STAT 512 Exam 2 Extra Practice

Kathleen Wendt

4/22/2020

Packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lme4)
```

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

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
library(lmerTest)
```

```
##
## Attaching package: 'lmerTest'
```

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

```
## The following object is masked from 'package:stats':
##
##     step
```

```r
library(emmeans)
```

Part 1: Plants

An investigator is interested in comparing the expression of a certain gene for plants grown under 2 different Conditions (Trt and Ctrl). 4 seedlings were randomly assigned to the Ctrl condition and 4 seedlings were randomly assigned to the Trt condition. So, there are 4 Plants per Condition. At the end of the study period, an RNA sample was obtained from each plant and split into triplicates (labeled Rep in the data). Each sample was analyzed using RT PCR and gene expression (Y) was measured. Due to the triplicates, there were 3 observations for each Plant for a total of 24 observations (2 x 4 x 3). Note: We will consider Condition to be fixed and Plant to be random. The data is available from Canvas as "Plants.csv"

```r
# read and prepare plant data
plant_data <- readr::read_csv("data/Plants.csv") %>%
  janitor::clean_names() %>%
  dplyr::mutate(condition = as.factor(condition),
                plant = as.factor(plant),
                rep = as.factor(rep))
```

```
## Parsed with column specification:
## cols(
##   Condition = col_character(),
##   Plant = col_character(),
##   Rep = col_character(),
##   Y = col_double()
## )
```

1A: Design

Are Plant and Condition crossed or nested? If nested, be sure to indicate the "direction" of the nesting.

Nested. Plant is nested within Condition.

1B: Nested mixed effects model

Considering your answer to A, fit an appropriate model. Include the variance parameter estimates and Type 3 ANOVA table in your assignment.

```r
# build mixed effects model with fixed condition and random plant
plant_mixed_lm <- lmer(y ~ condition + (1|condition:plant), data = plant_data)
# call summary on model to review variance parameter estimates
summary(plant_mixed_lm)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: y ~ condition + (1 | condition:plant)
##    Data: plant_data
##
## REML criterion at convergence: 85.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -1.5861 -0.6823  0.0094  0.5397  1.4766
##
## Random effects:
##  Groups          Name        Variance Std.Dev.
##  condition:plant (Intercept) 2.019    1.421
##  Residual                    1.429    1.196
## Number of obs: 24, groups:  condition:plant, 8
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)   10.4433     0.7899  6.0000  13.221 1.16e-05 ***
## conditionTrt  -1.8783     1.1171  6.0000  -1.681    0.144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## conditinTrt -0.707
```

```
# call anova on model for type 3 anova table - diff in gene expression?
anova(plant_mixed_lm, ddf = "Kenward-Roger")
```

```
## Type III Analysis of Variance Table with Kenward-Roger's method
##           Sum Sq Mean Sq NumDF DenDF F value Pr(>F)
## condition 4.0416  4.0416     1     6  2.8274 0.1437
```

1C: Conclusion

Using the model from part B, can we conclude there is a difference between the mean responses for the two Conditions? Provide an estimate of the difference and a p-value. Hint: Use emmeans.

```
# extract emmeans for mixed effects model
emmeans::emmeans(plant_mixed_lm, pairwise ~ condition)
```

```
## $emmeans
```

```
## Warning in format.default(nm[j], width = nchar(m[1, j]), just = "left"): partial
## argument match of 'just' to 'justify'
```

```
##  condition emmean   SE df lower.CL upper.CL
##  Ctrl       10.44 0.79  6     8.51     12.4
##  Trt         8.56 0.79  6     6.63     10.5
##
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.95
##
## $contrasts
```

```
## Warning in format.default(round(x$t.ratio, 3), nsmall = 3, sci = FALSE): partial
## argument match of 'sci' to 'scientific'
```

```
## Warning in format.default(round(x$p.value, 4), nsmall = 4, sci = FALSE): partial
## argument match of 'sci' to 'scientific'
```

```
## Warning in format.default(nm[j], width = nchar(m[1, j]), just = "left"): partial
## argument match of 'just' to 'justify'
```

```
## Warning in format.default(nm[j], width = nchar(m[1, j]), just = "left"): partial
## argument match of 'just' to 'justify'

## Warning in format.default(nm[j], width = nchar(m[1, j]), just = "left"): partial
## argument match of 'just' to 'justify'

##  contrast   estimate   SE df t.ratio p.value
##  Ctrl - Trt     1.88 1.12  6 1.681   0.1437
```

Fail to reject null hypothesis. No difference in mean response by condition with an estimated difference of 1.88 units, $p = 0.14$.

