EDUC 784: Regression

Peter Halpin

2022-01-12

# 1 About This Book

This “ebook” provides the course notes for EDUC 784. It is currently under development, so any feedback is appreciated (e.g., during class, via email, or the edit link in the header). This first chapter is just about how to use the book – the course content starts in Chapter 2.

## 1.1 Why this book?

There are few goals of moving from “slides + exercises” to an ebook.

* To integrate course content (text, code, examples, and exercises) into one format, rather than having multiple files to sort through on Sakai.
* To address the perennial problem of choosing a textbook for this course – rather than having a supplementary text, the goal is for this ebook to be the official course text.
* Most importantly, having a course text that is “at the right level” means that I can be more liberal in assigning readings as homework *before class*, so we can spend less class time “lecturing” and more class time discussing any questions you have about the readings, going through the examples in R together, and working on assignments.
* As a bonus, this book is another example cool things you can do with R. It’s written entirely in R – that is crazy, right??

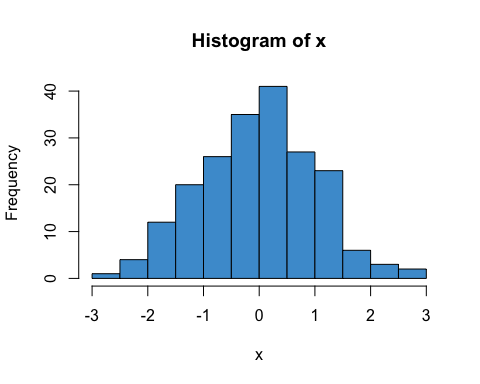
This book won’t replace Sakai (e.g., I still need to post assignments and grades somewhere). But, instead of the Sakai Lessons linking to a bunch of files on Sakai, they will link to this book.

## 1.2 How to use this book

The book is going to replace the lesson slides (powerpoint) and R exercises (Rmarkdown / HTML) familiar from EDUC 710. The main trick for doing this is called “code folding”.

An example of code folding is given below – you can see the histogram integrated into the text. Once you click on the button called “code” it will also show you the R code that produced the histogram. Note that you may need to scroll horizontally to see all of the text in the code window. Also note that when you hover your mouse over the code window, an icon appears in the top right corner – this lets you copy the full block of code with one click.

# Here is some R code. You don't have to look at it when reading the book, but it is here when you need it  
x <- rnorm(200)  
hist(x, col = "#4B9CD3")



So the basic workflow of how to use this book is as follows.

1. Before class, go through the assigned readings for conceptual understanding. You can skip all the code during your first reading.
2. We will go through the assigned readings again in class together, focusing this time on the code, as well as any questions you have.

Alright, let’s get to it!

# 2 Simple Regression

The focus of this course is linear regression with multiple predictors (AKA *multiple regression*), but we start by reviewing regression with one predictor (AKA *simple regression*). Figure 2.1 show the relationship between Grade 8 Reading Achievement and Socioeconomic Status (SES) using a sub sample of the 1988 National Educational Longitudinal Survey (NELS; see <https://nces.ed.gov/surveys/nels88/>).

# Load and attach the NELS88 data  
load("NELS.RData")  
attach(NELS)  
  
# Scatter plot  
plot(x = ses, y = achmat08, col = "#4B9CD3", ylab = "Reading Achievement (Grade 8)", xlab = "SES")  
  
# Run the regression model  
mod <- lm(achmat08 ~ ses)  
  
# Add the regression line to the plot  
abline(mod)

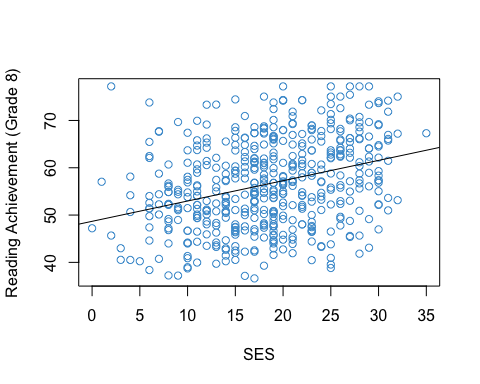


Figure 2.1: Reading Achievement and SES (NELS88).

