Stat 541 Homework #1

Kenny Flagg

February 1, 2016

1.
$$
\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2 &= \sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(y_{ij}^2-2y_{ij}\bar{y}_{..}+\bar{y}_{..}^2\right)\\
&= \sum_{i=1}^{a}\left(\sum_{j=1}^{n_i}y_{ij}^2-2\left(\sum_{j=1}^{n_i}y_{ij}\right)\bar{y}_{..}+\sum_{j=1}^{n_i}\bar{y}_{..}^2\right)\\
&= \sum_{i=1}^{a}\left(\sum_{j=1}^{n_i}y_{ij}^2-2y_{i.}\bar{y}_{..}+n_i\bar{y}_{..}^2\right)\\
&= \sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}^2-2\left(\sum_{i=1}^{a}y_{i.}\right)\bar{y}_{..}+\left(\sum_{i=1}^{a}n_i\right)\bar{y}_{..}^2\\
&= \sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}^2-2y_{..}\bar{y}_{..}+N\bar{y}_{..}^2\\
&= \sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}^2-2y_{..}\frac{y_{..}}{N}+N\left(\frac{y_{..}}{N}\right)^2\\
&= \sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}^2-2\frac{y_{..}^2}{N}+\frac{y_{..}^2}{N}\\
&= \sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}^2-\frac{y_{..}^2}{N}
\end{aligned}
$$

2. (a) Under H_0 the test statistic is $F_0 \sim \mathrm{F}(5, 24)$ so for each α we reject H_0 if:

Signifiance Level	Rejection Region
$\alpha = 0.01$	$F_0 \geq F_{0.01}(5, 24) = 3.90$
$\alpha = 0.05$	$F_0 \geq F_{0.05}(5, 24) = 2.62$
$\alpha = 0.10$	$F_0 \geq F_{0.10}(5, 24) = 2.10$

(b) First we need the overall total:

$$y_{..} = \sum_i \sum_j y_{ij} = \sum_i y_{i.} = 50 + 75 + 100 + 90 + 60 + 75 = 450$$

Next, find SS_T and SS_{Trt}:

$$SS_T = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = \sum_i \sum_j y_{ij}^2 - \frac{y_{..}^2}{N} = 7858 - \frac{450^2}{30} = 1108$$

$$SS_{Trt} = \sum_i \sum_j (\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_i \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N} = \frac{50^2}{5} + \frac{75^2}{5} + \frac{100^2}{5} + \frac{90^2}{5} + \frac{60^2}{5} + \frac{75^2}{5} - \frac{450^2}{30} = 340$$

Now we can fill in the ANOVA table:

Source of Variation	Sum Squares	d.f.	Mean Square	F-Ratio
Treatment	340	5	68	$F_0 = 2.125$
Error	768	24	32	
Total	1108	29		

(c) We would reject H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ for $\alpha = 0.10$ only.

(d) With all of the original observations, the treatment means were $\bar{y}_{1.} = 10$, $\bar{y}_{2.} = 15$, $\bar{y}_{3.} = 20$, $\bar{y}_{4.} = 18$, $\bar{y}_{5.} = 12$, and $\bar{y}_{6.} = 15$. Treatment 3 had the largest mean. With the one observation removed, the new mean for treatment 3 is $\bar{y}_{3.} = 16$ which is close to the means for treatments 2 and 6 and is no longer the largest mean. Thus there is less between-treatment variability so MS_{Trt} will decrease.

3. I ran the following SAS code:

```
data catalysts;
  input catalyst concentration @@;
datalines;
1 58.2 1 57.2 1 58.4 1 55.8 1 54.9
2 56.3 2 54.5 2 57.0 2 55.3
3 50.1 3 54.2 3 55.4
4 52.9 4 49.9 4 50.0 4 51.7
;

proc glm data = catalysts;
  class catalyst;
  model concentration = catalyst;
```

Here is the resulting ANOVA table:

```
                          The GLM Procedure

                  Dependent Variable: concentration

                                     Sum of
Source                     DF       Squares   Mean Square   F Value   Pr > F

Model                       3    85.6758333    28.5586111      9.92   0.0014

Error                      12    34.5616667     2.8801389

Corrected Total            15   120.2375000
```

4. (a) Yes, we can compute:

$$SS_E = \sum_i (n_i - 1)s_i^2$$

(b) No, we need the sample variance s^2 for the whole sample so we can compute:

$$SS_{Trt} = \left(\sum_i n_i - 1\right) s^2$$

(c) No, we need s^2 to find SS_{Trt} as above and then we can compute:

$$SS_T = SS_{Trt} + SS_E$$