1D: t-test

We will rerun the analysis using a different approach. We will start by averaging over the triplicates and run a two-sample t-test (assuming equal variance) with n = 4 observations per Condition. In your assignment, provide an estimate of the difference and a p-value. Note: Using option var.equal = TRUE, returns the two sample t-test assuming equal variance.

```
# calculate average response by condition and plant
plant_avg_data <- plant_data %>%
  dplyr::group_by(condition, plant) %>%
  dplyr::summarize(mean = mean(y))
# conduct paired samples t-test assuming equal variances
t.test(mean ~ condition,
       var.equal = TRUE,
       data = plant_avg_data)
```

```
##
##  Two Sample t-test
##
## data:  mean by condition
## t = 1.6815, df = 6, p-value = 0.1437
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8550527  4.6117193
## sample estimates:
## mean in group Ctrl  mean in group Trt
##           10.44333            8.56500
```

The estimate of the difference is 1.88 units, and the p-value is 0.14.

1E: Compare

Compare your result from part D to the test from part C. Are the results the same?

Yes.

1F: Independence

A colleague suggests that you "just do a two-sample t-test with n=12 observations per condition."

Would this analysis be appropriate? Justify your response.

No! Triplicates from the same plant cannot be considered independent observations.

1G: Independent samples t-test

Would the analysis from #6 give the same results as part C? Hint: Try it and compare the results!

```
# conduct indepndent samples t-test assuming equal variances
t.test(y ~ condition, var.equal = TRUE, data = plant_data)
```

```
##
##  Two Sample t-test
##
## data:  y by condition
## t = 2.621, df = 22, p-value = 0.0156
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.392085 3.364582
## sample estimates:
## mean in group Ctrl  mean in group Trt
##           10.44333            8.56500
```

No. The estimated difference matches parts C and D but the p-value = 0.0156.

Part 2: Seed weights

A split-plot experiment was conducted on sorghum with two treatment factors: plant Density and Hybrid. A total of 4 blocks were used in the study. Within each block, the four levels of plant Density (10, 15, 25 and 40 plants per meter of row) were randomly assigned to whole plots. Then within each whole plot, the three Hybrids (A, B, C) were randomly assigned to subplots. The response (Y) is the weight of the seed per plant in grams. The data are given in the file "SeedWeight.csv" on Canvas.

```
# read and prepare seed data
seed_data <- readr::read_csv("data/SeedWeight.csv") %>%
  janitor::clean_names() %>%
  dplyr::mutate(hybrid = as.factor(hybrid),
                block = as.factor(block),
                density = as.factor(density))
```

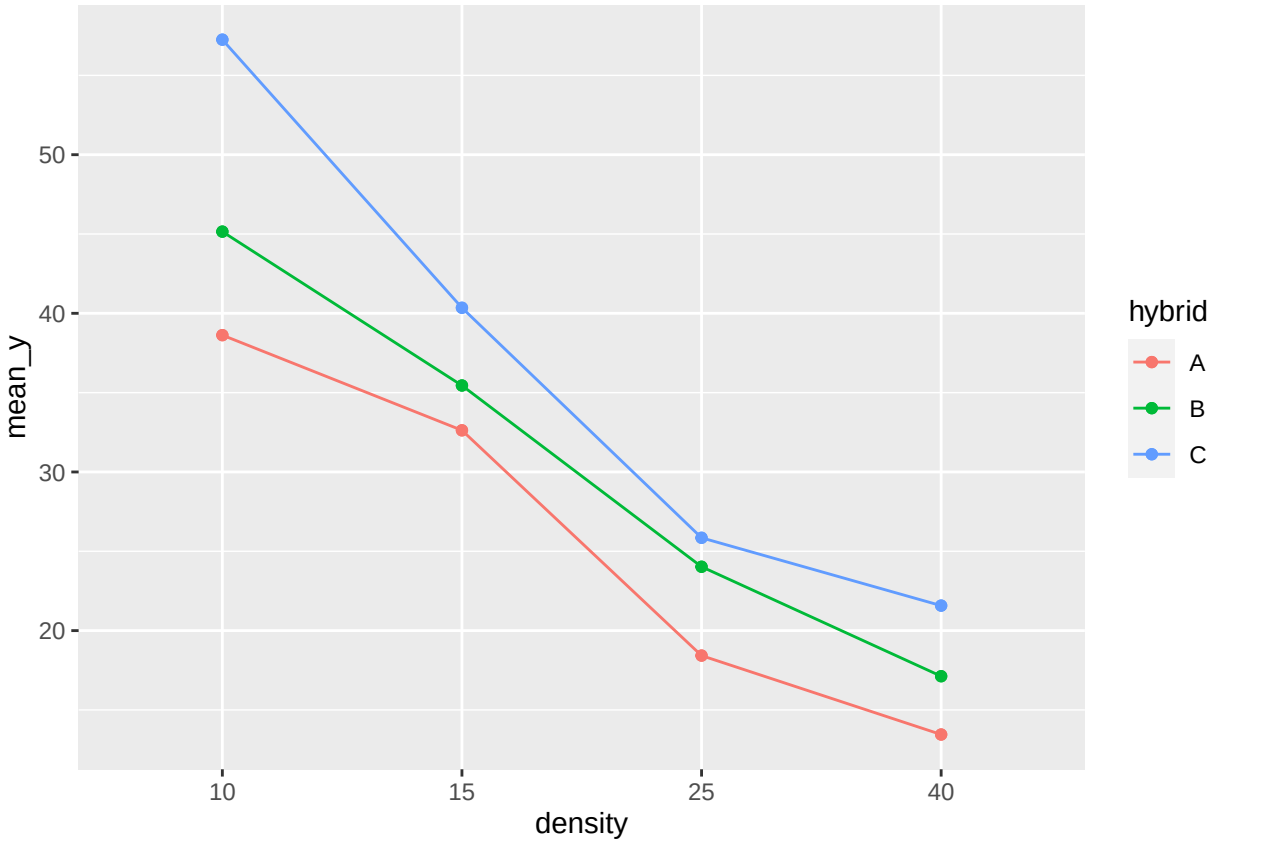
```
## Parsed with column specification:
## cols(
##   Hybrid = col_character(),
##   Block = col_double(),
##   Density = col_double(),
##   Y = col_double()
## )
```

