Introduction to Statistics

Contents

About this course

This course is an introduction to statistics, R and RStudio. Our primary aims are to introduce you to and help you become familiar with RStudio and quantitative methodologies critical to your development as an analyst.

By the end of the course, you should be able to understand fundamental research methods, apply them to real world problems and acquire competency in performing statistical functions using R.

Slides day 1

1 Introduction to R and RStudio

1.1 Learning objectives

In this session, we will have a look at R and RStudio. We will interact with both and use the various components of RStudio.

1.1.1 What is R?

R is an environment for statistical computing and graphics. RStudio is an editor or integrated development environment (IDE) that makes working with R much more comfortable.

To install R and RStudio on your computer, download both from the following sources:

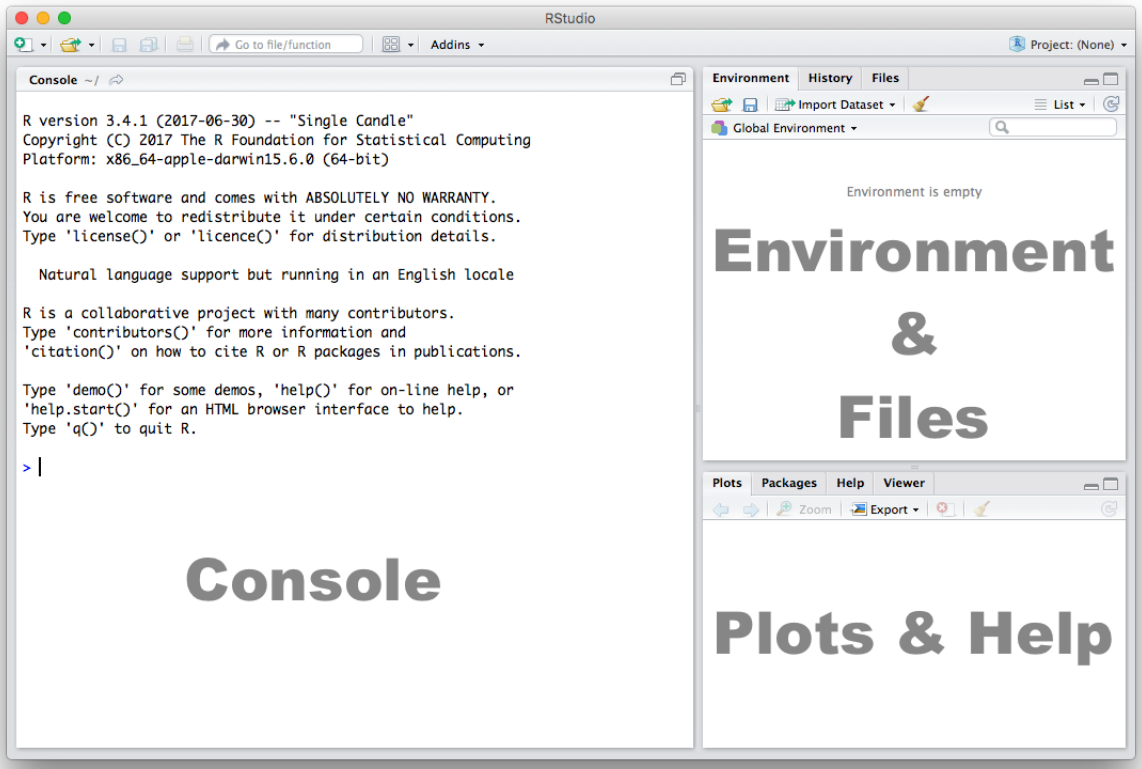
- Download R from The Comprehensive R Archive Network (CRAN)

- Download RStudio from RStudio.com

Keep both R and RStudio up to date. That means go online and check for newer versions. In case there are new versions, download those and re-install.

1.1.2 RStudio

Let's get acquainted with R. When you start RStudio for the first time, you'll see three panes:



1.1.3 Console

The Console in RStudio is the simplest way to interact with R. You can type some code at the Console and when you press ENTER, R will run that code. Depending on what you type, you may see some output in the Console or if you make a mistake, you may get a warning or an error message.

Let's familiarize ourselves with the console by using R as a simple calculator:

```
2 + 4
```

```
[1] 6
```

Now that we know how to use the `+` sign for addition, let's try some other mathematical operations such as subtraction (`-`), multiplication (`*`), and division (`/`).

```
10 - 4
```

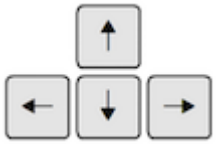
```
[1] 6
```

```
5 * 3
```

```
[1] 15
```

```
7 / 2
```

```
[1] 3.5
```



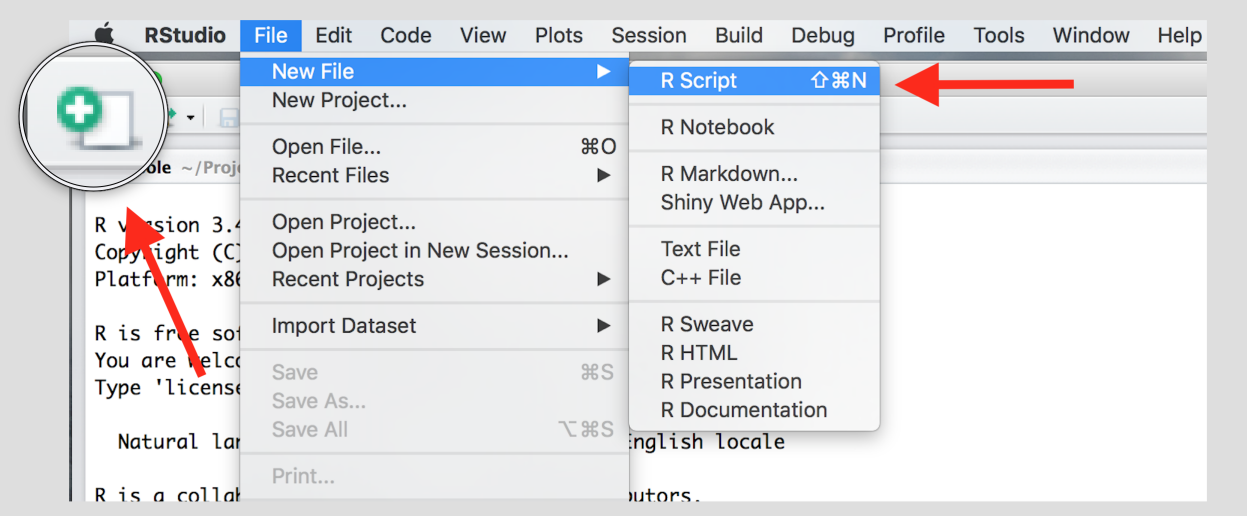
You can use the cursor or arrow keys on your keyboard to edit your code at the console:- Use the UP and DOWN keys to re-run something without typing it again- Use the LEFT and RIGHT keys to edit

Take a few minutes to play around at the console and try different things out. Don't worry if you make a mistake, you can't break anything easily!

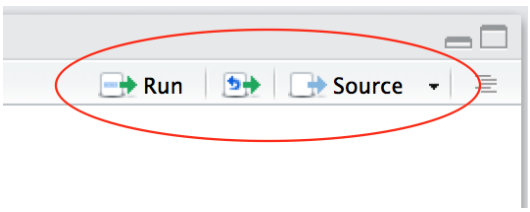
1.1.4 Scripts

The Console is great for simple tasks but if you're working on a project you would mostly likely want to save your work in some sort of a document or a file. Scripts in R are just plain text files that contain R code. You can edit a script just like you would edit a file in any word processing or note-taking application.

Create a new script using the menu or the toolbar button as shown below.



Once you've created a script, it is generally a good idea to give it a meaningful name and save it immediately. For our first session save your script as **seminar1.R**



Familiarize yourself with the script window in RStudio, and especially the two buttons labeled **Run** and **Source**

There are a few different ways to run your code from a script.

One line at a time	Place the cursor on the line you want to run and hit CTRL-ENTER or use the **Run** button
Multiple lines	Select the lines you want to run and hit CTRL-ENTER or use the **Run** button
Entire script	Use the **Source** button

2 Data Stuctures and Data Types

2.1 Seminar

In this session we introduce R-syntax, and data types.

2.1.1 Functions

Functions are a set of instructions that carry out a specific task. Functions often require some input and generate some output. For example, instead of using the `+` operator for addition, we can use the `sum` function to add two or more numbers.

```
sum(1, 4, 10)
```

```
[1] 15
```

In the example above, `1, 4, 10` are the inputs and 15 is the output. A function always requires the use of parenthesis or round brackets `()`. Inputs to the function are called **arguments** and go inside the brackets. The output of a function is displayed on the screen but we can also have the option of saving the result of the output. More on this later.

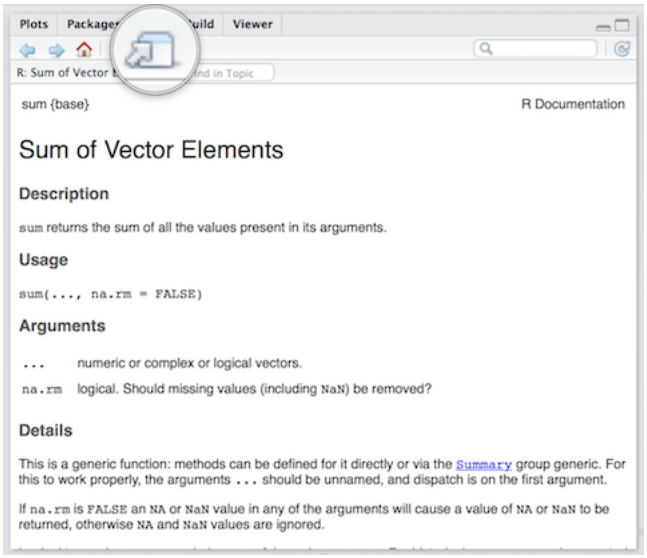
2.1.2 Getting Help

Another useful function in R is `help` which we can use to display online documentation. For example, if we wanted to know how to use the `sum` function, we could type `help(sum)` and look at the online documentation.

```
help(sum)
```

The question mark `?` can also be used as a shortcut to access online help.

```
?sum
```



Use the toolbar button shown in the picture above to expand and display the help in a new window.

Help pages for functions in R follow a consistent layout generally include these sections:

Description	A brief description of the function
Usage	The complete syntax or grammar including all arguments (inputs)
Arguments	Explanation of each argument
Details	Any relevant details about the function and its arguments
Value	The output value of the function
Examples	Example of how to use the function

2.1.3 The Assignment Operator

Now we know how to provide inputs to a function using parenthesis or round brackets `()`, but what about the output of a function?

