Assign. 1 STA 445

Zachary Hallemeyer

2024-02-22

```
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.

```
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.

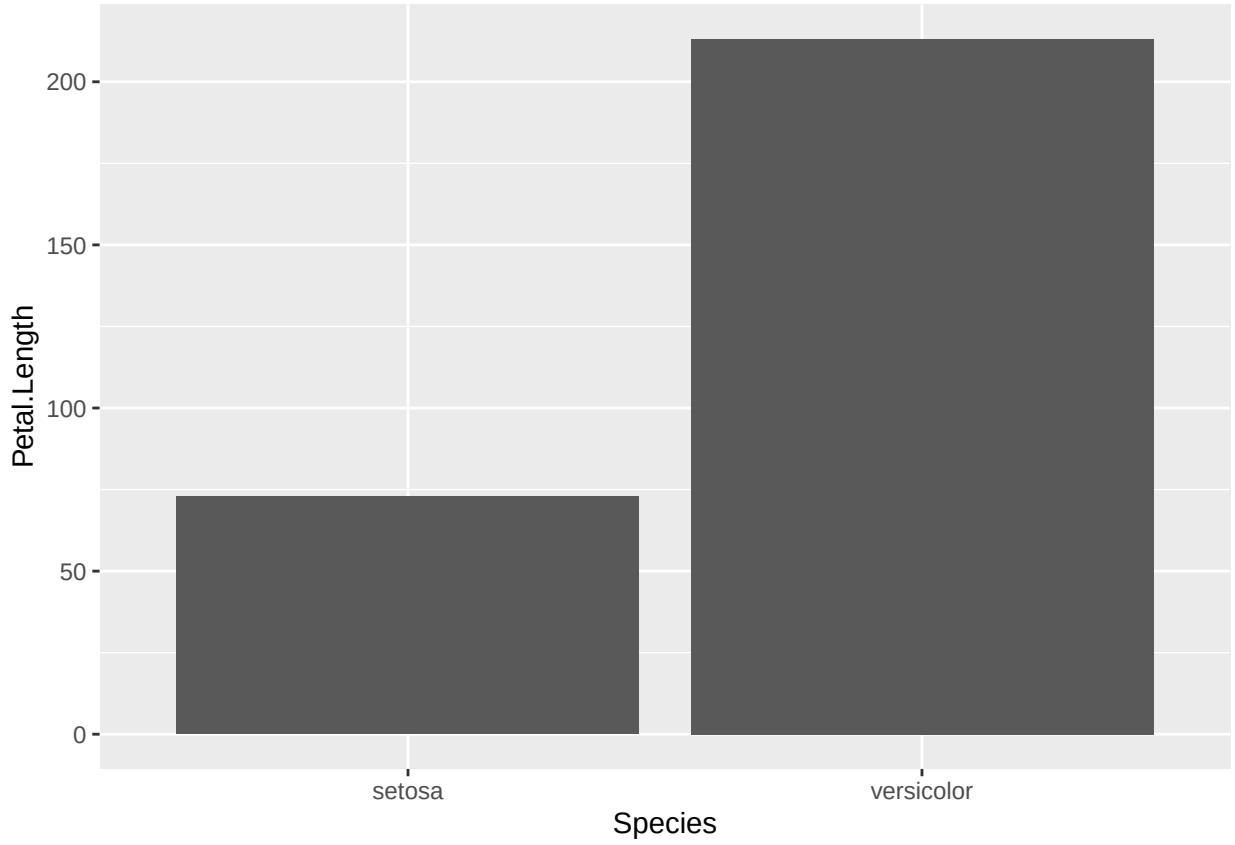
```
setOrVersi <- iris%>% filter(Species == 'setosa' | Species == 'versicolor')

slice_sample(setOrVersi, n=20)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1           6.9         3.1          4.9         1.5 versicolor
## 2           5.2         3.4          1.4         0.2     setosa
## 3           6.5         2.8          4.6         1.5 versicolor
## 4           4.8         3.4          1.6         0.2     setosa
## 5           5.4         3.7          1.5         0.2     setosa
## 6           5.7         2.8          4.5         1.3 versicolor
## 7           6.7         3.0          5.0         1.7 versicolor
```

```
## 8          6.3         2.3          4.4         1.3 versicolor
## 9          5.8         2.7          4.1         1.0 versicolor
## 10         5.7         3.8          1.7         0.3     setosa
## 11         5.2         3.5          1.5         0.2     setosa
## 12         6.3         2.5          4.9         1.5 versicolor
## 13         5.4         3.9          1.3         0.4     setosa
## 14         4.4         2.9          1.4         0.2     setosa
## 15         4.7         3.2          1.3         0.2     setosa
## 16         5.1         2.5          3.0         1.1 versicolor
## 17         4.8         3.1          1.6         0.2     setosa
## 18         4.4         3.0          1.3         0.2     setosa
## 19         6.0         2.2          4.0         1.0 versicolor
## 20         6.1         3.0          4.6         1.4 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(data=setOrVersi, aes(x=Species, y=Petal.Length)) + geom_bar(stat='identity')
```



