Assign. 1 STA 445

Myleen Maldonado

2024-02-27

```
library(tidyverse)
```

Directions:

This assignment covers chapter 5. Please show all work in this document and knit your final draft into a pdf. This is assignment is about statistical models, which will be helpful if you plan on taking STA 570, STA 371, or STA 571.

Problem 1: Two Sample t-test

##

a. Load the `iris` dataset.

```
slice_sample(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
## 1          7.9         3.8          6.4           2 virginica
```

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.

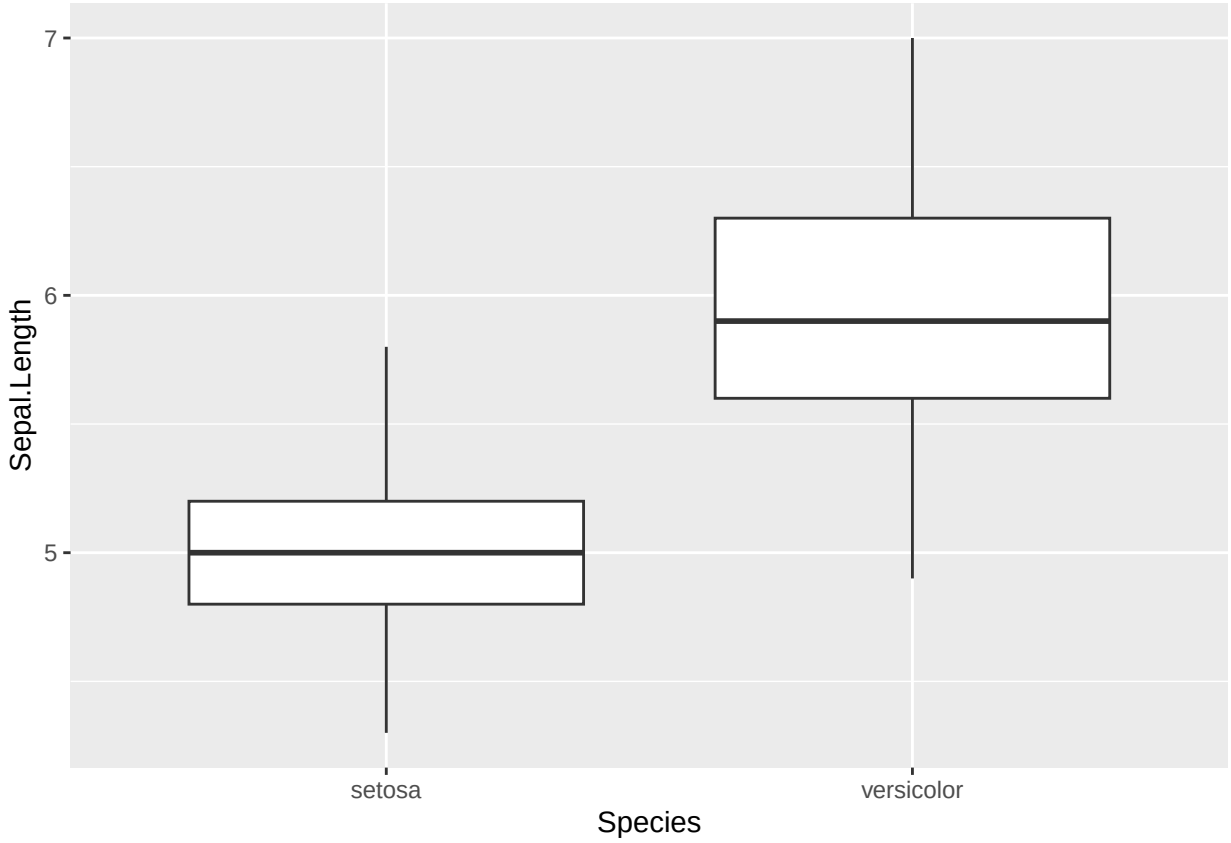
```
subset.iris <- iris %>%
  filter(Species=="setosa"| Species=="versicolor")
 slice_sample(subset.iris, n = 20)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1           6.6         3.0          4.4         1.4 versicolor
## 2           5.5         2.3          4.0         1.3 versicolor
## 3           5.0         3.3          1.4         0.2     setosa
## 4           5.6         2.9          3.6         1.3 versicolor
## 5           5.4         3.9          1.7         0.4     setosa
## 6           4.3         3.0          1.1         0.1     setosa
## 7           5.9         3.2          4.8         1.8 versicolor
## 8           6.6         2.9          4.6         1.3 versicolor
## 9           6.7         3.1          4.4         1.4 versicolor
## 10          4.4         3.2          1.3         0.2     setosa
## 11          5.8         2.7          4.1         1.0 versicolor
## 12          6.5         2.8          4.6         1.5 versicolor
## 13          4.9         3.1          1.5         0.1     setosa
## 14          5.1         3.5          1.4         0.3     setosa
## 15          5.0         3.5          1.3         0.3     setosa
## 16          5.2         2.7          3.9         1.4 versicolor
## 17          4.4         2.9          1.4         0.2     setosa
## 18          5.0         3.5          1.6         0.6     setosa
```

```
## 19          6.0         2.7          5.1         1.6 versicolor
## 20          6.3         2.5          4.9         1.5 versicolor
```

