BTRY 6020 Homework VIII

NAME: ANDRES CASTANO

NETID: AC986

DUE DATE: May 10, by 8:40 am

Question 1.

A study was conducted to investigate the effect of fertilization on the yield of a commercial variety of tomato. In this study, a completely randomized design was used as the experimental plot was homogeneous both in humidity and nutrient content. The treatments of interest in this study consisted of a control (E=no fertilizer used), very slow release (A), a slow release fertilizer (B), a moderate release fertilizer (C) and a fast release fertilizer (D). These treatment levels were each assigned at random to 20 plots with 15 plants per plot. Prior to planting, the researchers noticed big differences in plant heights. There was no record of when the seeds were planted, so nobody knew if the plants were the same age or not. They therefore decided to record information on mean plant height on each plot. The response variable of interest was the total weight of tomatoes harvested per plot. Data for this study can be downloaded from the course web site. The file name is Hwk8Q1DatSp17.

Answer the following questions.

A) Give a model for the analysis of covariance, explain each term in the model and formulate appropriate assumptions.

Here we could define a ANCOVA model with or without interaction. I will define it as an interaction model:

$$y_{ij} = \mu + \alpha_j + \beta(X_{ij} - \bar{X}) + \delta_j(X_{ij} - \bar{X}) + \epsilon_{ij}$$

Where y_{ij} is total weight of tomatoes in plot i with treatment j. μ is the grand overall mean of the response variable; α_j is the effect of the jth level of treatment; β is the main effect slope of mean plant height; δ_j is the interaction between treatment and mean plant height; and ϵ_{ij} is an error term. We assume that $y_{ij} \sim indN(\mu_{ij}, \sigma^2)$ and that $\epsilon_{ij} \sim i.i.dN(0, \sigma_\epsilon^2)$. The mean for mean plant height is computed over all the data.

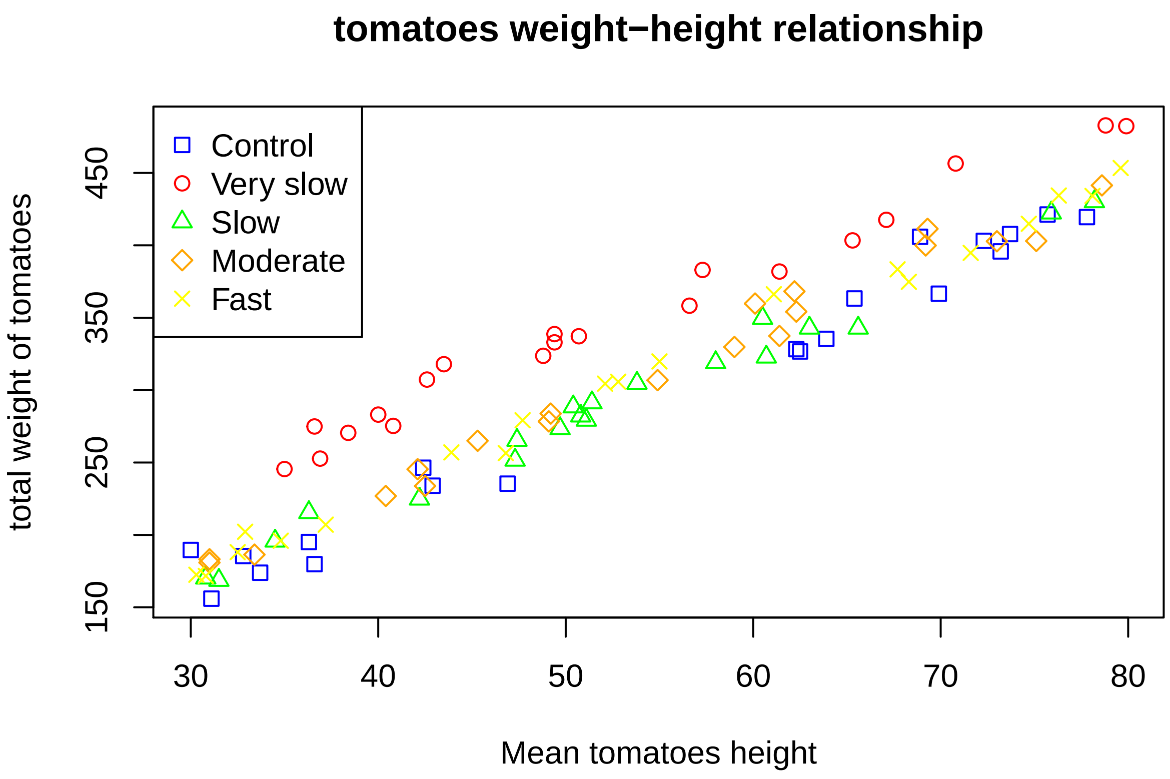
The constraints impose on the parameters are $\sum_j \alpha_j = 0$ and $\sum_j(\delta_j) = 0$.

