CodingNotesLinearModels

Mamata K C

2025-04-03

```r
#loading all necessary packages
library(ggplot2)
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 lubridate 1.9.3      v tibble    3.2.1
## v purrr     1.0.2      v tidyr     1.3.1
## -- 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
```

```r
library(multcomp)
```

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

```r
library(lme4)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
```

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

```
library(emmeans)
```

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

```
data("mtcars") #loading data
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_colour_gradient(low = "forestgreen", high = "black") +
  theme_classic()
```

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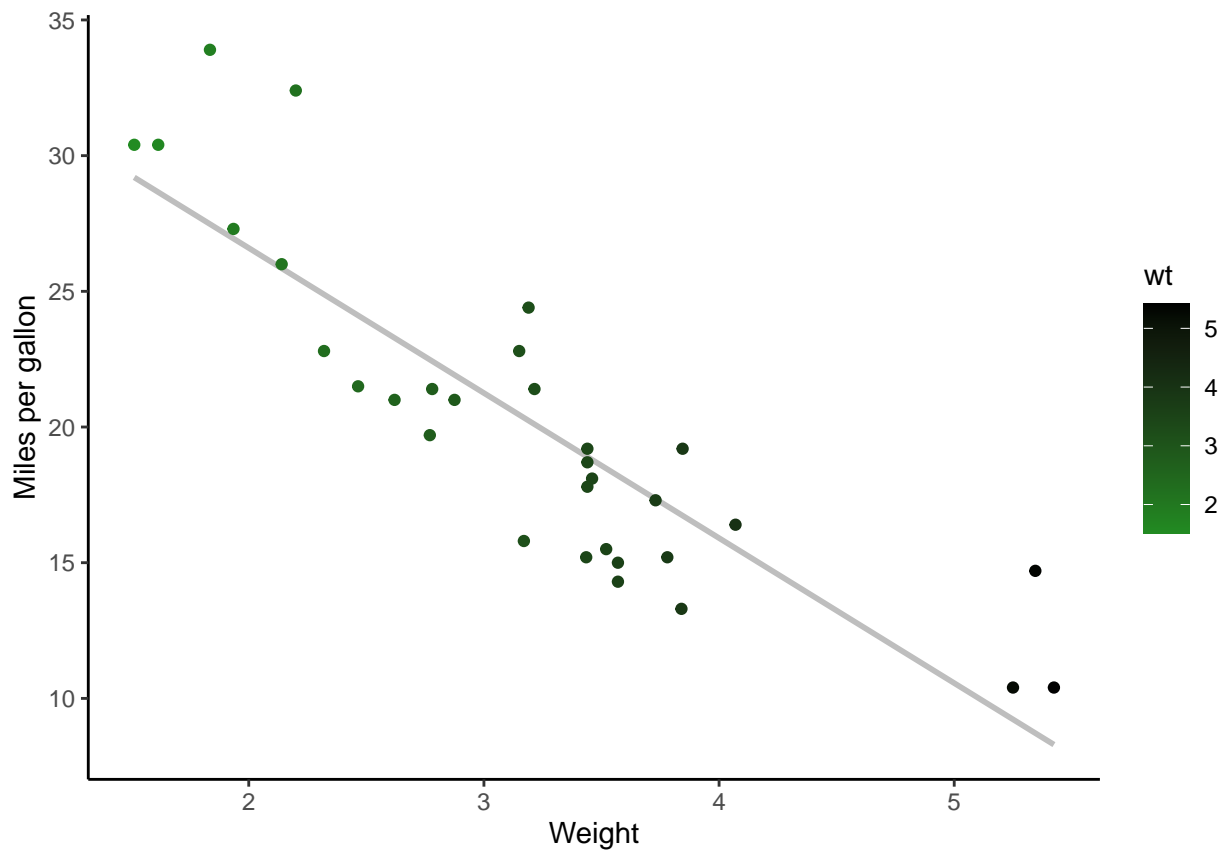

- to know if there is relationship between miles per gallon and weight, we run linear model which gives the estimate for our intercept.

```r
lm1 <- lm(mpg~wt, data = mtcars) #running linear model
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
anova(lm(mpg~wt, data = mtcars)) #running an anova (gives anova table)
```