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(subset.iris, aes(x=Species, y=Sepal.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=subset.iris, 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?

```
# We conclude that because p-value < 2.2e-16 , then we can conclude that they are similar.
```

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

```
t.test(data=subset.iris, 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
```

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=subset.iris, 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
```

g. What is the mean petal length for setosa?

```
# 1.462
```

h. What is the mean petal length for versicolor?

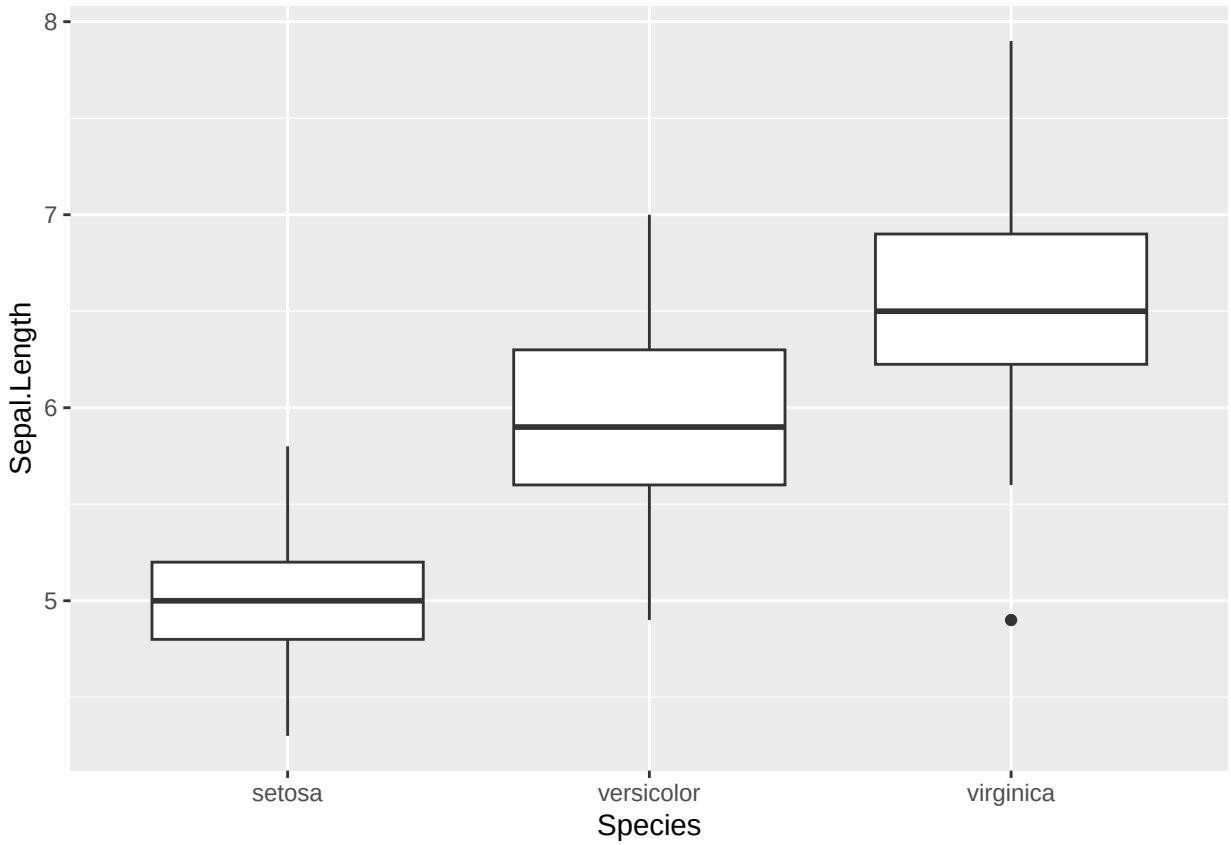
```
# 4.260
```

Problem 2: ANOVA

Use the iris data with all three species.

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(iris, aes(x=Species, y=Sepal.Length)) + geom_boxplot()
```




```
#yes it looks like there are differences between petal lengths
```

