Statistics

Maria Leticia

2023-04-04

### Question 1.

My project is to study the plant variability impact on soybean yield in Alabama.

The objectives of my project are:

1. Asses the soybean canopy plant-to-plant variability in the field.
2. Evaluate how this variability among plants in the same sowing row can impact soybean grain yield.

For that, we conducted a statewide assessment of plant in the field. The data was collected from commercial soybean fields around Alabama. For each field, 10 consecutive plants were collected in 3 distinct points within the field. The final data consists of the mean value for the 10 plants. To measure the variability among plants, I am using the standard deviation within the 10 plants collected per point. The data I am using for my project consists of plant growth parameters such as:

* Plant height (cm)
* Number of pods per plant
* Number of lateral branches per plant And grain yield production expressed as:
* Seed weight per plant (g)

My hypothesis is that non-uniform canopy among consecutive plants in the same sowing row can cause yield loss. While uniform plants can avoid yield losses.

To test my hypothesis, I will use the following statistical methods:

* Descriptive Statistic (Mean, Minimum, Maximum, Median)
* Pearson’s Correlation Test
* Linear Model

### Question 2.

library (ggplot2)  
library (lme4)

## Warning: package 'lme4' was built under R version 4.2.3

## Loading required package: Matrix

library(emmeans)

## Warning: package 'emmeans' was built under R version 4.2.3

library(multcomp)

## Warning: package 'multcomp' was built under R version 4.2.3

## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Warning: package 'TH.data' was built under R version 4.2.3

## Loading required package: MASS

##   
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':  
##   
## geyser

library (Hmisc)

## Warning: package 'Hmisc' was built under R version 4.2.3

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

library("PerformanceAnalytics")

## Warning: package 'PerformanceAnalytics' was built under R version 4.2.3

## Loading required package: xts

## Warning: package 'xts' was built under R version 4.2.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.2.3

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## ################################### WARNING ###################################  
## # We noticed you have dplyr installed. The dplyr lag() function breaks how #  
## # base R's lag() function is supposed to work, which breaks lag(my\_xts). #  
## # #  
## # If you call library(dplyr) later in this session, then calls to lag(my\_xts) #  
## # that you enter or source() into this session won't work correctly. #  
## # #  
## # All package code is unaffected because it is protected by the R namespace #  
## # mechanism. #  
## # #  
## # Set `options(xts.warn\_dplyr\_breaks\_lag = FALSE)` to suppress this warning. #  
## # #  
## # You can use stats::lag() to make sure you're not using dplyr::lag(), or you #  
## # can add conflictRules('dplyr', exclude = 'lag') to your .Rprofile to stop #  
## # dplyr from breaking base R's lag() function. #  
## ################################### WARNING ###################################

##   
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':  
##   
## legend

soy <- read.csv("Soy153new.csv")  
  
soymean <- soy[, c(3,5,7,9)]  
soysd <- soy [, c(4,6,8,10)]  
  
##Basic Stat  
summary(soymean)

## Height\_mean Pod\_mean Branch\_mean SW\_mean   
## Min. : 31.80 Min. : 7.70 Min. :0.000 Min. : 1.490   
## 1st Qu.: 58.10 1st Qu.:29.48 1st Qu.:1.200 1st Qu.: 7.195   
## Median : 74.70 Median :40.00 Median :2.000 Median :11.277   
## Mean : 72.64 Mean :42.85 Mean :2.292 Mean :12.149   
## 3rd Qu.: 85.80 3rd Qu.:54.62 3rd Qu.:3.200 3rd Qu.:15.835   
## Max. :117.90 Max. :93.50 Max. :6.600 Max. :39.931   
## NA's :1 NA's :1

summary(soysd)

## Height\_sd Pod\_sd Branch\_sd SW\_sd   
## Min. : 1.947 Min. : 3.098 Min. :0.0000 Min. : 0.6498   
## 1st Qu.: 4.040 1st Qu.:10.309 1st Qu.:0.9487 1st Qu.: 2.5179   
## Median : 5.547 Median :15.154 Median :1.3166 Median : 4.0923   
## Mean : 6.061 Mean :16.760 Mean :1.3633 Mean : 4.7397   
## 3rd Qu.: 7.902 3rd Qu.:21.192 3rd Qu.:1.7127 3rd Qu.: 5.7761   
## Max. :16.227 Max. :47.741 Max. :3.1340 Max. :23.2766   
## NA's :1 NA's :1

