Intro to Regression Analysis

Maria Tackett

2019-05-14

Contents

List of Tables

List of Figures

Chapter 1

Beginning of the Book

This is the introduction to the book.

Chapter 2

Introduction

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 2. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter **??**.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 2.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 2.1.

```
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2018) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

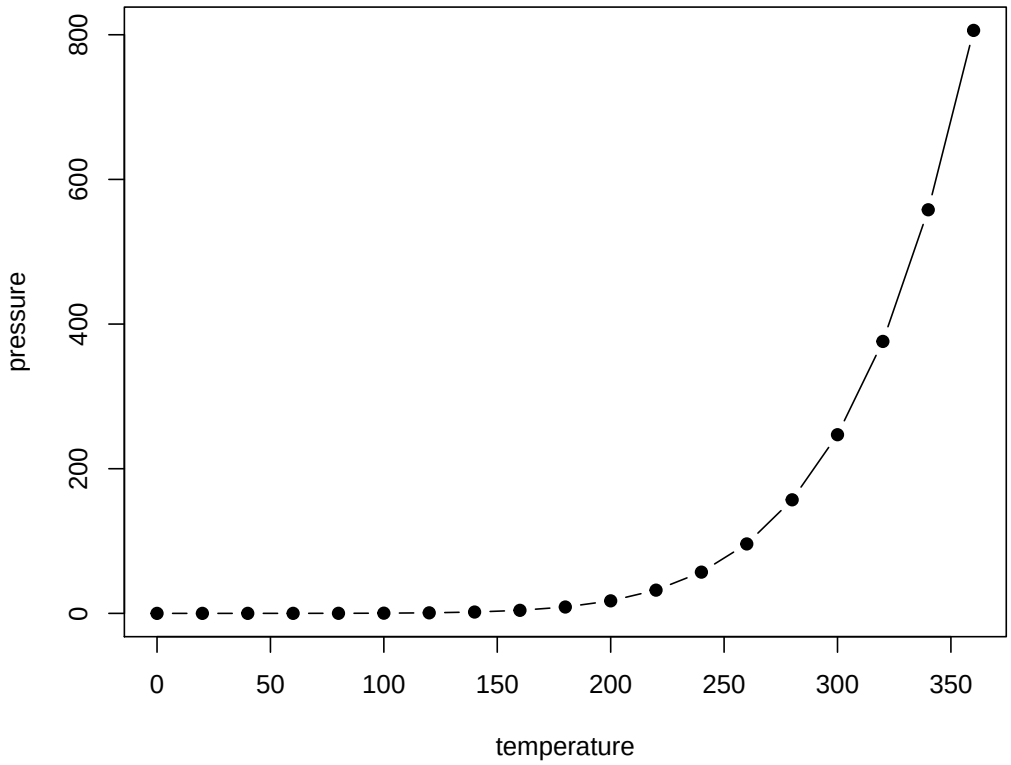

Figure 2.1: Here is a nice figure!

Table 2.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

Chapter 3

Getting Started

This is a chapter about getting started using R and GitHub.

Chapter 4

Simple Linear Regression

Computing Exercises

4.1 Computing: College Admissions

The primary goal of today's lab is to give you practice with some of the tools you will need to conduct regression analysis using R. An additional goal for today is for you to be introduced to your teams and practice collaborating using GitHub and RStudio.

4.1.1 Packages

We will use the following packages in today's lab.

In-Class Exercises

4.2 In-Class Exercise: Advertising Analysis

In this mini analysis, we will work with the `Advertising` data used in Chapters 2 and 3 of *Introduction to Statistical Learning*.

4.3 Data and packages

We start with loading the packages we'll use.

We will analyze the advertising and sales data for 200 markets. The variables we'll use are

- `tv`: total spending on TV advertising (in $thousands)
- `radio`: total spending on radio advertising (in $thousands)
- `newspaper`: total spending on newspaper advertising (in $thousands)
- `sales`: total sales (in $millions)

4.4 Analysis

We'll begin the analysis by getting quick view of the data:

Next, we can calculate summary statistics for each of the variables in the data set.

1. What type of advertising has the smallest median spending?
2. What type of advertising has the largest variation in spending?
3. Describe the shape of the distribution of `sales`.

We are most interested in understanding how advertising spending affect sales. One way to quantify the relationship between the variables is by calculating the correlation matrix.

1. What is the correlation between `radio` and `sales`? Interpret this value.
2. What type of advertising has the strongest linear relationship with `sales`?

Below are visualizations of `sales` versus each explanatory variable.

Since `tv` appears to have the strongest linear relationship with `sales`, let's calculate a simple linear regression model using these two variables.

1. Write the model equation.
2. Interpret the intercept in the context of the problem.
3. Interpret the slope in the context of the problem.

4.5 In-Class Exercise: Beer Data Analysis

In this analysis, we will analyze the relationship between the amount of alcohol (`PercentAlcohol`) and the caloric content (`CaloriesPer12Oz`) in domestic beers. Let `PercentAlcohol` be the predictor variable and `CaloriesPer12Oz` the response variable.

Due to limited class time, we will not do the exploratory data analysis in this example. In practice, however, you should always start with the exploratory data analysis.

You can add your answers to this R Markdown document.

