Assign. 1 STA 445

Connor Phillips

2/20/2024

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

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.

```
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.

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

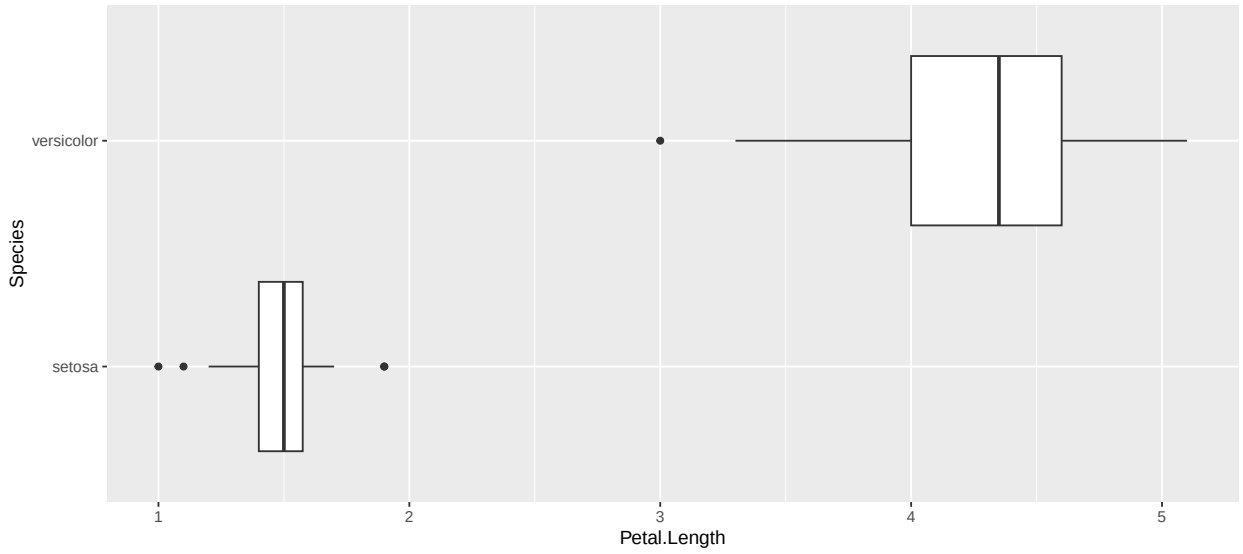
```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1           5.0         2.0          3.5         1.0 versicolor
## 2           5.5         3.5          1.3         0.2     setosa
## 3           5.7         2.8          4.1         1.3 versicolor
## 4           5.0         3.0          1.6         0.2     setosa
## 5           7.0         3.2          4.7         1.4 versicolor
## 6           6.3         2.5          4.9         1.5 versicolor
## 7           5.0         3.5          1.3         0.3     setosa
```

```
## 8          6.6         3.0          4.4         1.4 versicolor
## 9          5.7         2.6          3.5         1.0 versicolor
## 10         5.1         3.7          1.5         0.4     setosa
## 11         4.9         2.4          3.3         1.0 versicolor
## 12         4.7         3.2          1.6         0.2     setosa
## 13         5.7         2.9          4.2         1.3 versicolor
## 14         6.6         2.9          4.6         1.3 versicolor
## 15         6.0         2.7          5.1         1.6 versicolor
## 16         5.4         3.7          1.5         0.2     setosa
## 17         6.7         3.0          5.0         1.7 versicolor
## 18         4.6         3.1          1.5         0.2     setosa
## 19         5.1         2.5          3.0         1.1 versicolor
## 20         4.5         2.3          1.3         0.3     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?

Yes, the mean petal length of the setosas (1.5) is significantly smaller than the mean petal length of the versicolors (4.3).

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




```
iris.subset %>%
  group_by(Species) %>%
  summarise(mean.Length = mean(Petal.Length))
```

```
## # A tibble: 2 x 2
##   Species    mean.Length
##   <fct>            <dbl>
## 1 setosa            1.46
## 2 versicolor        4.26
```

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, 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 2.2e-16, which is very low and means we can reject the null hypothesis.

