Development of E-learning materials using R & R markdown

Ziv Shkedy, Bernard Osangir, Leyla Kodalci, Julia Duda

Hasselt University

Updated: 08/25

Visit us on Facebook ER-BioStat

GitHub https://github.com/eR-Biostat

twitter @erbiostat

Overview

- Starting point: first presentation:
  - Linear regression using R markdown.
  - Output development.

Slide 4-25

- From R markdown to development of E-learning course.
- From an output on my laptop to a course online.

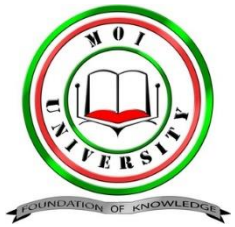
Slide 26-80

- Case study: 52-80

- Our approach: slides 81-127.

Rmd programs

- Basic analysis in R (10-16):
  - er_prog1_SA_2024.R.
- Basic HTML file (17-25):
  - er_prog2_SA_2024.Rmd.
- Advance HTML file (26-51):
  - er_prog3_SA_2024.Rmd.
- Case study (52-80):
  - er_prog4_VT_2024_V3a.Rmd.



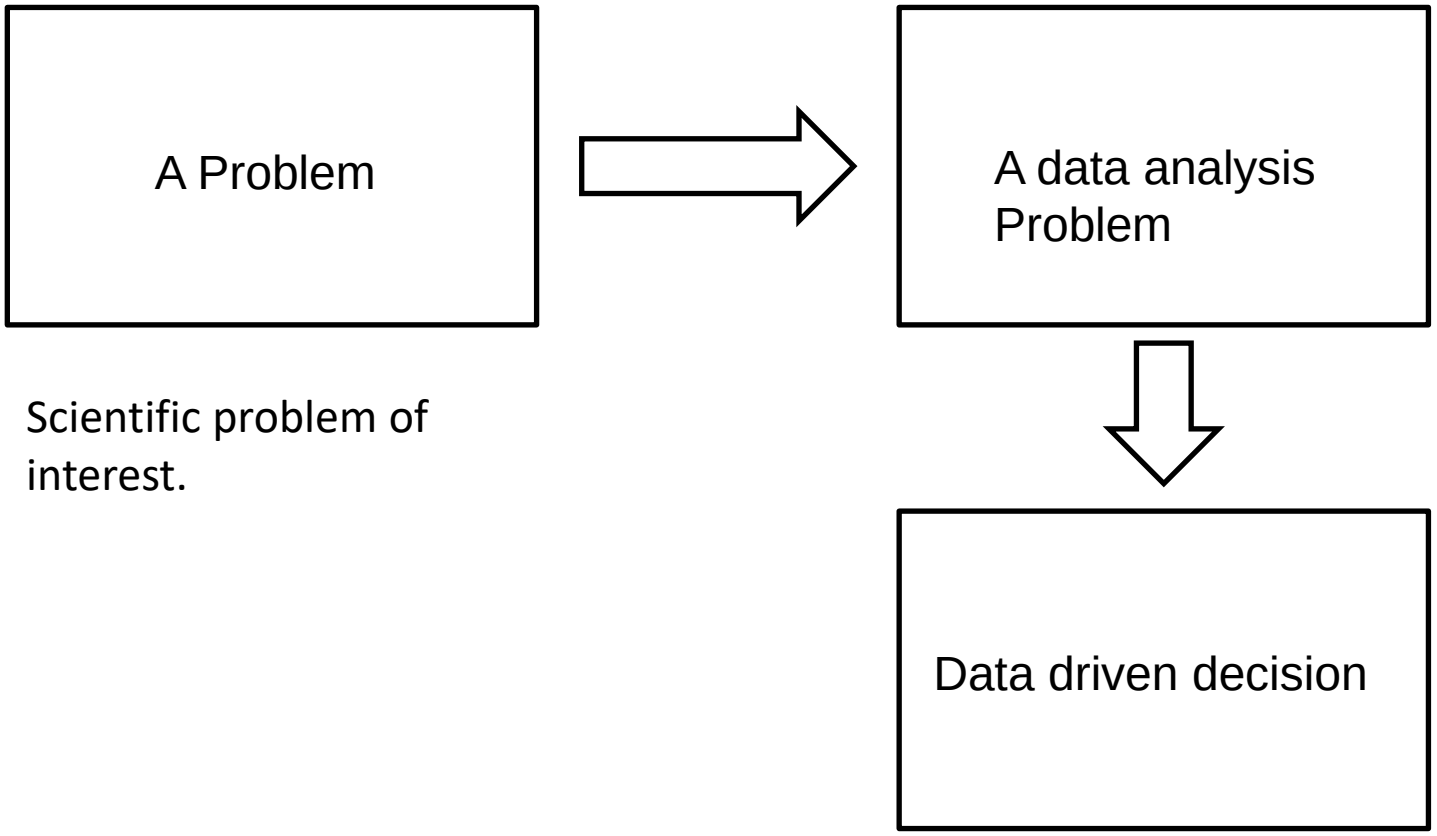
Short summary of the first session

Linear regression using R markdown

Slides 4-25: short summary of first lecture !!!

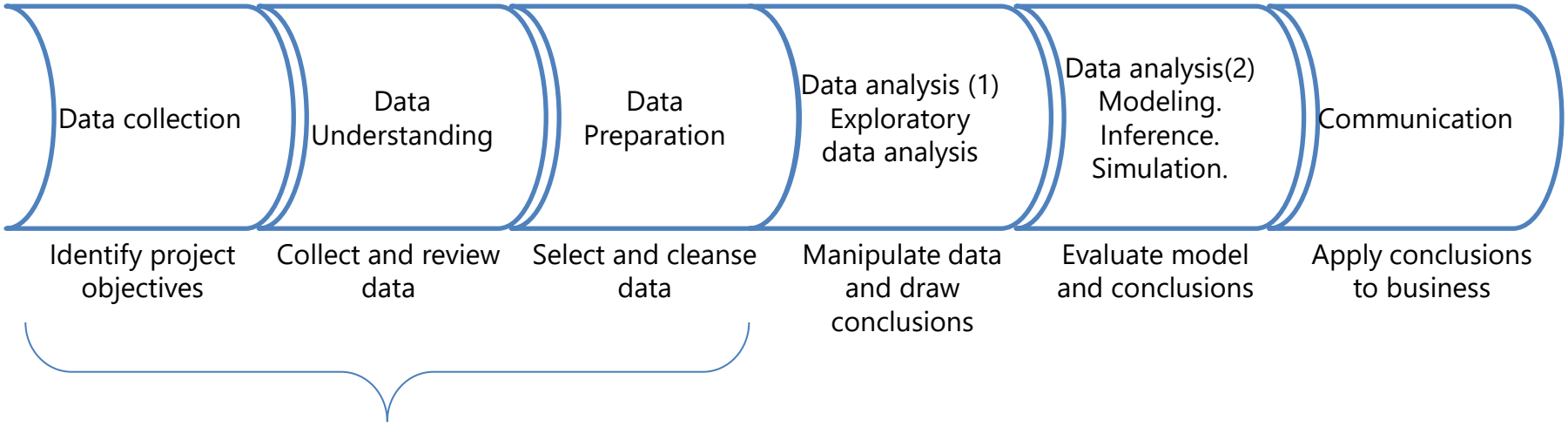
Steps in data analysis

- Data analysis approach in the course:



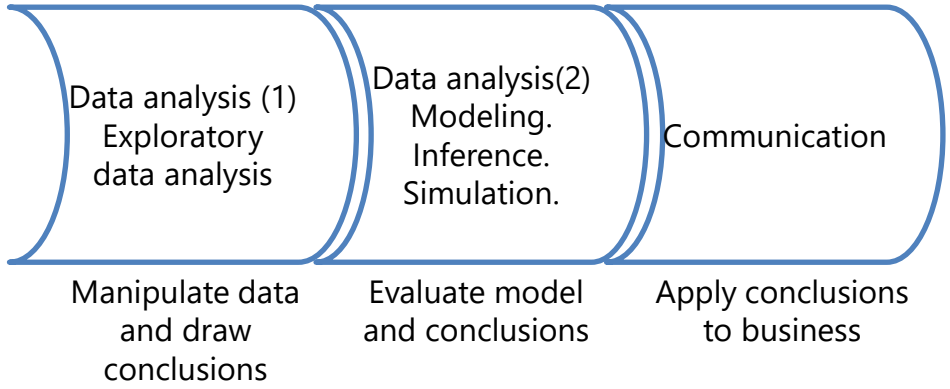
Steps in data analysis

- Steps related to data analysis:



Steps in data analysis

Modeling the association between the fuel consumption and the car's weight.



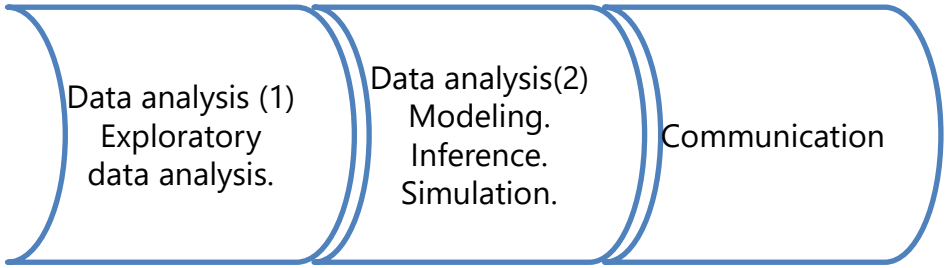
Scientific problem of interest:
how to model the association ?

Steps in data analysis

Modeling the association between the fuel consumption and the car's weight.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Methodology: simple linear regression.



Scatterplot

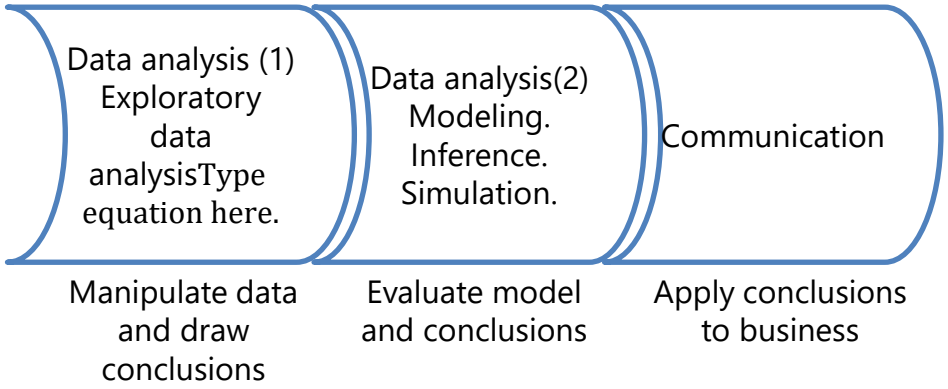
A linear regression model.

A report.

The solution

Steps in data analysis

Modeling the association between the fuel consumption and the car's weight.



$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Scatterplot.

Linear regression model.

A report.

Methodology: simple linear regression.

We "translate" the methodology to software usage

ggplot2()

lm()

R markdown to produce a HTML file.

We develop software to produce the solution and to communicate the solution

The `mtcars` data

Part 1

Analysis using basic R programming

R Program: er_prog1_SA_2024.R

The `mtcars` data in R

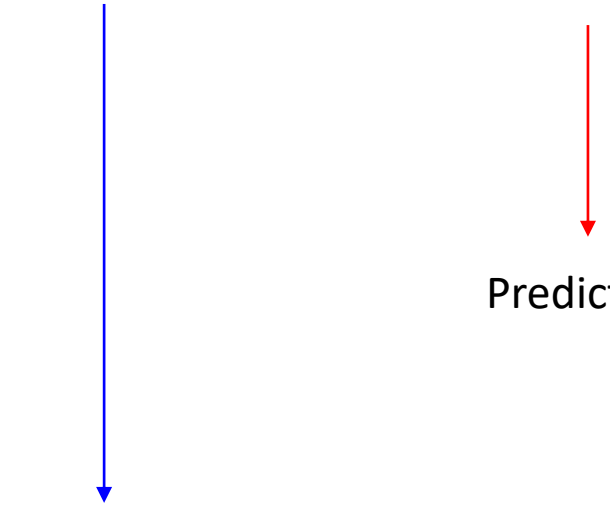
```
> dim(mtcars)
[1] 32 11
> names(mtcars)
 [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
[11] "carb"
```

The R object for the data: 32 observations and 11 variables.

Variables names:
mpg: mile per gallon – the response.
wt: car's weight – the predictor.

The `mtcars` data in R

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
>
```



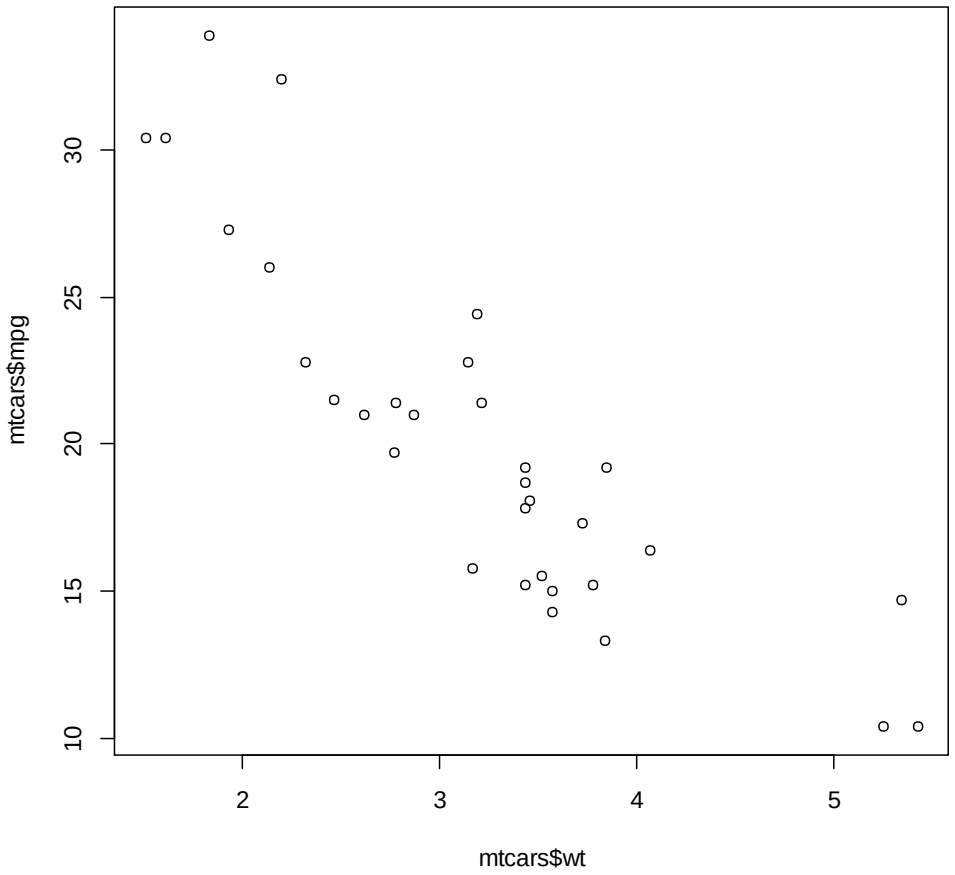
Response

Predictor

The `mtcars` data in R

```
> plot(mtcars$wt,mtcars$mpg)
> cor(mtcars$wt,mtcars$mpg)
[1] -0.8676594
```

- R functions to produce the plot and the correlation.
- Basic functions in R



Simple linear regression in R

```
lm(response~predictor)
```

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Simple linear regression in R: the `mtcars` data

```
>fit.lm=lm(mtcars$mpg~mtcars$wt)
```

The R object that contains the results of the fitted model.

$$mpg_i = \beta_0 + \beta_1 wt_i + \varepsilon_i$$

```
> summary(fit.lm)

Call:
lm(formula = mtcars$mpg ~ mtcars$wt)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
mtcars$wt    -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

- Output:
  - Parameter estimates etc.

R code for the analysis

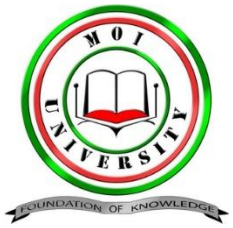
```
dim(mtcars)
names(mtcars)
head(mtcars)
plot(mtcars$wt,mtcars$mpg)
cor(mtcars$wt,mtcars$mpg)
fit.lm=lm(mtcars$mpg~mtcars$wt)
summary(fit.lm)
```

- Produce the plot.
- Calculate the correlation.
- Fit the model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Print the estimated model.

R Program: er_prog1_SA_2024.R



The `mtcars` data

Part 2

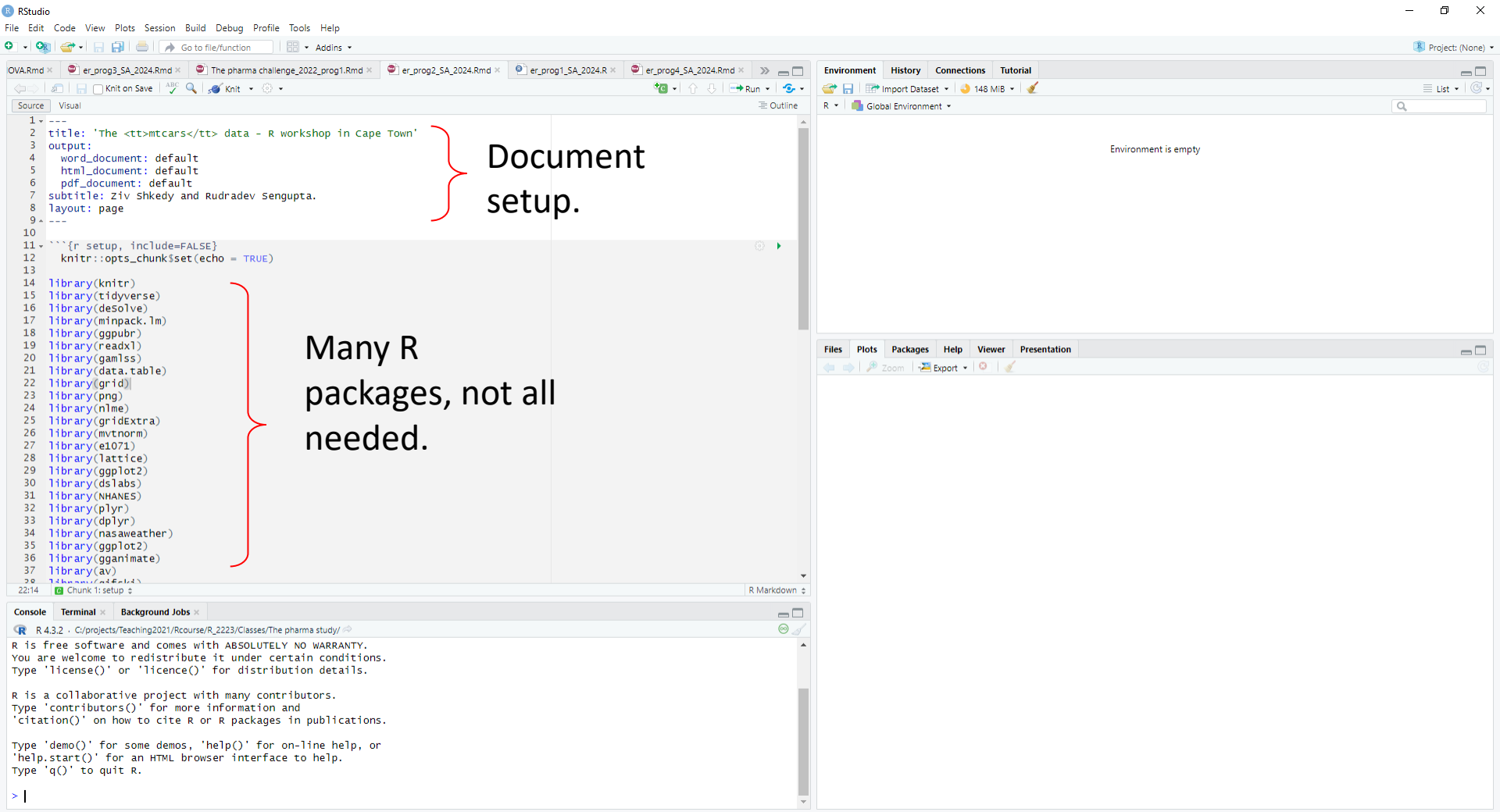
Analysis using basic Rmd programming

Rmd Program: er_prog2_SA_2024.Rmd

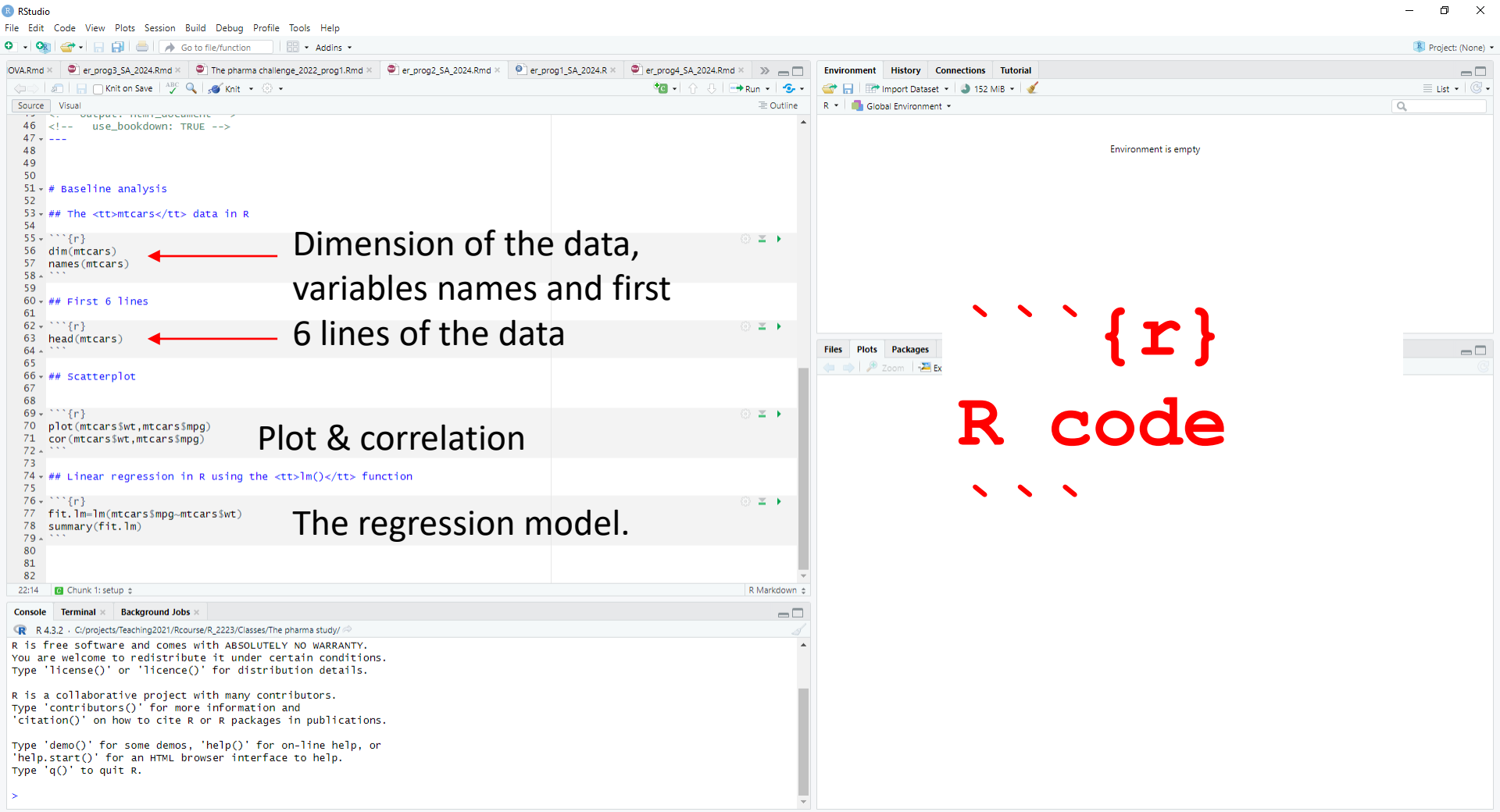
The output

- We run the same analysis as before.
- Use R markdown.
- Produce the possible output formats:
  - HTML.
  - PDF.
  - Word doc.

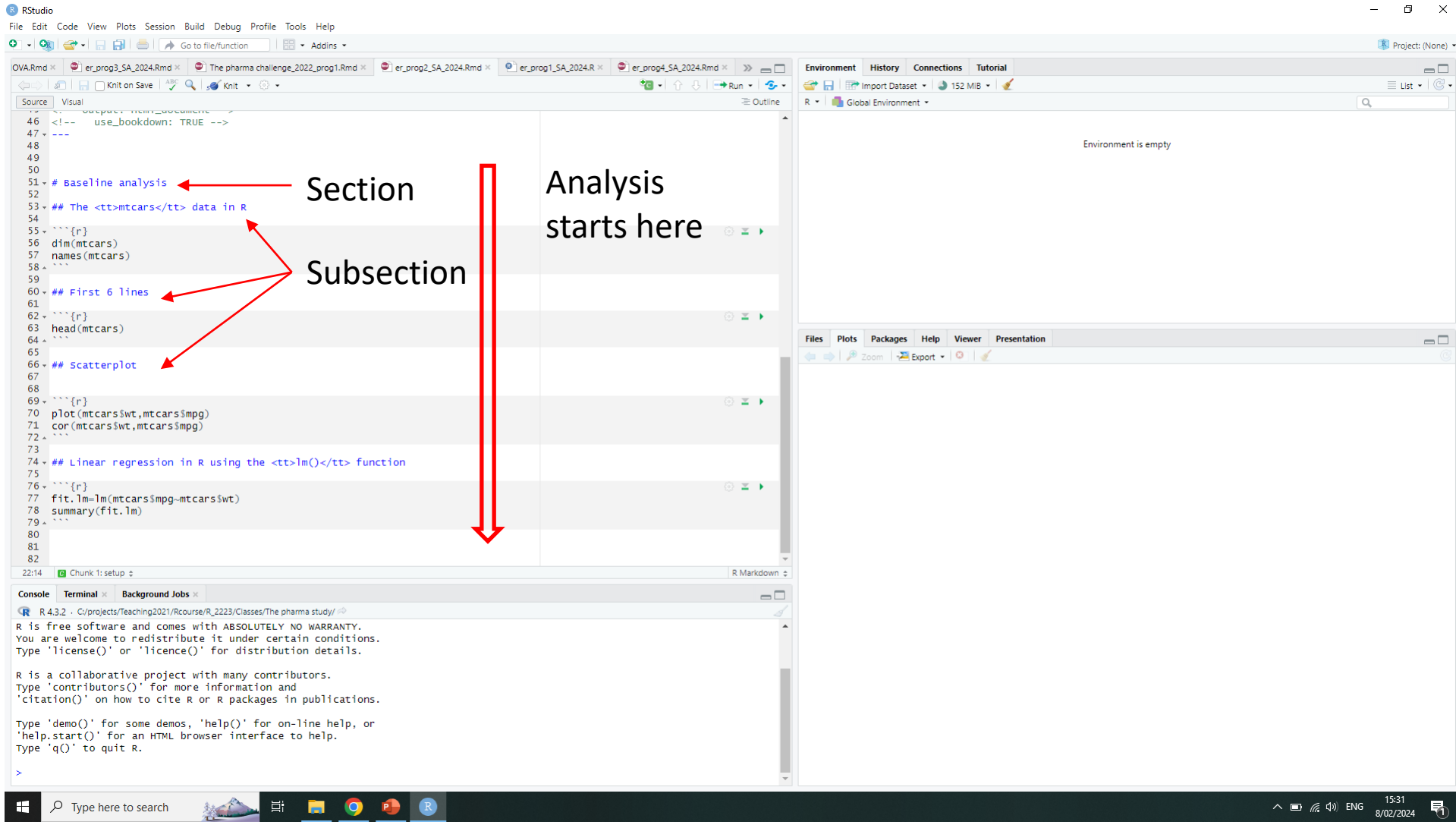
The Rmd program



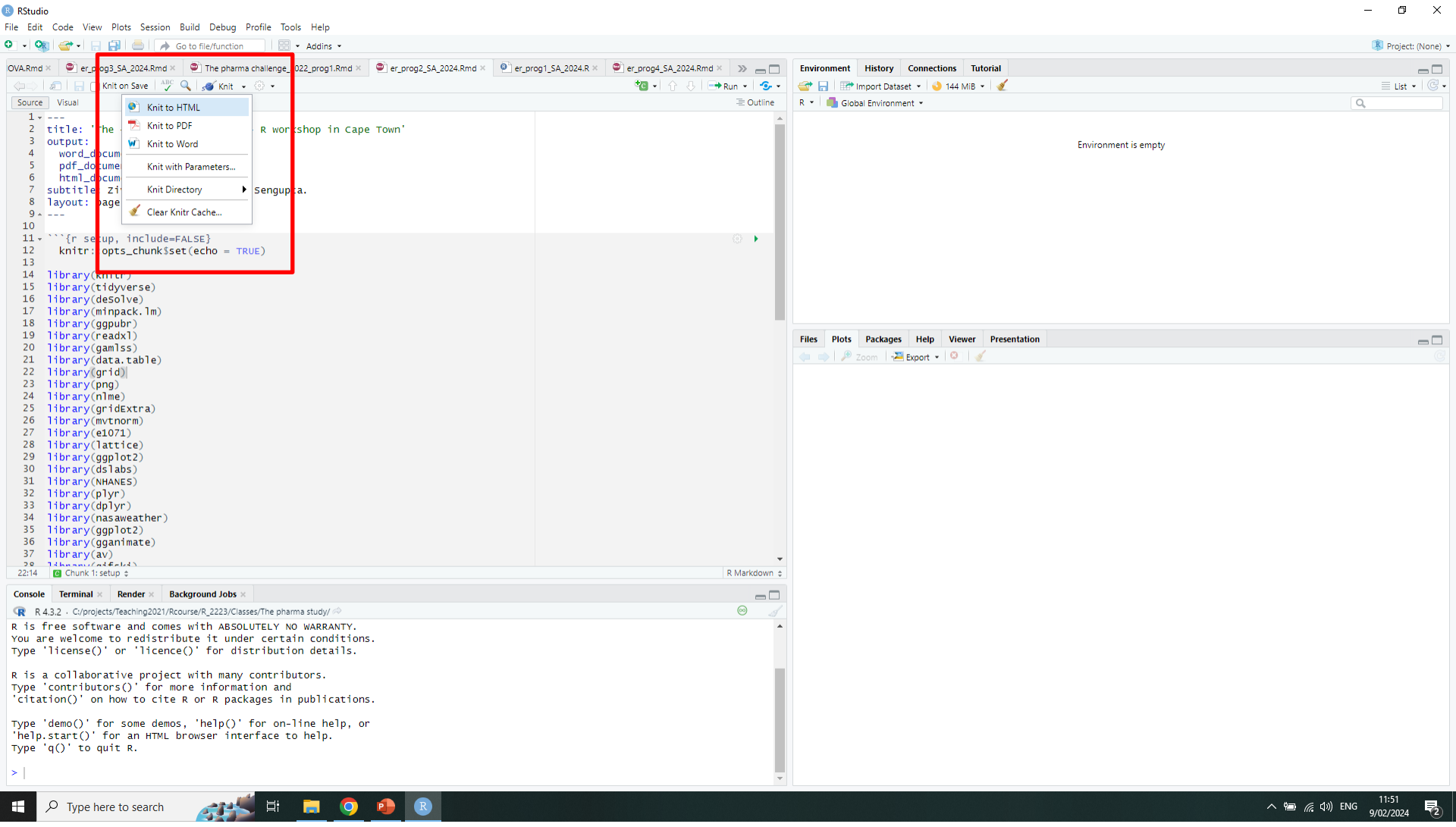
The Rmd program



The Rmd program



Choose the output



The HTML output

C:/projects/Teaching2021/Rcourse/R_2223/Classes/The pharma study/er_prog2_SA_2024.html

er_prog2_SA_2024.html | Open in Browser | Find | Publish

Section

Subsection

The `mtcars` data - R workshop in Cape Town

Ziv Shkedy and Rudradev Sengupta.

