Practice

Marcus Nunes

17 November, 2019

Contents

1 Instructions

Save a version of this file and call it `practice_answers.Rmd`. Remove this first section and practice what we learned today, writing a small report analyzing the `iris` dataset.

After your work is done, change the code `output: bookdown::html_document2` on line 5 to `output: bookdown::pdf_document2` in order to have a pdf. It is preferable to work with the HTML version of the report first to save time and create the pdf only after the analysis is complete.

2 Introduction

The Iris flower dataset is a well known multivariate dataset introduced in Fisher (1936). It contains 150 observations of three species of plants called *Iris setosa*, *Iris versicolor*, and *Iris virginica*. In this report we will analyze this dataset and show why it is interesting.

3 Descriptive Statistics

We will describe the most important features about the Iris flower dataset in this section.

Table 1: Regression model result when applied to data ‘iris‘.

Sepal.Length_mean	Sepal.Width_mean	Petal.Length_mean	Petal.Width_mean
5.8433	3.0573	3.758	1.1993

3.1 Tables

1. Create a table with the means for `iris` dataset. Describe your findings.

```
results <- iris %>%
  select_if(is.numeric) %>%
  summarise_all(list(mean = mean))

kable(results,
      format = "latex",
      booktabs = TRUE,
      caption = "Regression model result when applied to data `iris`.",
      digits = 4) %>%
  kable_styling(position = "center")
```

The variable with the highest mean is Sepal.Length, with 5.84cm.

3.2 Plots

2. Make scatter plots showing the relationships between all the numeric variables in this dataset. What you can see?

```
iris %>%
  select_if(is.numeric) %>%
  ggpairs()
```

It seems all the variables have linear relationships between them. Some are higher and some are lower, but they do seem linear, as we can see in Figure 1.

3. Plot the scatter plot between the two variables with the highest correlation, coloring the points according to the flower species. Is there anything special about this plot?

```
ggplot(iris, aes(x = Petal.Length, y = Petal.Width, colour = Species)) +
  geom_point() +
  labs(x = "Petal Length", y = "Petal Width")
```

2 shows that there are two main clusters in this dataset. One, to the left, with only *Iris setosa*, and other to the right, with both *Iris versicolor* and *virginica*.

4 Modeling

4. Fit a linear regression model between the two variables you found on question 3. Use the variable with the highest mean as the predictor variable.

