Homework10

```r
library(lme4)
```

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

```r
library(emmeans)
```

```
## Welcome to emmeans.
## Caution: You lose important information if you filter this package's results.
## See '? untidy'
```

```r
ds <- readxl::read_excel('/Users/kanoalindiwe/Downloads/Projects/playground/R/Quantitative Ecology/Data
ds <- ds[ds$species == "ohia", ]

plot <- as.factor(ds$plot)
species <- ds$species
dbh2 <- as.numeric(ds$dbh2)
comstat <- relevel(as.factor(ds$comstat), ref = "S")
growth <- ds$`growth rate`
```

Create two linear models that examine factors associated with ōhia growth.

Linear model with fixed effects only. For the first, run a linear model with "comstat" as the explanatory variable and "growth" as the response. Run this as a linear model (lm) using the *summary* function.
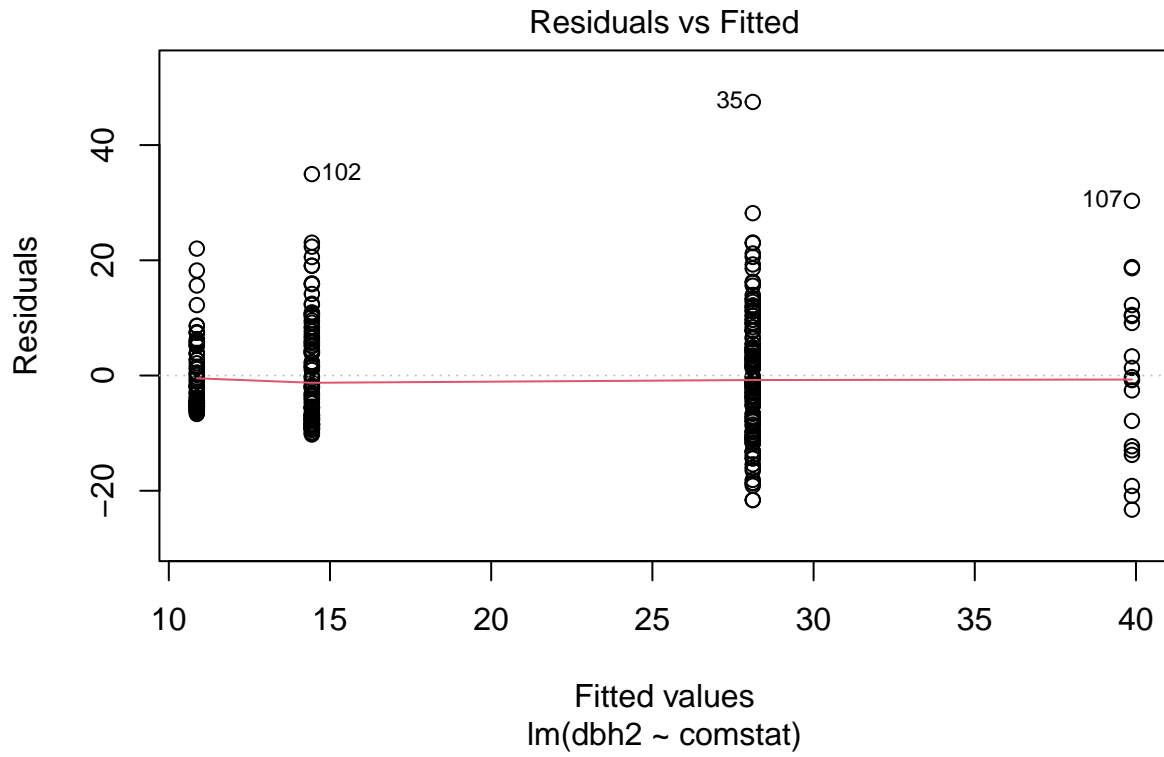
Asumptions

```r
# Test model
model_test <- lm(dbh2 ~ comstat)

# Linearity - The relationship between predictors and the response variable is linear.
plot(model_test, which = 1)
```