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

```
sepal.lgth <- lm(data=iris, Sepal.Length~Species-1)

sepal.lgth
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Species - 1, data = iris)
##
## Coefficients:
##     Speciessetosa  Speciesversicolor   Speciesvirginica
##             5.006              5.936              6.588
```

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

```
anova(sepal.lgth)
```

```
## Analysis of Variance Table
##
## Response: Sepal.Length
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Species     3 5184.9 1728.30  6521.7 < 2.2e-16 ***
## Residuals 147   39.0    0.27
## ---
## 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.

```
# it ends up being smaller than 2.2e-16 so that means that they are similiar.
```

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

```
summary(sepal.lgth)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Species - 1, data = iris)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6880 -0.3285 -0.0060  0.3120  1.3120
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## Speciessetosa       5.0060     0.0728   68.76   <2e-16 ***
## Speciesversicolor   5.9360     0.0728   81.54   <2e-16 ***
## Speciesvirginica    6.5880     0.0728   90.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5148 on 147 degrees of freedom
## Multiple R-squared:  0.9925, Adjusted R-squared:  0.9924
## F-statistic:  6522 on 3 and 147 DF,  p-value: < 2.2e-16
```

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

```
#  the mean petal length for species setosa is:
#  1.6880
```

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

```
#  the mean petal length for species versicolor is:
#  1.582

sepal.lgth <- lm(data=iris, Sepal.Length~Species)

