STA 303H1S / STA 1002 HS: Two-way Analysis of Variance Practice Problems
SOLUTIONS

1. The experimental unit – that which receives the treatment – is the platoon, not the individual soldier.

2. (a) It simply means that there are significant interactions. When there is non-additivity, the best strategy is to avoid talking about main effects.

   (b) If there is no significant interaction, it makes sense to talk about the main effects, that is, how the mean response differs among levels of each factor. If there is no interaction, the differences among levels of one factor are the same, regardless of the level of the other factor.

3. Find the $G \times H$ estimates of θ_{gh} by finding that values that minimize

$$S = \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{i=1}^{n_{gh}} (y_{ghi} - \theta_{gh})^2.$$

   Differentiating with respect to θ_{gh} gives

$$\frac{\partial S}{\partial \theta_{gh}} = -2 \sum_{i=1}^{n_{gh}} (y_{ghi} - \theta_{gh})$$

   and setting these derivatives equal to 0 gives

$$\hat{\theta}_{gh} = \overline{y}_{gh}.$$

4. (a) For i in the first level of factor 1 and the first level of factor 2, $\mathrm{E}(Y_i) = \beta_0 + \beta_1 + \beta_2 + \beta_3$.
   For i in the first level of factor 1 and the second level of factor 2, $\mathrm{E}(Y_i) = \beta_0 + \beta_1$.
   For i in the second level of factor 1 and the first level of factor 2, $\mathrm{E}(Y_i) = \beta_0 + \beta_2$.
   For i in the second level of factor 1 and the second level of factor 2, $\mathrm{E}(Y_i) = \beta_0$.

   (b) From question 3 and part (a),
   $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = \overline{y}_{11}$
   $\hat{\beta}_0 + \hat{\beta}_1 = \overline{y}_{12}$
   $\hat{\beta}_0 + \hat{\beta}_2 = \overline{y}_{21}$
   $\hat{\beta}_0 = \overline{y}_{22}$
   and solving gives the required expressions.

   (c) $\mathrm{Var}(\hat{\beta}_3) = \mathrm{Var}(\overline{Y}_{11} - \overline{Y}_{21} + \overline{Y}_{22} - \overline{Y}_{12}) = \sigma^2 \left(\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} + \frac{1}{n_{12}} \right)$

5. (a) An appropriate model is

$$Y_i = \beta_0 + \beta_1 I_{Treat,i} + \beta_2 I_{A,i} + \beta_3 I_{Treat,i} * I_{A,i} + e_i, \quad i = 1, \ldots, 112$$

   where Y_i is the reduction in depression for the ith patient, $I_{Treat,i}$ is 1 if the ith patient is in the treatment group and 0 if in the control group, $I_{A,i}$ is 1 if the ith patient is in hospital A and 0 if in hospital B, and e_i is the random error component.

In matrix terms, the model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ where $\mathbf{Y}$ is the vector of length 112 of the reductions in depression scores, $\mathbf{e}$ is the vector of length 112 of the random error terms, $\boldsymbol{\beta}$ is the vector of length 4 of the model parameters (excluding σ) and $\mathbf{X}$ is the 112×4 matrix with 1st column all 1's and 2nd column consisting of 1 if the ith row corresponds to an observation from the treatment group and 0 otherwise, 3rd column consisting of 1 if the ith row corresponds to an observation from Hospital A and 0 otherwise, and 4th column consisting of 1 if the ith observation is from both the treatment group and Hospital A and 0 otherwise.

(b) $\hat{\beta}_3 = 8.5 - 4.6 + (-1.7) - 2.3 = -0.1$

(c) The pooled estimate of the variance is

$$s_p^2 = \frac{24 * 1.3^2 + 24 * 0.9^2 + 30 * 1.8^2 + 30 * 1.1^2}{24 + 24 + 30 + 30} = 1.79$$

so the estimated variance of the estimated coefficient of the interaction term is

$$1.79 \left(\frac{1}{25} + \frac{1}{31} + \frac{1}{31} + \frac{1}{25} \right) = 0.26.$$

(d) Testing $H_0 : \beta_3 = 0$ versus $H_a : \beta_3 \neq 0$
Test statistic is $\frac{-0.1}{\sqrt{0.26}} = -0.196$
If $\beta_3 = 0$ this is an observation from a t-distribution with degrees of freedom $24 + 24 + 30 + 30 = 108$. From tables, the p-value is > 0.8 so the data are consistent with the coefficient of the interaction term being 0. That is, there is no evidence that the change in mean depression scores between the treatment and control groups differs with hospital.

