Assign. 1 STA 445

Sophia Kubisiak

2024-02-22

```
library(tidyverse)
library(ggplot2)
```

Problem 1: Two Sample t-test

a. Load the `iris` dataset.

```
data('iris')
```

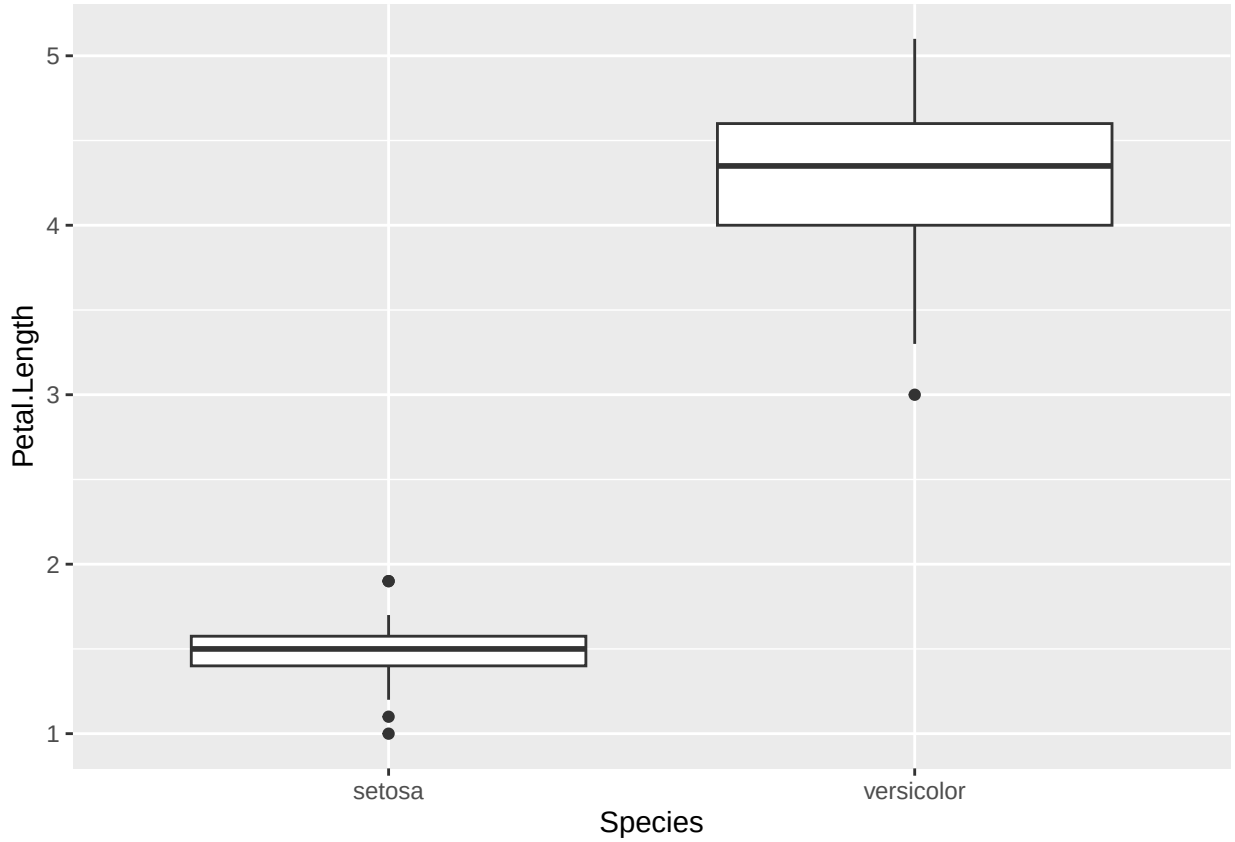
b. Create a subset of the data that just contains rows for the two species setosa and versicolor using filter. Use slice_sample to print out 20 random rows of the dataset.

```
iris.2 <- iris %>%
  filter(Species == 'setosa' | Species == 'versicolor')
slice_sample(iris.2 , n = 20)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1           5.3         3.7          1.5         0.2     setosa
## 2           6.6         3.0          4.4         1.4 versicolor
## 3           5.1         3.7          1.5         0.4     setosa
## 4           5.2         3.4          1.4         0.2     setosa
## 5           5.0         3.3          1.4         0.2     setosa
## 6           5.1         3.3          1.7         0.5     setosa
## 7           5.5         4.2          1.4         0.2     setosa
## 8           5.6         3.0          4.5         1.5 versicolor
## 9           5.8         2.7          3.9         1.2 versicolor
## 10          6.5         2.8          4.6         1.5 versicolor
## 11          6.1         2.8          4.0         1.3 versicolor
## 12          4.6         3.2          1.4         0.2     setosa
## 13          6.7         3.1          4.4         1.4 versicolor
## 14          7.0         3.2          4.7         1.4 versicolor
## 15          6.2         2.2          4.5         1.5 versicolor
## 16          5.0         3.2          1.2         0.2     setosa
## 17          6.4         3.2          4.5         1.5 versicolor
## 18          5.8         4.0          1.2         0.2     setosa
## 19          5.5         2.3          4.0         1.3 versicolor
## 20          5.0         3.0          1.6         0.2     setosa
```