sepal.lgth
```

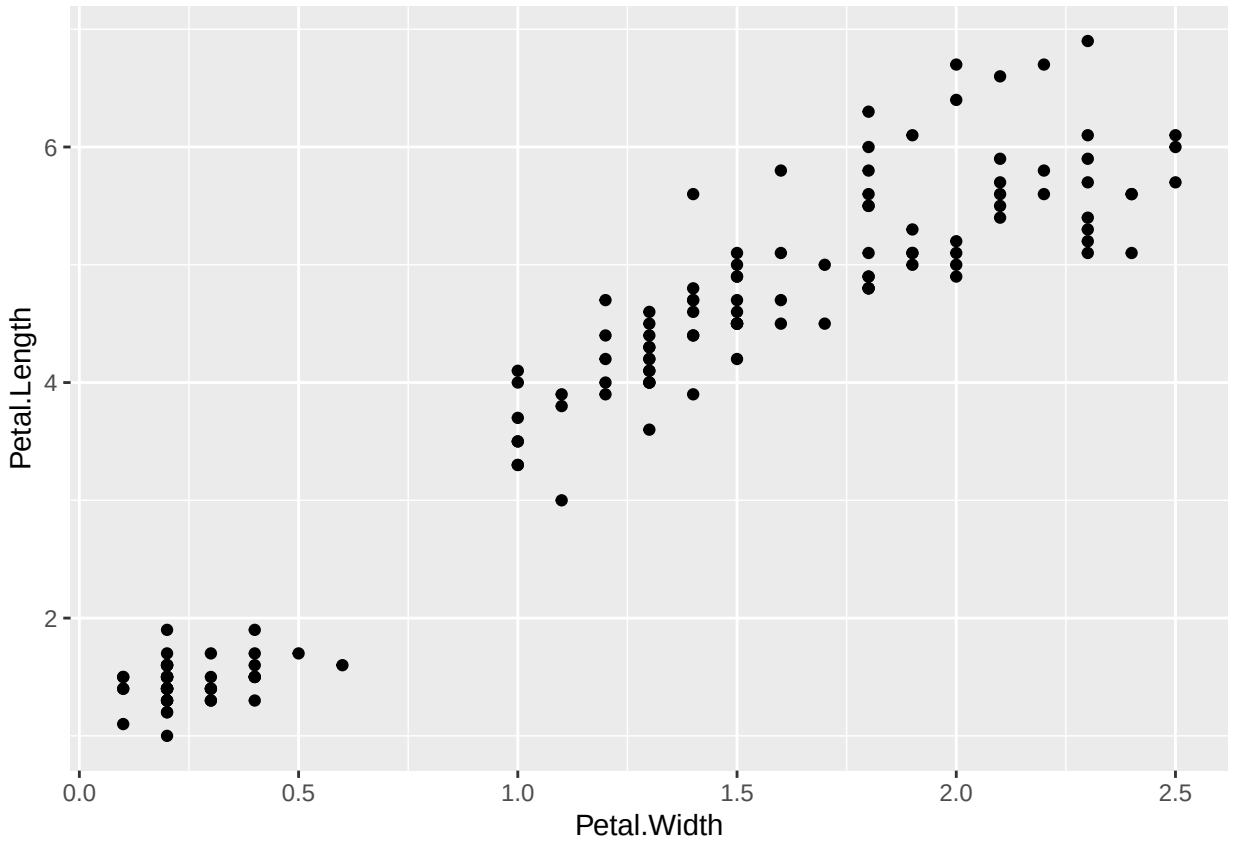
```
##
## Call:
## lm(formula = Sepal.Length ~ Species, data = iris)
##
## Coefficients:
##       (Intercept)  Speciesversicolor   Speciesvirginica
##             5.006              0.930              1.582
```

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(iris, aes(x=Petal.Width, y=Petal.Length)) +
  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.

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

petal.lm
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width * Species, data = iris)
##
## Coefficients:
##                   (Intercept)                    Petal.Width
##                        1.3276                         0.5465
##             Speciesversicolor                 Speciesvirginica
##                        0.4537                         2.9131
## Petal.Width:Speciesversicolor   Petal.Width:Speciesvirginica
##                        1.3228                         0.1008
```

c. What is the estimate of the slope parameter?

```
petal.lm
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width * Species, data = iris)
##
## Coefficients:
##                   (Intercept)                    Petal.Width
##                        1.3276                         0.5465
```

```
##             Speciesversicolor                Speciesvirginica
##                        0.4537                          2.9131
## Petal.Width:Speciesversicolor   Petal.Width:Speciesvirginica
##                        1.3228                          0.1008
```

```
# 1.3276
```

d. What is the estimate of the intercept parameter?

```
summary(petal.lm)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width * Species, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84099 -0.19343 -0.03686  0.16314  1.17065
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.3276     0.1309  10.139  < 2e-16 ***
## Petal.Width                     0.5465     0.4900   1.115   0.2666
## Speciesversicolor               0.4537     0.3737   1.214   0.2267
## Speciesvirginica                2.9131     0.4060   7.175 3.53e-11 ***
## Petal.Width:Speciesversicolor   1.3228     0.5552   2.382   0.0185 *
## Petal.Width:Speciesvirginica    0.1008     0.5248   0.192   0.8480
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3615 on 144 degrees of freedom
## Multiple R-squared:  0.9595, Adjusted R-squared:  0.9581
## F-statistic: 681.9 on 5 and 144 DF,  p-value: < 2.2e-16
```

```
# 0.1309
```

e. Use summary() to get additional information.

```
summary(petal.lm)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width * Species, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84099 -0.19343 -0.03686  0.16314  1.17065
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.3276     0.1309  10.139  < 2e-16 ***
## Petal.Width                     0.5465     0.4900   1.115   0.2666
## Speciesversicolor               0.4537     0.3737   1.214   0.2267
## Speciesvirginica                2.9131     0.4060   7.175 3.53e-11 ***
## Petal.Width:Speciesversicolor   1.3228     0.5552   2.382   0.0185 *
## Petal.Width:Speciesvirginica    0.1008     0.5248   0.192   0.8480
## ---
```