The strength and direction of the linear relationship between the two variables is summarized by their correlation (specifically, the Pearson product moment correlation). In this sample, the correlation is

options(scipen = 1, digits = 2)  
cor(achmat08, ses)

## [1] 0.32

This is a moderate, positive correlation between Reading Achievement and SES. The correlation means that eighth graders from more well-off families (higher SES) also tended to do better in reading (higher Reading Achievement). This relationship has been widely documented and discussed in education research (e.g., <https://www.apa.org/pi/ses/resources/publications/education>). What do you think about this relationship?

In terms of Figure 2.1, the magnitude (strength) of the correlation can be interpreted in terms of how close the points are to the line. If the correlation was exactly one, all of the points would be on the line. If the correlation was zero, the points would be scattered without any (linear) trend at all. See the review materials in section @ref{review} for more info about correlation.

The line in the figure can be represented mathematically as

where

* denotes Reading Achievement (the Y-value of the blue dots)
* denotes SES (the X-value of the blue dots)
* represented the values of on the line
* represents the regression intercept (the value of when )
* represents the regression slope (how much increases for each unit of increase in )

Note that is used to represent the actual data whereas represents the points on the line. The difference is called a *residual*. The residuals for a subset of the data points in Figure 1 are shown in pink in Figure 2.2

# Get predicted values from regression model  
yhat <- mod$fitted.values  
  
# select a subset of the data  
set.seed(10)  
index <- sample.int(500, 30)  
  
# plot again  
plot(x = ses[index], y = achmat08[index], ylab = "Reading Achievement (Grade 8)", xlab = "SES")  
abline(mod)  
  
# Add pink lines  
segments(x0 = ses[index], y0 = yhat[index] , x1 = ses[index], y1 = achmat08[index], col = 6, lty = 2)  
  
# Overwrite dots to make it look at bit better  
points(x = ses[index], y = achmat08[index], col = "#4B9CD3", pch = 16)

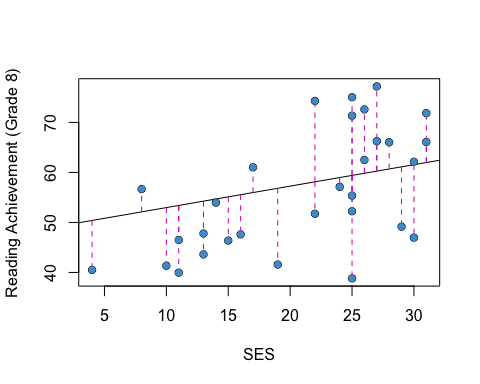


Figure 2.2: Residuals for a Subsample of Figure 1.

Intuitively, one approach to “fitting a line to the data” is to select the parameters of the line (its slope and intercept) to minimize the residuals. In ordinary least squares (OLS) regression, we minimize a related quantity, the sum of squared residuals:

where indexes the respondents in the sample. OLS regression is very widely used – people will assume you mean “OLS regression” when you say “regression”, unless you clarify otherwise.

Solving the minimization problem (i.e., doing the calculus) gives the following equations for the regression parameters

If you aren’t familiar with the symbols in these equations, check out the review materials in section ?? before moving on to the next section.

Note that if and are transformed to z-scores (i.e., to have mean of zero and variance of one – see review), then

So, regression, correlation, and covariance are all very closely related in simple regression. This is why we didn’t make a big deal about regression in EDUC 710 – its basically just the same thing as correlation. But when we get to multiple regression (i.e., more than one variable), we will see that relationship between regression and correlation (and covariance) gets more complicated.

## 2.1 Three uses of regression

Before moving onto more complicated regression models, let’s consider why we might be interested in them first place. As discussed in this section, regression has three main uses: \* Prediction (focus on ) \* Causation (focus on ) \* Explanation (focus on , which is defined below)

## 2.2 Regression for prediction

Prediction (etymology: “to make known beforehand”) means that we want to use to make a guess about . This use of regression makes the most sense when we know the value of before we know the value of .

When we are interested in using values of to make predictions about (yet unobserved) values of , we use our guess. This is why is called the “predicted value” of .

