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 are are significant main effects but no significant interaction?

3. A two-way analysis of variance model with G levels of one factor and H levels of the second factor can be thought of as a one-way analysis of variance with a factor with $G \times H$ levels. Let Y_{ghi} denote the response of the ith observation in the gth group of the first factor and hth 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 gth level of the first factor and the hth 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 ith 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 mth level of factor 1 and the nth 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 SAS output that follows, complete the ANOVA table (some numbers have been replaced with `X`'s).

```
 ----------------------------------------------------------
|                 |                  los                   |
|                 |----------------------------------------|
|                 |    Mean    |    Std     |     N        |
|-----------------+------------+------------+------------|
|region  |agegroup|            |            |            |
|--------+--------|            |            |            |
|1       |1       |        9.71|        0.82|        5.00|
|        |--------+------------+------------+------------|
|        |2       |       10.48|        1.74|       12.00|
|        |--------+------------+------------+------------|
|        |3       |       12.38|        3.52|       11.00|
|--------+--------+------------+------------+------------|
```

2	1		9.71	1.33	16.00
	2		10.01	0.86	9.00
	3		9.21	1.22	7.00
3	1		9.14	1.31	17.00
	2		8.97	1.20	7.00
	3		9.38	1.20	13.00
4	1		7.54	0.65	4.00
	2		8.95	0.88	7.00
	3		7.41	0.38	5.00

The GLM Procedure

Class Level Information

Class	Levels	Values
region	4	1 2 3 4
agegroup	3	1 2 3

Number of Observations Read	113
Number of Observations Used	113

The GLM Procedure

Dependent Variable: los

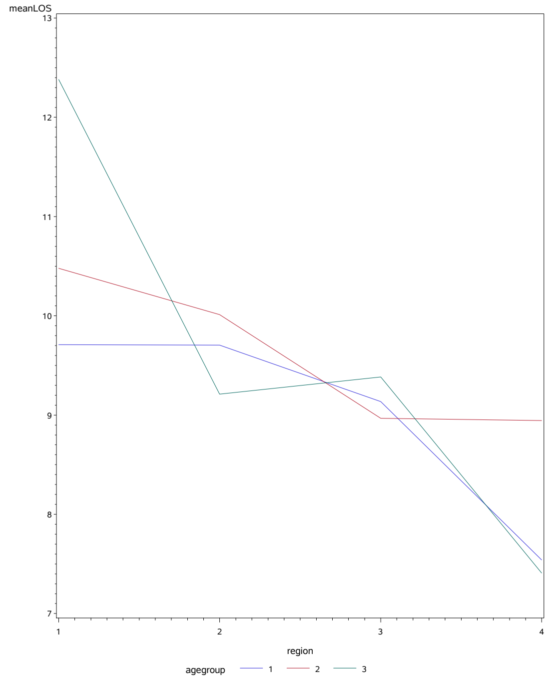
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	XX	147.9763195	13.4523927	XXXX	XXXXXX
Error	101	261.2340610	2.5864759		
Corrected Total	112	409.2103805			

R-Square	Coeff Var	Root MSE	los Mean
0.361614	16.66873	1.608252	9.648319

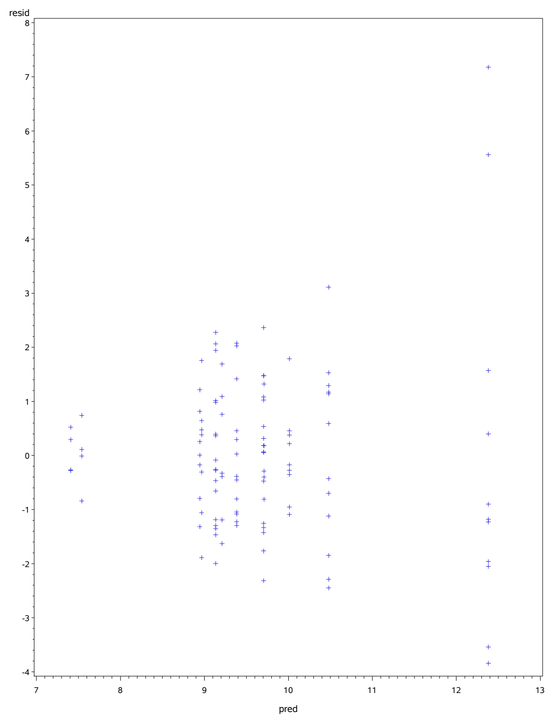
Source	DF	Type I SS	Mean Square	F Value	Pr > F
region	3	103.5541834	34.5180611	13.35	<.0001
agegroup	2	5.2461547	2.6230774	1.01	0.3664
region*agegroup	6	39.1759815	6.5293302	2.52	0.0256

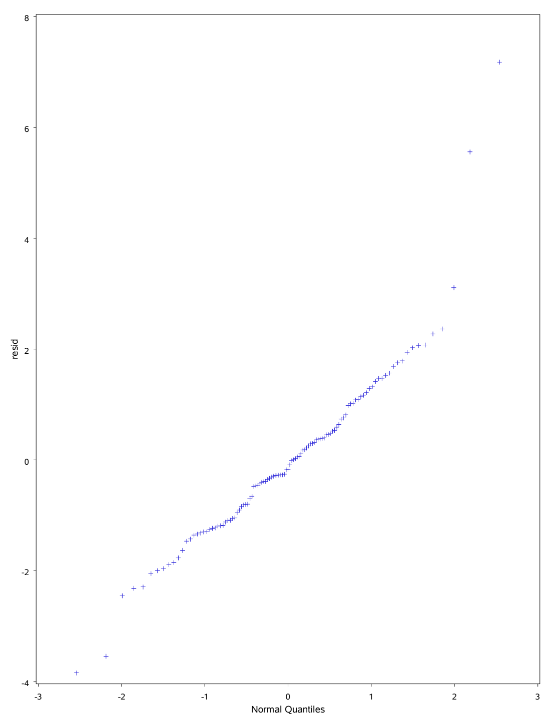
Source	DF	Type III SS	Mean Square	F Value	Pr > F
region	3	84.24919035	28.08306345	10.86	<.0001
agegroup	2	6.47626605	3.23813302	1.25	0.2903
region*agegroup	6	39.17598146	6.52933024	2.52	0.0256

(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?