We use the assignment operator `<-` for creating or updating objects. If we wanted to save the result of adding `sum(1, 4, 10)`, we would do the following:

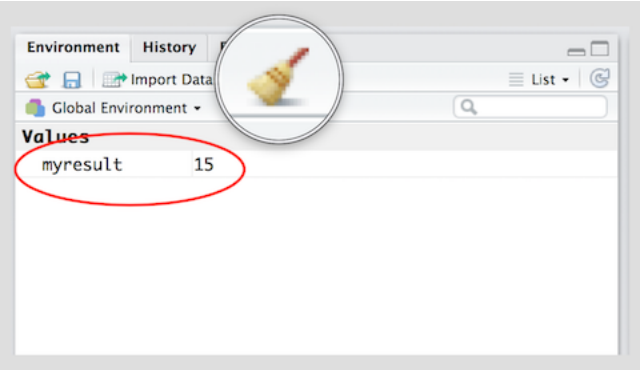
```
myresult <- sum(1, 4, 10)
```

The line above creates a new object called `myresult` in our environment and saves the result of the `sum(1, 4, 10)` in it. To see what's in `myresult`, just type it at the console:

```
myresult
```

```
[1] 15
```

Take a look at the **Environment** pane in RStudio and you'll see `myresult` there.



To delete all objects from the environment, you can use the **broom** button as shown in the picture above.

We called our object `myresult` but we can call it anything as long as we follow a few simple rules. Object names can contain upper or lower case letters (`A-Z`, `a-z`), numbers (`0-9`), underscores (`_`) or a dot (`.`) but all object names must start with a letter. Choose names that are descriptive and easy to type.

Good Object Names	Bad Object Names
result	a
myresult	x1
my.result	this.name.is.just.too.long
my_result	
data1	

2.1.4 Vectors and subsetting

A vector is one dimensional. It can contain one element in which case it is also called a scalar or many elements. We can add and multiply vectors. Think of a vector as a row or column in your excel spreadsheet.

To create a vector, we use the `c()` function, where c stands for collect. We start by creating a numeric vector.

```
vec1 <- c(10, 47, 99, 34, 21)
```

Creating a character vector works in the same way. We need to use quotation marks to indicate that the data type is textual data.

```
vec2 <- c("Emilia", "Martin", "Agatha", "James", "Luke", "Jacques")
```

Let's see how many elements our vector contains using the `length()` function.

```
length(vec1)
```

```
[1] 5
```

```
length(vec2)
```

```
[1] 6
```

We need one coordinate to identify a unique element in a vector. For instance, we may be interested in the first element of the vector only. We use square brackets `[]` to access a specific element. The number in square brackets is the vector element that we wish to see.

```
vec1[1]
```

```
[1] 10
```

To access all elements except the first element, we use the `-` operator

```
vec1[-1]
```

```
[1] 47 99 34 21
```

We can access elements 2 to 4 by using the colon `:` operator.

```
vec1[2:4]
```

```
[1] 47 99 34
```

We can access non-adjacent elements bu using the collect function `c()`.

```
vec1[c(2,5)]
```

```
[1] 47 21
```

Finally, we combine the `length()` function with the square brackets to access the last element in our vector.

```
vec1[ length(vec1) ]
```

```
[1] 21
```

2.1.5 Matrices

A matrix has two dimensions and stores data of the same type, e.g. numbers or text but never both. A matrix is always rectangular. Think of it as your excel spreadsheet - essentially, it is a data table.

We create a matrix using the `matrix()` function. We need to provide the following arguments:

```
mat1 <- matrix(
  data = c(99, 17, 19, 49, 88, 54),
  nrow = 2,
```

```
  ncol = 3,
  byrow = TRUE
)
```

Argument	Description
data	the data in the matrix
nrow	number of rows
ncol	number of columns
byrow	TRUE = matrix is filled rowwise

To display the matrix, we simply call the object by its name (in this case mat1).

```
mat1
```

```
     [,1] [,2] [,3]
[1,]   99   17   19
[2,]   49   88   54
```

To access a unique element in a matrix, we need 2 coordinates. First, a row coordinate and second, a column coordinate. We use square brackets and separate the coordinates with a comma `[ , ]`. The row coordinate goes before the comma and the column coordinate after.

We can access the the second row and third column like so:

```
mat1[2, 3]
```

```
[1] 54
```

To display an entire column, we specify the column we want to display and leave the row coordinate empty like so:

```
# display the 2nd column
mat1[ , 2]
```

```
[1] 17 88
```

Similarly, to display the entire second row, we specify the row coordinate but leave the column coordinate empty.

```
mat1[2, ]
```

```
[1] 49 88 54
```

2.1.6 Arrays

Arrays are similar to matrices but can contain more dimensions. You can think of an array as stacking multiple matrices. Generally, we refer to the rows, columns and layers in array. Let's create an array with 2 rows, 3 columns and 4 layers using the `array()` function.

```
arr1 <- array(
  data = c(1:24),
  dim = c(2, 3, 4)
)
```

To display the object, we call it by its name.

```
arr1
```

```
, , 1
```

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

, , 2

     [,1] [,2] [,3]
[1,]    7    9   11
[2,]    8   10   12

, , 3

     [,1] [,2] [,3]
[1,]   13   15   17
[2,]   14   16   18

, , 4

     [,1] [,2] [,3]
[1,]   19   21   23
[2,]   20   22   24
```

We can subset an array using the square brackets `[]`. To access a single element we need as many coordinates as our object has dimensions. Let's check the number of dimensions in our object first.

```
dim(arr1)
```

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

The `dim()` function informs us that we have 3 dimensions. The first is of length 2, the second of length 3 and the fourth of length 4.

Access the second column of the third layer on your own.

```
arr1[ , 2, 3]
```

```
[1] 15 16
```

3 Creating and Exploring Data Frames

3.1 Seminar

In this session we introduce data frames using R-syntax.

3.1.1 Clearing your workspace from previous work

To remove a specific object or in this case a matrix, we can use the following command.

```
ls()
```

```
[1] "chapter_header"
```

```
rm(mat1)
```

```
Warning in rm(mat1): object 'mat1' not found
```

Otherwise we can remove everything, which is good practice when starting new work. The same can also be achieved by clicking on the yellow broomstick in the Enviornemnt and Files Panel in R Studio (top right hand panel)

```
rm( list = ls() )
```

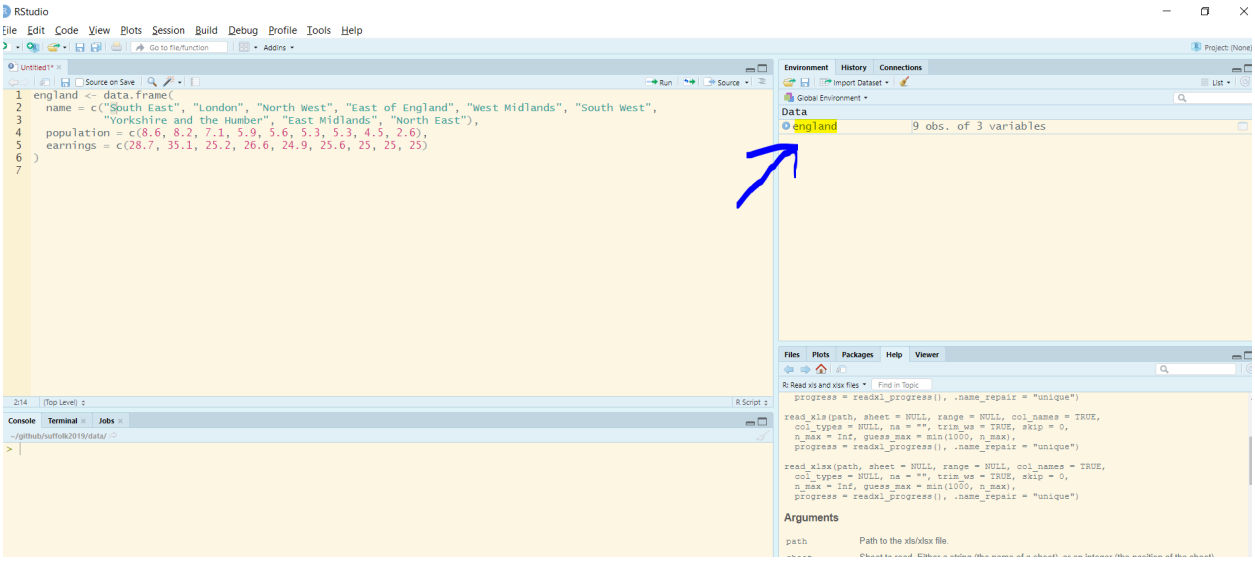
3.1.2 Creating data frames

A data frame is an object that holds data in a tabular format similar to how spreadsheets work. Variables are generally kept in columns and observations are in rows. Data frames are similar to matrices but they can store vectors of different types (e.g. numbers and text).

We start by creating a data frame with the `data.frame()` function. We will give each column a name (a variable name) followed by the `=` operator and the respective vector of data that we want to assign to that column.

```
england <- data.frame(
  name = c("South East", "London", "North West", "East of England", "West Midlands", "South West",
           "Yorkshire and the Humber", "East Midlands", "North East"),
  population = c(8.6, 8.2, 7.1, 5.9, 5.6, 5.3, 5.3, 4.5, 2.6),
  earnings = c(28.7, 35.1, 25.2, 26.6, 24.9, 25.6, 25, 25, 25)
)
```

We can also display the entire dataset in spreadsheet view by clicking on the object name in the environment window.



3.1.3 Working with data frames

we can display the entire dataset in spreadsheet view by clicking on the object name in the environment window.

Alternatively, you can call the object name to display the dataset in the console window. Let's do so:

```
england
```

```
                    name population earnings
1             South East        8.6     28.7
2                 London        8.2     35.1
3             North West        7.1     25.2
4       East of England        5.9     26.6
5         West Midlands        5.6     24.9
6             South West        5.3     25.6
7 Yorkshire and the Humber        5.3     25.0
```

```
8            East Midlands        4.5     25.0
9               North East        2.6     25.0
```

Often, datasets are too long to be viewed to in the console window. It is a good idea to look at the first couple of rows of a datasets to get an overview of its contents. We use the square brackets `[]` to view the first five rows and all columns.

```
england[1:5, ]
```

```
             name population earnings
1      South East        8.6     28.7
2          London        8.2     35.1
3      North West        7.1     25.2
4 East of England        5.9     26.6
5   West Midlands        5.6     24.9
```

v Columns in a dataframe have names. We will often need to know the name of a column/variable to access it. We use the `names()` function to view all variable names in a dataframe.