c. Create a box plot of the petal lengths for these two species using ggplot. Does it look like the mean petal length varies by species?

```
ggplot(data = iris.2 , aes(x = Species , y = Petal.Length))+
  geom_boxplot()
```



d. Do a two sample t-test using t.test to determine formally if the petal lengths differ. Note: The book uses the tidy function in the broom package to make the output "nice". I hate it! Please don't use tidy.

```
t.test(data=iris.2, Petal.Length~Species , conf.level = 0.9)
```

```
##
##  Welch Two Sample t-test
##
## data:  Petal.Length by Species
## t = -39.493, df = 62.14, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group setosa and group versicolor is not eq
## 90 percent confidence interval:
##  -2.916299 -2.679701
## sample estimates:
##     mean in group setosa mean in group versicolor
##                    1.462                    4.260
```

d. What is the p-value for the test? What do you conclude? The p-value is 2.2 x 10^(-16). So in conclusion, the data is statistically significant and we should reject the null hypothesis.

e. Give a 95% confidence interval for the difference in the mean petal lengths.

```
t.test(data=iris.2, Petal.Length~Species , conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  Petal.Length by Species
## t = -39.493, df = 62.14, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group setosa and group versicolor is not eq
## 95 percent confidence interval:
##  -2.939618 -2.656382
## sample estimates:
##     mean in group setosa mean in group versicolor
##                    1.462                    4.260
```

The 95% confidence interval is [-2.939618 , -2.656382]

f. Give a 99% confidence interval for the difference in mean petal lengths.
(Hint: type ?t.test. See that you can change the confidence level using the option conf.level)

```
t.test(data=iris.2, Petal.Length~Species , conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  Petal.Length by Species
## t = -39.493, df = 62.14, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group setosa and group versicolor is not eq
## 99 percent confidence interval:
##  -2.986265 -2.609735
## sample estimates:
##     mean in group setosa mean in group versicolor
##                    1.462                    4.260
```

The 99% confidence interval is [-2.986265 , -2.609735]

g. What is the mean petal length for setosa? The mean is 1.462.

h. What is the mean petal length for versicolor? The mean is 4.4260.

