BlakeLinearModels

Kylie Blake

2025-04-02

Contents


```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lme4)
```

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

```
library(emmeans) #modeled means
```

```
## Warning: package 'emmeans' was built under R version 4.4.3
```

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

```r
library(multcomp) #multiple comparison
```

```
## Warning: package 'multcomp' was built under R version 4.4.3
```

```
## Loading required package: mvtnorm
```

```
## Warning: package 'mvtnorm' was built under R version 4.4.3
```

```
## Loading required package: survival
## Loading required package: TH.data
```

```
## Warning: package 'TH.data' was built under R version 4.4.3
```

```
## Loading required package: MASS
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
##
##
## Attaching package: 'TH.data'
##
## The following object is masked from 'package:MASS':
##
##     geyser
```

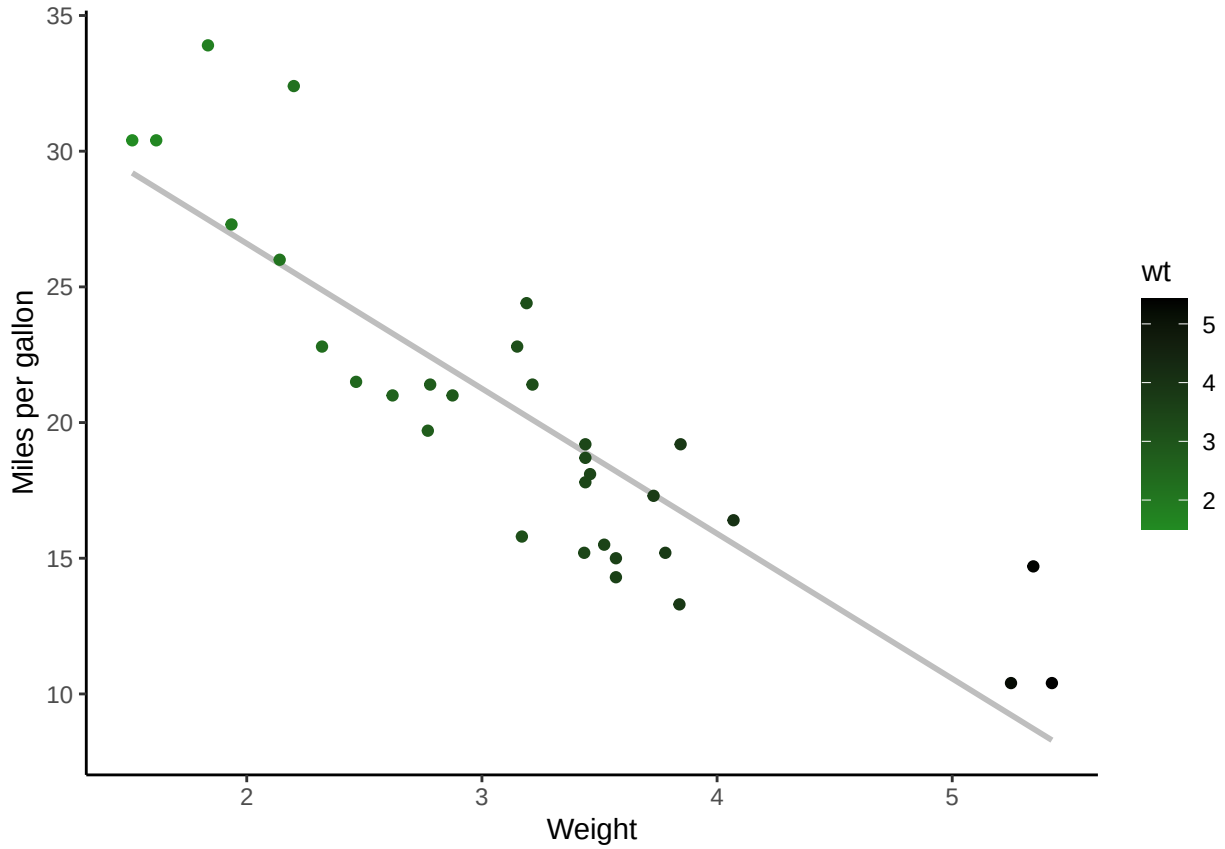
Continuous X and Continuous Y

Notes on Linear Regression: We run a linear model - or if it's a **continuous x variable and a continuous y variable**, we would call it a **regression**. If we consider it a **cause-and-effect** relationship, we may call it a **correlation.**

```r
data("mtcars")

ggplot(mtcars, aes(x=wt, y=mpg)) +
  geom_smooth(method = lm, se = FALSE, color = "grey")+
  geom_point(aes(color=wt)) +
  xlab("Weight") +
  ylab("Miles per gallon") +
