Econometrics with R at SciencesPo [WIP]

Florian Oswald and Jean-Marc Robin

2018-04-11

Contents

Chapter 1

Econometrics at ScPo - WORK IN PROGRESS

This is our book about Introductory Econometrics for 2nd year UGs at ScPo. We follow a practical approach with minimal reliance on maths and maximal reliance on practical learning.

On this first page we can explain a bit more about the course structure:

- Who is who
- how to communicate
- how many sessions etc

Figure 1.1:

Chapter 2

Introduction to R

2.1 Getting Started

R is both a programming language and software environment for statistical computing, which is *free* and *open-source*. To get started, you will need to install two pieces of software:

- R, the actual programming language.
  - Chose your operating system, and select the most recent version, 3.4.4.
- RStudio, an excellent IDE for working with R.
  - Note, you must have R installed to use RStudio. RStudio is simply an interface used to interact with R.

The popularity of R is on the rise, and everyday it becomes a better tool for statistical analysis. It even generated this book! (A skill you will learn in this course.) There are many good resources for learning R.

The following few chapters will serve as a whirlwind introduction to R. They are by no means meant to be a complete reference for the R language, but simply an introduction to the basics that we will need along the way. Several of the more important topics will be re-stressed as they are actually needed for analyses.

These introductory R chapters may feel like an overwhelming amount of information. You are not expected to pick up everything the first time through. You should try all of the code from these chapters, then return to them a number of times as you return to the concepts when performing analyses.

R is used both for software development and data analysis. We will operate in a grey area, somewhere between these two tasks. Our main goal will be to analyze data, but we will also perform programming exercises that help illustrate certain concepts.

RStudio has a large number of useful keyboard shortcuts. A list of these can be found using a keyboard shortcut – the keyboard shortcut to rule them all:

- On Windows: `Alt` + `Shift` + `K`
- On Mac: `Option` + `Shift` + `K`

The RStudio team has developed a number of "cheatsheets" for working with both R and RStudio. This particular cheatseet for Base R will summarize many of the concepts in this document.

When programming, it is often a good practice to follow a style guide. (Where do spaces go? Tabs or spaces? Underscores or CamelCase when naming variables?) No style guide is "correct" but it helps to be aware of what others do. The more import thing is to be consistent within your own code.

- Hadley Wickham Style Guide from Advanced R
- Google Style Guide

For this course, our main deviation from these two guides is the use of = in place of <-. (More on that later.)

2.2 Basic Calculations

To get started, we'll use R like a simple calculator.

Addition, Subtraction, Multiplication and Division

Math	R	Result
$3+2$	`3 + 2`	5
$3-2$	`3 - 2`	1
$3 \cdot 2$	`3 * 2`	6
$3/2$	`3 / 2`	1.5

Exponents

Math	R	Result
3^2	`3 ^ 2`	9
$2^{(-3)}$	`2 ^ (-3)`	0.125
$100^{1/2}$	`100 ^ (1 / 2)`	10
$\sqrt{100}$	`sqrt(100)`	10

Mathematical Constants

Math	R	Result
π	`pi`	3.1415927
e	`exp(1)`	2.7182818

Logarithms

Note that we will use ln and log interchangeably to mean the natural logarithm. There is no `ln()` in R, instead it uses `log()` to mean the natural logarithm.

Math	R	Result
$\log(e)$	`log(exp(1))`	1
$\log_{10}(1000)$	`log10(1000)`	3
$\log_2(8)$	`log2(8)`	3
$\log_4(16)$	`log(16, base = 4)`	2

Trigonometry

Math	R	Result
$\sin(\pi/2)$	`sin(pi / 2)`	1

Math	R	Result
cos(0)	`cos(0)`	1

2.3 Getting Help

In using R as a calculator, we have seen a number of functions: `sqrt()`, `exp()`, `log()` and `sin()`. To get documentation about a function in R, simply put a question mark in front of the function name, or call the function `help(function)` and RStudio will display the documentation, for example:

```
?log
?sin
?paste
?lm
help(lm)   # help() is equivalent
help(ggplot,package="ggplot2")  # show help from a certain package
```

Frequently one of the most difficult things to do when learning R is asking for help. First, you need to decide to ask for help, then you need to know *how* to ask for help. Your very first line of defense should be to Google your error message or a short description of your issue. (The ability to solve problems using this method is quickly becoming an extremely valuable skill.) If that fails, and it eventually will, you should ask for help. There are a number of things you should include when emailing an instructor, or posting to a help website such as Stack Overflow.

- Describe what you expect the code to do.
- State the end goal you are trying to achieve. (Sometimes what you expect the code to do, is not what you want to actually do.)
- Provide the full text of any errors you have received.
- Provide enough code to recreate the error. Often for the purpose of this course, you could simply email your entire `.R` or `.Rmd` file.
- Sometimes it is also helpful to include a screenshot of your entire RStudio window when the error occurs.

If you follow these steps, you will get your issue resolved much quicker, and possibly learn more in the process. Do not be discouraged by running into errors and difficulties when learning R. (Or any technical skill.) It is simply part of the learning process.

2.4 Installing Packages

R comes with a number of built-in functions and datasets, but one of the main strengths of R as an open-source project is its package system. Packages add additional functions and data. Frequently if you want to do something in R, and it is not available by default, there is a good chance that there is a package that will fulfill your needs.

To install a package, use the `install.packages()` function. Think of this as buying a recipe book from the store, bringing it home, and putting it on your shelf.

```
install.packages("ggplot2")
```

Once a package is installed, it must be loaded into your current R session before being used. Think of this as taking the book off of the shelf and opening it up to read.

```
library(ggplot2)
```

Once you close R, all the packages are closed and put back on the imaginary shelf. The next time you open R, you do not have to install the package again, but you do have to load any packages you intend to use by invoking `library()`.

2.5 Data Types

R has a number of basic data *types*.

- Numeric
  - Also known as Double. The default type when dealing with numbers.
  - Examples: `1`, `1.0`, `42.5`
- Integer
  - Examples: `1L`, `2L`, `42L`
- Complex
  - Example: `4 + 2i`
- Logical
  - Two possible values: `TRUE` and `FALSE`
  - You can also use `T` and `F`, but this is *not* recommended.
  - `NA` is also considered logical.
- Character
  - Examples: `"a"`, `"Statistics"`, `"1 plus 2."`

2.6 Data Structures

R also has a number of basic data *structures*. A data structure is either homogeneous (all elements are of the same data type) or heterogeneous (elements can be of more than one data type).

Dimension	**Homogeneous**	**Heterogeneous**
1	Vector	List
2	Matrix	Data Frame
3+	Array	

2.6.1 Vectors

Many operations in R make heavy use of **vectors**. Vectors in R are indexed starting at `1`. That is what the `[1]` in the output is indicating, that the first element of the row being displayed is the first element of the vector. Larger vectors will start additional rows with `[*]` where `*` is the index of the first element of the row.

Possibly the most common way to create a vector in R is using the `c()` function, which is short for "combine."" As the name suggests, it combines a list of elements separated by commas.

```
c(1, 3, 5, 7, 8, 9)
```

```
## [1] 1 3 5 7 8 9
```

Here R simply outputs this vector. If we would like to store this vector in a **variable** we can do so with the **assignment** operator `=`. In this case the variable `x` now holds the vector we just created, and we can access the vector by typing `x`.

```
x = c(1, 3, 5, 7, 8, 9)
x
```

```
## [1] 1 3 5 7 8 9
```

As an aside, there is a long history of the assignment operator in R, partially due to the keys available on the keyboards of the creators of the S language. (Which preceded R.) For simplicity we will use =, but know that often you will see <- as the assignment operator.

The pros and cons of these two are well beyond the scope of this book, but know that for our purposes you will have no issue if you simply use =. If you are interested in the weird cases where the difference matters, check out The R Inferno.

If you wish to use <-, you will still need to use =, however only for argument passing. Some users like to keep assignment (<-) and argument passing (=) separate. No matter what you choose, the more important thing is that you **stay consistent**. Also, if working on a larger collaborative project, you should use whatever style is already in place.

Because vectors must contains elements that are all the same type, R will automatically coerce to a single type when attempting to create a vector that combines multiple types.

```
c(42, "Statistics", TRUE)
```

```
## [1] "42"         "Statistics" "TRUE"
```

```
c(42, TRUE)
```

```
## [1] 42  1
```

Frequently you may wish to create a vector based on a sequence of numbers. The quickest and easiest way to do this is with the : operator, which creates a sequence of integers between two specified integers.

```
(y = 1:100)
```

```
##   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
##  [18]  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34
##  [35]  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51
##  [52]  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68
##  [69]  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85
##  [86]  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
```

Here we see R labeling the rows after the first since this is a large vector. Also, we see that by putting parentheses around the assignment, R both stores the vector in a variable called y and automatically outputs y to the console.

Note that scalars do not exists in R. They are simply vectors of length 1.

```
2
```

```
## [1] 2
```

If we want to create a sequence that isn't limited to integers and increasing by 1 at a time, we can use the seq() function.

```
seq(from = 1.5, to = 4.2, by = 0.1)
```

```
##  [1] 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3.0 3.1
## [18] 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9 4.0 4.1 4.2
```

We will discuss functions in detail later, but note here that the input labels from, to, and by are optional.

```
seq(1.5, 4.2, 0.1)
```

```
##  [1] 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3.0 3.1
## [18] 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9 4.0 4.1 4.2
```

Another common operation to create a vector is `rep()`, which can repeat a single value a number of times.

```
rep("A", times = 10)
```

```
##  [1] "A" "A" "A" "A" "A" "A" "A" "A" "A" "A"
```

The `rep()` function can be used to repeat a vector some number of times.

```
rep(x, times = 3)
```

```
##  [1] 1 3 5 7 8 9 1 3 5 7 8 9 1 3 5 7 8 9
```

We have now seen four different ways to create vectors:

- `c()`
- `:`
- `seq()`
- `rep()`

So far we have mostly used them in isolation, but they are often used together.

```
c(x, rep(seq(1, 9, 2), 3), c(1, 2, 3), 42, 2:4)
```

```
##  [1]  1  3  5  7  8  9  1  3  5  7  9  1  3  5  7  9  1  3  5  7  9  1  2
## [24]  3 42  2  3  4
```

The length of a vector can be obtained with the `length()` function.

```
length(x)
```

```
## [1] 6
```

```
length(y)
```

```
## [1] 100
```

2.6.1.1 Task

Let's try this out!

1. Create a vector of five ones, i.e. `[1,1,1,1,1] rep(1,5)`
2. Notice that the colon operator `a:b` is just short for *construct a sequence* ***from*** `a` ***to*** `b`. Create a vector the counts down from 10 to 0, i.e. it looks like `10,9,8,7,6,5,4,3,2,1,0`! `10:0`
3. the `rep` function takes additional arguments `times` (as above), and `each`, which tells you how often *each element* should be repeated (as opposed to the entire input vector). Use `rep` to create a vector that looks like this: `1 1 1 2 2 2 3 3 3 1 1 1 2 2 2 3 3 3` `rep(1:3,times=2,each=3)`

2.6.1.2 Subsetting

To subset a vector, i.e. to choose only some elements of it, we use square brackets, `[]`.

```
x
```

```
## [1] 1 3 5 7 8 9
```

```
x[1]
```

```
## [1] 1
```

```
x[3]
```

```
## [1] 5
```

We see that `x[1]` returns the first element, and `x[3]` returns the third element.

```
x[-2]
```

```
## [1] 1 5 7 8 9
```

We can also exclude certain indexes, in this case the second element.

```
x[1:3]
```

```
## [1] 1 3 5
```

```
x[c(1,3,4)]
```

```
## [1] 1 5 7
```

Lastly we see that we can subset based on a vector of indices.