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

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

The mean petal length of the setosas is 1.5.

h. What is the mean petal length for versicolor?

The mean petal length of the versicolors is 4.3.

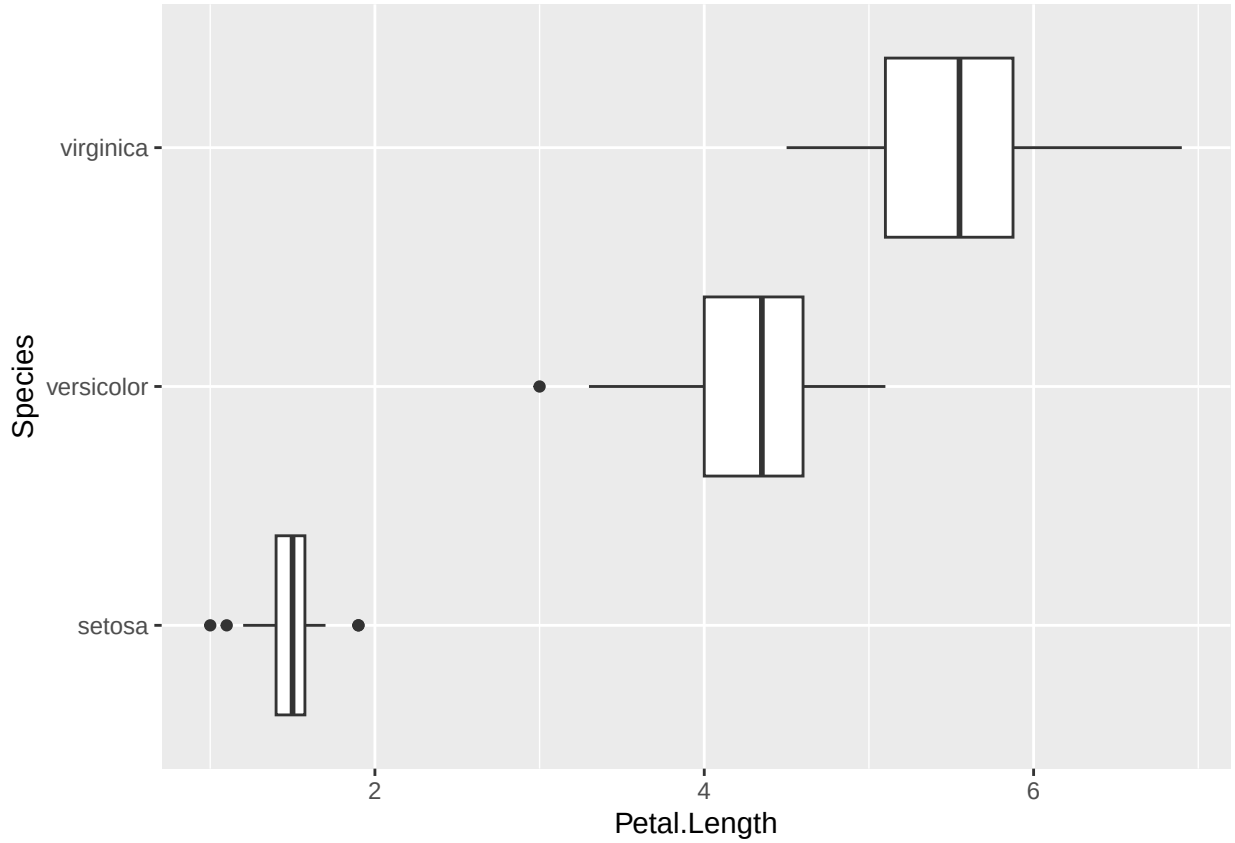
Problem 2: ANOVA

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, virginica has a higher mean pedal length than either of the previous two species.

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



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

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

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

```
anova(iris.model)
```

```
## 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, which is very low, and we can reject the null hypothesis.

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

```
summary(iris.model)
```

```
##
## 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 for the species setosa is 5.01.