Baseline analysis

The `mtcars` data in R

```
dim(mtcars)
```

```
## [1] 32 11
```

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

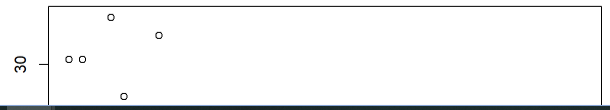
First 6 lines

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Scatterplot

```
plot(mtcars$wt,mtcars$mpg)
```



Type here to search

ENG 15:31 8/02/2024

The PDF output

er_prog2_SA_2024.pdf - Adobe Acrobat Reader (32-bit)

File Edit View Sign Window Help

Home Tools er_prog2_SA_2024.... Sign In

1 / 3 125%

The mtcars data - R workshop in Cape Town

Ziv Shkedy and Rudradev Sengupta.

Baseline analysis

The mtcars data in R

```
dim(mtcars)
```

```
## [1] 32 11
```

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

First 6 lines

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Scatterplot

```
plot(mtcars$wt,mtcars$mpg)
```

Search tools

Export PDF

Edit PDF

Create PDF

Comment

Combine Files

Organize Pages

Request E-signatures

Fill & Sign

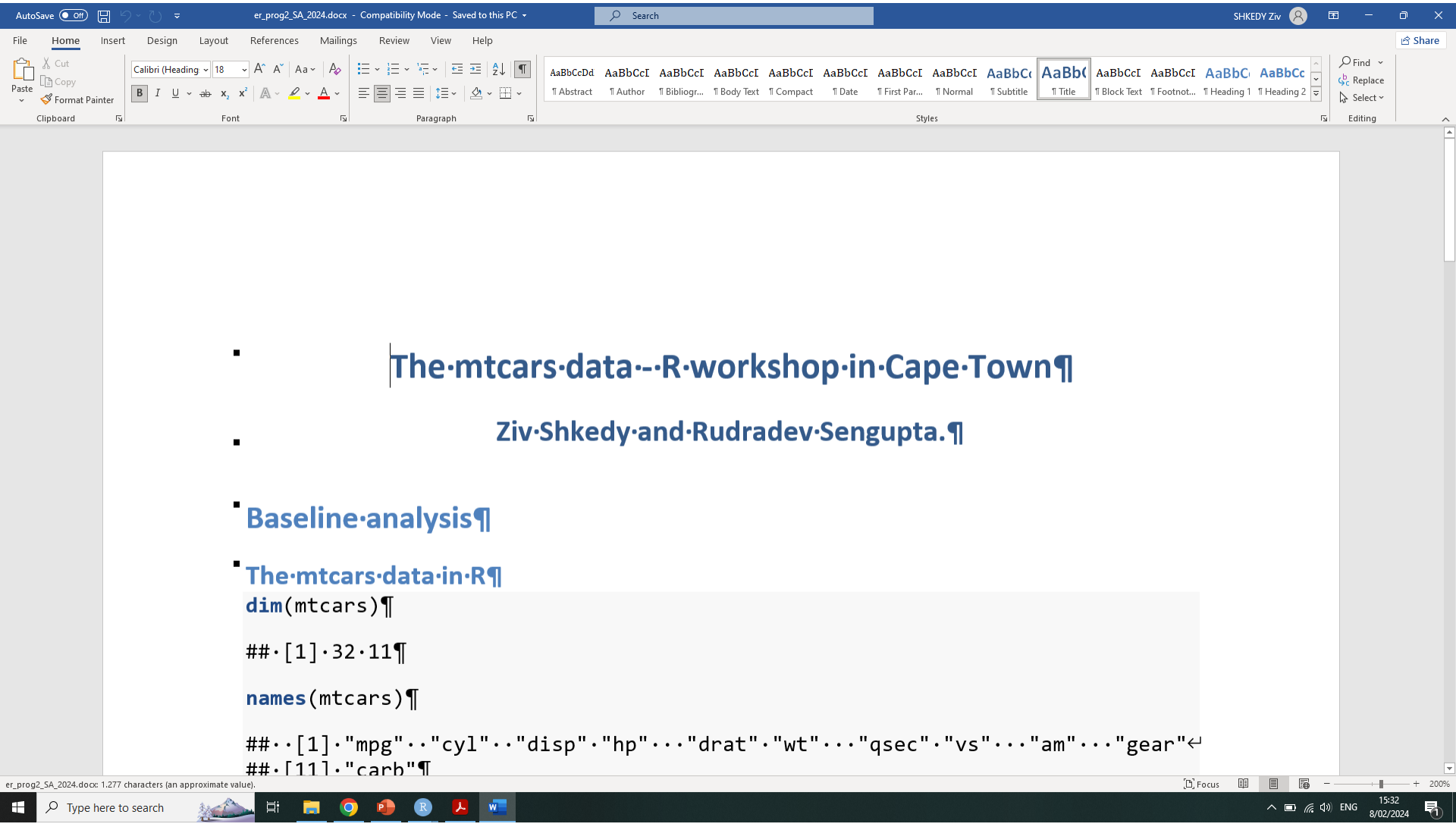
More Tools

Convert, edit and e-sign PDF forms & agreements

Free 7-Day Trial

Type here to search ENG 15:32 8/02/2024

The Word doc output



The `mtcars` data

Part 3

Advance HTML output

Rmd Program: er_prog3_SA_2024.Rmd

What do we cover in this part ?

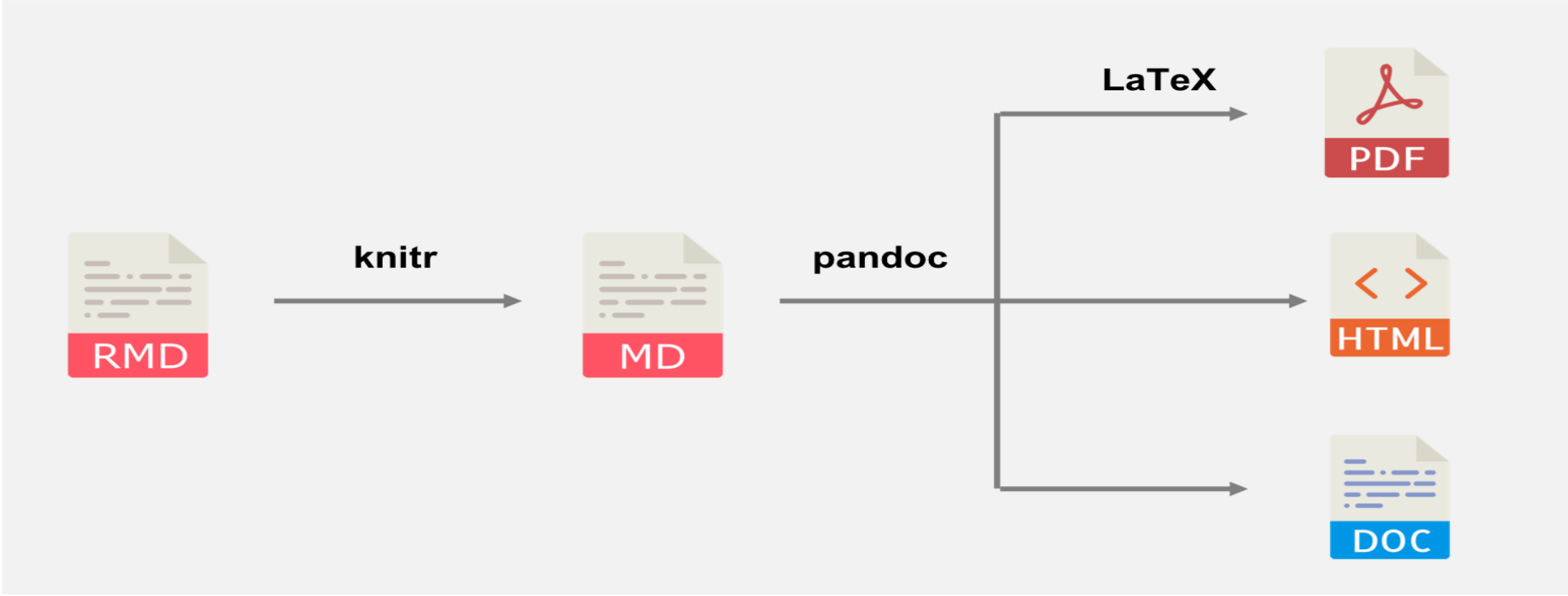
- We use the same example: the `mtcars` data and simple linear regression.
- How to produce a HTML file for the analysis we conducted before ?
- How to use the HTML file as a content for a website ?

Reproducible Research

- Aim: create an output in a different (highest) quality.
- Can be used to communicate the analysis' results with other people in the organization.
- Not all potential readers are interested on "how to do the analysis".
- We DO NOT aim to develop a report for the analysis but to provide a document from which the results can be seen and discuss by different people in the organization.

The Rmd file

- Analyses ⇨ high quality report.
- Rmarkdown – Different dynamic and statistic formats (html, pdf, word, books, dashboard, e.t.c).



Knitr: A General-Purpose Package for Dynamic Report Generation in r. https://yihui.org/knitr/.

The HTML output

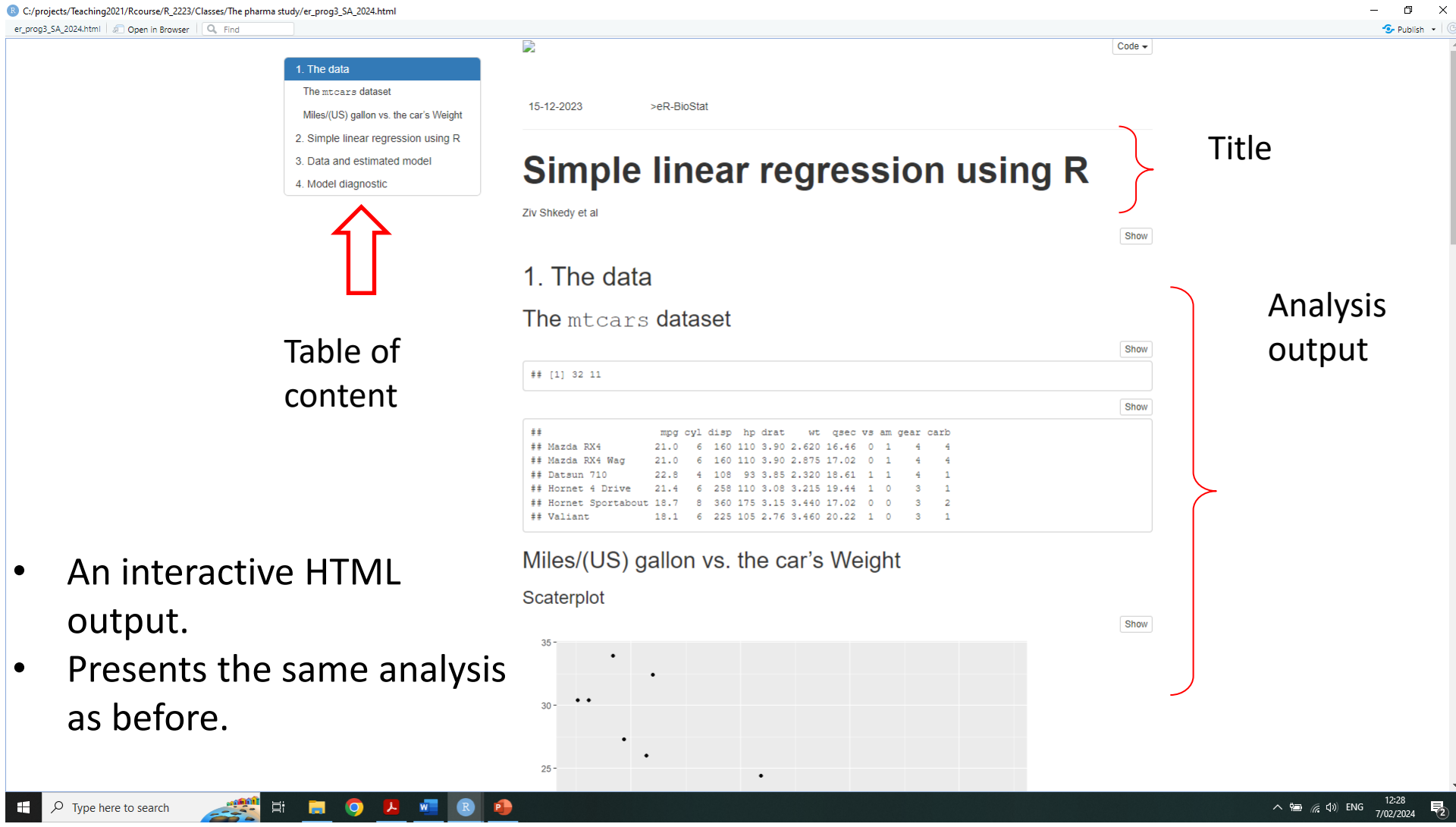


Table of content

Title

Analysis output

- An interactive HTML output.
- Presents the same analysis as before.

Part 3.1:
How to set up the HTML file ?

The Rmd file

- We use Rmd file to
  - Conduct the analysis.
  - Set up the document.
- We use html file to
  - Present & communicate the result.

Set up the document

Set up the HTML document:

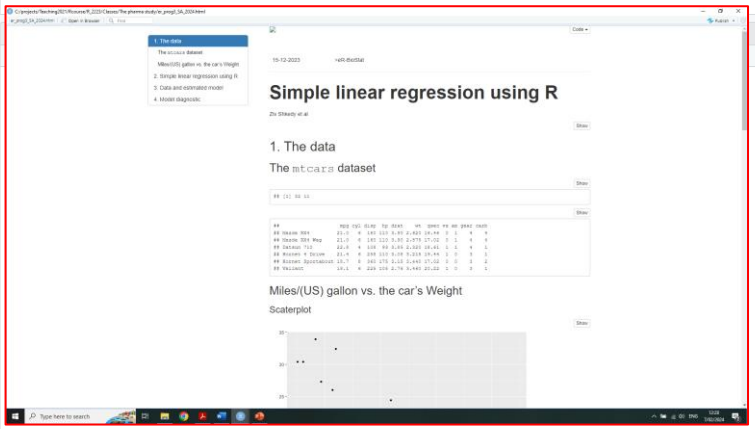
document_2

```
---
output:
  bookdown::html_document2:
    toc: TRUE
    toc_float: TRUE
    toc_depth: 2
    number_sections: no
    css: ./lib/stylesArial.css
    code_folding: hide

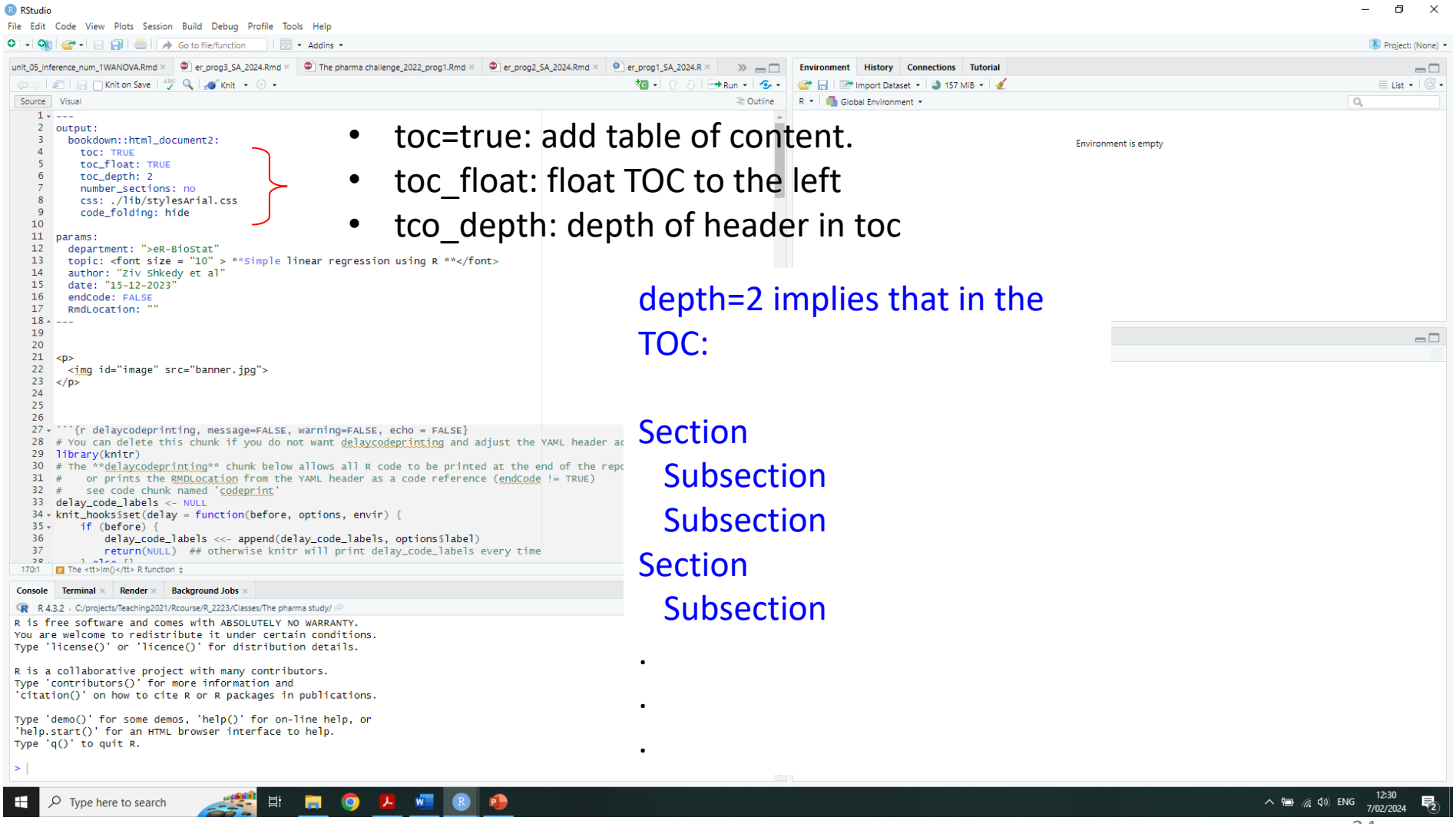
params:
  department: ">eR-BioStat"
  topic: <font size = "10" > **Simple linear regression using R **</font>
  author: "Ziv Shkedy et al"
  date: "15-12-2023"
  endCode: FALSE
  RmdLocation: ""
---


<p>
  <img id="image" src="banner.jpg">
</p>


```{r delaycodeprinting, message=FALSE, warning=FALSE, echo = FALSE}
# You can delete this chunk if you do not want delaycodeprinting and adjust the YAML header accordingly
library(knitr)
# The **delaycodeprinting** chunk below allows all R code to be printed at the end of the report (endCode = TRUE)
#    or prints the RMDLocation from the YAML header as a code reference (endCode != TRUE)
#    see code chunk named 'codeprint'
delay_code_labels <- NULL
knit_hooks$set(delay = function(before, options, envir) {
    if (before) {
        delay_code_labels <<- append(delay_code_labels, options$label)
        return(NULL)  ## otherwise knitr will print delay_code_labels every time
```


```

Set up the document



- toc=true: add table of content.
- toc_float: float TOC to the left
- tco_depth: depth of header in toc

depth=2 implies that in the TOC:

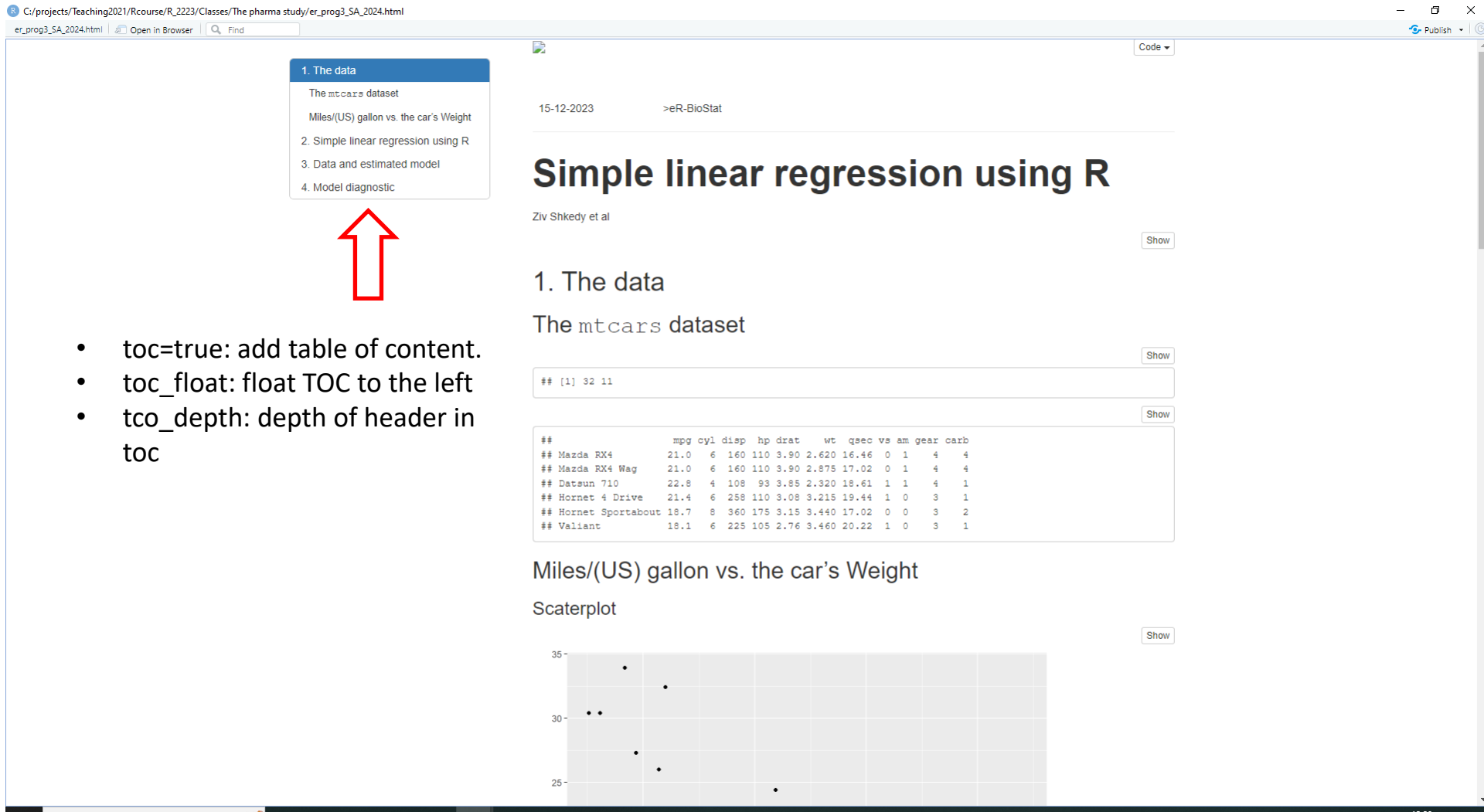
Section
  Subsection
  Subsection
Section
  Subsection

.

.

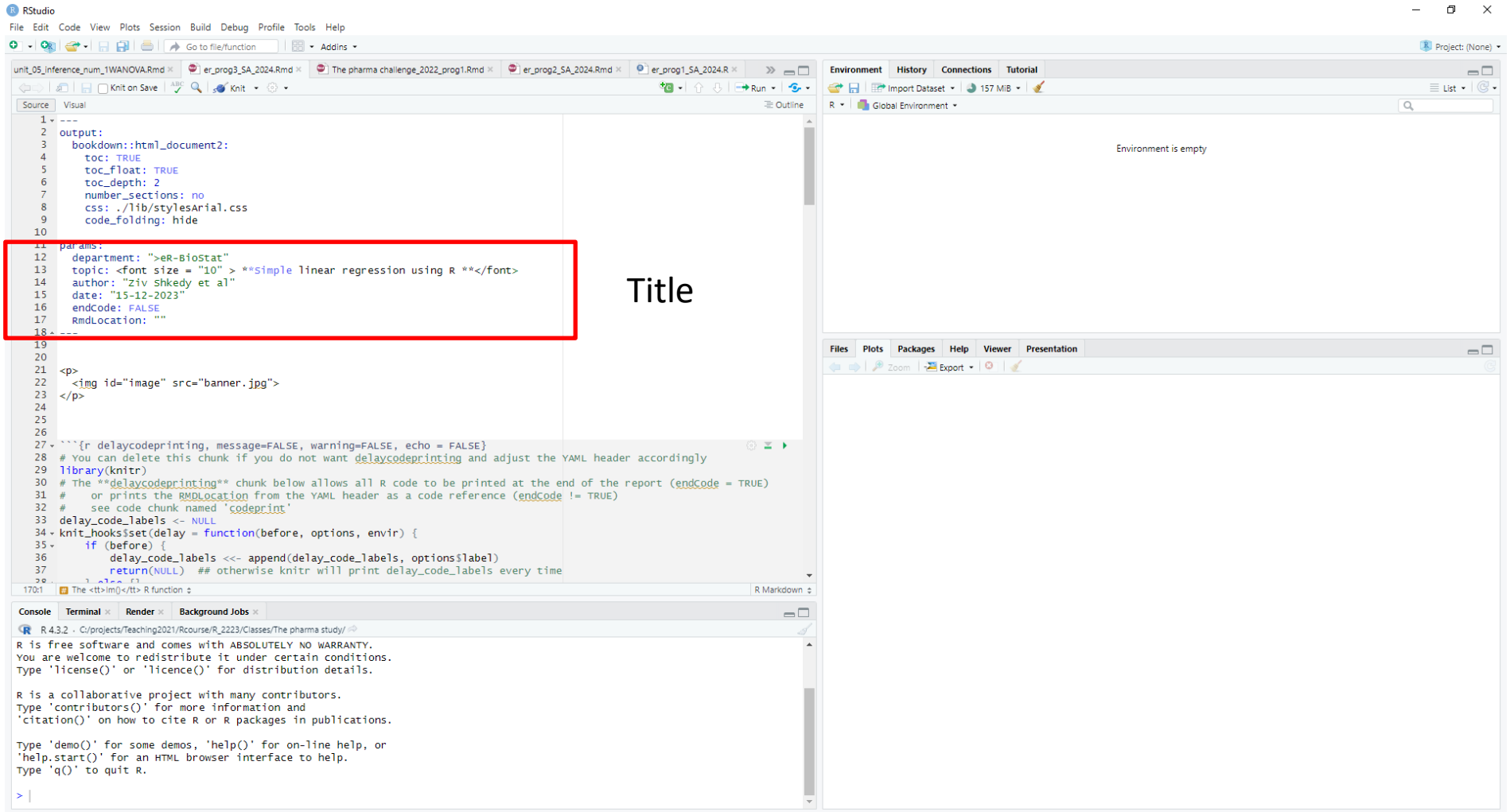
.