When making predictions, we usually want some additional information about how wrong we are likely to be. This information is provided by the prediction error variance (cite):

which leads to the t-statistic (and corresponding confidence intervals)

for the unknown value of , with degrees of freedom. The confidence intervals for the predictions in our sub set of the example in Figure 2.2 are depicted in Figure 2.3

# Using a different plotting function that adds prediction error bands easily  
ggplot2::ggplot(NELS[index, ], aes(x = ses, y = achmat08)) +   
 geom\_point(color='#3B9CD3', size = 2) +   
 geom\_smooth(method = lm, color = "grey35") +   
 ylab("Reading Achievement (Grade 8)") +   
 xlab("SES") +   
 theme\_bw()

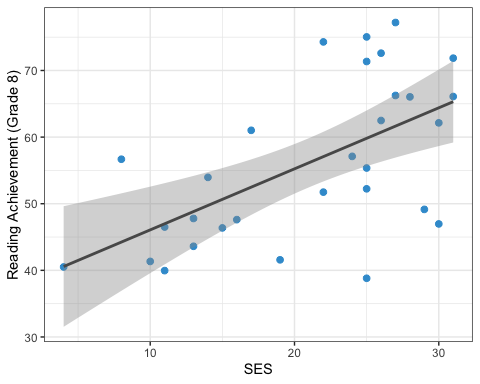


Figure 2.3: Prediction Error for Example Data.

Prediction was the original use of regression (see <https://en.wikipedia.org/wiki/Regression_toward_the_mean#History>). More recent methods developed in machine learning also focus mainly on prediction – although the methods used in machine learning are often more complicated than OLS regression, and the research context is usually quite different, the basic problem is the same. Machine learning has led to the use of out of sample predictions, rather than prediction error, as the main criterion for judging the quality of predictions made from a model. Machine learning has also introduced some new techniques for choosing which among many potential predictors ( variables) are most important for obtaining high quality predictions. We will touch on these topics later in the course, although, as mentioned, our main focus is OLS regression.

## 2.3 The lm function

The functionlm, short for “linear model”, can estimate linear regressions using OLS and provide a lot of useful output. The main argument that the user provieds to the lm function is a formula. For the simple regression of Y on X, a formula has the syntax

Y ~ X

Here Y denotes the outcome variable, the tilde ~ roughly means “equals”, and X is the predictor variable. We will see more complicated formulas as we go through the course. For more information on R’s formula syntax, see help(formula).

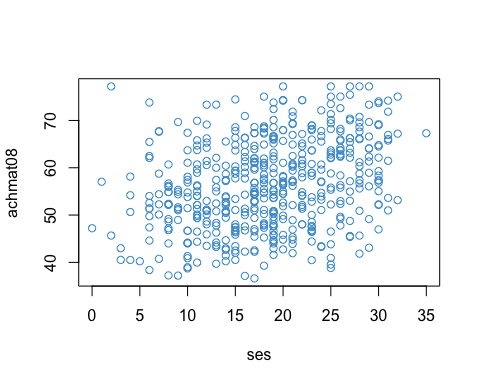
Let’s take a closer look using the following two variables from the NELS data (see Sakai resource folder for more information on the data).

* achmat08: eighth grade math achievement (percent correct on a math test)
* ses: a composite measure of socio-economic status, on a scale from 0-35

# Load and attach the data -- see last week's exercises for details   
load("NELS.RData")  
attach(NELS)

## The following objects are masked from NELS (pos = 3):  
##   
## absent12, achmat08, achmat10, achmat12, achrdg08, achrdg10,  
## achrdg12, achsci08, achsci10, achsci12, achsls08, achsls10,  
## achsls12, advmath8, alcbinge, algebra8, apoffer, approg, cigarett,  
## computer, cuts12, edexpect, excurr12, expinc30, famsize, gender,  
## homelang, hsprog, hwkin12, hwkout12, id, late12, marijuan, numinst,  
## nursery, parmarl8, region, schattrt, schtyp8, ses, slfcnc08,  
## slfcnc10, slfcnc12, tcherint, unitcalc, unitengl, unitmath, urban

plot(x = ses, y = achmat08, col = "#4B9CD3")



# Regress math achievement on SES  
mod <- lm(achmat08 ~ ses)  
  
# Print out the regression coefficients  
coef(mod)

## (Intercept) ses   
## 48.68 0.43

Let’s do some quick calcuations to check that the lm output corresponds the formulas for the slope and intercept we used in the lesson:

# Confirm that the slope from lm is just the covariance divided by the variance of X  
cov\_xy <- cov(achmat08, ses)  
s\_x <- var(ses)  
b <- cov\_xy / s\_x  
b

## [1] 0.43

# Confirm that the y-intercept is obtained from the two means and the slope  
xbar <- mean(ses)  
ybar <- mean(achmat08)  
  
a <- ybar - b \* xbar  
a

## [1] 49

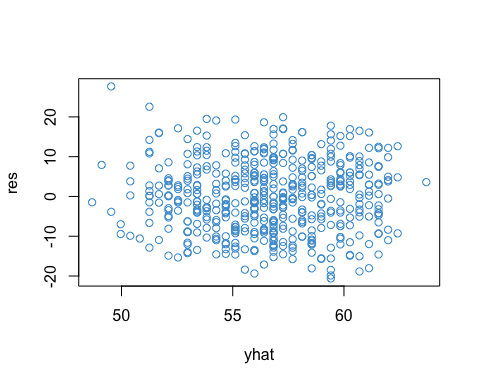
Let’s also check our interpretation of the parameters. If the answers to these questions are not clear, please make sure to ask in class or on the Sakai forum.

* What is the value of achmat08 when ses is equal to zero?
* How much do the predicted values of achmat08 increase for each unit of increase in ses?

## 2.4 Predicted values and residuals

The lm function also returns the residuals and the predicted values , which we can access using the $ operator.These will be useful later on for computing R-squared. For now we will just plot the two values to show that the predicted values and the residuals are uncorrelated, which was discussed in class.

yhat <- mod$fitted.values  
res <- mod$resid  
  
# In OLS, the predicted values and the residuals are uncorrelated:   
plot(yhat, res, col = "#4B9CD3")



cor(yhat, res)

## [1] -2.4e-16

## 2.5 Variance explained

Above we found out that the regression coefficient was 0.4-ish. Is that good? or average? or what?

One way to answer that question is by considering the amount of variation in that is associated with (or explained by) its relationship with . Recall from our lesson that one way to do this is via the variance decomposition

from which we can compute the proportion of variation in Y that is associated with the regression model

Let’s compute for our example.

# Compute the sums of squares  
ybar <- mean(achmat08)  
ss\_total <- sum((achmat08 - ybar)^2)  
ss\_reg <- sum((yhat - ybar)^2)  
ss\_res <- sum((achmat08 - yhat)^2)

# Check that SS\_total = SS\_reg + SS\_res  
ss\_total

## [1] 43527

ss\_reg + ss\_res

## [1] 43527

# Compute R-squared  
ss\_reg/ss\_total

## [1] 0.1

# Check that R-square is really equal to the square of the PPMC  
# Note -- this is only true for simple regression  
ppmc <- cor(achmat08, ses)  
ppmc^2

## [1] 0.1

## 2.6 Inference

At this point we can say that SES explained about 10% of the variation in eighth grade students’ math achievement, in our sample. However, we haven’t yet talked aboutstatistical inference, or how we can make conclusions about a population based on a sample from that population.

Let’s use the summary function to test the coefficients in our model.

mod <- lm(achmat08 ~ ses)  
summary(mod)

##   
## Call:  
## lm(formula = achmat08 ~ ses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.600 -6.552 -0.148 6.023 27.663   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 48.6780 1.1282 43.15 < 2e-16 \*\*\*  
## ses 0.4293 0.0573 7.49 3.1e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.9 on 498 degrees of freedom  
## Multiple R-squared: 0.101, Adjusted R-squared: 0.0995   
## F-statistic: 56.1 on 1 and 498 DF, p-value: 3.13e-13

In the table, the t-test and p-values are for the null hypothesis that the corresponding coefficient is zero in the population. We can see that the intercept and slope are both significantly different from zero.

The text below the table summarizes the output for R-square, including the F-test, it’s degrees of freedom, and the p-value. (We will talk about adjusted R-square later on.)