## Correlation  
cormean <- cor(soymean)  
corsd <- cor (soysd)  
  
##Mean  
rcormean <- rcorr(as.matrix(soymean))  
rcormean$P

## Height\_mean Pod\_mean Branch\_mean SW\_mean  
## Height\_mean NA 1.794824e-09 7.097212e-01 1.460054e-11  
## Pod\_mean 1.794824e-09 NA 0.000000e+00 0.000000e+00  
## Branch\_mean 7.097212e-01 0.000000e+00 NA 4.508941e-07  
## SW\_mean 1.460054e-11 0.000000e+00 4.508941e-07 NA

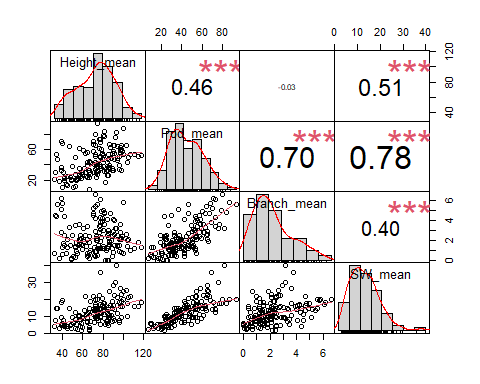
rcormean$r

## Height\_mean Pod\_mean Branch\_mean SW\_mean  
## Height\_mean 1.0000000 0.4636372 -0.0303349 0.5126448  
## Pod\_mean 0.4636372 1.0000000 0.6986956 0.7811380  
## Branch\_mean -0.0303349 0.6986956 1.0000000 0.3957268  
## SW\_mean 0.5126448 0.7811380 0.3957268 1.0000000

chart.Correlation(soymean, histogram=TRUE, pch=19)

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter



##SD  
rcorsd <- rcorr(as.matrix(soysd))  
rcorsd$P

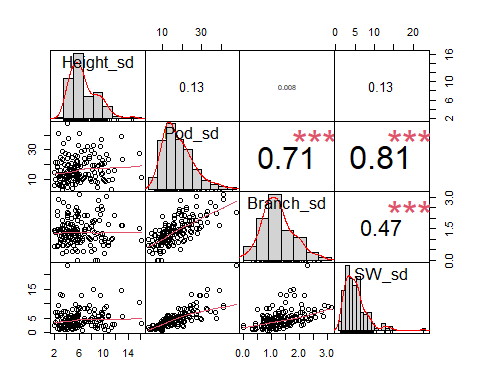
## Height\_sd Pod\_sd Branch\_sd SW\_sd  
## Height\_sd NA 0.1055708 9.218101e-01 1.061280e-01  
## Pod\_sd 0.1055708 NA 0.000000e+00 0.000000e+00  
## Branch\_sd 0.9218101 0.0000000 NA 1.445439e-09  
## SW\_sd 0.1061280 0.0000000 1.445439e-09 NA

rcorsd$r

## Height\_sd Pod\_sd Branch\_sd SW\_sd  
## Height\_sd 1.000000000 0.1317864 0.008000756 0.1315780  
## Pod\_sd 0.131786401 1.0000000 0.713196425 0.8061037  
## Branch\_sd 0.008000756 0.7131964 1.000000000 0.4660169  
## SW\_sd 0.131577991 0.8061037 0.466016863 1.0000000

chart.Correlation(soysd, histogram=TRUE, pch=19)

## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter



##Linear Models for the Means  
lm1 <- lm(Pod\_mean~Height\_mean, data = soy)  
summary(lm1)

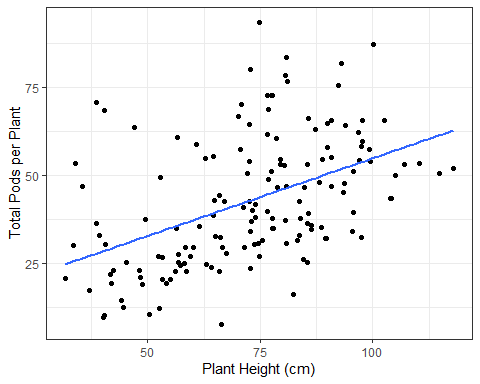
##   
## Call:  
## lm(formula = Pod\_mean ~ Height\_mean, data = soy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.278 -12.475 -4.558 8.743 49.774   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.74012 5.18178 2.073 0.0399 \*   
## Height\_mean 0.44099 0.06881 6.409 1.79e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 16.27 on 150 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.215, Adjusted R-squared: 0.2097   
## F-statistic: 41.07 on 1 and 150 DF, p-value: 1.795e-09

anova (lm1)