B) Plot the response variable weight against the covariate height using levels of the treatments as different plotting symbols. What relationship is there between these two variables?

```r
library(readxl)
data_tomatoes = read_excel("Hwk8Q1DatSp17.xlsx")
head(data_tomatoes)
```

```
##   Row Treatment Height Weight
## 1   1         E   63.9  335.4
## 2   2         E   46.9  235.4
## 3   3         E   68.9  405.9
## 4   4         E   75.7  421.3
## 5   5         E   62.3  328.3
## 6   6         E   42.9  234.0
```

```r
data_tomatoes$Treatment = as.factor(data_tomatoes$Treatment)
# coded scatterplot
firesymbols = c()
firecolors = c()
for (i in 1:(dim(data_tomatoes)[1])) {
#check if there is a fireplace, symbol 2 corresponds to triangles
if (data_tomatoes$Treatment[i] == "E") {
firesymbols[i] = 0
firecolors[i] = 'blue'
}
if (data_tomatoes$Treatment[i] == "A") {
firesymbols[i] = 1
firecolors[i] = 'red'
}
if (data_tomatoes$Treatment[i] == "B") {
firesymbols[i] = 2
firecolors[i] = 'green'
}
if (data_tomatoes$Treatment[i] == "C") {
firesymbols[i] = 5
firecolors[i] = 'orange'
}
if (data_tomatoes$Treatment[i] == "D") {
firesymbols[i] = 4
firecolors[i] = 'yellow'
}
}
plot(data_tomatoes$Height, data_tomatoes$Weight, pch = firesymbols, col = firecolors,
xlab = "Mean tomatoes height", ylab = "total weight of tomatoes",
main = "tomatoes weight-height relationship")
legend("topleft", legend = c("Control", "Very slow", "Slow", "Moderate", "Fast"),  pch = c(0,1,2,5,4), 
```



- There is a positive linear relationship between weight and height of the tomatoes.
- All the lines seem to be parallel for the different treatments. Even for all the treatments except very slow, the lines may be one above the other. This might suggest two things: 1) that the interaction between treatment and height is not relevant, and 2) that there are only differences between the other treatments (control, slow, moderate, and fast) vs Very Slow.

C) Based on ordinary one-way ANOVA perform a significance test for the equality of the five treatment means. State hypotheses, test statistic, p-value, and your conclusions.

The null hypothesis here is:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$$

The alternative hypotesis is:

$$H_A : Not H_0$$

The test statistic is defined as:

$$F = \frac{MSB}{MSE}$$

Where MSB is the mean square between treatments and MSE is the mean square error. The p-value is $P(F_{4,95} > \frac{MSB}{MSE})$ In our case the F statistic is 1.3708.

```
model1 = lm(Weight~Treatment, data=data_tomatoes)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: Weight
##           Df Sum Sq Mean Sq F value Pr(>F)
## Treatment  4  39818  9954.6  1.3708 0.2499
## Residuals 95 689866  7261.8
```

The p-value is 0.2499, Thus the p-value>0.05 (0.2499>0.05) and we dot not reject the null hypothesis. Therefore, there is not differences in the total weight between the different treatments.

D) Determine if the interaction between treatment and initial plant height is significant. State hypotheses, test statistic, p-value, and your conclusions.

