MLDS 401/IEMS 404-1 (Fall 2023): Lab 4 – 10/17/2023

3.7 (Derivation of the one-way ANOVA F-test using the extra SS method) In one-way ANOVA we have $k \geq 2$ groups (e.g., treatment groups) with n_i observations, $y_{i1}, \ldots, y_{in_i}$, from the ith group ($i = 1, \ldots, k$). Let $N = \sum n_i$ denote the total sample size. The data are usually modeled as

$$y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (i = 1, \ldots, k),$$

where the ε_{ij} are i.i.d. $N(0, \sigma^2)$ random errors, μ_i is the ith group mean, $\mu = \sum n_i \mu_i / N$ is the overall mean and $\alpha_i = \mu_i - \mu$ is the ith group "effect" subject to the linear restriction $\sum n_i \alpha_i = 0$. We want to test the overall null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ or equivalently $H_0 : \alpha_1 = \cdots = \alpha_k = 0$. It is easy to show that the LS estimates of the unknown parameters in the above linear model are as follows:

$$\widehat{\mu}_i = \overline{y}_i, \widehat{\mu} = \overline{\overline{y}} = \frac{\sum n_i \overline{y}_i}{N} \quad \text{and} \quad \widehat{\alpha}_i = \overline{y}_i - \overline{\overline{y}}.$$

(Alternatively, the overall mean μ may be defined as the unweighted average of the μ_i's: $\mu = \sum \mu_i / k$, in which case the $\alpha_i = \mu_i - \mu$ satisfy the linear restriction $\sum \alpha_i = 0$. The LS estimates of μ and the α_i will be different for this parameterization, but the estimates of the μ_i will be the same, namely $\widehat{\mu}_i = \overline{y}_i$.)

a) Show that the ANOVA identity (3.13) can be expressed as

$$\underbrace{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{\overline{y}})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{k} n_i (\overline{y}_i - \overline{\overline{y}})^2}_{\text{SSG}} + \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2}_{\text{SSE}},$$

where SST and SSE have their usual meanings and SSG is referred to as the SS between the groups.

b) Using the above identity, show that the extra SS F-test rejects H_0 at level α if

$$F = \frac{\text{SSG}/(k-1)}{\text{SSE}/(N-k)} > f_{k-1, N-k, \alpha}.$$

6.2 (Testing the difference between R_p^2 and R_q^2) Let R_p^2 and R_q^2 denote the R^2's for the full model with p predictors and a partial model with $q < p$ predictors. Show that the extra SS F-statistic equals

