Linear Models Homework

Katie Temple

2025-04-02

Table of Contents

# Linear Models

*I have taken Dr. Steury’s WILD7150 and Dr. Gitzen’s WILD7250 so this is all review for me* This tutorial is a brief overview/review on linear models. The following information is directly copied from Dr. Noel’s Canvas page: We want to estimate slope, intercept, and standard deviation. Additionally, we want to minimize the distance from each point to the line (minimize the sum of squared errors/SSE).The sum of squares of the regression (SSR) is the distance between the best fit line and the average. Adding these together equals the total sum of squares. Partitioning this into SSR and SSE is related to pvalue. The smaller the SSE and higher the SSR, the smaller the p-value. *I find Dr. Noel’s explanation to be slightly more helpful than steury’s*

In general: - as slope goes up p-value goes down - as sample size goes up p-value goes down - as noise (error) goes down p-value goes down

#load libraries  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ 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)

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

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

## Continuous X and Continuous Y

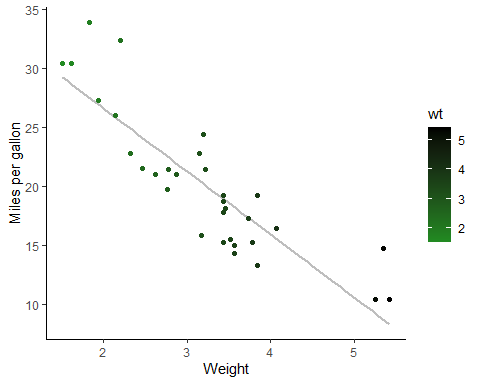
Using the pre-loaded mtcars data from r we will practice some basic linear regression.

The first step is always to visualize your data.

If you have two continuous variables, you will want a scatter plot

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 see confirm if there is a relationship or not, one can run a linear model(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).

lm(mpg ~ wt, data=mtcars) #output gives the predicted intercept/beta0

##   
## Call:  
## lm(formula = mpg ~ wt, data = mtcars)  
##   
## Coefficients:  
## (Intercept) wt   
## 37.285 -5.344

summary(lm(mpg~wt, data=mtcars))

##   
## 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

“Now, we have a pretty good idea that the slope is not equal to 0, and the intercept estimate is pretty good.

Another thing to pay attention to is the R squared. This tells you the variation in y explained by x. So, in our case, about 74% of the variation in y is explained by the x variable.”

anova(lm(mpg~wt, data=mtcars))

## 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

” a linear model and an ANOVA are essentially the same things, and the value reported in the ANOVA is the value of the linear regression or our slope parameter. This means that wt has a significant effect on miles per gallon.”

#Correlation  
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

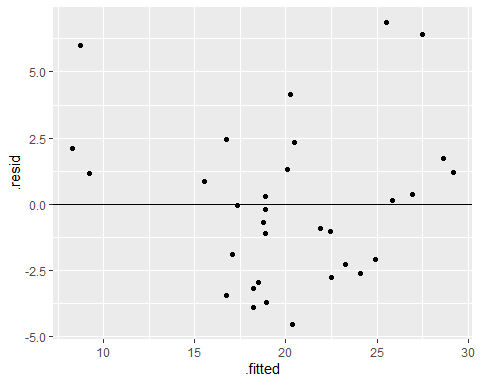
This gives a different r value, which is the correlation stat. The closer to -1 or 1, the stronger the correlation between the two variables

## Assumptions

1. y is continuous
2. error is normally distributed
3. relationship is linear
4. homoskedasticity
5. sigma (standard deviation) is consistent
6. independent samples/ no autocorrelation

The main takeaway is that you need to know how to read your data and if you want to look at the residual plot to diagnose violated assumoptions if any.

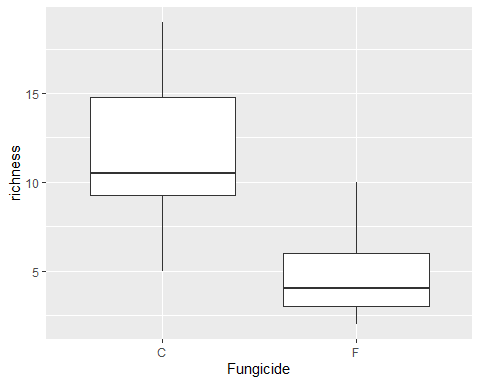
#below is how you can get the residuals from the linear model  
model <- lm(mpg~wt, data = mtcars)  
  
ggplot(model, aes(y = .resid, x = .fitted)) +  
 geom\_point() +  
 geom\_hline(yintercept = 0)



## Cateogorical Variables

Categorical x and continuous y- can preform a t-test

bull.rich <- read.csv("Bull\_richness(1).csv")  
  
#filter data to include one treatment and growth stage for demonstration of a t-test  
bull.rich%>%  
 filter(GrowthStage == "V8" & Treatment == "Conv.")%>%  
 ggplot(aes(x = Fungicide, y = richness))+  
 geom\_boxplot()



#To test if there is actually a meaningful difference between the two, you can use  
#a t test. H0=no difference in groups  
  
bull.rich.sub <- bull.rich %>%  
 filter(GrowthStage == "V8" & Treatment == "Conv.")  
  
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

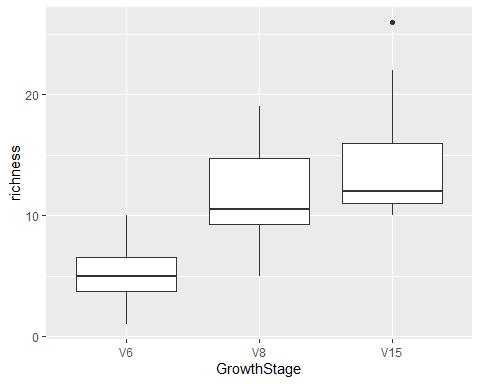
If we run an lm or an ANOVA, the results should be the same assuming there is equal variance. The p-value tells is that the slope of the line is different than 0.