Here, the null hypothesis is (using the notation defined in part A):

$$H_0 : all\delta_j = 0$$
$$H_A : NotH_0$$

The test statistic is:

$$F = \frac{\frac{SS(\alpha,\beta,\delta)-SS(\alpha,\beta)}{q}}{MSE_{full}}$$

Where $SS(\alpha, \beta, \delta)$ is the sum of squares of the model with all main effects and interactions; and $SS(\alpha, \beta)$ is the model with only main effects (treatments and slope), q are the additional parameters estimated in the model with interactions; MSE_{full} is the mean square error in the model with interactions and main effects.

```
model2 = lm(Weight~ Treatment*Height, data=data_tomatoes)
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: Weight
##                  Df Sum Sq Mean Sq   F value Pr(>F)
## Treatment         4  39818    9955   85.5258 <2e-16 ***
## Height            1 678975  678975 5833.4655 <2e-16 ***
## Treatment:Height  4    416     104    0.8928 0.4717
## Residuals        90  10475     116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value of the test is 0.4717, then $p - value > 0.05$. Therefore, we do not reject the null hypothesis of not interaction effects. In conclusion, the test support the idea of not interaction effects between treatments and height.

E) Regardless of your answer to Part D above, perform analysis of covariance and make and summarize your results.

here is not clear if we need to make the ANCOVA with or without interactions. I will do it including the interactions:

```
summary(model2)
```

```
##
## Call:
## lm(formula = Weight ~ Treatment * Height, data = data_tomatoes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2483  -7.7712  -0.7162   6.9003  30.3020
```

```
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         75.8012     9.5137   7.968 4.76e-12 ***
## TreatmentB         -69.5803    13.8160  -5.036 2.43e-06 ***
## TreatmentC         -67.4068    13.3679  -5.042 2.37e-06 ***
## TreatmentD         -66.2177    12.4305  -5.327 7.31e-07 ***
## TreatmentE         -80.2148    12.5064  -6.414 6.44e-09 ***
## Height               5.1554     0.1754  29.391  < 2e-16 ***
## TreatmentB:Height    0.2504     0.2565   0.976    0.332    
## TreatmentC:Height    0.3819     0.2420   1.578    0.118    
## TreatmentD:Height    0.3601     0.2257   1.595    0.114    
## TreatmentE:Height    0.3600     0.2251   1.599    0.113    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.79 on 90 degrees of freedom
## Multiple R-squared:  0.9856, Adjusted R-squared:  0.9842 
## F-statistic: 686.6 on 9 and 90 DF,  p-value: < 2.2e-16
```

```
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: Weight
##                  Df Sum Sq Mean Sq   F value Pr(>F)    
## Treatment         4  39818    9955   85.5258 <2e-16 ***
## Height            1 678975  678975 5833.4655 <2e-16 ***
## Treatment:Height  4    416     104    0.8928 0.4717    
## Residuals        90  10475     116                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

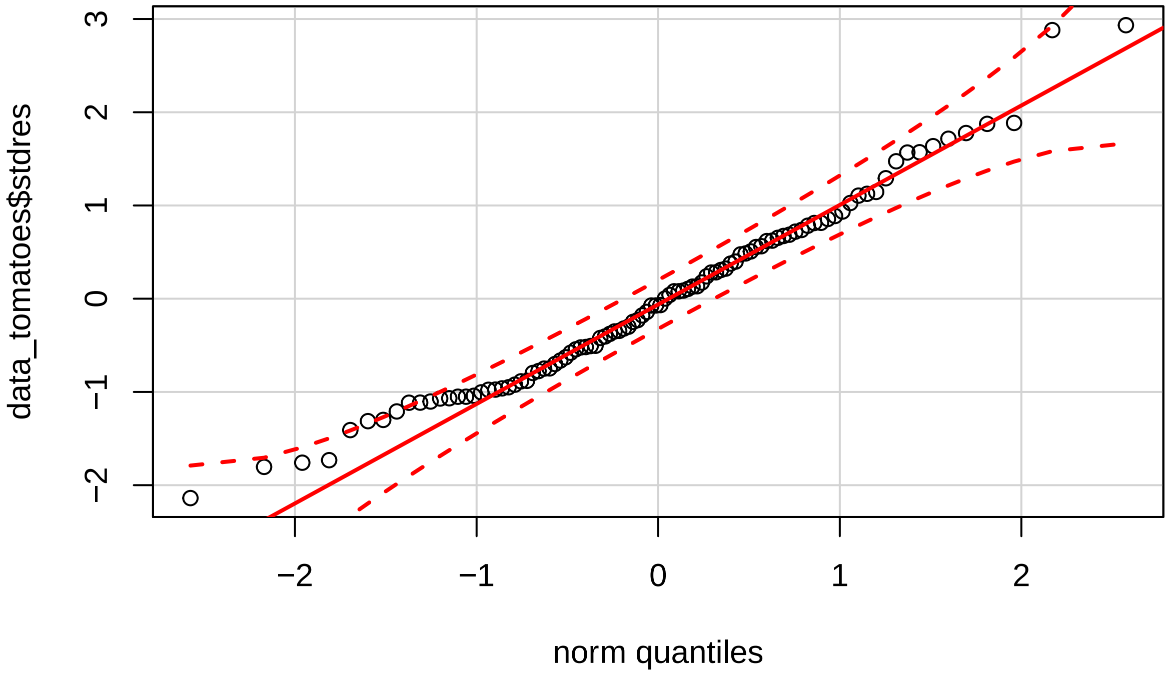
The results of the ANCOVA shows:

- There are differences in the in the mean weight between treatments.
- There is a positive and significate effect of the mean plant height on the tomatoes weight.
- The slope of height is not different among treatments.

F) Based on the model in Part A, make use of standardized residual plots to assess validity of the assumptions of independence, equal variance, and normality.

- The assumption of independence seems reasonable given the experimental design.
- The normality assumption could be checked with a q-q plot of the residuals as follows:

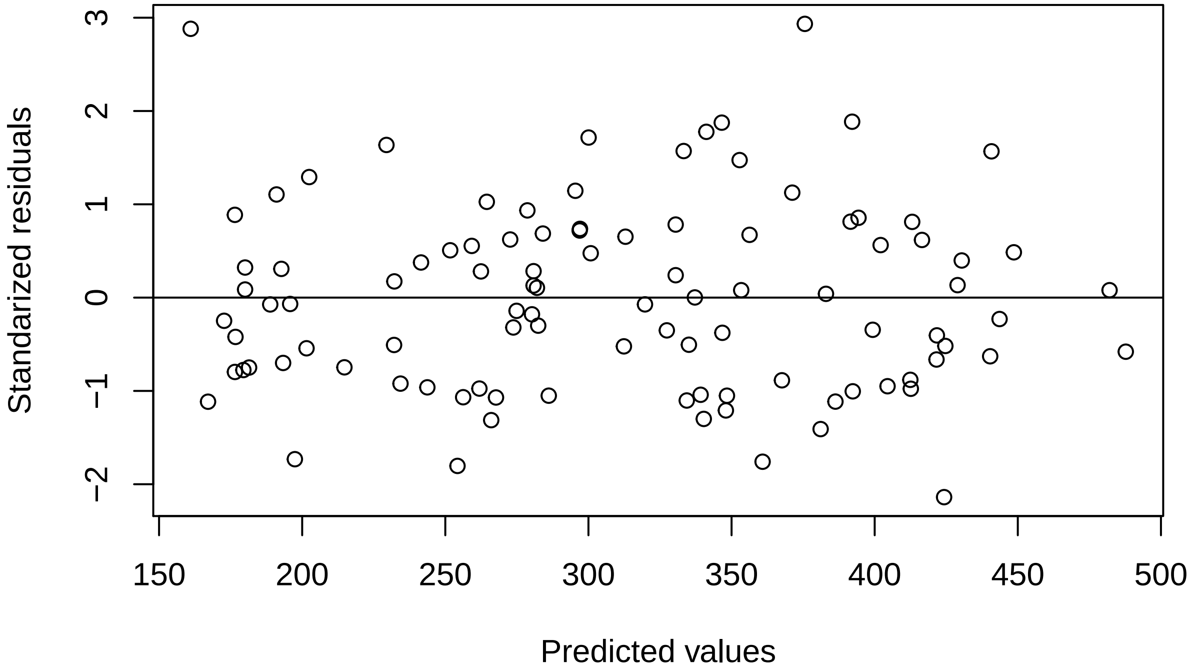
```
library(car)
data_tomatoes$stdres=rstandard(model2)
qqPlot(data_tomatoes$stdres)
```



The assumpation of normality holds.

- The assumption of equal variance could be checked with a plot of the residuals vs the fitted values as follows:

```
plot(model2$fitted.values, data_tomatoes$stdres, ylab="Standarized residuals", xlab="Predicted values",
```



The assumpation of equal variance seems reasonable.

