Assign. 1 STA 445

Paige Hawkinson

2024-02-20

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.

```
data("iris")
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

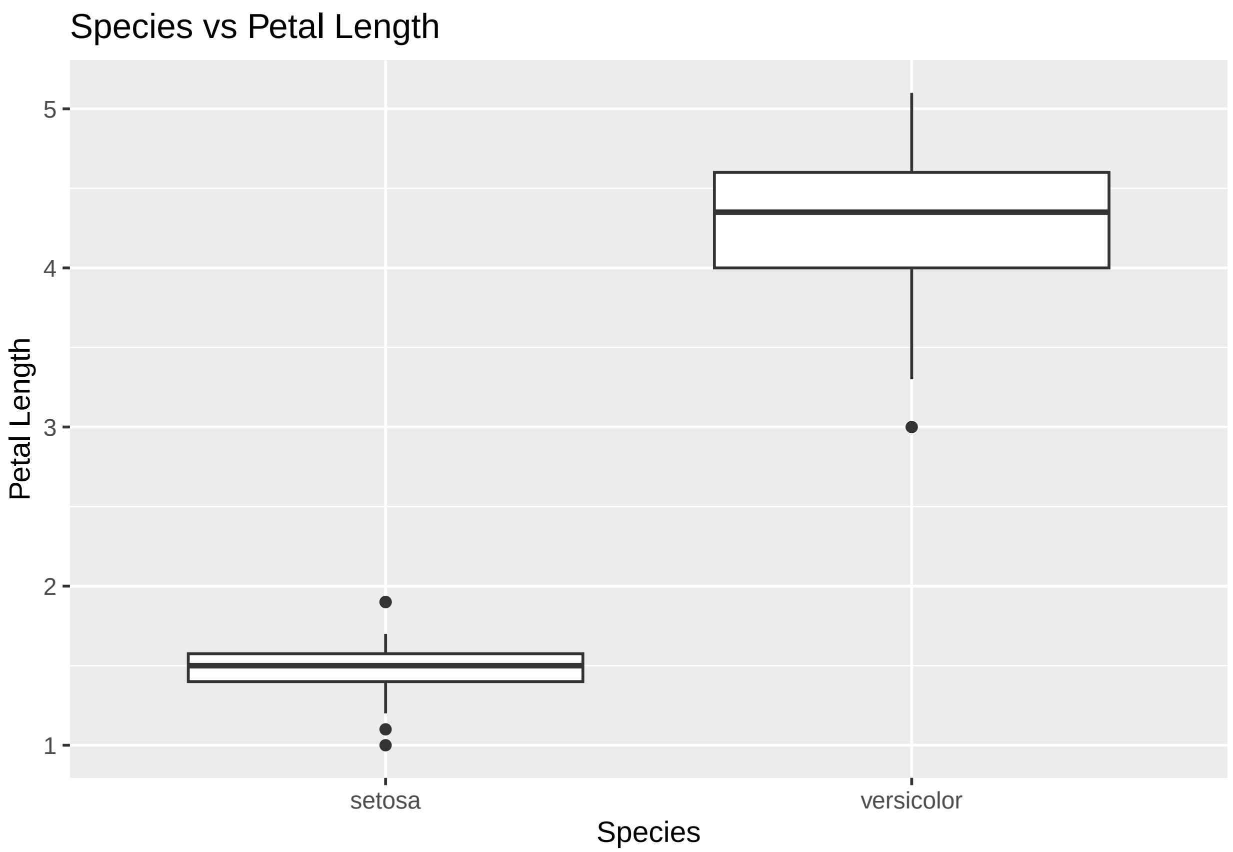
b. Create a subset of the data that just contains rows for the two species setosa and versicolor using filter. Use slice_sample to print out 20 random rows of the dataset.