The HTML file



- toc=true: add table of content.
- toc_float: float TOC to the left
- tco_depth: depth of header in toc

Titles, authors and dates



Title

Titles, authors and dates

C:/projects/Teaching2021/Rcourse/R_2223/Classes/The pharma study/er_prog3_SA_2024.html

er_prog3_SA_2024.html | Open in Browser | Find | Publish

1. The data
   - The mtcars dataset
   - Miles/(US) gallon vs. the car's Weight
2. Simple linear regression using R
3. Data and estimated model
4. Model diagnostic

Code

15-12-2023 >eR-BioStat

Simple linear regression using R

Ziv Shkedy et al

Show

Title

1. The data

The mtcars dataset

Show

```
## [1] 32 11
```

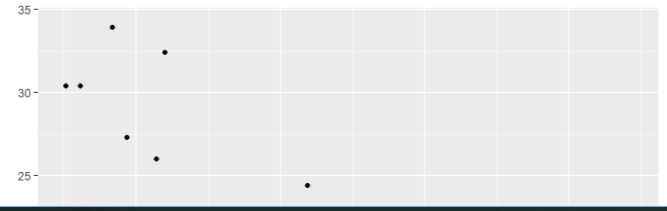
Show

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

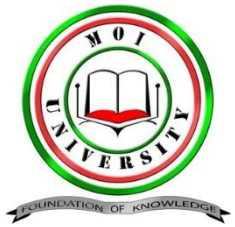
Miles/(US) gallon vs. the car's Weight

Scaterplot

Show

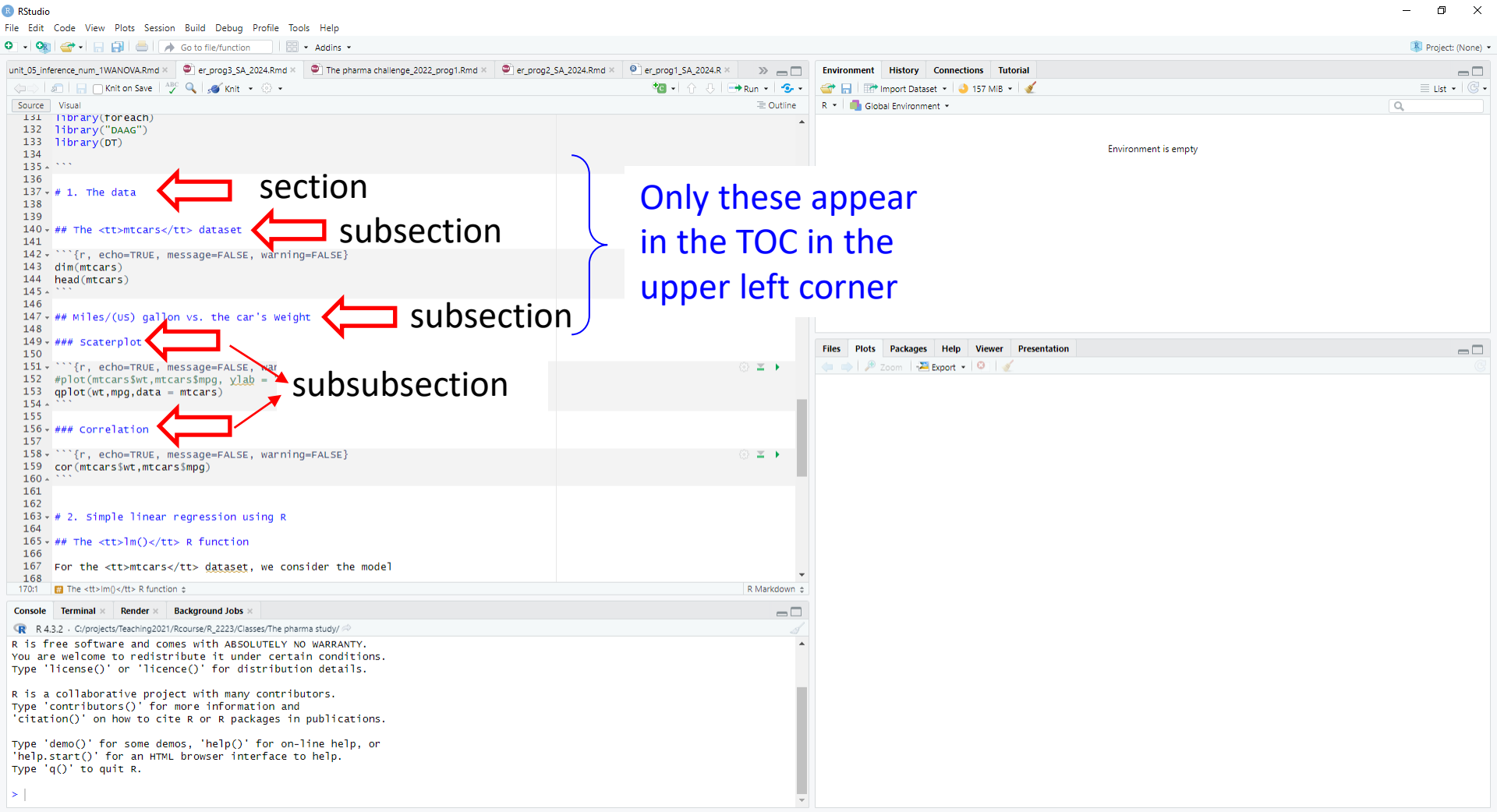


Type here to search ENG 12:28 7/02/2024



Part 3.2:
The HTML file and the Rmd program in details.

Section, subsection, subsubsection



Section, subsection, subsubsection

C:/projects/Teaching2021/Rcourse/R_2223/Classes/The pharma study/er_prog3_SA_2024.html

er_prog3_SA_2024.html | Open in Browser | Find | Publish

Depth=2
Only sections
and subsections

Ziv Shkedy et al

Show

1. The data section

The mtcars dataset subsection

Show

```
## [1] 32 11
```

Show

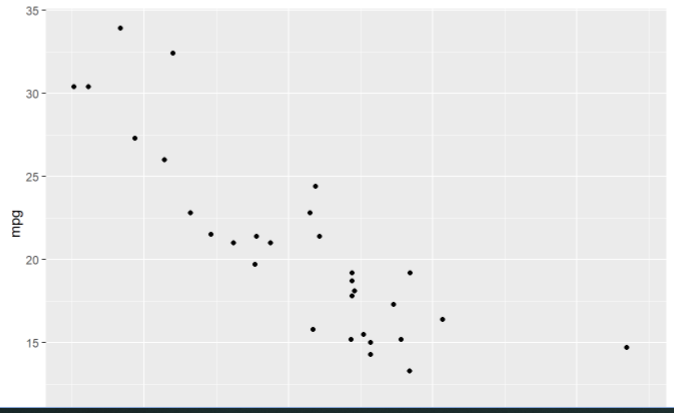
```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Miles/(US) gallon vs. the car's Weight subsection

Since depth=2, the subsubsection will not appears in the TOC

Scaterplot subsubsection

Show



Type here to search

ENG 12:37 7/02/2024

Analysis code

- The same as before.
- In addition to the code, we can add free text in the Rmd file.

The code for the analysis

C:/projects/Teaching2021/Rcourse/R_2223/Classes/The pharma study/er_prog3_SA_2024.html

er_prog3_SA_2024.html | Open in Browser | Find | Publish

Code
- Show All Code
- Hide All Code

15-12-2023 >eR-BioStat

Simple linear regression using R

Ziv Shkedy et al

1. The data

The mtcars dataset

Show

```
## [1] 32 11
```

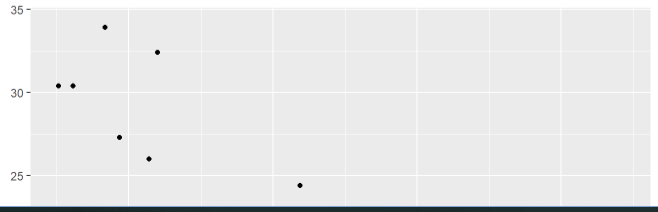
Show

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Miles/(US) gallon vs. the car's Weight

Scaterplot

Show



We can choose if we want to show the code or to hide the code.

Type here to search ENG 12:40 7/02/2024

Reading the external file

C:/projects/Teaching2021/Rcourse/R_2223/Classes/The pharma study/er_prog3_SA_2024.html

er_prog3_SA_2024.html | Open in Browser | Find | Publish

Code

15-12-2023 >eR-BioStat

Simple linear regression using R

Ziv Shkedy et al

Show

1. The data

The mtcars dataset

Show

```
## [1] 32 11
```

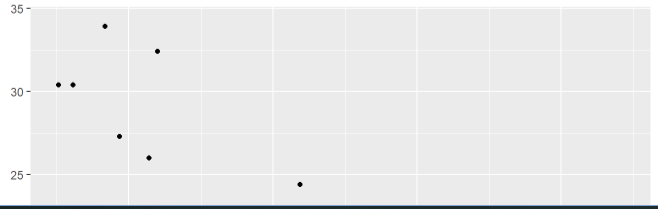
Show

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Miles/(US) gallon vs. the car's Weight

Scaterplot

Show

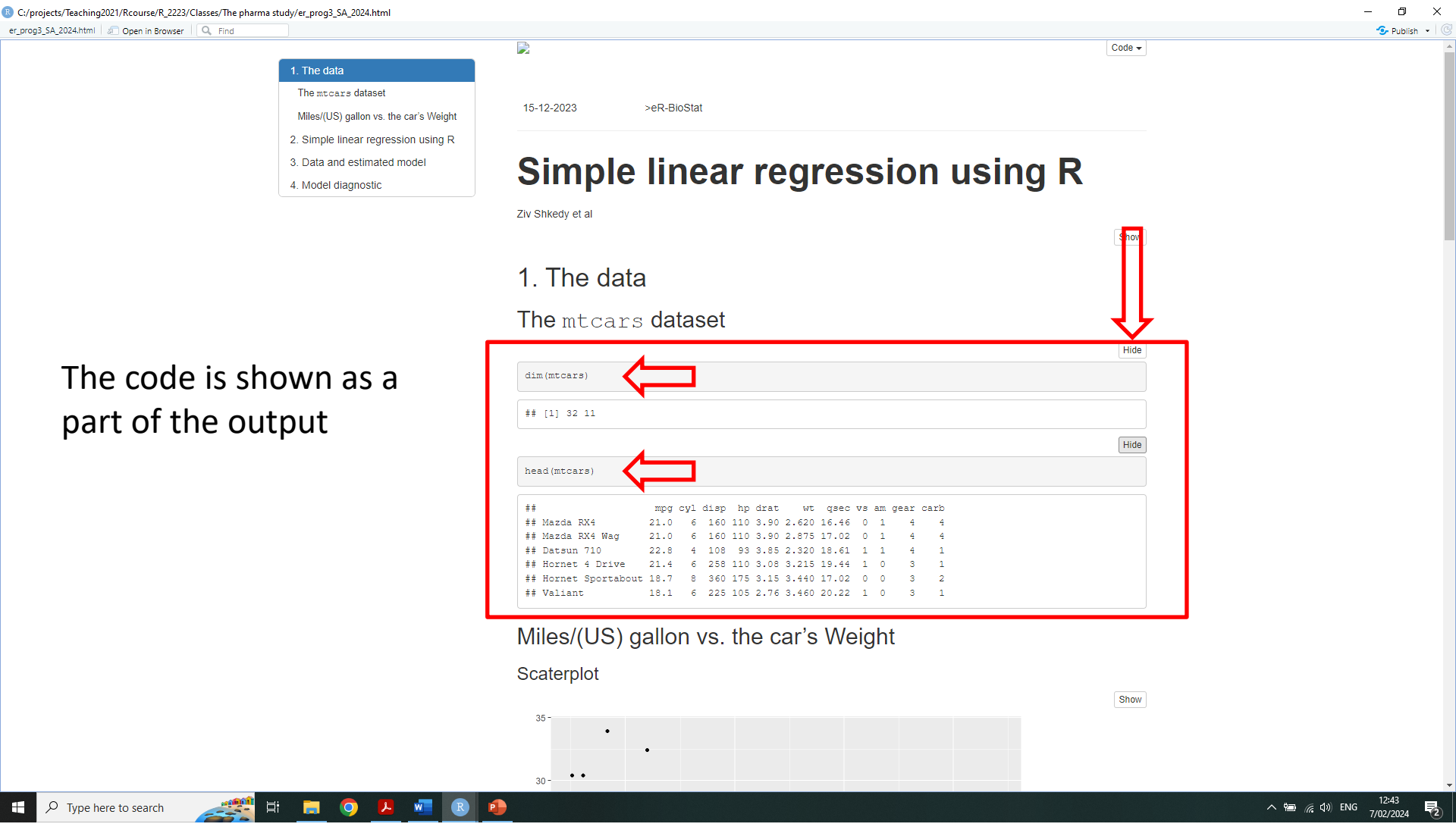


The code is not shown as a part of the output.

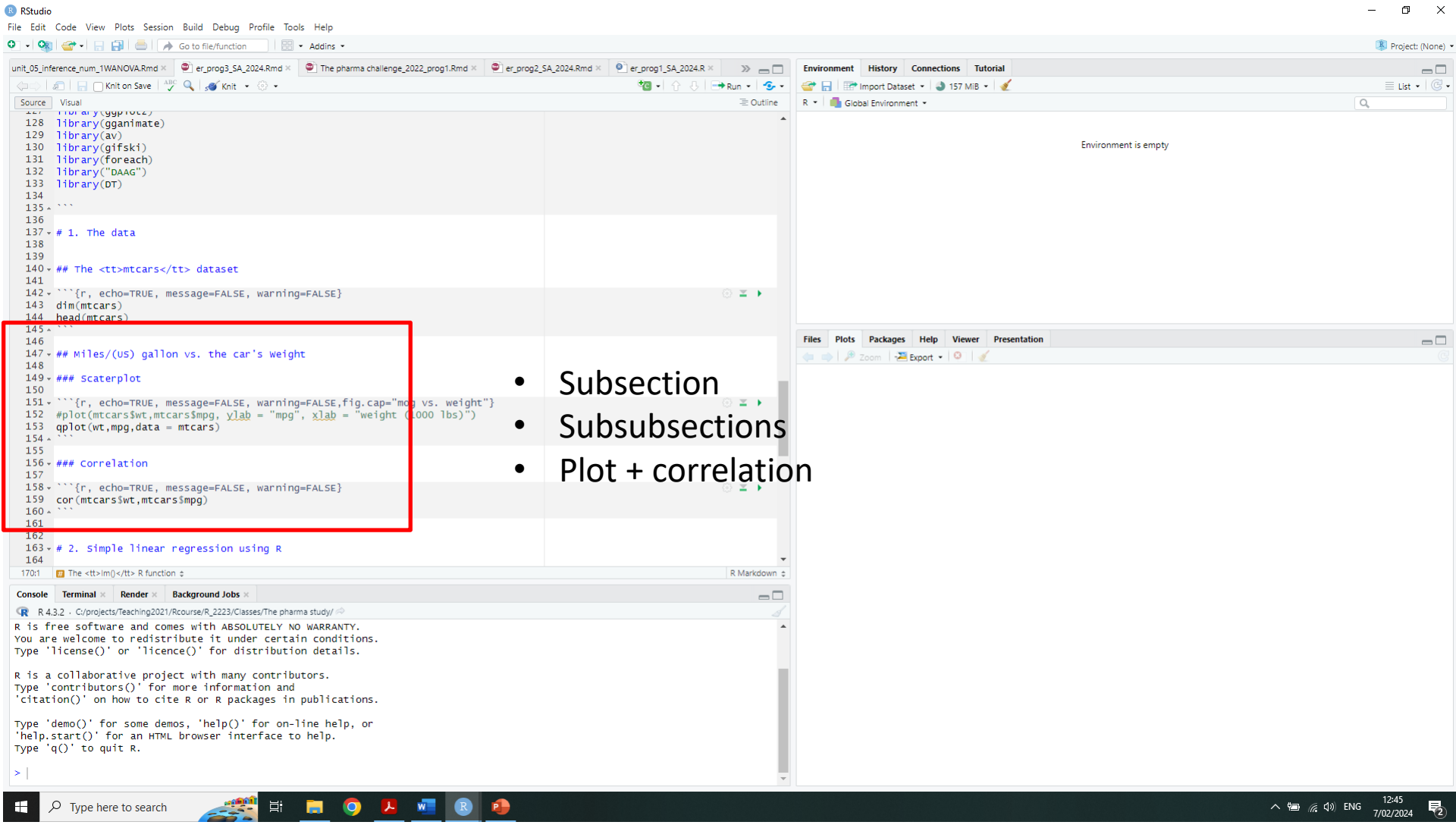
Type here to search ENG 12:42 7/02/2024

The HTML file

The code is shown as a part of the output



Code in the Rmd file



- Subsection
- Subsubsections
- Plot + correlation

The output in the HTML file

C:/projects/Teaching2021/Rcourse/R_2223/Classes/The pharma study/er_prog3_SA_2024.html

er_prog3_SA_2024.html | Open in Browser | Find | Publish

1. The data
   - The mtcars dataset
   - Miles/(US) gallon vs. the car's Weight
2. Simple linear regression using R
3. Data and estimated model
4. Model diagnostic

```
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Miles/(US) gallon vs. the car's Weight

Scaterplot

Show

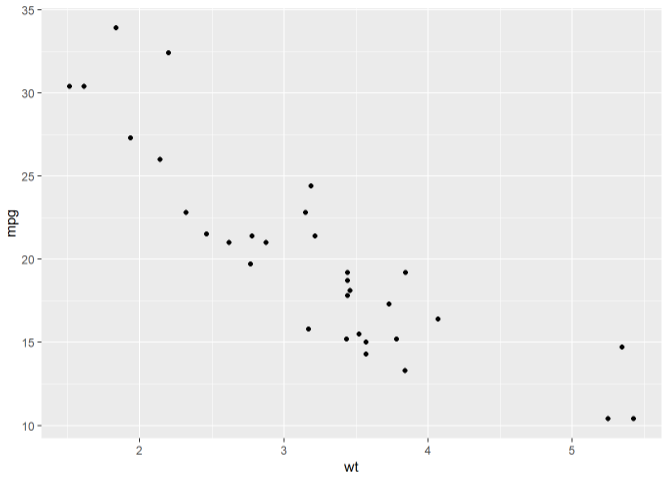


Figure 1: mog vs. weight

Correlation

Show

```
## [1] -0.8676594
```

2. Simple linear regression using R

The `lm()` R function

For the `mtcars` dataset, we consider the model

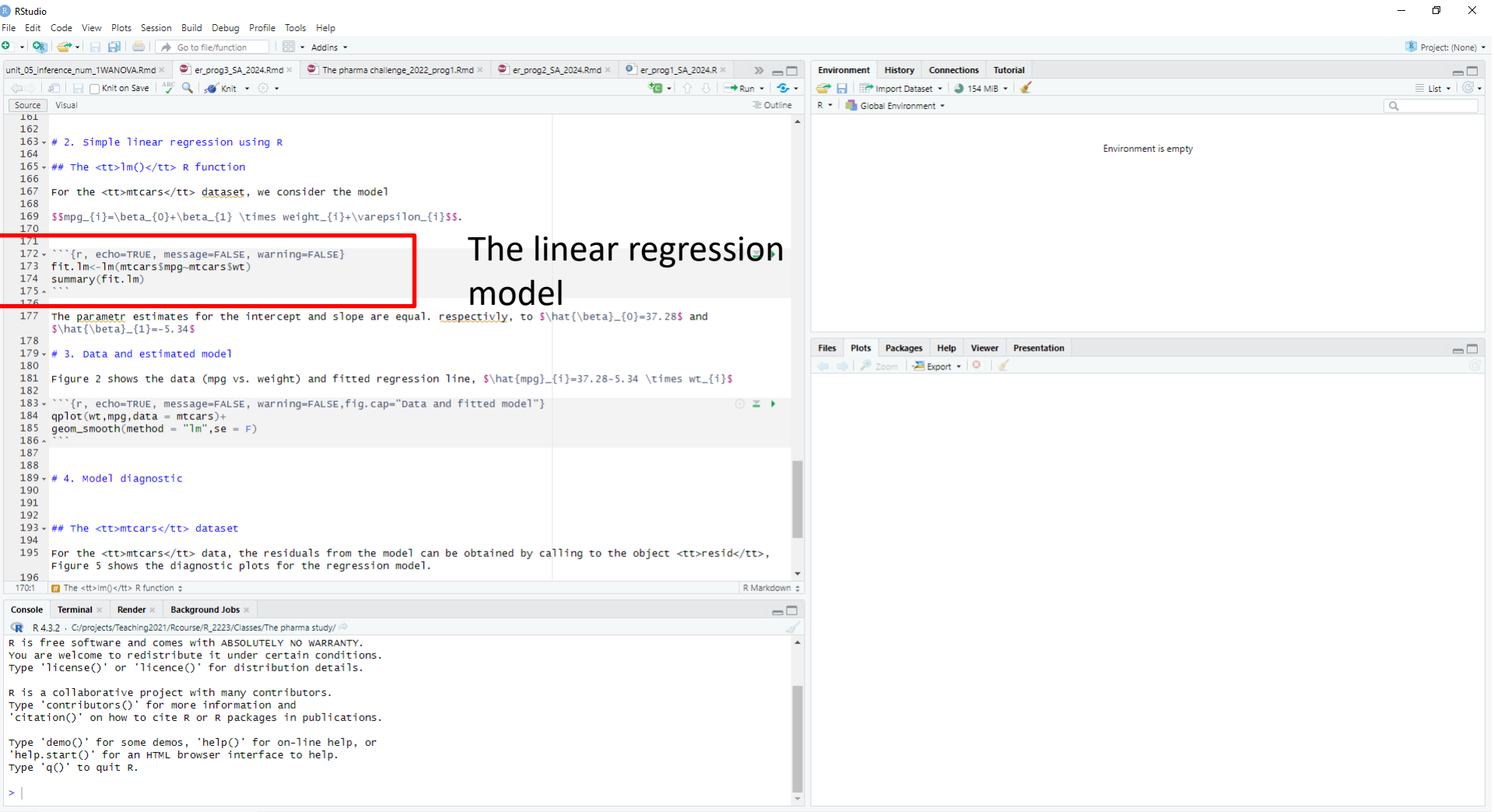
$$mpg_i = \beta_0 + \beta_1 \times weight_i + \varepsilon_i$$

- Subsection
- Subsubsections
- Plot + correlation

Type here to search

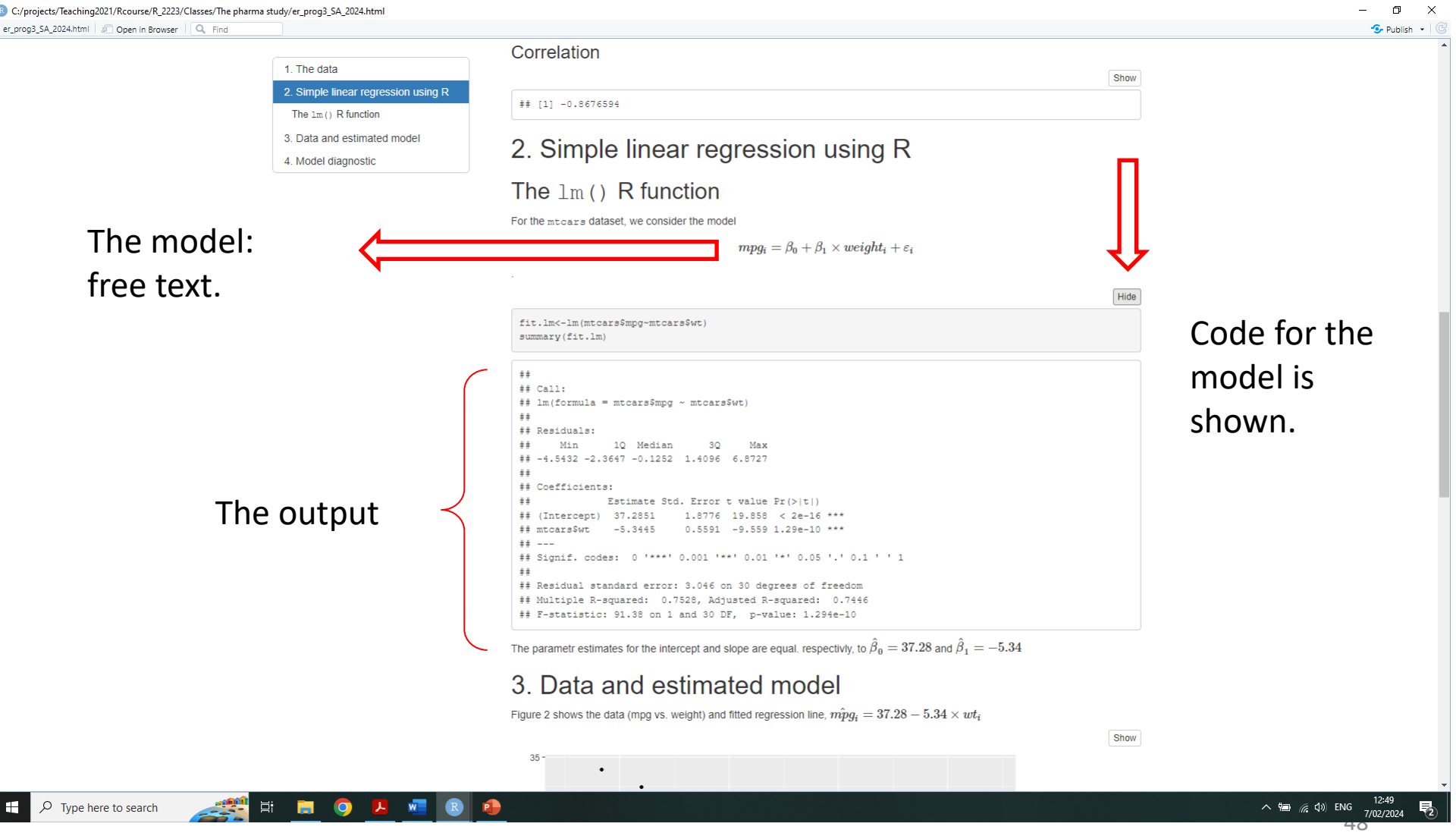
ENG 12:46 7/02/2024

Code in the Rmd file



The linear regression model

The output in the HTML file

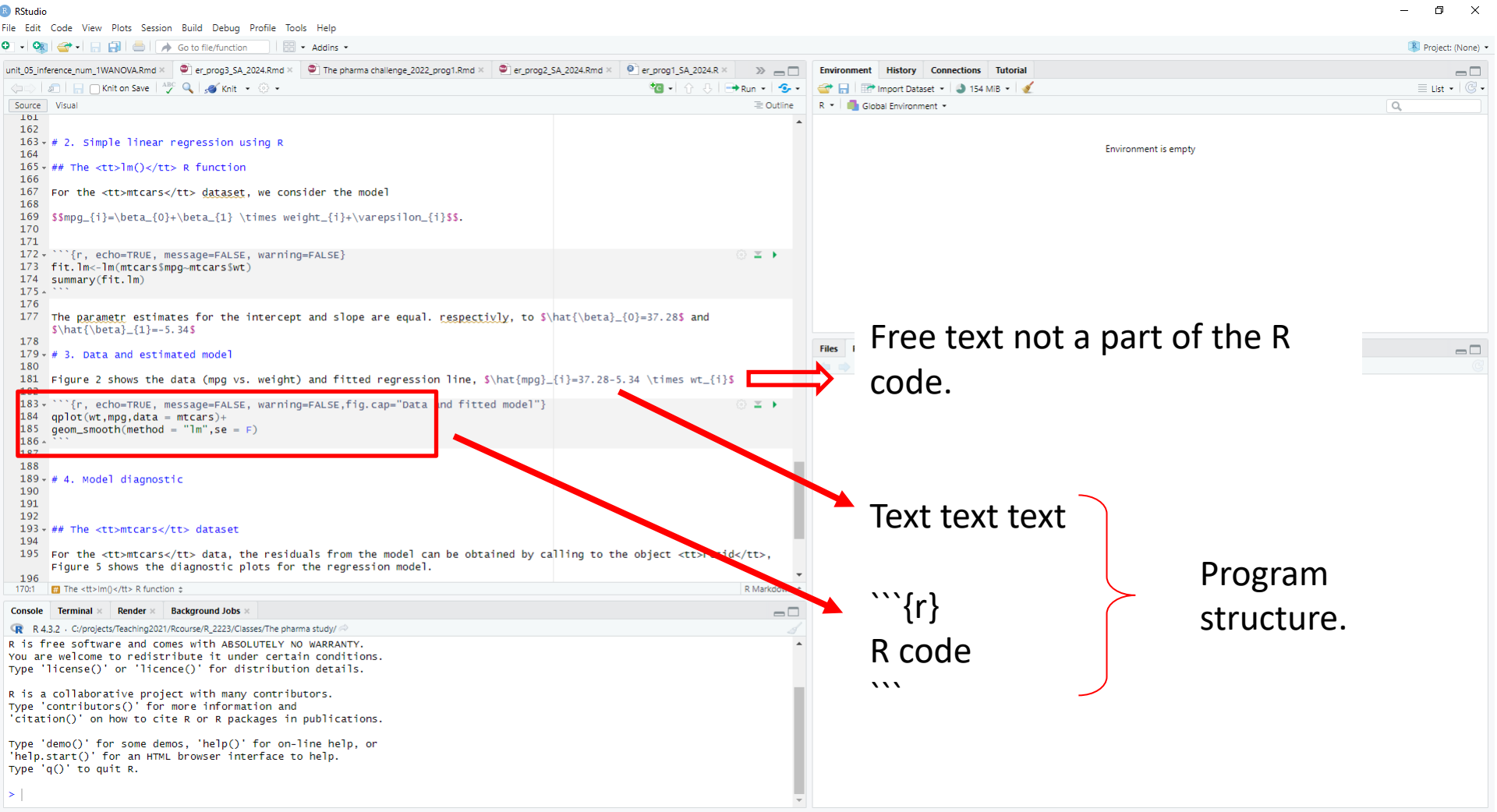


The model: free text.

The output

Code for the model is shown.