```
names(england)
```

```
[1] "name"       "population" "earnings"
```

We can access the earnings variable in multiple ways. First, we can use the `$` operator. We write the name of the dataset object, followed by the `$`, followed by the variable name like so:

```
england$population
```

```
[1] 8.6 8.2 7.1 5.9 5.6 5.3 5.3 4.5 2.6
```

We can also use the square brackets to access the earnings column.

```
england[, "population" ]
```

```
[1] 8.6 8.2 7.1 5.9 5.6 5.3 5.3 4.5 2.6
```

The square brackets are sometimes preferred because we could access multiple columns at once like so:

```
england[, c("name", "population") ]
```

```
                     name population
1               South East        8.6
2                   London        8.2
3               North West        7.1
4          East of England        5.9
5            West Midlands        5.6
6               South West        5.3
7 Yorkshire and the Humber        5.3
8            East Midlands        4.5
9               North East        2.6
```

You can also explore the variable further. For instance, calculate the average population in England and then select regions that have a population higher than this average. Again the dataset object is followed by the $and then the variable name like so

```
mean(england$population)
```

```
[1] 5.9
```

```
avg.pop = mean(england$population)
england [england$population > avg.pop,
         c("name", "population")]
```

```
             name population
1      South East        8.6
2          London        8.2
3      North West        7.1
4 East of England        5.9
```

Variables come in different types such as numbers, text, logical (true/false). We need to know the type of a variable because the type affects statistical analysis. We use the `str()` function to check the type of each variable in our dataset.

```
str(england)
```

```
'data.frame':   9 obs. of  3 variables:
 $ name      : Factor w/ 9 levels "East Midlands",..: 6 3 5 2 8 7 9 1 4
 $ population: num  8.6 8.2 7.1 5.9 5.6 5.3 5.3 4.5 2.6
 $ earnings  : num  28.7 35.1 25.2 26.6 24.9 25.6 25 25 25
```

The first variable in our dataset is a factor variable. Factors are categorical variables. Categories are mutually exclusive but they do not imply an ordering. For instance, "East of England" is not more or less than "West Midlands". The variables population and earnings are both numeric variables.

4 Data Import (from csv, txt, and excel) and Saving Data Frames

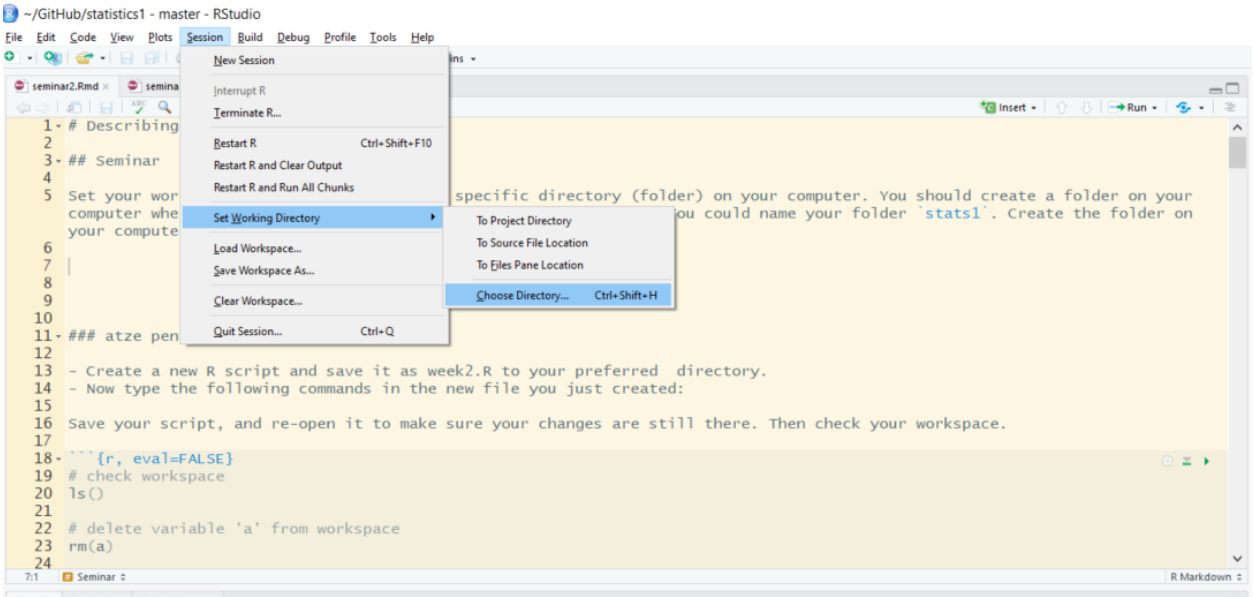
4.1 Seminar

In this section, we will learn how to check and set our working directory, load data in csv, txt, excel and R format, and then save our data frame.

4.1.1 Setting up

We set our working directory. R operates in specific directory (folder) on our computer. We create a folder on where we save our scripts. We name the folder `Stats101`. Let's create the folder on our computers now (in finder on Mac and explorer on Windows).

Now, we set our working directory to the folder, we just created like so:



Create a new R script and save it as day1.R to your `Stats101` directory.

At the beginning of each new script, we want to clear the workspace. The workspace is stored in working memory on our computer. If we do not clear it for a new script, it becomes too full over time. Our computer will slow down and it will become difficult for us to know which objects are stored in working memory.

Again, as we are starting a new script we should check the contents of our workspace like so:

```
# check workspace
ls()
```

Again to remove a specific object, we use the `rm()` function which stands for remove. Within the round brackets, we put the name of the object we want to remove. We could remove the object `a` like so:

```
# delete variable 'a' from workspace
rm(a)
```

At the beginning of each script, we should always clear the entire workspace. We can do so in the following way:

```
# delete everything from workspace
rm( list = ls() )
```

You can also clear text from the console window. To do so press Crtl+l on Windows or Command+l on Mac.

4.1.2 Loading data

Data comes in different file formats such as `.txt`, `.csv`, `.xlsx`, `.RData`, `.dta` and many more. To know the file type of a file right click on it and view preferences (in Windows explorer or Mac finder).

R can load files coming in many different file formats. To find out how to import a file coming in a specific format, it is usually a good idea to the google "R load file_format".

4.1.3 Importing a dataset in `.csv` format

One of the most common file types is `.csv` which means comma separated values. Columns are separated by commas and rows by line breaks.

Essentially, its best to work with this data format as it can be easily loaded into R, but also easy opening outside of R.

The dataset's name is "non_western_immigrants.csv". To load it, we use the `read.csv()` function.

```
dat1 <- read.csv("non_western_immigrants.csv")
```

4.1.4 Importing a dataset in Excel (xlsx) format

Another common file format is Microsoft's Excel `xlsx` format. We will load a dataset in this format now. To do so, we will need to install a package first. Packages are additional functions that we can add to R. A package is like an app on our phones.

We install the `readxl` package using `install.packages("readxl")`.

```
install.packages("readxl")
```

We only need to install a package once. It does not hurt to do it more often though, because every time we install, it will install the most recent version of the package.

Once a package is installed, we need to load it using the `library()` function.

```
library(readxl)
```

To load the excel file, we can now use the `read_excel()` function that is included in the `readxl` library. We need to provide the following arguments to the function:

Argument	Description
path	Filename of excel sheet
sheet	Sheet number to import

Now, let's load the file:

```
dat2 <- read_excel("non_western_immigrants.xlsx", sheet = 1)
```

4.1.5 Importing a dataset in RData format

The native file format of R is called `.RData`. To load files saved in this format, we use the `load()` function like so:

```
load(file = "non_western_immigrants.RData")
```

Notice that we usually need to assign the object we load to using the `<-` operator. The `load()` function is an exception where we do not need to do this.

4.1.6 Importing a dataset in .txt format.

Loading a dataset that comes in .txt format requires some additional information. The format is a text format and we need to know how the columns are separated. Usually it is enough to open the file in a word processor such as notepad to see how this is done. The most common ways to separate columns is by using commas or tabs but other separators such as for instance semicolons are sometimes also used.

In our example, columns are separated by semicolons. We use the `read.table()` function and provide the following arguments:

Argument	Description
file	Filename of excel sheet
sep	the symbol that separates columns
header	whether the first row contains variable names or not

```
dat3 <- read.table(file = "non_western_immigrants.txt", sep =";", header = TRUE)
```

4.1.7 Saving data frames

Datasets can be exported in many different file formats. We recommend exporting files as ".csv" files because csv is a very common file type. Such files can be handled by all statistical packages including Microsoft's Excel. We need to provide five arguments.

Argument	Description
x	The name of the object
file	The file name
sep	The symbol that separates columns
col.names	= TRUE saves the variable names (recommended)
row.names	= FALSE omits the row names (recommended)

Lets save the data we have open on pereceptions of non-western immigration. It is important to select a new file name, i.e. "newdata.csv", otherwise R overwrites the original dataframe and data may be lost. If updating a dataframe, it is good practice to save a file as V1.0, then V1.1 and so on. Let's try this below:

```
write.table(dat3,
            file = "non_western_immigrants_V1.1.csv",
            sep = ",",
            row.names = FALSE,
            col.names = TRUE
)
```

5 Descriptive Statistics

5.1 Seminar

Descriptive statistics are a good way to get a "feel" for the data and are an important first step for data analysis. Here we will explore two types: central tendency and dispersion.

Again, as good practice, let's first clear our workspace:

```
rm( list = ls() )
```

5.1.1 Central tendency

Central tendency explores the value of the observation at the center of a variable's distribution. This is the average or "typical" observation. What measure of central tendency you use depends on type of variable. In general, you use the following measures:

- Categorical variables - mode
- Ordinal variables - median
- Continuous variables - mean

As a recap:

- Categorical variables are unranked categories, such as political parties.
- Ordinal variables have categories that are ranked on a scale, i.e. council tax bands.
- Continuous variables have integer values and are simply not countable, i.e.income.
- Count variables represent countable data such as crime incidents
- Binary variables have only two categories i.e. employed or unemployed.

Let's first open some real data again:

```
dat1 <- read.csv("non_western_immigrants.csv", stringsAsFactors = FALSE)
dim (dat1)
```

```
[1] 1049   13
```

We can explore this in R by summarising our data: .