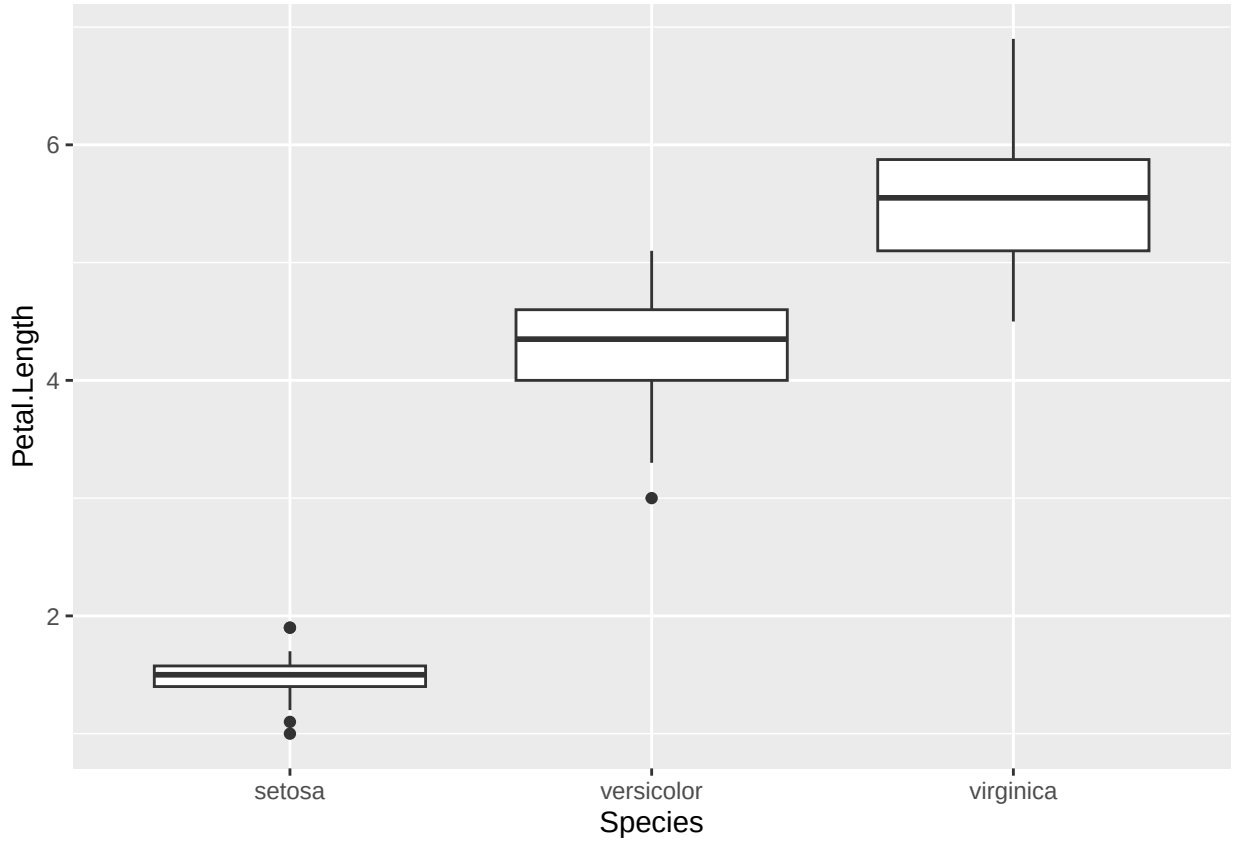
Problem 2: ANOVA

Use the iris data with all three species.

```
data(iris)
```

a. Create a box plot of the petal lengths for all three species using ggplot.Does it look like there are differences in the mean petal lengths?

```
ggplot(data=iris , aes(x=Species , y=Petal.Length))+
  geom_boxplot()
```



b. Create a linear model where sepal length is modeled by species. Give it an appropriate name.

```
irisflowers <- lm( Petal.Length ~ Sepal.Length * Species, data = iris )
```

c. Type anova(your model name) in a code chunk.

```
anova(irisflowers)
```

```
## Analysis of Variance Table
##
## Response: Petal.Length
##                      Df Sum Sq Mean Sq  F value    Pr(>F)
## Sepal.Length          1 352.87  352.87 5175.537 < 2.2e-16 ***
## Species               2  99.80   49.90  731.905 < 2.2e-16 ***
## Sepal.Length:Species  2   1.84    0.92   13.489 4.272e-06 ***
## Residuals           144   9.82    0.07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d. What is the p-value for the test? What do you conclude. The p-value is 2.2 x 10^(-16). This is statistically significant, so we reject the null hypothesis.

e. Type summary(your model name) in a code chunk.