We can use the confint function to obtain confidence intervals for the regression coefficients. Use help to find out more about the confint function.

confint(mod)

## 2.5 % 97.5 %  
## (Intercept) 46.46 50.89  
## ses 0.32 0.54

Be sure to remember the correct interpretation of confidence intervals: *there is a 95% chance that the interval includes the true parameter value* (not: there is a 95% chance that the parameter falls in the interval). For example, there is a 95% chance that the interval [.31, .54] includes the true regression coefficient for SES.

## 2.7 Power analysis

Power analyses should ideally be done prospectively – i.e., before the data are collected. Since this class will work with secondary data analyses, most of our analyses will be retrospective. But don’t let this mislead you about the importance of statistical power – you should always do a power analysis before collecting data!!

To do a power analsyes in R, we can install and load the pwr package. If you haven’t installed an R package before it’s pretty straight forward – but just ask course staff or a fellow student if you run into any issues.

# Install the package   
# Note you may have to select a mirror -- I suggest using Kansas (KS)   
install.packages("pwr")

# Load the package by using the library command  
library("pwr")

# Use the help menu to see what the package does  
help("pwr-package")

To do a power analysis for linear regression, it is common to use Cohen’s as the effect size:

Recall that is the proportion of variance in explained by the model, and so is the proportion of variance not explained by the model. Thus, can be interpreted as a signal to noise ratio.

In addition to the effect size, we need to know the degrees of freedom for the F-test of R-square. The pwr functions use the following notation:

* u is the degrees of freedom in the numerator of an F-test.
* v is the degrees of freedom in the denominator of an F-test.

In simple regression, u = 1 and v = N - 2.

### 2.7.1 Example: A prospective power analysis

How many observations would be required to detect an effet size of R-square = .1, using and power = .8? To find the answer, enter the provided information into the pwr.f2.test function, and the function will solve for the “missing piece” – in this case .

# Use the provided values of R2, alpha, power (and u = 1) to solve for v = N - 2  
R2 <- .1  
f2 <- R2/(1-R2)  
pwr.f2.test(u = 1, f2 = f2, sig.level = .05, power = .8)

##   
## Multiple regression power calculation   
##   
## u = 1  
## v = 71  
## f2 = 0.11  
## sig.level = 0.05  
## power = 0.8

In this example we find that . Since , so we know that a sample size of (rounded up to 73) is required to reject the null hypothesis that , when the true popuation value is , with a power of .8 and using a significance level of .05.

## 2.8 Exercises

Try the following (assume unless otherwise stated):

1. What sample size is required to detect an R-square of .25, with a power of .9?
2. What is the smallest R-square that can be detected with the NELS data (N = 500), using
3. What is the power of the NELS data to reject the null hypothesis when R-square = .05?

### 2.8.1 Answers

Don’t read these until you’ve tried the exercises!

# Q1  
pwr.f2.test(u = 1, f2 = .25/.75, power = .9, sig.level = .05)  
  
# Q2  
q2 <- pwr.f2.test(u = 1, v = 498, power = .8, sig.level = .05)  
q2$f2 / (1 + q2$f2)  
  
# Q3  
pwr.f2.test(u = 1, v = 498, f2 = .05/.95, sig.level = .05)

# 3 Parts

You can add parts to organize one or more book chapters together. Parts can be inserted at the top of an .Rmd file, before the first-level chapter heading in that same file.

Add a numbered part: # (PART) Act one {-} (followed by # A chapter)

Add an unnumbered part: # (PART\\*) Act one {-} (followed by # A chapter)

Add an appendix as a special kind of un-numbered part: # (APPENDIX) Other stuff {-} (followed by # A chapter). Chapters in an appendix are prepended with letters instead of numbers.

# 4 Footnotes and citations

## 4.1 Footnotes

Footnotes are put inside the square brackets after a caret ^[]. Like this one.[[1]](#footnote-46)

## 4.2 Citations

Reference items in your bibliography file(s) using @key.

For example, we are using the **bookdown** package (Xie [2021](#ref-R-bookdown)) (check out the last code chunk in index.Rmd to see how this citation key was added) in this sample book, which was built on top of R Markdown and **knitr** (Xie [2015](#ref-xie2015)) (this citation was added manually in an external file book.bib). Note that the .bib files need to be listed in the index.Rmd with the YAML bibliography key.