```
library(tidyverse)
iris.subset <- iris %>%
  filter(Species == "setosa" | Species == "versicolor")
slice_sample(iris.subset)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1          6.5         2.8          4.6         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?

```
library(ggplot2)
ggplot(data=iris.subset, aes(x=Species, y=Petal.Length)) +
  geom_boxplot() +
  labs( title='Species vs Petal Length') +
  labs( x="Species", y="Petal Length")
```



Based on the box plot, it looks like there is a difference in mean petal length

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.subset, Petal.Length ~ Species)
```

```
##
##  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 equ
## 95 percent confidence interval:
##  -2.939618 -2.656382
## 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 for the t-test is 2.2e-16, or essentially zero. The p-value means we can reject the null hypothesis and that there is enough evidence to show that there is a difference in mean petal lengths between setosa and versicolor.

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

We are 95% confident that the difference in mean petal lengths between setosa and versicolor is between -2.939618 cm and -2.656382 cm.

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.subset, 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 equ
## 99 percent confidence interval:
##  -2.986265 -2.609735
## sample estimates:
##     mean in group setosa mean in group versicolor
##                    1.462                    4.260
```

We are 99% confident that the difference of mean petal lengths between setosa and versicolor is between -2.986265 cm and -2.609735 cm.

g. What is the mean petal length for setosa?

The mean petal length for setosa is 1.462 cm.

h. What is the mean petal length for versicolor?

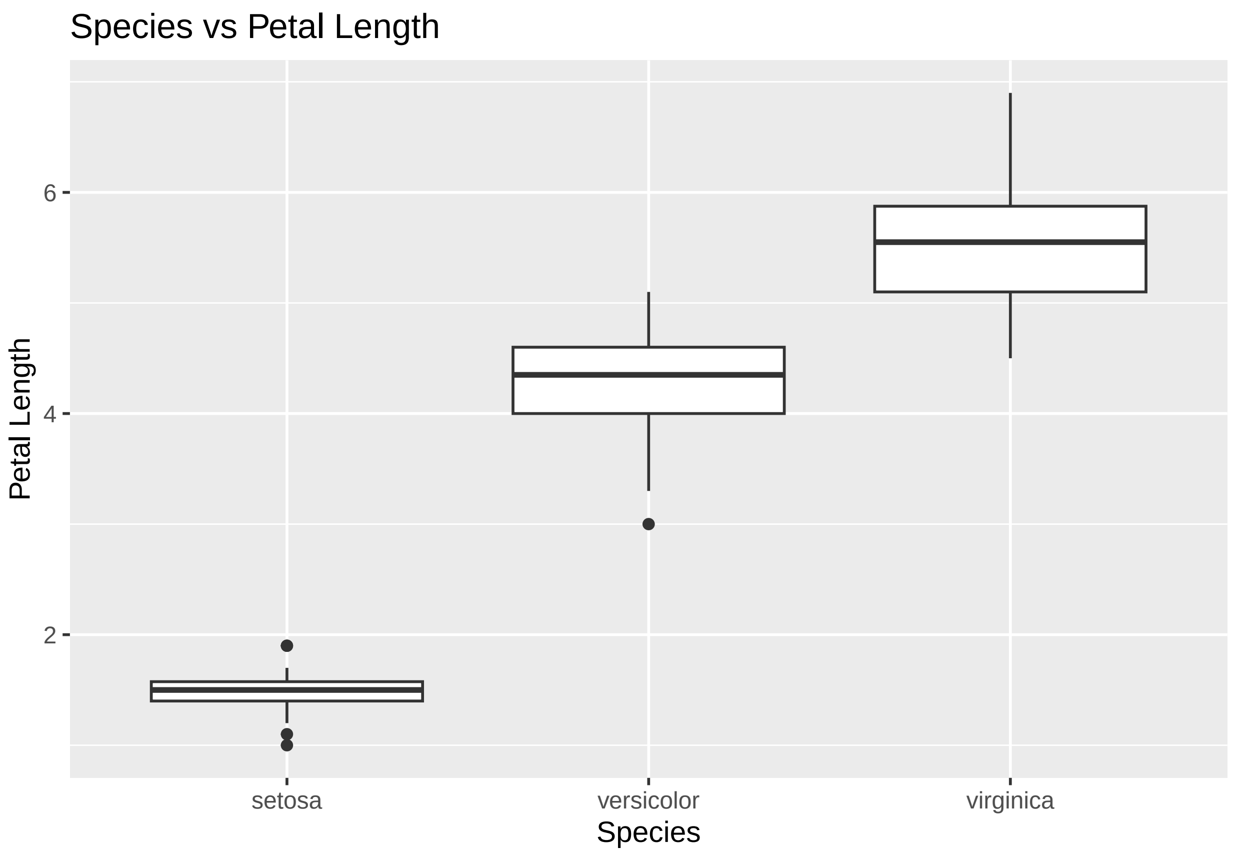
The mean petal length for versicolor is 4.260 cm.

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?