2A: Plot

Create an interaction plot (for Density and Hybrid) and include it in your assignment. For consistency, put `Density` on the horizontal axis.

```
# summarize mean response by density and hybrid
seed_avg_data <- seed_data %>%
  dplyr::group_by(density, hybrid) %>%
  dplyr::summarize(mean_y = mean(y))
# build interaction plot for density and hybrid
```

```
qplot(x = density,
      y = mean_y,
      group = hybrid,
      color = hybrid,
      data = seed_avg_data) +
  geom_line() +
  geom_point()
```



2B: Split-plot mixed effects model

Fit an appropriate model using `lmer()`. Include the variance component estimates and Type 3 ANOVA table in your assignment.

```
# build seed mixed effects split-plot model
seed_split_lm <- lmer(y ~ density*hybrid + (1|block) + (1|block:density),
                      data = seed_data)
summary(seed_split_lm)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: y ~ density * hybrid + (1 | block) + (1 | block:density)
##    Data: seed_data
##
## REML criterion at convergence: 246.2
##
## Scaled residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.70971 -0.39374 -0.01162  0.34474  2.22566
##
## Random effects:
##  Groups        Name        Variance Std.Dev.
##  block:density (Intercept)  9.007   3.001
##  block         (Intercept)  7.040   2.653
##  Residual                  24.820   4.982
## Number of obs: 48, groups:  block:density, 16; block, 4
##
## Fixed effects:
##                    Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)          38.625      3.196  22.850  12.084 2.10e-11 ***
## density15            -6.000      4.113  25.663  -1.459   0.1567
## density25           -20.200      4.113  25.663  -4.912 4.38e-05 ***
## density40           -25.175      4.113  25.663  -6.121 1.90e-06 ***
## hybridB               6.525      3.523  24.000   1.852   0.0763 .
## hybridC              18.625      3.523  24.000   5.287 2.01e-05 ***
## density15:hybridB    -3.700      4.982  24.000  -0.743   0.4649
## density25:hybridB    -0.925      4.982  24.000  -0.186   0.8543
## density40:hybridB    -2.850      4.982  24.000  -0.572   0.5726
## density15:hybridC   -10.900      4.982  24.000  -2.188   0.0387 *
## density25:hybridC   -11.200      4.982  24.000  -2.248   0.0340 *
## density40:hybridC   -10.500      4.982  24.000  -2.108   0.0457 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) dnst15 dnst25 dnst40 hybrdB hybrdC dn15:B dn25:B dn40:B
## density15   -0.643
## density25   -0.643  0.500
## density40   -0.643  0.500  0.500
## hybridB     -0.551  0.428  0.428  0.428
## hybridC     -0.551  0.428  0.428  0.428  0.500
## dnsty15:hyB  0.390 -0.606 -0.303 -0.303 -0.707 -0.354
## dnsty25:hyB  0.390 -0.303 -0.606 -0.303 -0.707 -0.354  0.500
## dnsty40:hyB  0.390 -0.303 -0.303 -0.606 -0.707 -0.354  0.500  0.500
## dnsty15:hyC  0.390 -0.606 -0.303 -0.303 -0.354 -0.707  0.500  0.250  0.250
## dnsty25:hyC  0.390 -0.303 -0.606 -0.303 -0.354 -0.707  0.250  0.500  0.250
## dnsty40:hyC  0.390 -0.303 -0.303 -0.606 -0.354 -0.707  0.250  0.250  0.500
##             dn15:C dn25:C
## density15
## density25
## density40
## hybridB
## hybridC
## dnsty15:hyB
## dnsty25:hyB
## dnsty40:hyB
## dnsty15:hyC
## dnsty25:hyC  0.500
## dnsty40:hyC  0.500  0.500
```

```r
anova(seed_split_lm, ddf = "Kenward-Roger")
```

```
## Warning in attr(Xorig, "contrast"): partial match of 'contrast' to 'contrasts'
```

```
## Type III Analysis of Variance Table with Kenward-Roger's method
##                 Sum Sq Mean Sq NumDF DenDF F value    Pr(>F)
## density        3078.27 1026.09     3     9 41.3407 1.370e-05 ***
## hybrid          881.41  440.70     2    24 17.7558 1.851e-05 ***
## density:hybrid  207.51   34.58     6    24  1.3934    0.2577
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2C: Pairwise comparisons of density

You should find that the Density*Hybrid interaction is not significant. Use `emmeans()` to run pairwise comparisons of Density levels (averaging over Hybrids) and comparisons of Hybrids (averaging over Density levels).