G) What multiple comparison method that we've used would be more appropriate to compare each treatment to the control treatment? Use such a method at an overall error rate ??? = 0.05 and state carefully your conclusions.

Since we are interested in compare each treatment to the control treatment keeping an overall error rate of 0.05, the Dunnett's multiple comparison method is reasonable.

```
library(multcomp)
```

```
## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

```
model3 = aov(Weight~ Treatment*Height, data=data_tomatoes)
factors=table(data_tomatoes$Treatment)
#cont = contrMat(factors, base=5) # this function help us to establish treatment E as the control group
cont = contrMat(factors, type=c("Dunnett"), base=5) # this function help us to establish treatment E as
summary(glht(model3, linfct = mcp(Treatment=cont)))
```

```
## Warning in mcp2matrix(model, linfct = linfct): covariate interactions found
## -- default contrast might be inappropriate

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: aov(formula = Weight ~ Treatment * Height, data = data_tomatoes)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## A - E == 0    80.21      12.51   6.414   <1e-04 ***
## B - E == 0    10.63      12.89   0.825    0.839
## C - E == 0    12.81      12.41   1.032    0.707
## D - E == 0    14.00      11.40   1.228    0.569
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

The results suggest that there is only difference between the control and treatment A (very slow release). This result is supported by the graphical evidence that we have shown in part B: the mean weight for treatment A is greater than for the control. There is not differences in the mean weight for the other treatments vs the control.

Question 2.

A field trial is run to test the productivity of three different varieties of strawberries in an experimental field station in New York State. Four equally sized fields are available for use, and one-third of each field is planted in one variety of strawberry, the density of plants kept the same throughout the trial. Each 1/3 field is then randomly assigned one of the three varieties in such a way that each field has all three varieties planted within it. The Yield of strawberries (kg) over a two week period is then recorded for each 1/3 field.

A) Give a model statement for this experiment. Define each term and any constraints or conditions that are attached to it.

I consider that there is no reason to consider a random effects model is this experiment. The study seems to be interested in the three varieties of strawberries and the blocks are not chosen randomly but just for convenience (availability). Therefore, the model for this experiment is just a two-way fixed effects ANOVA:

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

Where y_{ij} is the response for observation with treatments i (var=A,B,..,C) in block j (block=1,2,3,..,4). In our case y_{ij} is the Yield of strawberries (kg) over a two week period for observation in treatments i and block j. μ is the grand overall mean of the response variable; α_i is the effect of the ith variety of strawberry; β_j is the effect of the jth block; and ϵ_{ij} is an error term. We also assume that $y_{ij} \sim indN(\mu_{ij}, \sigma^2)$ and that $\epsilon_{ij} \sim i.i.dN(0, \sigma_\epsilon^2)$. The constraints impose on the parameters are $\sum_i \alpha_i = 0$ and $\sum_j \beta_j = 0$.

B) Test to see if there are any differences between the yields of the various strawberries. State hypotheses, test statistic, p-value, and conclusions.

```r
library(readxl)
data_berries = read_excel("Hwk8Q2DatSp17.xlsx")
head(data_berries)
```

```
##   Yield Var Blk
## 1  10.1   A   1
## 2   6.3   B   1
## 3   8.4   C   1
## 4  10.8   A   2
## 5   6.9   B   2
## 6   9.4   C   2
```

```r
data_berries$Var = as.factor(data_berries$Var)
data_berries$Blk = as.factor(data_berries$Blk)
```

The null hypothesis here is:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$$

The alternative hypotesis is:

$$H_A : Not H_0$$

The test statistic is defined as:

$$F = \frac{MSB}{MSE} = 58.1875$$

Where MSB is the mean square between treatments (variety of strawberries) and MSE is the mean square error. The p-value is $P(F_{2,8} > 58.1875)$. We get the value of the test in R as follows:

```
model1_berries = aov(Yield~ Var + Blk, data=data_berries)
anova(model1_berries)
```

```
## Analysis of Variance Table
##
## Response: Yield
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## Var        2 35.582 17.7908 147.2345 7.963e-06 ***
## Blk        3  1.722  0.5742   4.7517   0.05011 .
## Residuals  6  0.725  0.1208
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is 0.00001712, Thus the p-value<0.05 (0.00001712>0.05) and we do reject the null hypothesis. Therefore, there are differences in the yield between the different strawberries.

C) Use Tukey's HSD to find the varieties of strawberries which produce different yields. Whiuch would you recommend to be used in New York State (based solely on this yield criteria).

