STATS 203: HOMEWORK SOLUTION #2

THANKS TO ROHAN TANDON, RONGZHI LU

Question 1

We are given a one-way ANOVA model

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

The mean sum-of-treatment-squares term is

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

and we assume that each group i has n observations, i.e. $n_i = n$.

(a). From the definition of a one-way ANOVA model, we see immediately that

$$\begin{aligned}
\bar{Y}_{i\cdot} &= \frac{1}{n}\sum_{j=1}^{n} Y_{ij} = \frac{1}{n}\sum_{j=1}^{n}\left(\mu + \alpha_i + \epsilon_{ij}\right) = \mu + \alpha_i + \bar{\epsilon}_{i\cdot} \\
\bar{Y}_{\cdot\cdot} &= \frac{1}{r}\sum_{i=1}^{r}\left(\frac{1}{n}\sum_{j=1}^{n} Y_{ij}\right) = \frac{1}{r}\sum_{i=1}^{r}\left[\frac{1}{n}\sum_{j=1}^{n}\left(\mu + \alpha_i + \epsilon_{ij}\right)\right] = \frac{1}{r}\sum_{i=1}^{r}\left(\mu + \alpha_i + \bar{\epsilon}_{i\cdot}\right)
\end{aligned}$$

where $\bar{\epsilon}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n}\epsilon_{ij}$. Then clearly, we see that

$$\begin{aligned}
\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} &= \left(\mu + \alpha_i + \bar{\epsilon}_{i\cdot}\right) - \frac{1}{r}\sum_{i=1}^{r}\left(\mu + \alpha_i + \bar{\epsilon}_{i\cdot}\right) \\
&= \left(\mu + \alpha_i + \bar{\epsilon}_{i\cdot}\right) - \mu - \frac{1}{r}\left(\sum_{i=1}^{r}\alpha_i\right) + \frac{1}{r}\sum_{i=1}^{r}\frac{1}{n}\sum_{j=1}^{n}\epsilon_{ij} \\
&= \left(\mu + \alpha_i + \bar{\epsilon}_{i\cdot}\right) - \mu - \frac{1}{r}\cdot 0 + \frac{1}{nr}\sum_{i=1}^{r}\sum_{j=1}^{n}\epsilon_{ij} \\
&= \mu + \alpha_i + \bar{\epsilon}_{i\cdot} - \mu + \bar{\epsilon}_{\cdot\cdot} \\
&= \alpha_i + \left(\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}\right)
\end{aligned}$$

where we have used the "identifiability" assumption of the one-way ANOVA model that $\sum_{i=1}^{r}\alpha_i = 0$. ■

Date: Submitted on February 4, 2010.

(b). Using the result from part (a), we have that $\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} = \alpha_i + (\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot})$. Hence, it follows directly from the quadratic expansion that

$$\begin{aligned}
\sum_{i=1}^{r} \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 &= \sum_{i=1}^{r} \left[\alpha_i + (\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot})\right]^2 \\
&= \sum_{i=1}^{r} \left[\alpha_i^2 + 2\alpha_i \left(\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}\right) + \left(\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}\right)^2\right] \\
&= \sum_{i=1}^{r} \alpha_i^2 + \sum_{i=1}^{r} 2\alpha_i \left(\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}\right) + \sum_{i=1}^{r} \left(\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}\right)^2 \\
&= \sum_{i=1}^{r} \alpha_i^2 + 2\sum_{i=1}^{r} \alpha_i \left(\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}\right) + \sum_{i=1}^{r} \left(\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}\right)^2
\end{aligned}$$

which proves the desired result. ■

(c). By multiple applications of the property of the linearity of the expectation, we see that