  scale_color_gradient(low= "forestgreen", high="black")+
  theme_classic()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#to calculate that linear regression in the above plot

lm1 <- lm(mpg~wt, data= mtcars) #lm = linear model -- When we run this it gives us an output of the pre
#gives the intercept and slope of wt

#We can now run summary of this linear model to output some summary statistics
summary(lm1)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

```r
#the slope is not equal to 0, and the intercept estimate is pretty good. R squared tells you the variat

anova(lm1) #regression, saying weight is significant
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## wt         1 847.73  847.73  91.375 1.294e-10 ***
## Residuals 30 278.32    9.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

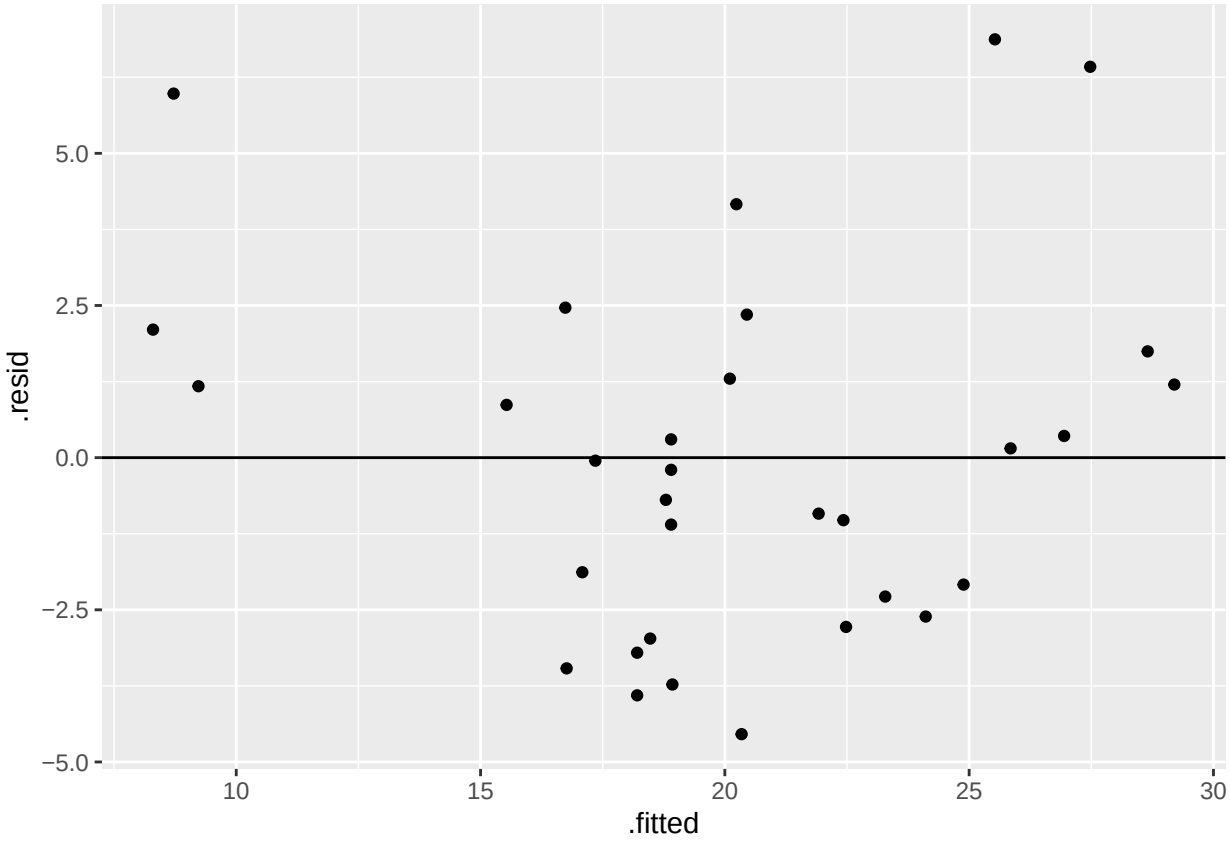
```r
#P-value is same in ANOVA as with the summary; linear model and an ANOVA are essentially the same thing

cor.test(mtcars$wt, mtcars$mpg)
```

```
##
##  Pearson's product-moment correlation
##
## data:  mtcars$wt and mtcars$mpg
## t = -9.559, df = 30, p-value = 1.294e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9338264 -0.7440872
## sample estimates:
##        cor
## -0.8676594
```

```r
#p-value is the same again! This gives you a different r value, which is the correlation statistic. The

ggplot(lm1, aes(y=.resid, x=.fitted)) +
  geom_point()+
  geom_hline(yintercept=0)
```