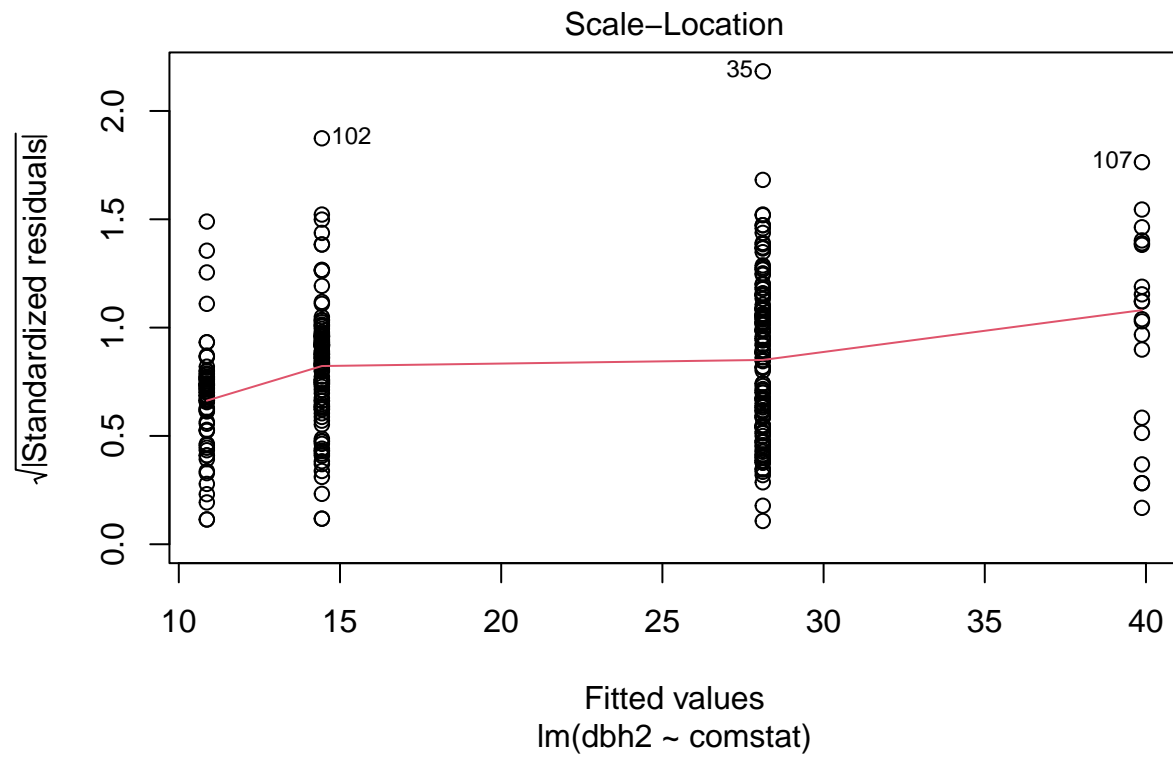## Residuals vs Fitted

Fitted values
lm(dbh2 ~ comstat)

```
# Pass, does not look skewed

# Independence - Observations (and residuals) are independent of each other.
# Pass, assumed to be good.

# Homoscedasticity - The variance of residuals is constant across all fitted values.
