STA 303H1F: Two-way Analysis of Variance Practice Problems

1. In the Pygmalion example from lecture, why are the average scores of the platoon used as the response variable, rather than the scores of the individual soldiers?

2. In two-way analysis of variance,

   (a) What does it mean when there are significant interactions but no significant main effects? ("Main effects" are the effects of the factors considered on their own.)

   (b) What does it mean when there are significant main effects but no significant interaction?

3. Two-way tables with G levels of one factor and H levels of the second factor can be anlayzed using one-way analysis of variance with a factor with $G \times H$ levels. Let Y_{ghi} denote the response of the i^{th} observation in the g^{th} group of the first factor and h^{th} group of the second factor, with

$$\mathrm{E}(Y_{ghi}) = \theta_{gh}$$

   for $g = 1, \ldots, G$, $h = 1, \ldots, H$, and $i = 1, \ldots, n_{gh}$ where n_{gh} is the number of observations in the g^{th} level of the first factor and the h^{th} level of the second factor. The least squares solutions can found by minimizing

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

   with respect to θ_{gh} for $g = 1, \ldots, G$ and $h = 1, \ldots, H$.

   Show that the least squares solutions is

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

   where

$$\overline{y}_{gh} = \frac{1}{n_{gh}}\sum_{i=1}^{n_{gh}} y_{ghi}.$$

4. Consider the model for a two-way analysis of variance with two levels of each factor (a 2×2 classification

$$Y_i = \beta_0 + \beta_1 I_{\text{factor 1},i} + \beta_2 I_{\text{factor 2},i} + \beta_3 I_{\text{factor 1},i} I_{\text{factor 2},i} + e_i$$

   where $I_{\text{factor 1},i} = 1$ if the i^{th} observation is in the first group of factor 1 and is 0 otherwise.

(a) What are the expected values of Y_i for each of the 4 groups means?

(b) Use the result of question 3 to show that the least squares estimate of the coefficients are

$$\begin{aligned}
b_0 &= \overline{y}_{22} \\
b_1 &= \overline{y}_{12} - \overline{y}_{22} \\
b_2 &= \overline{y}_{21} - \overline{y}_{22} \\
b_3 &= \overline{y}_{11} - \overline{y}_{21} + \overline{y}_{22} - \overline{y}_{12}
\end{aligned}$$

where $\overline{y}_{mn}$ is the mean of observations for the m^{th} level of factor 1 and the n^{th} level of factor 2.

(c) Under the assumption that the Y's are uncorrelated with variance σ^2, what is the variance of b_3?

5. (The scenario for this question is taken from Kleinbaum *et al.* Chapter 20, Question 7.)
The effect of a new antidepressant drug on reducing the severity of depression was studied in manic-depressive patients at two state mental hospitals. In each hospital all such patients were randomly assigned to either a treatment (new drug) or a control (old drug) group. The results of this experiment are summarized in the following table; a high mean score indicates more lowering in depression level than does a low mean score.

	Group	
Hospital	Treatment	Control
A	$n = 25,\ \overline{y} = 8.5,\ s = 1.3$	$n = 31,\ \overline{y} = 4.6,\ s = 1.8$
B	$n = 25,\ \overline{y} = 2.3,\ s = 0.9$	$n = 31,\ \overline{y} = -1.7,\ s = 1.1$

(a) Write an appropriate linear model for analysing these data, both with and without the use of matrices.

(b) Use the results of question 4 to find a numeric value for the coefficient of the interaction term.

(c) Estimate the variance of the coefficient of the interaction term.

(d) Test the hypothesis of no interaction.

6. The data for this question were taken from the appendix of Kutner *et al.* (the SENIC data). The dependent variable is length of stay (variable name `los` in output below) in hospital for patients. In this

question the effects of geographic region (variable name `region`, 4 categories where 1=North East, 2=North Central, 3=South, and 4=West) and age of patient are to be studied. For this question, age has been classified into three categories (variable name `agegroup` where 1=under 52.0 years, 2=52.0 - under 55.0 years, 3=55.0 years or more).

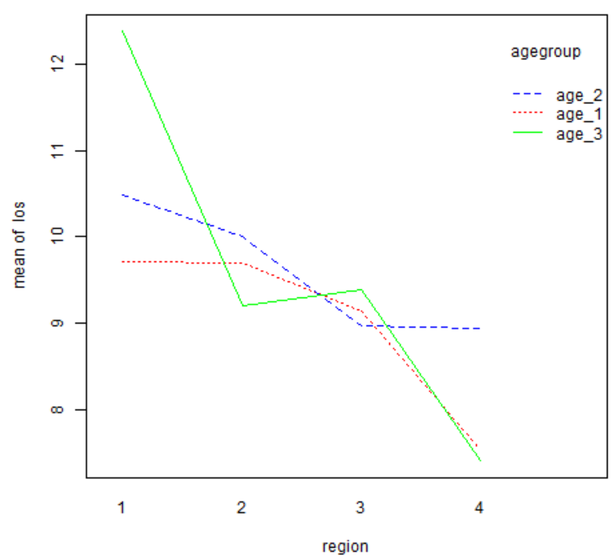
(a) Write the linear model including interactions for analysing these data, both with and without the use of matrices, using indicator variables coded as 0 or 1.