Code in the Rmd file



The output in the HTML file

C:/projects/Teaching2021/Rcourse/R_2223/Classes/The pharma study/er_prog3_SA_2024.html

er_prog3_SA_2024.html | Open in Browser | Find | Publish

1. The data
2. Simple linear regression using R
3. Data and estimated model
4. Model diagnostic

```
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

The parametr estimates for the intercept and slope are equal. respectivly, to $\hat{\beta}_0 = 37.28$ and $\hat{\beta}_1 = -5.34$

3. Data and estimated model

Free text

Figure 2 shows the data (mpg vs. weight) and fitted regression line, $\hat{mpg}_i = 37.28 - 5.34 \times wt_i$

Show

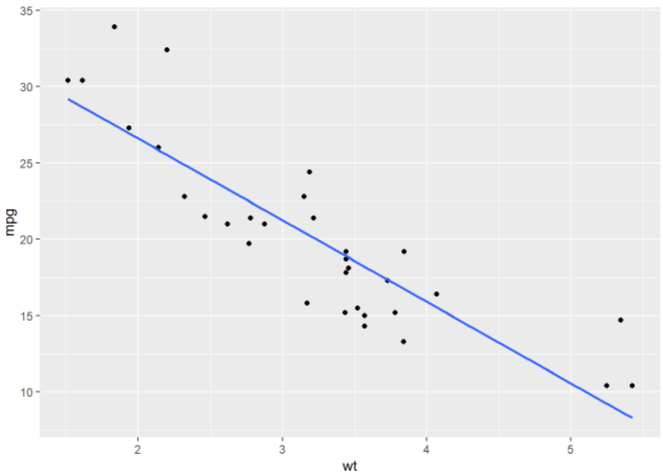


Figure 2: Data and fitted model

4. Model diagnostic

The `mtcars` dataset

For the `mtcars` data, the residuals from the model can be obtained by calling to the object `resid`, Figure 5 shows the diagnostic plots for the regression model.

Show

```
##          1          2          3          4          5          6          7
## -2.2826106 -0.9197704 -2.0859521  1.2973499 -0.2001440 -0.6932545 -3.9053627
```

Type here to search

ENG 12:50 7/02/2024

Short discussion

- R Studio + R markdown:
- Easy to use.
- Text + code.
- Output:
  - Standard: HTML, PDF, DOC.
  - Advanced: HTML.

- So far: simple analysis that produce an output.
- How do we create a course ?

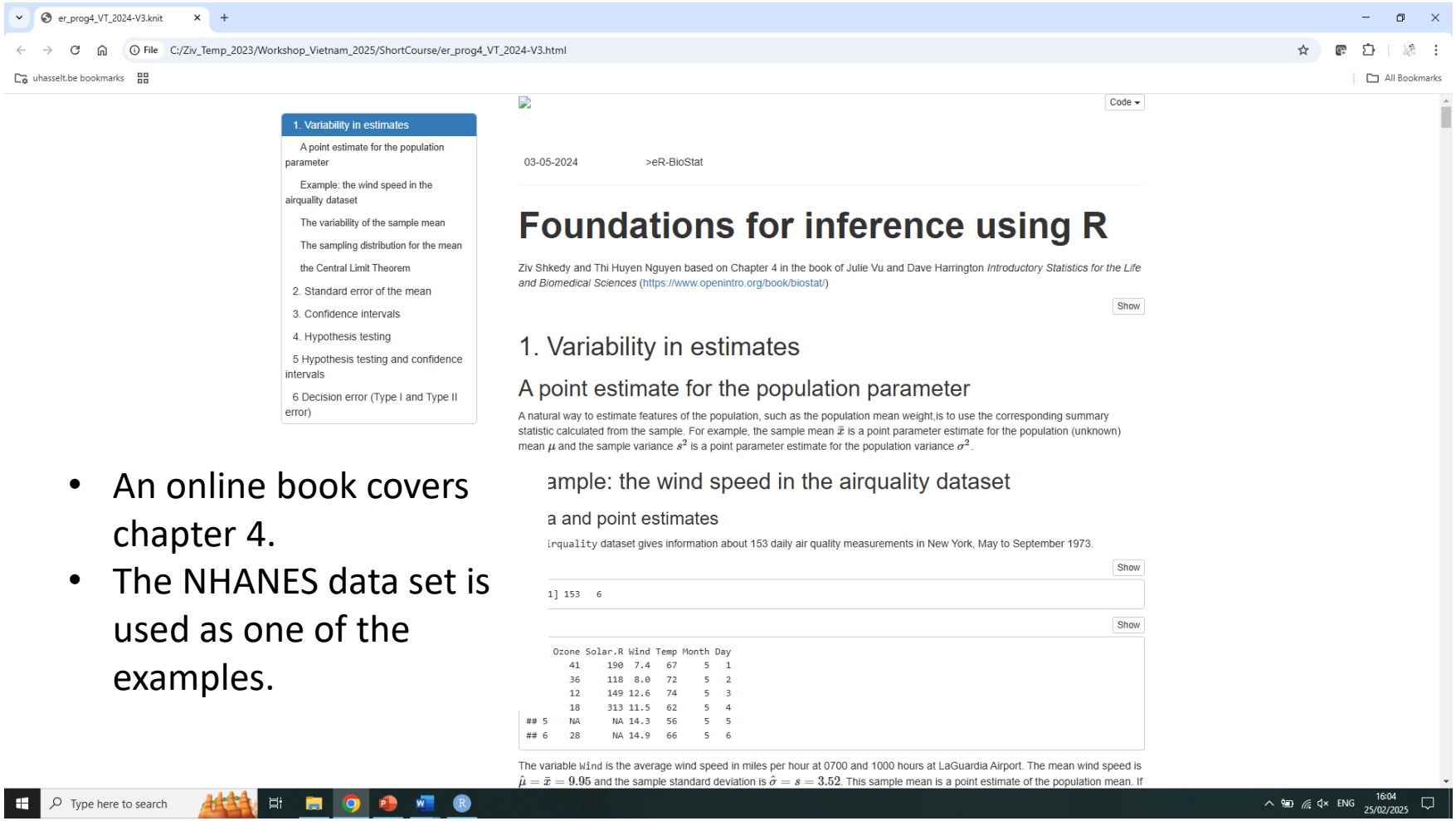
Part 4

Case study: The NHANES dataset

The number of sleep hours per night

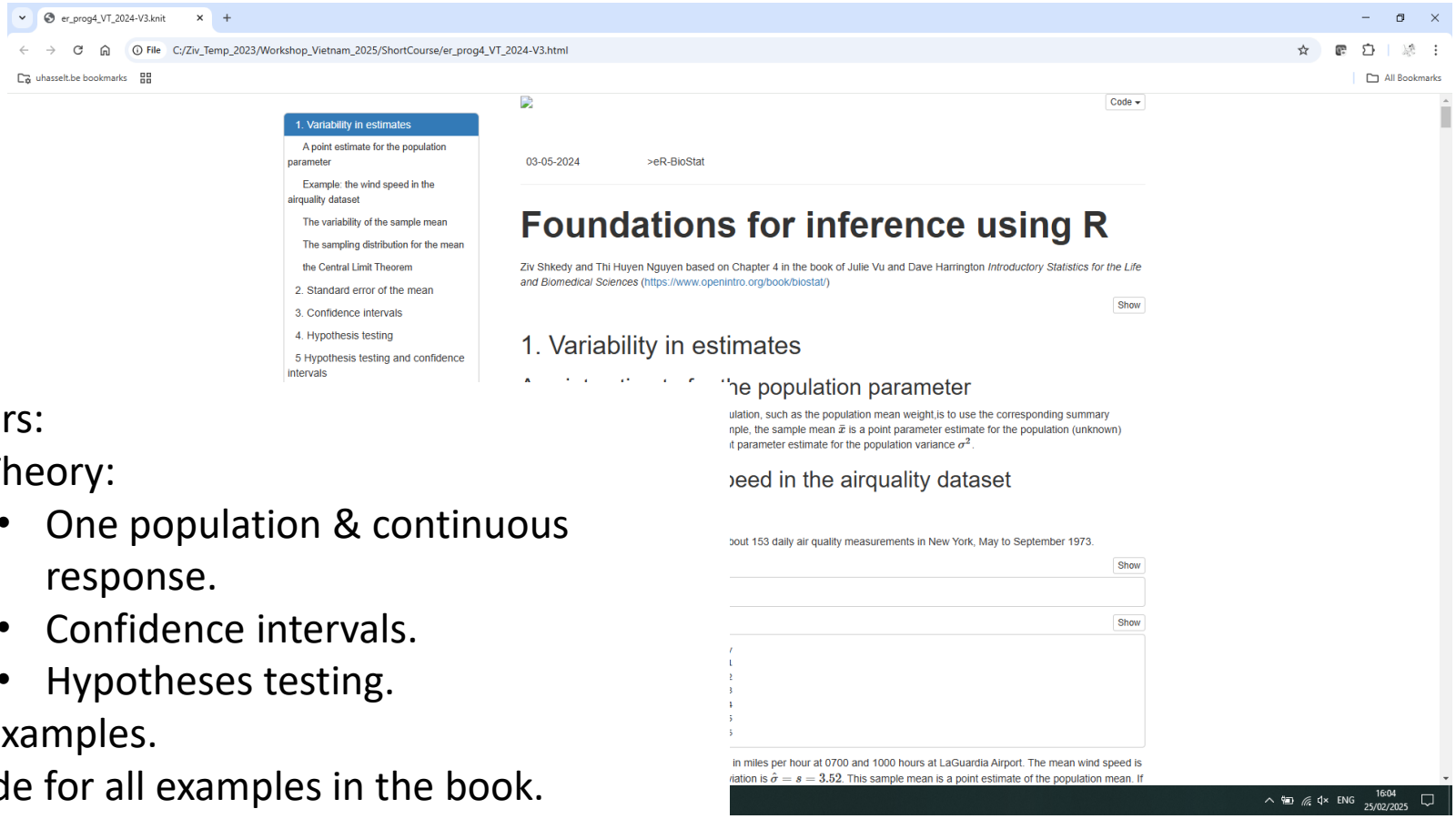
Rmd Program: er_prog4_VT_2024_V3a.Rmd.

The HTML book



- An online book covers chapter 4.
- The NHANES data set is used as one of the examples.

The HTML book



- Covers:
  - Theory:
    - One population & continuous response.
    - Confidence intervals.
    - Hypotheses testing.
  - Examples.
- R code for all examples in the book.

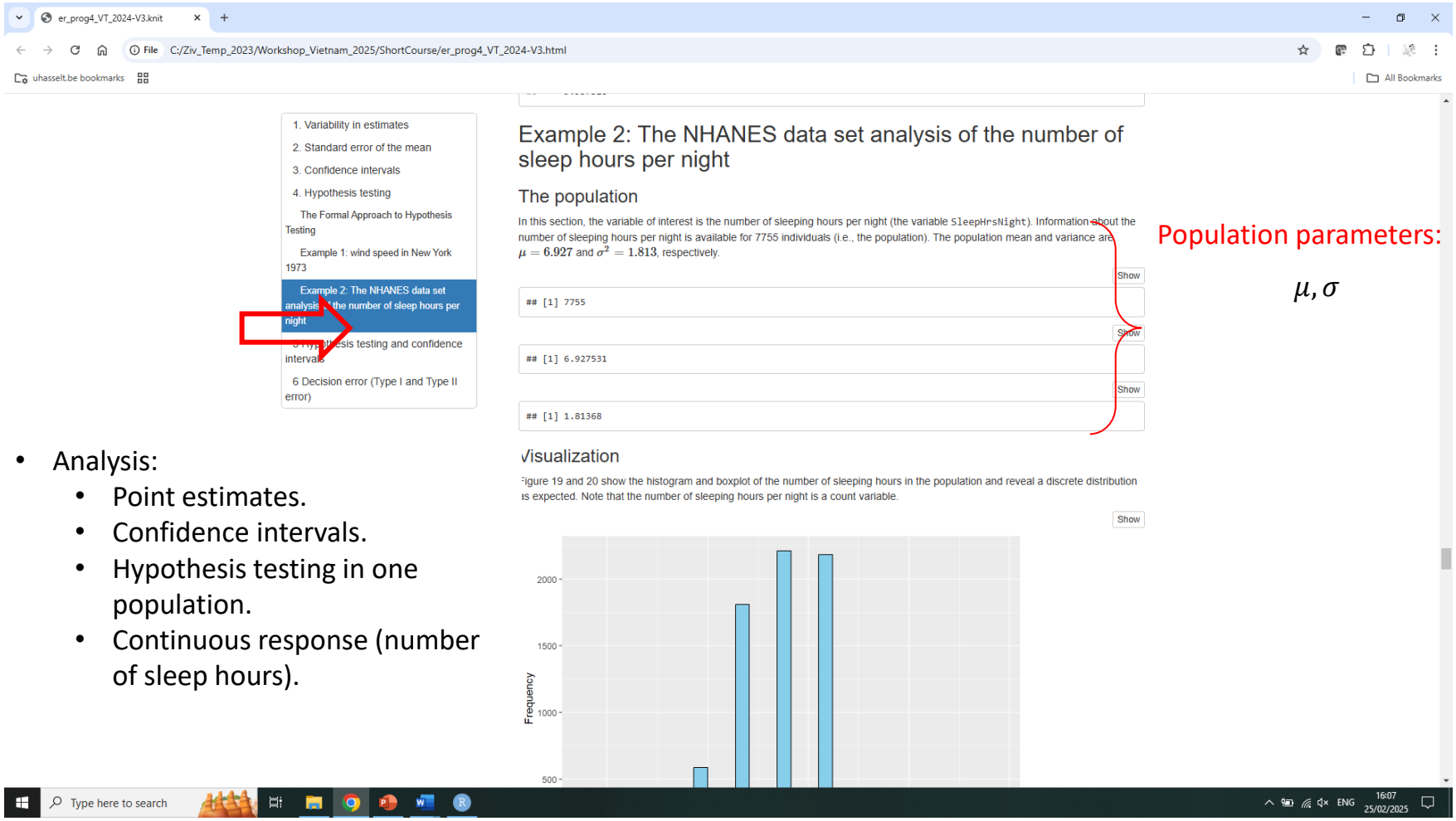
The NHANES data set

- The NHANES dataset consists of data from the US National Health and Nutrition Examination Study.
- Information about 76 variables is available for 10000 individuals included in the study.
- The 10000 individuals are considered as the population.

The NHANES data set: analysis of the number of sleep hours per night

- The variable of interest is the number of sleeping hours per night (the variable `SleepHrsNight`).
- Continuous variable.
- Information about the number of sleeping hours per night is available for 7755 individuals (i.e., the population).

The NHANES data set: analysis of the number of sleep hours per night



Population parameters:

μ, σ

- Analysis:
  - Point estimates.
  - Confidence intervals.
  - Hypothesis testing in one population.
  - Continuous response (number of sleep hours).

The NHANES data set analysis of the number of sleep hours per night

The population

In this section, the variable of interest is the number of sleeping hours per night (the variable `SleepHrsNight`). Information about the number of sleeping hours per night is available for 7755 individuals (i.e., the population). The population mean and variance are $\mu = 6.927$ and $\sigma^2 = 1.813$, respectively.

Hide

```
library(NHANES)
data(NHANES)
#dim(NHANES)
sleep<-na.omit(NHANES$SleepHrsNight)
length(sleep)
```

```
## [1] 7755
```

Hide

```
mean(sleep)
```

Population mean

```
## [1] 6.927531
```

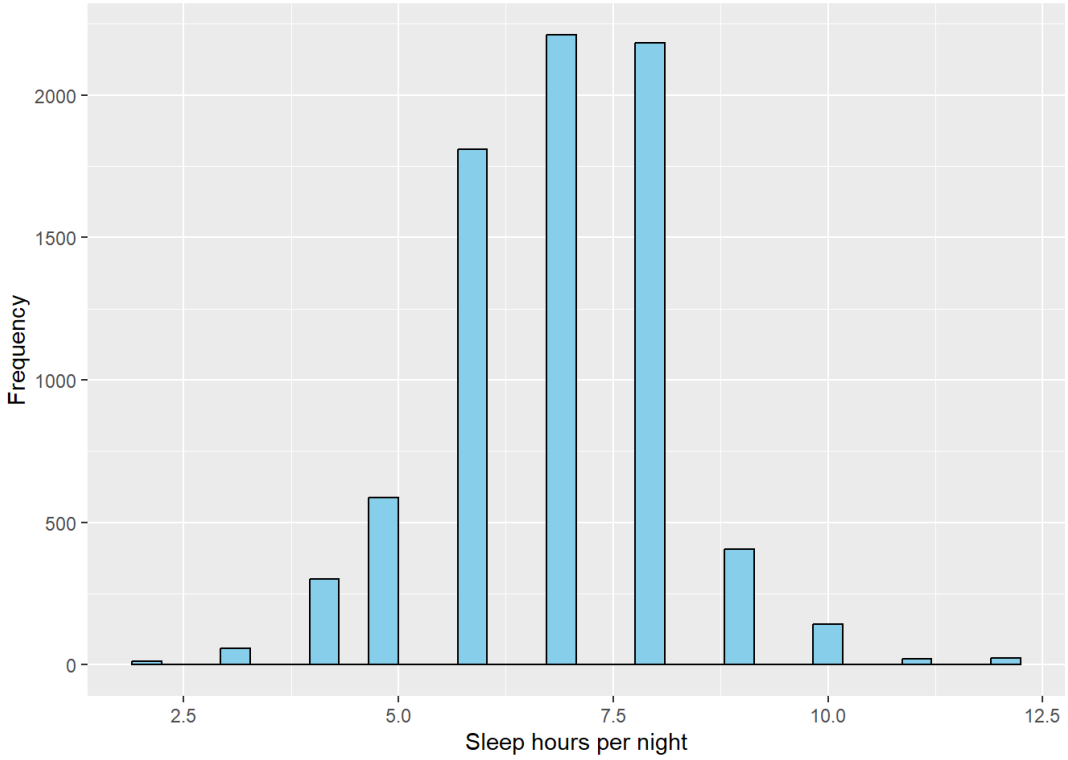
Hide

Population variance

```
var(sleep)
```

```
## [1] 1.81368
```

The number of sleep hours per night in the population



$$n = 7755$$
$$\mu = 6.927$$
$$\sigma = 1.813$$

Visualization

1. Variability in estimates
2. Standard error of the mean
3. Confidence intervals
4. Hypothesis testing

The Formal Approach to Hypothesis Testing

Example 1: wind speed in New York 1973

Example 2: The NHANES data set analysis of the number of sleep hours per night

5 Hypothesis testing and confidence intervals

6 Decision error (Type I and Type II error)

Hide

```
ggplot(NHANES, aes(x = SleepHrsNight)) +
  geom_histogram(fill = "skyblue", color = "black")+
  ylab("Frequency")+
  xlab("Sleep hours per night")
```

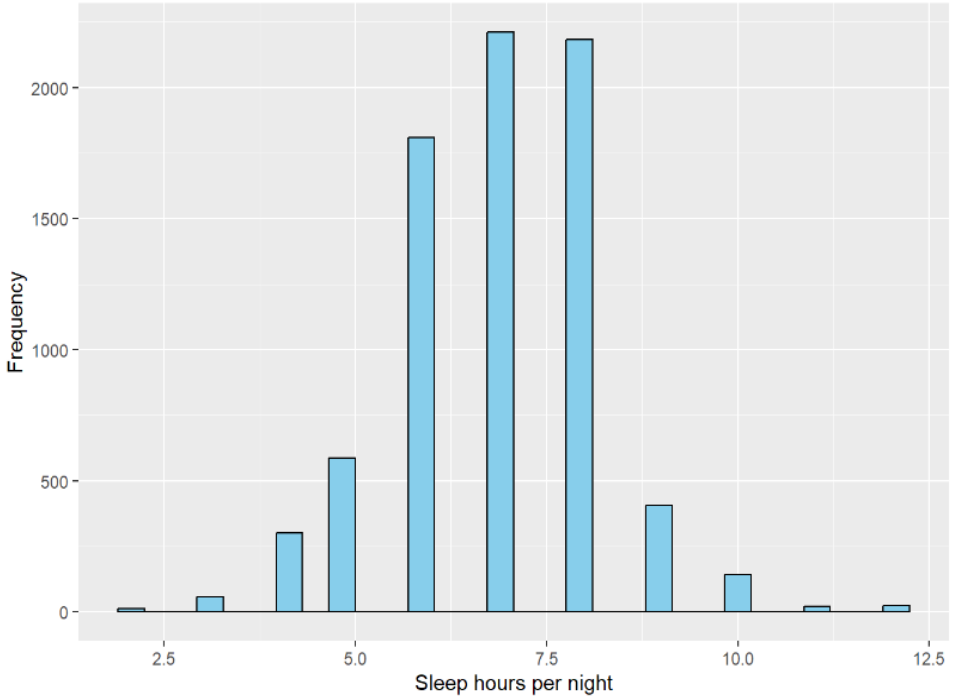


Figure 19: Histogram of sleep hours per night.

Case study: The NHANES dataset

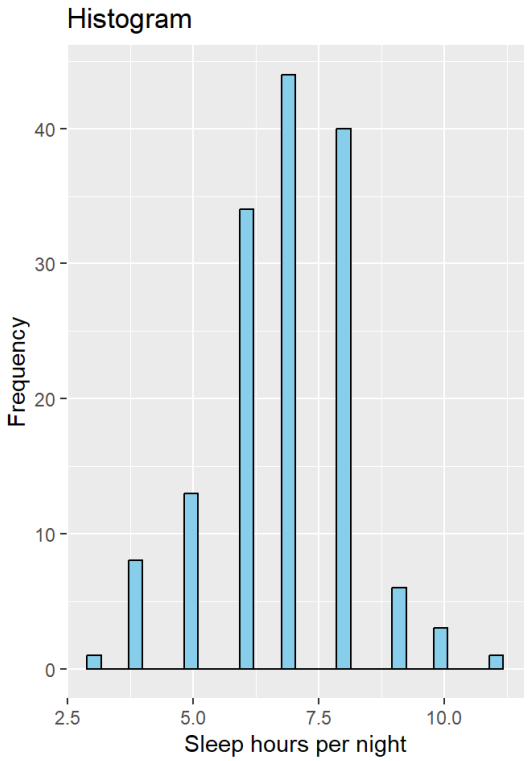
The number of sleep hours per night

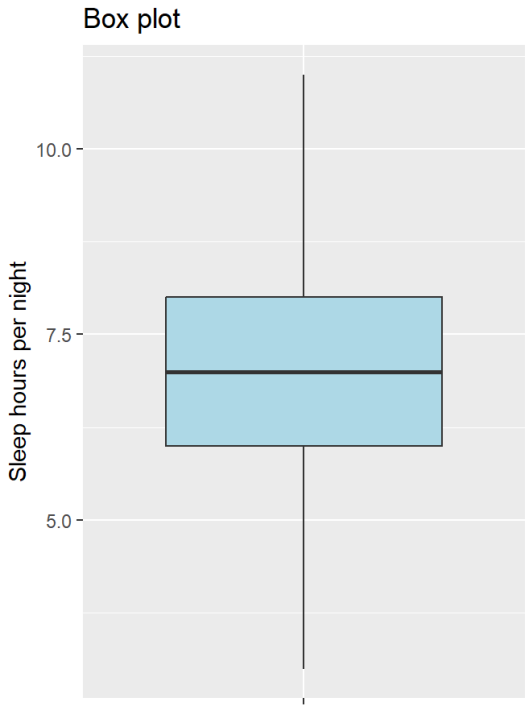
Point estimates

A random sample from the population

- Population size: 7755.
- We draw a random sample from the population.
- Sample size: 150.

A random sample from the population





- A random sample from the population:

$$n = 150$$
$$\bar{x} = 6.846$$
$$s^2 = 1.862$$

A random sample from the population

A random sample of size 150 from the population

We draw a sample of 150 indivuduals from the population ($n = 150$). The point estimates for the sample are $\bar{x} = 6.8466$ and $\sigma^2 = 1.8622$.

Hide

```
set.seed(456789)
x.sleep<-sample(na.omit(NHANES$SleepHrsNight),size=150,replace=FALSE)
length(x.sleep)
```

```
## [1] 150
```

Show

```
## [1] 6.846667
```

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Hide

```
var(x.sleep)
```

```
## [1] 1.862237
```

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Sample mean and variance.

A random sample from the population

Hide

```
box_sleep = ggplot(data.frame(SleepHrsNight = x.sleep), aes(x = "", y = SleepHrsNight)) +
  geom_boxplot(fill = "lightblue")+
  xlab("")+
  ylab("Sleep hours per night")+
  ggtitle("Box plot")
```

Hide

```
grid.arrange(hist_sleep, box_sleep, ncol = 2)
```

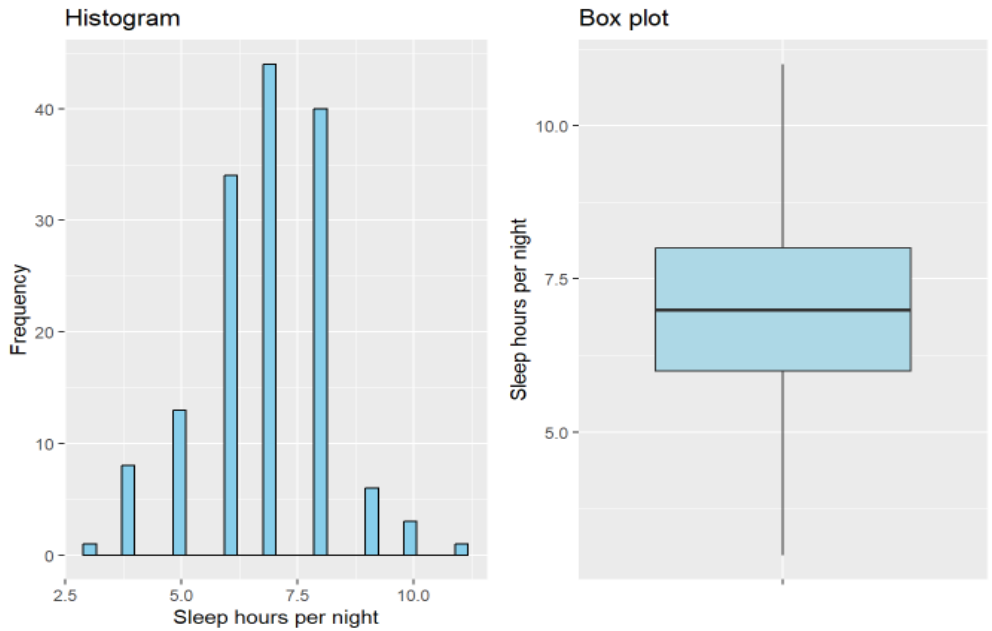


Figure 21: Histogram and box plot of sleep hours per night in the sample.

Case study: The NHANES dataset

The number of sleep hours per night

Confidence interval for the population mean

Confidence interval for the population mean (Case 2)

If $X \sim F$

Then : $\bar{X} \sim N(\mu, \frac{S^2}{n})$

and $T_{\bar{X}} = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)$

X has an unknown distribution, but we have a large sample (n> 30)

$E(X) = \mu$

$Var(X) = \sigma^2$

The same as case 1 but we replace σ^2 by S^2.

C.I. for case 2

Step 1: example, choose 1-α = 0.95

Step 2: case 2, so : $\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$ or $\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)$

Step 3: critical points: -1.96 and 1.96
(the same as in Case 1, since we are still using the standard normal distribution function)

$\bar{X}$

Step 4: Calculate the point estimator (s) (and possibly s^2)

C.I. for case 2

Step 5: In the same manner as in Case 1:

The (1-α) CI for μ is :

$$\left[\bar{x} - z\sqrt{\frac{\sigma^2}{n}}, \bar{x} + z\sqrt{\frac{\sigma^2}{n}}\right] \quad \text{or} \quad \left[\bar{x} - z\sqrt{\frac{s^2}{n}}, \bar{x} + z\sqrt{\frac{s^2}{n}}\right]$$

Example for case 2

- Suppose X = number of sleep hours per night.
- X has an unknown distribution with unknown variance.
- But large sample (n = 150 >> 30).

The 95% CI for μ : the mean number of sleep hours per night in the population.

Step 1: choose confidence level 1-α = 0.95

Step 2: case 2, so : $\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)$

Step 3: critical points: -1.96 and 1.96

(the same as in Case 1, since we are still using the standard normal distribution function).

Example for case 2

Step 4 : Calculate the point estimators:

$$\bar{x} = 6.8466 \quad \text{and} \quad s^2 = 1.8622$$

Step 5 : In the same manner as in Case 1:

The (1-α) CI for μ is :

$$\left[\bar{x} - z\sqrt{\frac{s^2}{n}}, \bar{x} + z\sqrt{\frac{s^2}{n}}\right]$$

$$\left[6.8466 - 1.96\sqrt{\frac{1.8622}{150}}, 6.8466 + 1.96\sqrt{\frac{1.8622}{150}}\right]$$

[6.6283,7.0650]

Example for case 2

- A 95% CI for the population mean μ of the number of sleep hours per night [6.6283, 7.0650]

- **Interpretations:**
  - Based on our sample, we are 95% confident that the true mean of number of sleep hours per night lie between 6.6283 and 7.0650.

A 95% C.I. for the mean sleep hours per night

A 95% C.I for the mean sleep hours per night

The sample standard deveation and the standard error of the sample mean are equal to 1.3646 and 0.1114, respectivly.

Hide