```
summary( dat1)
```

```
    IMMBRIT       over.estimate         RSex            RAge
 Min.   :  0.00   Min.   :0.0000   Min.   :1.000   Min.   :17.00
 1st Qu.: 10.00   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:36.00
 Median : 25.00   Median :1.0000   Median :2.000   Median :49.00
 Mean   : 29.03   Mean   :0.7235   Mean   :1.544   Mean   :49.75
 3rd Qu.: 40.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:62.00
 Max.   :100.00   Max.   :1.0000   Max.   :2.000   Max.   :99.00
    Househld         paper          WWWhourspW       religious
 Min.   :1.000   Min.   :0.0000   Min.   :  0.000   Min.   :0.0000
 1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:  0.000   1st Qu.:0.0000
 Median :2.000   Median :0.0000   Median :  2.000   Median :0.0000
```

```
 Mean   :2.392   Mean   :0.4538   Mean   :  5.251   Mean   :0.4929
 3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:  7.000   3rd Qu.:1.0000
 Max.   :8.000   Max.   :1.0000   Max.   :100.000   Max.   :1.0000
  employMonths        urban        health.good        HHInc
 Min.   :  1.00   Min.   :1.000   Min.   :0.000   Min.   : 1.000
 1st Qu.: 72.00   1st Qu.:2.000   1st Qu.:2.000   1st Qu.: 6.000
 Median : 72.00   Median :3.000   Median :2.000   Median : 9.000
 Mean   : 86.56   Mean   :2.568   Mean   :2.044   Mean   : 9.586
 3rd Qu.: 72.00   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:13.000
 Max.   :600.00   Max.   :4.000   Max.   :3.000   Max.   :17.000
   party_self
 Min.   :1.000
 1st Qu.:1.000
 Median :2.000
 Mean   :3.825
 3rd Qu.:7.000
 Max.   :7.000
```

However, this can be less useful when you have many variables in the data frame. So instead you can summzarise the variables of interest by subsetting:

```
summary (dat1[ , c("IMMBRIT", "over.estimate")])
```

```
    IMMBRIT       over.estimate
 Min.   :  0.00   Min.   :0.0000
 1st Qu.: 10.00   1st Qu.:0.0000
 Median : 25.00   Median :1.0000
 Mean   : 29.03   Mean   :0.7235
 3rd Qu.: 40.00   3rd Qu.:1.0000
 Max.   :100.00   Max.   :1.0000
```

Here the mean and median is reported. Again the median is relevant for ordinal values, while the mean is best for continuous, count and binary variables.

Large differences between the median and mean could indicate that a variable is skewed; when a variable has extreme values or many 0s which drags the mean either side of the median. For example, we often refer to "median incomes," even though income is a continous variable. Here the median is used because the mean is dragged to the right by extremely wealthy outliers, while most incomes are in fact clustered as the lower end. The median income in the UK for 2014/15, before housing costs, was £473 per week. But for the mean, weekly income is placed at £581.

You may have noticed that the mode is not reported. To gain the mode you can create a frequency table and then use the order function to display the most common categories first. This is done by using the table and order functions:

```
a <- table(dat1$party_self)
a[order(a, decreasing = TRUE)]
```

```
  7   1   2   6   5   4   3
383 284 280  32  31  23  16
```

As you can see, category 7 is the most common category, with the number of observations reported below the category.

5.1.2 Dispersion

Dispersion measures the spread of values within a variable. Again, the precise measure of dispersion depends on the level of measurement:

- Categorical/binary variables - proportion of each category
- Ordinal variables - range or interquartile range
- Continuous/count variables - variance or the standard deviation

Categorical variables and proportional percentages

Let's first start with categorical variables (not this is also applicable for binary variables). Here we should again look at the a frequency table to understand the proportion of each category, in this case the proportion of observations for each political party:

```
#absolute frequency
table1 <- table (dat1$party_self)
```

```
#and the percentages, rounded by two decimal spaces and then multiplied by 100
round (table1 /sum(table1), 2)*100
```

```
 1  2  3  4  5  6  7
27 27  2  2  3  3 37
```

This variable does not have labels for the categories, but by looking at the codebook of this data set we know the categories are as followed: 1 - Conservatives 2 - Labour 3 - SNP 4 - Green 5 - UKIP 6 - BNP 7 - Other

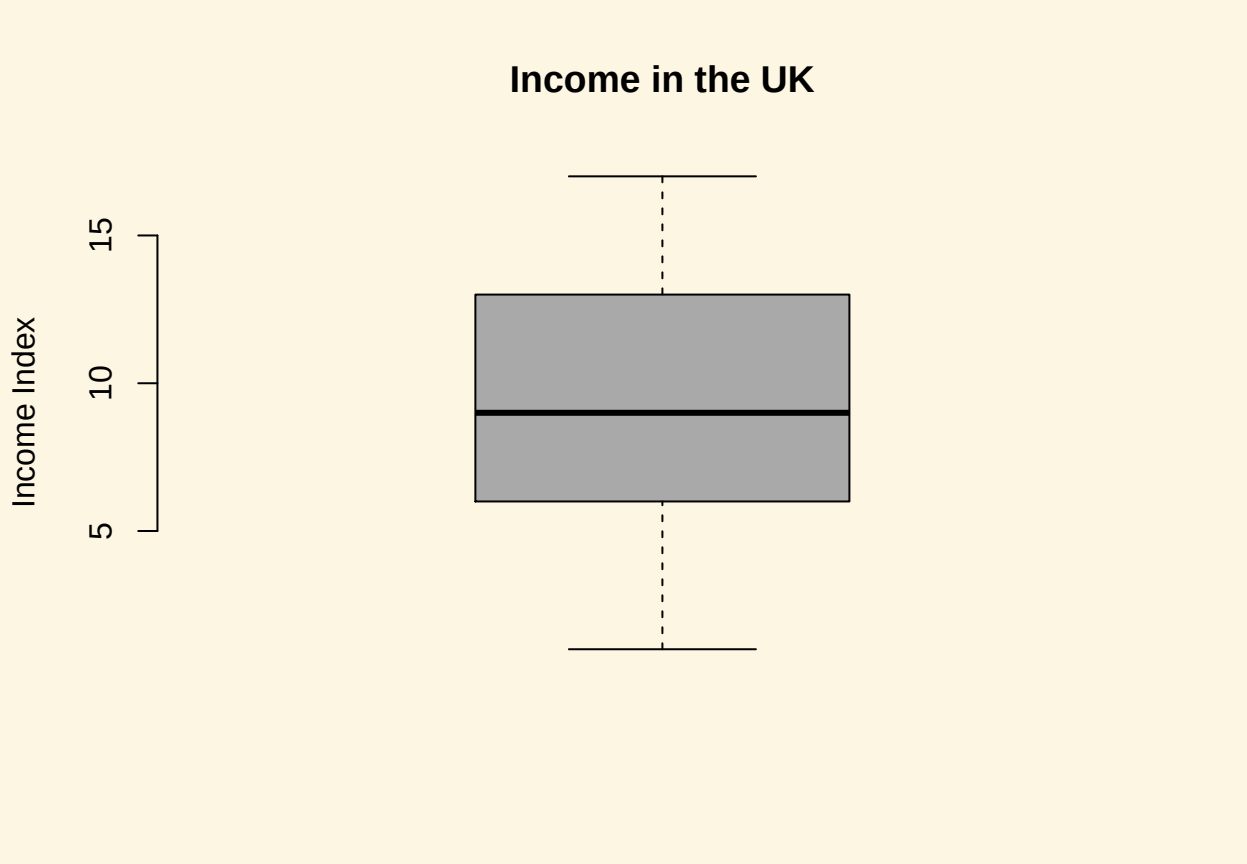
Perhaps this is not the best sample of our population, since there is no category for the Liberal Democrats!

Ordinal variables and the interquartile range

For ordinal variables, the most useful way to explore dispersion is by looking at the interquartile range. This can be visualised by plotting a simple box plot.

Note that with plots in R we have to provide arguments to design the plot, which are seperated by commas. Below we select: i) the variable of interest within the data frame (using the $ symbol), ii) add labels for the graph and y-axis, iii) remove the frame around the plot using the frame.plot function iv) choose the colour.

```
boxplot(
  dat1$HHInc,
  main = "Income in the UK",
  ylab = "Income Index",
  frame.plot = FALSE,
  col = "darkgray"
)
```



The box that is displayed within the plot represents the location of 50% of the data, between the 25% and 75% quartiles. The line within the box tells us the median. The two lines outside of the box shows us the data that falls within 1.5 times the quartiles.

Continuous variables and the standard deviation

For continuous (and count) variables we can look at the standard deviation. This tells us where 68% of the data can be found, since 68% of the data falls within one standard deviation of the mean. This can be done in R with a simple function. Here we are looking at British perceptions of non-Western immigration:

```
sd(dat1$IMMBRIT)
```

```
[1] 21.06331
```

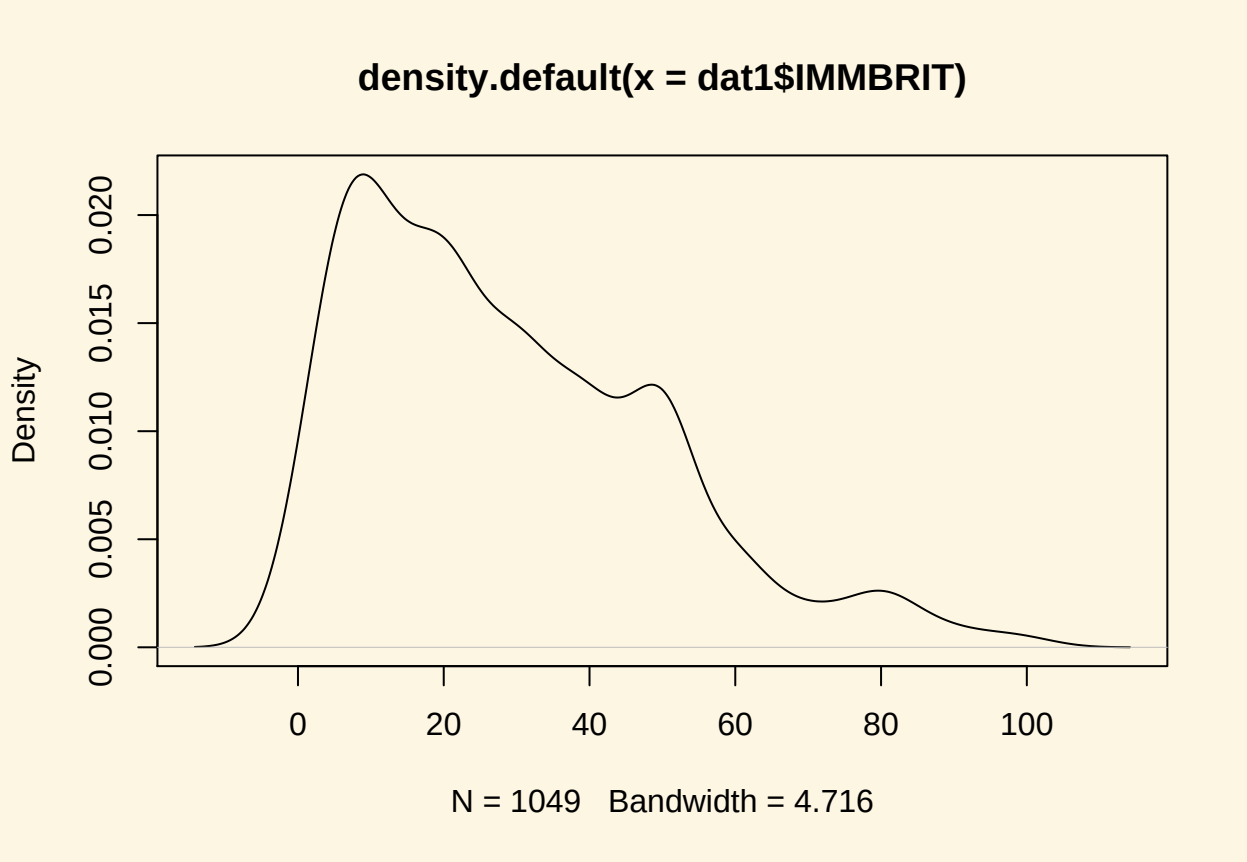
We can then compare this to the mean, which is another way to see if a continous variable is skewed:

```
mean(dat1$IMMBRIT)
```

```
[1] 29.03051
```

With the standard deviation at 21 and the mean at 29, we can see there is a large gap which indicates skewedness. But we can also simply visualise this by plotting the density of the variable:

```
plot(density(dat1$IMMBRIT))
```



As we can see most of the data is in the left hand side of the graph with fewer larger values on the far right of the graph. This drags the mean to the right which explains the large difference with the standard deviation. In this case it would be more appropriate to use the median (as is the case with income in the example above). When we run the median below, it is much closer to the centre of the data:

```
median(dat1$IMMBRIT)
```

```
[1] 25
```

To clear the graphs from our workspace, we can use the dev.off function with brackets:

```
dev.off()
```

```
null device
          1
```

6 Correlations and differences in means

6.1 Seminar

In this session, we will cover bi-variate relationships, that is relationships between two variables. For relationships between two continuous variables, we will look at correlations and plots and for relationships between a continuous dependent variable and a binary independent variable, we will look at differences in means.

6.1.1 Bi-variate relationships

In data analysis we are mostly interested in understanding the relationships between two (or more) variables. For instance, between a predictor (independent variable) and an outcome of interest (dependent variable). Summary statistics are often the first steps in trying to understand relationships between variables.

Initially we might want to visualise a relationship between an independent variable and a dependent variable (outcome). For example, we might want to see how two categories are related to an outcome of interest.

Again, as good practice, let's first clear our workspace before we start:

```
rm( list = ls() )
```

Let's open the data again:

```
dat1 <- read.csv("non_western_immigrants.csv", stringsAsFactors = FALSE)
dim (dat1)
```

```
[1] 1049   13
```

Here we will compare if there are differences in perceptions of non-Western immigrants between those that identify with the Convervatives and those that support the Labour party.

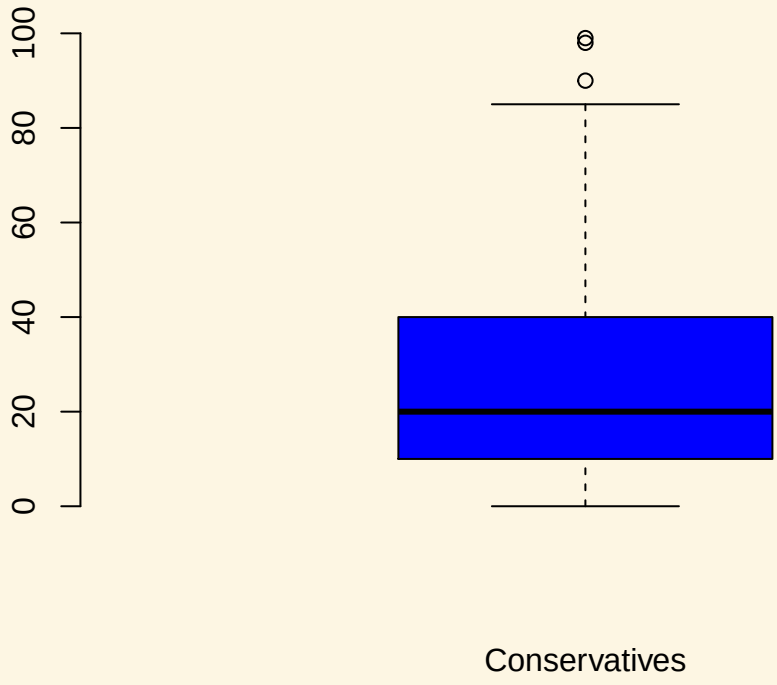
We need to first create two new dummy variables from the categorical variable that lists identified political parties. These two binary variables will tell us observations where someone identifies with the Conservatives (1=yes and 0 = no) and the same for the Labour party (1=yes and 0=no). Here we have to use the ifelse function:

```
dat1$conservatives <- ifelse(dat1$party_self == 1, yes = 1, no = 0)
dat1$labour <- ifelse(dat1$party_self == 2, yes = 1, no = 0)
```

So we create new variables within our dataframe using categories from the party_self variable. == 1 is the category we are choosing. If the ifelse statement is true, they do identify with that party, we code yes as a 1 and no as a 0.

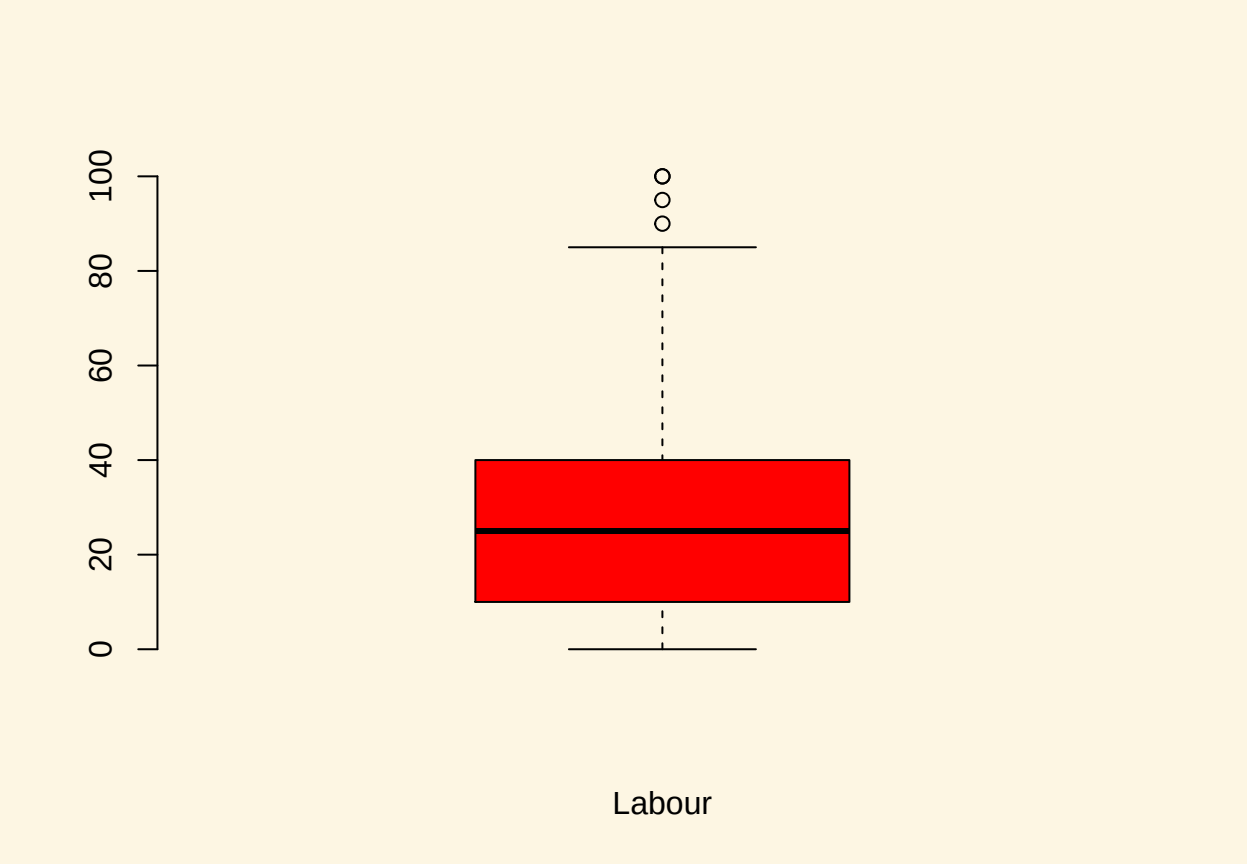
We can now run plots for the Conservatives and Labour party. We need to select the arguments as before, including appropriate colours for the parties. Here we also set the ylim to show a scale of the data from 0-100. First for the Conservative party:

```
boxplot(
  dat1$IMMBRIT[dat1$conservatives==1],
  frame.plot = FALSE,
  xlab = "Conservatives",
  ylim = c(0, 100),
  col = "blue"
)
```



and then for the Labour party:

```
boxplot(
  dat1$IMMBRIT[dat1$labour==1],
  frame.plot = FALSE,
  xlab = "Labour",
  ylim = c(0, 100),
  col = "red"
)
```



Overall, there appears to be very little difference between the Conservatives and Labour distributions in regards to perceptions of non-Western immigrants. The dots at the top of the scale represent outlier cases, where perceptions overestimate the number of immigrants.

The higher values for the Labour party could be due to high Labour support in diverse urban centres such as London. Yet, these are just samples and a snapshot of the possible relationship.Samples are never the same and vary. For instance, crime is not statistic and will vary over time and space.

Can we be sure that this result did not simply occur by chance?

This brings us to the importance of hypothesis testing.

6.1.2 Hypothesis testing

Because relationships can occur by chance, we engage in hyp

Hypotheis testing is central to statistics. We form our own hypothesis about a potential relationship between variables and then compare this against what is known as the null hypothesis. The null hypothesis represents the notion that nothing special has occured and that the relationship we are observing could only have occured by chance, i.e. getting heads when flipping a coin.

In the lecture slides, we cover this by exploring the Lady Tea Tasting Test (Fisher, 1925). When running statistics in R and other statistical packages, most models report a t-statistic and p-value, which we then informs us of whether we should reject the null hypothesis.

The t-statistic is easy to interpret. While we know 1 standard deviation from the mean represents 68% of the distribution, more than 1.96 deviations is more than 95% of the distribution. Therefore, if the t-statistic is higher than an absolute value of 1.96 (either 1.96 or -1.96), then the estimate we observe falls outside 95% of possible resampled estimations and is simply too rare to have occured by chance.

The p-value is based on the t-statistic and provides the actually probability that a estimate occured by chance. In most fields of research, the conventional threshold for a p-value is less than 0.05.This implies we will incorrectly reject the null hypothesis less than 5% of the time. A p-value of 0.01 implies we will incorrectly reject the null hypothesis 1% of the time. So the lower the p-value, the more confidence we can have in our estimate. If we reject the null hypothesis then the relationship we observe is statistically significant.

Let's move on to a t-test to assess and outcome and whether two categories are statistically significant from one another. Here we explore difference between two estimated means.

6.1.3 T-test (difference in means)

We are interested in whether there is a difference in income between countries that have an history of colonialisation and those that do not. Put more formally, we are interested in the difference between two conditional means. Recall that a conditional mean is the mean in a subpopulation, such as the mean of income given that the country was a victim of colonialisation (conditional mean 1).

The t-test is the appropriate test-statistic. Our interval-level dependent variable is `wdi_gdpc` which is GDP per capita taken from the World Development Indicators of the World Bank. Our binary independent variable is `h_j`.

Let's check the summary statistics of our dependent variable GDP per captia using the `summary()`. It returns several descriptive statistics as well as the number of NA observations (missing values). Missing values mean that we have no information on the correct value of the variable for an observation. Missing values may be missing for many different reasons. We need to be aware of missings because we cannot calculate with missings.

Lets bring the data in:]

```
dat2 <- read.csv("QoG2012.csv", stringsAsFactors = FALSE)
```

```
summary(dat2$wdi_gdpc)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  226.2  1768.0  5326.1 10184.1 12976.5 63686.7      16
```

We use the `which()` function to identify the row-numbers of the countries in our dataset that have a colonial history. The code below returns the row index numbers of countries with a history of colonialisation.

```
which(dat2$former_col == 1)
```

```
  [1]   3   5   6   8  11  12  13  15  17  18  20  21  22  23  24  26  27
 [18]  29  30  32  33  34  35  36  39  40  41  42  43  45  46  48  50  51
 [35]  52  53  54  56  58  61  62  64  66  67  69  70  71  72  73  74  77
 [52]  78  80  84  85  88  89  92  94  95  96  99 103 104 105 106 107 109
 [69] 110 111 116 117 118 119 120 123 125 126 127 129 130 131 132 133 134
 [86] 135 136 137 140 142 145 146 147 148 150 152 154 155 156 158 160 161
[103] 162 164 165 166 169 172 173 174 175 176 179 180 183 185 187 188 190
[120] 191 192 194
```

Now, we can explore conditional means. Below we access the variable that we want (`wdi_gdpc`) with the dollar sign and the rows in square brackets. The code below returns the per capita wealth of the countries with a colonial history:

```
mean( dat2$wdi_gdpc[dat2$former_col == 1], na.rm = TRUE)
```

```
[1] 6599.714
```

Now, go ahead and find the mean per capita wealth of countries without a colonial history:

```
mean( dat2$wdi_gdpc[dat2$former_col == 0], na.rm = TRUE)
```

```
[1] 16415.39
```

There is a clear numeric difference. Countries with a colonial history do appear to be significantly poorer.

However, we know that samples are subject to sampling variability. We therefore need to quantify the uncertainty that results from variable samples. To assess whether we can be reasonably sure that the difference between the estimates of wealth is not due to a strange sample or by chance, we carry out the t test to assess if there is a statistically significant difference between these populations:

```
t.test(dat2$wdi_gdpc[dat2$former_col == 1], dat2$wdi_gdpc[dat2$former_col == 0],
       mu = 0, alt = "two.sided", conf = 0.95)
```

```
    Welch Two Sample t-test

data:  dat2$wdi_gdpc[dat2$former_col == 1] and dat2$wdi_gdpc[dat2$former_col == 0]
t = -5.0603, df = 101.69, p-value = 1.866e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13663.313  -5968.043
sample estimates:
mean of x mean of y
 6599.714 16415.392
```

Let's interpret the results we get from `t.test()`. In our example the question is: Do countries with a former colonial history have different mean income levels than countries without colonial history?

The dependent variable is income, while the indepedent variable is where a country has a colonial history.

We know that when the t-statistic is greater then 1.96 or than -1.96, then we can safely reject the null hypothesis, i.e. it is less than 5% likely the estimate occured by chance. The p-value gives us a more direct probability of us wrongly rejecting the null hypothesis of `1.866e-06`which means a p-value of `0.000001866`!!!

In the next line you see the 95% confidence interval because we specified `conf=0.95`. If you were to take 100 samples and in each you checked the means of the two groups, 95 times the difference in means would be within the interval you see there.

At the very bottom you see the means of the dependent variable by the two groups of the independent variable. These are the conditional means that we estimated above. By minusing the $6599.7 from $16414.4, we know that former colonial countries are on average $9814.70 worst off!

6.1.4 Relationships between continuous variables

One way to explore the relationship between two continous variables is to use correlation. This is a widely used as a summary statistic. Correlation is a measure of **linear** association. It can take values between -1 and +1. Where -1 is a perfect negative relationship, 0 is a no relationship and +1 is a perfect positive relationship. 0 means there is no relationship, as would be expected if the null hypothesis was true.

We estimate the correlation coefficient in the following:

```
cor(y = dat2$undp_hdi, x = dat2$wbgi_cce, use = "complete.obs")
```

```
[1] 0.6821114
```

Argument	Description
`x`	The independent variable that you want to correlate.
`y`	The dependent variable that you want to correlate.
`use`	How R should handle missing values. `use="complete.obs"` will use only those rows where neither `x` nor `y` is missing.

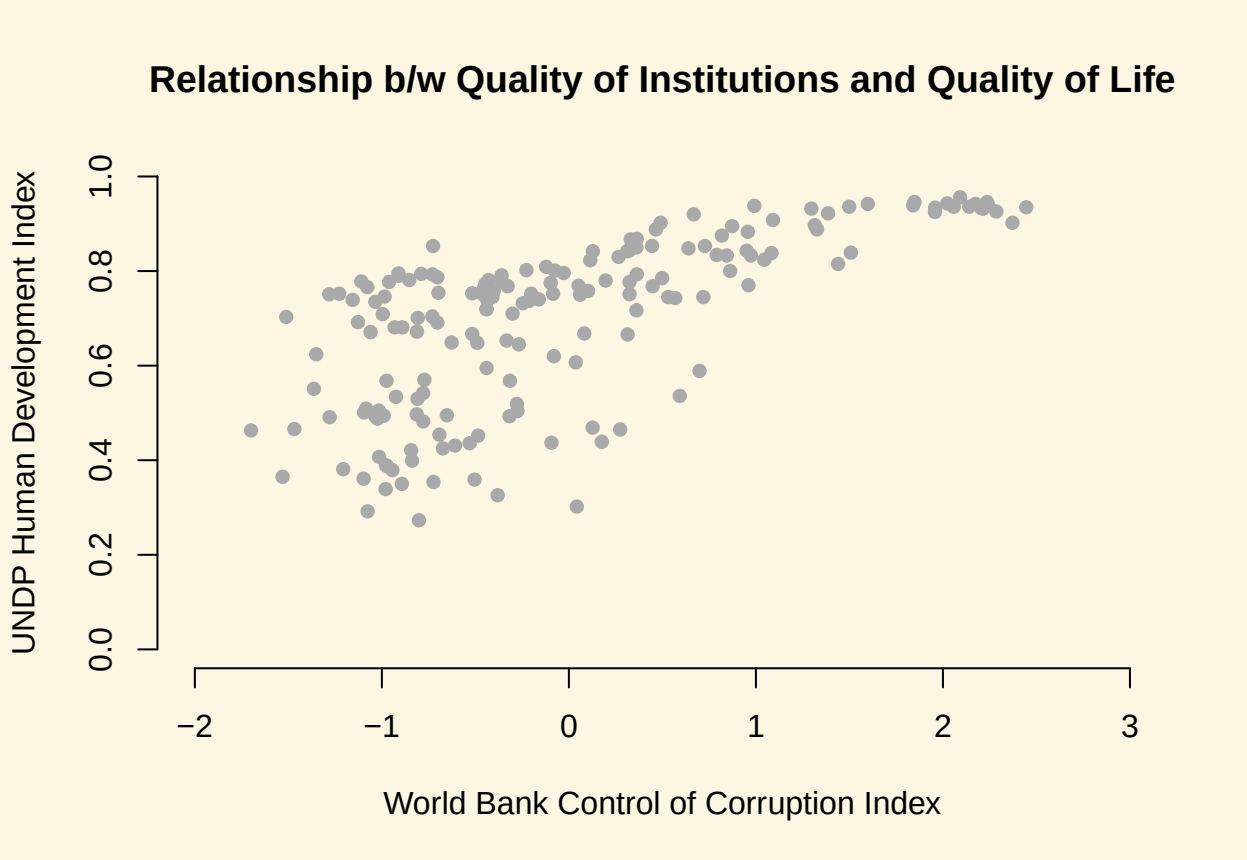
The interpretation of the correlation coefficient is that there is a positive relationship. However, the correlation coefficients does not tell us the magnitude of the relationship.

Another weakness is that it is a measure of linear association only. That means, there could be a curvilinear (or non-linear) realtionship which we miss (for instance a u-shaped relationship).

Often the best way to get a sense for the relationship beetween two continuous variables is visual inspection using a scatter plot. The human development index measures the quality of life and the World Bank Control of Corruption Index is a measure for the quality if institutions. We want to evaluate whether the two variables are related at all. Here, we could form the hypothesis that better insitutions improve the quality of life of citizens.

To investigate this relationship, we construct a scatterplot.

```
plot(
  x = dat2$wbgi_cce,
  y = dat2$undp_hdi,
  xlim = c(-2, 3),
  ylim = c(0, 1),
  frame = FALSE,
  xlab = "World Bank Control of Corruption Index",
  ylab = "UNDP Human Development Index",
  main = "Relationship b/w Quality of Institutions and Quality of Life",
  pch = 16,
  col = "darkgray"
)
```



The plot will give you a good idea whether about whether these two variables are related or not. Sometimes, the correleation coefficient is reported.

Overall, summary statistics are an important step in understanding the relationship between variables and whether such relationships occur by chance. However, this does not tell us much about the magnitude of the relationships we seek to explore. For this we must switch our attention to linear regression, which provides coefficients that can be interpreted as the extent to which a independent variable impacts a dependent variable. Linear regression also allows us to control for other independent variables which may also explain our dependent variable, i.e. alternative explainations of the outcome we are exploring.

7 Regression

7.1 Seminar

In this section, we will cover regression models. We will first introduce the bivariate linear regression model. We will then move to linear models with multiple independent variables.

7.1.1 Bivariate linear regression

We will use a dataset collected by the US census bureau that contains several socioeconomic indicators.