```
library(ggplot2)
ggplot(data=iris, aes(x=Species, y=Petal.Length)) +
  geom_boxplot() +
  labs( title='Species vs Petal Length') +
  labs( x="Species", y="Petal Length")
```



Based on the box plot form above, there seems to be a difference in mean petal lengths among the three different species.

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

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

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

```
anova(iris.mod.1)
```

```
## 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.

The p-value is 2.2e-16 or essentially zero. This p-value means that we can reject the null hypothesis and that there is enough evidence to show that there is a difference in mean sepal length between species.

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

```
summary(iris.mod.1)
```

```
##
## 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 sepal length for the species setosa?

The mean sepal length for setosa is 5.0060 cm.

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

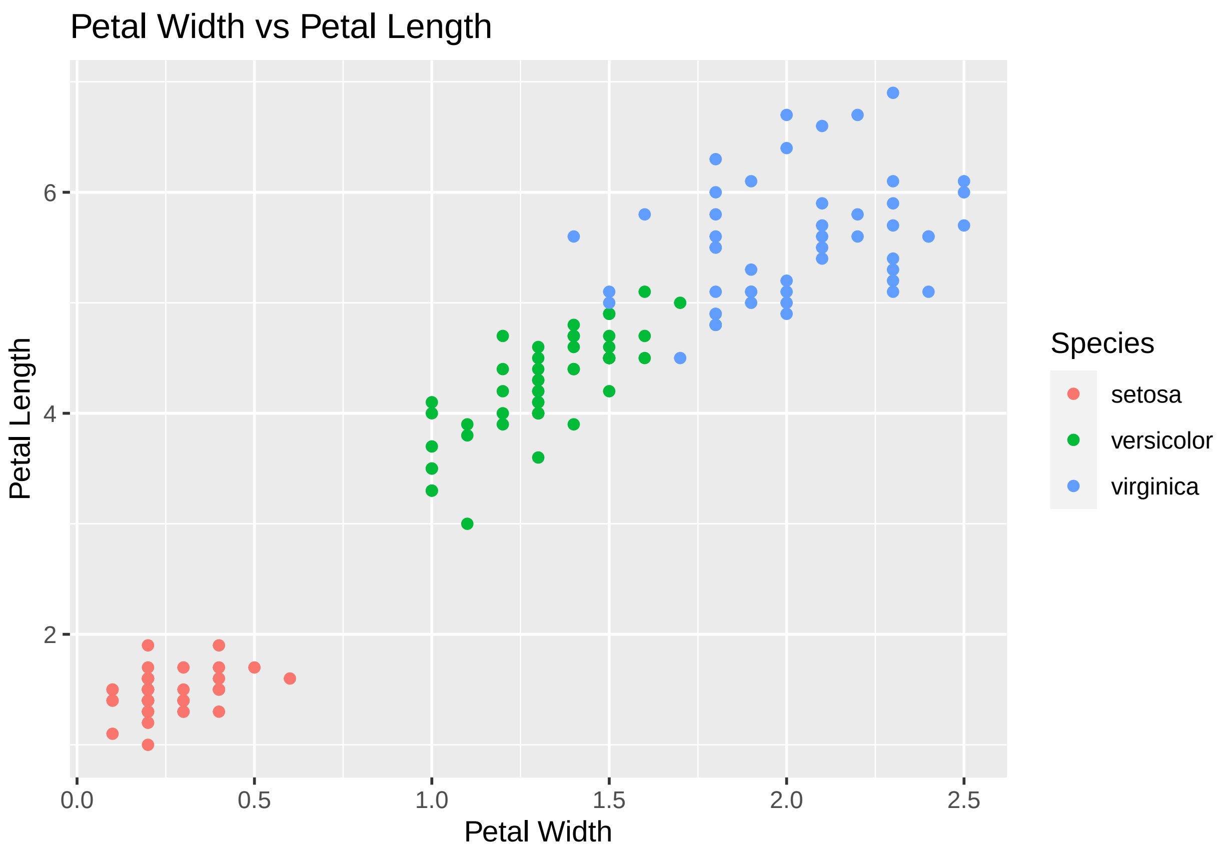
The mean sepal length for versicolor is 5.9360 cm.

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, color=Species)) +
  geom_point() +
  labs( title='Petal Width vs Petal Length') +
  labs( x="Petal Width", y="Petal Length")
```



