STAT 510 Homework 9
Due Date: 11:00 A.M., Wednesday, April 10

1. An experiment was conducted to assess the effect of a virus infection on two plant genotypes (labeled $G1$ and $G2$). Plants were grown in a growth chamber with one plant per pot. A total of 18 pots – 6 containing plants of genotype $G1$ and 12 containing plants of genotype $G2$ – were arranged in the growth chamber using a completely randomized design. On each plant, one leaf was randomly selected for infection with the virus, and another leaf was randomly selected for infection with a control substance. One week after infection, a device was used to measure the color of each leaf. The measurement device returned a continuous score, where high values of the score are associated with healthy, dark green leaves and low values are associated with pale, unhealthy leaves. Let y_{ijk} be the score for genotype Gi ($i = 1, 2$), infection j ($j = 1$ for control and $j = 2$ for virus), and plant k ($k = 1, \ldots, 6$ for genotype $G1$ and $k = 7, \ldots, 18$ for genotype $G2$). Suppose

$$y_{ijk} = \mu_{ij} + p_k + e_{ijk},$$

where $\mu_{11}, \mu_{12}, \mu_{21}$, and μ_{22} are unknown parameters, $p_k \sim N(0, \sigma_p^2)$ for all k, $e_{ijk} \sim N(0, \sigma_e^2)$ for all i, j, k, all random effects and errors are mutually independent, and σ_p^2 and σ_e^2 are unknown variance components. `R` code and output are provided after parts (a) through (e) below. Answer parts (a) through (e) using whatever parts of the `R` code and output you judge to be useful.

   (a) Provide the value of a test statistic that can be used to test for a genotype main effect.

   (b) Provide the value of a test statsitic that can be used to test for an infection main effect.

   (c) Provide the value of a test statistic that can be used to test for genotype $\times$ infection interaction.

   (d) Estimate σ_e^2.

   (e) Estimate σ_p^2.

```
> #The data are stored in a data frame d.
> #The columns labeled Control and Virus give the response
> #for the leaf infected with the control substance and
> #virus, respectively.
>
> d
   Plant Genotype Control Virus
1      1       G1    96.7  88.8
2      2       G1    90.6  79.1
3      3       G1    84.7  75.8
4      4       G1    92.7  81.0
5      5       G1    91.1  83.2
6      6       G1    78.3  76.7
7      7       G2    81.6  76.6
8      8       G2    77.8  87.0
9      9       G2    89.6  81.5
10    10       G2    93.8  85.5
11    11       G2    84.7  87.4
12    12       G2    87.1  77.7
13    13       G2    72.7  68.6
```

```
14    14       G2    79.1  80.2
15    15       G2    77.6  81.7
16    16       G2    72.2  74.9
17    17       G2    74.8  81.9
18    18       G2    83.4  73.5
> y=as.vector(t(cbind(d$Control,d$Virus)))
> geno=factor(rep(1:2,c(12,24)))
> infection=factor(rep(1:2,18))
> y
 [1] 96.7 88.8 90.6 79.1 84.7 75.8 92.7 81.0 91.1 83.2 78.3 76.7 81.6 76.6 77.8
[16] 87.0 89.6 81.5 93.8 85.5 84.7 87.4 87.1 77.7 72.7 68.6 79.1 80.2 77.6 81.7
[31] 72.2 74.9 74.8 81.9 83.4 73.5
> geno
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
Levels: 1 2
> infection
 [1] 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
Levels: 1 2
>
> anova(lm(y~geno+infection+geno:infection))
Analysis of Variance Table

Response: y
              Df  Sum Sq Mean Sq F value  Pr(>F)
geno           1  157.53  157.53  4.2404 0.04769 *
infection      1  126.19  126.19  3.3967 0.07461 .
geno:infection 1   91.35   91.35  2.4590 0.12669
Residuals     32 1188.79   37.15
>
> avg=(d$Control+d$Virus)/2
> summary(lm(avg~0+d$Genotype))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
d$GenotypeG1   84.892      2.169   39.14   <2e-16 ***
d$GenotypeG2   80.454      1.534   52.46   <2e-16 ***

Residual standard error: 5.313 on 16 degrees of freedom
Multiple R-squared:  0.9963,    Adjusted R-squared:  0.9958
F-statistic:  2142 on 2 and 16 DF,  p-value: < 2.2e-16

> diff=d$Control-d$Virus
> summary(lm(diff~1))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.744      1.569   2.386   0.0289 *
```

```
Residual standard error: 6.658 on 17 degrees of freedom

> summary(lm(diff~0+d$Genotype))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
d$GenotypeG1    8.250      2.439   3.383   0.0038 **
d$GenotypeG2    1.492      1.724   0.865   0.3998

Residual standard error: 5.974 on 16 degrees of freedom
Multiple R-squared:  0.4325,    Adjusted R-squared:  0.3615
F-statistic: 6.096 on 2 and 16 DF,  p-value: 0.01076
```