```
summary(irisflowers)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Sepal.Length * Species, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68611 -0.13442 -0.00856  0.15966  0.79607
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      0.8031     0.5310   1.512    0.133
## Sepal.Length                     0.1316     0.1058   1.244    0.216
## Speciesversicolor               -0.6179     0.6837  -0.904    0.368
## Speciesvirginica                -0.1926     0.6578  -0.293    0.770
## Sepal.Length:Speciesversicolor   0.5548     0.1281   4.330 2.78e-05 ***
## Sepal.Length:Speciesvirginica    0.6184     0.1210   5.111 1.00e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2611 on 144 degrees of freedom
## Multiple R-squared:  0.9789, Adjusted R-squared:  0.9781
## F-statistic:  1333 on 5 and 144 DF,  p-value: < 2.2e-16
```

f. What is the mean petal length for the species setosa?

```
irisflowers.2 <- lm( Petal.Length ~ Species-1, data = iris )
summary(irisflowers.2)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Species - 1, data = iris)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.260 -0.258  0.038  0.240  1.348
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## Speciessetosa      1.46200    0.06086   24.02   <2e-16 ***
## Speciesversicolor  4.26000    0.06086   70.00   <2e-16 ***
## Speciesvirginica   5.55200    0.06086   91.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4303 on 147 degrees of freedom
## Multiple R-squared:  0.9895, Adjusted R-squared:  0.9892
## F-statistic:  4600 on 3 and 147 DF,  p-value: < 2.2e-16
```

The mean petal length for species setosa is 1.462.

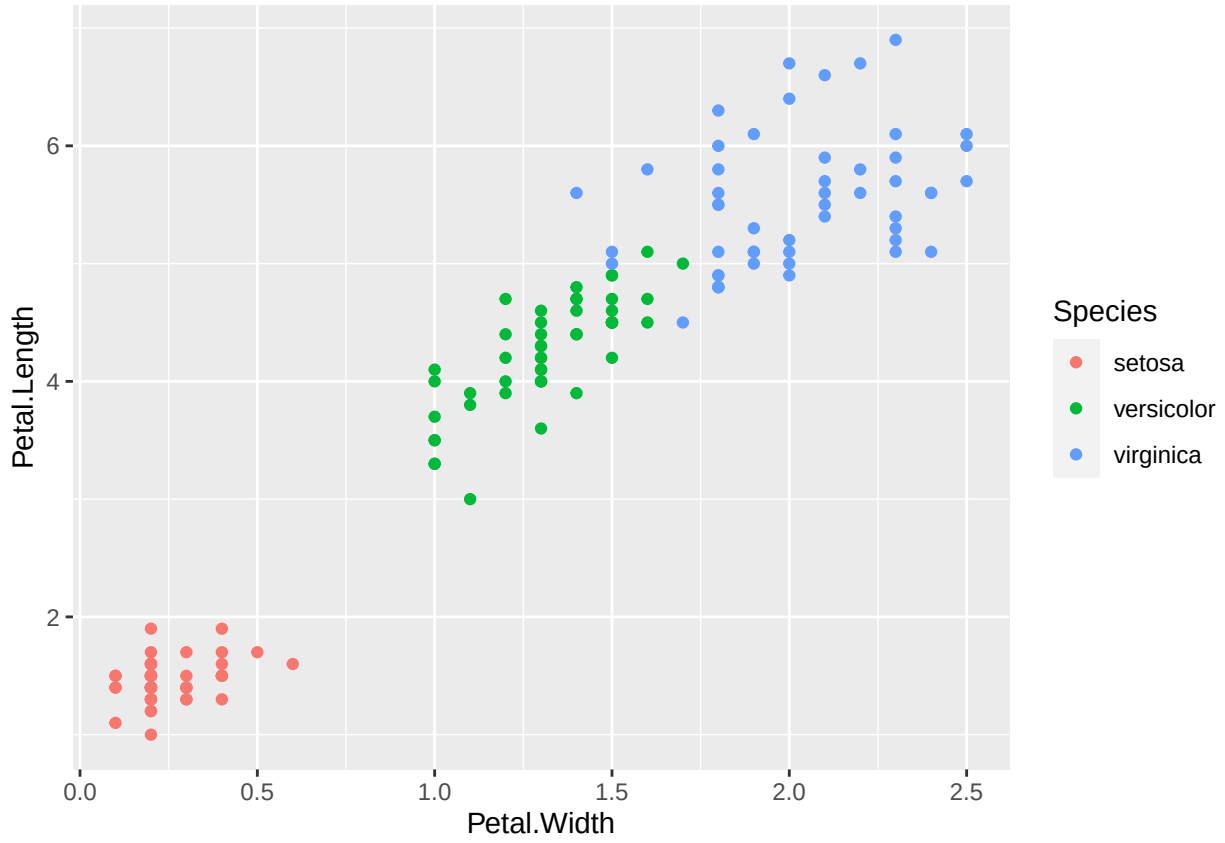
g. What is the mean petal length for the species versicolor? The mean length for species versicolor is 4.260.

Problem 3: Regression

Can we describe the relationship between petal length and petal width?

a. Create a scatterplot with petal length on the y-axis and petal width on the x-axis using ggplot.