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

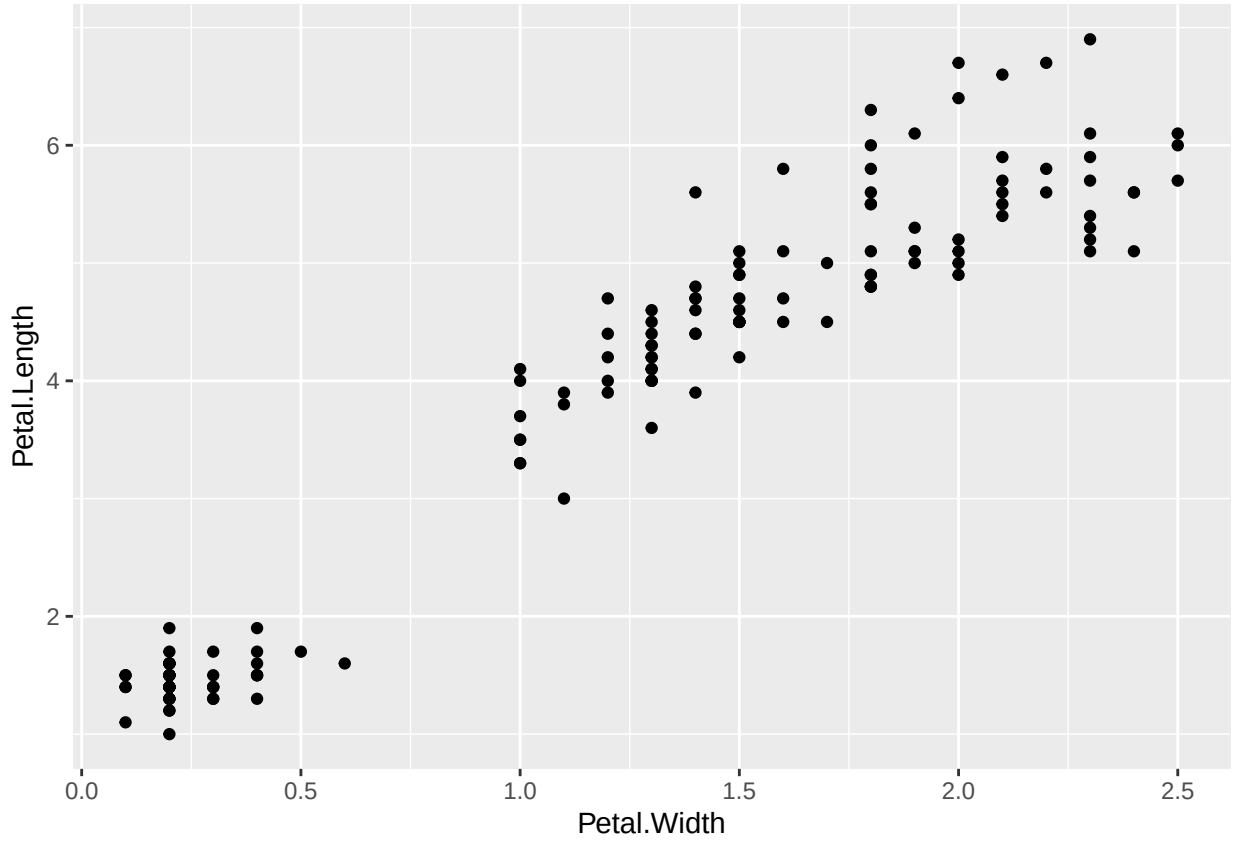
The mean petal length for the species versicolor is 5.94.

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.model = lm(data=iris, Petal.Length ~ Petal.Width)
petal.model
```

```
##
## 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?

The estimate of the slope parameter is 2.230.

d. What is the estimate of the intercept parameter?