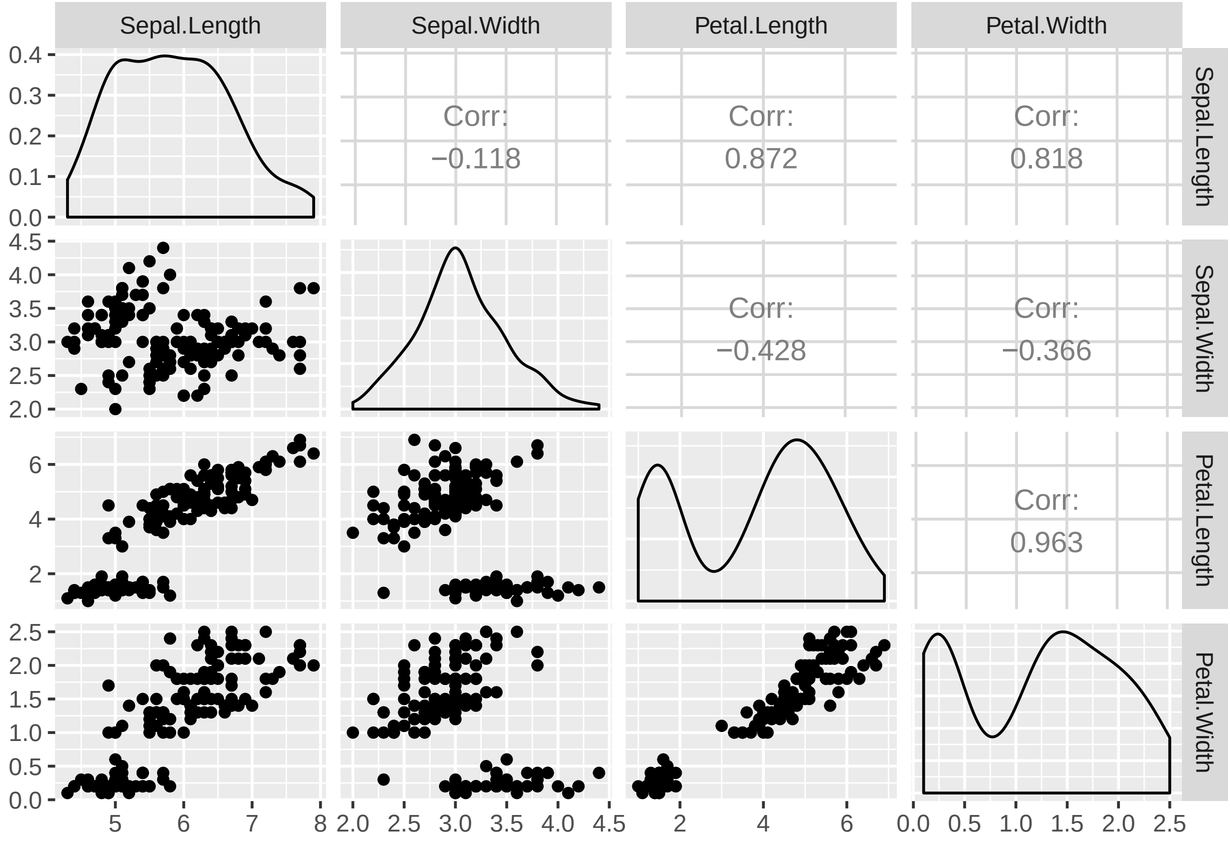

Figure 1: Exploratory data analysis for the iris dataset.

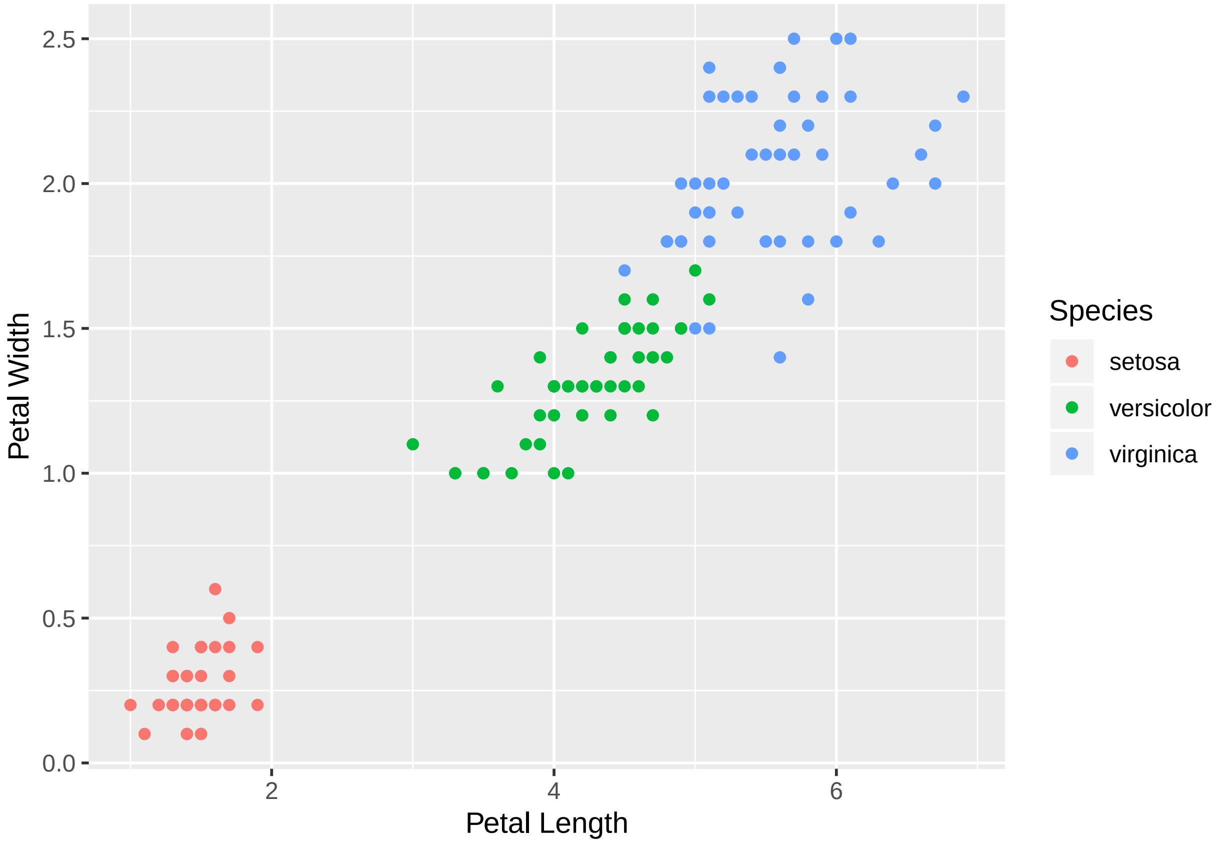

Figure 2: Scatter plot of iris dataset.