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

```
##
##  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
```

- The value of multiple R squared gives the idea that 75% of the variation in miles per gallon is due to weight.
- linear model, ANOVA or correlation, all are basically the same thing, gives same p value and all are based on regression.

```
bull.rich <- read.csv("Bull_richness.csv") #loading data
bull.rich %>%
  filter(GrowthStage == "V8" & Treatment == "Conv.") %>%
  ggplot(aes(x = Fungicide, y = richness)) +
  geom_boxplot()
```
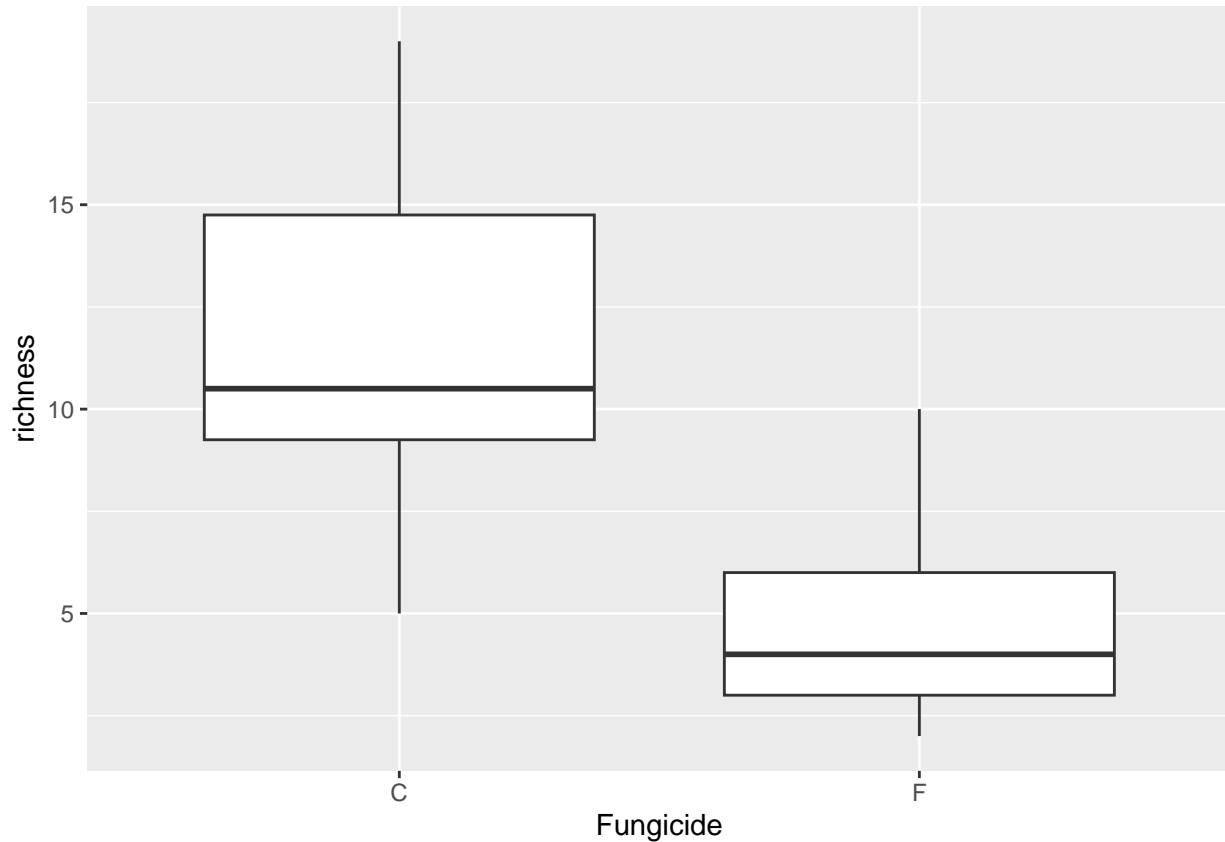

- Again, looking at the boxplot, it looks like there is a difference between the treatments so we can do
  t.test (ANOVA with two groups).

```
bull.rich.sub <- bull.rich %>%
  filter(GrowthStage == "V8" & Treatment == "Conv.") #subset 1

t.test(richness~Fungicide, data = bull.rich.sub) #t test
```

```
##
##  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
```

```r
summary(lm(richness~Fungicide, data = bull.rich.sub)) #linear model
```

```
##
## 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
```

```r
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
```