##Assumptions Notes from Dr. Noel: In general, there are several assumptions in a regression, linear model, ANOVA, or whatever you want to call it.

They are:

- y is continuous
- error is normally distributed
- relationship is linear
- homoskedasticity
- sigma is consistent
- independent samples However, regression/anovas/linear models are generally robust enough to moderate departures from the assumptions. So, do the assumptions matter... yeah, but only if they are terrible. However, in most cases, if one assumption is violated, it won't change the result too much. If anything, it will increase the p-value, and your conclusion is more conservative at that point.

Yes, there are assumption tests, but we will not do them. You just need to know how to read your data and if you want look at the residual plot to diagnose your violated assumptions.

In our example above we can simply get the residuals from our linear model like this

Categorical variables

```
bull.rich <- read.csv("CodingChallenges/Bull_richness.csv")

bull.rich.sub <-  bull.rich %>%
```

```
  filter(GrowthStage =="V8" & Treatment == "Conv.")

#t-test allows us to test and see if there's a difference in richness depending on the fungicide used
t.test(richness ~ Fungicide, data = bull.rich.sub)
```

```
##
##  Welch Two Sample t-test
##
## data:  richness by Fungicide
## t = 4.8759, df = 17.166, p-value = 0.0001384
## alternative hypothesis: true difference in means between group C and group F is not equal to 0
## 95 percent confidence interval:
##   4.067909 10.265425
## sample estimates:
## mean in group C mean in group F
##       11.750000        4.583333
```

```
#null hypothesis which is diff means = 0 so trying to prove that wrong, We have evidence means are diff

summary(lm(richness~Fungicide, data = bull.rich.sub)) #other way to look at summary stats
```

```
##
## Call:
## lm(formula = richness ~ Fungicide, data = bull.rich.sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7500 -1.7500 -0.6667  2.2500  7.2500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.750      1.039  11.306 1.24e-10 ***
## FungicideF    -7.167      1.470  -4.876 7.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.6 on 22 degrees of freedom
## Multiple R-squared:  0.5194, Adjusted R-squared:  0.4975
## F-statistic: 23.77 on 1 and 22 DF,  p-value: 7.118e-05
```

```
anova(lm(richness~Fungicide, data=bull.rich.sub))
```