The RStudio Visual Markdown Editor can also make it easier to insert citations: <https://rstudio.github.io/visual-markdown-editing/#/citations>

# 5 Blocks

## 5.1 Equations

Here is an equation.

You may refer to using \@ref(eq:binom), like see Equation (5.1).

## 5.2 Theorems and proofs

Labeled theorems can be referenced in text using \@ref(thm:tri), for example, check out this smart theorem ??.

For a right triangle, if denotes the *length* of the hypotenuse and and denote the lengths of the **other** two sides, we have

Read more here <https://bookdown.org/yihui/bookdown/markdown-extensions-by-bookdown.html>.

## 5.3 Callout blocks

The R Markdown Cookbook provides more help on how to use custom blocks to design your own callouts: <https://bookdown.org/yihui/rmarkdown-cookbook/custom-blocks.html>

# 6 Sharing your book

## 6.1 Publishing

HTML books can be published online, see: <https://bookdown.org/yihui/bookdown/publishing.html>

## 6.2 404 pages

By default, users will be directed to a 404 page if they try to access a webpage that cannot be found. If you’d like to customize your 404 page instead of using the default, you may add either a \_404.Rmd or \_404.md file to your project root and use code and/or Markdown syntax.

## 6.3 Metadata for sharing

Bookdown HTML books will provide HTML metadata for social sharing on platforms like Twitter, Facebook, and LinkedIn, using information you provide in the index.Rmd YAML. To setup, set the url for your book and the path to your cover-image file. Your book’s title and description are also used.

This gitbook uses the same social sharing data across all chapters in your book- all links shared will look the same.

Specify your book’s source repository on GitHub using the edit key under the configuration options in the \_output.yml file, which allows users to suggest an edit by linking to a chapter’s source file.

Read more about the features of this output format here:

<https://pkgs.rstudio.com/bookdown/reference/gitbook.html>

Or use:

?bookdown::gitbook

# 7 Cross-references

Cross-references make it easier for your readers to find and link to elements in your book.

## 7.1 Chapters and sub-chapters

There are two steps to cross-reference any heading:

1. Label the heading: # Hello world {#nice-label}.
   * Leave the label off if you like the automated heading generated based on your heading title: for example, # Hello world = # Hello world {#hello-world}.
   * To label an un-numbered heading, use: # Hello world {-#nice-label} or {# Hello world .unnumbered}.
2. Next, reference the labeled heading anywhere in the text using \@ref(nice-label); for example, please see Chapter 7.
   * If you prefer text as the link instead of a numbered reference use: [any text you want can go here](#cross).

## 7.2 Captioned figures and tables

Figures and tables *with captions* can also be cross-referenced from elsewhere in your book using \@ref(fig:chunk-label) and \@ref(tab:chunk-label), respectively.

See Figure 7.1.

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

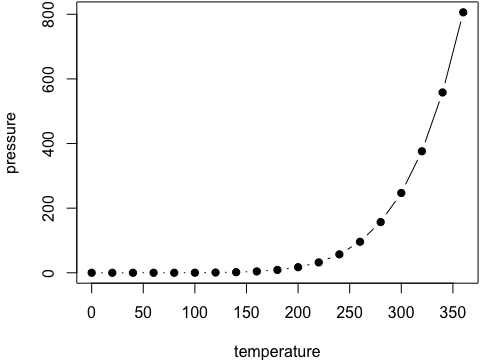


Figure 7.1: Here is a nice figure!

Don’t miss Table 7.1.

knitr::kable(  
 head(pressure, 10), caption = 'Here is a nice table!',  
 booktabs = TRUE  
)

Table 7.1: Here is a nice table!

|  |  |
| --- | --- |
| temperature | pressure |
| 0 | 0.00 |
| 20 | 0.00 |
| 40 | 0.01 |
| 60 | 0.03 |
| 80 | 0.09 |
| 100 | 0.27 |
| 120 | 0.75 |
| 140 | 1.85 |
| 160 | 4.20 |
| 180 | 8.80 |

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. <http://yihui.org/knitr/>.

———. 2021. *Bookdown: Authoring Books and Technical Documents with R Markdown*. <https://CRAN.R-project.org/package=bookdown>.

1. This is a footnote. [↑](#footnote-ref-46)