```r
# compare density levels, averaging over hybrid
emmeans::emmeans(seed_split_lm, pairwise ~ density)
```

```
## NOTE: Results may be misleading due to involvement in interactions
```

```
## $emmeans
```

```
## Warning in format.default(nm[j], width = nchar(m[1, j]), just = "left"): partial
## argument match of 'just' to 'justify'
```

```
##  density emmean   SE   df lower.CL upper.CL
##  10        47.0 2.47 9.59     41.5     52.5
##  15        36.1 2.47 9.59     30.6     41.7
##  25        22.8 2.47 9.59     17.2     28.3
##  40        17.4 2.47 9.59     11.9     22.9
##
## Results are averaged over the levels of: hybrid
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.95
##
## $contrasts
```

```
## Warning in format.default(round(x$t.ratio, 3), nsmall = 3, sci = FALSE): partial
## argument match of 'sci' to 'scientific'
```

```
## Warning in format.default(round(x$p.value, 4), nsmall = 4, sci = FALSE): partial
## argument match of 'sci' to 'scientific'
```

```
## Warning in format.default(nm[j], width = nchar(m[1, j]), just = "left"): partial
## argument match of 'just' to 'justify'
```

```
## Warning in format.default(nm[j], width = nchar(m[1, j]), just = "left"): partial
## argument match of 'just' to 'justify'
```

```
## Warning in format.default(nm[j], width = nchar(m[1, j]), just = "left"): partial
## argument match of 'just' to 'justify'
```

```
##  contrast estimate   SE df t.ratio p.value
##  10 - 15     10.87 2.94  9  3.697  0.0212
##  10 - 25     24.24 2.94  9  8.247  0.0001
```

```
##  10 - 40     29.62 2.94  9 10.079  <.0001
##  15 - 25     13.38 2.94  9  4.550  0.0062
##  15 - 40     18.76 2.94  9  6.382  0.0006
##  25 - 40      5.38 2.94  9  1.831  0.3202
##
## Results are averaged over the levels of: hybrid
## P value adjustment: tukey method for comparing a family of 4 estimates
```

```
# compare hybrid levels, averaging over density
emmeans::emmeans(seed_split_lm, pairwise ~ hybrid)
```

```
## NOTE: Results may be misleading due to involvement in interactions

## $emmeans

## Warning in format.default(nm[j], width = nchar(m[1, j]), just = "left"): partial
## argument match of 'just' to 'justify'

##  hybrid emmean   SE   df lower.CL upper.CL
##  A        25.8 1.97 5.49     20.9     30.7
##  B        30.4 1.97 5.49     25.5     35.4
##  C        36.3 1.97 5.49     31.3     41.2
##
## Results are averaged over the levels of: density
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.95
##
## $contrasts

## Warning in format.default(round(x$t.ratio, 3), nsmall = 3, sci = FALSE): partial
## argument match of 'sci' to 'scientific'

## Warning in format.default(round(x$p.value, 4), nsmall = 4, sci = FALSE): partial
## argument match of 'sci' to 'scientific'