All of the above are subsetting a vector using a vector of indexes. (Remember a single number is still a vector.) We could instead use a vector of logical values.

```
z = c(TRUE, TRUE, FALSE, TRUE, TRUE, FALSE)
z
```

```
## [1]  TRUE  TRUE FALSE  TRUE  TRUE FALSE
```

```
x[z]
```

```
## [1] 1 3 7 8
```

2.6.2 Vectorization

One of the biggest strengths of `R` is its use of vectorized operations. This means, operations which work on - and are optimized for - entire vectors.

```
x = 1:10  # a vector
x + 1     # add scalar
```

```
##  [1]  2  3  4  5  6  7  8  9 10 11
```

```
2 * x     # multiply all elts by 2
```

```
##  [1]  2  4  6  8 10 12 14 16 18 20
```

```
2 ^ x     # take 2 to the x as exponents
```

```
##  [1]    2    4    8   16   32   64  128  256  512 1024
```

```
sqrt(x)   # compute the square root of all elements in x
```

```
##  [1] 1.000000 1.414214 1.732051 2.000000 2.236068 2.449490 2.645751
##  [8] 2.828427 3.000000 3.162278
```

```
log(x)
```

```
##  [1] 0.0000000 0.6931472 1.0986123 1.3862944 1.6094379 1.7917595 1.9459101
##  [8] 2.0794415 2.1972246 2.3025851
```

We see that when a function like `log()` is called on a vector `x`, a vector is returned which has applied the function to each element of the vector `x`.

2.6.3 Logical Operators

Operator	Summary	Example	Result
`x < y`	x less than y	`3 < 42`	TRUE
`x > y`	x greater than y	`3 > 42`	FALSE
`x <= y`	x less than or equal to y	`3 <= 42`	TRUE
`x >= y`	x greater than or equal to y	`3 >= 42`	FALSE
`x == y`	xequal to y	`3 == 42`	FALSE
`x != y`	x not equal to y	`3 != 42`	TRUE
`!x`	not x	`!(3 > 42)`	TRUE
`x | y`	x or y	`(3 > 42) | TRUE`	TRUE
`x & y`	x and y	`(3 < 4) & ( 42 > 13)`	TRUE

In R, logical operators are vectorized.

```
x = c(1, 3, 5, 7, 8, 9)
```

```
x > 3
```

```
## [1] FALSE FALSE  TRUE  TRUE  TRUE  TRUE
```

```
x < 3
```

```
## [1]  TRUE FALSE FALSE FALSE FALSE FALSE
```

```
x == 3
```

```
## [1] FALSE  TRUE FALSE FALSE FALSE FALSE
```

```
x != 3
```

```
## [1]  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
```

```
x == 3 & x != 3
```

```
## [1] FALSE FALSE FALSE FALSE FALSE FALSE
```

```
x == 3 | x != 3
```

```
## [1] TRUE TRUE TRUE TRUE TRUE TRUE
```

This is extremely useful for subsetting.

```
x[x > 3]
```

```
## [1] 5 7 8 9
```

```
x[x != 3]
```

```
## [1] 1 5 7 8 9
```

- coercion

```
sum(x > 3)
```

```
## [1] 4
```

```
as.numeric(x > 3)
```

```
## [1] 0 0 1 1 1 1
```

Here we see that using the `sum()` function on a vector of logical `TRUE` and `FALSE` values that is the result of `x > 3` results in a numeric result. `R` is first automatically coercing the logical to numeric where `TRUE` is `1` and `FALSE` is `0`. This coercion from logical to numeric happens for most mathematical operations.

```
which(x > 3)
```

```
## [1] 3 4 5 6
```

```
x[which(x > 3)]
```

```
## [1] 5 7 8 9
```

```
max(x)
```

```
## [1] 9
```

```
which(x == max(x))
```

```
## [1] 6
```

```
which.max(x)
```

```
## [1] 6
```

2.6.4 More Vectorization

```
x = c(1, 3, 5, 7, 8, 9)
y = 1:100
```

```
x + 2
```

```
## [1]  3  5  7  9 10 11
```

```
x + rep(2, 6)
```

```
## [1]  3  5  7  9 10 11
```

```
x > 3
```

```
## [1] FALSE FALSE  TRUE  TRUE  TRUE  TRUE
```

```
x > rep(3, 6)
```

```
## [1] FALSE FALSE  TRUE  TRUE  TRUE  TRUE
```

```
x + y
```

```
## Warning in x + y: longer object length is not a multiple of shorter object
## length
```

```
##   [1]   2   5   8  11  13  15   8  11  14  17  19  21  14  17  20  23  25
##  [18]  27  20  23  26  29  31  33  26  29  32  35  37  39  32  35  38  41
##  [35]  43  45  38  41  44  47  49  51  44  47  50  53  55  57  50  53  56
##  [52]  59  61  63  56  59  62  65  67  69  62  65  68  71  73  75  68  71
##  [69]  74  77  79  81  74  77  80  83  85  87  80  83  86  89  91  93  86
##  [86]  89  92  95  97  99  92  95  98 101 103 105  98 101 104 107
```

```
length(x)
```

```
## [1] 6
```

```
length(y)
```

```
## [1] 100
```

```
length(y) / length(x)
```

```
## [1] 16.66667
```

```
(x + y) - y
```

```
## Warning in x + y: longer object length is not a multiple of shorter object
## length
```

```
##   [1] 1 3 5 7 8 9 1 3 5 7 8 9 1 3 5 7 8 9 1 3 5 7 8 9 1 3 5 7 8 9 1 3 5 7 8
##  [36] 9 1 3 5 7 8 9 1 3 5 7 8 9 1 3 5 7 8 9 1 3 5 7 8 9 1 3 5 7 8 9 1 3 5 7
##  [71] 8 9 1 3 5 7 8 9 1 3 5 7 8 9 1 3 5 7 8 9 1 3 5 7 8 9 1 3 5 7
```

```
y = 1:60
x + y
```

```
##  [1]  2  5  8 11 13 15  8 11 14 17 19 21 14 17 20 23 25 27 20 23 26 29 31
## [24] 33 26 29 32 35 37 39 32 35 38 41 43 45 38 41 44 47 49 51 44 47 50 53
## [47] 55 57 50 53 56 59 61 63 56 59 62 65 67 69
```

```
length(y) / length(x)
```

```
## [1] 10
```

```
rep(x, 10) + y
```

```
##  [1]  2  5  8 11 13 15  8 11 14 17 19 21 14 17 20 23 25 27 20 23 26 29 31
## [24] 33 26 29 32 35 37 39 32 35 38 41 43 45 38 41 44 47 49 51 44 47 50 53
## [47] 55 57 50 53 56 59 61 63 56 59 62 65 67 69
```

```
all(x + y == rep(x, 10) + y)
```

```
## [1] TRUE
```

```
identical(x + y, rep(x, 10) + y)
```

```
## [1] TRUE
```

```
# ?any
# ?all.equal
```

2.6.5 Matrices

R can also be used for **matrix** calculations. Matrices have rows and columns containing a single data type. In a matrix, the order of rows and columns is important. (This is not true of *data frames*, which we will see later.)

Matrices can be created using the `matrix` function.

```
x = 1:9
x
```

```
## [1] 1 2 3 4 5 6 7 8 9
```

```
X = matrix(x, nrow = 3, ncol = 3)
X
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    7
```

```
## [2,]    2    5    8
## [3,]    3    6    9
```

Note here that we are using two different variables: lower case `x`, which stores a vector and capital `X`, which stores a matrix. (Following the usual mathematical convention.) We can do this because `R` is case sensitive.

By default the `matrix` function reorders a vector into columns, but we can also tell `R` to use rows instead.

```
Y = matrix(x, nrow = 3, ncol = 3, byrow = TRUE)
Y
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
## [3,]    7    8    9
```

We can also create a matrix of a specified dimension where every element is the same, in this case `0`.

```
Z = matrix(0, 2, 4)
Z
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    0    0    0    0
## [2,]    0    0    0    0
```

Like vectors, matrices can be subsetted using square brackets, `[]`. However, since matrices are two-dimensional, we need to specify both a row and a column when subsetting.

```
X
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    7
## [2,]    2    5    8
## [3,]    3    6    9
```

```
X[1, 2]
```

```
## [1] 4
```

Here we accessed the element in the first row and the second column. We could also subset an entire row or column.

```
X[1, ]
```

```
## [1] 1 4 7
```

```
X[, 2]
```

```
## [1] 4 5 6
```

We can also use vectors to subset more than one row or column at a time. Here we subset to the first and third column of the second row.

```
X[2, c(1, 3)]
```

```
## [1] 2 8
```

Matrices can also be created by combining vectors as columns, using `cbind`, or combining vectors as rows, using `rbind`.

```
x = 1:9
rev(x)
```

```
## [1] 9 8 7 6 5 4 3 2 1
```

```
rep(1, 9)
```

```
## [1] 1 1 1 1 1 1 1 1 1
```

```
rbind(x, rev(x), rep(1, 9))
```

```
##   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
## x    1    2    3    4    5    6    7    8    9
##      9    8    7    6    5    4    3    2    1
##      1    1    1    1    1    1    1    1    1
```

```
cbind(col_1 = x, col_2 = rev(x), col_3 = rep(1, 9))
```

```
##       col_1 col_2 col_3
##  [1,]     1     9     1
##  [2,]     2     8     1
##  [3,]     3     7     1
##  [4,]     4     6     1
##  [5,]     5     5     1
##  [6,]     6     4     1
##  [7,]     7     3     1
##  [8,]     8     2     1
##  [9,]     9     1     1
```

When using `rbind` and `cbind` you can specify "argument" names that will be used as column names.