```
n<-length(x.sleep)
SD.x<-sqrt(var(x.sleep))
SD.x
```

```
## [1] 1.364638
```

Standard deviation

Hide

```
SE<-SD.x/sqrt(n)
SE
```

$$SE = \sqrt{\frac{s^2}{n}}$$

```
## [1] 0.1114222
```

Standard error of the sample mean

For the sample, the error margin for a 95% confidence interval is $m = 1.96 \times SE = 1.96 \times 0.1114$ and the confidence interval is given by

$$\bar{x} \pm m = 6.8466 \pm 0.2183 = (6.628279, 7.065054).$$

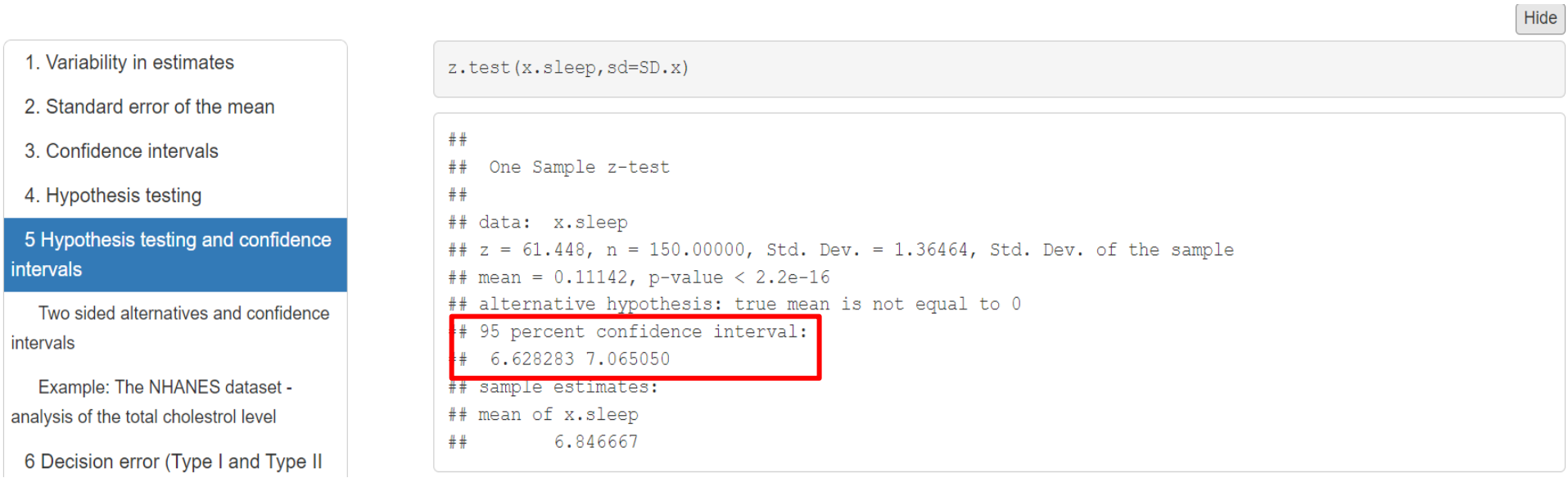
Hide

```
LL<-mean(x.sleep)-1.96*SE
UL<-mean(x.sleep)+1.96*SE
c(LL,UL)
```

```
## [1] 6.628279 7.065054
```

A 95% C.I. for the mean sleep hours per night

- A 95% Confidence interval for the population mean using the R function `z.test()`.



Case study: The NHANES dataset

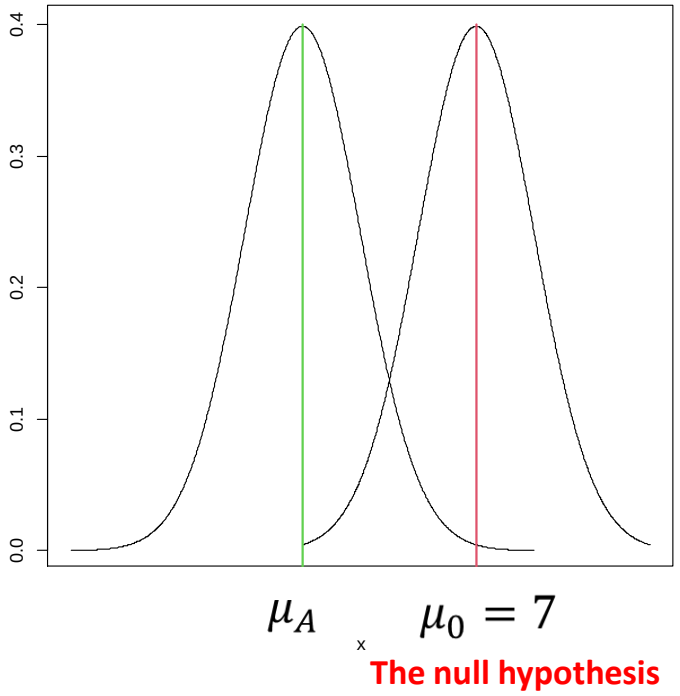
The number of sleep hours per night

Hypotheses testing

Test of hypothesis: a one sided test

$$H_0: \mu = 7$$
$$H_A: \mu < 7$$

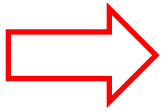
- We test the null hypothesis versus a one sided alternative.
- In our case, under the alternative the mean is smaller than 7 (but not specified).



Test statistic

$$t=\frac{\bar{x}-\mu_0}{\sqrt{\frac{1.3646^2}{150}}}=-1.3761$$

The population variance σ^2 is unknown but…n=150.



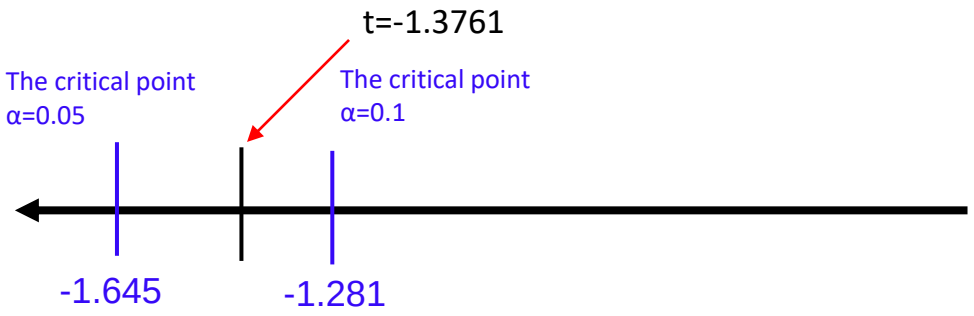
$$\frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)$$

The critical points and the test statistic

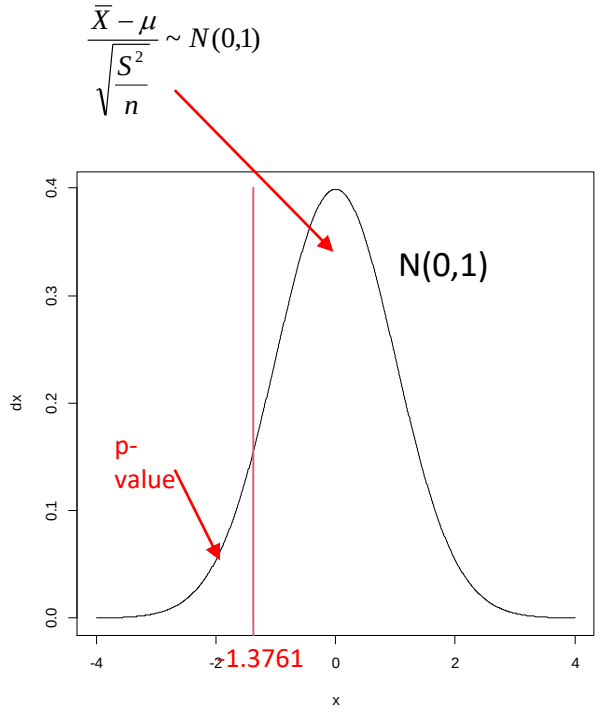
For one sided test and α=0.05, Z=-1.645.

For one sided test and α=0.1, Z=-1.281.

For α=0.1 We reject H_0 : -1.3761< -1.281.



p-value



$H_0: \mu = 7$

$H_A: \mu < 7$

$P(Z < -1.3761)$=0.08439

- For α=0.05, we DO NOT reject the null hypothesis.
- For α=0.1, we reject the null hypothesis.

Hypothesis testing

Hypothesis testing

We wish to test the null hypothesis $\mu = 7$ aginst a one sided alternative $H_1 : \mu < 7$. This can be done using the argument `alternative = 'less'` in the function `z.test`. Note that we assume that in the population, $\sigma = 1.3646$. As can be seen in the panel below, for the sample, the mean number of sleeping hours is equal to $\bar{x} = 6.8466$ and the test statistic is equal to -1.3761. The p-=0.08439 > 0.05$. We cannot reject the null hypothesis and conclude that $\mu = 7$.

Hide