```
ggplot(data = iris, aes(x=Petal.Width , y=Petal.Length , col = Species))+
  geom_point()
```



b. Create a linear model to model petal length with petal width (length is the response variable and width is the explanatory variable) using lm.

```
irispetals <- lm(Petal.Length~Petal.Width , data = iris)
```

c. What is the estimate of the slope parameter? The estimate slope parameter is 2.22994.
d. What is the estimate of the intercept parameter? The estimate intercept parameter is 1.08356.
e. Use summary() to get additional information.

```
summary(irispetals)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33542 -0.30347 -0.02955  0.25776  1.39453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.08356    0.07297   14.85   <2e-16 ***
## Petal.Width  2.22994    0.05140   43.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4782 on 148 degrees of freedom
## Multiple R-squared:  0.9271, Adjusted R-squared:  0.9266
## F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

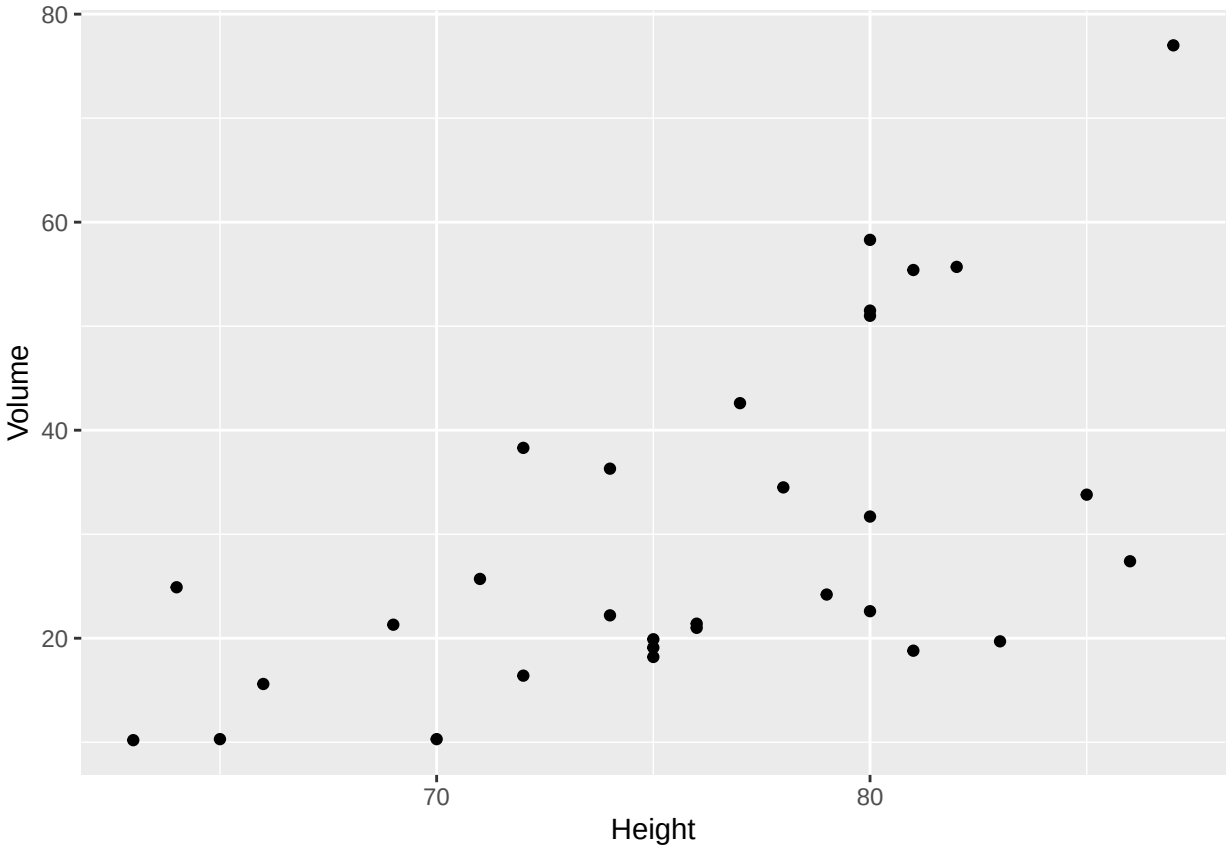
Problem 4: Modeling Trees

Using the `trees` data frame that comes pre-installed in R, follow the steps below to fit the regression model that uses the tree `Height` to explain the `Volume` of wood harvested from the tree.

```
data = 'trees'
```

a. Create a scatterplot of the data using ggplot.

```
trees%>%
  ggplot(aes(x=Height , y=Volume))+
  geom_point()
```



b. Fit a `lm` model using the command `model <- lm(Volume ~ Height, data=trees)`.

```
modeltrees <- lm(Volume ~ Height, data=trees)
```

c. Print out the table of coefficients with estimate names, estimated value, standard error, and upper and lower 95% confidence intervals.

```
summary(modeltrees)
```

```
##
## Call:
## lm(formula = Volume ~ Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.274  -9.894  -2.894  12.068  29.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.1236    29.2731  -2.976 0.005835 **
## Height        1.5433     0.3839   4.021 0.000378 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 29 degrees of freedom
```

```
## Multiple R-squared:  0.3579, Adjusted R-squared:  0.3358
## F-statistic: 16.16 on 1 and 29 DF,  p-value: 0.0003784
```

```
confint(modeltrees , conf.level = .95)
```

```
##                   2.5 %     97.5 %
## (Intercept) -146.993871 -27.253357
## Height         0.758249   2.328451
```