2. Researchers created a device to test the effectiveness of helmets at reducing the stress caused by head impacts. The device includes a head-shaped sensor on which a helmet can be placed, as well as a striking weight that can produce impacts to the front or side of a helmet placed on the sensor. The intensity of each impact can be controlled by the researchers. When an impact is delivered, a measurement of the amount of stress experienced by the head-shaped sensor is recorded. A measurement of 0 indicates no stress, while a measurement of 100 indicates stress high enough to cause serious brain injury.

   The researchers used the device to test a total of 10 helmets consisting of 5 helmets of type 1 and 5 helmets of type 2. The 10 helmets were tested in random order. When each helmet was tested, it was struck a total of 4 times: once with low impact to the front, once with high impact to the front, once with low impact to the side, and once with high impact to the side. The order of the 4 impacts was determined separately for each helmet using the following procedure. A fair coin was flipped. If the result of the flip was heads, the first two impacts were front impacts and the last two impacts were side impacts. If the result of the flip was tails, the first two impacts were side impacts and the last two impacts were front impacts. For the first two impacts, the coin was flipped again. If the result of the flip was heads, the first impact was at low intensity and the second was at high intensity. If the result of the flip was tails, the first impact was at high intensity and the second at low intensity. A coin was flipped a third time to determine the order of the impact intensities for the third and fourth impacts so that each order (low and then high vs. high and then low) was equally likely.

   Let $i = 1, 2$ index helmet types 1 and 2. Let $j = 1, \ldots, 5$ index helmets nested within helmet types. Let $k = 1, 2$ index the direction of impact, with $k = 1$ for front and $k = 2$ for side. Let $\ell = 1, 2$ index the intensity of the impact, with $\ell = 1$ for low and $\ell = 2$ for high. Let $y_{ijk\ell}$ be the stress measurement for the corresponding values of i, j, k, and ℓ. For $i = 1, 2$, $j = 1, \ldots, 5$, $k = 1, 2$, and $\ell = 1, 2$, consider the model

$$y_{ijk\ell} = \mu_{ik\ell} + a_{ij} + b_{ijk} + e_{ijk\ell}, \tag{1}$$

   where the $\mu_{ik\ell}$ values are unknown parameters, $a_{ij} \sim N(0, \sigma_a^2)$, $b_{ijk} \sim N(0, \sigma_b^2)$, $e_{ijk\ell} \sim N(0, \sigma_e^2)$, and all random terms are independent. Model (1) was fit to the dataset, and the following ANOVA table was obtained. Because we have a balanced experimental design, the type I and type III sums of squares are the same, and the lines of the ANOVA table can be reordered in a variety of ways without changing the results.

Source	Sum of Squares	Expected Mean Square
Type	226	
Direction	255	
Intensity	8910	
Type×Direction	207	
Type×Intensity	2	
Direction× Intensity	7	
Type×Direction×Intensity	9	
Helmet(Type)	254	$4\sigma_a^2 + 2\sigma_b^2 + \sigma_e^2$
Direction×Helmet(Type)	114	$2\sigma_b^2 + \sigma_e^2$
Error	59	σ_e^2
C. Total	10043	

(a) We learned a shortcut for expressing sums of squares in summation notation that works for balanced designs like the one considered here. Use that shortcut to express the sum of squares for Direction × Intensity using summation notation.

(b) Compute a t statistic that can be used to test $H_0 : \bar{\mu}_{1\cdot\cdot} = \bar{\mu}_{2\cdot\cdot}$

(c) The statistic in part (b) has a noncentral t distribution. Provide an expression for the noncentrality parameter in terms of model (1) parameters.

