Assign. 1 STA 445

Levi Mault

2024-04-16

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.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2

## Warning: package 'ggplot2' was built under R version 4.3.3

## Warning: package 'tidyr' was built under R version 4.3.3

## Warning: package 'readr' was built under R version 4.3.3

## Warning: package 'purrr' was built under R version 4.3.3

## Warning: package 'dplyr' was built under R version 4.3.3

## Warning: package 'stringr' was built under R version 4.3.3

## Warning: package 'lubridate' was built under R version 4.3.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

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.

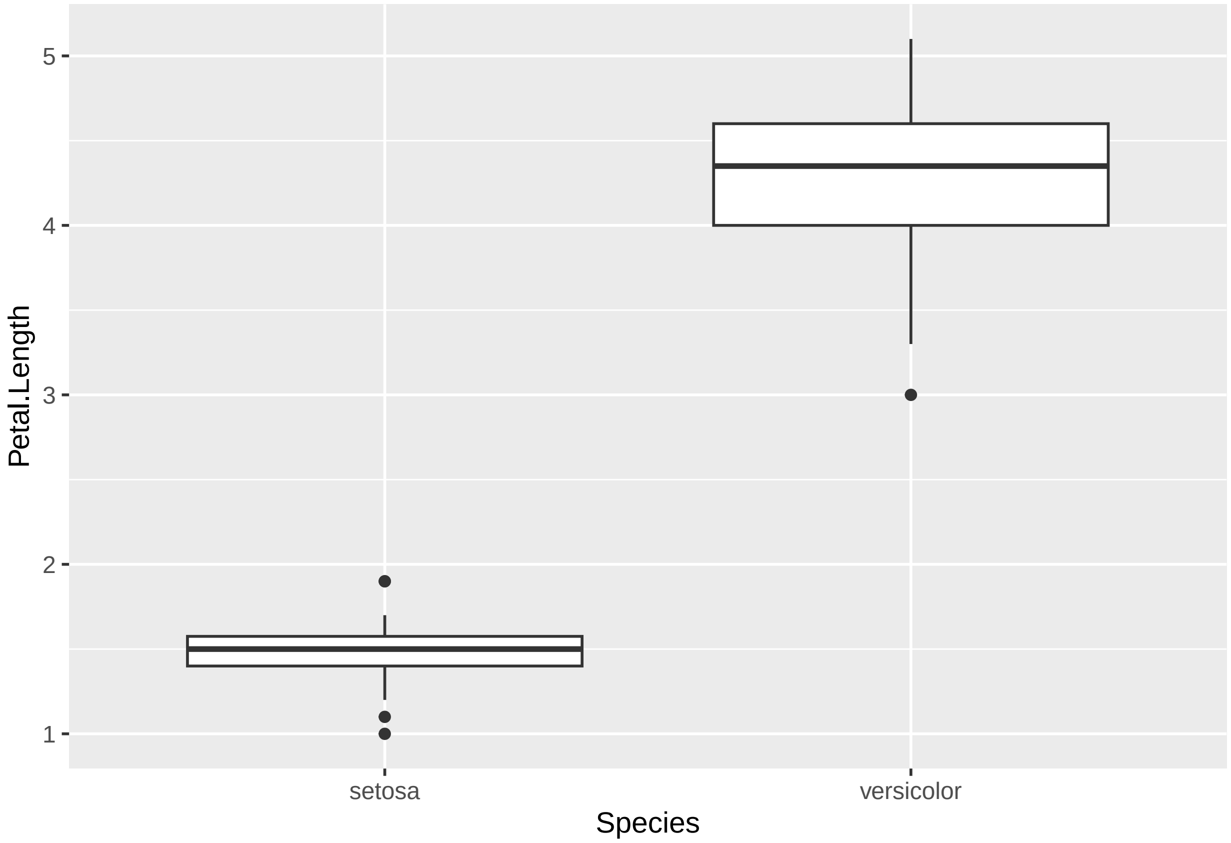
```
iris.filtered <- iris %>%
  filter( Species == "setosa" | Species == "versicolor")

slice_sample(iris.filtered, n=20)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1           5.4         3.4          1.5         0.4     setosa
## 2           6.7         3.1          4.7         1.5 versicolor
## 3           6.4         2.9          4.3         1.3 versicolor
## 4           5.5         2.6          4.4         1.2 versicolor
## 5           5.1         3.3          1.7         0.5     setosa
## 6           5.0         3.6          1.4         0.2     setosa
## 7           5.7         2.6          3.5         1.0 versicolor
## 8           5.1         3.5          1.4         0.3     setosa
## 9           5.5         3.5          1.3         0.2     setosa
## 10          4.6         3.4          1.4         0.3     setosa
## 11          4.9         3.1          1.5         0.1     setosa
## 12          6.8         2.8          4.8         1.4 versicolor
## 13          6.1         2.8          4.7         1.2 versicolor
## 14          5.4         3.9          1.7         0.4     setosa
## 15          6.7         3.0          5.0         1.7 versicolor
## 16          5.5         4.2          1.4         0.2     setosa
## 17          5.0         3.4          1.5         0.2     setosa
## 18          5.5         2.3          4.0         1.3 versicolor
## 19          6.1         2.8          4.0         1.3 versicolor
## 20          4.6         3.1          1.5         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.filtered, aes(x = Species, y = Petal.Length)) +
  geom_boxplot()
```



I

would say; yes, yes it does vary.

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.filtered, 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 resulting p-value of the t-test is 2.2e-16. Thus, we reject the null hypothesis. This provides conclusive evidence that petal lengths differ depending on the species.

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

```
petal.length.mod <- lm(data=iris.filtered, Petal.Length ~ Species)

confint(petal.length.mod)
```

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)

g. What is the mean petal length for setosa?

h. What is the mean petal length for versicolor?

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?

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

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

d. What is the p-value for the test? What do you conclude.

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

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

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

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.

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.

c. What is the estimate of the slope parameter?

d. What is the estimate of the intercept parameter?

e. Use summary() to get additional information.

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.

b. Fit a `lm` model using the command `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.

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.

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

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