## Analysis of Variance Table  
##   
## Response: Pod\_mean  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Height\_mean 1 10877 10877.3 41.073 1.795e-09 \*\*\*  
## Residuals 150 39724 264.8   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

ggplot(lm1, aes(y = Pod\_mean, x = Height\_mean)) +  
 ylab("Total Pods per Plant") +  
 xlab ("Plant Height (cm)") +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 theme\_bw()

## `geom\_smooth()` using formula = 'y ~ x'



lm2 <- lm(SW\_mean~Height\_mean, data = soy)  
summary(lm2)

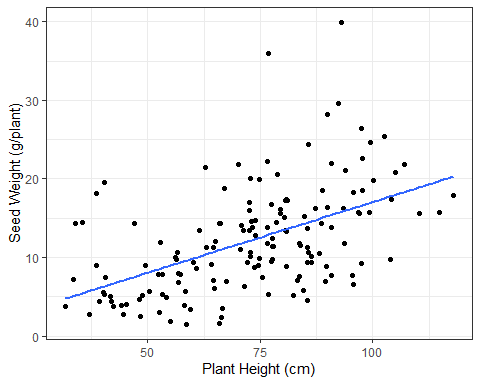
##   
## Call:  
## lm(formula = SW\_mean ~ Height\_mean, data = soy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.8354 -3.6204 -0.7847 2.7953 24.1322   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.95783 1.85353 -0.517 0.606   
## Height\_mean 0.17999 0.02461 7.313 1.46e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.821 on 150 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.2628, Adjusted R-squared: 0.2579   
## F-statistic: 53.47 on 1 and 150 DF, p-value: 1.46e-11

anova (lm2)

## Analysis of Variance Table  
##   
## Response: SW\_mean  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Height\_mean 1 1812.0 1811.95 53.474 1.46e-11 \*\*\*  
## Residuals 150 5082.7 33.88   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

ggplot(lm2, aes(y = SW\_mean, x = Height\_mean)) +  
 ylab("Seed Weight (g/plant)") +  
 xlab ("Plant Height (cm)") +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 theme\_bw()

## `geom\_smooth()` using formula = 'y ~ x'



lm3 <- lm(Pod\_mean~Branch\_mean, data = soy)  
summary(lm3)

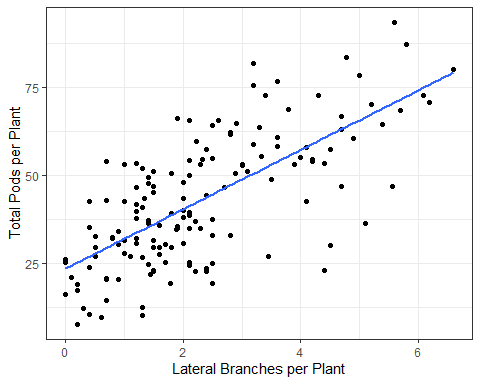
##   
## Call:  
## lm(formula = Pod\_mean ~ Branch\_mean, data = soy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -37.602 -7.631 -0.178 9.299 31.408   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 23.5327 1.9352 12.16 <2e-16 \*\*\*  
## Branch\_mean 8.4248 0.7043 11.96 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.14 on 150 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.4882, Adjusted R-squared: 0.4848   
## F-statistic: 143.1 on 1 and 150 DF, p-value: < 2.2e-16

anova (lm3)

## Analysis of Variance Table  
##   
## Response: Pod\_mean  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Branch\_mean 1 24702 24702.4 143.07 < 2.2e-16 \*\*\*  
## Residuals 150 25899 172.7   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

ggplot(lm3, aes(y = Pod\_mean, x = Branch\_mean)) +  
 ylab("Total Pods per Plant") +  
 xlab ("Lateral Branches per Plant") +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 theme\_bw()

## `geom\_smooth()` using formula = 'y ~ x'



lm4 <- lm(SW\_mean~Pod\_mean, data = soy)  
summary(lm4)

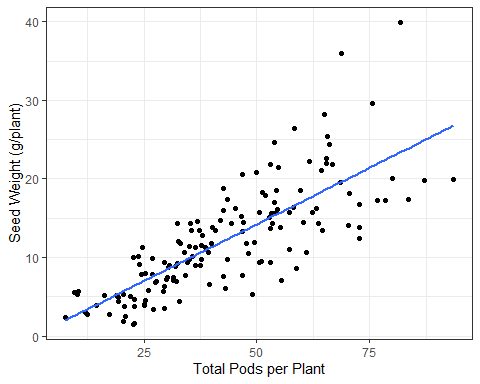
##   
## Call:  
## lm(formula = SW\_mean ~ Pod\_mean, data = soy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.7220 -2.5442 -0.0884 2.4390 16.5234   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.20724 0.87645 -0.236 0.813   
## Pod\_mean 0.28834 0.01882 15.323 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.233 on 150 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.6102, Adjusted R-squared: 0.6076   
## F-statistic: 234.8 on 1 and 150 DF, p-value: < 2.2e-16

anova (lm4)