Table 2: Regression model result when applied to data ‘iris‘.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.3631	0.0398	-9.1312	0
Petal.Length	0.4158	0.0096	43.3872	0

```
fit <- lm(Petal.Width ~ Petal.Length, data = iris)
kable(summary(fit)$coefficients,
      format = "latex",
      booktabs = TRUE,
      caption = "Regression model result when applied to data `iris`.",
      digits = 4) %>%
  kable_styling(position = "center")
```

5. Plot the regression line on the scatter plot.

```
ggplot(iris, aes(x = Petal.Length, y = Petal.Width)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, colour = "black") +
  labs(x = "Petal Length", y = "Petal Width")
```

6. Make a boxplot comparing the observations of the `Sepal.Width` variable between the three plant species. Do you think there is a group whose mean is different from the others?

```
ggplot(iris, aes(x=Species, y=Sepal.Width)) +
  geom_boxplot() +
  labs(x="Species", y="Sepal Width")
```

According to Figure 4, it seems *Iris setosa* has a higher average than the other two species.

7. Test the hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_0 : \text{at least one pair } \mu_i \neq \mu_j, \text{ if } i \neq j$$

where μ_i is the mean for the variable `Sepal.Width` for the groups

- $i = 1$ (setosa)
- $i = 2$ (versicolor)
- $i = 3$ (virginica)

What is your conclusion?

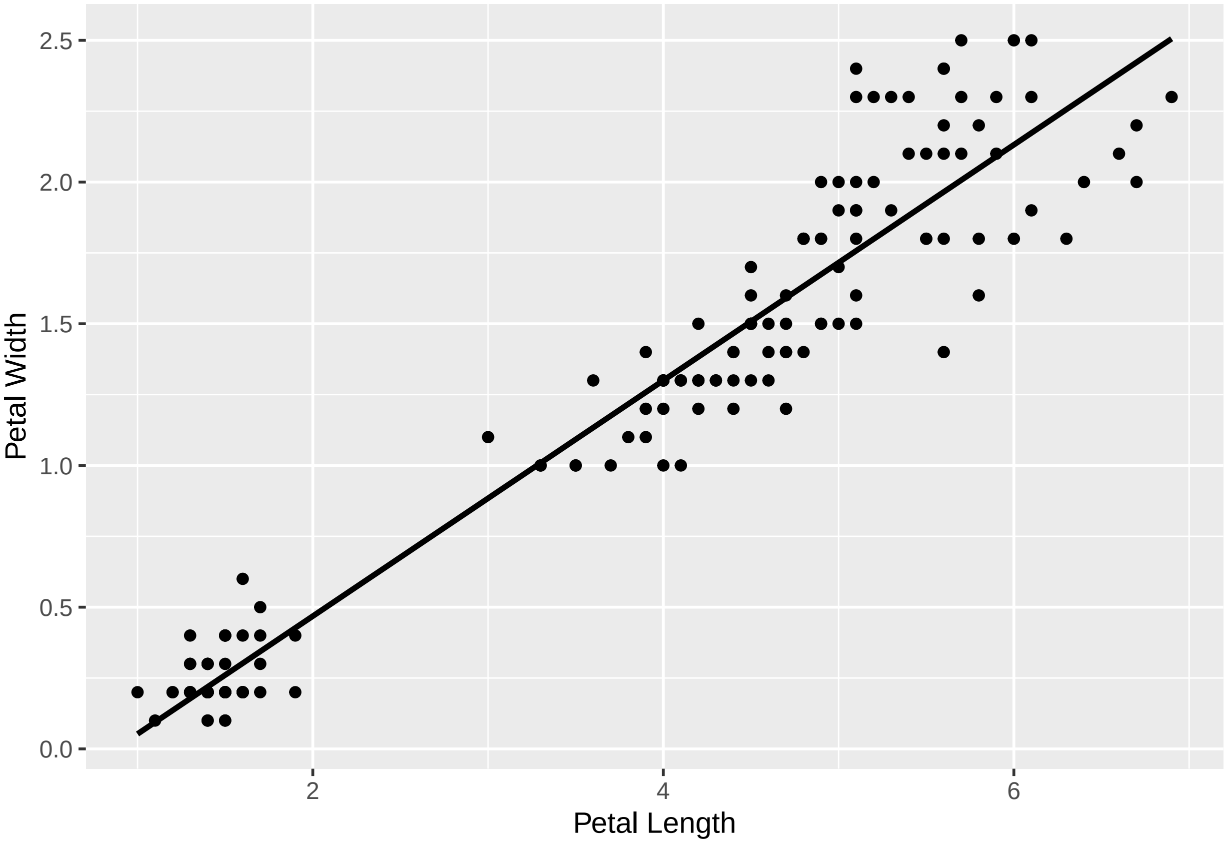

Figure 3: Linear regression model fitted to iris dataset.