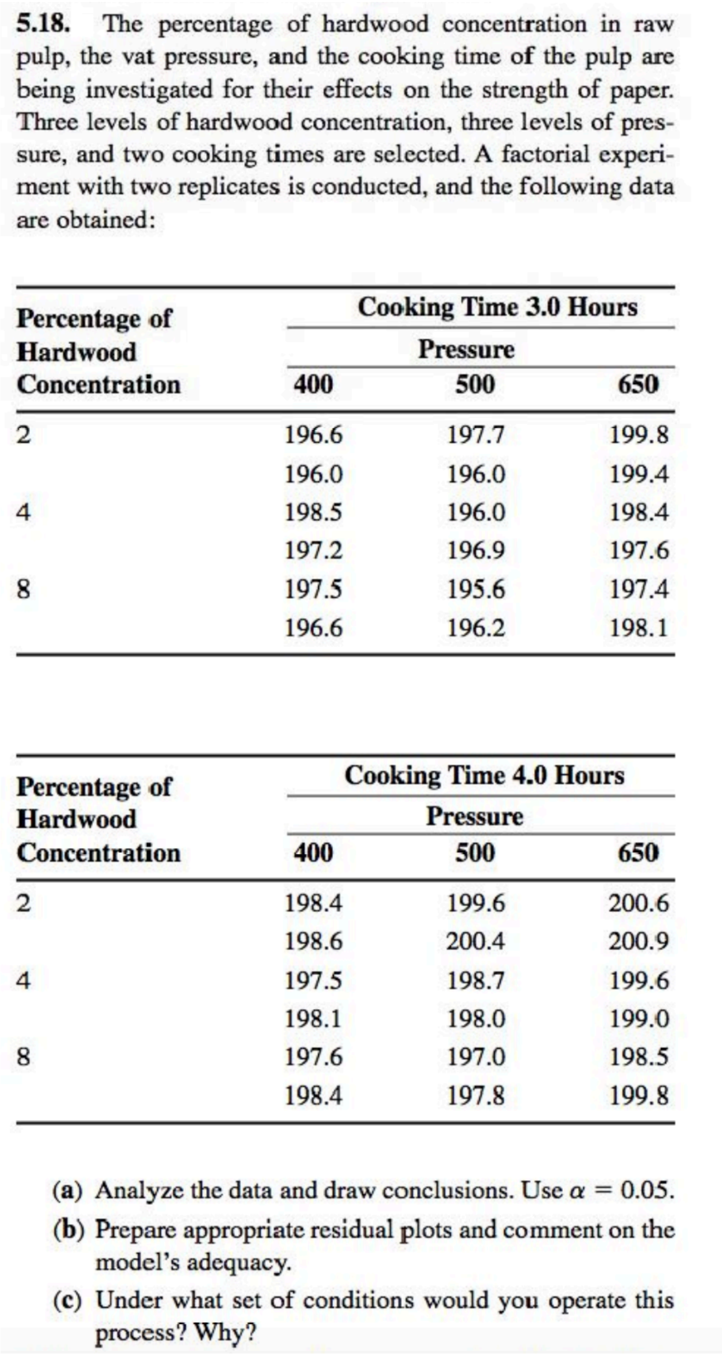
$$F = \frac{(R_p^2 - R_q^2)/(p-q)}{(1 - R_p^2)/[n - (p+1)]}.$$

Suppose that $n = 26, q = 3$ and $p = 5$. Further suppose that $R_p^2 = 0.90$ and $R_q^2 = 0.80$. Test whether the increase in R^2 from the partial model to the full model is statistically significant at the 1% level.

3.9 (Alternate coding of categorical variables) Refer to Example 3.14 and the data in Table 3.8. Suppose that the Gender is coded as $x_1 = -1$ for females and $x_1 = +1$ for males. Similarly, Race is coded as $x_2 = -1$ for non-Whites and $x_2 = +1$ for Whites. What are the new values of $\beta_0, \beta_1, \beta_2$ and β_3? Interpret them.

Table 3.8 Average salaries by gender and race

		Race	
		Non-White	White
Gender	Female	$40K	$50K
	Male	$45K	$65K



5.18. The percentage of hardwood concentration in raw pulp, the vat pressure, and the cooking time of the pulp are being investigated for their effects on the strength of paper. Three levels of hardwood concentration, three levels of pressure, and two cooking times are selected. A factorial experiment with two replicates is conducted, and the following data are obtained:

Percentage of Hardwood Concentration	Cooking Time 3.0 Hours		
	Pressure		
	400	500	650
2	196.6	197.7	199.8
	196.0	196.0	199.4
4	198.5	196.0	198.4
	197.2	196.9	197.6
8	197.5	195.6	197.4
	196.6	196.2	198.1

Percentage of Hardwood Concentration	Cooking Time 4.0 Hours		
	Pressure		
	400	500	650
2	198.4	199.6	200.6
	198.6	200.4	200.9
4	197.5	198.7	199.6
	198.1	198.0	199.0
8	197.6	197.0	198.5
	198.4	197.8	199.8

(a) Analyze the data and draw conclusions. Use $\alpha = 0.05$.

(b) Prepare appropriate residual plots and comment on the model's adequacy.

(c) Under what set of conditions would you operate this process? Why?