plot(model_test, which = 3)
```
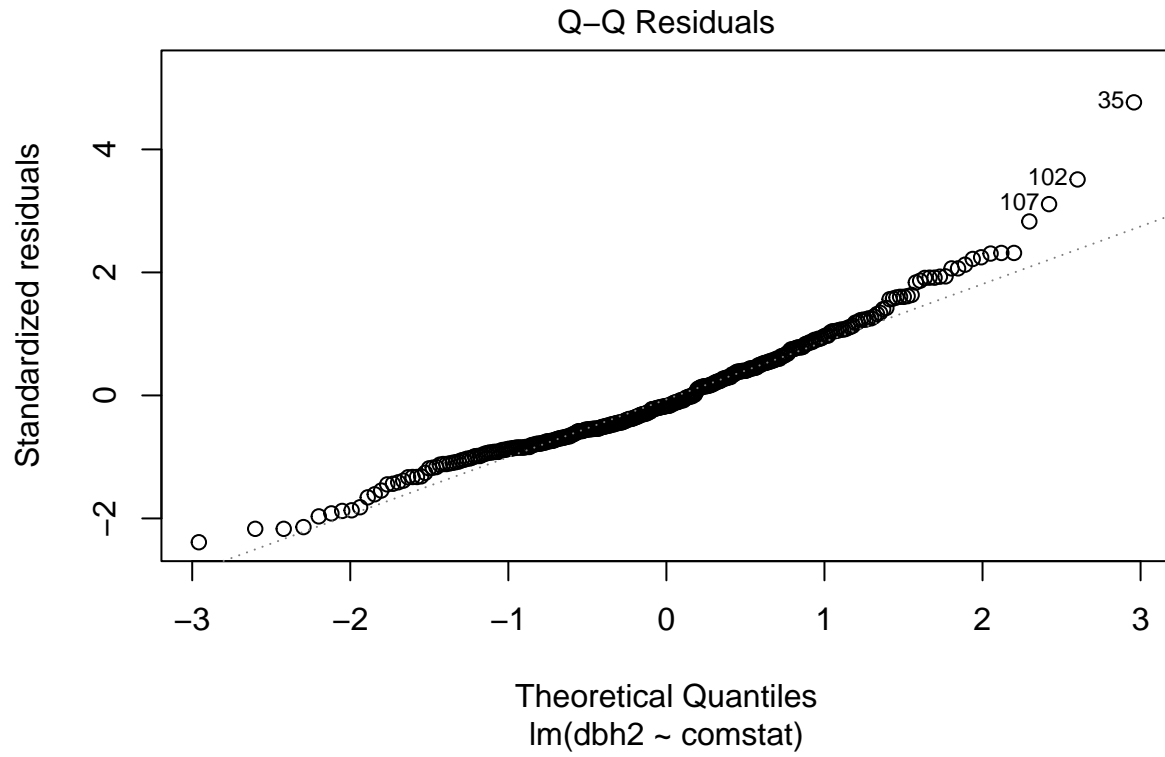
Scale–Location

lm(dbh2 ~ comstat)

```
# Pass, looks relatively flat

# Normality of Errors - Residuals are approximately normally distributed.
plot(model_test, which = 2)
```