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.

```
iris.mod.2 <- lm(data=iris, Petal.Length ~ Petal.Width)
iris.mod.2
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width, data = iris)
##
## Coefficients:
## (Intercept)  Petal.Width
##       1.084        2.230
```

c. What is the estimate of the slope parameter?

Estimate of slope parameter = 2.230

d. What is the estimate of the intercept parameter?

Estimate of intercept parameter = 1.084

e. Use summary() to get additional information.

```
summary(iris.mod.2)
```

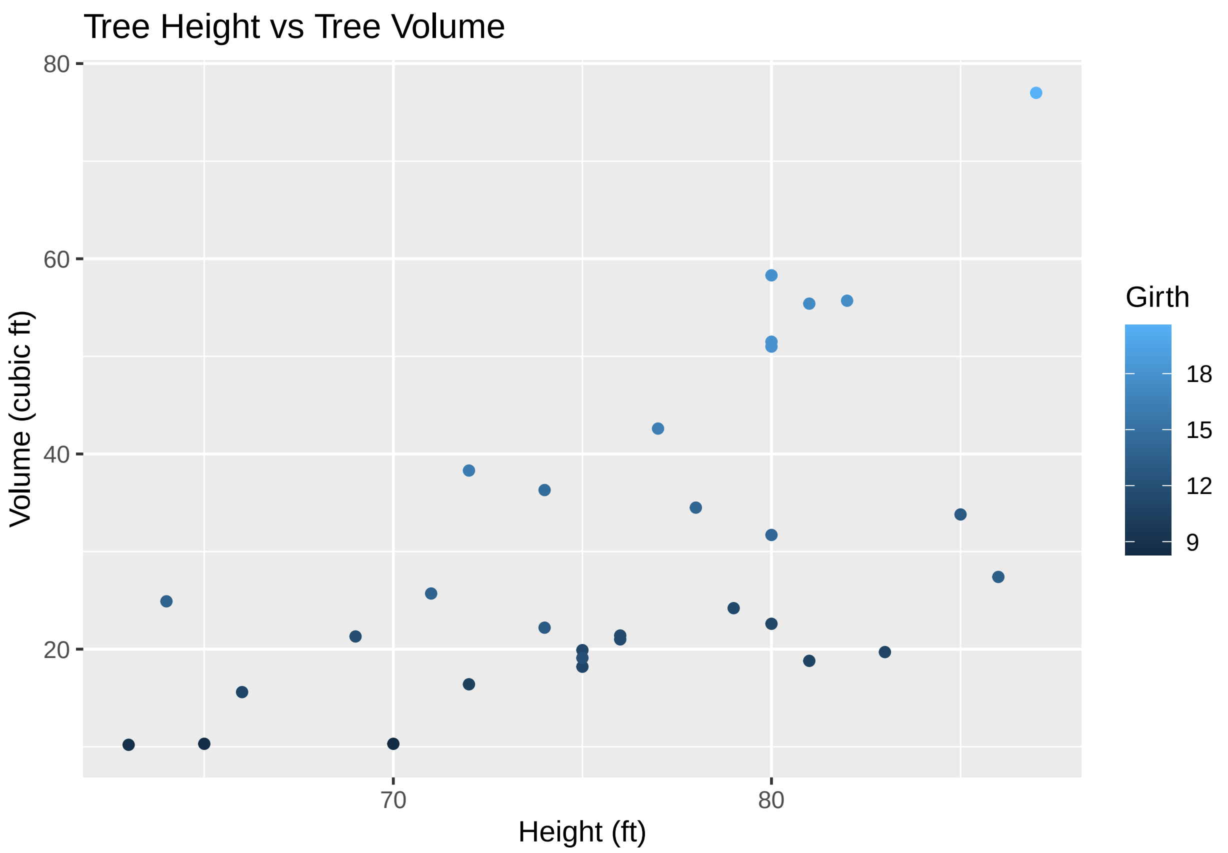
```
## 
## Call:
## lm(formula = Petal.Length ~ Petal.Width, data = iris)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.33542 -0.30347 -0.02955  0.25776  1.39453 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.08356    0.07297   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.

a. Create a scatterplot of the data using ggplot.