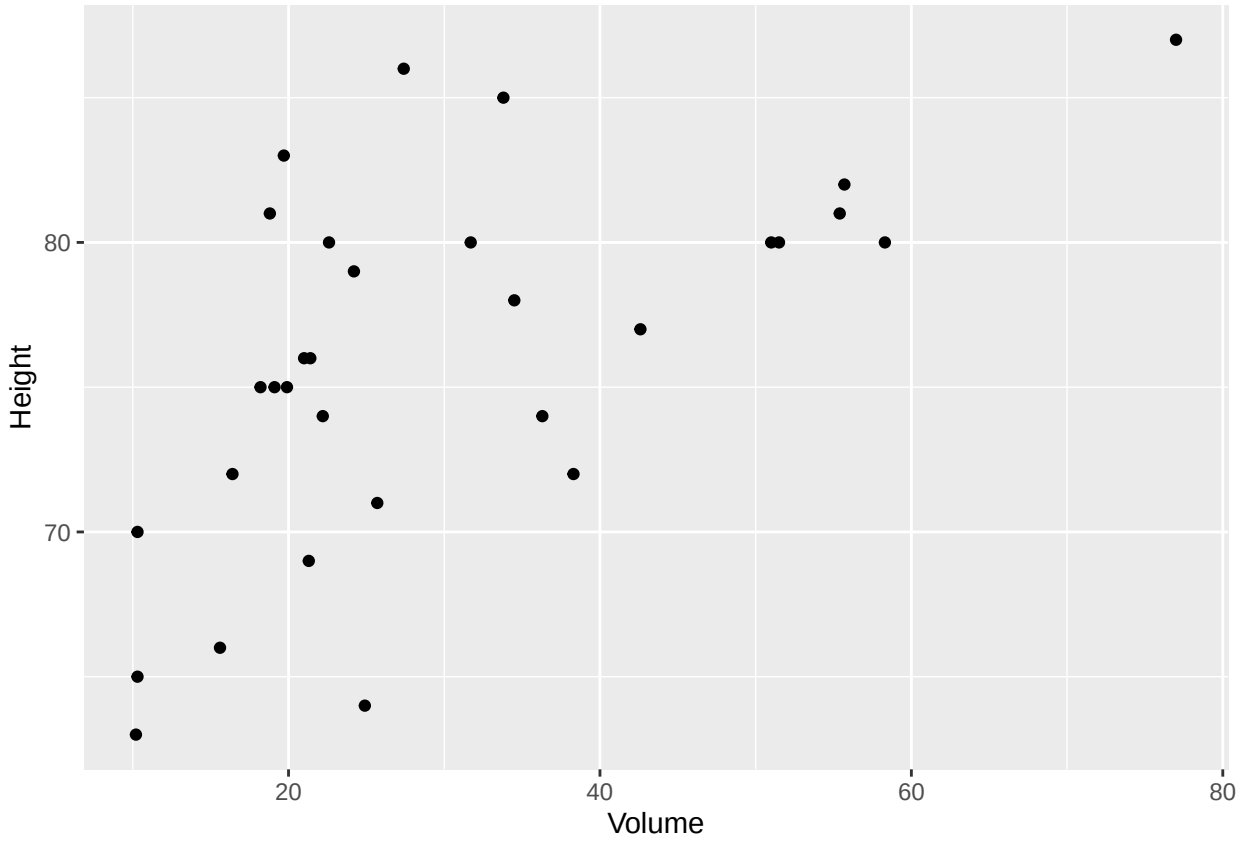
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3615 on 144 degrees of freedom
## Multiple R-squared:  0.9595, Adjusted R-squared:  0.9581
## F-statistic: 681.9 on 5 and 144 DF,  p-value: < 2.2e-16
```

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.

a. Create a scatterplot of the data using ggplot.

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



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

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

```
##
## Call:
## lm(formula = Volume ~ Height, data = trees)
##
## Coefficients:
## (Intercept)       Height
##     -87.124        1.543
```

c. Print out the table of coefficients with estimate names, estimated value, standard error, and upper and

lower 95% confidence intervals.

