STA 303H1S: One-way Analysis of Variance Practice Problems

1. (Adapted from question 1 in chapter 17 of Kleinbaum *et al.*)
   Five treatments for fever blisters, including a placebo, were randomly assigned to 30 patients, with 6 patients receiving each treatment. The data are the numbers of days from initial appearance of the blisters until healing is complete. The analysis is available at the course website.

   (a) Write the linear model fit to these data. Define all terms.

   (b) Write the linear model in matrix terms.

   (c) What is the estimate of the intercept? What practical quantity does it estimate?

   (d) What are the estimates of the other regression coefficients? What practical quantity do they estimate?

   (e) Do the effects of the five treatments differ significantly?

   (f) Post-hoc pairwise comparisons of the means has been carried out. What do you conclude? Which of the procedures carried out is most appropriate in this example?

   (g) Does there appear to be any violation of the model assumptions?

2. In the Spock conspiracy trial example, for the comparison of Spock's judge versus all of the other judges considered as one group, we conducted the analysis two ways: (1) using a two-sample t-test, and (2) using analysis of variance (a linear model). Show that the square of the two-sample t-test test statistic is the test statistic for the analysis of variance F-test.

3. The table below gives the number of observations, mean and standard deviation for the percentage of women for each of the seven Boston judges from the Spock conspiracy trial example. Use these summary statistics to construct the ANOVA (analysis of variance) table.

Judge	Number of observations	Mean	Standard deviation
A	5	34.12	11.94
B	6	33.62	6.58
C	9	29.10	4.59
D	2	27.00	3.82
E	6	26.97	9.01
F	9	26.80	5.97
Spock's	9	14.62	5.04

4. Consider the one-way analysis of variance model for all seven judges:

$$Y_i = \beta_0+\beta_1 I_{A,i}+\beta_2 I_{B,i}+\beta_3 I_{C,i}+\beta_4 I_{D,i}+\beta_5 I_{E,i}+\beta_6 I_{F,i}+e_i, \;\; i = 1, \ldots, N$$

where, for example, $I_{A,i} = 1$ if the ith venire belongs to judge A and is 0 otherwise, Y_i is the percentage of women in the ith venire, and N is the total number of observations for all seven judges. Assume that the e_i's are uncorrelated with constant variance σ^2.

(a) Show that

$$\text{Var}(b_1) = \sigma^2 \left(\frac{1}{n_{Spock's}} + \frac{1}{n_A} \right)$$

where b_1 is the least squares estimate for β_1 and $n_{Spock's}$ and n_A are the number of observations for Spock's judge and judge A, respectively.

(b) Show that

$$\frac{1}{N-G} \left\{ \sum_{(Spock's)} (Y_i - \overline{Y}_{Spock's})^2 + \sum_{(A)} (Y_i - \overline{Y}_A)^2 + \sum_{(B)} (Y_i - \overline{Y}_B)^2 + \sum_{(C)} (Y_i - \overline{Y}_C)^2 + \sum_{(D)} (Y_i - \overline{Y}_D)^2 + \sum_{(E)} (Y_i - \overline{Y}_E)^2 + \sum_{(F)} (Y_i - \overline{Y}_F)^2 \right\}$$

is an unbiased estimate of σ^2, where, for example, $\overline{Y}_A$ is the mean of the observations for judge A, $\sum_{(A)}$ denotes the summation over observations for judge A only, and G is the number of judges (7).

5. Here is an alternative way of creating a linear model for the analysis of Spock's judge versus the other 6 judges considered as one group. Consider the model

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, \ldots, N$$

where

$$x_i = \begin{cases} -\frac{1}{n_{Spock's}} & \text{if the } i\text{th venire is for Spock's judge} \\ +\frac{1}{n_{Other}} & \text{if the } i\text{th venire is for one of the other judges} \end{cases}$$

where $n_{Spock's}$ is the number of venires for Spock's judge and n_{Other} is the number of venires for all of the other judges. Show that the least squares solutions for this model are

$$\begin{aligned} b_0 &= \overline{y} \\ b_1 &= \left(\overline{y}_{Other} - \overline{y}_{Spock's}\right) / \left(\frac{1}{n_{Spock's}} + \frac{1}{n_{Other}}\right). \end{aligned}$$

(Using x_i as defined above, rather than dummy variables, is called "effect coding".)

6. Multiple subscripts are often used in analysis of variance models. For example, suppose there are N observations, and each observation is classified into one of G groups. We can denote the observations by Y_{gi} using two subscripts, the first denoting the group g to which the observation belongs, $g = 1, \ldots, G$, and the second indexing the observation within the group, $i = 1, \ldots, n_g$, where n_g is the number of observations in the gth group.

   We can define the linear model as

   $$Y_{gi} = \theta_g + e_{gi}$$

   and fit it by minimizing

   $$\sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{gi} - \theta_g)^2$$

   with respect to $\theta_1, \ldots, \theta_G$.

   Show that the minimizing values are $\hat{\theta}_g = \overline{y}_g$, where, $\overline{y}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} y_{gi}$.

7. For the one-way classification with G groups, the total sum of squares can be written as

   $$\sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{gi} - \overline{y})^2 = \sum_{g=1}^{G} n_g (\overline{y}_g - \overline{y})^2 + \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{gi} - \overline{y}_g)^2.$$

where y_{gi} is defined in question 6. This is easily shown by first adding and subtracting the predicted value of the observation in the total sum of squares and expanding that result. Verify that

$$\sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{gi}-\overline{y})^2 = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{gi}-\overline{y}+\hat{y}_{gi}-\hat{y}_{gi})^2 = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(\hat{y}_{gi}-\overline{y})^2+\sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{gi}-\hat{y}_{gi})^2.$$

8. The analysis of variance table below is for the analysis comparing the means of all seven judges in the Spock conspiracy trial example. Many cells in the table have been left blank. Fill in the blank cells using only the other entries and the total number of observations (which is 46). Estimate the p-value as accurately as possible using R, or from an F-table. (You can use the F-table at

   `http://www.sjsu.edu/faculty/gerstman/StatPrimer/F-table.pdf`

   if you don't have a text handy.)

Source	df	SS	MS	F
Model	6	1927.1	_____	6.72
Error	_____	_____	_____	
Total	_____	_____		

   (*Note:* Of course you can get the answers by looking at the R output for the example that is posted on the course website. But make sure you can do this with just the information given for this question.)