```
ggplot(data=trees, aes(x=Height, y=Volume, color=Girth)) +
  geom_point() +
  labs( title='Tree Height vs Tree Volume') +
  labs( x="Height (ft)", y="Volume (cubic ft)")
```



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

```
model <- 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(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
```

```
confint(model)
```

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

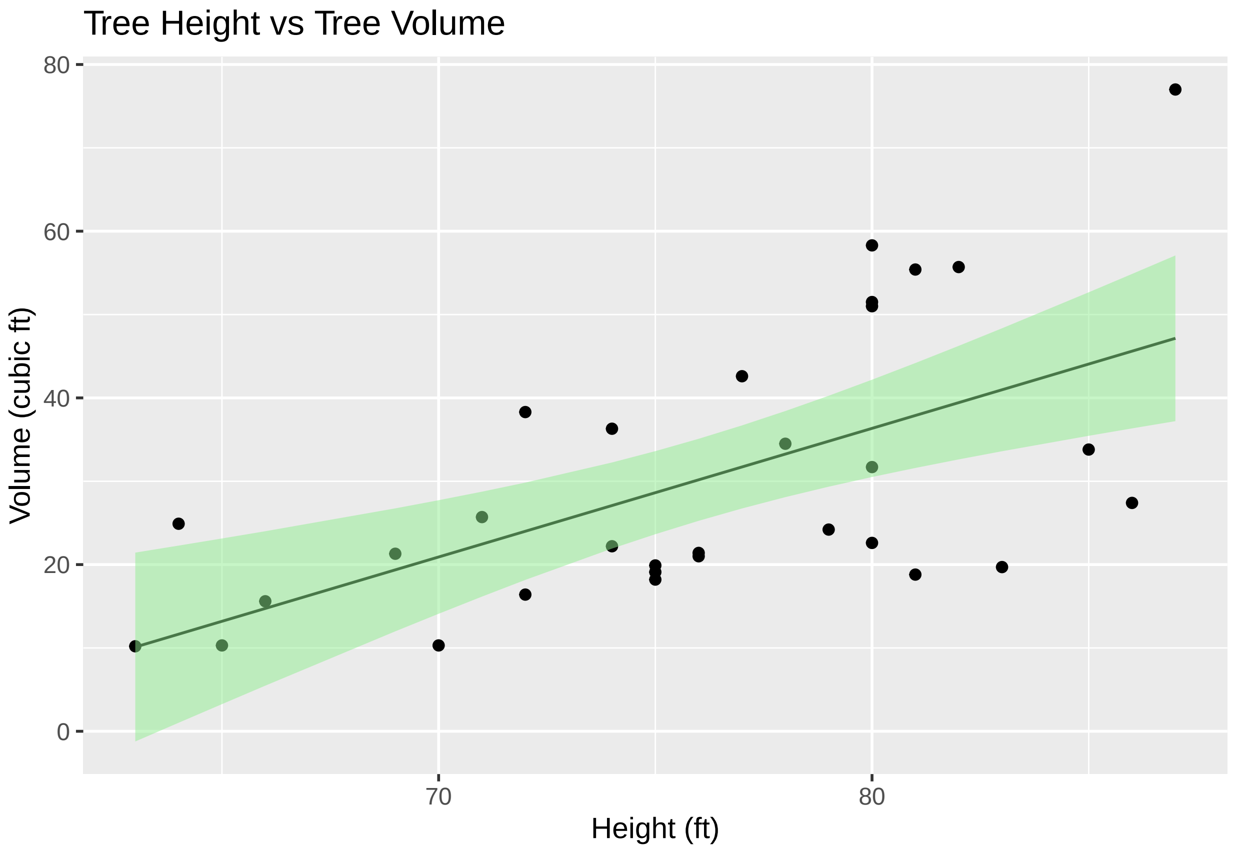
d.Add the model fitted values to the **trees** data frame along with the regression model confidence intervals. Note: the book does this in a super convoluted way. Don't follow the model in the book. Instead try cbind.