$$\begin{aligned}
E\left[2\sum_{i=1}^{r} \alpha_i \left(\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}\right)\right] &= 2 \cdot E\left[\sum_{i=1}^{r} \alpha_i \left(\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}\right)\right] \\
&= 2 \cdot \sum_{i=1}^{r} E\left[\alpha_i \left(\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}\right)\right] \\
&= 2 \cdot \sum_{i=1}^{r} E\left[\alpha_i \bar{\epsilon}_{i\cdot} - \alpha_i \bar{\epsilon}_{\cdot\cdot}\right] \\
&= 2 \cdot \sum_{i=1}^{r} \left(E\left[\alpha_i \bar{\epsilon}_{i\cdot}\right] - E\left[\alpha_i \bar{\epsilon}_{\cdot\cdot}\right]\right) \\
&= 2 \cdot \sum_{i=1}^{r} \left(E\left[\alpha_i \cdot \frac{1}{n}\sum_{j=1}^{n} \epsilon_{ij}\right] - E\left[\alpha_i \cdot \frac{1}{r}\sum_{i=1}^{r} \frac{1}{n}\sum_{j=1}^{n} \epsilon_{ij}\right]\right) \\
&= 2 \cdot \sum_{i=1}^{r} \left(\alpha_i \cdot \frac{1}{n}\sum_{j=1}^{n} E\left[\epsilon_{ij}\right] - \alpha_i \cdot \frac{1}{r}\sum_{i=1}^{r} \frac{1}{n}\sum_{j=1}^{n} E\left[\epsilon_{ij}\right]\right) \qquad \text{since } \{\alpha\}_{i=1}^{n} \text{ are constant} \\
&= 2 \cdot \sum_{i=1}^{r} \left(\alpha_i \cdot \frac{1}{n}\sum_{j=1}^{n} 0 - \alpha_i \cdot \frac{1}{r}\sum_{i=1}^{r} \frac{1}{n}\sum_{j=1}^{n} 0\right) \qquad \text{since } \epsilon_{ij} \sim N\left(0, \sigma^2\right) \\
&= 2 \cdot \sum_{i=1}^{r} \left(\alpha_i \cdot 0 - \alpha_i \cdot 0\right) = 2 \cdot \sum_{i=1}^{r} 0 = 0
\end{aligned}$$

which proves the desired result. ■

(d). Given that $E\left[\sum_{i=1}^{r}\left(\bar{\epsilon}_{i\cdot}-\bar{\epsilon}_{\cdot\cdot}\right)^2\right]=\frac{(r-1)\sigma^2}{n}$, we can now use parts (a) - (c) to show that by the linearity of the expectation,

$$\begin{aligned}
E\left[MSTR\right] &= E\left[\frac{\sum_{i=1}^{r} n_i\left(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot}\right)^2}{r-1}\right] \\
&= E\left[\frac{n\sum_{i=1}^{r}\left(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot}\right)^2}{r-1}\right] \qquad \text{since } n_i = n\ \forall i \\
&= E\left[\frac{n\sum_{i=1}^{r}\left[\alpha_i^2+\left(\bar{\epsilon}_{i\cdot}-\bar{\epsilon}_{\cdot\cdot}\right)^2\right]}{r-1}\right] \qquad \text{by parts (b) and (c)} \\
&= E\left[\frac{n\sum_{i=1}^{r}\alpha_i^2}{r-1}\right]+\left(\frac{1}{r-1}\right)\cdot E\left[n\sum_{i=1}^{r}\left(\bar{\epsilon}_{i\cdot}-\bar{\epsilon}_{\cdot\cdot}\right)^2\right] \\
&= \frac{n\sum_{i=1}^{r}\alpha_i^2}{r-1}+\left(\frac{1}{r-1}\right)\cdot\frac{n(r-1)\sigma^2}{n} \\
&= \sigma^2+\frac{n\sum_{i=1}^{r}\alpha_i^2}{r-1}
\end{aligned}$$

which proves the desired result. ■

Question 2

Our full model is $Y_{ij}=\mu+\alpha_1 I_{above}+\alpha_2 I_{average}+\alpha_3 I_{below}+\epsilon_{ij}$.

The null hypothesis is $\alpha_2-\alpha 1=\alpha_3-\alpha_2$, which is equivalent to $\alpha_3=2\alpha_2-\alpha_1$.

Therefore our reduced model is

$$Y_{ij}=\mu+\alpha_i(I_{above}-I_{below})+\alpha_2(I_{average}+2I_{below})+\epsilon_{ij}$$

The p value from the F-test is **0.5947**. Therefore we could not reject null hypothesis.

Question 3

(a). To ascertain whether the variable I should be left in the model, we consider an ANOVA test with the following nested models:

$$\begin{aligned}
\text{Full Model: } V &= \beta_0+\beta_1 I+\beta_2 D+\beta_3 W+\beta_4\left(G\cdot I\right)+\beta_5 P+\beta_6 N+\epsilon \\
\text{Reduced Model: } V &= \beta_0+\beta_2 D+\beta_3 W+\beta_5 P+\beta_6 N+\beta_7 G:I+\epsilon
\end{aligned}$$

The null hypothesis of the ANOVA test is that the Reduced Model (RM) is adequate in explaining the response variable, versus the alternative that the Reduced Model is inadequate, favoring instead the Full Model. Thus, we conduct an F-test, and the p value is **0.2591**.