- Again, all of the tests give same result.

```r
bull.rich.sub2 <- bull.rich %>%
  filter(Fungicide == "C" & Treatment == "Conv." & Crop == "Corn") #subset 2

bull.rich.sub2$GrowthStage <- factor(bull.rich.sub2$GrowthStage, levels = c("V6", "V8", "V15"))

ggplot(bull.rich.sub2, aes(x = GrowthStage, y = richness)) +
  geom_boxplot() #plotting the subset data
```
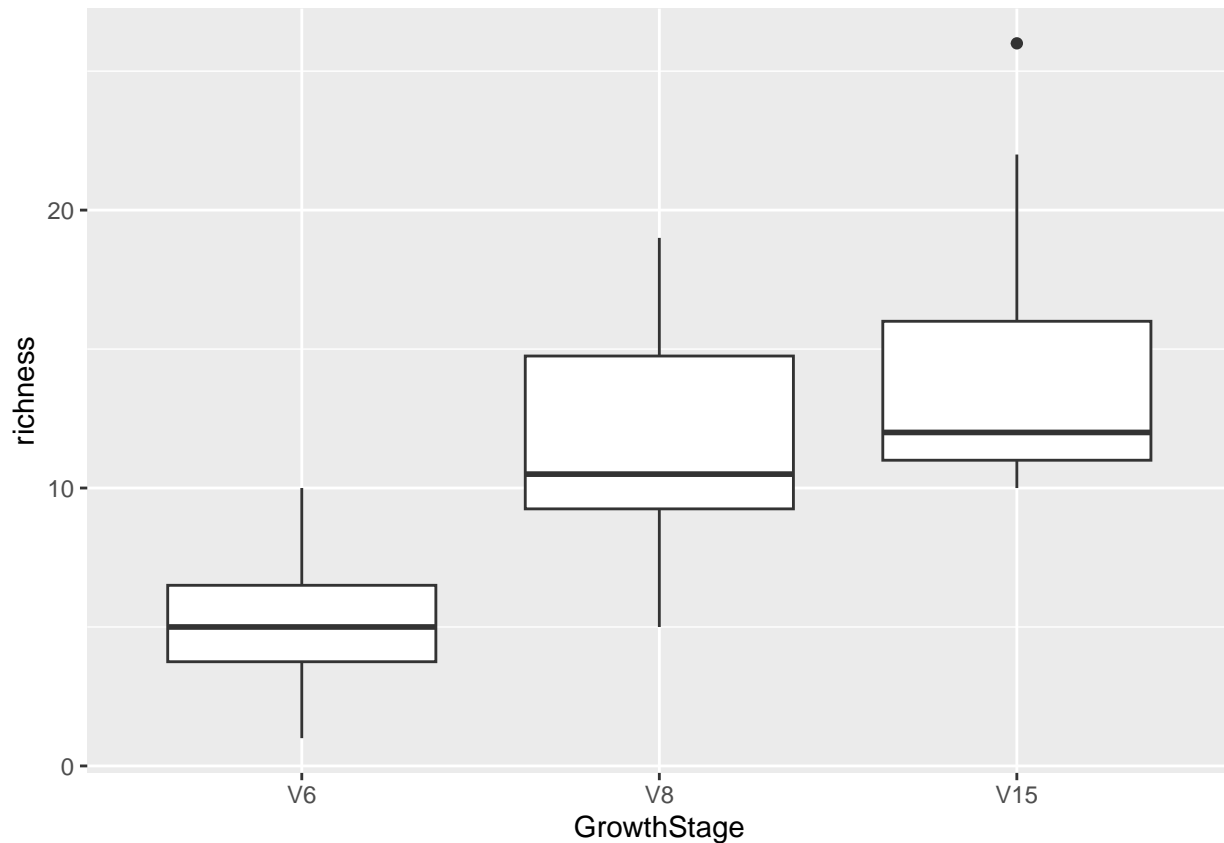
```
lm.growth <- lm(richness ~ GrowthStage, data = bull.rich.sub2) #linear model
summary(lm.growth)
```

```
##
## 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)      5.250      1.208   4.348 0.000124 ***
## GrowthStageV8    6.500      1.708   3.806 0.000581 ***
## GrowthStageV15   9.167      1.708   5.368 6.23e-06 ***
## ---
## 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
```

```
#ANOVAs
anova(lm.growth)
```

```
## Analysis of Variance Table
##
## Response: richness
##             Df Sum Sq Mean Sq F value    Pr(>F)
## GrowthStage  2 533.56 266.778  15.247 2.044e-05 ***
## Residuals   33 577.42  17.497
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(richness ~ GrowthStage, data = bull.rich.sub2))
```

```
##             Df Sum Sq Mean Sq F value    Pr(>F)
## GrowthStage  2  533.6   266.8   15.25 2.04e-05 ***
## Residuals   33  577.4    17.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- All the above resutls have same p value again.
- To find out which groups are different from each other, we can do post hoc test.