```
communities <- read.csv("communities.csv")
```

The dataset includes 38 variables but we're only interested in a handful at the moment.

Variable	Description
PctUnemployed	proportion of citizens in each community who are unemployed
PctNotHSGrad	proportion of citizens in each community who failed to finish high-school
population	proportion of adult population living in cities

If we summarize these variables with the `summary()` function, we will see that they are both measured as proportions (they vary between 0 and 1):

```
summary(communities$PctUnemployed)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.2200  0.3200  0.3635  0.4800  1.0000
```

```
summary(communities$PctNotHSGrad)
```

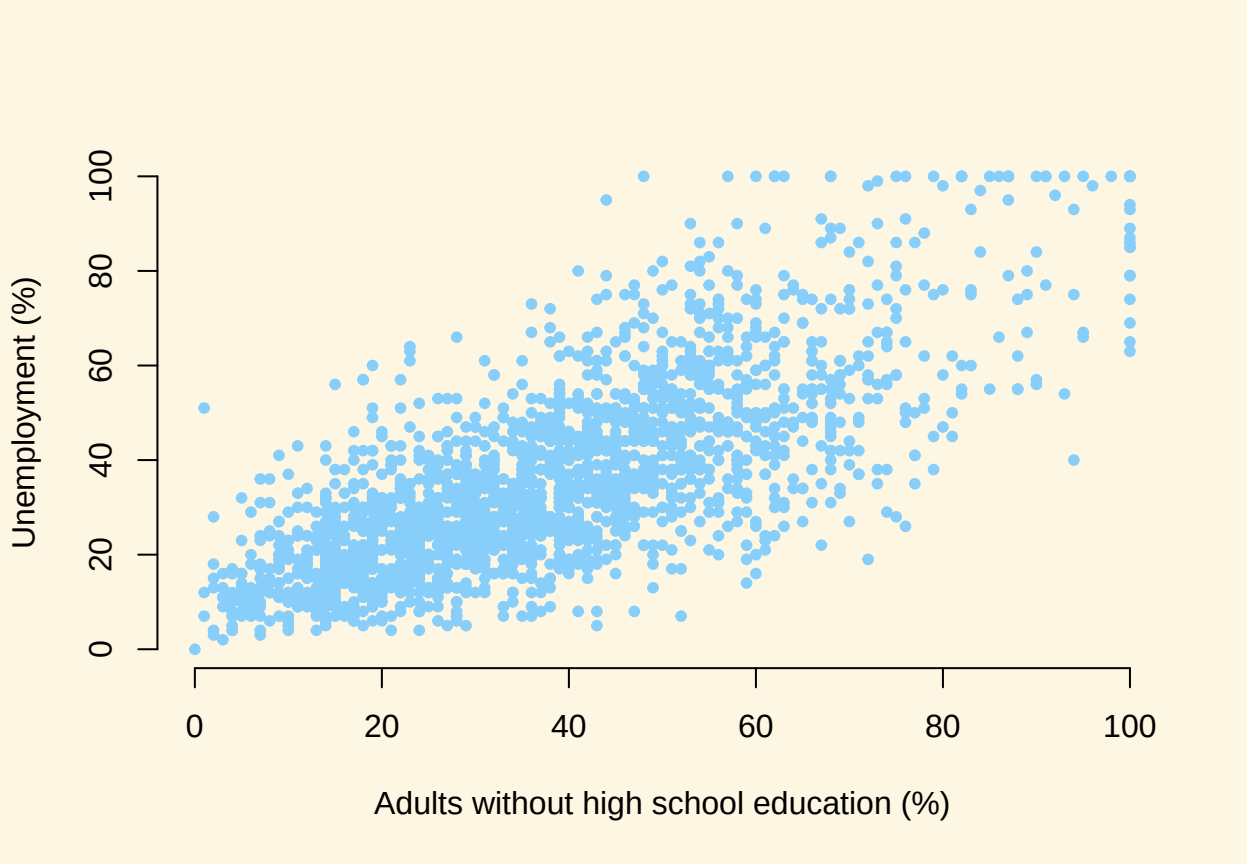
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.2300  0.3600  0.3833  0.5100  1.0000
```

It will be a little easier to interpret the regression output if we convert these to percentages rather than proportions. We can do this with the following lines of code:

```
communities$PctUnemployed <- communities$PctUnemployed * 100
communities$PctNotHSGrad <- communities$PctNotHSGrad * 100
```

We can begin by drawing a scatterplot with the percentage of unemployed people on the y-axis and the percentage of adults without high-school education on the x-axis.

```
plot(
  x = communities$PctNotHSGrad,
  y = communities$PctUnemployed,
  xlab = "Adults without high school education (%)",
  ylab = "Unemployment (%)",
  frame.plot = FALSE,
  pch = 20,
  col = "LightSkyBlue"
)
```



From looking at the plot, what is the association between the unemployment rate and lack of high-school level education?

In order to answer that question empirically, we will run a linear regression using the `lm()` function in R. The `lm()` function needs to know a) the relationship we're trying to model and b) the dataset for our observations. The two arguments we need to provide to the `lm()` function are described below.

Argument	Description
`formula`	The `formula` describes the relationship between the dependent and independent variables, for example `dependent.variable ~ independent.variable` In our case, we'd like to model the relationship using the formula: `PctUnemployed ~ PctNotHSGrad`
`data`	This is simply the name of the dataset that contains the variable of interest. In our case, this is the merged dataset called `communities`.

For more information on how the `lm()` function works, type help(lm) in R.

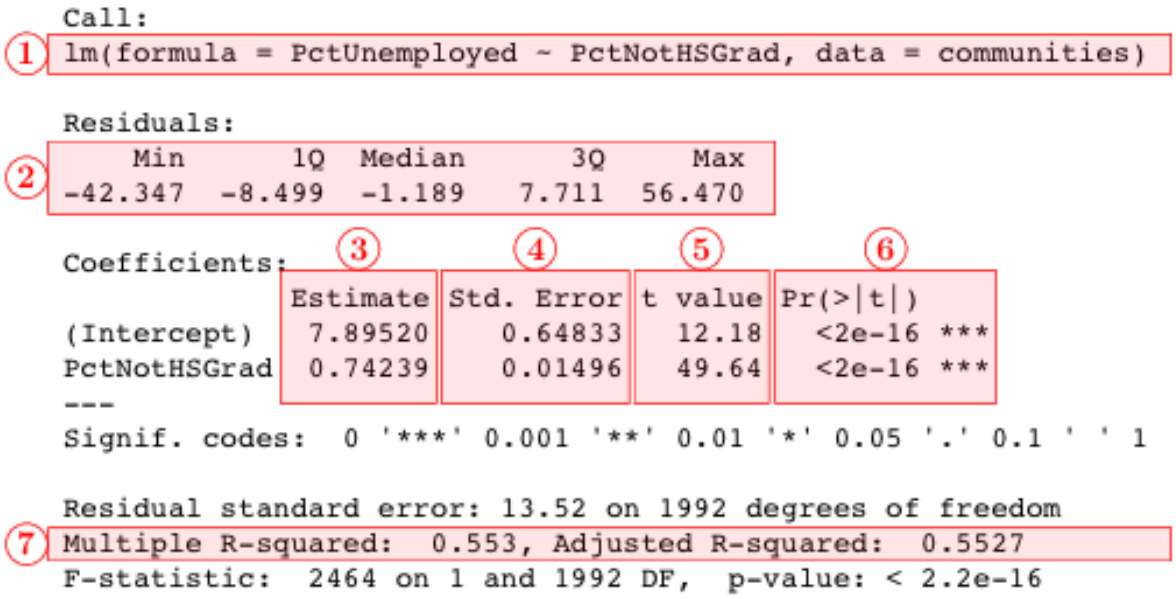
```
model1 <- lm(PctUnemployed ~ PctNotHSGrad, data = communities)
```

7.1.2 Interpreting Regression Output

The `lm()` function has modeled the relationship between `PctUnemployed` and `PctNotHSGrad` and we've saved it in an object called `model1`. Let's use the `summary()` function to see what this linear model looks like.

```
summary(model1)
```

The output from `summary()` might seem overwhelming at first so let's break it down one item at a time.