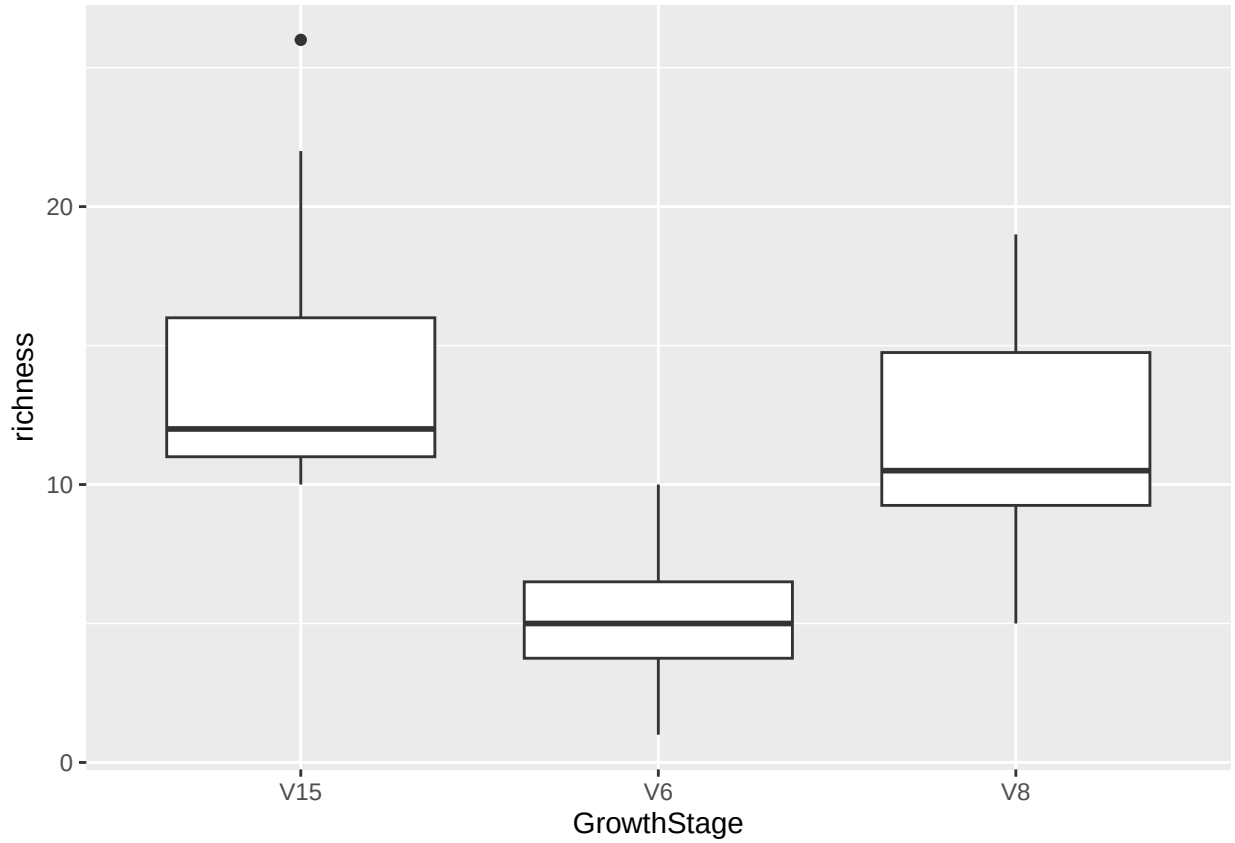
```
## Analysis of Variance Table
##
## Response: richness
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Fungicide  1 308.17 308.167  23.774 7.118e-05 ***
## Residuals 22 285.17  12.962
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Assuming equal variance in groups and performing a two-sample t-test is the same result as a linear mo
```