R can then be used to perform matrix calculations.

```
x = 1:9
y = 9:1
X = matrix(x, 3, 3)
Y = matrix(y, 3, 3)
X
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    7
## [2,]    2    5    8
## [3,]    3    6    9
```

```
Y
```

```
##      [,1] [,2] [,3]
## [1,]    9    6    3
## [2,]    8    5    2
## [3,]    7    4    1
```

```
X + Y
```

```
##      [,1] [,2] [,3]
## [1,]   10   10   10
## [2,]   10   10   10
## [3,]   10   10   10
```

```
X - Y
```

```
##      [,1] [,2] [,3]
## [1,]   -8   -2    4
## [2,]   -6    0    6
## [3,]   -4    2    8
```

```
X * Y
```

```
##      [,1] [,2] [,3]
## [1,]    9   24   21
## [2,]   16   25   16
## [3,]   21   24    9
```

```
X / Y
```

```
##           [,1]      [,2]     [,3]
## [1,] 0.1111111 0.6666667 2.333333
## [2,] 0.2500000 1.0000000 4.000000
## [3,] 0.4285714 1.5000000 9.000000
```

Note that `X * Y` is not matrix multiplication. It is element by element multiplication. (Same for `X / Y`). Instead, matrix multiplication uses `%*%`. Other matrix functions include `t()` which gives the transpose of a matrix and `solve()` which returns the inverse of a square matrix if it is invertible.

```
X %*% Y
```

```
##      [,1] [,2] [,3]
## [1,]   90   54   18
## [2,]  114   69   24
## [3,]  138   84   30
```

```
t(X)
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
## [3,]    7    8    9
```

```
Z = matrix(c(9, 2, -3, 2, 4, -2, -3, -2, 16), 3, byrow = TRUE)
Z
```

```
##      [,1] [,2] [,3]
## [1,]    9    2   -3
## [2,]    2    4   -2
## [3,]   -3   -2   16
```

```
solve(Z)
```

```
##             [,1]        [,2]       [,3]
## [1,]  0.12931034 -0.05603448 0.01724138
## [2,] -0.05603448  0.29094828 0.02586207
## [3,]  0.01724138  0.02586207 0.06896552
```

To verify that `solve(Z)` returns the inverse, we multiply it by `Z`. We would expect this to return the identity matrix, however we see that this is not the case due to some computational issues. However, `R` also has the `all.equal()` function which checks for equality, with some small tolerance which accounts for some computational issues. The `identical()` function is used to check for exact equality.

```
solve(Z) %*% Z
```

```
##              [,1]          [,2]         [,3]
## [1,] 1.000000e+00 -6.245005e-17 0.000000e+00
## [2,] 8.326673e-17  1.000000e+00 5.551115e-17
## [3,] 2.775558e-17  0.000000e+00 1.000000e+00
```

```
diag(3)
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
```

```
all.equal(solve(Z) %*% Z, diag(3))
```

```
## [1] TRUE
```

R has a number of matrix specific functions for obtaining dimension and summary information.

```
X = matrix(1:6, 2, 3)
X
```

```
##      [,1] [,2] [,3]
## [1,]    1    3    5
## [2,]    2    4    6
```

```
dim(X)
```

```
## [1] 2 3
```

```
rowSums(X)
```

```
## [1]  9 12
```

```
colSums(X)
```

```
## [1]  3  7 11
```

```
rowMeans(X)
```

```
## [1] 3 4
```

```
colMeans(X)
```

```
## [1] 1.5 3.5 5.5
```

The `diag()` function can be used in a number of ways. We can extract the diagonal of a matrix.

```
diag(Z)
```

```
## [1]  9  4 16
```

Or create a matrix with specified elements on the diagonal. (And `0` on the off-diagonals.)

```
diag(1:5)
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
## [2,]    0    2    0    0    0
## [3,]    0    0    3    0    0
## [4,]    0    0    0    4    0
## [5,]    0    0    0    0    5
```

Or, lastly, create a square matrix of a certain dimension with `1` for every element of the diagonal and `0` for the off-diagonals.

```
diag(5)
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
```

```
## [2,]    0    1    0    0    0
## [3,]    0    0    1    0    0
## [4,]    0    0    0    1    0
## [5,]    0    0    0    0    1
```

Calculations with Vectors and Matrices

Certain operations in R, for example %*% have different behavior on vectors and matrices. To illustrate this, we will first create two vectors.

```
a_vec = c(1, 2, 3)
b_vec = c(2, 2, 2)
```

Note that these are indeed vectors. They are not matrices.

```
c(is.vector(a_vec), is.vector(b_vec))
```

```
## [1] TRUE TRUE
```

```
c(is.matrix(a_vec), is.matrix(b_vec))
```

```
## [1] FALSE FALSE
```

When this is the case, the %*% operator is used to calculate the **dot product**, also know as the **inner product** of the two vectors.

The dot product of vectors $\boldsymbol{a} = [a_1, a_2, \cdots a_n]$ and $\boldsymbol{b} = [b_1, b_2, \cdots b_n]$ is defined to be

$$\boldsymbol{a} \cdot \boldsymbol{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots a_n b_n.$$

```
a_vec %*% b_vec # inner product
```

```
##      [,1]
## [1,]   12
```

```
a_vec %o% b_vec # outer product
```

```
##      [,1] [,2] [,3]
## [1,]    2    2    2
## [2,]    4    4    4
## [3,]    6    6    6
```

The %o% operator is used to calculate the **outer product** of the two vectors.

When vectors are coerced to become matrices, they are column vectors. So a vector of length n becomes an $n \times 1$ matrix after coercion.

```
as.matrix(a_vec)
```

```
##      [,1]
## [1,]    1
## [2,]    2
## [3,]    3
```

If we use the %*% operator on matrices, %*% again performs the expected matrix multiplication. So you might expect the following to produce an error, because the dimensions are incorrect.

```
as.matrix(a_vec) %*% b_vec
```

```
##      [,1] [,2] [,3]
## [1,]    2    2    2
## [2,]    4    4    4
## [3,]    6    6    6
```

At face value this is a 3×1 matrix, multiplied by a 3×1 matrix. However, when `b_vec` is automatically coerced to be a matrix, `R` decided to make it a "row vector", a 1×3 matrix, so that the multiplication has conformable dimensions.

If we had coerced both, then `R` would produce an error.

```
as.matrix(a_vec) %*% as.matrix(b_vec)
```

Another way to calculate a *dot product* is with the `crossprod()` function. Given two vectors, the `crossprod()` function calculates their dot product. The function has a rather misleading name.

```
crossprod(a_vec, b_vec)  # inner product
```

```
##      [,1]
## [1,]   12
```

```
tcrossprod(a_vec, b_vec)  # outer product
```

```
##      [,1] [,2] [,3]
## [1,]    2    2    2
## [2,]    4    4    4
## [3,]    6    6    6
```

These functions could be very useful later. When used with matrices X and Y as arguments, it calculates

$$X^\top Y.$$

When dealing with linear models, the calculation

$$X^\top X$$

is used repeatedly.

```
C_mat = matrix(c(1, 2, 3, 4, 5, 6), 2, 3)
D_mat = matrix(c(2, 2, 2, 2, 2, 2), 2, 3)
```

This is useful both as a shortcut for a frequent calculation and as a more efficient implementation than using `t()` and `%*%`.

```
crossprod(C_mat, D_mat)
```

```
##      [,1] [,2] [,3]
## [1,]    6    6    6
## [2,]   14   14   14
## [3,]   22   22   22
```

```
t(C_mat) %*% D_mat
```

```
##      [,1] [,2] [,3]
## [1,]    6    6    6
## [2,]   14   14   14
## [3,]   22   22   22
```

```
all.equal(crossprod(C_mat, D_mat), t(C_mat) %*% D_mat)
```

```
## [1] TRUE
```

```
crossprod(C_mat, C_mat)
```

```
##      [,1] [,2] [,3]
## [1,]    5   11   17
## [2,]   11   25   39
## [3,]   17   39   61
```

```
t(C_mat) %*% C_mat
```

```
##      [,1] [,2] [,3]
## [1,]    5   11   17
## [2,]   11   25   39
## [3,]   17   39   61
```

```
all.equal(crossprod(C_mat, C_mat), t(C_mat) %*% C_mat)
```

```
## [1] TRUE
```

2.6.6 Lists

A list is a one-dimensional heterogeneous data structure. So it is indexed like a vector with a single integer value, but each element can contain an element of any type. Lists are extremely useful and versatile objects, so make sure you understand their useage:

```
# creation
list(42, "Hello", TRUE)
```

```
## [[1]]
## [1] 42
##
## [[2]]
## [1] "Hello"
##
## [[3]]
## [1] TRUE
```

```
ex_list = list(
  a = c(1, 2, 3, 4),
  b = TRUE,
  c = "Hello!",
  d = function(arg = 42) {print("Hello World!")},
  e = diag(5)
)
```

Lists can be subset using two syntaxes, the `$` operator, and square brackets `[]`. The `$` operator returns a named **element** of a list. The `[]` syntax returns a **list**, while the `[[]]` returns an **element** of a list.

- `ex_list[1]` returns a list contain the first element.
- `ex_list[[1]]` returns the first element of the list, in this case, a vector.

```
# subsetting
ex_list$e
```

```
##      [,1] [,2] [,3] [,4] [,5]
```

```
## [1,]    1    0    0    0    0
## [2,]    0    1    0    0    0
## [3,]    0    0    1    0    0
## [4,]    0    0    0    1    0
## [5,]    0    0    0    0    1
```

```
ex_list[1:2]
```

```
## $a
## [1] 1 2 3 4
##
## $b
## [1] TRUE
```

```
ex_list[1]
```

```
## $a
## [1] 1 2 3 4
```

```
ex_list[[1]]
```

```
## [1] 1 2 3 4
```

```
ex_list[c("e", "a")]
```

```
## $e
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
## [2,]    0    1    0    0    0
## [3,]    0    0    1    0    0
## [4,]    0    0    0    1    0
## [5,]    0    0    0    0    1
##
## $a
## [1] 1 2 3 4
```

```
ex_list["e"]
```

```
## $e
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
## [2,]    0    1    0    0    0
## [3,]    0    0    1    0    0
## [4,]    0    0    0    1    0
## [5,]    0    0    0    0    1
```

```
ex_list[["e"]]
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
## [2,]    0    1    0    0    0
## [3,]    0    0    1    0    0
## [4,]    0    0    0    1    0
## [5,]    0    0    0    0    1
```

```
ex_list$d
```

```
## function(arg = 42) {print("Hello World!")}
```

```
ex_list$d(arg = 1)
```

```
## [1] "Hello World!"
```

2.6.7 Data Frames

We have previously seen vectors and matrices for storing data as we introduced R. We will now introduce a **data frame** which will be the most common way that we store and interact with data in this course.

```
example_data = data.frame(x = c(1, 3, 5, 7, 9, 1, 3, 5, 7, 9),
                          y = c(rep("Hello", 9), "Goodbye"),
                          z = rep(c(TRUE, FALSE), 5))
```

Unlike a matrix, which can be thought of as a vector rearranged into rows and columns, a data frame is not required to have the same data type for each element. A data frame is a **list** of vectors. So, each vector must contain the same data type, but the different vectors can store different data types.

```
example_data
```

```
##    x       y     z
## 1  1   Hello  TRUE
## 2  3   Hello FALSE
## 3  5   Hello  TRUE
## 4  7   Hello FALSE
## 5  9   Hello  TRUE
## 6  1   Hello FALSE
## 7  3   Hello  TRUE
## 8  5   Hello FALSE
## 9  7   Hello  TRUE
## 10 9 Goodbye FALSE
```

Unlike a list which has more flexibility, the elements of a data frame must all be vectors, and have the same length. Again, we access any given column with the $ operator:

```
example_data$x
```

```
##  [1] 1 3 5 7 9 1 3 5 7 9
```

```
all.equal(length(example_data$x),
          length(example_data$y),
          length(example_data$z))
```

```
## [1] TRUE
```

```
str(example_data)
```

```
## 'data.frame':    10 obs. of  3 variables:
##  $ x: num  1 3 5 7 9 1 3 5 7 9
##  $ y: Factor w/ 2 levels "Goodbye","Hello": 2 2 2 2 2 2 2 2 2 1
##  $ z: logi  TRUE FALSE TRUE FALSE TRUE FALSE ...
```

```
nrow(example_data)
```

```
## [1] 10
```

```
ncol(example_data)
```

```
## [1] 3
```

```
dim(example_data)
```

```
## [1] 10  3
```

The `data.frame()` function above is one way to create a data frame. We can also import data from various file types in into `R`, as well as use data stored in packages.