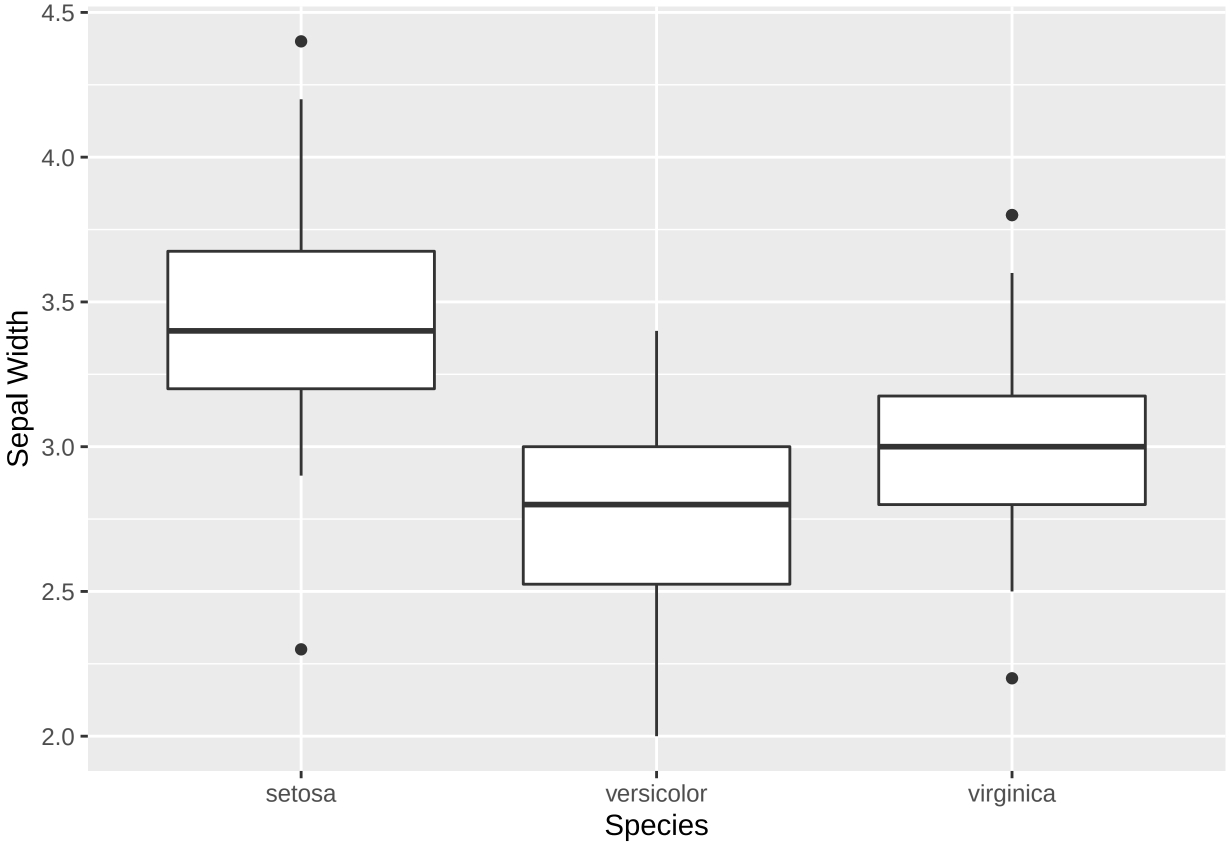

Figure 4: Boxplot comparing the means of sepal width for the iris dataset.

Table 3: ANOVA results.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species	2	11.3449	5.6725	49.16	0
Residuals	147	16.9620	0.1154		

```r
fit.anova <- aov(Sepal.Width ~ Species, data = iris)
results.anova <- summary(fit.anova)
options(knitr.kable.NA = "")
kable(results.anova[[1]],
      format = "latex",
      booktabs = TRUE,
      caption = "ANOVA results.",
      digits = 4) %>%
  kable_styling(position = "center")
```

According to Table 3, p-value $= 0$. Therefore, we reject H_0 at level $\alpha = 0,05$. Therefore, there is at least one pair $\mu_i \neq \mu_j$, if $i \neq j$.

5 Conclusion

As we could see, this dataset is great to practice statistical concepts.

References

Fisher, R. A. 1936. “The Use of Multiple Measurements in Taxonomic Problems.” *Annals of Eugenics* 7 (7): 179–88.