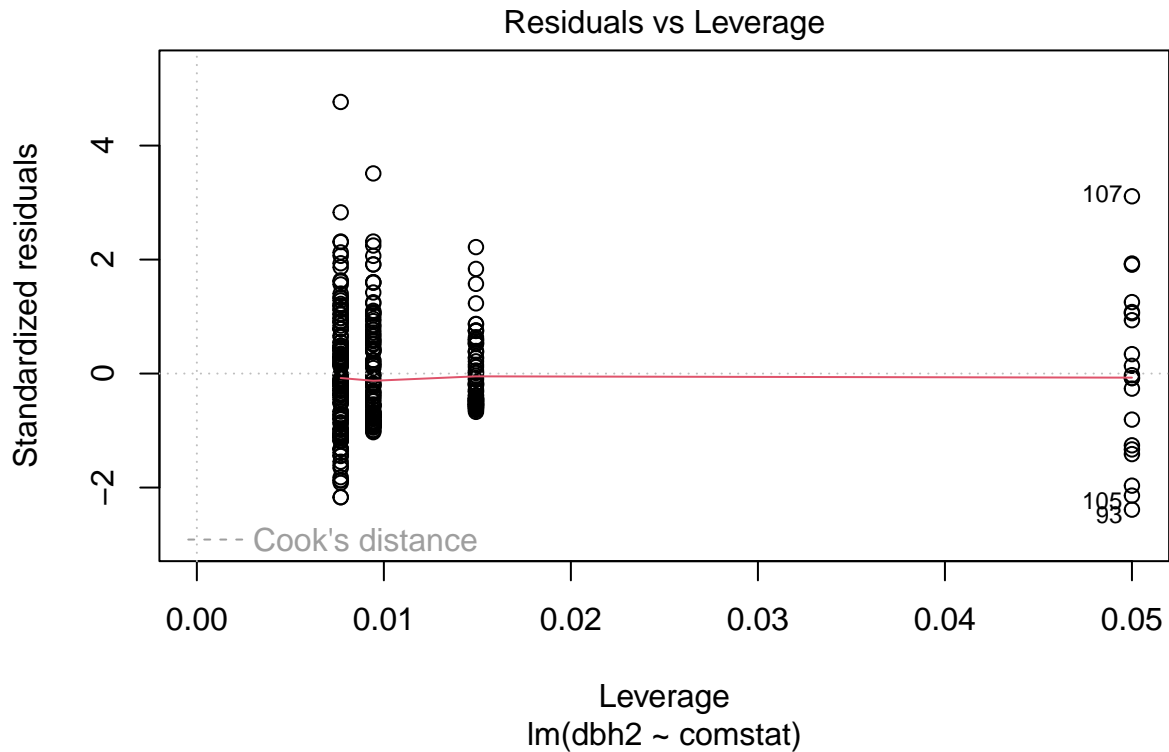## Q-Q Residuals

Theoretical Quantiles
lm(dbh2 ~ comstat)

```
# Pass, relatively normal

# No Multicollinearity - Predictor variables are not highly correlated with each other.
# Pass, only one factor variable input so not required

# No Significant Outliers or Influential Points - Extreme values do not unduly affect the model.
plot(model_test, which = 5)
```

Residuals vs Leverage

Leverage
lm(dbh2 ~ comstat)

```
# Pass, does not pass cooks distance
```

Model

```
model1 <- lm(dbh2 ~ comstat)
summary(model1)
```

```
##
## Call:
## lm(formula = dbh2 ~ comstat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.275  -6.939  -1.715   5.646  47.485
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.870      1.222   8.894   <2e-16 ***
## comstatC      17.244      1.504  11.462   <2e-16 ***
## comstatD      29.005      2.549  11.379   <2e-16 ***
## comstatI       3.569      1.561   2.286   0.0229 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10 on 319 degrees of freedom
```

```
## Multiple R-squared:  0.4404, Adjusted R-squared:  0.4352
## F-statistic:  83.7 on 3 and 319 DF,  p-value: < 2.2e-16
```