```
mean(x.sleep)
```

```
## [1] 6.846667
```

Hide

```
sqrt(var(x.sleep))
```

```
## [1] 1.364638
```

Hide

```
z.test(x.sleep,mu=7, 1.364638, alternative = 'less')
```

```
##
##  One Sample z-test
##
## data:  x.sleep
## z = -1.3761, n = 150.00000, Std. Dev. = 1.36464, Std. Dev. of the
## sample mean = 0.11142, p-value = 0.08439
## alternative hypothesis: true mean is less than 7
## 95 percent confidence interval:
##     -Inf 7.02994
## sample estimates:
## mean of x.sleep
##        6.846667
```

Development of E-learning materials using R markdown

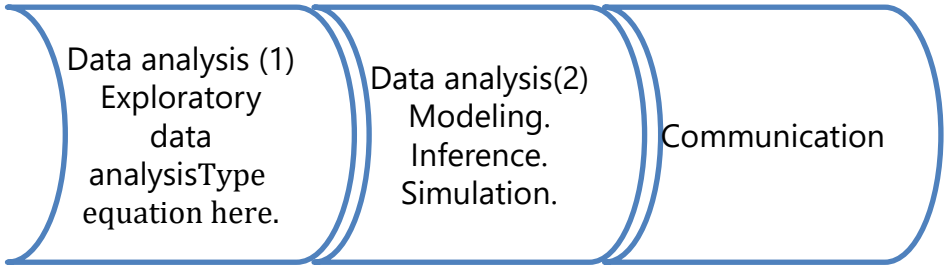
Part 5

The course online (1)

Steps in data analysis

Modeling the association between the fuel consumption and the car's weight.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



Boxplot by treatment group.

A simple linear regression.

A report.

Methodology: simple linear regression.

We "translate" the methodology to software usage

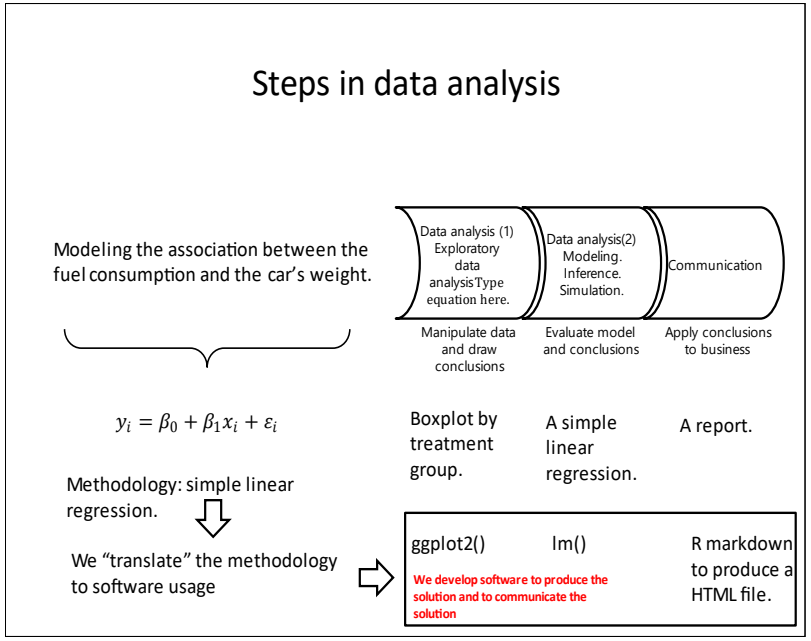
ggplot2()

lm()

R markdown to produce a HTML file.

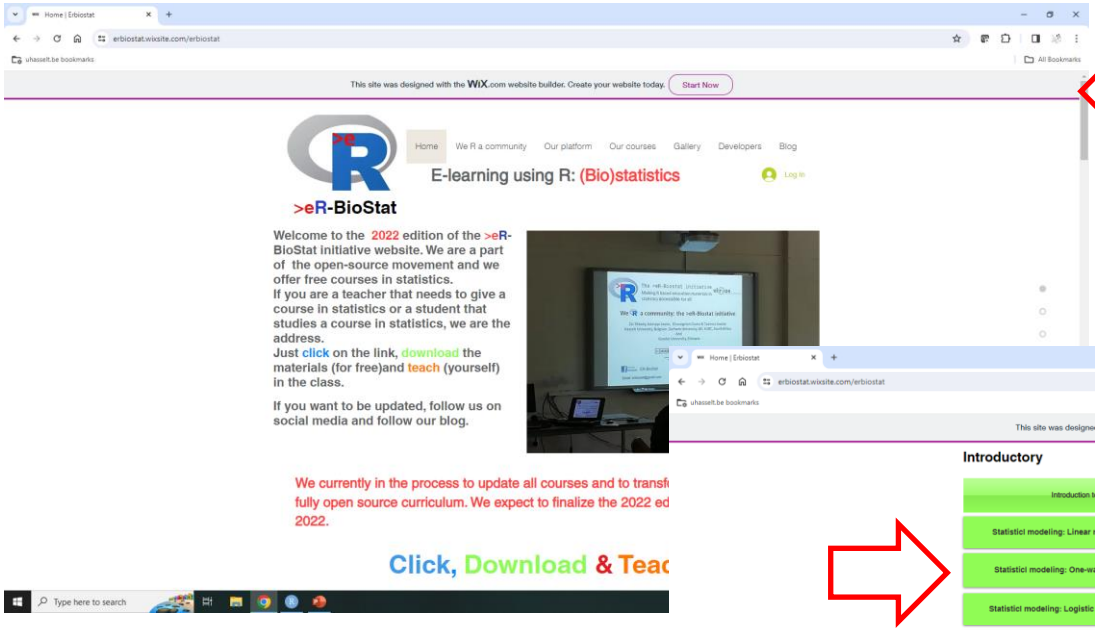
We develop software to produce the solution and to communicate the solution

Developing a course about linear regression



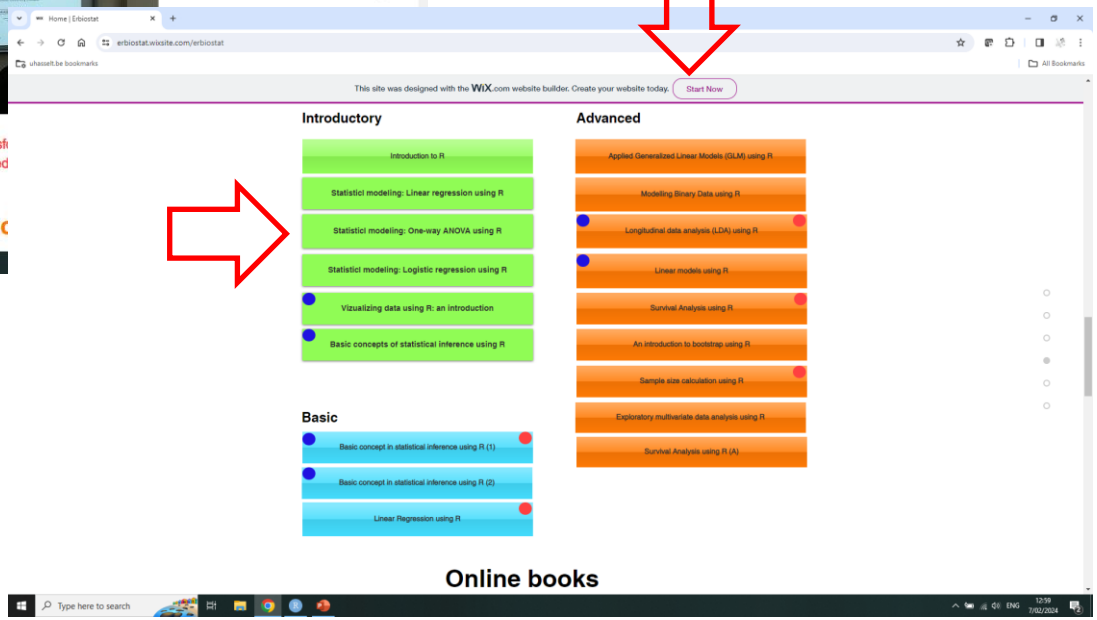
- Suppose that we do not need to produce a report for an analysis but…
- Our aim: development of education materials for a course about linear regression.

Introduction to statistical modeling using R



>eR-Biostat website.

List of courses.



https://erbiostat.wixsite.com/erbiostat

Introduction to statistical modeling using R

Home | Erbiostat | Home | IntroStatMod

erbiostat.wixsite.com/introstatmod

uhasselt.be bookmarks | All Bookmarks

Introduction to Statistical modeling using R

Home About Topics Online tutorials Contact

>eR-BioStat

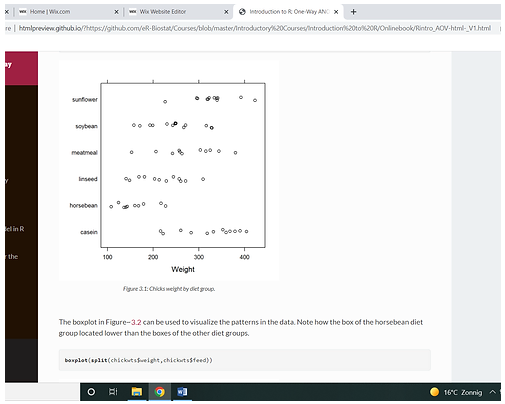
This group of courses is focused on statistical modelling and covers the following topics:

- Simple linear regression using R.
- One-way ANOVA using R.
- Logistic Regression using R.

The courses can be given together, as three parts of a course about statistical modelling or separately as a part of a specific course in statistical modelling.

The courses were developed at an undergraduate level courses (for both statistician and non statisticians) .

This is an open source course and all source files used to produce the slides are available online (in PP, Tex or Rmd formats).



About

General information about the course and course materials and the study methods used in the course.

Topics

In this page, the course is presented in a typical slides format. The course

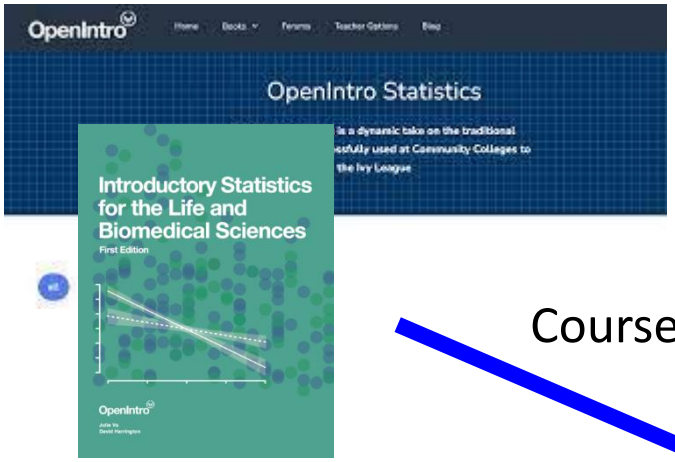
Online tutorials

In this page, suppurating online tutorials are given in different formats. The online

Type here to search ENG 13:00 7/02/2024

https://erbiostat.wixsite.com/introstatmod

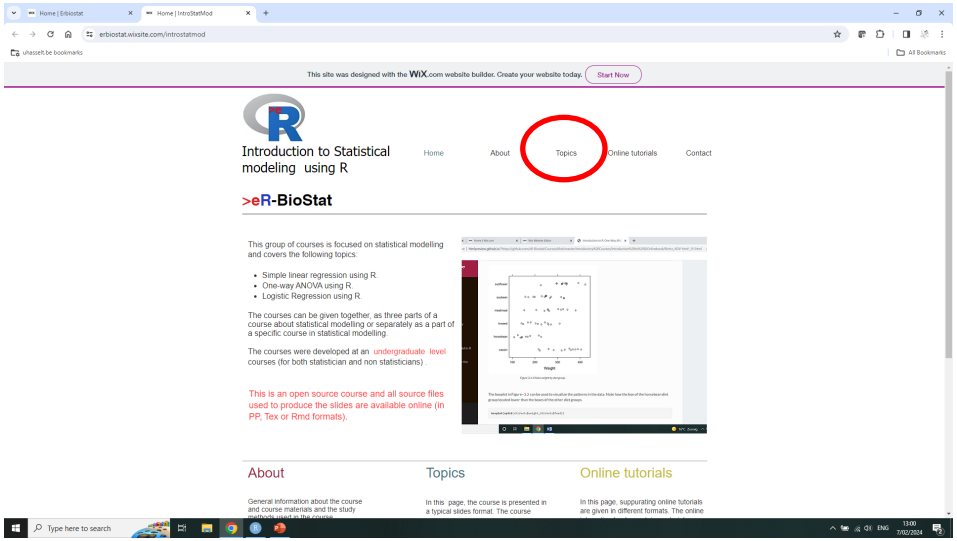
Introduction to statistical modeling using R



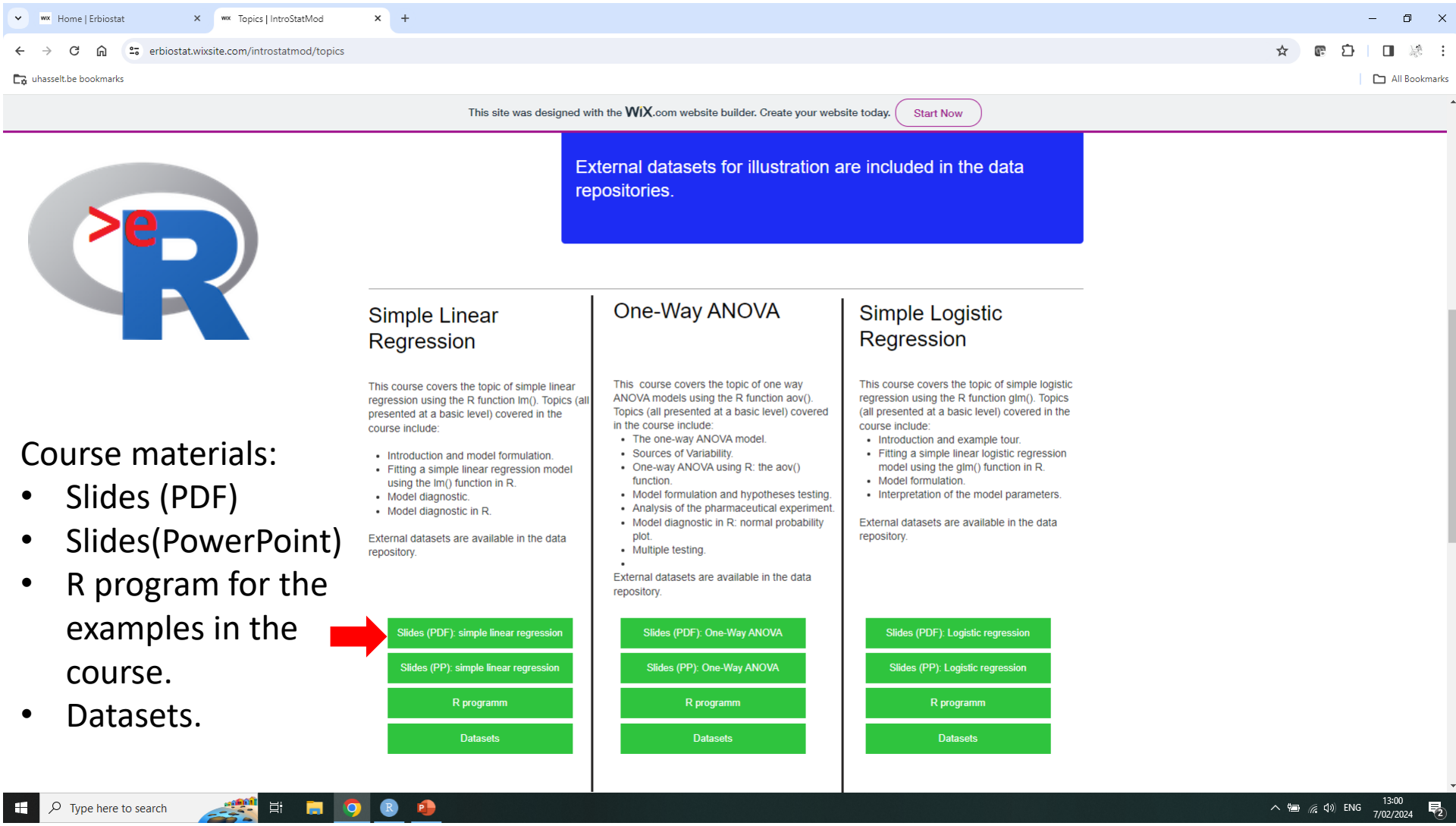
Course II

Course materials from two sources

Course I



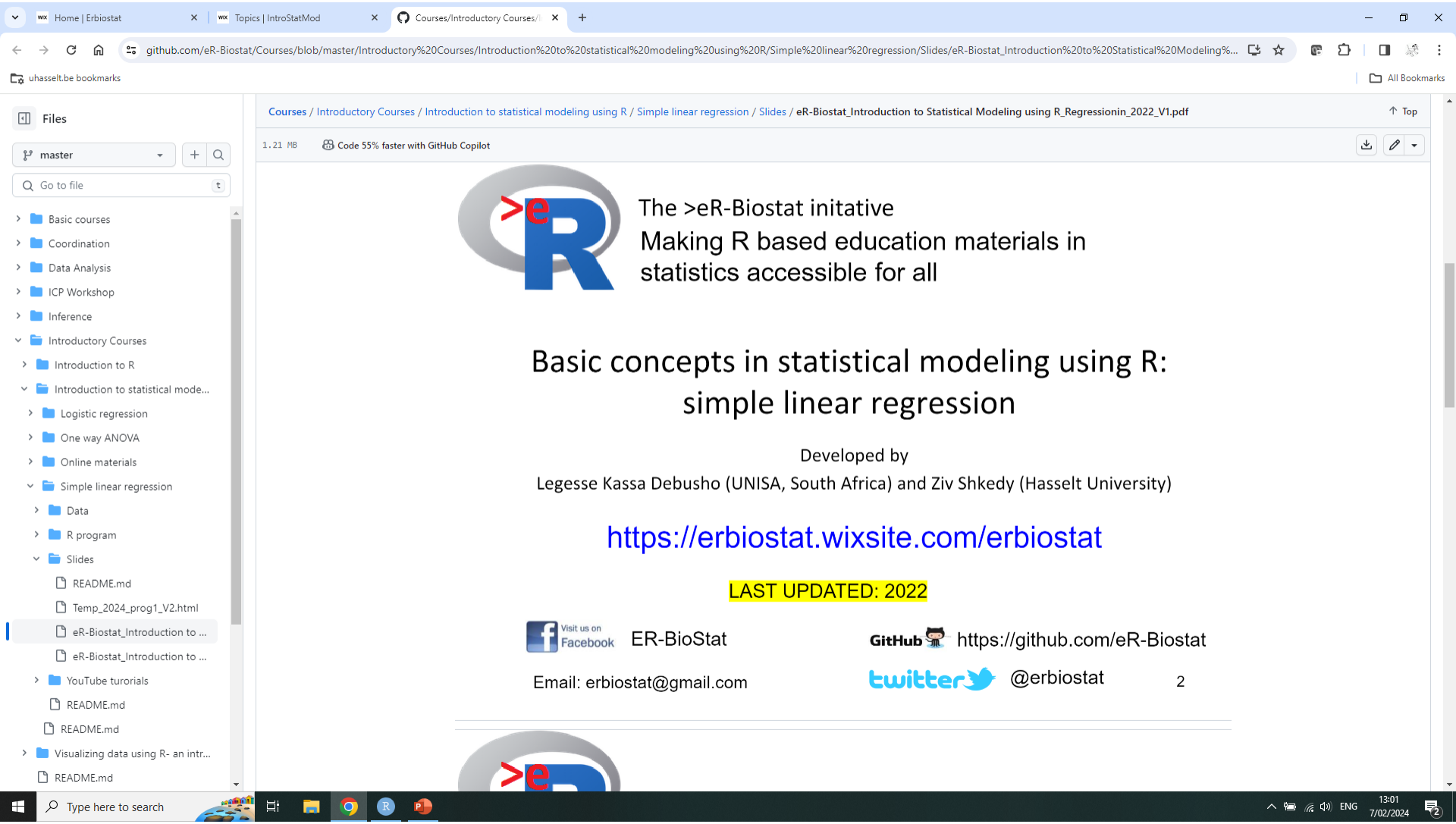
Course I



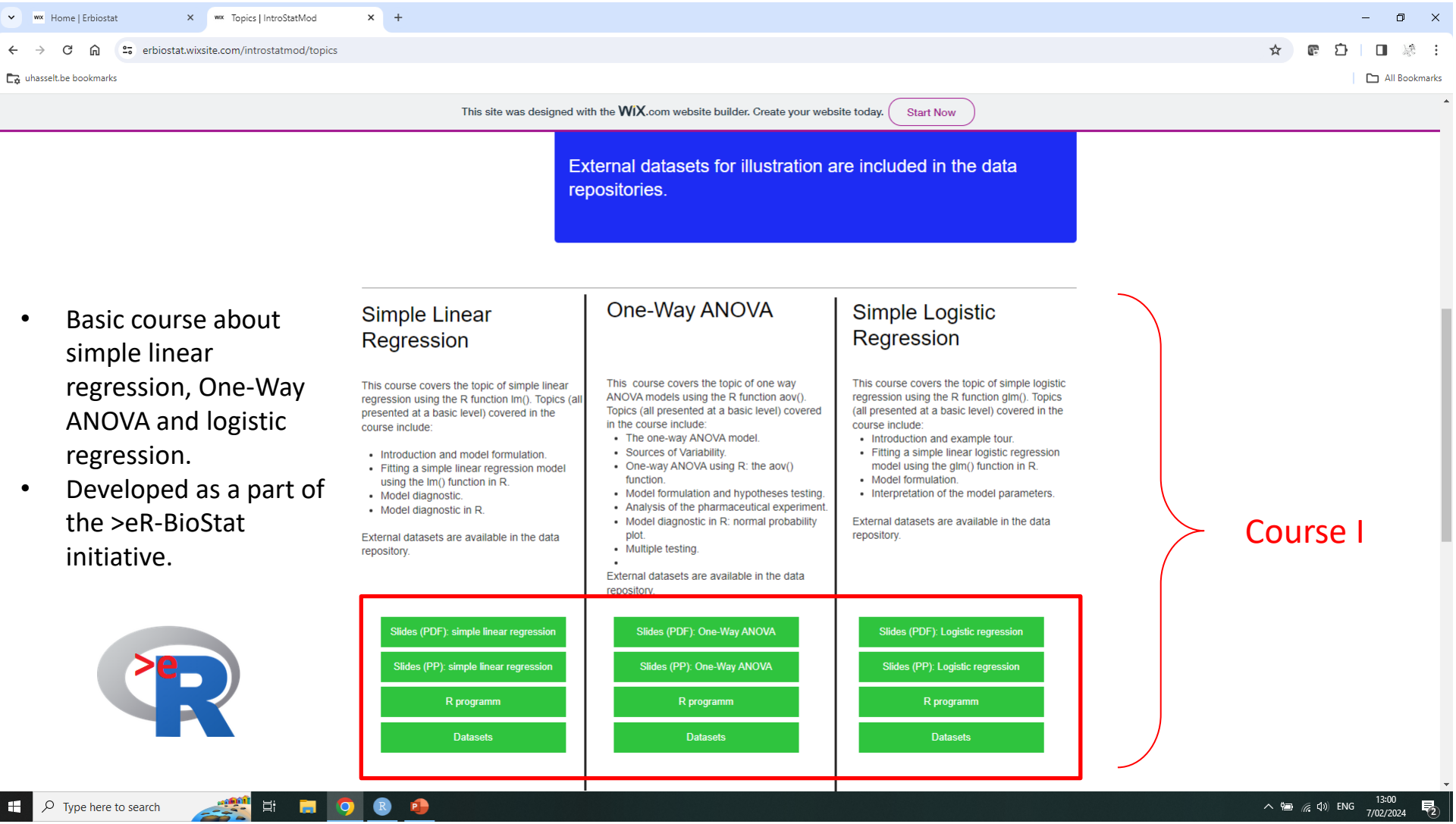
Course materials:

- Slides (PDF)
- Slides(PowerPoint)
- R program for the examples in the course.
- Datasets.

Examples of the slide

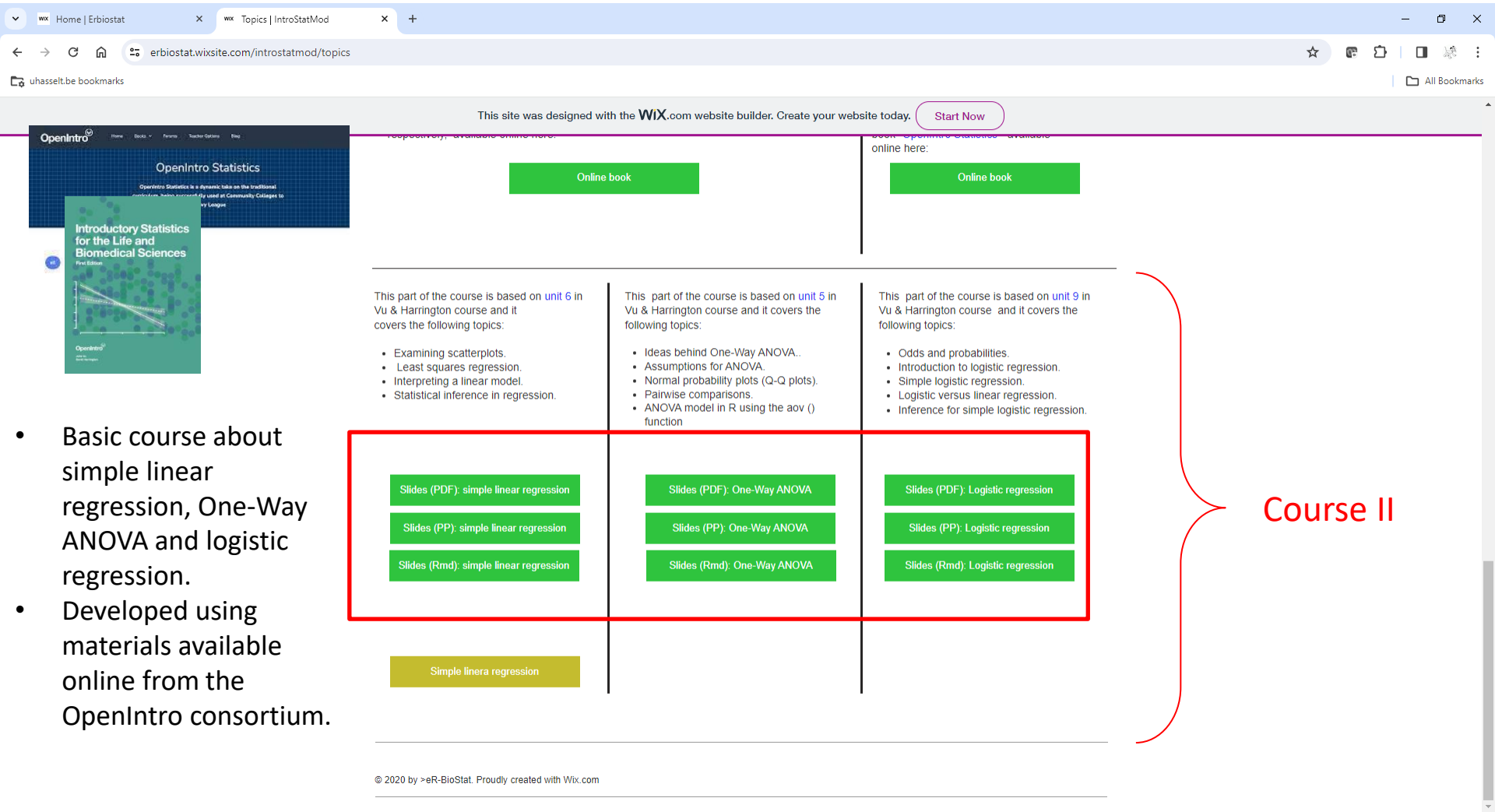


Course materials



- Basic course about simple linear regression, One-Way ANOVA and logistic regression.
- Developed as a part of the >eR-BioStat initiative.

Course materials



- Basic course about simple linear regression, One-Way ANOVA and logistic regression.
- Developed using materials available online from the OpenIntro consortium.

Course II

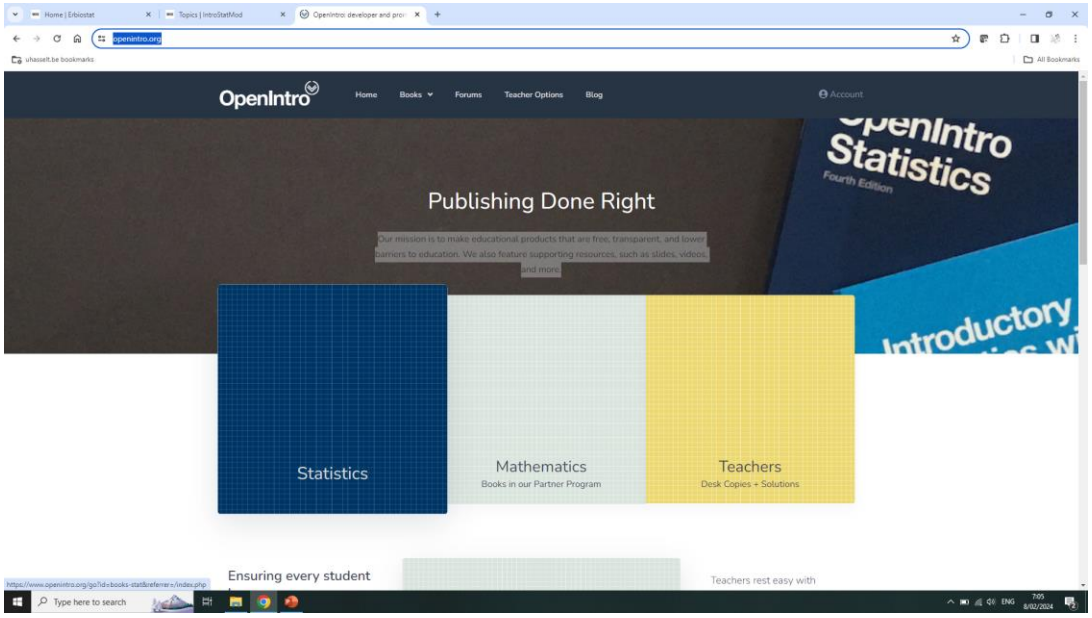
openintro.org

- The OpenIntro project was founded in 2009 to improve the quality and availability of education by producing:
  - Exceptional books.
  - Teaching tools.
- Free to use and easy to modify.
- “Our inaugural effort is *OpenIntro Statistics*. Probability is optional, inference is key, and we feature real data whenever possible”.
- Files for the entire book are freely available at openintro.org.

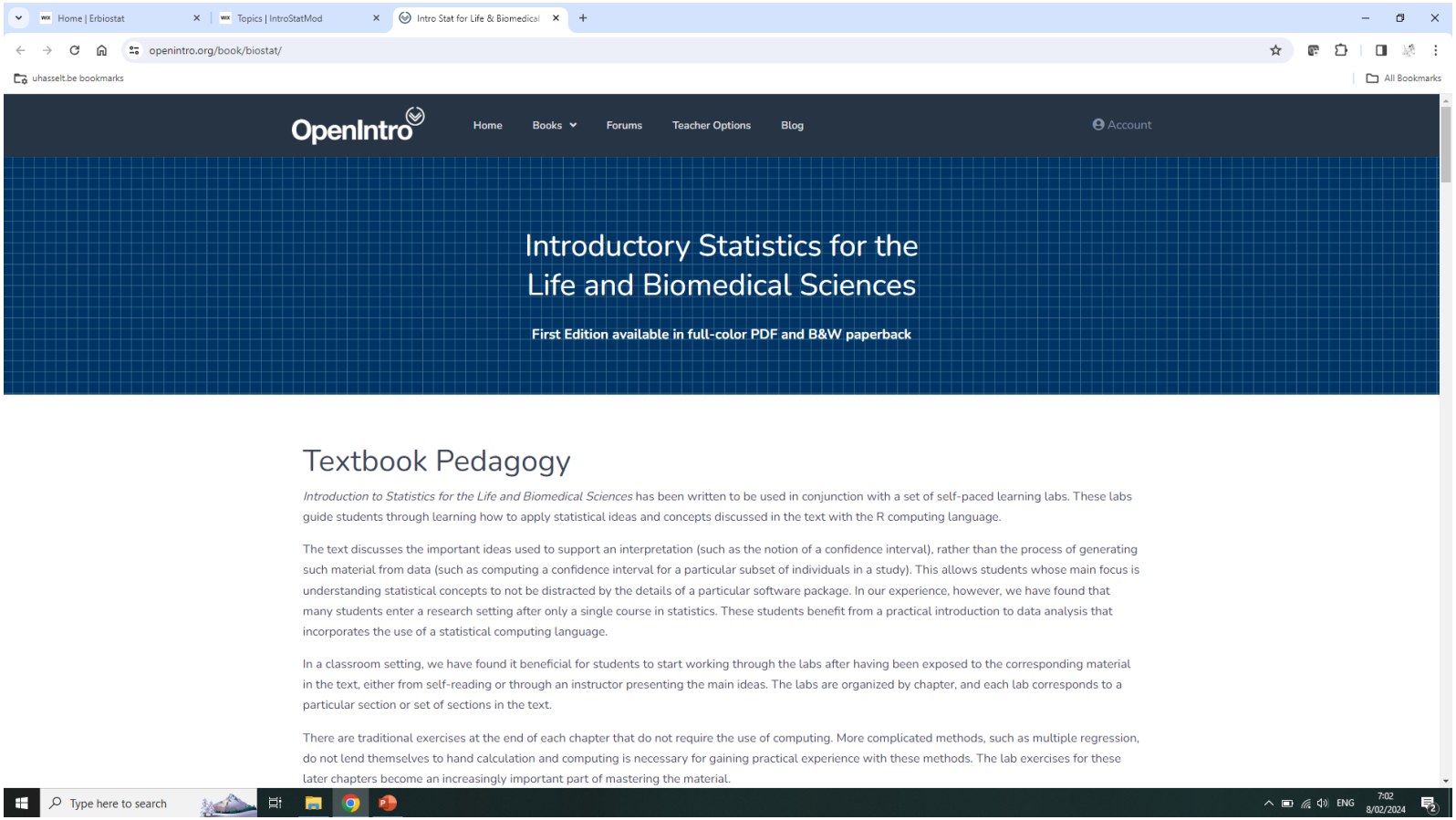
openintro.org

https://www.openintro.org/

“Our mission is to make educational products that are free, transparent, and lower barriers to education. We also feature supporting resources, such as slides, videos, and more.”

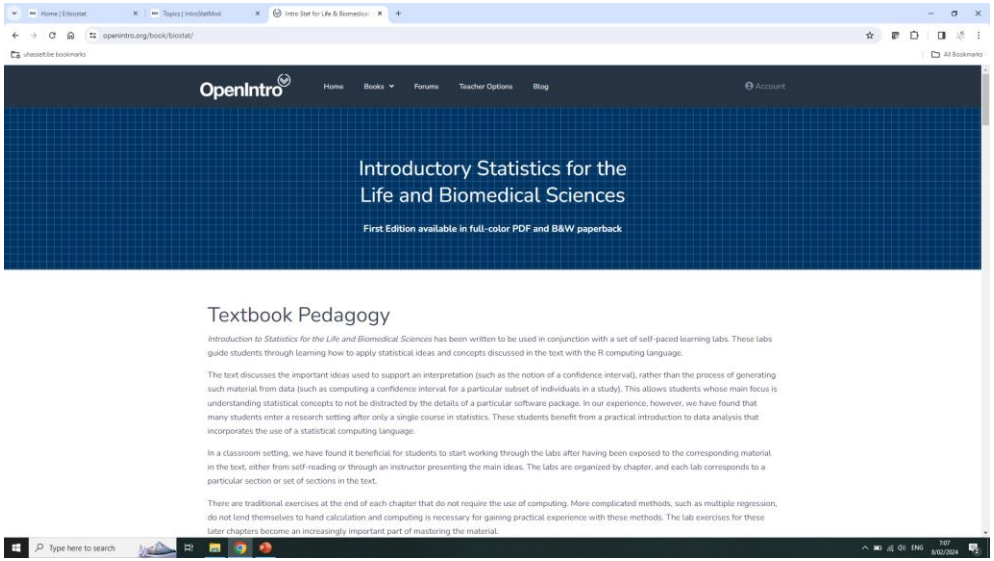


Introductory Statistics for the Life and Biomedical Sciences



https://www.openintro.org/book/biostat/

Introductory Statistics for the Life and Biomedical Sciences



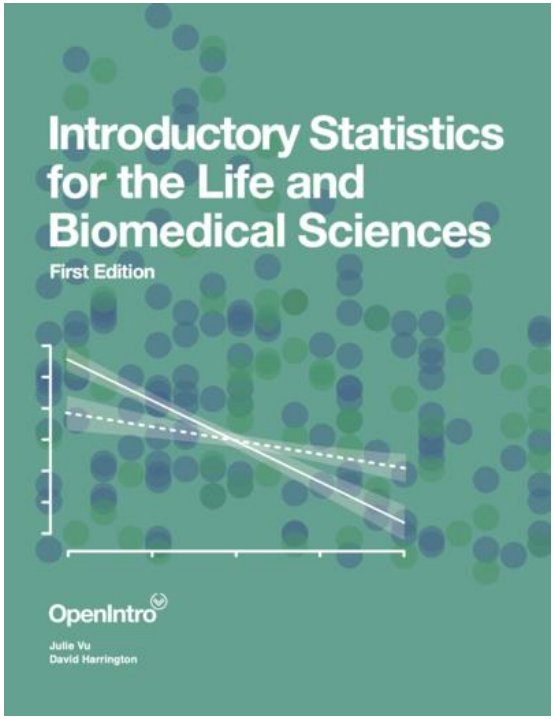
OpenIntro Biostatistics: developing open source materials for teaching and learning applied statistics with R.

Developed by Dave Harrington and Julie Vu.

Developed for students in health or life sciences:

- Motivated more by application than theory.
- No previous statistics courses.
- Familiarity with basic algebra.
- No or limited experience with computing.

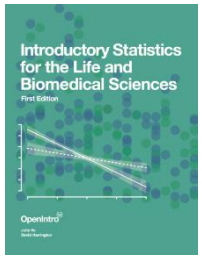
Introductory Statistics for the Life and Biomedical Sciences



- Written by Dave Harrington and Julie Vu.
- Available on line (for free) on PDF format.
- Focus on theory and practice.
- Undergraduate/graduate levels.
- Available (for free) online:
  - Source files for the book in.
  - Presentations and practical sessions (labs).

Reference for Course I & II

Reference for course I & II



Home | Erbiostat | Topics | IntroStatMod

erbiostat.wixsite.com/introstatmod/topics

uhasselt.be bookmarks

All Bookmarks

This site was designed with the WiX.com website builder. Create your website today. Start Now

Online book

online here:

Online book

This part of the course is based on unit 6 in Vu & Harrington course and it covers the following topics:

- Examining scatterplots.
- Least squares regression.
- Interpreting a linear model.
- Statistical inference in regression.

Slides (PDF): simple linear regression

Slides (PP): simple linear regression

Slides (Rmd): simple linear regression

Simple linera regression

This part of the course is based on unit 5 in Vu & Harrington course and it covers the following topics:

- Ideas behind One-Way ANOVA..
- Assumptions for ANOVA.
- Normal probability plots (Q-Q plots).
- Pairwise comparisons.
- ANOVA model in R using the aov () function

Slides (PDF): One-Way ANOVA

Slides (PP): One-Way ANOVA

Slides (Rmd): One-Way ANOVA

This part of the course is based on unit 9 in Vu & Harrington course and it covers the following topics:

- Odds and probabilities.
- Introduction to logistic regression.
- Simple logistic regression.
- Logistic versus linear regression.
- Inference for simple logistic regression.

Slides (PDF): Logistic regression

Slides (PP): Logistic regression

Slides (Rmd): Logistic regression

Course II

© 2020 by >eR-BioStat. Proudly created with Wix.com

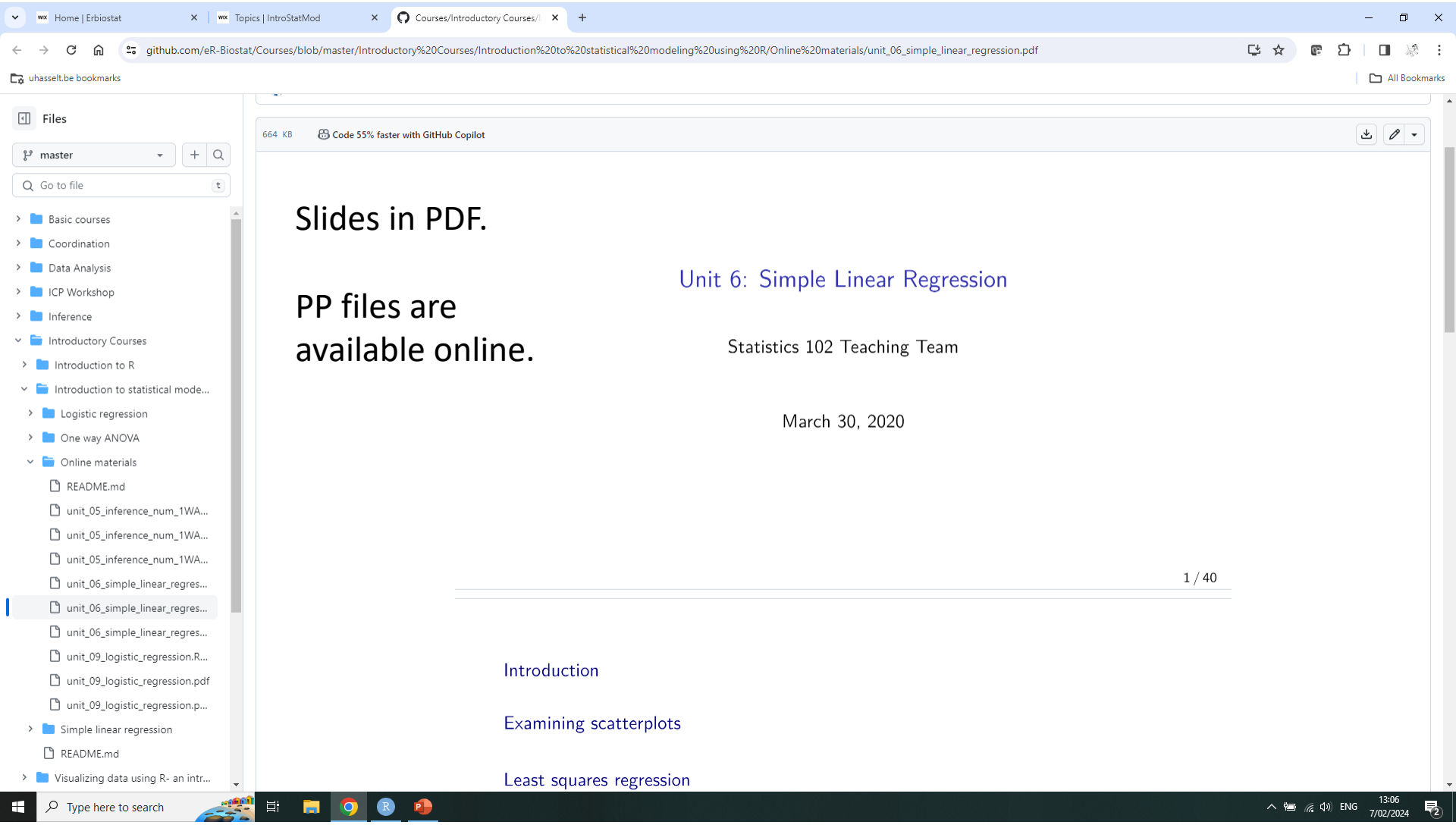
Type here to search

ENG 13:02 7/02/2024

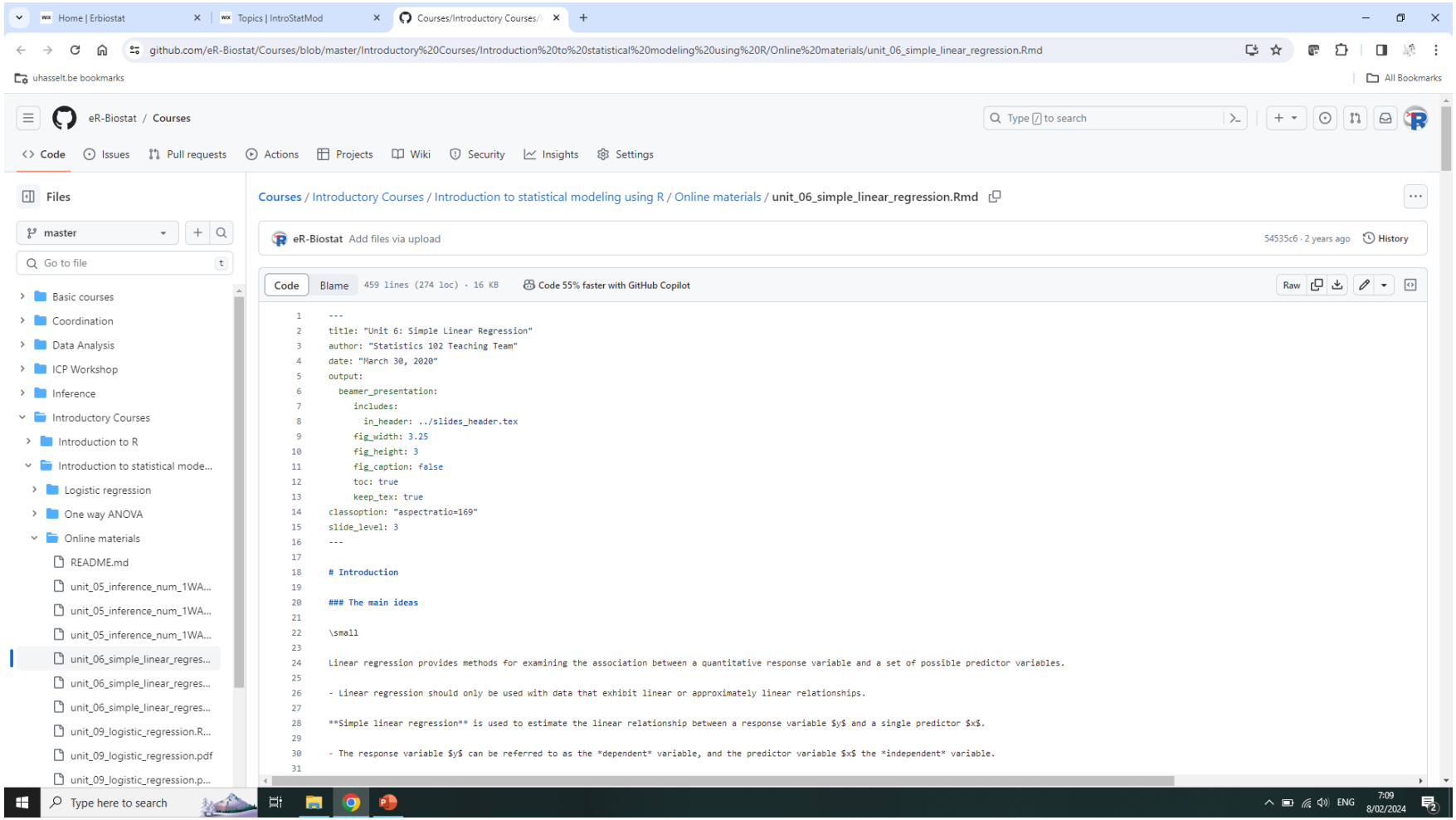
Course II: Slides

Slides in PDF.

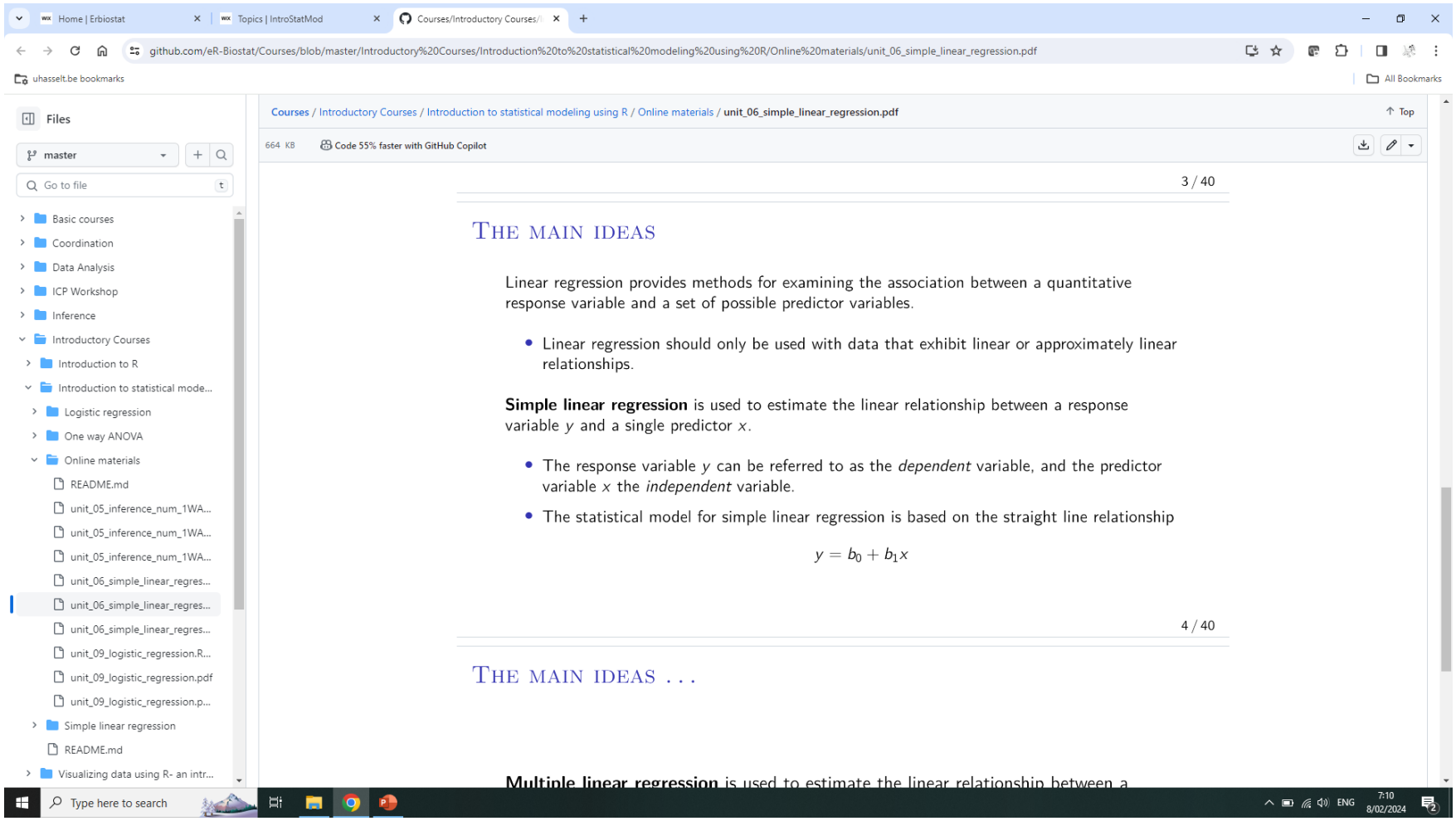
PP files are available online.



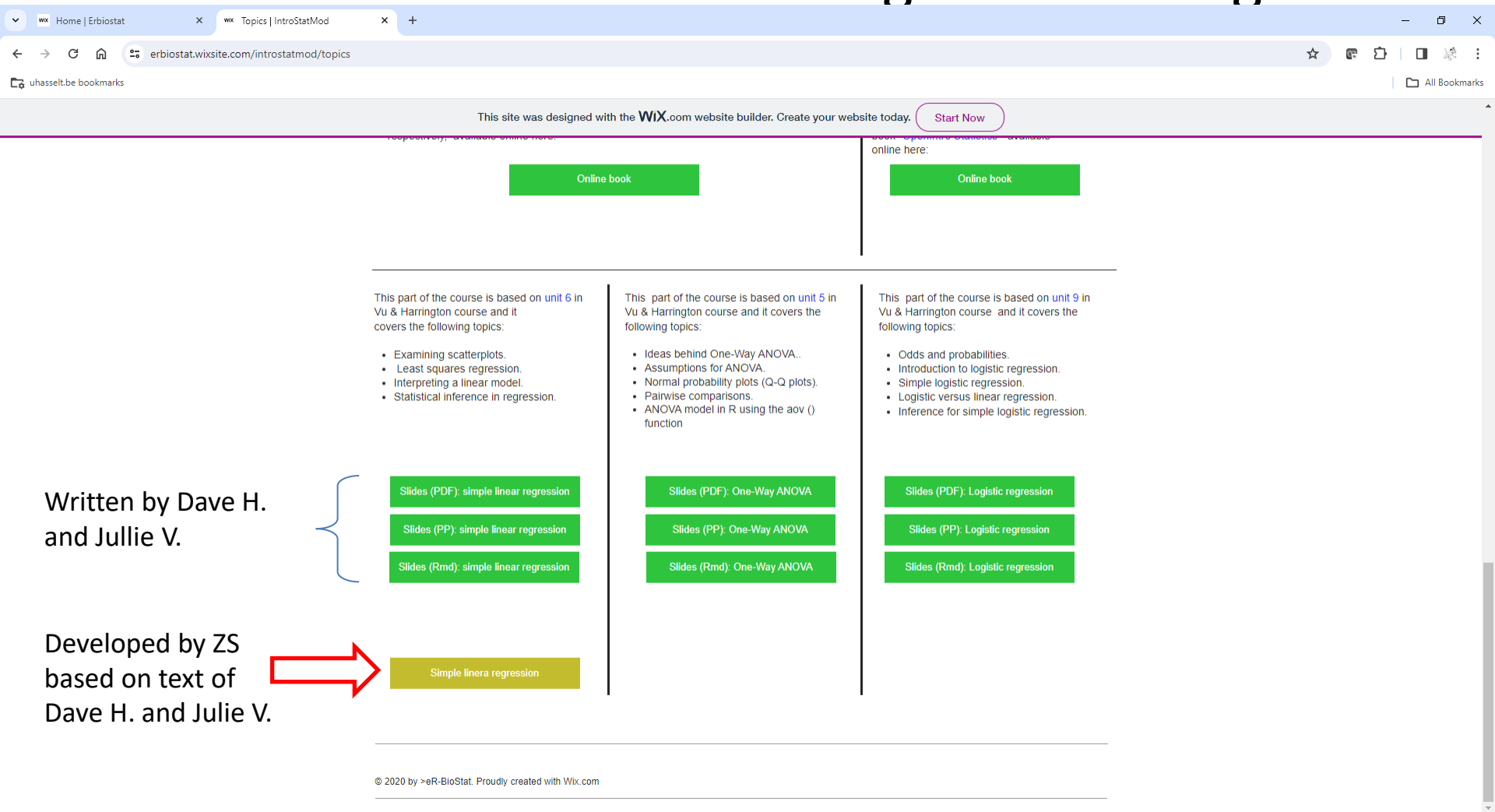
Course II: The Rmd file for the slides



Course II: A PDF output



Course I + II:
online text about linear regression using R



Written by Dave H. and Jullie V.

Developed by ZS based on text of Dave H. and Julie V.

A part of the website of the course

Home | Erbiostat | Topics | IntroStatMod | Temp_2024_prog1_V2.knit

htmlpreview.github.io/?https://github.com/eR-Biostat/Courses/blob/master/Introductory%20Courses/Introduction%20to%20statistical%20modeling%20using%20R/Simple%20linear%20regression/Slides/Temp_2024_prog1_V2.html

uhasselt.be bookmarks | All Bookmarks

Code

15-12-2023 >eR-BioStat

Simple linear regression using R

Ziv Shkedy et al