```
trees.w.pred <- cbind(trees, predict(model, interval="confidence"))
trees.w.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
## 7   11.0     66   15.6 14.73747  5.471607 24.00333
## 8   11.0     75   18.2 28.62762 23.644217 33.61102
## 9   11.1     80   22.6 36.34437 30.506556 42.18218
## 10  11.2     75   19.9 28.62762 23.644217 33.61102
## 11  11.3     79   24.2 34.80102 29.345254 40.25678
## 12  11.4     76   21.0 30.17097 25.249799 35.09214
## 13  11.4     76   21.4 30.17097 25.249799 35.09214
## 14  11.7     69   21.3 19.36752 11.990482 26.74456
## 15  12.0     75   19.1 28.62762 23.644217 33.61102
## 16  12.9     74   22.2 27.08427 21.918668 32.24987
## 17  12.9     85   33.8 44.06112 35.450370 52.67186
## 18  13.3     86   27.4 45.60447 36.338602 54.87033
## 19  13.7     71   25.7 22.45422 16.159183 28.74926
## 20  13.8     64   24.9 11.65077  1.021703 22.27984
## 21  14.0     78   34.5 33.25767 28.092067 38.42327
## 22  14.2     80   31.7 36.34437 30.506556 42.18218
## 23  14.5     74   36.3 27.08427 21.918668 32.24987
## 24  16.0     72   38.3 23.99757 18.159758 29.83538
## 25  16.3     77   42.6 31.71432 26.730917 36.69772
## 26  17.3     81   