#ANOVA

```r
#Filter dataset for richness in diff crop growth stages in control and soybean in conventional manageme
bull.rich.sub2 <- bull.rich %>%
  filter(Fungicide == "C" & Treatment == "Conv." & Crop == "Corn")

ggplot(bull.rich.sub2, aes(x=GrowthStage, y=richness))+
  geom_boxplot()
```




```r
lm3 <- lm(richness~ GrowthStage, data = bull.rich.sub2)
summary(lm(richness~ GrowthStage, data = bull.rich.sub2))
```

```
##
## Call:
## lm(formula = richness ~ GrowthStage, data = bull.rich.sub2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.750 -2.625 -1.000  2.250 11.583
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)     14.417      1.208  11.939 1.60e-13 ***
## GrowthStageV6   -9.167      1.708  -5.368 6.23e-06 ***
## GrowthStageV8   -2.667      1.708  -1.562    0.128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.183 on 33 degrees of freedom
## Multiple R-squared:  0.4803, Adjusted R-squared:  0.4488
## F-statistic: 15.25 on 2 and 33 DF,  p-value: 2.044e-05
```

```
#overall effect of growth stage on richness is significant! yes model is significant and is all a linea

#To understand which groups are different from each other, run ANOVA, report ANOVA table for sig factor
```

Pos-hoc is basically individual t-tests across groups. The most versatile way to do this is with the packages emmeans, and multcomp. The lsmeans are the least squared means - the means estimated by the linear model. This contrasts the arithmetic means, which are the means calculated or the average.

```
#Post-hoc test

#modeled means using emmeans
lsmeans <- emmeans(lm3, ~GrowthStage)
lsmeans #gives us the means of each growth stage and their upper and lower confidence intervals
```

```
##  GrowthStage emmean   SE df lower.CL upper.CL
##  V15          14.42 1.21 33    11.96    16.87
##  V6            5.25 1.21 33     2.79     7.71
##  V8           11.75 1.21 33     9.29    14.21
##
## Confidence level used: 0.95
```

```
#Set up compact letter display
results_lsmeans <-  cld(lsmeans, alpha = 0.05, details = TRUE)
results_lsmeans #outputs which groups are diff from another, groups that dont share same numbers are st
```

```
## $emmeans
##  GrowthStage emmean   SE df lower.CL upper.CL .group
##  V6            5.25 1.21 33     2.79     7.71  1
##  V8           11.75 1.21 33     9.29    14.21   2
##  V15          14.42 1.21 33    11.96    16.87   2
##
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 3 estimates
## significance level used: alpha = 0.05
## NOTE: If two or more means share the same grouping symbol,
##       then we cannot show them to be different.
##       But we also did not show them to be the same.
##
## $comparisons
##  contrast estimate   SE df t.ratio p.value
##  V8 - V6      6.50 1.71 33   3.806  0.0016
##  V15 - V6     9.17 1.71 33   5.368  <.0001
```

```
##  V15 - V8     2.67 1.71 33   1.562  0.2763
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

#Interaction Term

```
bull.rich.sub3 <-  bull.rich %>%
  filter(Treatment== "Conv."& Crop == "Corn")

#We can test interactions between factors within a linear model using the * between factor
lm.interaction <- lm(richness~GrowthStage * Fungicide, data = bull.rich.sub3)
#other way to write this: lm.inter <- lm(richness ~ GrowthStage + Fungicide + GrowthStage:Fungicide, da
summary(lm.interaction)
```

```
##
## Call:
## lm(formula = richness ~ GrowthStage * Fungicide, data = bull.rich.sub3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5000 -2.4167 -0.4167  2.0625 11.5833
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                14.4167     1.1029  13.072  < 2e-16 ***
## GrowthStageV6              -9.1667     1.5597  -5.877 1.51e-07 ***
## GrowthStageV8              -2.6667     1.5597  -1.710   0.0920 .
## FungicideF                 -0.9167     1.5597  -0.588   0.5587
## GrowthStageV6:FungicideF   -0.3333     2.2057  -0.151   0.8803
## GrowthStageV8:FungicideF   -6.2500     2.2057  -2.834   0.0061 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.82 on 66 degrees of freedom
## Multiple R-squared:  0.5903, Adjusted R-squared:  0.5593
## F-statistic: 19.02 on 5 and 66 DF,  p-value: 1.144e-11
```

