothesis that the recovery times of patients in the below and above average fitness group have the same absolute deviation from the average group. Write out explicitly the null hypothesis and the way you conducted the test.

(3) Presidential Election Data (1916-1996): This data set was kindly provided by Professor Ray Fair of Yale University, who has found that the proportion of votes obtained by a presidential candidate in a United States presidential election can be predicted accurately by three macroeconomic variables, incumbency, and a variable which indicates whether the election was held during or just after a war (variables listed in table below). The variables considered are given

in the data file. All growth rates are annual rates in percentage points. Consider fitting the initial model

$$V = \beta_0 + \beta_1 I + \beta_2 D + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \epsilon$$

to the data. Using ANOVA tests, answer the following questions:

(a) Do we need to keep the variable I in the above model?

(b) Do we need to keep the interaction variable $G \cdot I$ in the above model?

(Data file: Election.txt)

YEAR	Election Year
V	Democratic share of the two-party presidential vote.
I	1 if incumbent is democratic, -1 if incumbent is republican.
D	1 if democratic incumbent is running for election, -1 if republican incumbent is running for election, 0 otherwise.
W	1 for elections of 1920, 1944, and 1948, 0 otherwise.
G	Growth rate of rea per capita GDP in the first three quarters of election year.
P	Absolute value of the growth rate of the GDP deflator in the first 15 quarters of the administration.
N	Number of quarters in the first 15 quarters of the administration in which the growth rate of real per capita GDP is greater than 3.2%.

(4) A research studied the sodium content in lager beer by selecting at random six brands from the large number of brands of U.S. and Canadian beers sold in a metropolitan area. The researcher then chose eight 12-ounce cans or bottles of each selected brand at random from retail outlets in the area and measured the sodium content (in milligrams) of each can or bottle. Assume a random effects ANOVA model.

(a) Test whether or not the mean sodium content is the same in all brands sold in the metropolitan area; use $\alpha = 0.1$. State the null and alternative hypothesis, and show the ANOVA table. What is the P-value of the test?

(b) Estimate the mean sodium content for all brands; give a 99 percent confidence interval.

(data file Beer.txt)