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=setOrVersi, 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?

p-value < 2.2e-16

The petal length differs between species

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

```
t.test(data = setOrVersi, 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 equ
## 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=setOrVersi, Petal.Length ~ Species, conf.level=.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
```

g. What is the mean petal length for setosa?

```
mean( (iris %>% filter(Species == 'setosa') )$Petal.Length )
```

```
## [1] 1.462
```

h. What is the mean petal length for versicolor?

```
mean( (iris %>% filter(Species == 'versicolor') )$Petal.Length )
```

```
## [1] 4.26
```

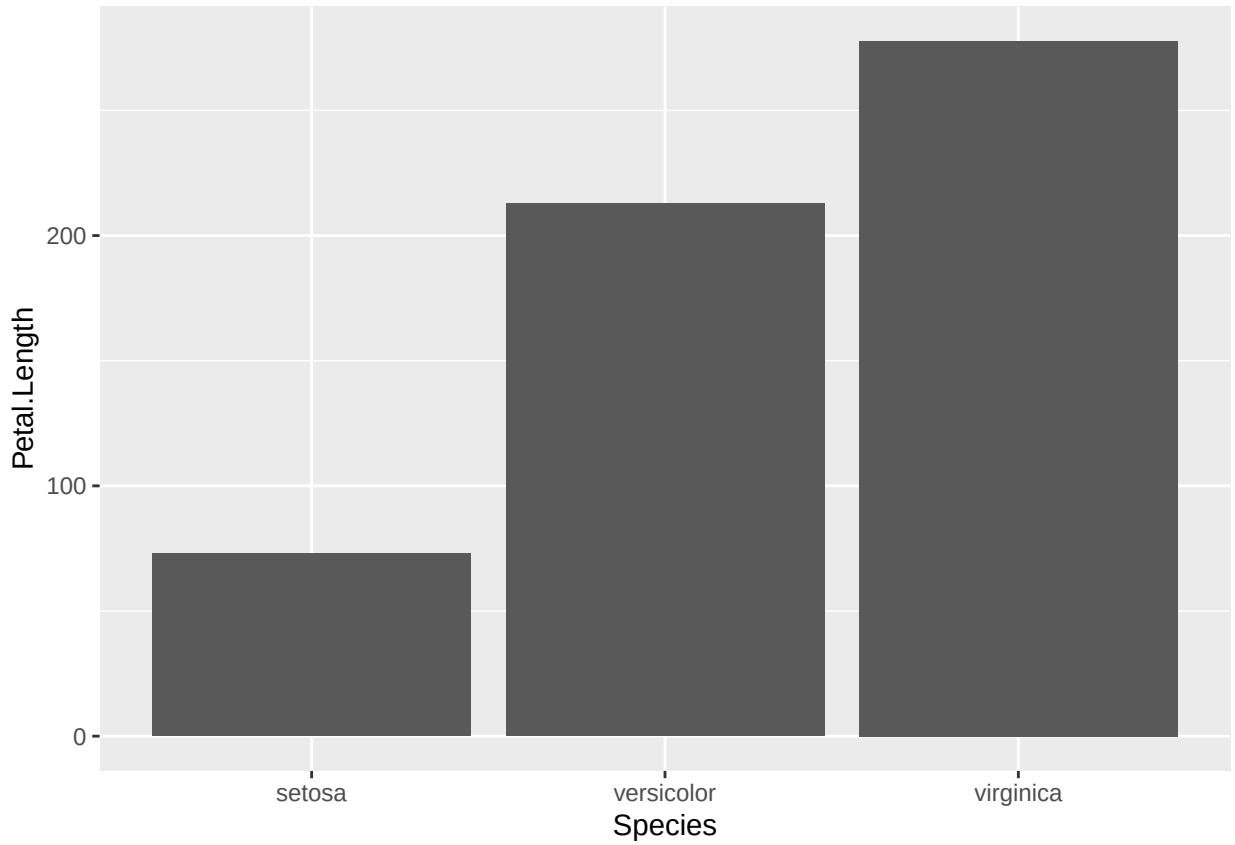
Problem 2: ANOVA

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?