The estimate of the intercept parameter is 1.084.

e. Use summary() to get additional information.

```
summary(petal.model)
```

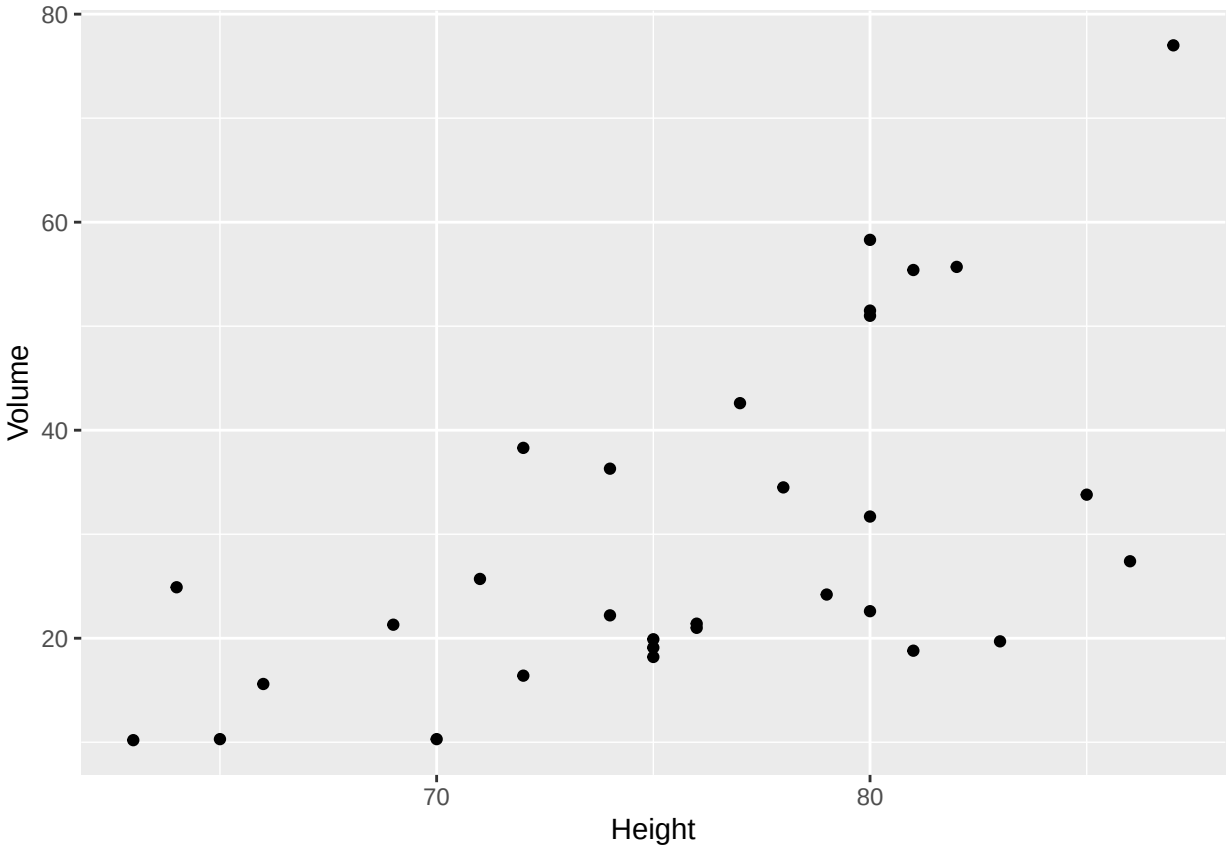
```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33542 -0.30347 -0.02955  0.25776  1.39453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.08356    0.07297   14.85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



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

```
tree.model = lm(data=trees, Volume ~ Height)
tree.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(tree.model)$coef
```

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -87.12361 29.2731221 -2.976232 0.0058346689
## Height        1.54335  0.3838693  4.020509 0.0003783823
```

```
confint(tree.model, level=0.95)
```