## ANOVAS

Continuous y and multinomial categorical x (more than 2 groups).

In the following example, filter the dataset to only ave richness in different crop growth stages in the control samples in soybean in conventional management. There are three growth stages.

bull.rich.sub2 <- bull.rich %>%  
 filter(Fungicide == "C" & Treatment == "Conv." & Crop == "Corn")  
  
#Visualize the data  
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()



#can run an ANOVA as it's pretty much the same as a regression  
lm.growth <- lm(richness ~ GrowthStage, data = bull.rich.sub2)  
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

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

To find out which is different from the other groups, we can do a post-hoc test. “We can do post-hoc tests to find out. The simplest way to think of this is individual t-tests across groups. The most versatile way to do this is with the packages means and multcomp.

“The lsmeans are the least squared means - the means estimated by the linear model. This contrasts the arithmetic means, which are the emmeans calculated or the average.”

#install.packages("multcompView")  
library(multcompView)  
lsmeans <- emmeans(lm.growth, ~GrowthStage) # estimate lsmeans of variety within siteXyear  
Results\_lsmeans <- cld(lsmeans, alpha = 0.05, reversed = TRUE, details = TRUE) # contrast with Tukey ajustment by default.   
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

## Interaction Terms

Interaction between factors/does the effect of one variable depend on the other.

Ex: Does Fungicide depend on time? Can use \* between factors in an lm.

bull.rich.sub3 <- bull.rich %>%  
 filter(Treatment == "Conv." & Crop == "Corn")  
  
bull.rich.sub3$GrowthStage <- factor(bull.rich.sub3$GrowthStage, levels = c("V6", "V8", "V15"))  
  
#two ways to write the iteraction factor demonstrated below  
lm.inter <- lm(richness ~ GrowthStage + Fungicide + GrowthStage:Fungicide, data = bull.rich.sub3)  
summary(lm.inter) #significant terms

##   
## Call:  
## lm(formula = richness ~ GrowthStage + Fungicide + 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

lm(richness ~ GrowthStage\*Fungicide, data = bull.rich.sub3)

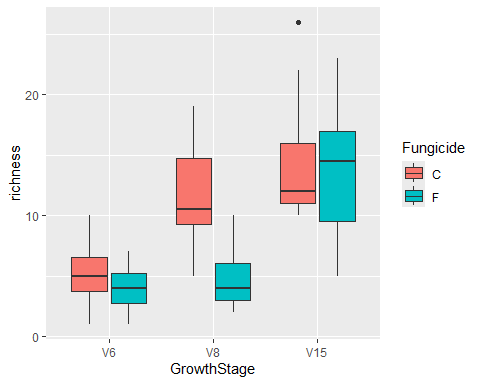
##   
## Call:  
## lm(formula = richness ~ GrowthStage \* Fungicide, data = bull.rich.sub3)  
##   
## Coefficients:  
## (Intercept) GrowthStageV8   
## 5.2500 6.5000   
## GrowthStageV15 FungicideF   
## 9.1667 -1.2500   
## GrowthStageV8:FungicideF GrowthStageV15:FungicideF   
## -5.9167 0.3333

anova(lm.inter)

## 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

I think the explanation provided word for word by Dr. Noel is the most straightforward and is therefore copied below. “So the interaction term is significant. In general you want to look at the most complex interaction that is significant and look at that first. But this means that the effect of fungicide is dependent on the growthstage, or in other words at least one comparison of fungicide within growthstage or growthstage within fungicide is significant.”

#Effect of fungicide over the levels of growthstage  
bull.rich.sub3 %>%  
 ggplot(aes(x = GrowthStage, y = richness, fill = Fungicide)) +  
 geom\_boxplot()



#Can do a post-hoc test within the levels of growth stage to confirm what is being   
#shown in the plot (that V8 affected overall richness more so than the others and that by #v15 they have recovered.)  
  
lsmeans <- emmeans(lm.inter, ~Fungicide|GrowthStage) # estimate lsmeans of variety within siteXyear  
Results\_lsmeans <- cld(lsmeans, alpha = 0.05, reversed = TRUE, details = TRUE) # contrast with Tukey ajustment  
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  
## 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

## Mixed Effects Models

“In mixed-effects models, we have fixed and random effects terms. The random effects term is something that affects the variation in y. A fixed effect is something that affects the mean of y. There are NO rules that determine what is a fixed or random effect. You have to justify that in your experimental design. Another way to think about this would be factors that you do not really care about affecting the mean but still want to account for that variation in your model so you can generalize over that factor. A common way that random effects are included in experimental designs is called the blocking factor.”

Common fixed effects for ag: treatment, species, gene.

Common random effects (blocking factor): year, replicate, trial, individuals, and fields

lme0 <- lm(richness ~ GrowthStage\*Fungicide, data = bull.rich.sub3)  
  
#use the lme4 package to include a random effect  
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

In the above code we included the replicate variation (spatial field variation) as an error term and the estimate of sigma/standard dev. due to the error that is associated with Rep. To understand what this does, look at the Betas in the original model.

summary(lme0)

##   
## 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

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

“For some terms, it actually makes the standard error of our betas go down, which means we have better estimates of our betas, and we are more confident that we are estimating them correctly in our linear models, which means we have more power to detect differences.”