```
TukeyHSD(model1_berries)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Yield ~ Var + Blk, data = data_berries)
##
## $Var
##       diff       lwr        upr     p adj
## B-A -4.125 -4.879176 -3.3708242 0.0000066
## C-A -1.300 -2.054176 -0.5458242 0.0044487
## C-B  2.825  2.070824  3.5791758 0.0000643
##
## $Blk
##           diff        lwr         upr     p adj
## 2-1  0.7666667 -0.2158470  1.74918034 0.1232180
## 3-1 -0.2333333 -1.2158470  0.74918034 0.8423948
## 4-1  0.3666667 -0.6158470  1.34918034 0.5992709
## 3-2 -1.0000000 -1.9825137 -0.01748632 0.0465784
## 4-2 -0.4000000 -1.3825137  0.58251368 0.5376017
## 4-3  0.6000000 -0.3825137  1.58251368 0.2494088
```

As we can see, all the differences calculated are significant at 0.05 of significance level. This result means that all the varieties produce different mean yields. However, we are interested in the one that produce the higher yield; based on the comparisons we can see that the strawberry to be recommended to be used in New York is the variety A because compared with either variety B or C produce a higher yield.

D) What are the the values of the two compnents of variance in this study? What proportion of the variance of yield is attributable to the differences in fields?

In this question, you assume that I have define a model in part A with some of the factors being random. Well, I did not define neither factor to be random and I have given my reasoning to do it. However, since in this question you are looking for another source of variability (field), I will run an ANOVA treating field as a random effect. This means that we have two component of variance: σ^2_ϵ and σ^2_{blocks}.

```
library('lme4')
```

```
## Loading required package: Matrix
```

```
model2_berries= lmer(Yield~ Var + (1|Blk), data=data_berries, REML=F)
summary(model2_berries)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Yield ~ Var + (1 | Blk)
##    Data: data_berries
##
##      AIC      BIC   logLik deviance df.resid
##     21.5     23.9     -5.7     11.5        7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.70450 -0.33109  0.08301  0.46652  1.20209
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Blk      (Intercept) 0.11333  0.3367
##  Residual             0.09063  0.3010
## Number of obs: 12, groups:  Blk, 4
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  10.3000     0.2258   45.61
## VarB         -4.1250     0.2129  -19.38
## VarC         -1.3000     0.2129   -6.11
##
## Correlation of Fixed Effects:
##      (Intr) VarB
## VarB -0.471
## VarC -0.471  0.500
```

The variance for the error $\sigma^2_\epsilon = 0.09063$ and the variance due to blocks $\sigma^2_{blocks} = 0.11333$. Therefore, the proportion of variance in yield due to blocks is:

$$\%Var_{blocks} = \frac{\sigma^2_{blocks}}{\sigma^2_{blocks} + \sigma^2_\epsilon} = \frac{0.11333}{0.20396} = 0.5556 \approx 55\%$$

E) Use library(lme4) and library(LmerTest) to test to see if the random effects are statistically significant in this model. (Note: There is some debate amongst statisticians that this procedure is appropriate when using REML. Some say it is OK to use REML with this test IF THE FIXED EFFECTS ARE THE SAME IN BOTH THE FULL AND REDUCED MODELS, which how it is stated in our text.)

```
library(nlme)
```

```
##
## Attaching package: 'nlme'

## The following object is masked from 'package:lme4':
##
##     lmList
```

```
modelA = lme(Yield~ Var, random=~1|Blk, data=data_berries, method="ML")
modelB = gls(Yield~ Var, data=data_berries, method = "ML")
anova(modelB,modelA)
```

```
##        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## modelB     1  4 22.97645 24.91608 -7.488225
## modelA     2  5 21.47625 23.90079 -5.738126 1 vs 2 3.500197  0.0614
```

The results show that at 0.05 of significance level, we do not reject the null hypothesis ($H_0 : \sigma^2_{blocks} = 0$). p-value>0.05 (0.0614>0.05). Therefore, the random effect of block is not significant. This result is surprising given the proportion of variability due to blocks (part E).

F) State which of our observations are correlated in this situation.

Since the random effect of blocks is not significant, we have evidence to say that strawberries with the same block for the random predictor will not be correlated.