```
## load/install libraries
.libPaths(c("./Rpackages",.libPaths()))
library(knitr)
library(tidyverse)
library(deSolve)
library(minpack.lm)
library(ggpubr)
library(readxl)
library(gamlss)
library(data.table)
library(grid)
library(png)
library(nlme)
library(gridExtra)
library(mvtnorm)
library(e1071)
library(lattice)
library(ggplot2)
library(dslabs)
library(NHANES)
library(plyr)
library(dplyr)
library(nasaweather)
library(ggplot2)
library(gganimate)
library(av)
library(gifski)
library(foreach)
library("DAAG")
library(DT)
```

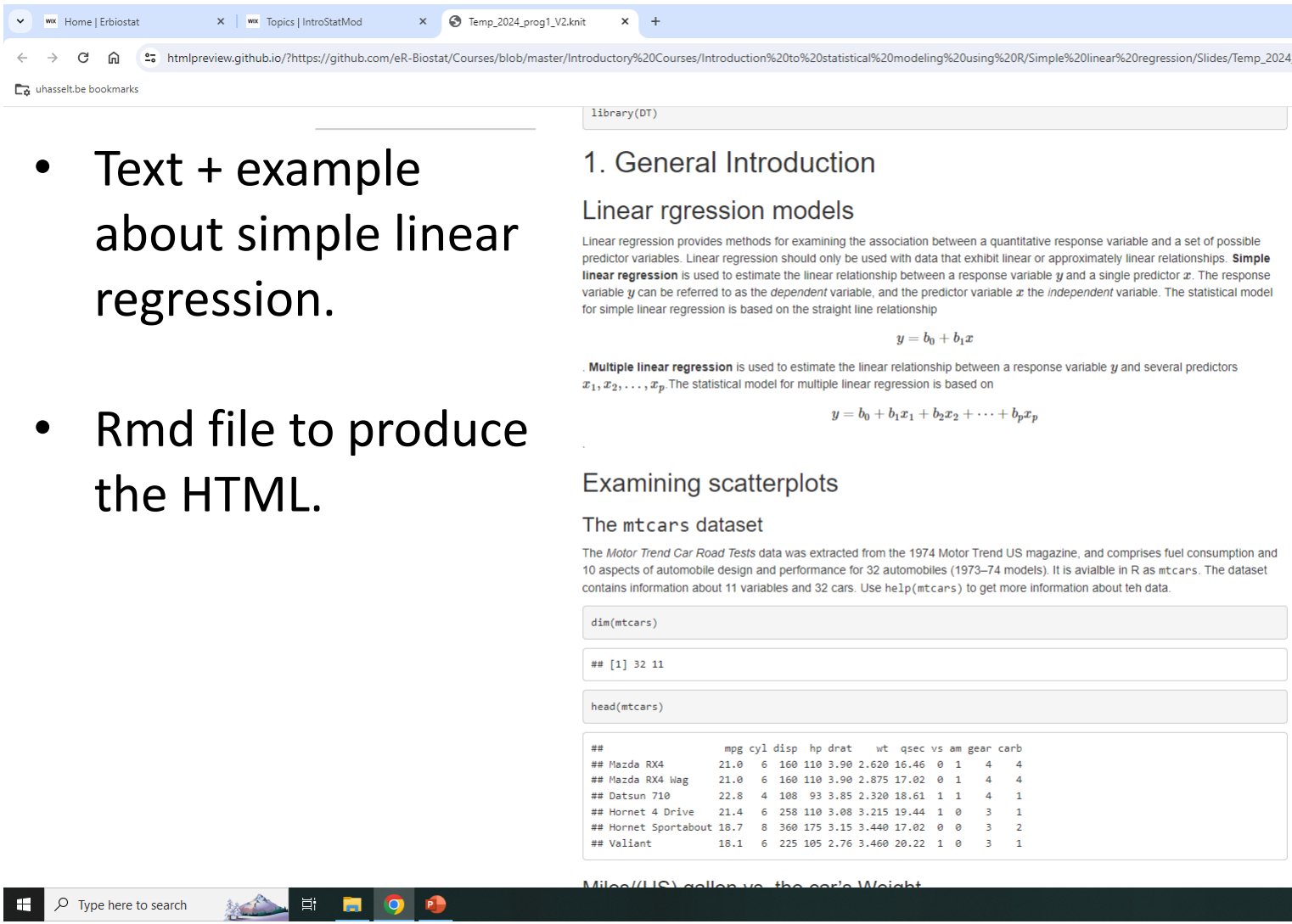
1. General Introduction

Linear regression models

Type here to search | ENG 7:11 8/02/2024

A part of the website of the course

- Text + example about simple linear regression.
- Rmd file to produce the HTML.



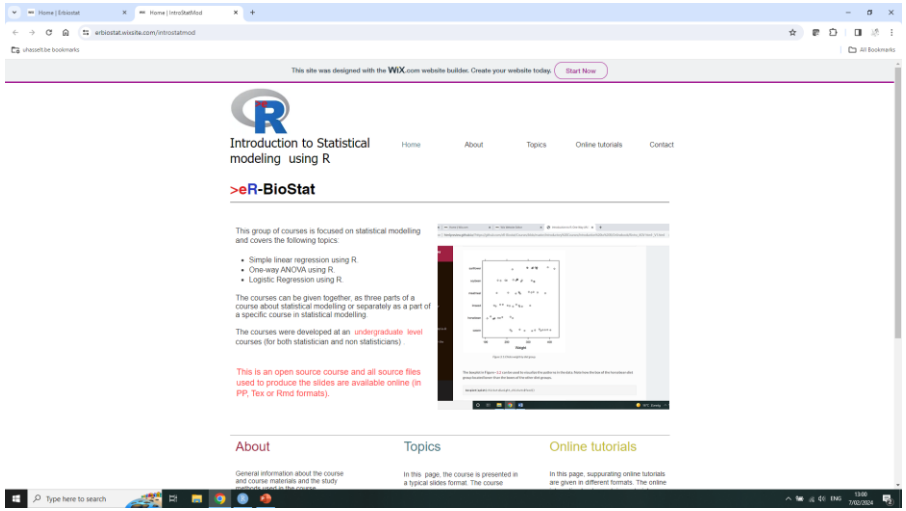
Development of E-learning materials using R markdown

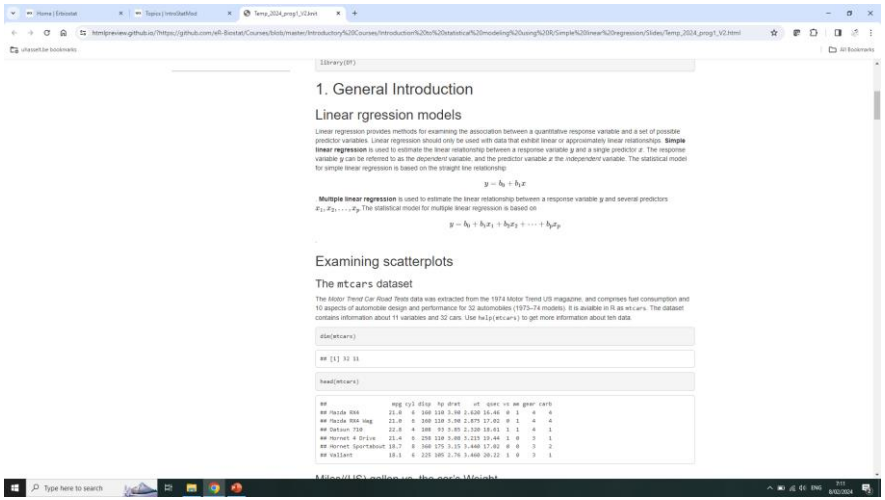
Part 6: developing the content

R Program: er_prog4_SA_2024.Rmd

Content on the website

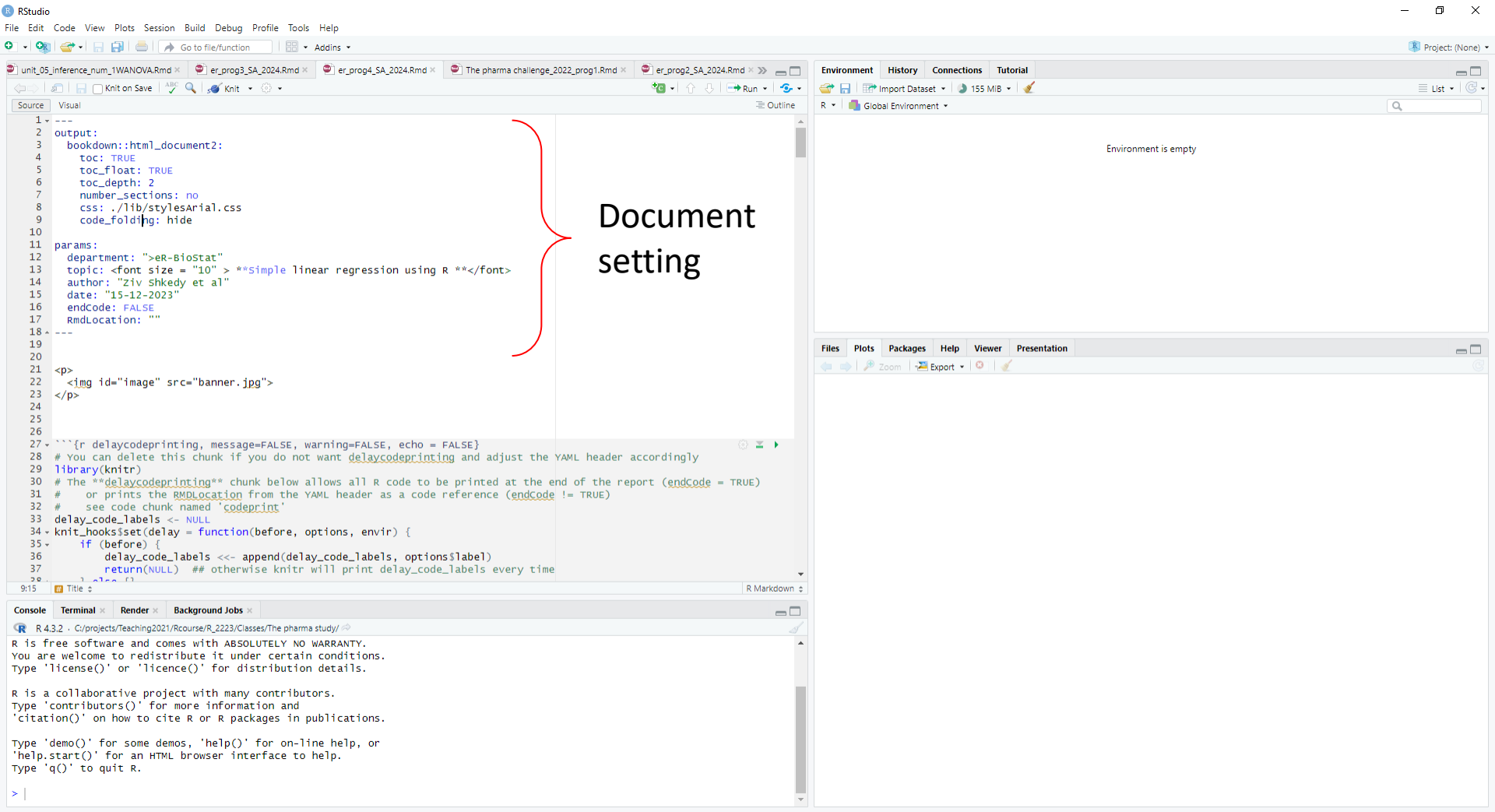
- The Rmd file for the online book: the program and output.





Can be produced using the Rmd program: er_prog4_SA_2024.Rmd

The Rmd program



The HTML output

C:/projects/Teaching2021/Rcourse/R_2223/Classes/The pharma study/er_prog4_SA_2024.html

er_prog4_SA_2024.html | Open in Browser | Find | Publish

Code

15-12-2023 >eR-BioStat

Simple linear regression using R

Ziv Shkedy et al

Show

1. General Introduction

Linear rgression models

Linear regression provides methods for examining the association between a quantitative response variable and a set of possible predictor variables. Linear regression should only be used with data that exhibit linear or approximately linear relationships. **Simple linear regression** is used to estimate the linear relationship between a response variable y and a single predictor x. The response variable y can be referred to as the *dependent* variable, and the predictor variable x the *independent* variable. The statistical model for simple linear regression is based on the straight line relationship

$$y = b_0 + b_1 x$$

. **Multiple linear regression** is used to estimate the linear relationship between a response variable y and several predictors $x_1, x_2, \ldots, x_p$.The statistical model for multiple linear regression is based on

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

.

Examining scatterplots

The `mtcars` dataset

The *Motor Trend Car Road Tests* data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). It is avialble in R as `mtcars`. The dataset contains information about 11 variables and 32 cars. Use `help(mtcars)` to get more information about teh data.

Show

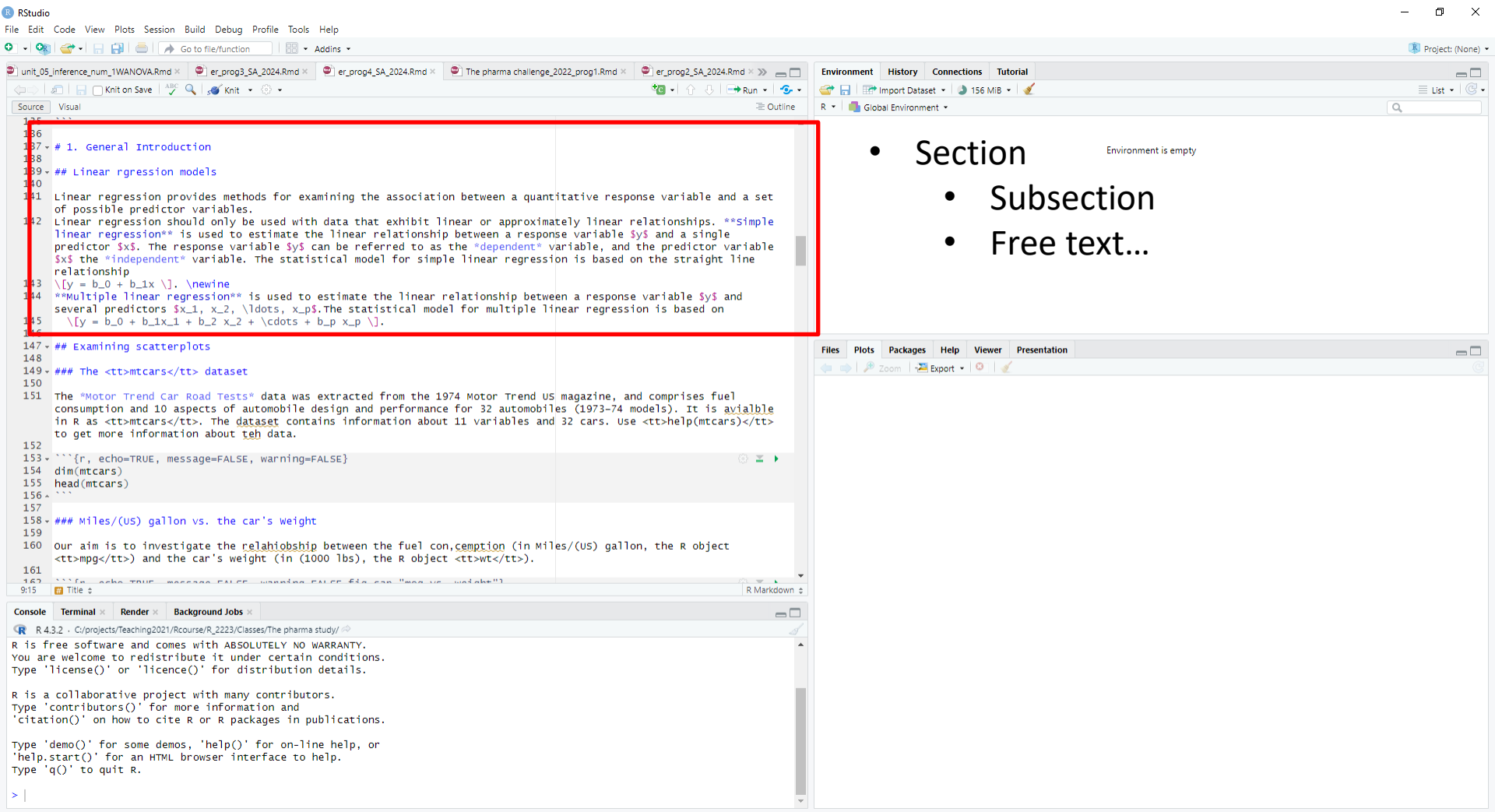
```
## [1] 32 11
```

Show

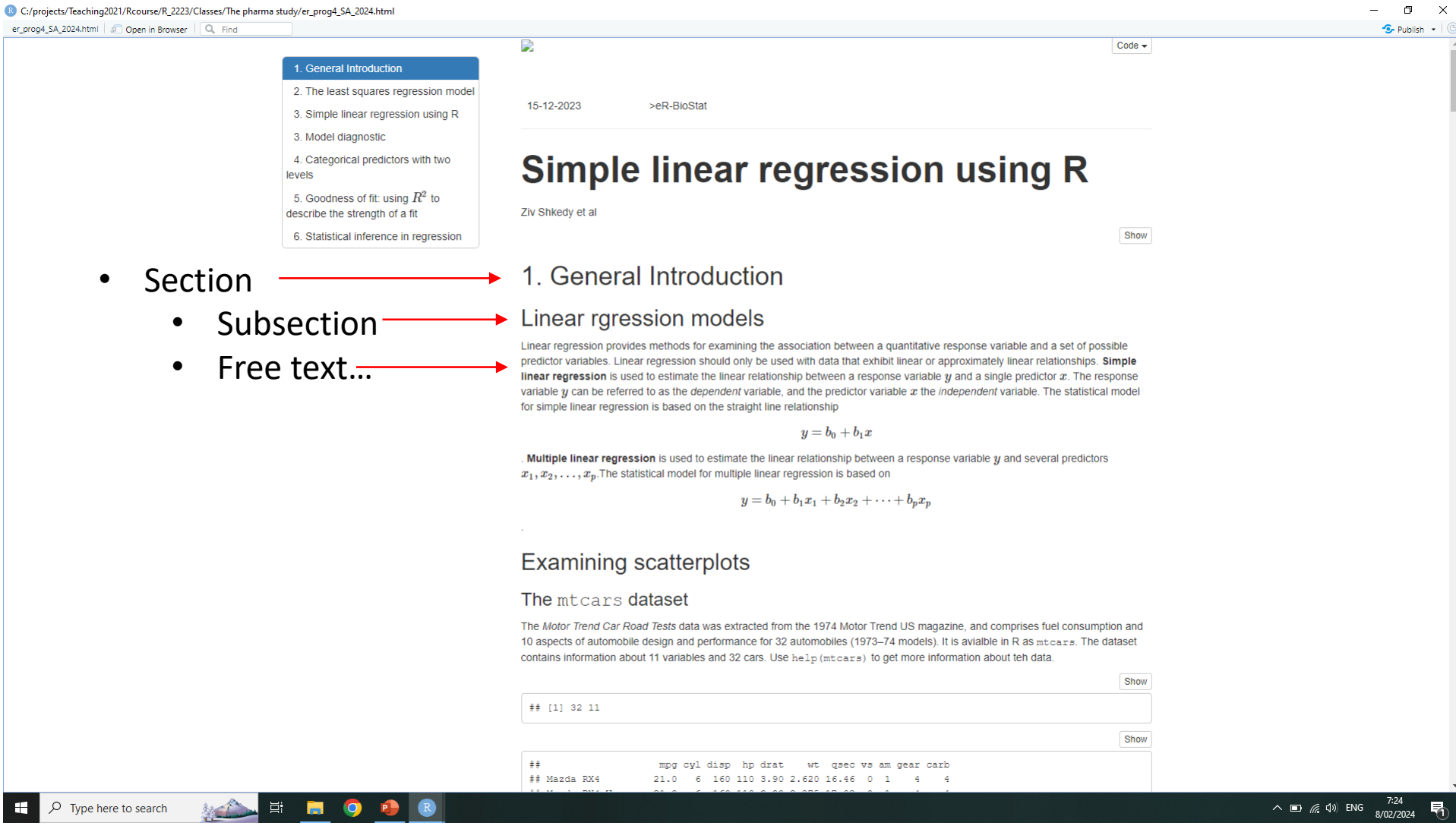
```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
```

Type here to search ENG 7:24 8/02/2024

The Rmd program

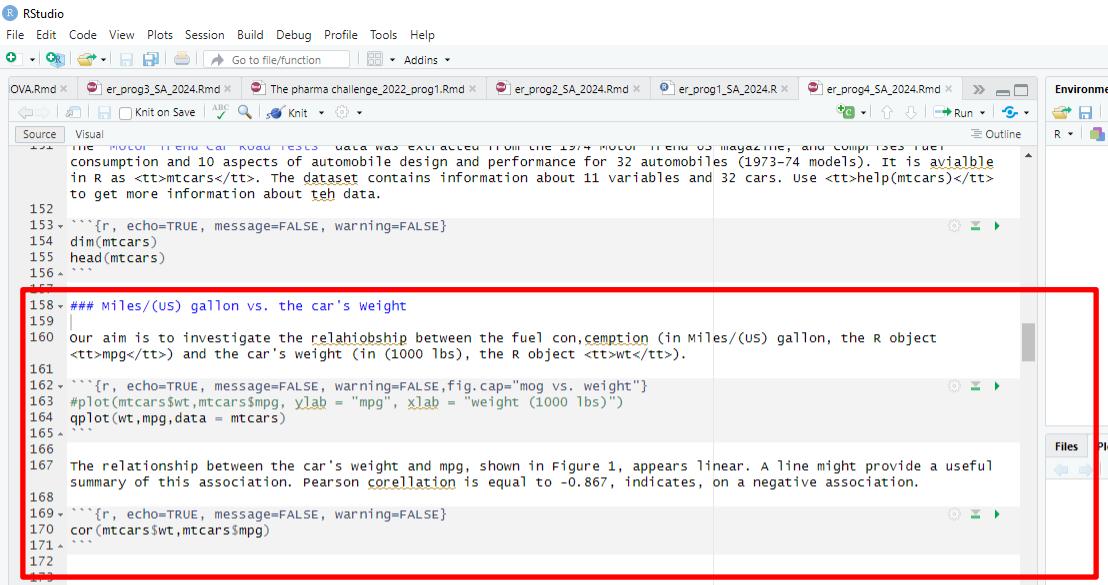


The HTML output



- Section
  - Subsection
  - Free text…

The Rmd program



- Analysis in R:
  - Scatterplot + correlation.

The HTML output

C:/projects/Teaching2021/Rcourse/R_2223/Classes/The pharma study/er_prog4_SA_2024.html

er_prog4_SA_2024.html | Open in Browser | Find | Publish


```
## Datsun 710        22.8   4  108   93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Miles/(US) gallon vs. the car's Weight

Our aim is to investigate the relahiobship between the fuel con,cemption (in Miles/(US) gallon, the R object `mpg`) and the car's weight (in (1000 lbs), the R object `wt`).

Show

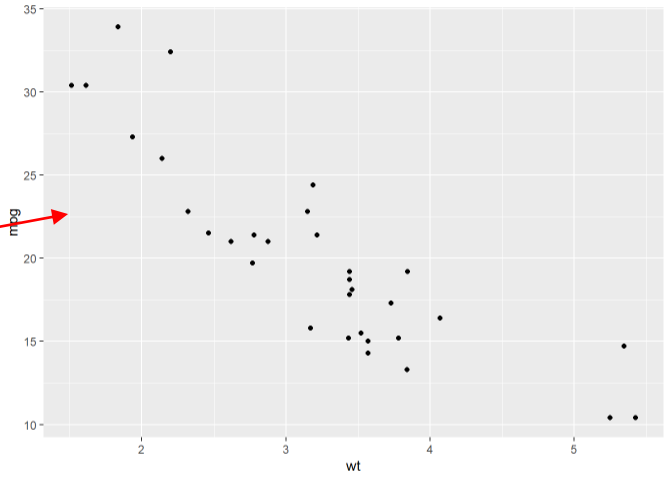


Figure 1: mog vs. weight

The relationship between the car's weight and mpg, shown in Figure 1, appears linear. A line might provide a useful summary of this association. Pearson corellation is equal to -0.867, indicates, on a negative association.

Show

```
## [1] -0.8676594
```

2. The least squares regression model

Model assumptions

There are 4 assumptions that should be satisfied for a line to be considered a reasonable approximation for a relationship shown in a

- Analysis in R:
  - Scatterplot.
  - Correlation.

- The example:
  - `mtcars` data.
  - Not in Dave & Julie files.

Text without the code.

