STAT 510 Homework 9
Due Date: 11:00 A.M., Wednesday, March 29

1. This question refers to slide set 16 on simulating data from a split plot experiment. The R code presented in that slide set is available at

   https://dnett.github.io/S510/16SplitPlot.R

   Questions (a) through (e) below are about the true values of the parameters used to simulate the data. They are not questions about the data and do not require any data analysis.

   (a) Make a table that shows the true expected value of the response for each combination of genotype and fertilizer that was used to simulate the data.

   (b) Is the null hypothesis of no genotype main effects true based on the simulation settings?

   (c) Is the null hypothesis of no fertilizer main effects true based on the simulation settings?

   (d) Is the null hypothesis of no genotype $\times$ fertilizer interactions true based on the simulation settings?

   (e) According to the model used to simulate the data, the mean response is a quadratic function of the amount of fertilizer for each genotype. Give the quadratic equation for each genotype and plot these three quadratic functions on a single plot.

   For the remainder of this problem, base all your answers on the fit of the model given in slides 26 through 30 of slide set 14. This is the model that can be fit in R using the `lme` code on slide 12 of slide set 16 or using the `lmer` code on slide 24 of slide set 16.

   (f) Use the simulated data to compute a confidence interval with approximate coverage 95% for the difference between the treatment mean for genotype 1 and fertilizer 1 and the treatment mean for genotype 1 and fertilizer 2.

   (g) Does the confidence interval computed in part (f) contain the true value of the difference between the treatment mean for genotype 1 and fertilizer 1 and the treatment mean for genotype 1 and fertilizer 2?

   (h) Use the simulated data to compute a confidence interval with approximate coverage 95% for the difference between the treatment mean for genotype 1 and fertilizer 1 and the treatment mean for genotype 2 and fertilizer 1.

   (i) Does the confidence interval computed in part (h) contain the true value of the difference between the treatment mean for genotype 1 and fertilizer 1 and the treatment mean for genotype 2 and fertilizer 1?

   (j) The first line of the ANOVA table on Slide 14 of Slide Set 16 is labeled "(Intercept)". Presumably, this is supposed to be a test of the null hypothesis that says the intercept parameter is zero. The denominator degrees of freedom are given by R as 27, which is the split-plot error degrees of freedom. I don't think this makes sense. To see why, determine an appropriate standard error for the intercept estimate and find its degrees of freedom.

2. Reconsider Homework 8 Problem 4. Suppose data from that experiment have been collected and stored in the vector `y` in `R`. Suppose `GH`, `WL`, and `GENO` are factors in `R` corresponding to the experimental factors greenhouse, watering level, and genotype, respectively. Consider the follow `R` commands and output.

```
o=lm(y~GH*WL*GENO)
anova(o)

Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq
GH          3  113.3    37.8
WL          2  321.8   160.9
GENO        1    2.5     2.5
GH:WL       6  116.4    19.4
GH:GENO     3   11.7     3.9
WL:GENO     2   75.1    37.5
GH:WL:GENO  6   14.5     2.4
```

Suppose our goal is to analyze the data under the assumption that MODEL 1 is correct. Use information in the output to compute three F-statistics that can be used to test for

(a) watering level main effects,

(b) genotype main effects, and

(c) watering level by genotype interactions, respectively.

3. 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
```