```
summary(model)
```

```
##
## 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
```

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.

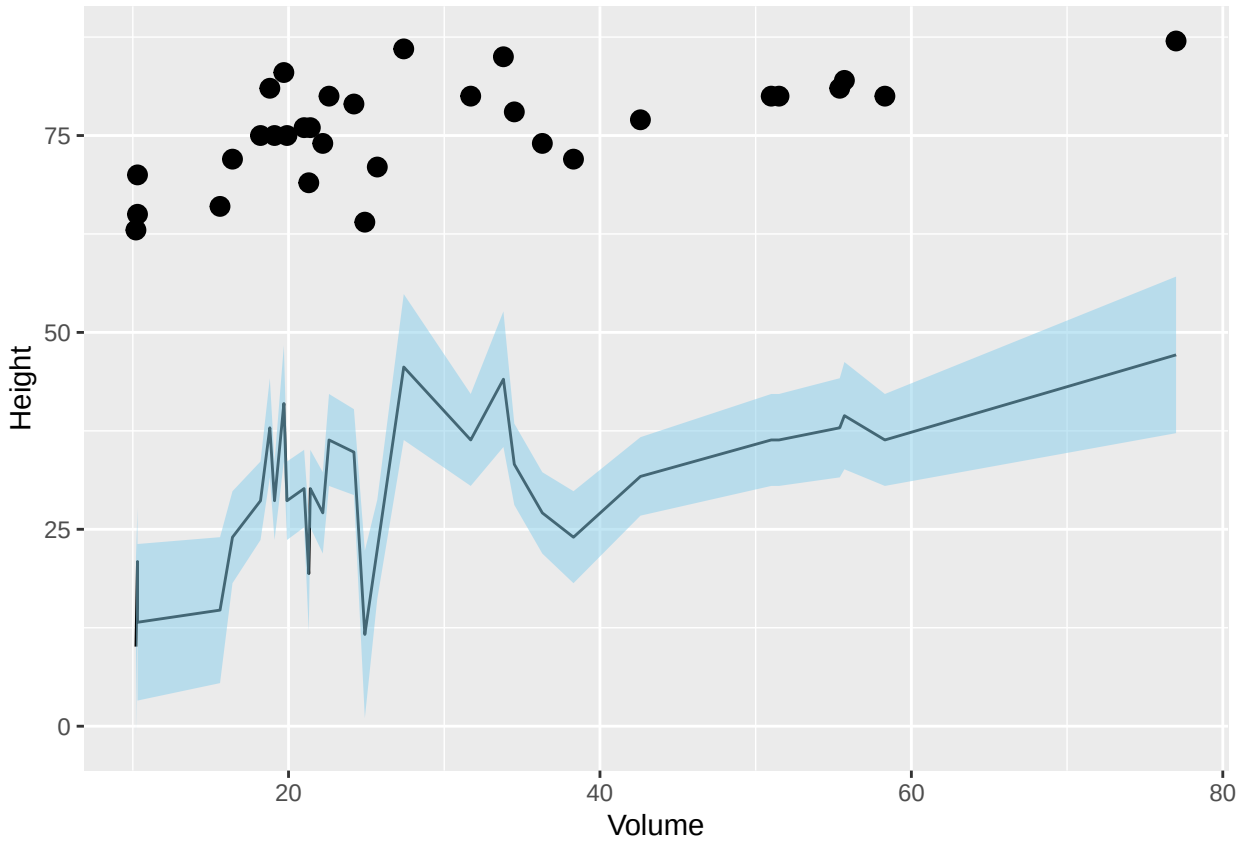
```
trees.pred <- cbind(trees, predict(model, interval="confidence"))