Yes, there seems to be difference between petal lengths between species

```
ggplot(data=iris, aes(x=Species, y=Petal.Length)) + geom_bar(stat='identity')
```



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

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

irisModel
```

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

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

```
anova(irisModel)
```

```
## Analysis of Variance Table
##
## Response: Sepal.Length
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Species     2 63.212  31.606  119.26 < 2.2e-16 ***
## Residuals 147 38.956   0.265
## ---
## 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.

P-value < 2.2e-16

There is a difference in Sepal.Length between species.

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

```
summary(irisModel)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Species, 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|)
## (Intercept)         5.0060     0.0728  68.762  < 2e-16 ***
## Speciesversicolor   0.9300     0.1030   9.033 8.77e-16 ***
## Speciesvirginica    1.5820     0.1030  15.366  < 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:  0.6187, Adjusted R-squared:  0.6135
## F-statistic: 119.3 on 2 and 147 DF,  p-value: < 2.2e-16
```

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

```
mean( (iris %>% filter(Species == 'setosa') )$Sepal.Length )
```

```
## [1] 5.006
```

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

```
mean( (iris %>% filter(Species == 'versicolor') )$Sepal.Length )
```

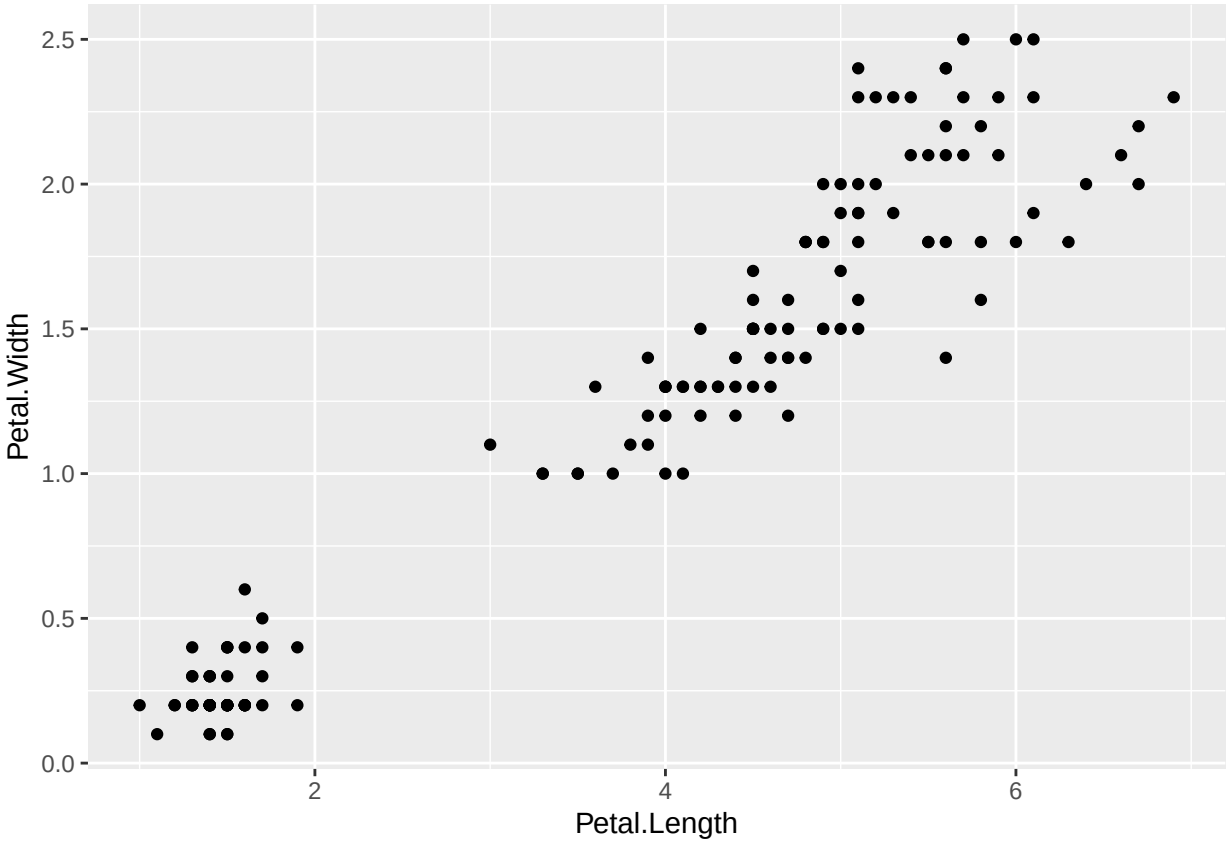
```
## [1] 5.936
```

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.Length, y=Petal.Width)) + 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.