## Analysis of Variance Table  
##   
## Response: SW\_mean  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Pod\_mean 1 4207.0 4207.0 234.79 < 2.2e-16 \*\*\*  
## Residuals 150 2687.7 17.9   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

ggplot(lm4, aes(y = SW\_mean, x = Pod\_mean)) +  
 ylab("Seed Weight (g/plant)") +  
 xlab ("Total Pods per Plant") +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 theme\_bw()

## `geom\_smooth()` using formula = 'y ~ x'



lm5 <- lm(SW\_mean~Branch\_mean, data = soy)  
summary(lm5)

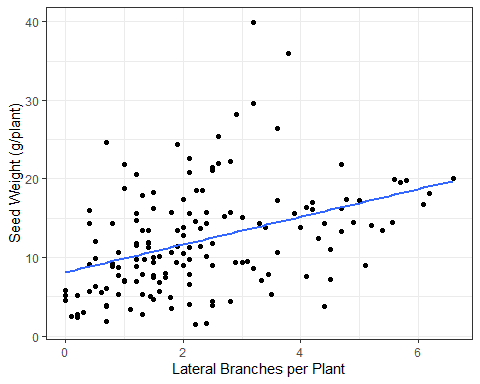
##   
## Call:  
## lm(formula = SW\_mean ~ Branch\_mean, data = soy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.0479 -3.7906 -0.7709 2.3730 26.1850   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.1098 0.9170 8.844 2.35e-15 \*\*\*  
## Branch\_mean 1.7613 0.3337 5.277 4.51e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.226 on 150 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.1566, Adjusted R-squared: 0.151   
## F-statistic: 27.85 on 1 and 150 DF, p-value: 4.509e-07

anova (lm5)

## Analysis of Variance Table  
##   
## Response: SW\_mean  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Branch\_mean 1 1079.7 1079.70 27.852 4.509e-07 \*\*\*  
## Residuals 150 5815.0 38.77   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

ggplot(lm5, aes(y = SW\_mean, x = Branch\_mean)) +  
 ylab("Seed Weight (g/plant)") +  
 xlab ("Lateral Branches per Plant") +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 theme\_bw()

## `geom\_smooth()` using formula = 'y ~ x'



##Linear Models for the standard deviations  
lm6 <- lm(Pod\_sd~Branch\_sd, data = soy)  
summary(lm6)

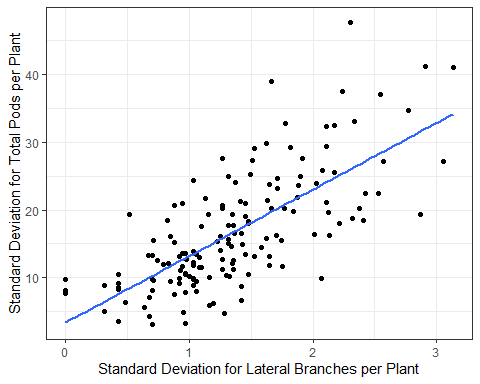
##   
## Call:  
## lm(formula = Pod\_sd ~ Branch\_sd, data = soy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.8669 -4.3375 -0.5356 3.5666 21.7655   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.3555 1.1823 2.838 0.00517 \*\*   
## Branch\_sd 9.8358 0.7893 12.461 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.05 on 150 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.5086, Adjusted R-squared: 0.5054   
## F-statistic: 155.3 on 1 and 150 DF, p-value: < 2.2e-16

anova (lm6)

## Analysis of Variance Table  
##   
## Response: Pod\_sd  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Branch\_sd 1 5683.6 5683.6 155.28 < 2.2e-16 \*\*\*  
## Residuals 150 5490.3 36.6   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

ggplot(lm6, aes(y = Pod\_sd, x = Branch\_sd)) +  
 ylab("Standard Deviation for Total Pods per Plant") +  
 xlab ("Standard Deviation for Lateral Branches per Plant") +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 theme\_bw()

## `geom\_smooth()` using formula = 'y ~ x'



lm7 <- lm(Pod\_sd~SW\_sd, data = soy)  
summary(lm7)