The example data above can also be found here as a .csv file. To read this data into `R`, we would use the `read_csv()` function from the `readr` package. Note that `R` has a built in function `read.csv()` that operates very similarly. The `readr` function `read_csv()` has a number of advantages. For example, it is much faster reading larger data. It also uses the `tibble` package to read the data as a tibble.

```
library(readr)
example_data_from_csv = read_csv("data/example-data.csv")
```

This particular line of code assumes that the file `example_data.csv` exists in a folder called `data` in your current working directory.

```
example_data_from_csv
```

```
## # A tibble: 10 x 3
##        x y       z
##    <int> <chr>   <lgl>
##  1     1 Hello   TRUE
##  2     3 Hello   FALSE
##  3     5 Hello   TRUE
##  4     7 Hello   FALSE
##  5     9 Hello   TRUE
##  6     1 Hello   FALSE
##  7     3 Hello   TRUE
##  8     5 Hello   FALSE
##  9     7 Hello   TRUE
## 10     9 Goodbye FALSE
```

A tibble is simply a data frame that prints with sanity. Notice in the output above that we are given additional information such as dimension and variable type.

The `as_tibble()` function can be used to coerce a regular data frame to a tibble.

```
library(tibble)
example_data = as_tibble(example_data)
example_data
```

```
## # A tibble: 10 x 3
##        x y       z
##    <dbl> <fct>   <lgl>
##  1    1. Hello   TRUE
##  2    3. Hello   FALSE
##  3    5. Hello   TRUE
##  4    7. Hello   FALSE
##  5    9. Hello   TRUE
##  6    1. Hello   FALSE
##  7    3. Hello   TRUE
##  8    5. Hello   FALSE
##  9    7. Hello   TRUE
## 10    9. Goodbye FALSE
```

Alternatively, we could use the "Import Dataset" feature in RStudio which can be found in the environment

window. (By default, the top-right pane of RStudio.) Once completed, this process will automatically generate the code to import a file. The resulting code will be shown in the console window. In recent versions of RStudio, `read_csv()` is used by default, thus reading in a tibble.

Earlier we looked at installing packages, in particular the `ggplot2` package. (A package for visualization. While not necessary for this course, it is quickly growing in popularity.)

```
library(ggplot2)
```

Inside the `ggplot2` package is a dataset called `mpg`. By loading the package using the `library()` function, we can now access `mpg`.

When using data from inside a package, there are three things we would generally like to do:

- Look at the raw data.
- Understand the data. (Where did it come from? What are the variables? Etc.)
- Visualize the data.

To look at the data, we have two useful commands: `head()` and `str()`.

```
head(mpg, n = 10)
```

```
## # A tibble: 10 x 11
##    manufacturer model    displ  year   cyl trans   drv     cty   hwy fl
##    <chr>        <chr>    <dbl> <int> <int> <chr>   <chr> <int> <int> <chr>
##  1 audi         a4         1.80  1999     4 auto(l~ f        18    29 p
##  2 audi         a4         1.80  1999     4 manual~ f        21    29 p
##  3 audi         a4         2.00  2008     4 manual~ f        20    31 p
##  4 audi         a4         2.00  2008     4 auto(a~ f        21    30 p
##  5 audi         a4         2.80  1999     6 auto(l~ f        16    26 p
##  6 audi         a4         2.80  1999     6 manual~ f        18    26 p
##  7 audi         a4         3.10  2008     6 auto(a~ f        18    27 p
##  8 audi         a4 quat~   1.80  1999     4 manual~ 4        18    26 p
##  9 audi         a4 quat~   1.80  1999     4 auto(l~ 4        16    25 p
## 10 audi         a4 quat~   2.00  2008     4 manual~ 4        20    28 p
## # ... with 1 more variable: class <chr>
```

The function `head()` will display the first `n` observations of the data frame. The `head()` function was more useful before tibbles. Notice that `mpg` is a tibble already, so the output from `head()` indicates there are only `10` observations. Note that this applies to `head(mpg, n = 10)` and not `mpg` itself. Also note that tibbles print a limited number of rows and columns by default. The last line of the printed output indicates with rows and columns were omitted.

```
mpg
```

```
## # A tibble: 234 x 11
##    manufacturer model    displ  year   cyl trans   drv     cty   hwy fl
##    <chr>        <chr>    <dbl> <int> <int> <chr>   <chr> <int> <int> <chr>
##  1 audi         a4         1.80  1999     4 auto(l~ f        18    29 p
##  2 audi         a4         1.80  1999     4 manual~ f        21    29 p
##  3 audi         a4         2.00  2008     4 manual~ f        20    31 p
##  4 audi         a4         2.00  2008     4 auto(a~ f        21    30 p
##  5 audi         a4         2.80  1999     6 auto(l~ f        16    26 p
##  6 audi         a4         2.80  1999     6 manual~ f        18    26 p
##  7 audi         a4         3.10  2008     6 auto(a~ f        18    27 p
##  8 audi         a4 quat~   1.80  1999     4 manual~ 4        18    26 p
##  9 audi         a4 quat~   1.80  1999     4 auto(l~ 4        16    25 p
## 10 audi         a4 quat~   2.00  2008     4 manual~ 4        20    28 p
## # ... with 224 more rows, and 1 more variable: class <chr>
```

The function `str()` will display the "structure" of the data frame. It will display the number of **observations** and **variables**, list the variables, give the type of each variable, and show some elements of each variable. This information can also be found in the "Environment" window in RStudio.

```
str(mpg)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    234 obs. of  11 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

It is important to note that while matrices have rows and columns, data frames (tibbles) instead have observations and variables. When displayed in the console or viewer, each row is an observation and each column is a variable. However generally speaking, their order does not matter, it is simply a side-effect of how the data was entered or stored.

In this dataset an observation is for a particular model-year of a car, and the variables describe attributes of the car, for example its highway fuel efficiency.

To understand more about the data set, we use the `?` operator to pull up the documentation for the data.

```
?mpg
```

R has a number of functions for quickly working with and extracting basic information from data frames. To quickly obtain a vector of the variable names, we use the `names()` function.

```
names(mpg)
```

```
##  [1] "manufacturer" "model"        "displ"        "year"
##  [5] "cyl"          "trans"        "drv"          "cty"
##  [9] "hwy"          "fl"           "class"
```

To access one of the variables **as a vector**, we use the `$` operator.

```
mpg$year
```

```
##   [1] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 2008 1999 1999 2008
##  [15] 2008 1999 2008 2008 2008 2008 2008 1999 2008 1999 1999 2008 2008 2008
##  [29] 2008 2008 1999 1999 1999 2008 1999 2008 2008 1999 1999 1999 1999 2008
##  [43] 2008 2008 1999 1999 2008 2008 2008 2008 1999 1999 2008 2008 2008 1999
##  [57] 1999 1999 2008 2008 2008 1999 2008 1999 2008 2008 2008 2008 2008 2008
##  [71] 1999 1999 2008 1999 1999 1999 2008 1999 1999 1999 2008 2008 1999 1999
##  [85] 1999 1999 1999 2008 1999 2008 1999 1999 2008 2008 1999 1999 2008 2008
##  [99] 2008 1999 1999 1999 1999 1999 2008 2008 2008 2008 1999 1999 2008 2008
## [113] 1999 1999 2008 1999 1999 2008 2008 2008 2008 2008 2008 2008 1999 1999
## [127] 2008 2008 2008 2008 1999 2008 2008 1999 1999 1999 2008 1999 2008 2008
## [141] 1999 1999 1999 2008 2008 2008 2008 1999 1999 2008 1999 1999 2008 2008
## [155] 1999 1999 1999 2008 2008 1999 1999 2008 2008 2008 2008 1999 1999 1999
## [169] 1999 2008 2008 2008 2008 1999 1999 1999 1999 2008 2008 1999 1999 2008
## [183] 2008 1999 1999 2008 1999 1999 2008 2008 1999 1999 2008 1999 1999 1999
## [197] 2008 2008 1999 2008 1999 1999 2008 1999 1999 2008 2008 1999 1999 2008
```

```
## [211] 2008 1999 1999 1999 1999 2008 2008 2008 2008 1999 1999 1999 1999 1999
## [225] 1999 2008 2008 1999 1999 2008 2008 1999 1999 2008
```

```
mpg$hwy
```

```
##   [1] 29 29 31 30 26 26 27 26 25 28 27 25 25 25 25 24 25 23 20 15 20 17 17
##  [24] 26 23 26 25 24 19 14 15 17 27 30 26 29 26 24 24 22 22 24 24 17 22 21
##  [47] 23 23 19 18 17 17 19 19 12 17 15 17 17 12 17 16 18 15 16 12 17 17 16
##  [70] 12 15 16 17 15 17 17 18 17 19 17 19 19 17 17 17 16 16 17 15 17 26 25
##  [93] 26 24 21 22 23 22 20 33 32 32 29 32 34 36 36 29 26 27 30 31 26 26 28
## [116] 26 29 28 27 24 24 24 22 19 20 17 12 19 18 14 15 18 18 15 17 16 18 17
## [139] 19 19 17 29 27 31 32 27 26 26 25 25 17 17 20 18 26 26 27 28 25 25 24
## [162] 27 25 26 23 26 26 26 26 25 27 25 27 20 20 19 17 20 17 29 27 31 31 26
## [185] 26 28 27 29 31 31 26 26 27 30 33 35 37 35 15 18 20 20 22 17 19 18 20
## [208] 29 26 29 29 24 44 29 26 29 29 29 29 23 24 44 41 29 26 28 29 29 29 28
## [231] 29 26 26 26
```

We can use the `dim()`, `nrow()` and `ncol()` functions to obtain information about the dimension of the data frame.

```
dim(mpg)
```

```
## [1] 234  11
```

```
nrow(mpg)
```

```
## [1] 234
```

```
ncol(mpg)
```

```
## [1] 11
```

Here `nrow()` is also the number of observations, which in most cases is the *sample size*.

Subsetting data frames can work much like subsetting matrices using square brackets, `[,]`. Here, we find fuel efficient vehicles earning over 35 miles per gallon and only display `manufacturer`, `model` and `year`.

```
mpg[mpg$hwy > 35, c("manufacturer", "model", "year")]
```

```
## # A tibble: 6 x 3
##   manufacturer model       year
##   <chr>        <chr>      <int>
## 1 honda        civic       2008
## 2 honda        civic       2008
## 3 toyota       corolla     2008
## 4 volkswagen   jetta       1999
## 5 volkswagen   new beetle  1999
## 6 volkswagen   new beetle  1999
```

An alternative would be to use the `subset()` function, which has a much more readable syntax.

```
subset(mpg, subset = hwy > 35, select = c("manufacturer", "model", "year"))
```

Lastly, we could use the `filter` and `select` functions from the `dplyr` package which introduces the *pipe operator* `%>%` from the `magrittr` package. A *pipe* is a concept from the Unix world, where it means to take the output of some command, and pass it on to another command. This way, one can construct a *pipeline* of commands. We will see more of this in chapter 3. For additional info on the pipe operator in R, you might be interested in this tutorial.

```
library(dplyr)
mpg %>%
```

```
  filter(hwy > 35) %>%
  select(manufacturer, model, year)
```

```
## # A tibble: 6 x 3
##   manufacturer model       year
##   <chr>        <chr>      <int>
## 1 honda        civic       2008
## 2 honda        civic       2008
## 3 toyota       corolla     2008
## 4 volkswagen   jetta       1999
## 5 volkswagen   new beetle  1999
## 6 volkswagen   new beetle  1999
```

All three approaches produce the same results. Which you use will be largely based on a given situation as well as user preference.

When subsetting a data frame, be aware of what is being returned, as sometimes it may be a vector instead of a data frame. Also note that there are differences between subsetting a data frame and a tibble. A data frame operates more like a matrix where it is possible to reduce the subset to a vector. A tibble operates more like a list where it always subsets to another tibble.

2.7 Programming Basics

2.7.1 Control Flow

In R, the if/else syntax is:

```
if (...) {
  some R code
} else {
  more R code
}
```

For example,

```
x = 1
y = 3
if (x > y) {
  z = x * y
  print("x is larger than y")
} else {
  z = x + 5 * y
  print("x is less than or equal to y")
}
```

```
## [1] "x is less than or equal to y"
```

```
z
```

```
## [1] 16
```

R also has a special function `ifelse()` which is very useful. It returns one of two specified values based on a conditional statement.