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

c. What is the estimate of the slope parameter?

```
summary(petalModel)$coef["Petal.Width", "Estimate"]
```

```
## [1] 2.22994
```

d. What is the estimate of the intercept parameter?

```
summary(petalModel)$coef["(Intercept)", "Estimate"]
```

```
## [1] 1.083558
```

e. Use summary() to get additional information.

```
summary(petalModel)
```

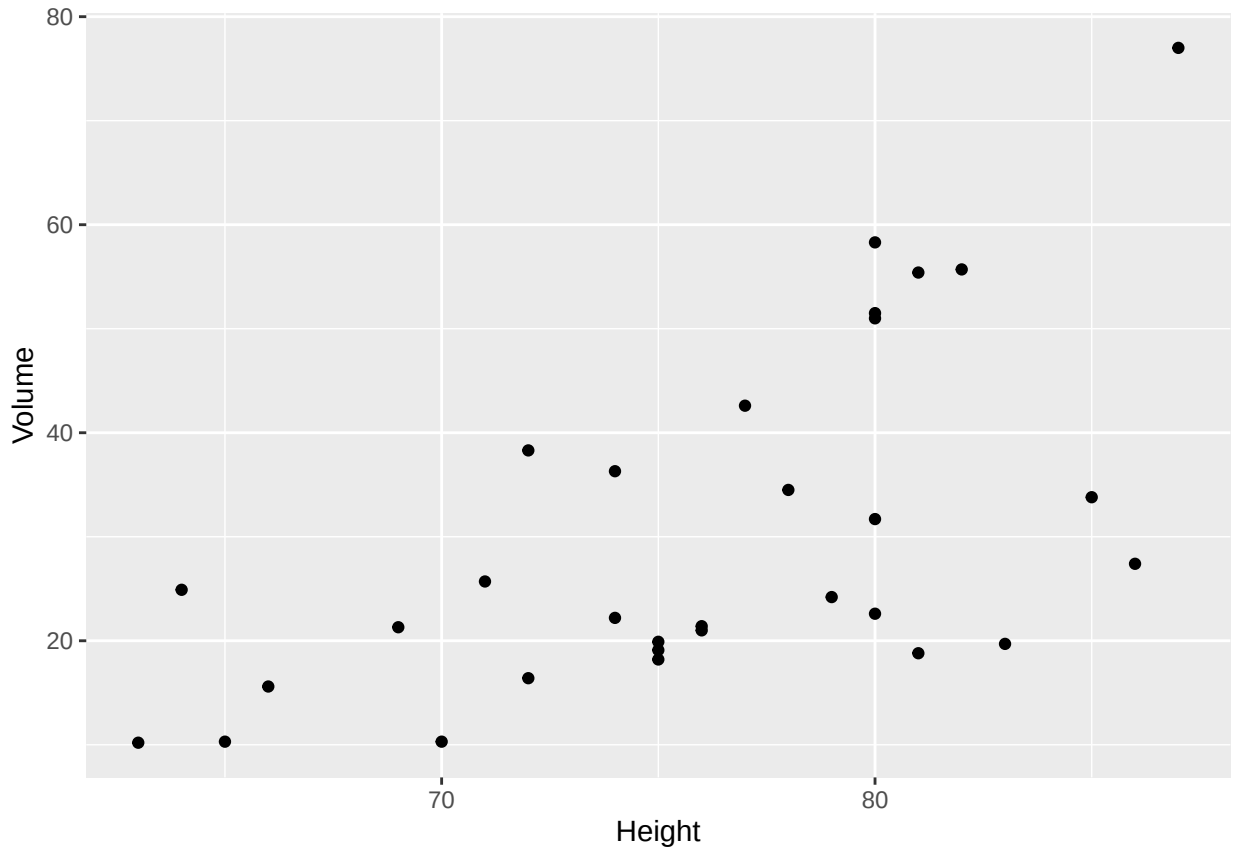
```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33542 -0.30347 -0.02955  0.25776  1.39453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.08356    0.07297   14.85   <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.