```
lsmeans <- emmeans(lm.growth, ~GrowthStage)
Results_lsmeans <- cld(lsmeans, alpha = 0.05, reversed = TRUE, details = TRUE)
Results_lsmeans
```

```
## $emmeans
##  GrowthStage emmean   SE df lower.CL upper.CL .group
##  V15          14.42 1.21 33    11.96    16.87  1
##  V8           11.75 1.21 33     9.29    14.21  1
##  V6            5.25 1.21 33     2.79     7.71   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
```

- emmeans gives means of our model distribution for each categories and their confidence interval.
- also outputs which group are different from each other.

```
bull.rich.sub3 <- bull.rich %>%
  filter(Treatment == "Conv." & Crop == "Corn")  #subset3
```

```
bull.rich.sub3$GrowthStage <- factor(bull.rich.sub3$GrowthStage, levels = c("V6", "V8", "V15"))
```

```
lm.interactions <- lm(richness ~ GrowthStage*Fungicide, data = bull.rich.sub3) #interaction between grou
summary(lm.interactions)
```

```
##
## 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)                 5.2500     1.1029   4.760 1.10e-05 ***
## GrowthStageV8               6.5000     1.5597   4.167 9.15e-05 ***
## GrowthStageV15              9.1667     1.5597   5.877 1.51e-07 ***
## FungicideF                 -1.2500     1.5597  -0.801  0.42575
## GrowthStageV8:FungicideF   -5.9167     2.2057  -2.682  0.00923 **
## GrowthStageV15:FungicideF   0.3333     2.2057   0.151  0.88034
## ---
## 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.interactions)
```

```
## 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
```

```
lsmeans <- emmeans(lm.interactions, ~Fungicide|GrowthStage) # effect of fungicide within growthstage
Results_lsmeans <- cld(lsmeans, alpha = 0.05, reversed = TRUE, details = TRUE)
Results_lsmeans
```

```
## $emmeans
## GrowthStage = V6:
##  Fungicide emmean  SE df lower.CL upper.CL .group
##  C           5.25 1.1 66     3.05     7.45  1
##  F           4.00 1.1 66     1.80     6.20  1
##
## GrowthStage = V8:
##  Fungicide emmean  SE df lower.CL upper.CL .group
```

8

```
## C            11.75 1.1 66     9.55    13.95  1
## F             4.58 1.1 66     2.38     6.79  2
##
## GrowthStage = V15:
##  Fungicide emmean  SE df lower.CL upper.CL .group
## C            14.42 1.1 66    12.21    16.62  1
## F            13.50 1.1 66    11.30    15.70  1
##
## 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 = 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
##
## GrowthStage = V15:
##  contrast estimate   SE df t.ratio p.value
##  C - F       0.917 1.56 66   0.588  0.5587
```