head(trees.pred)
```

```
##   Girth Height Volume      fit       lwr      upr
## 1   8.3     70   10.3 20.91087 14.098550 27.72319
## 2   8.6     65   10.3 13.19412  3.254288 23.13395
## 3   8.8     63   10.2 10.10742 -1.223363 21.43821
## 4  10.5     72   16.4 23.99757 18.159758 29.83538
## 5  10.7     81   18.8 37.88772 31.592680 44.18275
## 6  10.8     83   19.7 40.97442 33.597379 48.35145
```

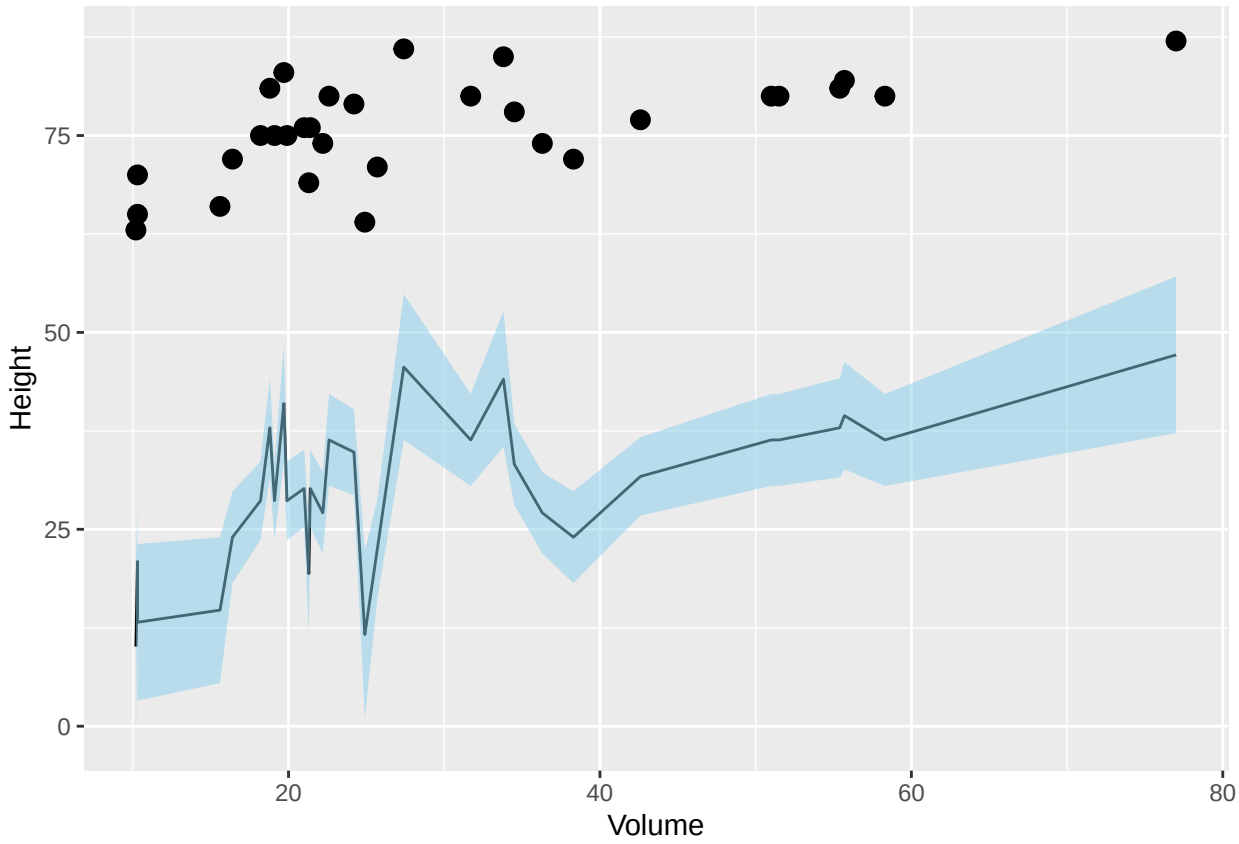
e. Graph the data and fitted regression line and uncertainty ribbon.

```
ggplot(data = trees.pred, aes(x=Volume, y=Height)) +
  geom_point(size=3) +
  geom_line(aes(y=fit)) +
  geom_ribbon(aes(ymin=lwr, ymax=upr), alpha=0.5, fill = "skyblue")
```



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

```
ggplot(data = trees.pred, aes(x=Volume, y=Height)) +
  geom_point(size=3) +
  geom_line(aes(y=fit)) +
  geom_ribbon(aes(ymin=lwr, ymax=upr), alpha=0.5, fill = "skyblue")
```




```
# annotate("r-squared value is:" )
```