```
data(trees)

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



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

```
treeModel <- 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(treeModel)
```

```
##
## 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|)
```

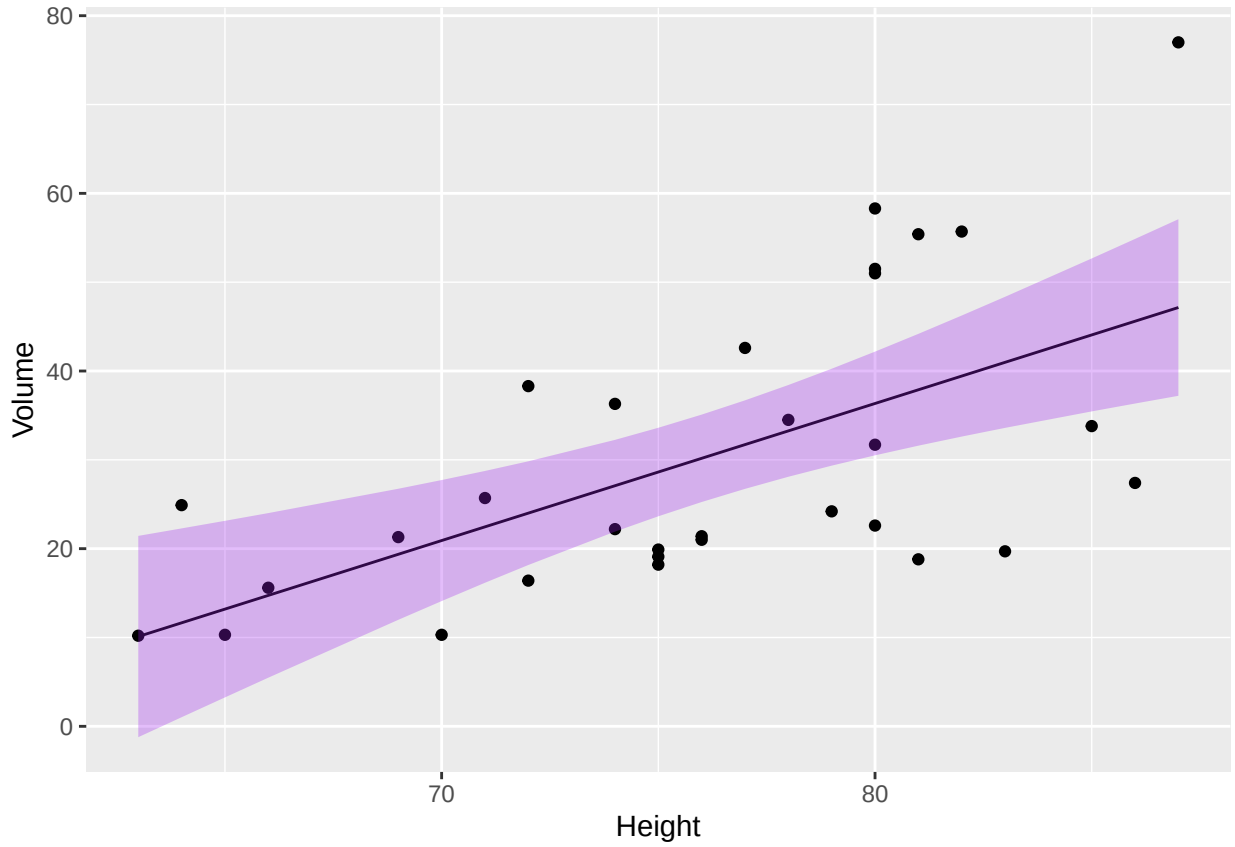
```
## (Intercept) -87.1236    29.2731  -2.976 0.005835 **
## Height        1.5433     0.3839   4.021 0.000378 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 29 degrees of freedom
## Multiple R-squared:  0.3579, Adjusted R-squared:  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.

```
newTrees <- cbind(  trees, predict(treeModel, interval = "confidence") )
```

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

```
ggplot(data=newTrees, aes(x=Height, y=Volume)) +
        geom_point() +
        geom_line(aes(y = fit)) +
        geom_ribbon(aes(ymin = lwr, ymax = upr), alpha = 0.3, fill = "purple")
```



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

```
ggplot(data=newTrees, aes(x=Height, y=Volume)) +
        geom_point() +
        geom_line(aes(y = fit)) +
        geom_ribbon(aes(ymin = lwr, ymax = upr), alpha = 0.3, fill = "purple") +
        annotate('label', label=paste("R-Squared: ", round(summary(treeModel)$r.squared), 2), x=66, y=75
```