**1. Calculate a regression model to describe the relationship between `PercentAlcohol` and `CaloriesPer12Oz`. Display the model output.*

2. Does it make sense to interpret the intercept? Why or why not?

There are non-alocoholic beers, so it is possible to have a meaningful interpretation of the intercept. In our data, however, there are very few beers with less than 3% alcoholic content, so it would not be wise to interpret the intercept. It is not safe to assume the same relationship between `PercentAlcohol` and `CaloriesPer12Oz` hold for beers with 0% alcohol; this would be extrapolation.

3. Interpret the 95% confidence interval for the slope in the context of the data.

We are 95% confident that the interval (26.557, 30.620) contains the true population slope for `PercentAlcohol`. This means we are 95% confident that for every 1% increase in alcohol content, the number of calories (per 12 oz) is expected to increase between 26.557 and 30.620 calories.

4. Find the critical value, t^*, used to calculate the 95% confidence interval. The code below is a guide; uncomment and complete the lines of code to calculate and display the critical value.

The critical value used to calculate the 95% confident interval for the slope is _____.

5. Interpret the test statistic in the context of the data

The estimated slope of 28.577 is 27.78 standard errors above the hypothesized mean of 0, assuming there is no linear relationship between percent alcohol and calories in domestic beers.

6. How was the p-value calculated? Fill in the code below to calculate the p-value. The code below is a guide; uncomment and complete the lines of code to calculate and display the p-value.

The p-value is _____. Given there is no linear relationship between `PercentAlcohol` and `CaloriesPer12Oz`, the probability of obtaining a test statistic with magnitude _____ or more extreme is _____.

7. Fill in the code below to calculate the predicted calories and corresponding 90% interval for a single beer with alcohol content of 4.3%.

8. Fill in the code below to calculate the predicted calories and corresponding 90% interval for the subset of beers with alcohol content of 4.3%.

Chapter 5

Analysis of Variance

Computing Assignments

Chapter 6

Multiple Linear Regression

Computing Assignments

In-Class Exericses

Chapter 7

(PART*) Math Notes

Chapter 8

Model Selection

Computing Assignments

In-Class Exericses

Math Notes

Chapter 9

Logistic Regression

Computing Assignments

In-Class Exericse

Chapter 10

Multinomial Logistic Regression

Computing Assignments

In-Class Exericses

Chapter 11

Special Topics

Computing Assignments

In-Class Exericses

Chapter 12

Data Sets

Data Set	Description	Chapter	Original Source
advertising.csv	test	test	test
airbnb_basic.csv	test	test	test
airbnb_details.csv	test	test	test
beer.csv	test	test	test
evals_mod.csv	test	test	test
fivethirtyeight-recent-grads.R	test	test	test
framingham.csv	test	test	test
gss2016.csv	test	test	test
KingCountyHouses.csv	test	test	test
ncbreweries.csv	test	test	test
recent-grads.csv	test	test	test
sesame.csv	test	test	test
sis.csv	test	test	test
spotify.csv	test	test	test
test_songs.csv	test	test	test

Chapter 13

References

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2018). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.9.