(b) In the R output that follows, complete the ANOVA table (some numbers have been replaced with `X`'s).

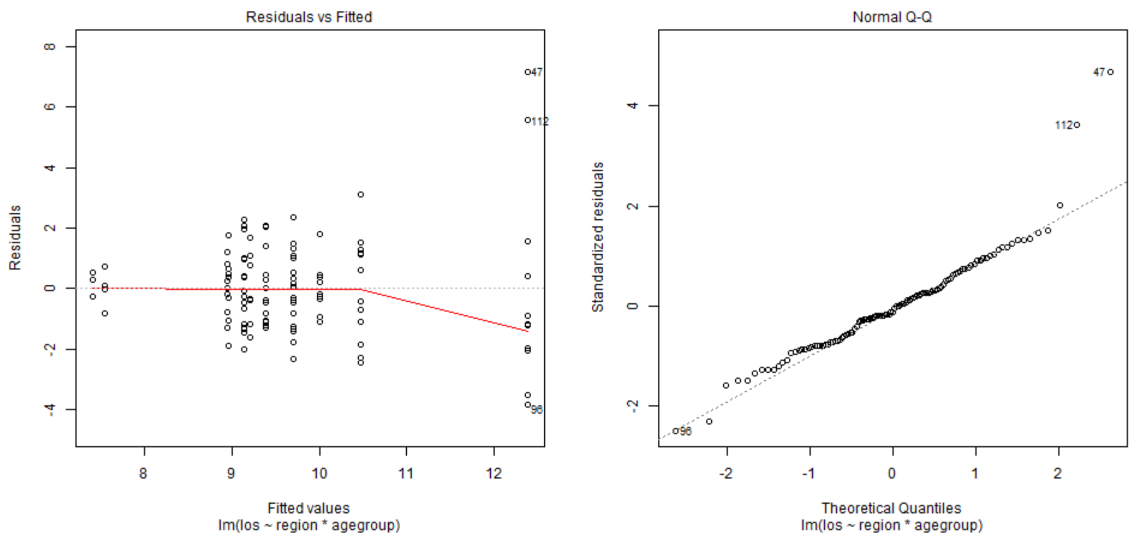
```
> with(senic, tapply(los, list(region, agegroup), mean))
     age_1     age_2     age_3
1 9.710000 10.479167 12.380909
2 9.705625 10.012222  9.210000
3 9.135882  8.967143  9.384615
4 7.540000  8.945714  7.408000
> with(senic, tapply(los, list(region, agegroup), sd))
      age_1     age_2     age_3
1 0.8177714 1.7396993 3.5231732
2 1.3338464 0.8604763 1.2217337
3 1.3074118 1.1992458 1.1955934
4 0.6494613 0.8753448 0.3803551
> with(senic, tapply(los, list(region, agegroup), length))
  age_1 age_2 age_3
1     5    12    11
2    16     9     7
3    17     7    13
4     4     7     5


> fit <- lm(los ~ region*agegroup, data= senic)
> anova(fit)
Analysis of Variance Table

Response: los
                 Df  Sum Sq Mean Sq F value    Pr(>F)
region            3 103.554  34.518 13.3456 2.095e-07
agegroup          2   5.246   2.623  1.0142    0.3664
region:agegroup   6  39.176   6.529  2.5244    0.0256
Residuals       101 261.234   2.586
> summary(fit)
```

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            9.710000   0.719232  13.501  < 2e-16
region2               -0.004375   0.823984  -0.005  0.99577
region3               -0.574118   0.818194  -0.702  0.48449
region4               -2.170000   1.078849  -2.011  0.04695
agegroupage_2          0.769167   0.856058   0.898  0.37106
agegroupage_3          2.670909   0.867427   3.079  0.00267
region2:agegroupage_2 -0.462569   1.087141  -0.425  0.67138
region3:agegroupage_2 -0.937906   1.120034  -0.837  0.40435
region4:agegroupage_2  0.636548   1.322479   0.481  0.63132
region2:agegroupage_3 -3.166534   1.132952  -2.795  0.00621
region3:agegroupage_3 -2.422176   1.050493  -2.306  0.02317
region4:agegroupage_3 -2.802909   1.384321  -2.025  0.04553

Residual standard error: 1.608 on 101 degrees of freedom
Multiple R-squared:  0.3616,Adjusted R-squared:  0.2921
F-statistic: (XX) on (XX) and 101 DF,  p-value: (XX)
```

(c) What do you conclude? Is your conclusion consistent with the plot of means below?



(d) Below are plots of the residuals versus predicted values and a normal quantile plot of the residuals. What do you conclude from them?


Residuals vs Fitted
Residuals
Fitted values
lm(los ~ region * agegroup)
47
112
96
Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(los ~ region * agegroup)
47
112
96