```
anova(lm.interaction)
```

```
## Analysis of Variance Table
##
## Response: richness
##                       Df  Sum Sq Mean Sq F value    Pr(>F)
## GrowthStage            2 1065.58  532.79 36.5027 2.113e-11 ***
## Fungicide              1  174.22  174.22 11.9363 0.0009668 ***
## GrowthStage:Fungicide  2  148.36   74.18  5.0823 0.0088534 **
## Residuals             66  963.33   14.60
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

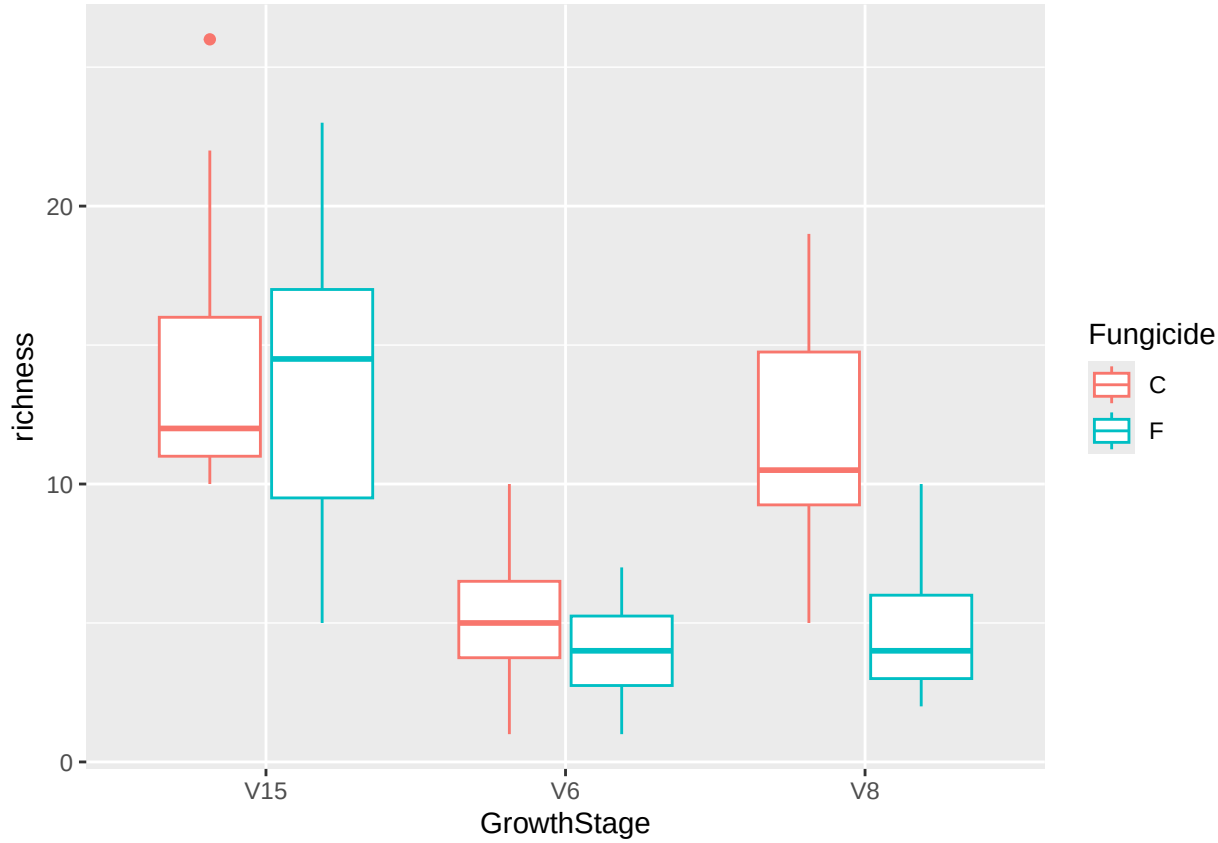
```
#ANOVA tells us that bc fungicide is sig it depends on when it was applied (Growth stage)
```

```r
#How to parse out some of the other meaning behind the variables:

 # estimate lsmeans of variety within siteXyear
lsmeans <- emmeans(lm.interaction, ~Fungicide|GrowthStage) #| = within, see Fungicide effect WITHIN (|)
results_lsmeans <-  cld(lsmeans, alpha = 0.05, details = TRUE)
results_lsmeans #shows us the comparison of means between fungicide and control at different growth sta
```

```