```
ifelse(4 > 3, 1, 0)
```

```
## [1] 1
```

The real power of `ifelse()` comes from its ability to be applied to vectors.

```
fib = c(1, 1, 2, 3, 5, 8, 13, 21)
ifelse(fib > 6, "Foo", "Bar")
```

```
## [1] "Bar" "Bar" "Bar" "Bar" "Bar" "Foo" "Foo" "Foo"
```

Now a `for` loop example,

```
x = 11:15
for (i in 1:5) {
  x[i] = x[i] * 2
}

x
```

```
## [1] 22 24 26 28 30
```

Note that this `for` loop is very normal in many programming languages, but not in `R`. In `R` we would not use a loop, instead we would simply use a vectorized operation.

```
x = 11:15
x = x * 2
x
```

```
## [1] 22 24 26 28 30
```

2.7.2 Functions

So far we have been using functions, but haven't actually discussed some of their details.

```
function_name(arg1 = 10, arg2 = 20)
```

To use a function, you simply type its name, followed by an open parenthesis, then specify values of its arguments, then finish with a closing parenthesis.

An **argument** is a variable which is used in the body of the function. Specifying the values of the arguments is essentially providing the inputs to the function.

We can also write our own functions in `R`. For example, we often like to "standardize" variables, that is, subtracting the sample mean, and dividing by the sample standard deviation.

$$\frac{x - \bar{x}}{s}$$

In `R` we would write a function to do this. When writing a function, there are three thing you must do.

- Give the function a name. Preferably something that is short, but descriptive.
- Specify the arguments using `function()`
- Write the body of the function within curly braces, `{}`.

```
standardize = function(x) {
  m = mean(x)
  std = sd(x)
  result = (x - m) / std
  result
}
```

Here the name of the function is `standardize`, and the function has a single argument `x` which is used in the body of function. Note that the output of the final line of the body is what is returned by the function. In this case the function returns the vector stored in the variable `results`.

To test our function, we will take a random sample of size `n = 10` from a normal distribution with a mean of `2` and a standard deviation of `5`.

```
(test_sample = rnorm(n = 10, mean = 2, sd = 5))
```

```
##  [1]   2.7665009  -0.5246944   7.6859092   2.5269809  10.7340675
##  [6] -15.5489238  -0.5346690   4.3940025   3.6148929  11.8296569
```

```
standardize(x = test_sample)
```

```
##  [1]  0.009395804 -0.419330954  0.650221334 -0.021805202  1.047288927
##  [6] -2.376458535 -0.420630289  0.221401905  0.119911391  1.190005617
```

This function could be written much more succinctly, simply performing all the operations on one line and immediately returning the result, without storing any of the intermediate results.

```
standardize = function(x) {
  (x - mean(x)) / sd(x)
}
```

When specifying arguments, you can provide default arguments.

```
power_of_num = function(num, power = 2) {
  num ^ power
}
```

Let's look at a number of ways that we could run this function to perform the operation `10^2` resulting in `100`.

```
power_of_num(10)
```

```
## [1] 100
```

```
power_of_num(10, 2)
```

```
## [1] 100
```

```
power_of_num(num = 10, power = 2)
```

```
## [1] 100
```

```
power_of_num(power = 2, num = 10)
```

```
## [1] 100
```

Note that without using the argument names, the order matters. The following code will not evaluate to the same output as the previous example.

```
power_of_num(2, 10)
```

```
## [1] 1024
```

Also, the following line of code would produce an error since arguments without a default value must be specified.

```
power_of_num(power = 5)
```

To further illustrate a function with a default argument, we will write a function that calculates sample variance two ways.

By default, the function will calculate the unbiased estimate of σ^2, which we will call s^2.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x - \bar{x})^2$$

It will also have the ability to return the biased estimate (based on maximum likelihood) which we will call $\hat{\sigma}^2$.

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x - \bar{x})^2$$

```
get_var = function(x, biased = FALSE) {
  n = length(x) - 1 * !biased
  (1 / n) * sum((x - mean(x)) ^ 2)
}
```

```
get_var(test_sample)
```

```
## [1] 58.93132
```

```
get_var(test_sample, biased = FALSE)
```

```
## [1] 58.93132
```

```
var(test_sample)
```

```
## [1] 58.93132
```

We see the function is working as expected, and when returning the unbiased estimate it matches R's built in function `var()`. Finally, let's examine the biased estimate of σ^2.

```
get_var(test_sample, biased = TRUE)
```

```
## [1] 53.03818
```

Chapter 3

Working With Data

In this chapter we will first learn some basic concepts that help summarizing data. Then, we will tackle a real-world task and read, clean, and summarize data from the web.

3.1 Summary Statistics

`R` has built in functions for a large number of summary statistics. For numeric variables, we can summarize data with the center and spread. Make sure to have loaded the `ggplot2` library to be able to access the `mpg` dataset as introduced in section 2.6.7.

Central Tendency

Measure	`R`	Result
Mean	`mean(mpg$cty)`	16.8589744
Median	`median(mpg$cty)`	17

Spread

Measure	`R`	Result
Variance	`var(mpg$cty)`	18.1130736
Standard Deviation	`sd(mpg$cty)`	4.2559457
IQR	`IQR(mpg$cty)`	5
Minimum	`min(mpg$cty)`	9
Maximum	`max(mpg$cty)`	35
Range	`range(mpg$cty)`	9, 35

Categorical

For categorical variables, counts and percentages can be used for summary.

```
table(mpg$drv)
```

```
##
##   4   f   r
## 103 106  25
```

```
table(mpg$drv) / nrow(mpg)
```

```
##
##         4         f         r
## 0.4401709 0.4529915 0.1068376
```

3.2 Plotting

Now that we have some data to work with, and we have learned about the data at the most basic level, our next tasks is to visualize the data. Often, a proper visualization can illuminate features of the data that can inform further analysis.

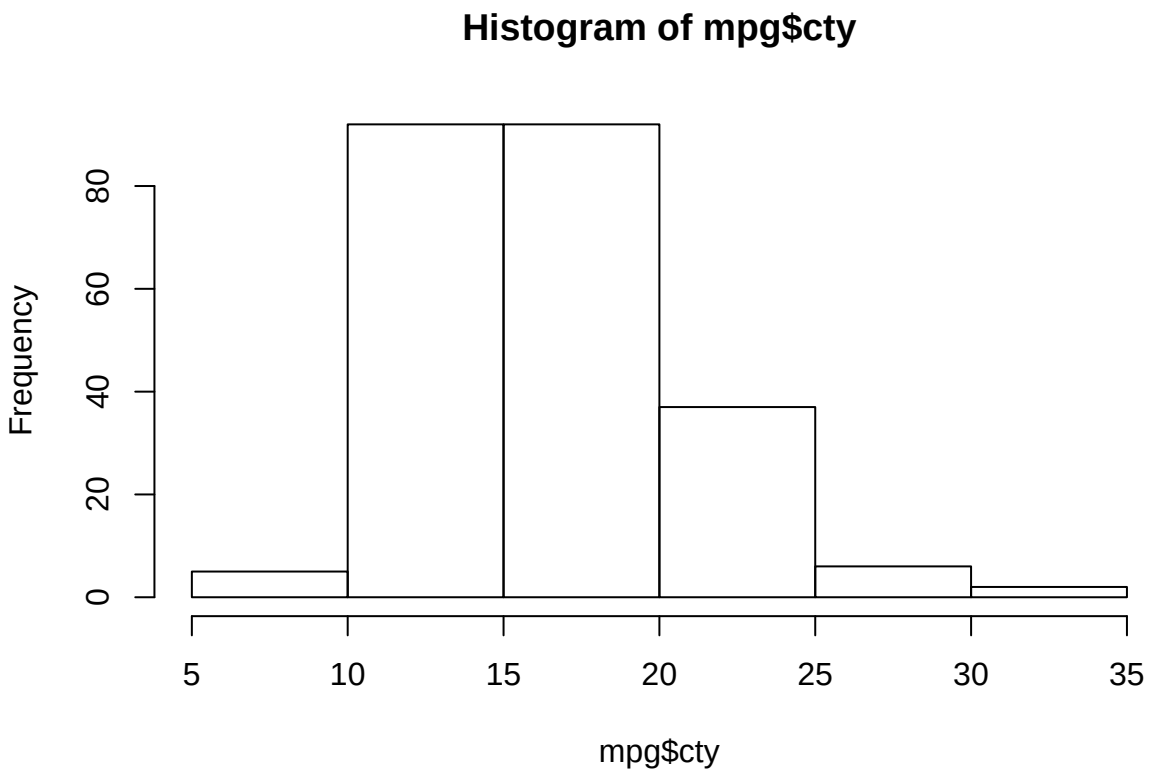
We will look at four methods of visualizing data that we will use throughout the course:

- Histograms
- Barplots
- Boxplots
- Scatterplots

3.2.1 Histograms

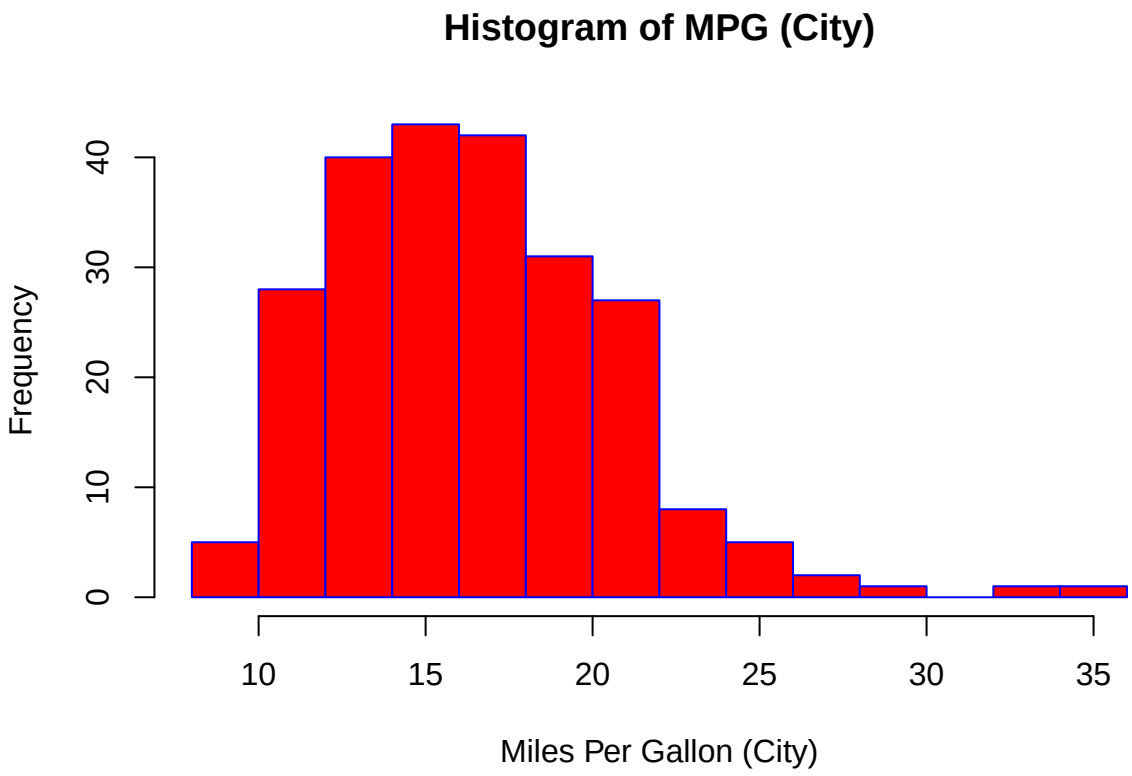
When visualizing a single numerical variable, a **histogram** will be our go-to tool, which can be created in `R` using the `hist()` function.