6. (a) An appropriate model is

$$\begin{aligned} Y_i \quad = \quad & \beta_0 + \beta_1 I_{NE,i} + \beta_2 I_{NC,i} + \beta_3 I_{S,i} + + \beta_4 I_{age1,i} + \beta_5 I_{age2,i} \\ & + \beta_6 I_{NE,i} * I_{age1,i} + \beta_7 I_{NE,i} * I_{age2,i} + \beta_8 I_{NC,i} * I_{age1,i} + \beta_9 I_{NC,i} * I_{age2,i} \\ & + \beta_{10} I_{S,i} * I_{age1,i} + \beta_{11} I_{S,i} * I_{age2,i} + e_i, \qquad i = 1, \ldots, 113 \end{aligned}$$

where Y_i is the length of stay for the ith patient, $I_{NE,i}$ is 1 if the ith patient is in the North East Region and 0 otherwise, etc., $I_{age1,i}$ is 1 if the ith patient is in the 1st age group and 0 otherwise, etc., and e_i is the random error component.
In matrix terms, the model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ where $\mathbf{Y}$ is the vector of length 113 of the lengths of stay, $\mathbf{e}$ is the vector of length 113 of the random error terms, $\boldsymbol{\beta}$ is the vector of length 12 of the model parameters (excluding σ) and $\mathbf{X}$ is the 113×12 matrix with 1st column all 1's, 2nd, 3rd and 4th columns consisting of 1 if the ith row corresponds to an observation from the North East, North Central, South (respectively) regions and 0 otherwise, 5th and 6th columns consisting of 1 if the ith row corresponds to an observation from age group 1 or 2 (respectively) and 0 otherwise, and 7th through 12th columns consisting of 1 if the ith observation is from both the relevant region and age group and 0 otherwise.

(b) The DF for the model is $3 + 2 + 6 = 11$
The F value is $13.4523927/2.5864759 = 5.201$
From the F table with 10 and 60 degrees of freedom (used to estimate 11 and 101 degrees of freedom), the p-value is $< .001$.

(c) There is some evidence of an age group - region interaction ($p = 0.0256$). So differences in the mean length of stay among age groups differ with regions. (Since there is a significant interaction, it doesn't make sense to talk about the individual effects of age group or region.)

The interaction is evident in the plot of means. The mean length of stay is similar in regions 2 (North Central) and 3 (South) for all age groups. However in region 1 (North East), the oldest age group has the highest mean length of stay and in region 4 (West) the middle age group has the highest mean length of stay.

(d) Variance is increasing with predicted value, so a transformation of `los` (eg., log or square root) is appropriate. Thus, inferences are not accurate. If we were to look at the normal quantile plot without knowing about the increasing variance, we would conclude that the distribution of the residuals has heavier tails than a normal distribution. However, this perceived problem may be a consequence of the non-constant variance.

(e) Since there is a significant interaction, post-hoc tests must compare means in all $3 \times 4 = 12$ combinations of explanatory variables.

The mean length of stay for the highest age group in region 1 (North East) is higher than all other means. Looking at the Tukey-adjusted p-values for the pairwise comparisons of means, the mean length of stay for this group does not differ significantly from the mean length of stay for the other two age groups in region 1 but it is significantly different than the mean length of stay in all other age groups in all other regions (except the evidence is only weak for age group 2 in region 2). There is also evidence of a difference in the means of length of stay between age group 2 in region 1 and age group 3 in region 4. (This last may be of lesser interest since it does not give a direct comparison between an age group or region.)