Linear model with fixed and random effects. Because reviewers may wonder if the plot that the trees are growing in may have an effect on the growth of the trees, also make a linear mixed effects model as above but with plot as a random factor. Be sure to examine the assumption of normality of the response and log transform if necessary (for both models).
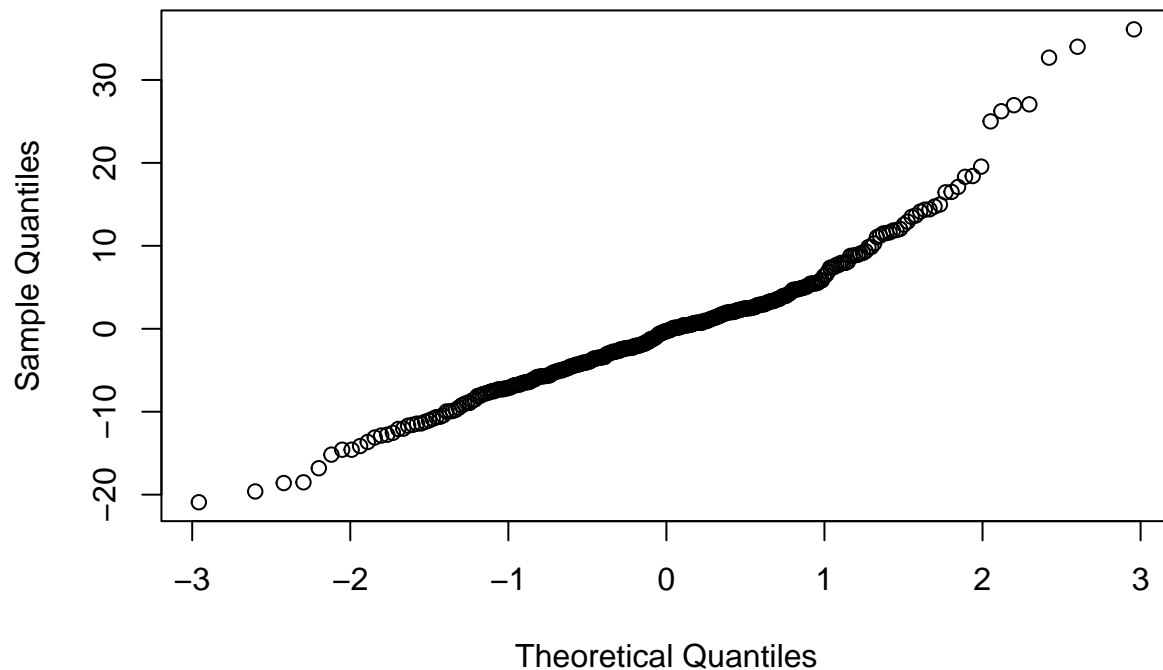
```
model_test1 <- lmer(dbh2 ~ comstat + (1 | plot))

# Homoscedasticity – The variance of residuals is constant across all fitted values.
plot(fitted(model_test1, sqrt(abs(residuals(model_test1)))))
```
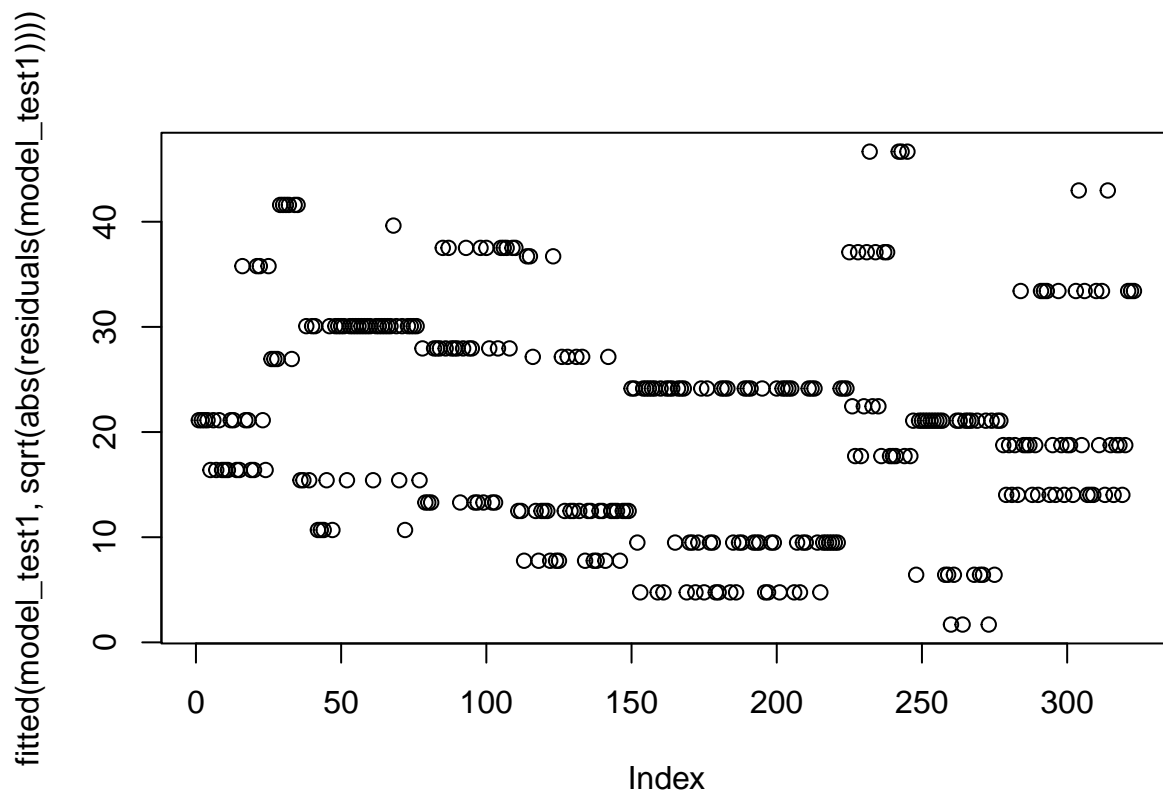


```
# Pass, looks relatively flat

# Normality of Errors – Residuals are approximately normally distributed.
qqnorm(residuals(model_test1))
```