```
hist(mpg$cty)
```



The histogram function has a number of parameters which can be changed to make our plot look much nicer. Use the `?` operator to read the documentation for the `hist()` to see a full list of these parameters.

```
hist(mpg$cty,
     xlab   = "Miles Per Gallon (City)",
     main   = "Histogram of MPG (City)", # main title
     breaks = 12,   # how many breaks?
     col    = "red",
     border = "blue")
```

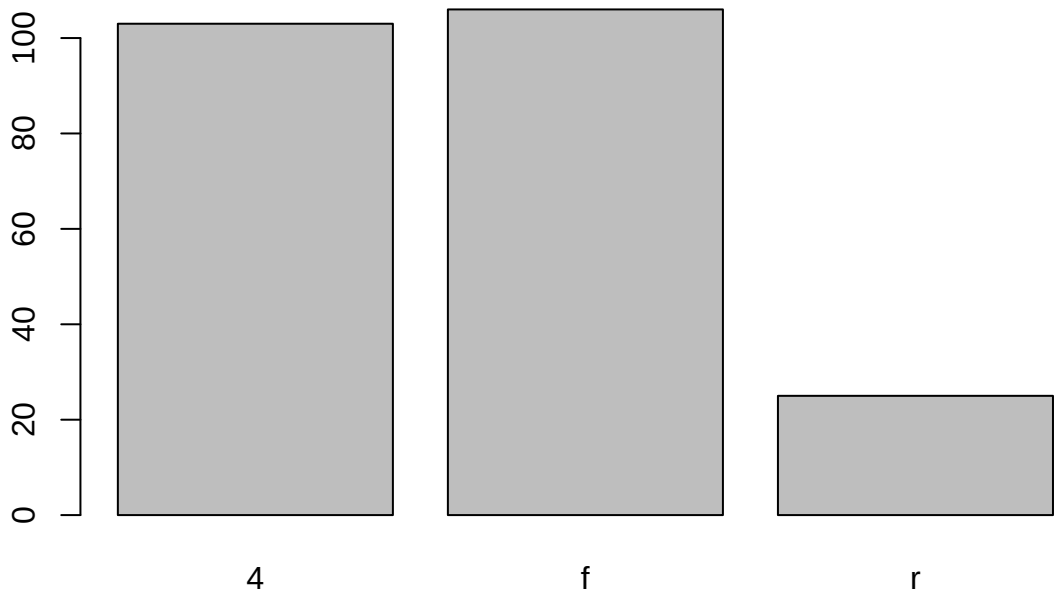


Importantly, you should always be sure to label your axes and give the plot a title. The argument `breaks` is specific to `hist()`. Entering an integer will give a suggestion to `R` for how many bars to use for the histogram. By default `R` will attempt to intelligently guess a good number of `breaks`, but as we can see here, it is sometimes useful to modify this yourself.

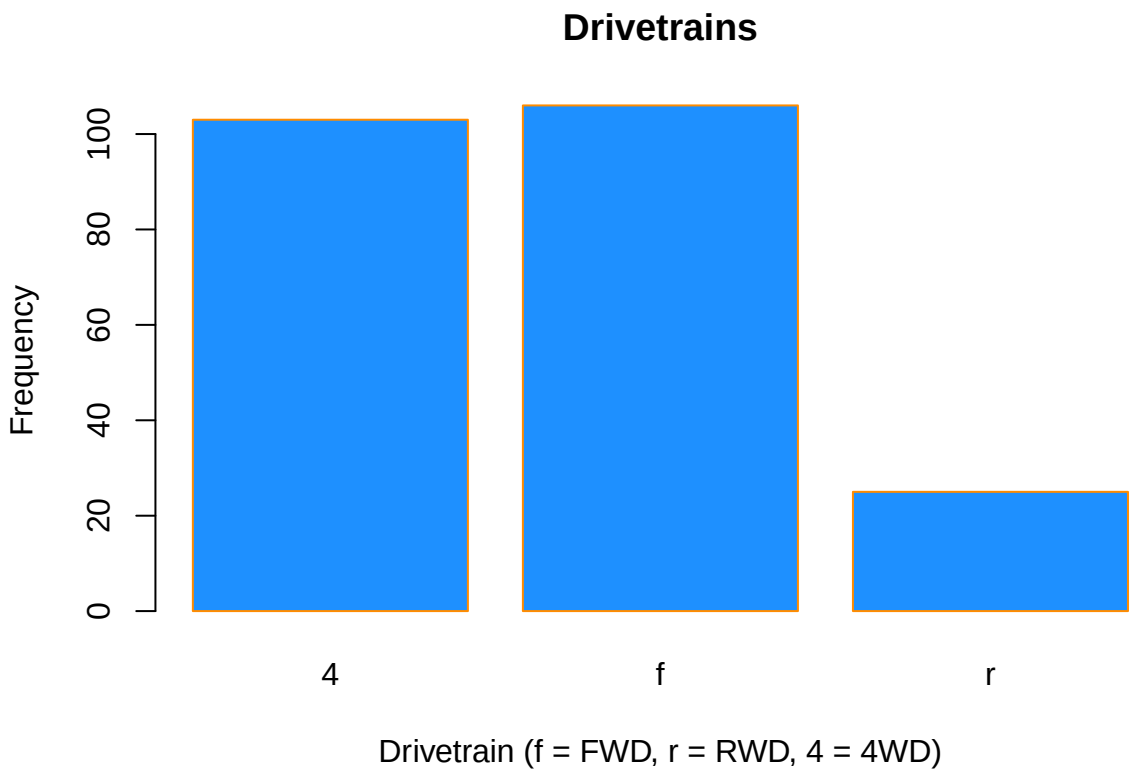
3.2.2 Barplots

Somewhat similar to a histogram, a barplot can provide a visual summary of a categorical variable, or a numeric variable with a finite number of values, like a ranking from 1 to 10.

```
barplot(table(mpg$drv))
```




```
barplot(table(mpg$drv),
        xlab   = "Drivetrain (f = FWD, r = RWD, 4 = 4WD)",
        ylab   = "Frequency",
        main   = "Drivetrains",
        col    = "dodgerblue",
        border = "darkorange")
```



3.2.3 Boxplots

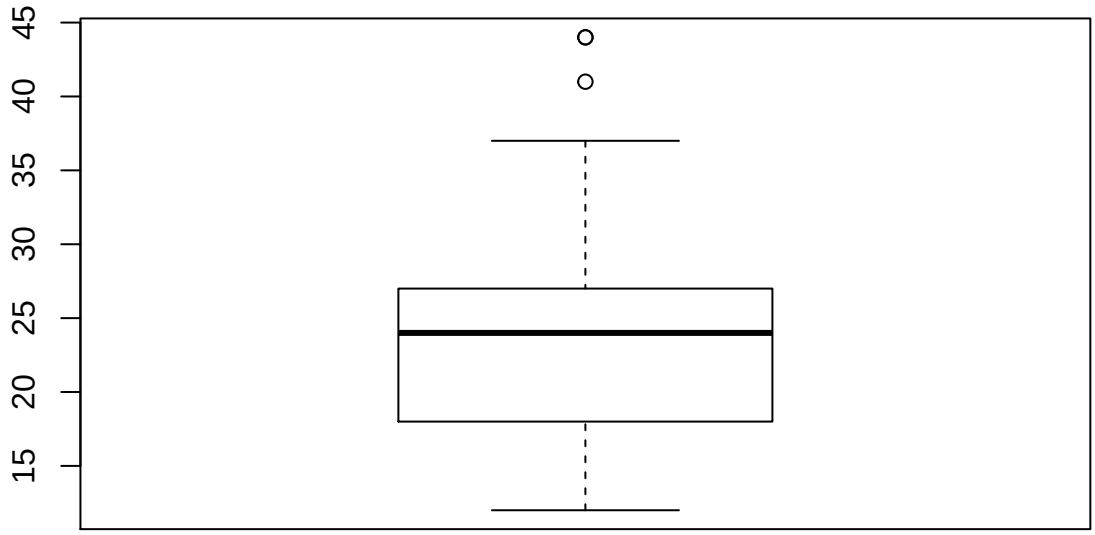
To visualize the relationship between a numerical and categorical variable, we will use a **boxplot**. In the `mpg` dataset, the `drv` variable takes a small, finite number of values. A car can only be front wheel drive, 4 wheel drive, or rear wheel drive.

```
unique(mpg$drv)
```

```
## [1] "f" "4" "r"
```

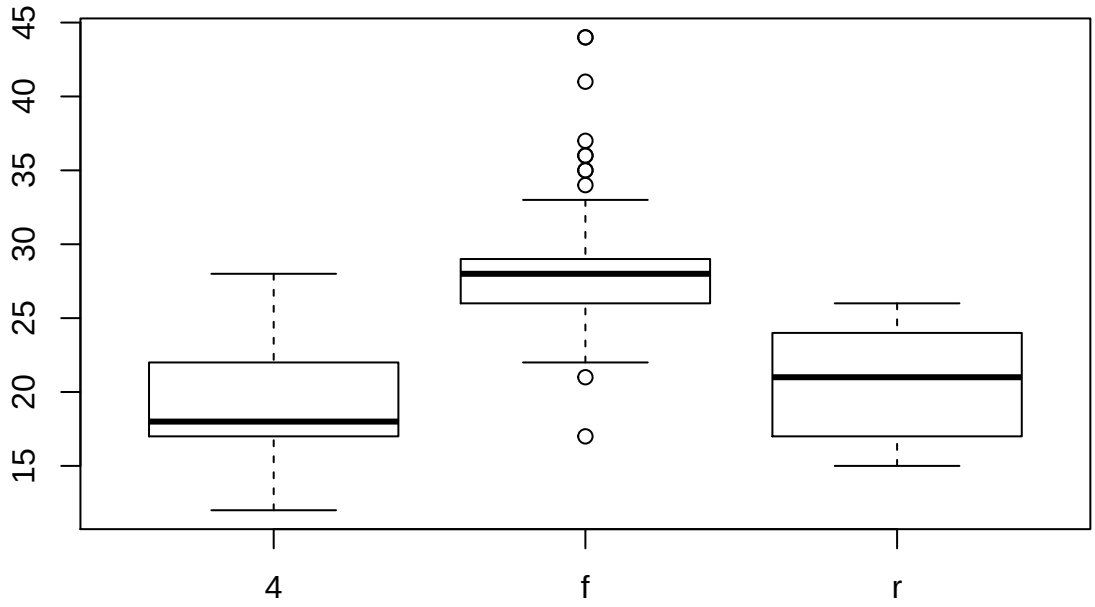
First note that we can use a single boxplot as an alternative to a histogram for visualizing a single numerical variable. To do so in `R`, we use the `boxplot()` function.