55.4 37.88772 31.592680 44.18275
## 27  17.5     82   55.7 39.43107 32.618747 46.24339
## 28  17.9     80   58.3 36.34437 30.506556 42.18218
## 29  18.0     80   51.5 36.34437 30.506556 42.18218
## 30  18.0     80   51.0 36.34437 30.506556 42.18218
## 31  20.6     87   77.0 47.14782 37.207982 57.08765
```

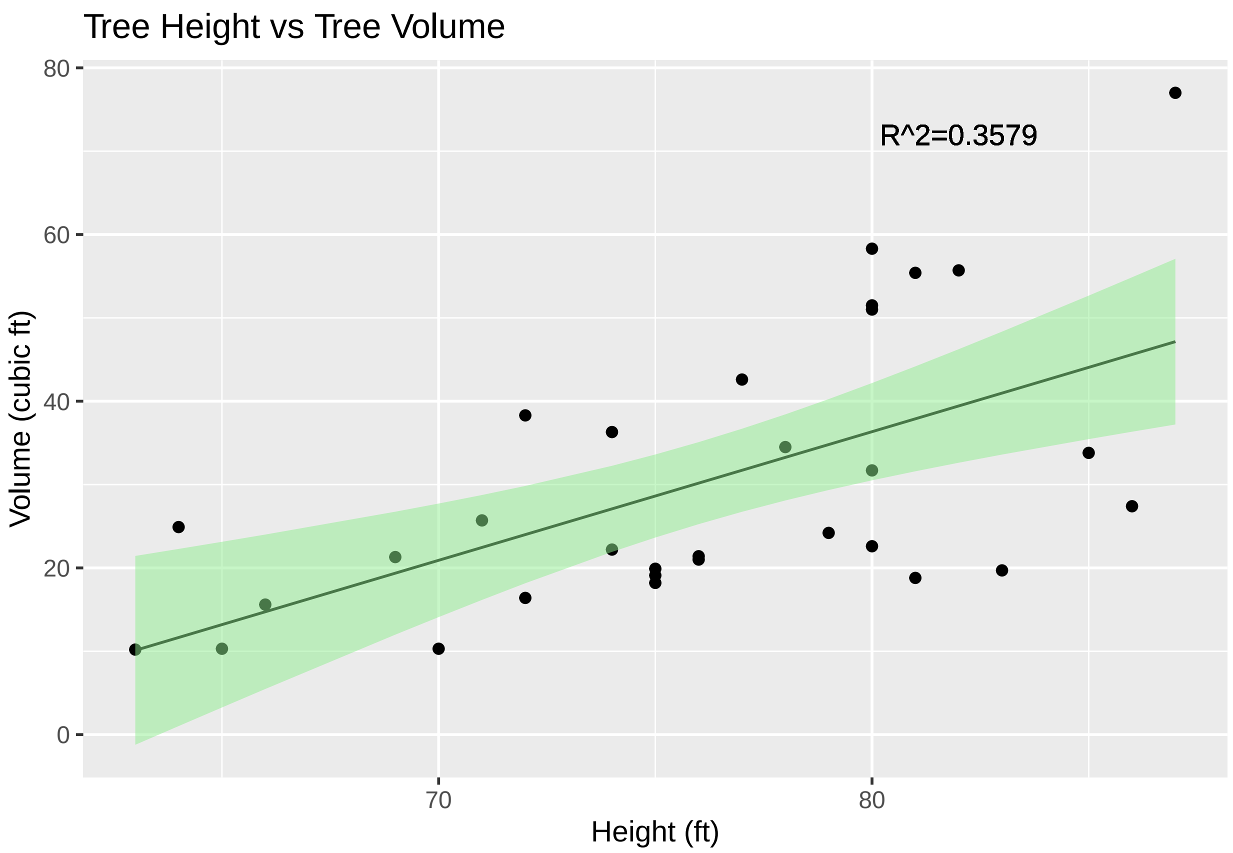
e. Graph the data and fitted regression line and uncertainty ribbon.

```
ggplot(data = trees.w.pred, aes(x=Height, y=Volume)) +
  geom_point() +
  geom_line(aes(y=fit)) +
  geom_ribbon(aes(ymin=lwr, ymax=upr), alpha=0.5, fill = "lightgreen") +
  labs( title='Tree Height vs Tree Volume') +
  labs( x="Height (ft)", y="Volume (cubic ft)")
```



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

```
ggplot(data = trees.w.pred, aes(x=Height, y=Volume)) +
  geom_point() +
  geom_line(aes(y=fit)) +
  geom_ribbon(aes(ymin=lwr, ymax=upr), alpha=0.5, fill = "lightgreen") +
  geom_text(data = trees.w.pred, aes(x=82, y=72, label="R^2=0.3579")) +
  labs( title='Tree Height vs Tree Volume') +
  labs( x="Height (ft)", y="Volume (cubic ft)")
```


Tree Height vs Tree Volume
R^2=0.3579
Volume (cubic ft)
80
60
40
20
0
70
80
Height (ft)