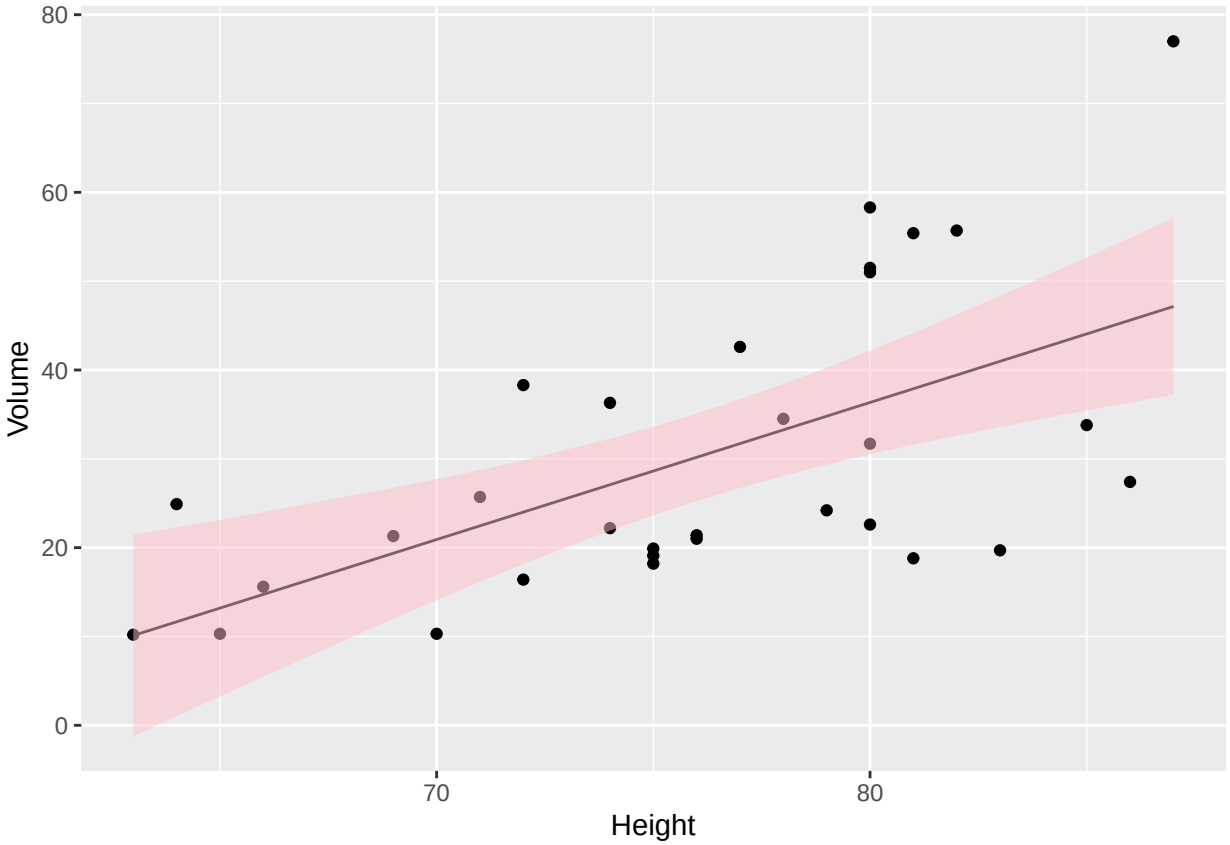
d.Add the model fitted values to the `trees` data frame along with the regression model confidence intervals. Note: the book does this in a super convoluted way. Don't follow the model in the book. Instead try cbind.

```
trees.2 <- cbind(predict(modeltrees , interval="confidence") , trees)
summary(trees.2)
```

```
##       fit             lwr              upr            Girth           Height
##  Min.   :10.11   Min.   :-1.223   Min.   :21.44   Min.   : 8.30   Min.   :63
##  1st Qu.:24.00   1st Qu.:18.160   1st Qu.:29.84   1st Qu.:11.05   1st Qu.:72
##  Median :30.17   Median :25.250   Median :35.09   Median :12.90   Median :76
##  Mean   :30.17   Mean   :23.466   Mean   :36.88   Mean   :13.25   Mean   :76
##  3rd Qu.:36.34   3rd Qu.:30.507   3rd Qu.:42.18   3rd Qu.:15.25   3rd Qu.:80
##  Max.   :47.15   Max.   :37.208   Max.   :57.09   Max.   :20.60   Max.   :87
##      Volume
##  Min.   :10.20
##  1st Qu.:19.40
##  Median :24.20
##  Mean   :30.17
##  3rd Qu.:37.30
##  Max.   :77.00
```

e. Graph the data and fitted regression line and uncertainty ribbon.