```
boxplot(mpg$hwy)
```



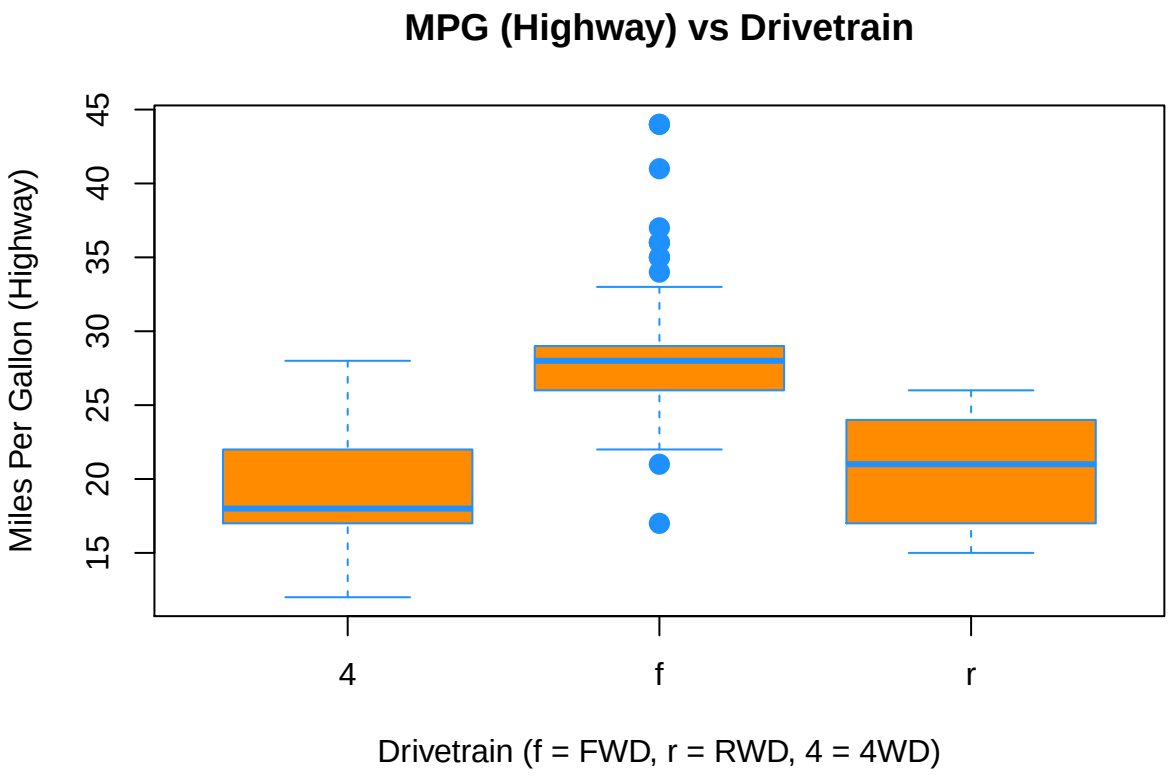
However, more often we will use boxplots to compare a numerical variable for different values of a categorical variable.

```
boxplot(hwy ~ drv, data = mpg)
```



Here used the `boxplot()` command to create side-by-side boxplots. However, since we are now dealing with two variables, the syntax has changed. The `R` syntax `hwy ~ drv, data = mpg` reads "Plot the `hwy` variable against the `drv` variable using the dataset `mpg`." We see the use of a `~` (which specifies a formula) and also a `data =` argument. This will be a syntax that is common to many functions we will use in this course.

```
boxplot(hwy ~ drv, data = mpg,
     xlab   = "Drivetrain (f = FWD, r = RWD, 4 = 4WD)",
     ylab   = "Miles Per Gallon (Highway)",
     main   = "MPG (Highway) vs Drivetrain",
     pch    = 20,
     cex    = 2,
     col    = "darkorange",
     border = "dodgerblue")
```

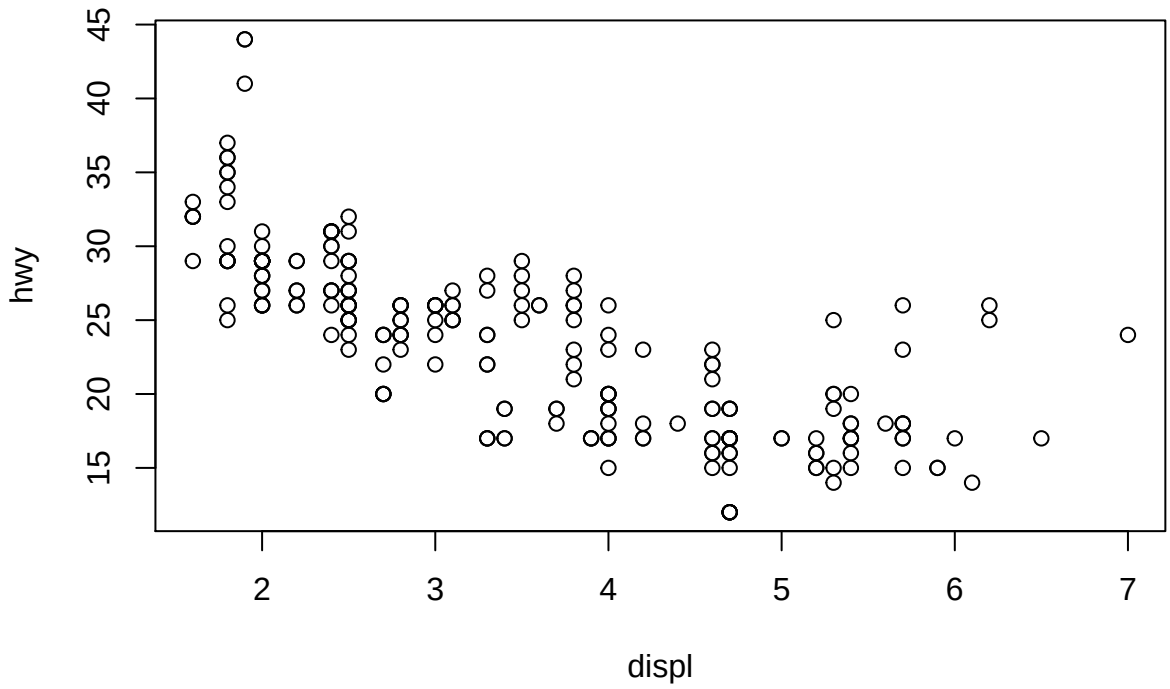


Again, `boxplot()` has a number of additional arguments which have the ability to make our plot more visually appealing.

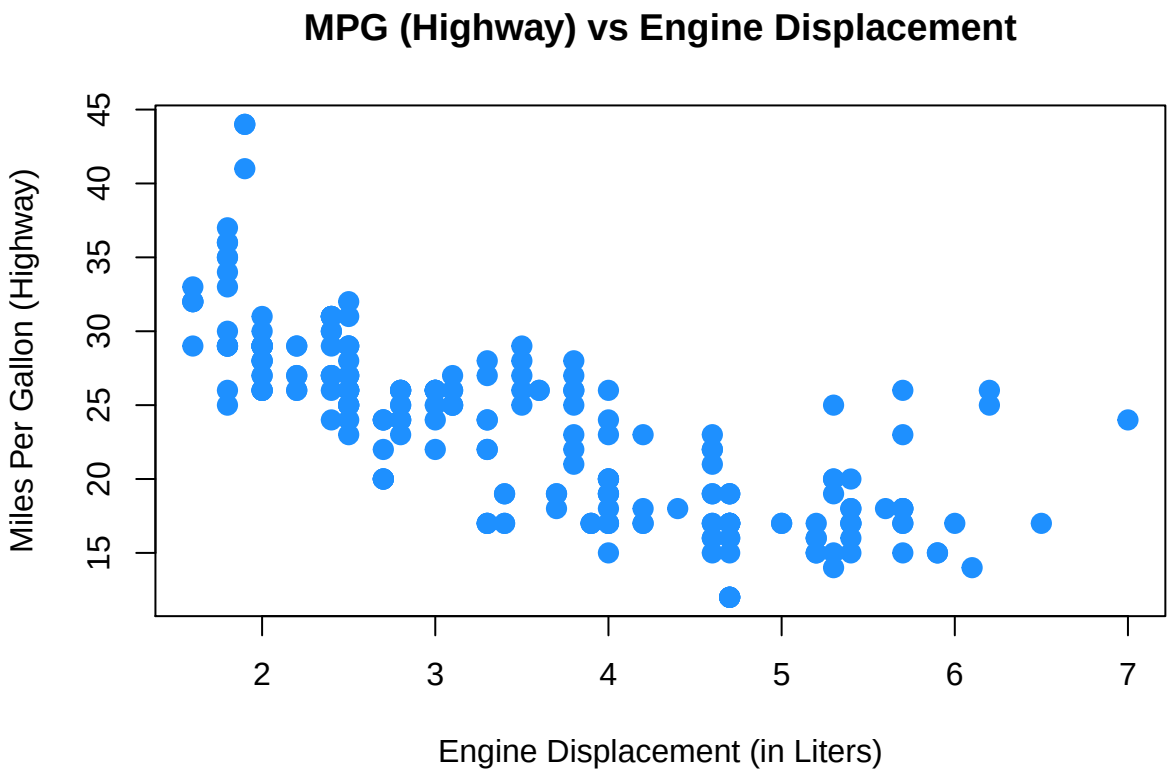
3.2.4 Scatterplots

Lastly, to visualize the relationship between two numeric variables we will use a **scatterplot**. This can be done with the `plot()` function and the `~` syntax we just used with a boxplot. (The function `plot()` can also be used more generally; see the documentation for details.)

```
plot(hwy ~ displ, data = mpg)
```



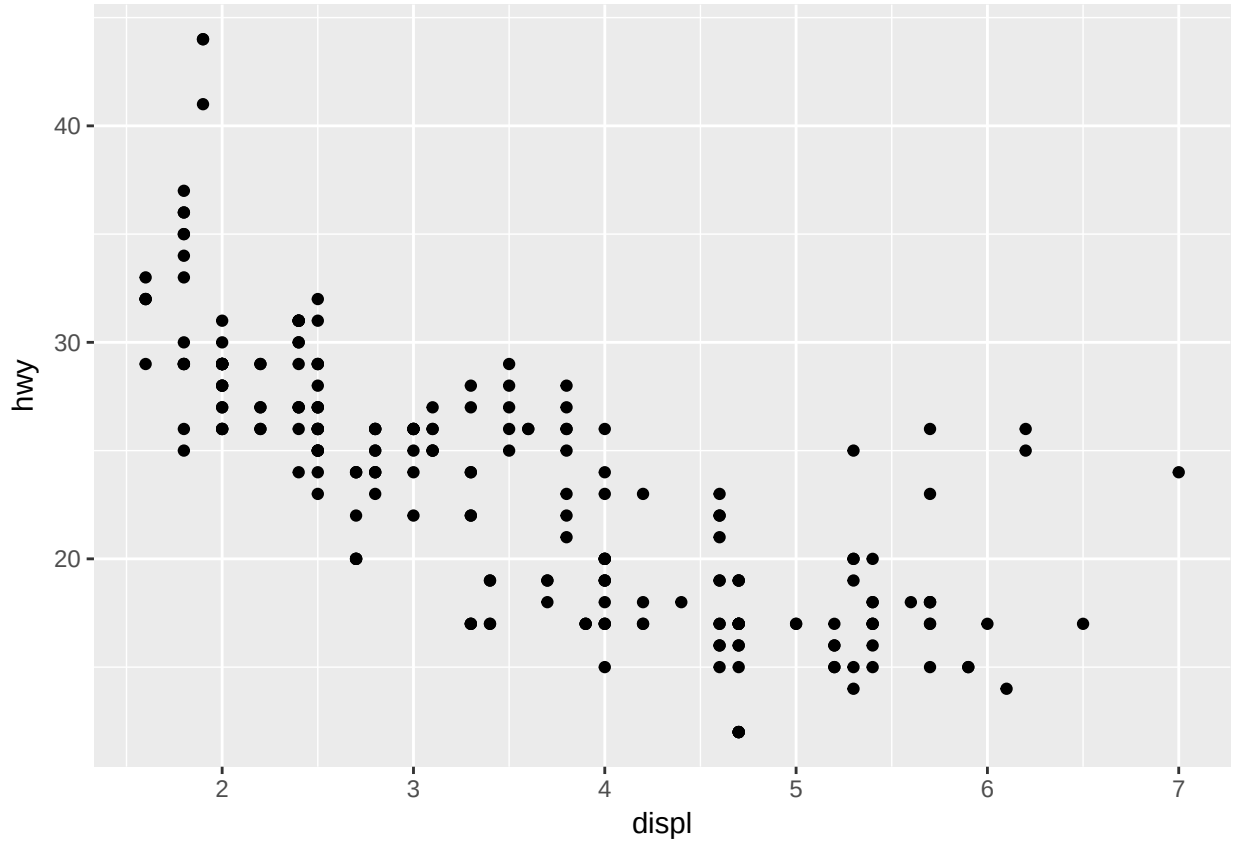