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

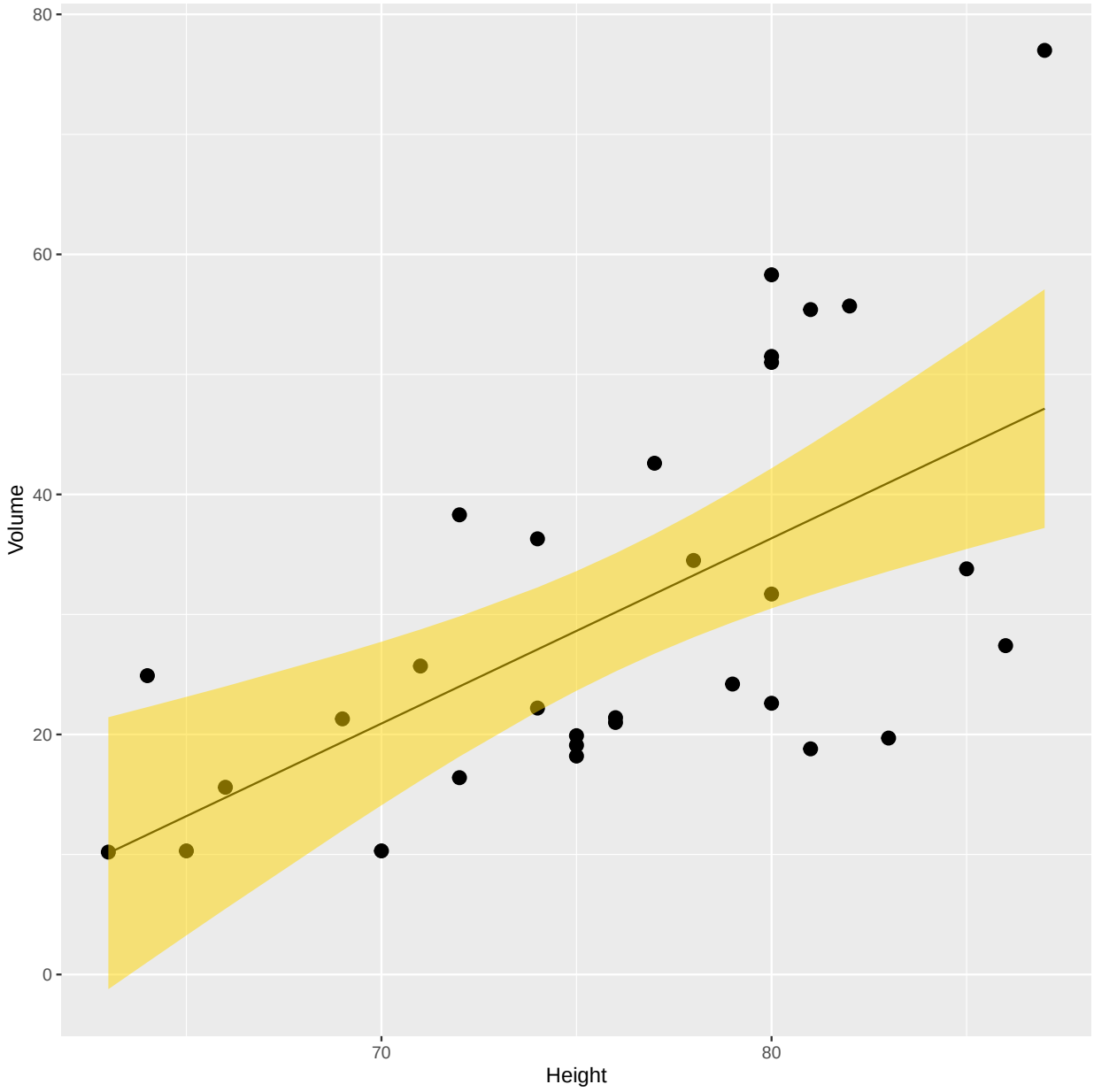
d.Add the model fitted values to the `trees` data frame along with the regression model confidence intervals. Note: the book does this in a super convoluted way. Don't follow the model in the book. Instead try cbind.

```
trees.w.pred = cbind(trees, predict(tree.model, interval = "confidence"))
head(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
```