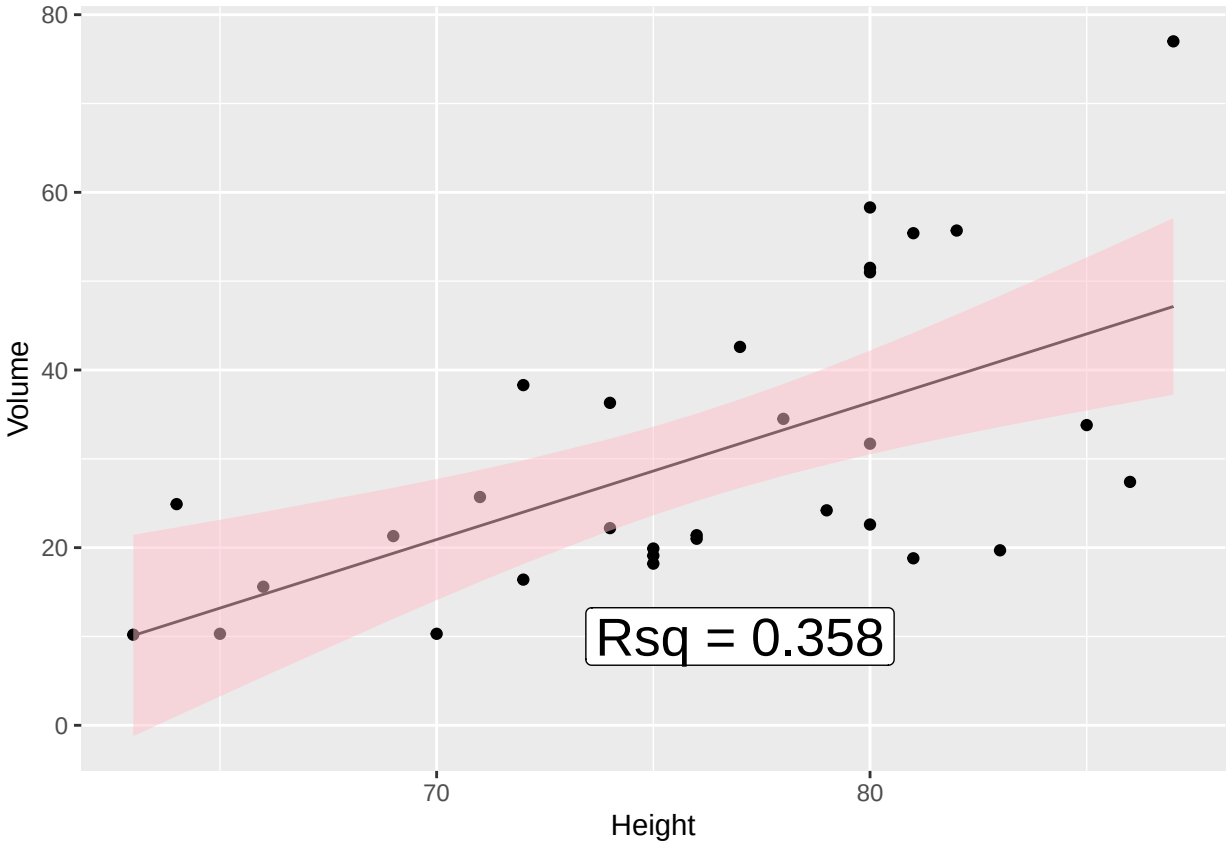
```
ggplot(data = trees.2 , aes(x=Height, y=Volume))+
  geom_point()+
  geom_line(aes(y=fit))+
  geom_ribbon(aes(ymin=lwr, ymax=upr), alpha=0.5, fill = "pink")
```



f. Add the R-squared value as an annotation to the graph using annotate.

```
Rsq_string <-
  broom::glance(modeltrees) %>%
  select(r.squared) %>%
  mutate(r.squared = round(r.squared, digits=3)) %>%
  mutate(r.squared = paste('Rsq =', r.squared)) %>%
  pull(r.squared)

ggplot(data = trees.2 , aes(x=Height, y=Volume))+
  geom_point()+
  geom_line(aes(y=fit))+
  geom_ribbon(aes(ymin=lwr, ymax=upr), alpha=0.5, fill = "pink")+
  annotate('label', label=Rsq_string, x=77, y=10, size=7)
```


80
60
40
20
0
Volume
70
80
Height
Rsq = 0.358