```
plot(hwy ~ displ, data = mpg,
     xlab = "Engine Displacement (in Liters)",
     ylab = "Miles Per Gallon (Highway)",
     main = "MPG (Highway) vs Engine Displacement",
     pch  = 20,
     cex  = 2,
     col  = "dodgerblue")
```



3.2.5 `ggplot`

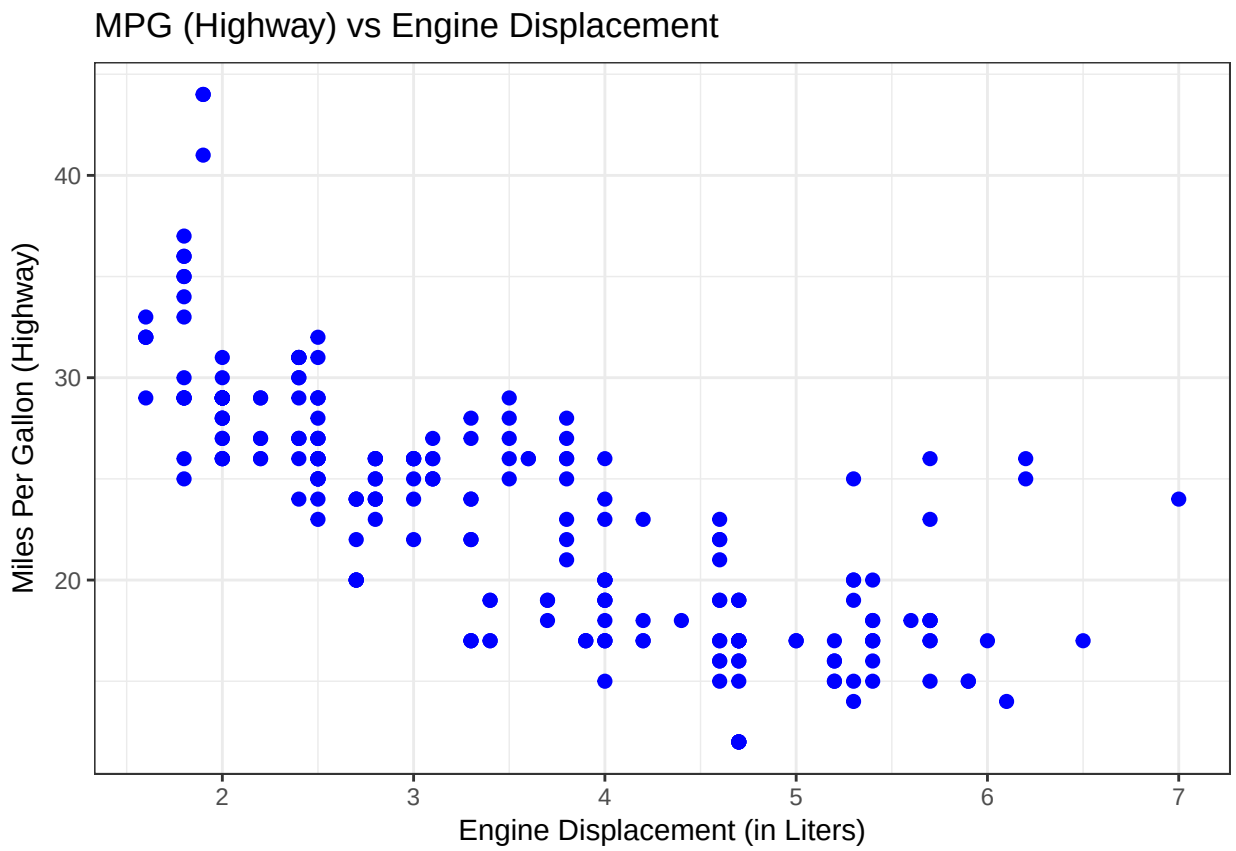
All of the above plots could also have been generate using the `ggplot` function from the already loaded `ggplot2` package. Which function you use is up to you, but sometimes a plot is easier to build in base R (like in the `boxplot` example maybe), sometimes the other way around.

```
ggplot(data=mpg,aes(x=displ,y=hwy)) + geom_point()
```



`ggplot` is impossible to describe in brief terms, so please look at the package's website which provides excellent guidance. We will from time to time use ggplot in this book, so try to familiarize yourself with it. Let's quickly demonstrate how one could customize that first plot:

```
ggplot(data=mpg,aes(x=displ,y=hwy)) +   # ggplot() makes base plot
  geom_point(color="blue",size=2) +     # how to show x and y?
  scale_y_continuous(name="Miles Per Gallon (Highway)") +  # name of y axis
  scale_x_continuous(name="Engine Displacement (in Liters)") + # x axis
  theme_bw() +    # change the background
  ggtitle("MPG (Highway) vs Engine Displacement")   # add a title
```



3.2.6 `dplyr`

The `dplyr` package is widely used tool to work with data in `R`. We will briefly look at it's functionality now.

Chapter 4

Linear Regression

1. linear regression - stop at R-squared
    1. different data: missing variable
    2. non-linear realtionship
2. scatter plot
    1. label observations
3. how do the data come to us? spreadsheet
4. approx link x and y by a line
5. OLS gives the best line for this
    1. $y_i = a + bx_i$. find a,b s.t. dist is minimal
    2. write out sum of least-squares and call it MSE: u_1 + u_2 + … / N
6. plot fitted values - see imperfect approximation
7. R-squared: goodness of fit / measure of goodness
    1. 1 - sum of squared errors / SST
    2. how much of total variance is explained by the model?
8. regression on mean
9. How come there are residuals?
    1. measurement error?
    2. there is more to this than just x
    3. misspecification
10. There is statistical uncertainty about those estimates
11. plot a second data set with a less clear interpretation
    1. do you *really* think there is a linear relationship?
    2. SE tells us whethe rwe really think this is a positive slope
    3. poor R2 and large standard error
    4. How **confident** are you about this relationship? Is there enought data?
    5. SE is ameasure of precision depending on N

4.1 Try to find the Slope!

```
knitr::include_url("https://gallery.shinyapps.io/simple_regression/")
```

Chapter 5

Standard Errors

1. Standard Errors
    1. True data
    2. play with N and see how errors behave

Chapter 6

Multiple Regression

- Same as simple regression but more variables
- income ~ age is not linear

```
knitr::include_app("https://gallery.shinyapps.io/multi_regression/",height="1600px")
```

Chapter 7

Categorial Variables: Dummies and Interactions

1. if you only have educ categories, the estimate is like the conditional mean by educ
2. if you have too many categories you will have empty bins

Chapter 8

Quantile Regression

1. before you were modelling the mean. the average link
2. now what happens to **outliers**? how robust is the mean to that
3. what about the entire distribution of this?

Chapter 9

Panel Data

- scanner data on breakfast cereals, (Q_{it}, D_{it})
- why does D vary with Q
- pos relation ship
- don't observe the group identity!
- unobserved het alpha is correlated with Q
- within group estimator
- what if you don't have panel data?

Chapter 10

Instrumental Variables

1. wage = educ
2. educ = ability
3. ols estimator captures your ability
4. you predict x by z and you only keep the variation that is in the error

Chapter 11

Logit and Probit

Chapter 12

Principal Component Analysis

Chapter 13

Notes

this creates a library for the used R packages.

13.1 Book usage

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter **??**. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter 4.

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 13.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 13.1.

```
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).

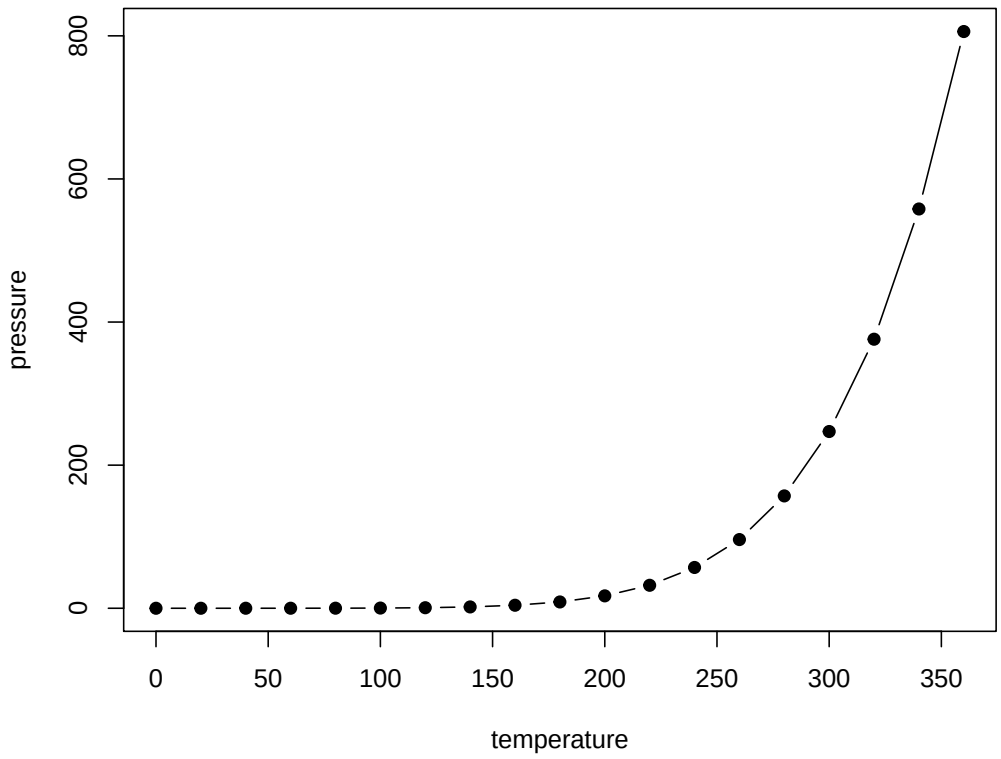


Figure 13.1: Here is a nice figure!

Table 13.1: 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

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 0.7.7.