## Warning in format.default(nm[j], width = nchar(m[1, j]), just = "left"): partial
## argument match of 'just' to 'justify'

## Warning in format.default(nm[j], width = nchar(m[1, j]), just = "left"): partial
## argument match of 'just' to 'justify'

## Warning in format.default(nm[j], width = nchar(m[1, j]), just = "left"): partial
## argument match of 'just' to 'justify'

##  contrast estimate   SE df t.ratio p.value
##  A - B       -4.66 1.76 24 -2.643  0.0366
##  A - C      -10.47 1.76 24 -5.947  <.0001
##  B - C       -5.82 1.76 24 -3.303  0.0081
##
## Results are averaged over the levels of: density
## P value adjustment: tukey method for comparing a family of 3 estimates
```

2D: Accuracy

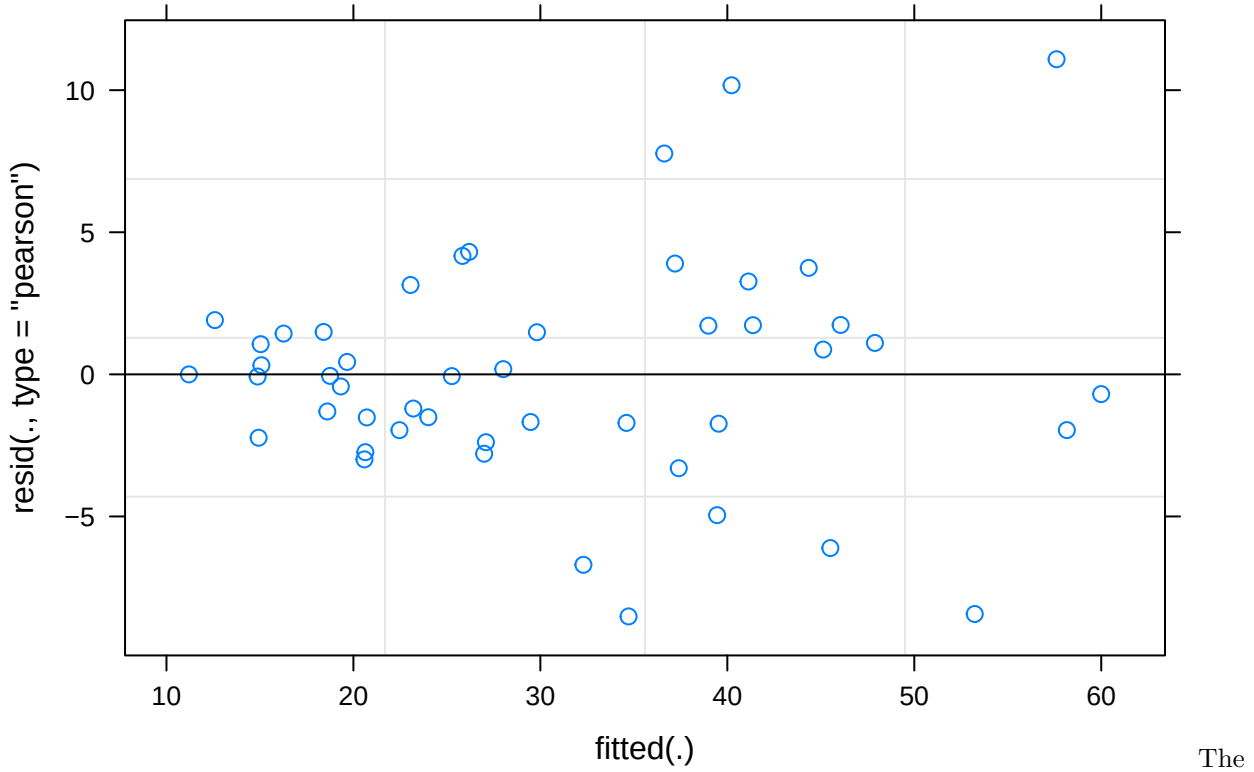
On slide 61 (Random2 notes), it says "the accuracy on factor B (sub-plot factor) is better than the accuracy of comparisons on factor A (whole-plot factor)". Looking at your results from part B, what evidence do we see to support this statement (for this dataset).

The standard error when comparing densities is larger than the SE for hybrids.

2E:

Use `plot()` to generate the plot of residuals versus fitted values. You do not have to include this plot in your assignment, just comment on what you see.

```
plot(seed_split_lm)
```



The plot of residuals vs. fitted values shows a megaphone shape. This indicates that the assumption of equal variance is violated. We could try a transformation.