## $emmeans
## GrowthStage = V15:
##  Fungicide emmean  SE df lower.CL upper.CL .group
##  F          13.50 1.1 66    11.30    15.70  1
##  C          14.42 1.1 66    12.21    16.62  1
##
## GrowthStage = V6:
##  Fungicide emmean  SE df lower.CL upper.CL .group
##  F           4.00 1.1 66     1.80     6.20  1
##  C           5.25 1.1 66     3.05     7.45  1
##
## GrowthStage = V8:
##  Fungicide emmean  SE df lower.CL upper.CL .group
##  F           4.58 1.1 66     2.38     6.79  1
##  C          11.75 1.1 66     9.55    13.95   2
##
## Confidence level used: 0.95
## significance level used: alpha = 0.05
## NOTE: If two or more means share the same grouping symbol,
##       then we cannot show them to be different.
##       But we also did not show them to be the same.
##
## $comparisons
## GrowthStage = V15:
##  contrast estimate   SE df t.ratio p.value
##  C - F       0.917 1.56 66   0.588  0.5587
##
## GrowthStage = V6:
##  contrast estimate   SE df t.ratio p.value
##  C - F       1.250 1.56 66   0.801  0.4258
##
## GrowthStage = V8:
##  contrast estimate   SE df t.ratio p.value
##  C - F       7.167 1.56 66   4.595  <.0001
```

```r
#To see our interaction terms
ggplot(bull.rich.sub3, aes(x=GrowthStage, y=richness, color=Fungicide)) +
  geom_boxplot()
```