(d) Compute the value of an unbiased estimator for σ_a^2

(e) The best linear unbiased estimator of $\bar{\mu}_{12\cdot} - \bar{\mu}_{11\cdot}$ is equal to 0.5 for this dataset. Provide a 95% confidence interval for $\bar{\mu}_{12\cdot} - \bar{\mu}_{11\cdot}$

(f) Compute a standard error for the best linear unbiased estimator of $\mu_{121} - \mu_{111}$

3. Researchers were interested in studying the effect of two drugs on weight gain in pigs. Researchers set up a full-factorial design with two factors: *drug* (drug 1 vs. drug 2) and *dose* (0 vs. 10 milligrams of drug per kilogram of pig body weight at the time of treatment). Originally, eight pigs were used in the experiment with two pigs assigned to each of the four combinations of *drug* and *dose* using a completely randomized design. However, one pig had to be removed from the experiment for reasons unrelated to the experiment. The data are provided in the following `R` code and output. There is one row for each pig in the `R data.frame d`. The variables `drug` and `dose` in `d` are factors, and `y` in `d` is weight gain in kilograms over the period of study. These weight gains are presented as integers to make calculations easier. As indicated in the last line of the `R` code, the researchers fit a cell-means model to this dataset with one unrestricted mean for each of the four combinations of the levels of the factors *drug* and *dose*.

```
> d
  drug dose  y
1    1    0  6
2    1    0  2
3    1   10 12
4    1   10  6
5    2    0  4
6    2   10 16
7    2   10 10
> o = lm(y ~  drug + dose + drug:dose, data = d)
```

(a) Without using a computer, find the Type II sum of squares for `drug`.

(b) Without using a computer, find the Type III sum of squares for `drug`.

4. A plant scientist was interested in comparing two plant genotypes (1 vs. 2) grown in soil of varying moisture levels (1 = low vs. 2 = medium vs. 3 = high). An experiment was conducted in a greenhouse with one table, eight trays, and 24 pots of soil. The table in the greenhouse held the eight trays with three pots on each tray. In each pot, one seed of genotype 1 and one seed of genotype 2 were planted. The three soil moisture levels were assigned, completely at random, to the three pots in each tray. The response of interest is a quantitative measurement of overall plant health that was calculated for each plant 42 days after planting. For $i = 1, 2, 3$, $j = 1, 2$, and $k = 1, \ldots, 8$, let y_{ijk} be the response for soil moisture level i, genotype j, and tray k. Suppose

$$y_{ijk} = \mu_{ij} + t_k + p_{ik} + e_{ijk}, \tag{2}$$

where μ_{ij} is an unknown real-valued parameter, $t_k \sim N(0, \sigma_t^2)$, $p_{ik} \sim N(0, \sigma_p^2)$, $e_{ijk} \sim N(0, \sigma_e^2)$, and all the random effects and errors are independent. Table 1 contains the values of $\bar{y}_{ij.}$ for $i = 1, 2, 3$ and $j = 1, 2$.

Table 1. $\bar{y}_{ij.}$ for $i = 1, 2, 3$ and $j = 1, 2$.

	Genotype 1	Genotype 2
Soil Moisture Level 1	4.2	5.6
Soil Moisture Level 2	6.3	6.1
Soil Moisture Level 3	9.4	8.8

The researchers obtained an ANOVA table that included one line for each of the factors *soil moisture*, *genotype*, and *tray* and all possible interaction among these factors. Mean squares and expected mean squares are provided in Table 2 for a subset of the lines in the ANOVA table obtained by the researchers.

Table 2. Mean Squares and Expected Mean Square for Selected Lines of the ANOVA Table

Source	Mean Square	Expected Mean Square
Tray	17.9	$6\sigma_t^2 + 2\sigma_p^2 + \sigma_e^2$
Tray × Soil Moisture	5.3	$2\sigma_p^2 + \sigma_e^2$
Tray × Genotype	3.3	σ_e^2
Tray × Soil Moisture × Genotype	3.9	σ_e^2

(a) Determine the value of the test statistic you would use to test $H_0 : \bar{\mu}_{2.} = \bar{\mu}_{3.}$.

(b) The low, medium, and high soil moisture levels were maintained by adding either 2, 10, or 18 ounces of water per week to the soil of each pot (i.e., 2 ounces for pots with low soil moisture level, 10 ounces for pots with medium soil moisture level, and 18 ounces for pots with high soil moisture level). Determine the degrees of freedom for the test statistic you would use to test whether the genotype 1 means (μ_{11}, μ_{21}, and μ_{31}) are linear in the amount of water added to the soil. In other words, determine the degrees of freedom for the test statistic you would use to test the null hypothesis that says the points $(2, \mu_{11})$, $(10, \mu_{21})$, and $(18, \mu_{31})$ all fall on a single line.