(e) Below is output from the `lsmeans` statement of `proc glm`. Why are means given for 12 groups rather than 3 or 4? What do you conclude?

The GLM Procedure
Least Squares Means

region	agegroup	los LSMEAN	LSMEAN Number
1	1	9.7100000	1
1	2	10.4791667	2
1	3	12.3809091	3
2	1	9.7056250	4
2	2	10.0122222	5
2	3	9.2100000	6
3	1	9.1358824	7
3	2	8.9671429	8
3	3	9.3846154	9
4	1	7.5400000	10
4	2	8.9457143	11
4	3	7.4080000	12

Least Squares Means for effect region*agegroup
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: los

i/j	1	2	3	4	5	6
1		0.3711	0.0027	0.9958	0.7369	0.5966
2	0.3711		0.0056	0.2107	0.5118	0.1002
3	0.0027	0.0056		<.0001	0.0014	<.0001
4	0.9958	0.2107	<.0001		0.6483	0.4980
5	0.7369	0.5118	0.0014	0.6483		0.3246
6	0.5966	0.1002	<.0001	0.4980	0.3246	
7	0.4845	0.0290	<.0001	0.3115	0.1892	0.9185
8	0.4320	0.0508	<.0001	0.3133	0.2002	0.7781
9	0.7014	0.0922	<.0001	0.5941	0.3703	0.8173
10	0.0469	0.0020	<.0001	0.0178	0.0120	0.1007

Least Squares Means for effect region*agegroup
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: los

i/j	7	8	9	10	11	12
1	0.4845	0.4320	0.7014	0.0469	0.4189	0.0258
2	0.0290	0.0508	0.0922	0.0020	0.0477	0.0005
3	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
4	0.3115	0.3133	0.5941	0.0178	0.2996	0.0063
5	0.1892	0.2002	0.3703	0.0120	0.1912	0.0045
6	0.9185	0.7781	0.8173	0.1007	0.7591	0.0585
7		0.8157	0.6755	0.0772	0.7929	0.0372
8	0.8157		0.5810	0.1599	0.9802	0.1009
9	0.6755	0.5810		0.0475	0.5618	0.0215
10	0.0772	0.1599	0.0475		0.1662	0.9029

Least Squares Means for effect region*agegroup
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: los

i/j	1	2	3	4	5	6
11	0.4189	0.0477	<.0001	0.2996	0.1912	0.7591
12	0.0258	0.0005	<.0001	0.0063	0.0045	0.0585

Least Squares Means for effect region*agegroup
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: los

i/j	7	8	9	10	11	12
11	0.7929	0.9802	0.5618	0.1662		0.1056
12	0.0372	0.1009	0.0215	0.9029	0.1056	

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Least Squares Means for effect region*agegroup
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: los

i/j	1	2	3	4	5	6
1		0.9990	0.1019	1.0000	1.0000	1.0000
2	0.9990		0.1825	0.9822	1.0000	0.8819
3	0.1019	0.1825		0.0027	0.0606	0.0049
4	1.0000	0.9822	0.0027		1.0000	0.9999
5	1.0000	1.0000	0.0606	1.0000		0.9976
6	1.0000	0.8819	0.0049	0.9999	0.9976	
7	0.9999	0.5428	<.0001	0.9970	0.9743	1.0000
8	0.9997	0.7076	0.0016	0.9971	0.9787	1.0000
9	1.0000	0.8640	0.0009	1.0000	0.9990	1.0000
10	0.6847	0.0817	<.0001	0.4100	0.3176	0.8830

Least Squares Means for effect region*agegroup
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: los

i/j	7	8	9	10	11	12
1	0.9999	0.9997	1.0000	0.6847	0.9996	0.5091
2	0.5428	0.7076	0.8640	0.0817	0.6891	0.0246
3	<.0001	0.0016	0.0009	<.0001	0.0014	<.0001
4	0.9970	0.9971	1.0000	0.4100	0.9962	0.2010

5	0.9743	0.9787	0.9990	0.3176	0.9752	0.1557
6	1.0000	1.0000	1.0000	0.8830	1.0000	0.7480
7		1.0000	1.0000	0.8219	1.0000	0.6157
8	1.0000		1.0000	0.9578	1.0000	0.8834
9	1.0000	1.0000		0.6883	1.0000	0.4591
10	0.8219	0.9578	0.6883		0.9621	1.0000

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Least Squares Means for effect region*agegroup
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: los

i/j	1	2	3	4	5	6
11	0.9996	0.6891	0.0014	0.9962	0.9752	1.0000
12	0.5091	0.0246	<.0001	0.2010	0.1557	0.7480

Least Squares Means for effect region*agegroup
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: los

i/j	7	8	9	10	11	12
11	1.0000	1.0000	1.0000	0.9621		0.8926
12	0.6157	0.8834	0.4591	1.0000	0.8926	