(b). To ascertain whether the interaction variable $(G\cdot I)$ should be left in the model, we consider an ANOVA test with the following nested models:

$$\begin{aligned}
\text{Full Model: } V &= \beta_0+\beta_1 I+\beta_2 D+\beta_3 W+\beta_4\left(G\cdot I\right)+\beta_5 P+\beta_6 N+\epsilon \\
\text{Reduced Model: } V &= \beta_0+\beta_1 I+\beta_2 D+\beta_3 W+\beta_5 P+\beta_6 N+\epsilon
\end{aligned}$$

The null hypothesis of the ANOVA test is that the Reduced Model (RM) is adequate in explaining the response variable, versus the alternative that the Reduced Model is inadequate, favoring instead the Full Model. Thus, we conduct an F-test, with our statistic given by:

$$F = \frac{SSE_{RM} - SSE_{FM}}{SSE_{RM}/(n-p-1)} = 29.932$$

from the R code below, it is clear that $SSE_{RM} = 0.0743$ and $SSE_{FM} = 0.0236$, which has a $F_{(1,16)}$ distribution under the null hypothesis.

Thus, we have $P\left(F_{(1,16)} > 29.932\right) = 8.24 \times 10^{-5} < 0.05$. Hence, we can reject the null hypothesis at the 5% level that the reduced model is sufficient and conclude that the interaction variable $(G \cdot I)$ should remain in the model. ■

Question 4

(a). To test the hypothesis of that the mean sodium content is the same in all brands sold in the metropolitan area, we use the one-way ANOVA random effects model. In this experiment, the "brand" is random. Hence, we are dealing with a model of the form

$$Y_{ij} \sim \mu + \alpha_i + \varepsilon_{ij} \qquad 1 \leq i \leq r,\, 1 \leq j \leq n_i$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$, $\alpha_i \sim N(0, \sigma_\alpha^2)$. Here, μ is the population mean, σ^2 is the measurement variance, and σ_α^2 is the variation in sodium content in beer. Given our experiment design, we can state the null hypothesis as

$$H_0 : \sigma_\alpha^2 = 0 \quad = \quad 0$$

which translates to saying that there is no random effect in the brand of beer. The alternative hypothesis states $H_1 : \sigma_\alpha^2 \neq 0$ which translates to saying that there exists a random effect of the beer brand. Hence, we conduct the one way ANOVA test, yielding an F statistic given by

$$F = \frac{SSE_{RM} - SSE_{FM}}{SSE_{RM}/(n-p-1)} = 238.71$$

from the R code below, it is clear that $SSE_{RM} = 854.53$ and $SSE_{FM} = 30.7$, which has a $F_{(5,42)}$ distribution under the null hypothesis.

Thus, we have $P\left(F_{(5,42)} > 238.71\right) = 2.2 \times 10^{-16} < 0.1$. Hence, we can reject the null hypothesis at the 10% level that the mean sodium content is the same in all brands sold in the metropolitan area.

(b). The estimated mean for sodium content for all brands can be obtained directly from the results of the regression, since $E(\bar{Y}_{\cdot\cdot}) = \mu$. A straightforward R command shows that this estimated value is 17.629. The variance of the mean can be calulated by

$$Var(\bar{Y}_{\cdot\cdot}) = \frac{n\sigma_\alpha^2 + \sigma^2}{rn} = \frac{SSTR}{(r-1)rn}$$

So the 99% confidence interval is given by

$$E(\bar{Y}_{\cdot\cdot}) \pm t_{(r-1,\alpha/2)} \cdot \sqrt{\frac{SSTR}{(r-1)rn}}$$

As seen by the code below, the 99% confidence interval is (10.02076,25.23757).

```
# STATS 203-Assignment #2 code
#(2)
rehab = read.table("http://www-stat.stanford.edu/~nzhang/203_web/Data/Rehab.txt",header=T)
above = as.numeric(rehab$Fitness=="ABOVE")
ave = as.numeric(rehab$Fitness=="AVERAGE")
below = as.numeric(rehab$Fitness=="BELOW")
rehab.lm = lm(rehab$Time~-1+above+ave+below)
rehab.rlm = lm(rehab$Time~-1+I(above-below)+I(ave+2*below))
anova(rehab.lm,rehab.rlm)


#(3.a)
election = read.table("http://www-stat.stanford.edu/~nzhang/203_web/Data/Election.txt",header=T)
attach(election)
I = as.factor(I)
D = as.factor(D)
W = as.factor(W)
elect.lm = lm(V~I+D+W+G:I+P+N)
elect.rlm = lm(V~D+W+G:I+P+N)
anova(elect.lm,elect.rlm)


#(3.b)
elect.rlm2 = lm(V~I+D+W+P+N)
anova(elect.lm,elect.rlm2)



#(4.a)
beer = read.table("http://www-stat.stanford.edu/~nzhang/203_web/Data/Beer.txt",header=T)
attach(beer)
brand = as.factor(beer$brand)
lm = lm(sodium~brand,data = beer)
anova(lm)

#(4.b)
mean(sodium)

beer.anova=anova(lm)
se.sodium = sqrt(beer.anova$Mean[1]/48)
b = qt(0.995,5)*se.sodium
c(mean(sodium)-b,mean(sodium)+b)
```