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

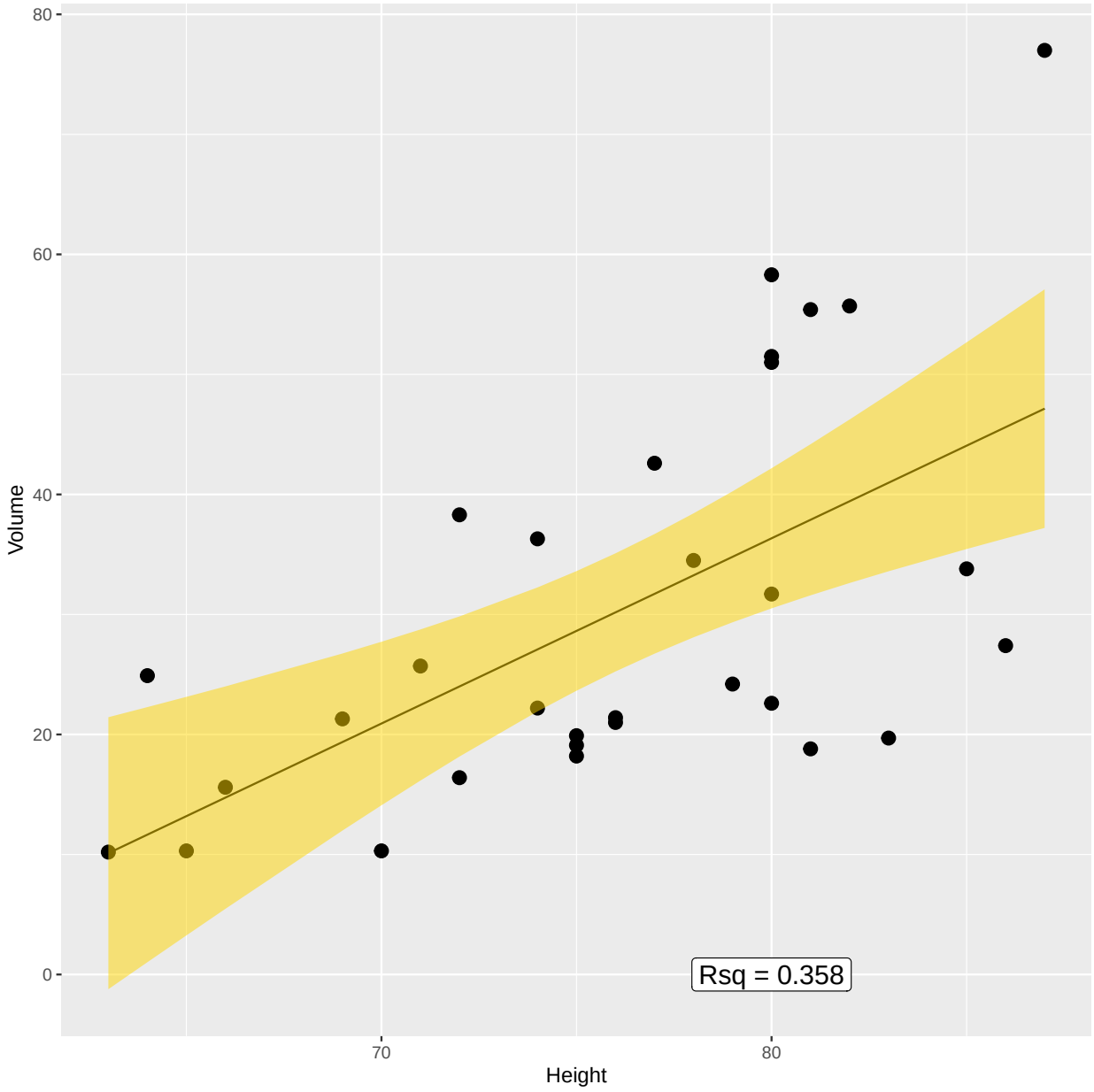
```
ggplot(data = trees.w.pred, aes(x=Height, y=Volume)) +
  geom_point(size=3) +
  geom_line(aes(y=fit)) +
  geom_ribbon(aes(ymin=lwr, ymax=upr), alpha=0.5, fill = "gold")
```



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

```
tree.r.squared = summary(tree.model)$r.squared
tree.r.sq = paste('Rsq =', round(tree.r.squared, digits = 3))

ggplot(data = trees.w.pred, aes(x=Height, y=Volume)) +
  geom_point(size=3) +
  geom_line(aes(y=fit)) +
  geom_ribbon(aes(ymin=lwr, ymax=upr), alpha=0.5, fill = "gold") +
  annotate('label', label=tree.r.sq, x=80, y=0, size=5)
```


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