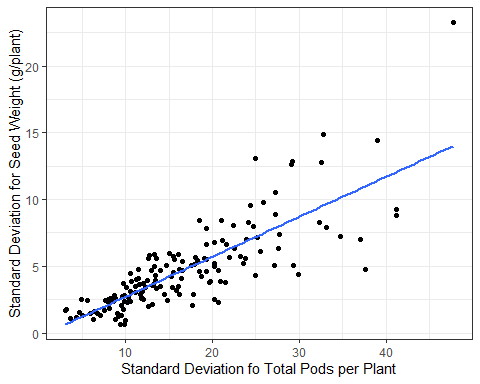
##   
## Call:  
## lm(formula = Pod\_sd ~ SW\_sd, data = soy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.886 -3.352 -1.103 1.961 20.728   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.4720 0.7429 8.712 5.1e-15 \*\*\*  
## SW\_sd 2.1705 0.1301 16.683 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.108 on 150 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.6498, Adjusted R-squared: 0.6475   
## F-statistic: 278.3 on 1 and 150 DF, p-value: < 2.2e-16

anova (lm7)

## Analysis of Variance Table  
##   
## Response: Pod\_sd  
## Df Sum Sq Mean Sq F value Pr(>F)   
## SW\_sd 1 7260.8 7260.8 278.33 < 2.2e-16 \*\*\*  
## Residuals 150 3913.1 26.1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

ggplot(lm7, aes(y = SW\_sd, x = Pod\_sd)) +  
 ylab("Standard Deviation for Seed Weight (g/plant)") +  
 xlab ("Standard Deviation fo Total Pods per Plant") +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 theme\_bw()

## `geom\_smooth()` using formula = 'y ~ x'



lm8 <- lm(Branch\_sd~SW\_sd, data = soy)  
summary(lm8)

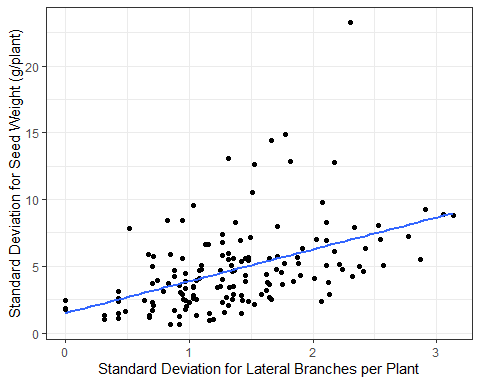
##   
## Call:  
## lm(formula = Branch\_sd ~ SW\_sd, data = soy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.15115 -0.37550 -0.08789 0.35972 1.43802   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.93154 0.08054 11.567 < 2e-16 \*\*\*  
## SW\_sd 0.09099 0.01410 6.451 1.45e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5537 on 150 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.2172, Adjusted R-squared: 0.212   
## F-statistic: 41.61 on 1 and 150 DF, p-value: 1.445e-09

anova (lm8)

## Analysis of Variance Table  
##   
## Response: Branch\_sd  
## Df Sum Sq Mean Sq F value Pr(>F)   
## SW\_sd 1 12.759 12.7586 41.613 1.445e-09 \*\*\*  
## Residuals 150 45.990 0.3066   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

ggplot(lm8, aes(y = SW\_sd, x = Branch\_sd)) +  
 ylab("Standard Deviation for Seed Weight (g/plant)") +  
 xlab ("Standard Deviation for Lateral Branches per Plant") +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 theme\_bw()

## `geom\_smooth()` using formula = 'y ~ x'



### Question 3.

For the descriptive statistics: In Alabama fields, we had an average of soybean plants with 72.6 cm height, 43 pods per plant, 2 lateral branches, and seed weight per plant of 12.1 grams.

In Alabama fields, we can see a variability in soybean height ranging from 1.9 to 16.2 cm, with an average of 6 cm between plants. Number of pods per plants can variate from 3 to 47.7, with an average of varibility of 16.76 pods per plant. Branches variates from 0 to 3 between plants, with an average of 1 branch variability. Finally, the seed weight between plants can range from 0.6 to 23.3 grams with an average of variability 4.7 grams.

For the correlations: The plant height is significantly correlated to the number of pods per plant and seed weight per plant. Taller plants results in more pods and more yielding. Number of pods per plant is significantly correlated to the number of lateral branches and seed weight per plant. More lateral branches results in more pods and yielding per plant. Also, number of lateral branches has a significant effect on the seed weight per plant.

Higher the variability in number of lateral branches and pods between plants, higher the variability in yielding per plant. Therefore, lateral branches and number of pods play a significant role in uniformity of the plant canopy and therefore they are the factors that could most interfere in grain yield.