Type here to search | ENG | 7:26 8/02/2024

The HTML output

C:/projects/Teaching2021/Rcourse/R_2223/Classes/The pharma study/er_prog4_SA_2024.html

er_prog4_SA_2024.html | Open in Browser | Find | Publish


```
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Miles/(US) gallon vs. the car's Weight

Our aim is to investigate the relahiobship between the fuel con,cemption (in Miles/(US) gallon, the R object `mpg`) and the car's weight (in (1000 lbs), the R object `wt`).

Hide

```
#plot(mtcars$wt,mtcars$mpg, ylab = "mpg", xlab = "weight (1000 lbs)")
qplot(wt,mpg,data = mtcars)
```

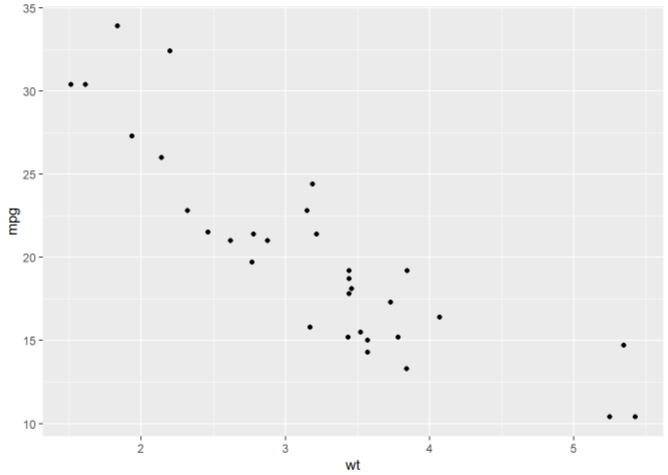


Figure 1: mog vs. weight

The relationship between the car's weight and mpg, shown in Figure 1, appears linear. A line might provide a useful summary of this association. Pearson corellation is equal to -0.867, indicates, on a negative association.

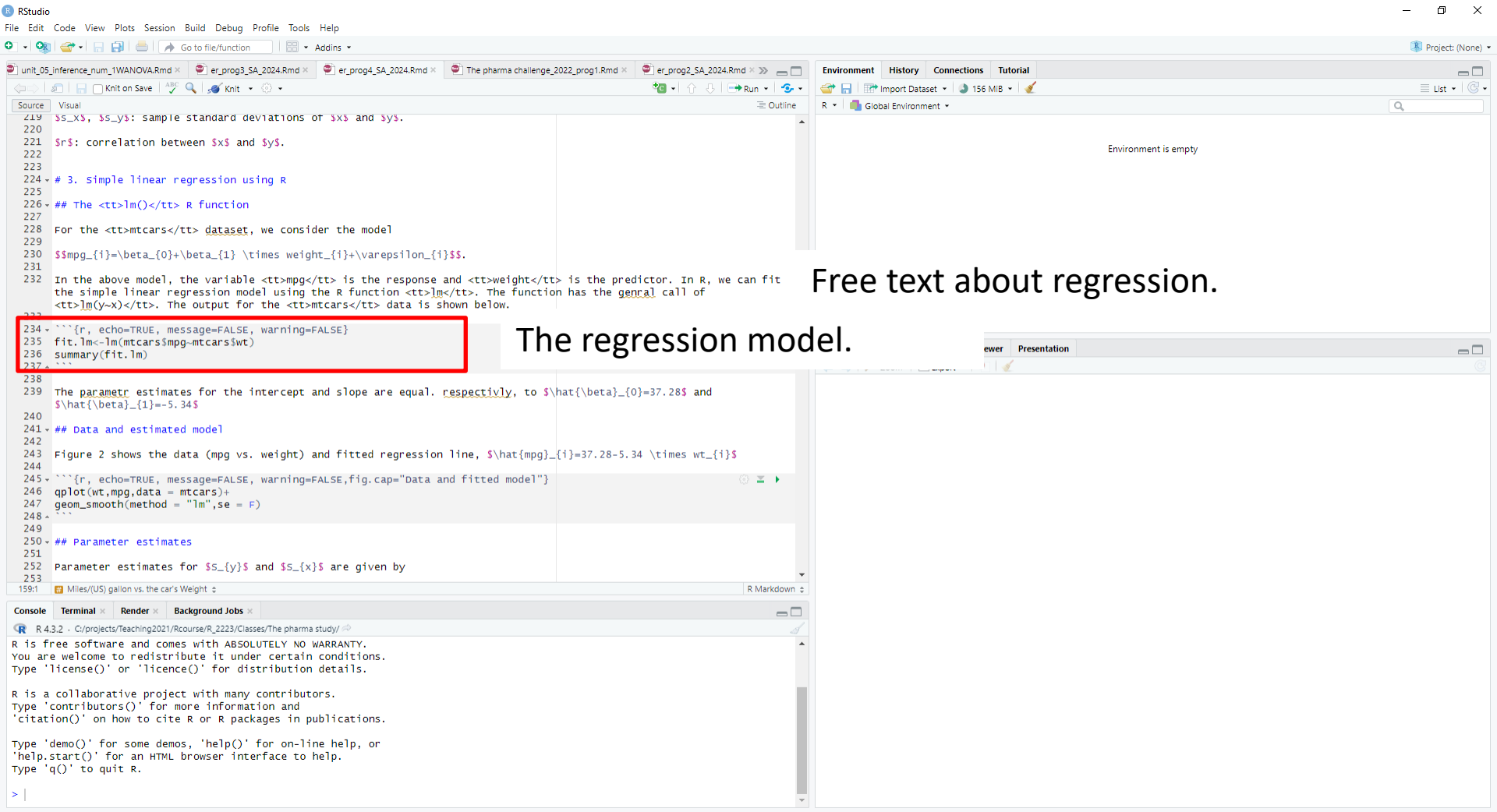
Show

```
## [1] -0.8676594
```

2. The least squares regression model

Text with the code.

The Rmd program



Free text about regression.

The regression model.

The HTML output

C:/projects/Teaching2021/Rcourse/R_2223/Classes/The pharma study/er_prog4_SA_2024.html

er_prog4_SA_2024.html | Open in Browser | Find | Publish

r: correlation between x and y.

3. Simple linear regression using R

The `lm()` R function

For the `mtcars` dataset, we consider the model

$$mpg_i = \beta_0 + \beta_1 \times weight_i + \varepsilon_i$$

In the above model, the variable `mpg` is the response and `weight` is the predictor. In R, we can fit the simple linear regression model using the R function `lm`. The function has the genral call of `lm(y~x)`. The output for the `mtcars` data is shown below.

Hide

```
fit.lm<-lm(mtcars$mpg~mtcars$wt)
summary(fit.lm)
```

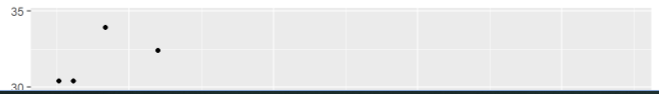
```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars$wt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## mtcars$wt    -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

The parametr estimates for the intercept and slope are equal. respectivly, to $\hat{\beta}_0 = 37.28$ and $\hat{\beta}_1 = -5.34$

Data and estimated model

Figure 2 shows the data (mpg vs. weight) and fitted regression line, $\widehat{mpg}_i = 37.28 - 5.34 \times wt_i$

Show



- The regression model:
  - Code.
  - Output.

Free text about regression.

Type here to search | ENG | 7:27 8/02/2024

Short summary

- Use Rmd to create the document.
- Upload the document online as part of the course.
- Text: written by Dave & Jullie (and available online).
- Examples written by Ziv (and available online).

Development of E-learning materials using R markdown

The >eR-BioStat approach

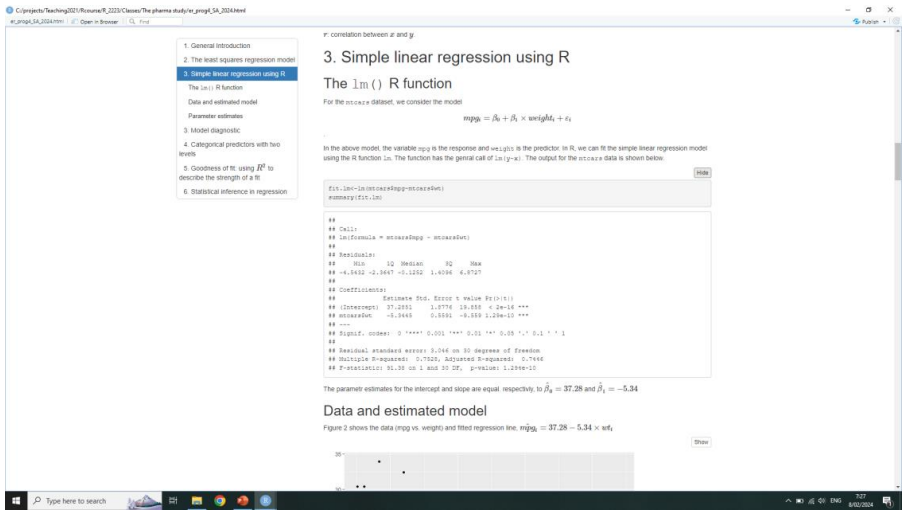
How can we create a course online??

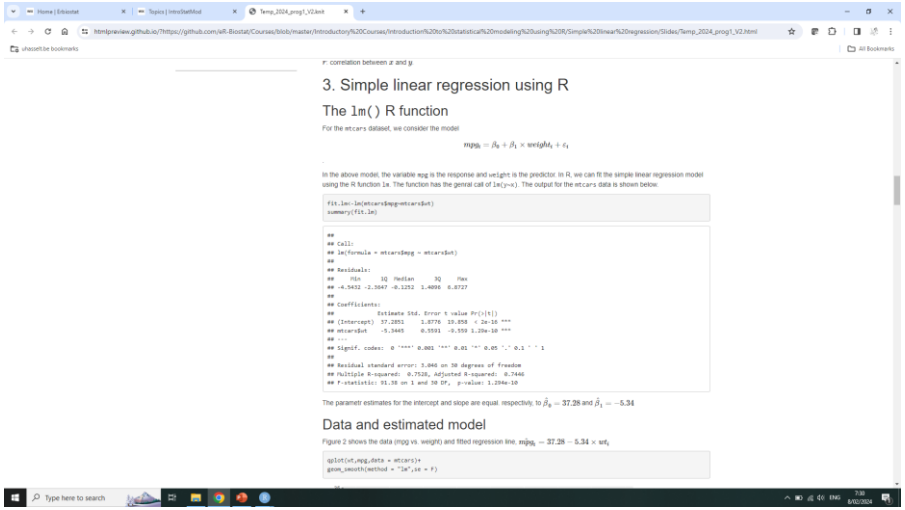
The >eR-BioStat approach

- Content development:
  - R/R markdown.
- Storage:
  - Github.
- Website development:
  - WIX.

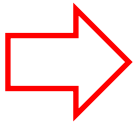
All have publicly available and free versions.

From a laptop to a website..



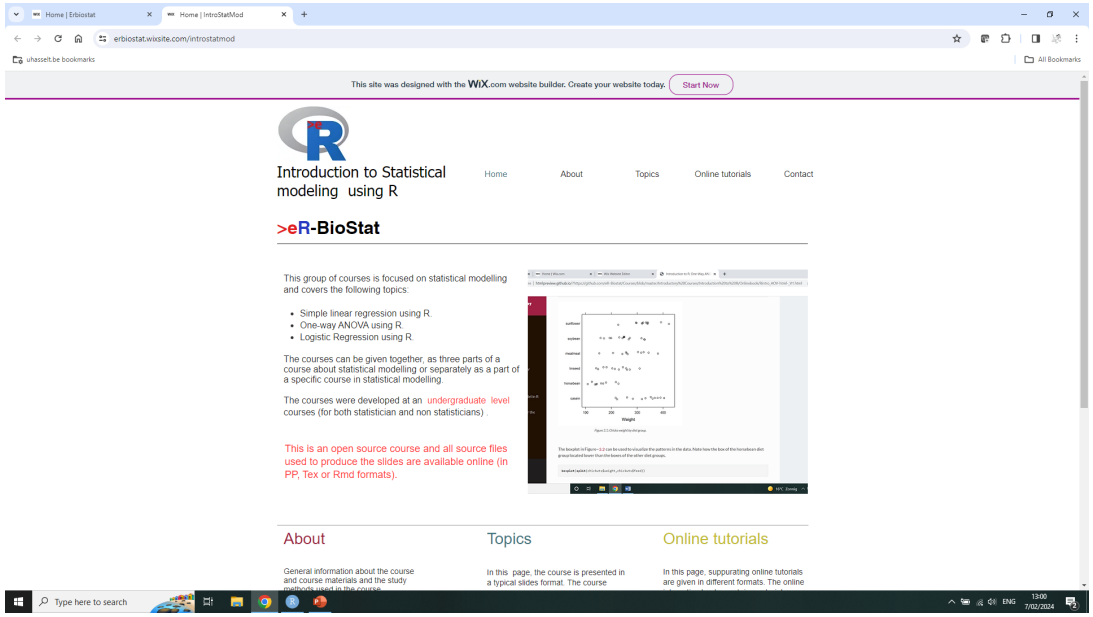


The HTML file on the laptop….
Can be produced using the Rmd program: er_prog4_SA_2024.Rmd.



The HTML file as a part of the course's website.

Introduction to statistical modeling using R



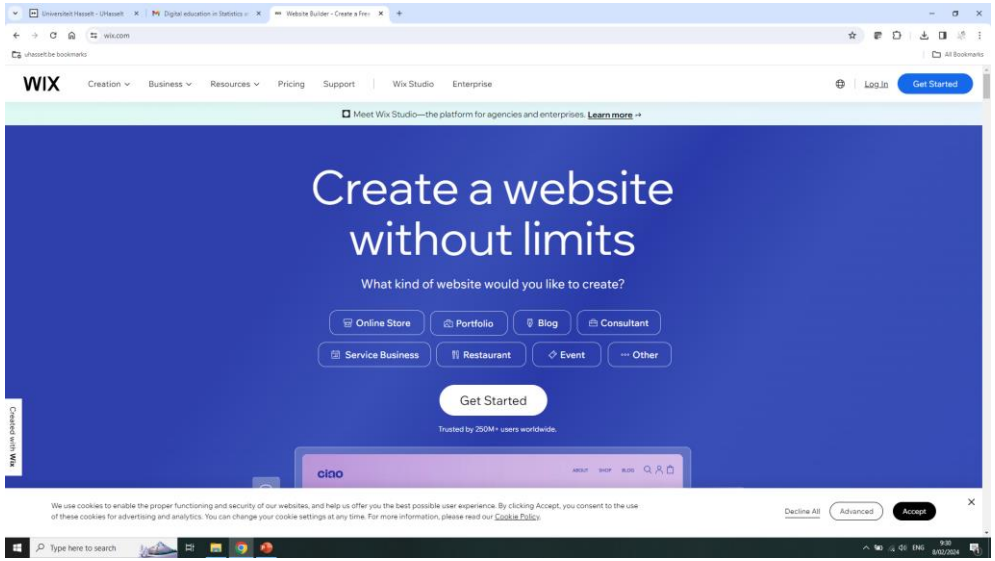
- Website.
- Storage of materials:
  - Slides.
  - Programs.
  - Datasets.
  - Etc.
- Software for the analysis ?
- Storage space & cost ?

https://erbiostat.wixsite.com/introstatmod

- Our approach: bring costs to zero…

How can I develop a website for my course ?

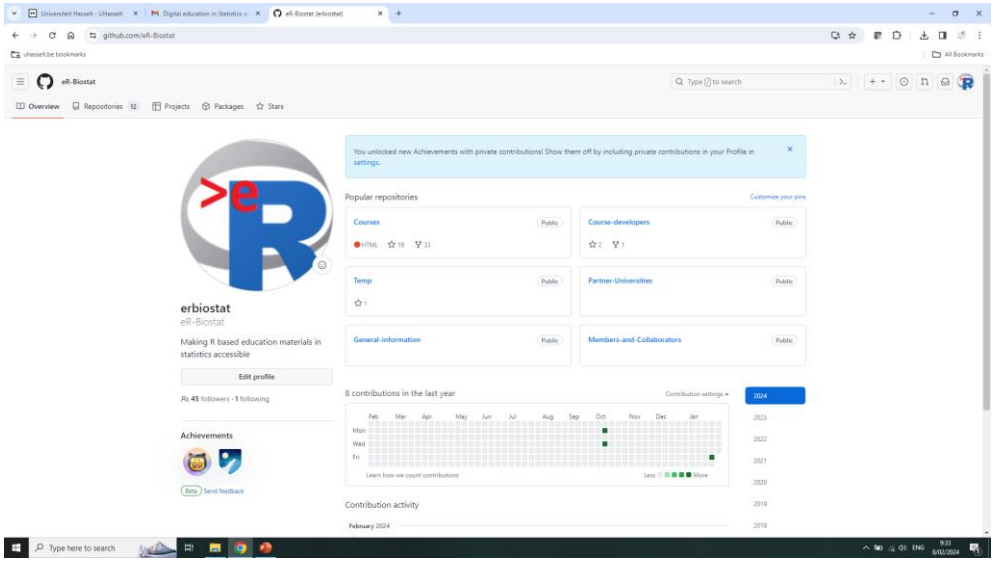
- Websites for the courses were developed using WIX.
  - Free.
  - Easy to use and learn.



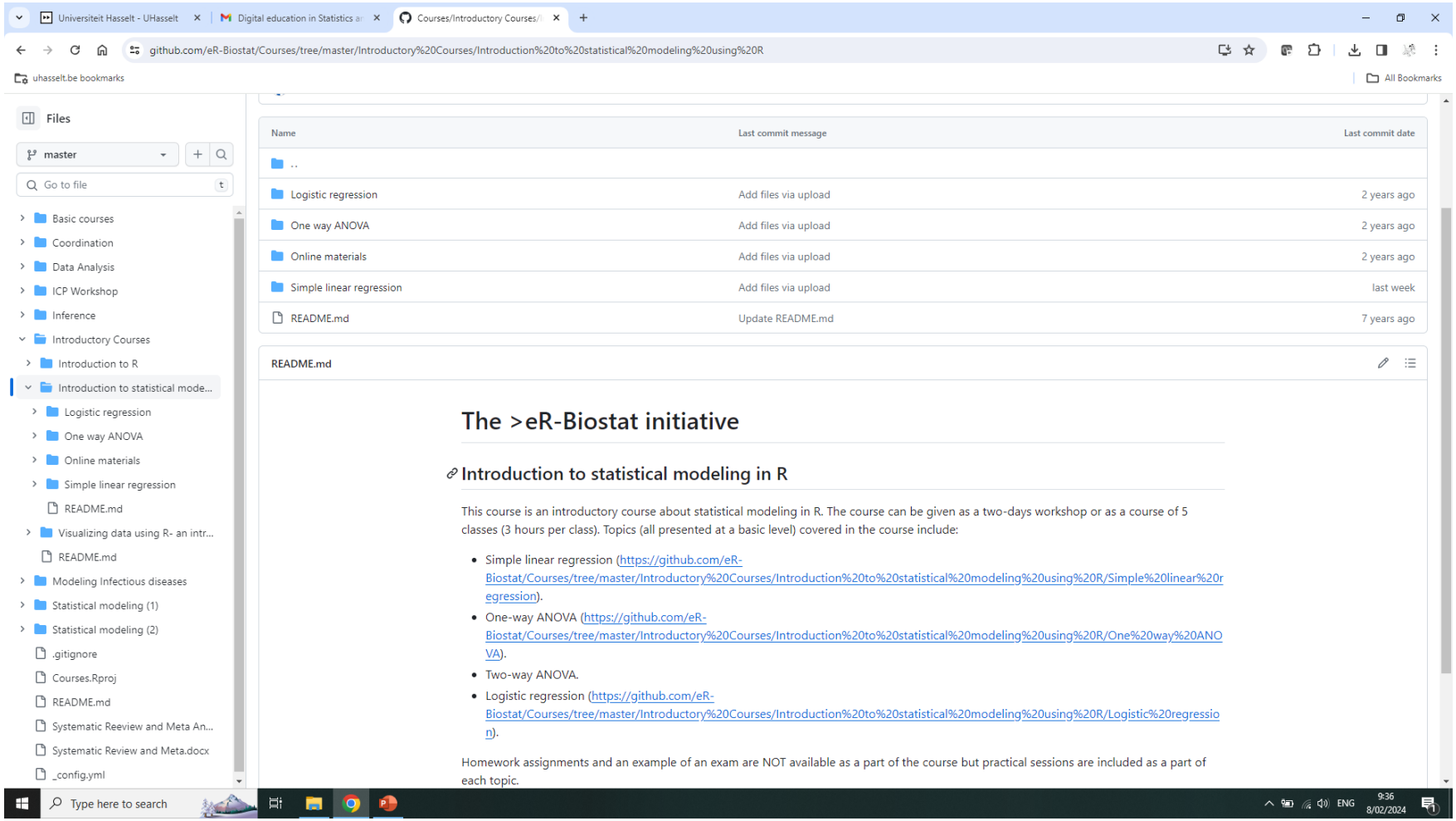
https://www.wix.com/

Where can I store the course materials (slides, programs, notes…..?

- All course materials are stored on Github:
  - Slides.
  - Programs.
  - Datasets.
  - R markdown programs.
  - HTML files….
- Free and unlimited.



https://github.com/eR-Biostat

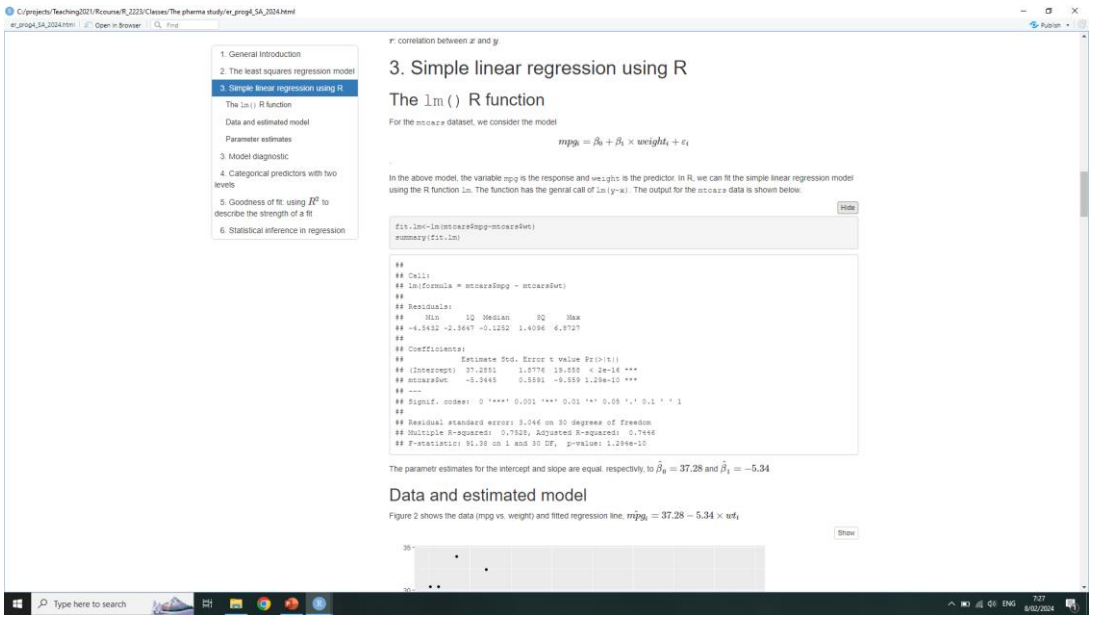


https://github.com/eR-Biostat/Courses/tree/master/Introductory%20Courses

Software for the analysis

- We use R but……

Summary



- Content produced using R markdown.
- Store online on GitHub.
- Presented online in a website developed using WIX.
- Data analysis using R.

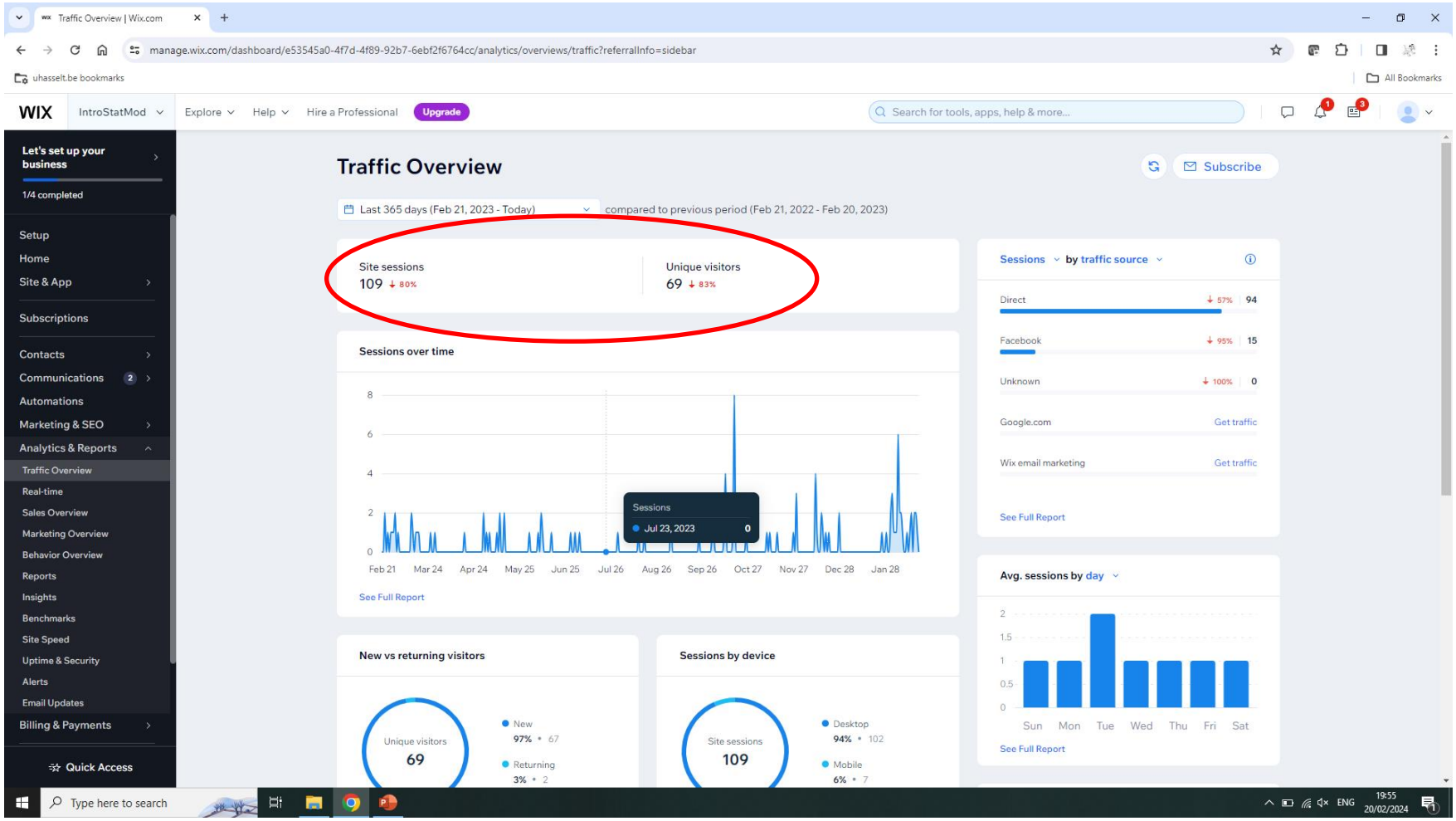
- Costs=0 !!!

What is available to the users ?

- Who are the users ? Teachers & students & others.
- What is available ? Everything.
- An open source approach:
  - Slides.
  - R programs for examples.
  - R programs for the slides.
  - PowerPoints files.
  - HTML files.

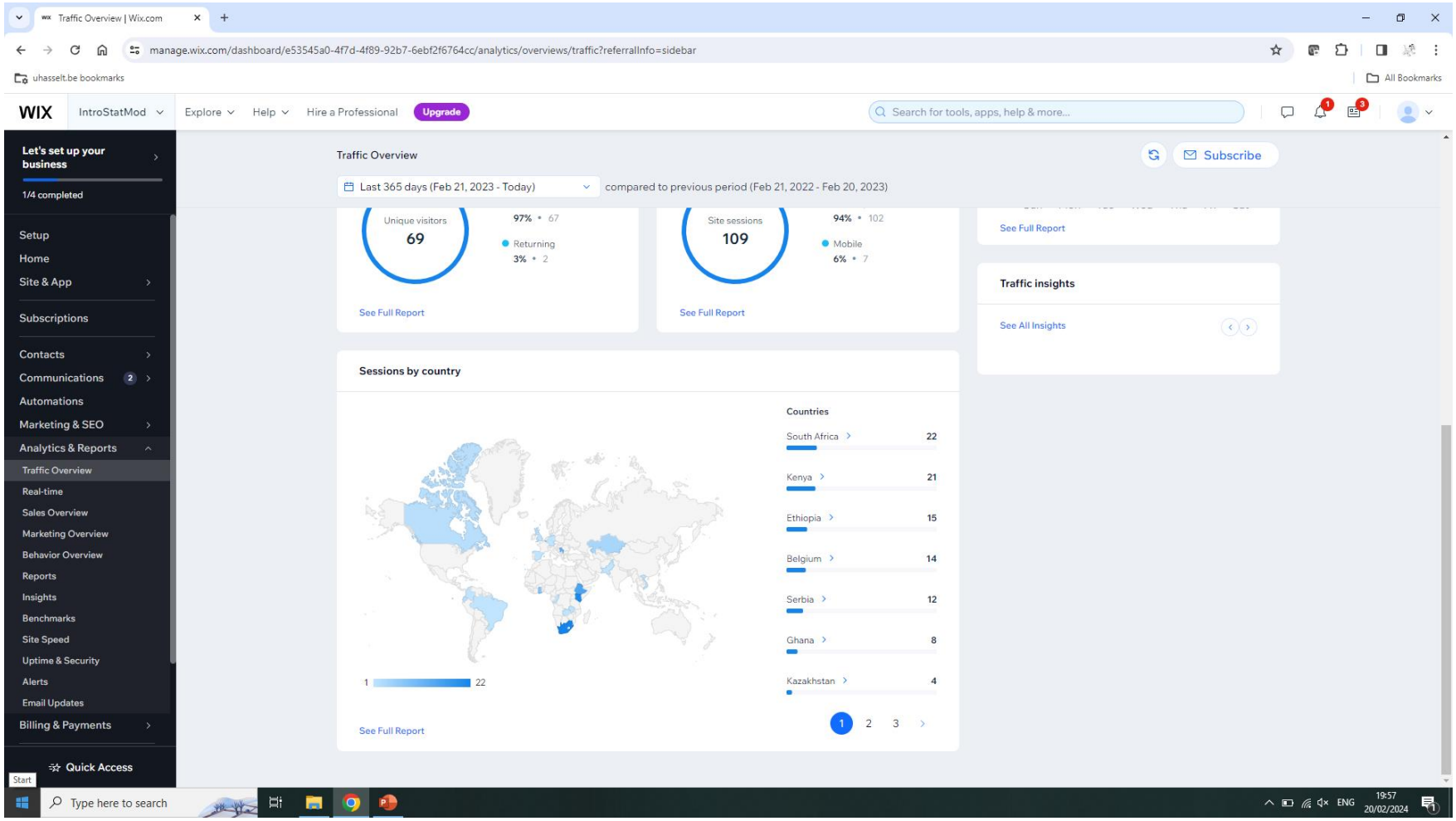
Users

21/02/23-20/02/24



Users

21/02/23-20/02/24



Discussion

- R Studio + R markdown:
- Easy to use.
- Text + code.
- Output:
  - Standard: HTML, PDF, DOC.
  - Advanced: HTML.