#Mixed Effect Models

- Things that you want to generalize variation
- Fixed effect: if we care about effect of variables interactions

Common fixed effects

- Treatment
- Species
- gene

Common random effects (blocking factor) - Year - replicate - trial - Individuals - Fields

```
# Load the lme4 package to use linear mixed-effects models
library(lme4)

# Fit a linear mixed-effects model with 'richness' as the response variable
# 'GrowthStage' and 'Fungicide' as fixed effects, and 'Rep' as a random effect
lm.interaction2 <- lmer(richness ~ GrowthStage * Fungicide + (1 | Rep), data = bull.rich.sub3)

# Display the summary of the fitted model, which includes estimates of fixed effects,
# random effects, t-values, and standard errors
summary(lm.interaction2)
```

```
## Linear mixed model fit by REML ['lmerMod']
```

```
## Formula: richness ~ GrowthStage * Fungicide + (1 | Rep)
##    Data: bull.rich.sub3
##
## REML criterion at convergence: 378.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4664 -0.5966 -0.1788  0.6257  2.9101
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Rep      (Intercept)  0.7855  0.8863
##  Residual             13.9533  3.7354
## Number of obs: 72, groups:  Rep, 4
##
## Fixed effects:
##                            Estimate Std. Error t value
## (Intercept)                 14.4167     1.1658  12.366
## GrowthStageV6               -9.1667     1.5250  -6.011
## GrowthStageV8               -2.6667     1.5250  -1.749
## FungicideF                  -0.9167     1.5250  -0.601
## GrowthStageV6:FungicideF    -0.3333     2.1566  -0.155
## GrowthStageV8:FungicideF    -6.2500     2.1566  -2.898
##
## Correlation of Fixed Effects:
##             (Intr) GrwSV6 GrwSV8 FngcdF GSV6:F
## GrowthStgV6 -0.654
## GrowthStgV8 -0.654  0.500
## FungicideF  -0.654  0.500  0.500
## GrwthSV6:FF  0.462 -0.707 -0.354 -0.707
## GrwthSV8:FF  0.462 -0.354 -0.707 -0.707  0.500
```

```
# Display the summary of fitted model to compare
summary(lm.interaction)
```

```
##
## Call:
## lm(formula = richness ~ GrowthStage * Fungicide, data = bull.rich.sub3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5000 -2.4167 -0.4167  2.0625 11.5833
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 14.4167     1.1029  13.072  < 2e-16 ***
## GrowthStageV6               -9.1667     1.5597  -5.877 1.51e-07 ***
## GrowthStageV8               -2.6667     1.5597  -1.710   0.0920 .
## FungicideF                  -0.9167     1.5597  -0.588   0.5587
## GrowthStageV6:FungicideF    -0.3333     2.2057  -0.151   0.8803
## GrowthStageV8:FungicideF    -6.2500     2.2057  -2.834   0.0061 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.82 on 66 degrees of freedom
## Multiple R-squared:  0.5903, Adjusted R-squared:  0.5593
## F-statistic: 19.02 on 5 and 66 DF,  p-value: 1.144e-11
```

```
# Note: The summary does not provide p-values directly; focus on t-values and standard errors instead

# Use the 'emmeans' package to calculate estimated marginal means (least-squares means)
# for 'Fungicide' within each 'GrowthStage'
lsmeans <- emmeans(lm.interaction2, ~ Fungicide | GrowthStage)

# Display the estimated marginal means, along with their upper and lower confidence intervals
lsmeans
```

```
## GrowthStage = V15:
##  Fungicide emmean   SE   df lower.CL upper.CL
##  C          14.42 1.17 28.1    12.03    16.80
##  F          13.50 1.17 28.1    11.11    15.89
##
## GrowthStage = V6:
##  Fungicide emmean   SE   df lower.CL upper.CL
##  C           5.25 1.17 28.1     2.86     7.64
##  F           4.00 1.17 28.1     1.61     6.39
##
## GrowthStage = V8:
##  Fungicide emmean   SE   df lower.CL upper.CL
##  C          11.75 1.17 28.1     9.36    14.14
##  F           4.58 1.17 28.1     2.20     6.97
##
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.95
```

```
# Use the 'cld' function to perform pairwise comparisons of the estimated marginal means
# and determine which groups are significantly different from each other
results_lsmeans <- cld(lsmeans, alpha = 0.05, details = TRUE)

# Display the results of the pairwise comparisons, indicating groups that do not share the same letters
results_lsmeans
```

```
## $emmeans
## GrowthStage = V15:
##  Fungicide emmean   SE   df lower.CL upper.CL .group
##  F          13.50 1.17 28.1    11.11    15.89  1
##  C          14.42 1.17 28.1    12.03    16.80  1
##
## GrowthStage = V6:
##  Fungicide emmean   SE   df lower.CL upper.CL .group
##  F           4.00 1.17 28.1     1.61     6.39  1
##  C           5.25 1.17 28.1     2.86     7.64  1
##
## GrowthStage = V8:
##  Fungicide emmean   SE   df lower.CL upper.CL .group
##  F           4.58 1.17 28.1     2.20     6.97  1
##  C          11.75 1.17 28.1     9.36    14.14   2
```

```
##
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.95
## significance level used: alpha = 0.05
## NOTE: If two or more means share the same grouping symbol,
##       then we cannot show them to be different.
##       But we also did not show them to be the same.
##
## $comparisons
## GrowthStage = V15:
##  contrast estimate   SE df t.ratio p.value
##  C - F       0.917 1.52 63   0.601  0.5499
##
## GrowthStage = V6:
##  contrast estimate   SE df t.ratio p.value
##  C - F       1.250 1.52 63   0.820  0.4155
##
## GrowthStage = V8:
##  contrast estimate   SE df t.ratio p.value
##  C - F       7.167 1.52 63   4.700  <.0001
##
## Degrees-of-freedom method: kenward-roger
```