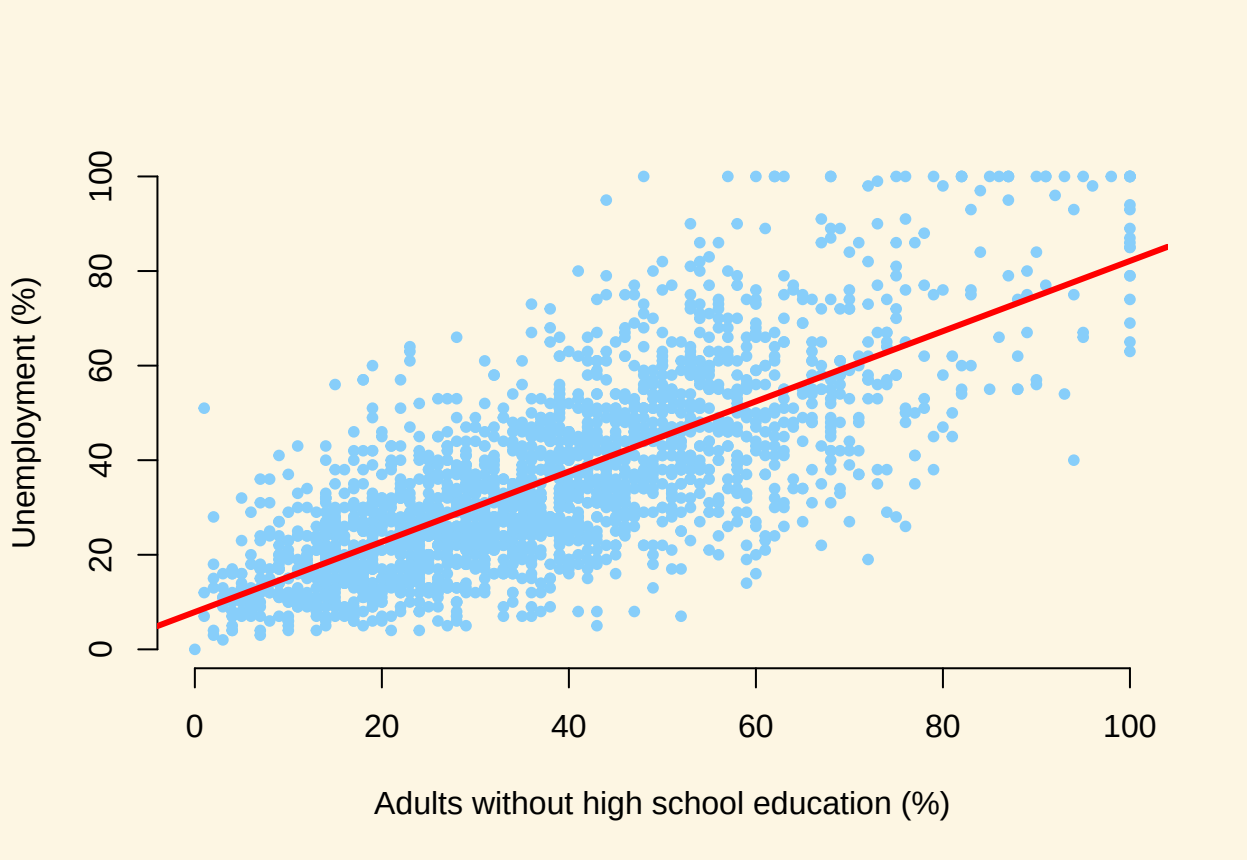


#	Item	Description
1	*formula*	The *formula* describes the relationship between the dependent and independent variables
2	*residuals*	The differences between the observed values and the predicted values are called *residuals*.
3	*coefficients*	The *coefficients* for all the *independent* variables and the intercept. Using the *coefficients* we can write down the relationship between the *dependent* and the *independent* variables as: `PctUnemployed` = 7.8952023 + (0.7423853 * `PctNotHSGrad`) This tells us that for each unit increase in the variable `PctNotHSGrad`, the `PctUnemployed` increases by 0.7423853.
4	*standard error*	The *standard error* estimates the standard deviation of the sampling distribution of the coefficients in our model. We can think of the *standard error* as the measure of precision for the estimated coefficients.
5	*t-statistic*	The *t-statistic* is obtained by dividing the *coefficients* by the *standard error*.
6	*p-value*	The *p-value* for each of the coefficients in the model. Recall that according to the null hypotheses, the value of the coefficient of interest is zero. The *p-value* tells us whether can can reject the null hypotheses or not.
7	R^2 and *adj-*R^2	tell us how much of the variance in our model is accounted for by the *independent* variable. The *adjusted* R^2 is always smaller than R^2 as it takes into account the number of *independent* variables and degrees of freedom.

Now let's add a regression line to the scatter plot using the `abline()` function.

First we run the same `plot()` function as before, then we overlay a line with `abline()`:

```r
plot(
  x = communities$PctNotHSGrad,
  y = communities$PctUnemployed,
  xlab = "Adults without high school education (%)",
  ylab = "Unemployment (%)",
  frame.plot = FALSE,
  pch = 20,
  col = "LightSkyBlue"
)

abline(model1, lwd = 3, col = "red")
```



We can see by looking at the regression line that it matches the coefficients we estimated above. For example, when `PctNotHSGrad` is equal to zero (i.e. where the line intersects the Y-axis), the predicted value for `PctUnemployed` seems to be above 0 but below 10. This is good, as the *intercept* coefficient we estimated in the regression was 7.8952023.

Similarly, the coefficient for the variable `PctNotHSGrad` was estimated to be 0.7423853, which implies that a one point increase in the percentage of citizens with no high-school education is associated with about 0.7423853 of a point increase in the percentage of citizens who are unemployed. The line in the plot seems to reflect this: it is upward sloping, so that higher levels of the no high-school variable are associated with higher levels of unemployment, but the relationship is not quite 1-to-1. That is, for each additional percentage point of citzens without high school education, the percentage of citizens who are unemployed increases by a little less than one point.

7.1.3 Multivariate linear regression

We might be interested in other variables that could explain the outcome. For instance, unemployment may be explained whether an community has a high urban population. For example, a more urban population may have better access to jobs.

Again we need to change the scale of a variable of interest - urban populationx, so that it is more interpretable.

```
communities$PctUrban <- communities$population *100
```

Now lets add this to the model.

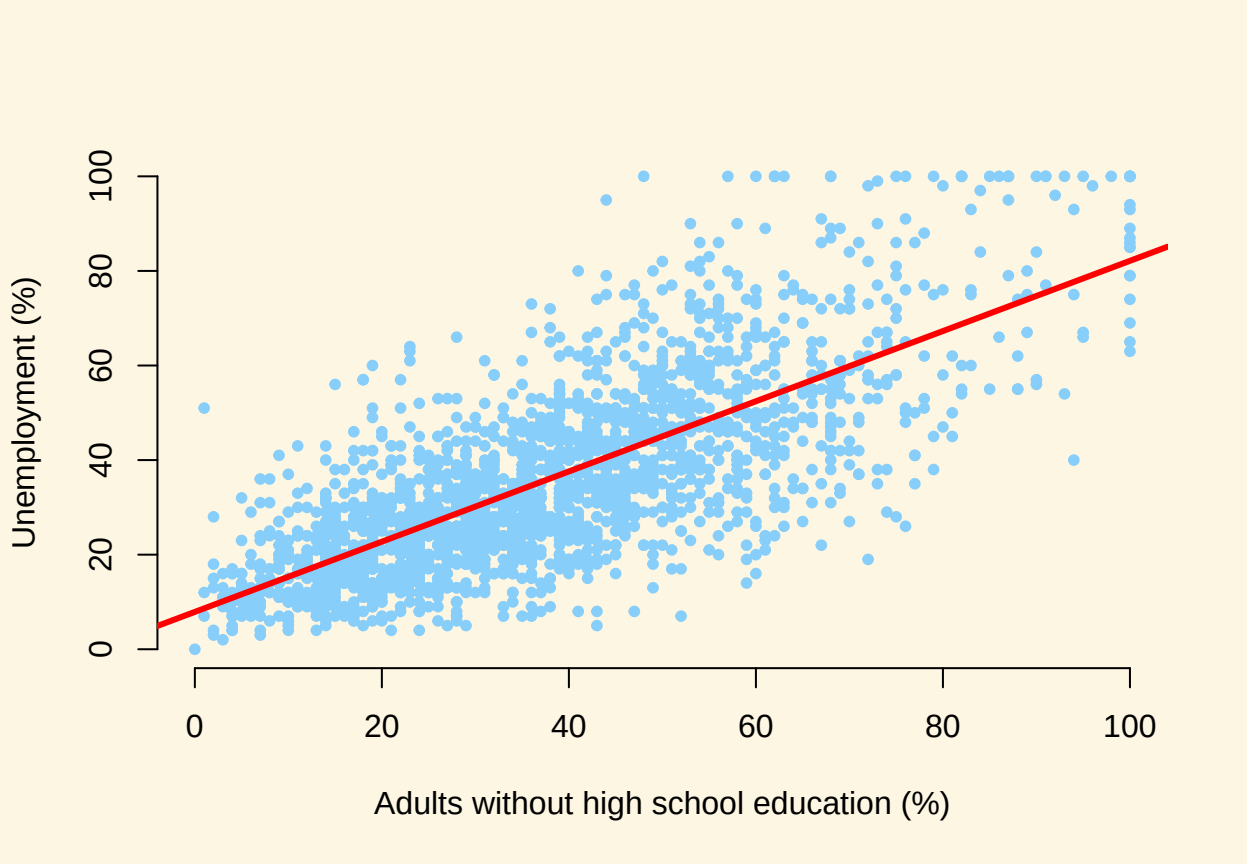
```
model2 <- lm(PctUnemployed ~ PctNotHSGrad + PctUrban, data = communities)
summary(model2)
```

```
Call:
lm(formula = PctUnemployed ~ PctNotHSGrad + PctUrban, data = communities)

Residuals:
    Min      1Q  Median      3Q     Max
-41.203  -8.372  -1.274   7.399  57.474

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.17377    0.64909  11.052  < 2e-16 ***
PctNotHSGrad  0.73651    0.01480  49.748  < 2e-16 ***
PctUrban      0.16436    0.02362   6.957 4.69e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.36 on 1991 degrees of freedom
Multiple R-squared:  0.5636,    Adjusted R-squared:  0.5631
F-statistic:  1286 on 2 and 1991 DF,  p-value: < 2.2e-16
```

The coefficient for urban population is positive, which suggests a higher urban population increases unemployment. A one unit change in population increases unemployment by 1.64%. Because this variable is at the same scale, we can compare this with the coefficient concerning the percentage of people without high school education. The p-value for this coefficient is also extremely low, meaning the estimate is statistically significant. Finally, we can see that the adjusted R-square has moved from 0.5527 to 0.5631 after adding the additional variable. This suggests the model is able to explain a bit more variation and has slightly more predictive power than the bivariate regression.

Let's plot the regression line:

```
plot(
  x = communities$PctNotHSGrad,
  y = communities$PctUnemployed,
  xlab = "Adults without high school education (%)",
  ylab = "Unemployment (%)",
  frame.plot = FALSE,
  pch = 20,
  col = "LightSkyBlue"
)
abline(model1, lwd = 3, col = "red")
```



As you can see from the plot, there are strange observations at the top of the y-axis and x-axis. This is strange as no observations with no education are likely to have 100% employment. These outliers or incorrect values will bias the results, so lets run another model without these high values using subsetting:

```
model3 <- lm(PctUnemployed ~ PctNotHSGrad + PctUrban,
             data = communities [communities$PctUnemployed < 100
                          & communities$PctNotHSGrad <100 , ])
summary(model3)
```

```
Call:
lm(formula = PctUnemployed ~ PctNotHSGrad + PctUrban, data = communities[communities$PctUnemployed <
    100 & communities$PctNotHSGrad < 100, ])

Residuals:
    Min      1Q  Median      3Q     Max
-38.832  -8.043  -1.301   7.264  55.962

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.61368    0.64421  13.371  < 2e-16 ***
PctNotHSGrad  0.68359    0.01528  44.745  < 2e-16 ***
PctUrban      0.17304    0.02269   7.626 3.77e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.78 on 1948 degrees of freedom
```

```
Multiple R-squared:  0.5204,    Adjusted R-squared:   0.52
F-statistic:  1057 on 2 and 1948 DF,  p-value: < 2.2e-16
```

This drops the coefficient slightly for the level of education attainment. What this means is the effect (or the slope of the regression line) on unemployment is reduced. However, the result is still statistically significant as the p-value is less than 0.05.

7.1.4 Additional Resources

- Linear Regression - Interactive App