Normal Q–Q Plot


```
# Pass, relatively normal
```

Model

```
model2 <- lmer(dbh2 ~ comstat + (1 | plot))
summary(model2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: dbh2 ~ comstat + (1 | plot)
##
## REML criterion at convergence: 2305.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4835 -0.5960 -0.0348  0.3969  4.2868
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  plot     (Intercept) 46.10    6.790
##  Residual             70.92    8.421
## Number of obs: 323, groups:  plot, 9
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   11.535      2.520   4.578
```

```
## comstatC        19.385        1.352  14.336
## comstatD        28.947        2.323  12.459
## comstatI         4.733        1.339   3.535
##
## Correlation of Fixed Effects:
##          (Intr) cmsttC cmsttD
## comstatC -0.352
## comstatD -0.215  0.391
## comstatI -0.335  0.632  0.376
```

```
confint(model2, method = "Wald")
```

```
##                  2.5 %     97.5 %
## .sig01              NA         NA
## .sigma              NA         NA
## (Intercept)  6.596534 16.473796
## comstatC    16.735144 22.035633
## comstatD    24.393270 33.500412
## comstatI     2.109131  7.356688
```

```
posthoc <- emmeans(model2, pairwise ~ comstat, adjust = "tukey")
posthoc
```

```
## $emmeans
##  comstat emmean   SE    df lower.CL upper.CL
##  S         11.5 2.52 10.92     5.98     17.1
##  C         30.9 2.41  9.09    25.49     36.4
##  D         40.5 3.04 22.63    34.18     46.8
##  I         16.3 2.43  9.40    10.82     21.7
##
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate   SE  df t.ratio p.value
##  S - C     -19.39 1.35 315 -14.309  <.0001
##  S - D     -28.95 2.33 315 -12.434  <.0001
##  S - I      -4.73 1.34 313  -3.533  0.0027
##  C - D      -9.56 2.19 315  -4.369  0.0001
##  C - I      14.65 1.16 314  12.673  <.0001
##  D - I      24.21 2.21 315  10.971  <.0001
##
## Degrees-of-freedom method: kenward-roger
## P value adjustment: tukey method for comparing a family of 4 estimates
```

1) Report and interpret the results of the linear model. (Note that the output for each level of comstat will be reported as contrasts of that level in relation to the reference level - refer to "ANCOVA exercise" we did in class a few weeks ago with a similar dataset).

The model was set to have a baseline of "suppressed", which results in other levels indicating increases in growth compared to the harshest growing conditions. Codominant (slope $= 17.244 \pm 1.504$, $t(319) = 11.462$, $p < 2e\text{-}16$), dominant (slope $= 29.005 \pm 2.549$, $t(319) = 11.379$, $p < 2e\text{-}16$), and intermediate (slope $= 3.569$

± 1.561, t(319) = 2.286, p = 0.0229) all show statistically significant differences from baseline. The complete model (dbh2 = 10.870 + 17.244×(C) + 29.005×(D) + 3.569×(I)) was statistically significant (R2 = 0.4404, Adjusted R2 = 0.4352, F(3, 319) = 83.7, p < 2.2e-16) indicating competitive status has an effect on dbh.

2) Report the results of the mixed-model, including:

a) the fixed effects estimates,

Suppressed was still chosen as the baseline. Codominant (slope = 19.385 ± 1.352, t = 14.336), dominant (slope = 28.947 ± 2.323, t = 12.459), and intermediate (slope = 4.733 ± 1.339, t = 3.535) all show statistically significant differences from the baseline.

b) the confidence limits,

The 95% confidence interval for the intercept (suppressed trees) ranged from 6.6 to 16.4. Codominant trees showed an estimated difference of 19.3 cm/year above suppressed trees, with a 95% confidence interval of 16.7 to 22.0. Dominant trees showed an estimated difference of 28.9 cm/year, with a 95% confidence interval of 24.3 to 33.5. Intermediate trees showed an estimated difference of 4.7 cm/year, with a 95% confidence interval of 2.1 to 7.3.

c) the variance explained by the fixed effects and by the random effects,

The mixed model indicated substantial variation among plots. The plot (random variable) accounted for a variance of 46.1 (SD = 6.7). The residual plot variation, representing within-plot variation, was 70.9 (SD = 8.4). Thus, a notable portion of the total variability in growth rates was explained by differences among plots.

d) the post-hoc comparisons.

The posthoc showed significant differences between all competitive status categories. Suppressed trees grew significantly more slowly than codominant (p < 0.0001), dominant (p < 0.0001), and intermediate trees (p = 0.0027). Codominant trees differed significantly from both dominant trees (p = 0.0001) and intermediate trees (p < 0.0001). Dominant trees also grew significantly faster than intermediate trees (p < 0.0001).

3) How did the results of the two models differ? To what extent (if any) did running the mixed model clarify the relationship between the explanatory and response variables?

Both the linear and mixed models both show significant differences between the competitive status categories. The mixed model showed that much of the variation was between plots (variance 46.1 (SD = 6.7)) which indicates differences in the ecology of the plots. This variation could be further explored with more envirnmental variables.