```
ggplot(bull.rich.sub3, aes(x = GrowthStage, y = richness, color = Fungicide)) +
  geom_boxplot()
```
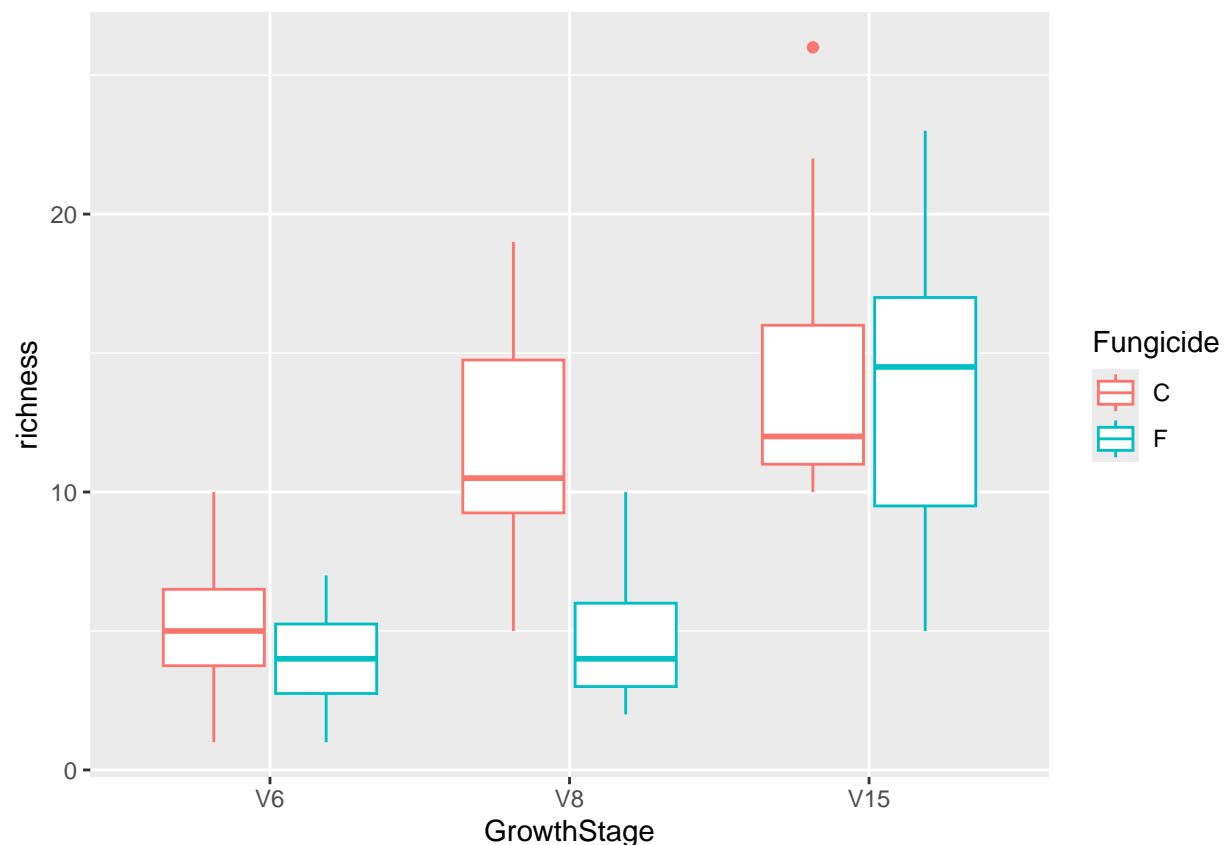
```r
lme1 <- lmer(richness ~ GrowthStage*Fungicide + (1|Rep), data = bull.rich.sub3)
summary(lme1)
```

```
## 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)       5.2500     1.1658   4.503
## GrowthStageV8     6.5000     1.5250   4.262
## GrowthStageV15    9.1667     1.5250   6.011
## FungicideF       -1.2500     1.5250  -0.820
```

```
## GrowthStageV8:FungicideF   -5.9167     2.1566  -2.743
## GrowthStageV15:FungicideF   0.3333     2.1566   0.155
##
## Correlation of Fixed Effects:
##             (Intr) GrwSV8 GrSV15 FngcdF GSV8:F
## GrowthStgV8 -0.654
## GrwthStgV15 -0.654  0.500
## FungicideF  -0.654  0.500  0.500
## GrwthSV8:FF  0.462 -0.707 -0.354 -0.707
## GrwtSV15:FF  0.462 -0.354 -0.707 -0.707  0.500
```

- Here, Rep used as blocking factor
- Rep is random effect.
- Results don't change drastically but we get better ability to determine the effect using mixed effect model.