

## HOMEWORK 2

## MATH 484-564, REGRESSION

DUE SEPTEMBER 13TH 2024, FRIDAY, 11:59PM, SEE THE SUBMISSION INSTRUCTIONS ON CANVAS.

1. (6 points) Please install R and R Studio. Refer to Figures 1 and 2 below for guidance. You do not need to provide proof of installation. If you have answered at least one question from Questions 2, 4, or 5, you will automatically receive these 6 points.

### Installation of R and RStudio

Go to

<https://www.rstudio.com/products/rstudio/download/>

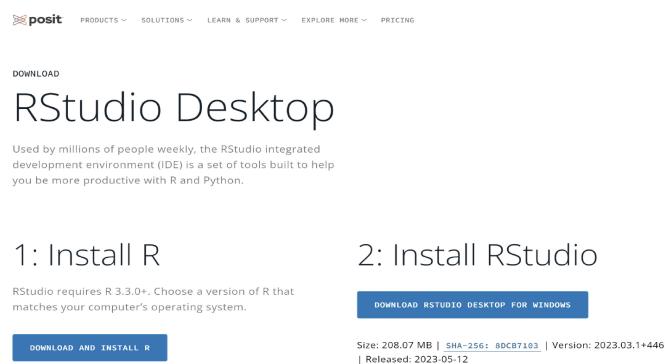


Figure 1: Download R and RStudio by following the instructions on the website.

### RStudio Interface

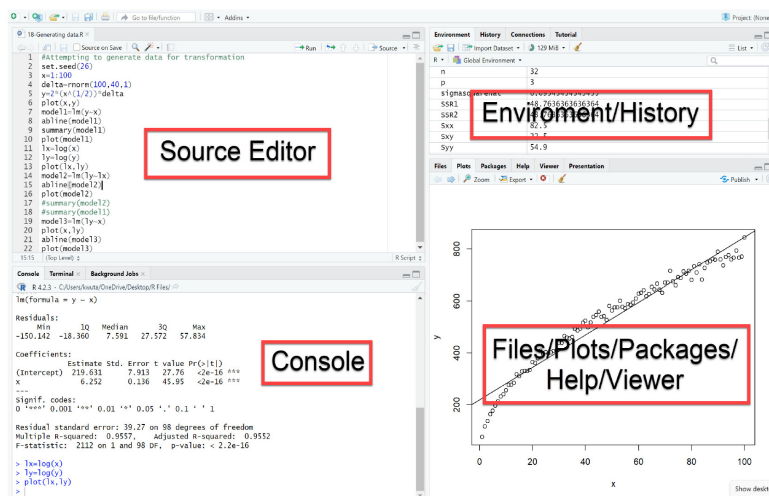


Figure 2: This is how the the dashboard of R looks like.

2. (6 points) For this and some future assignments, you will need to use R Markdown to submit your answers. R Markdown is a format that allows you to combine code and narrative in one document. Please follow the instructions below.

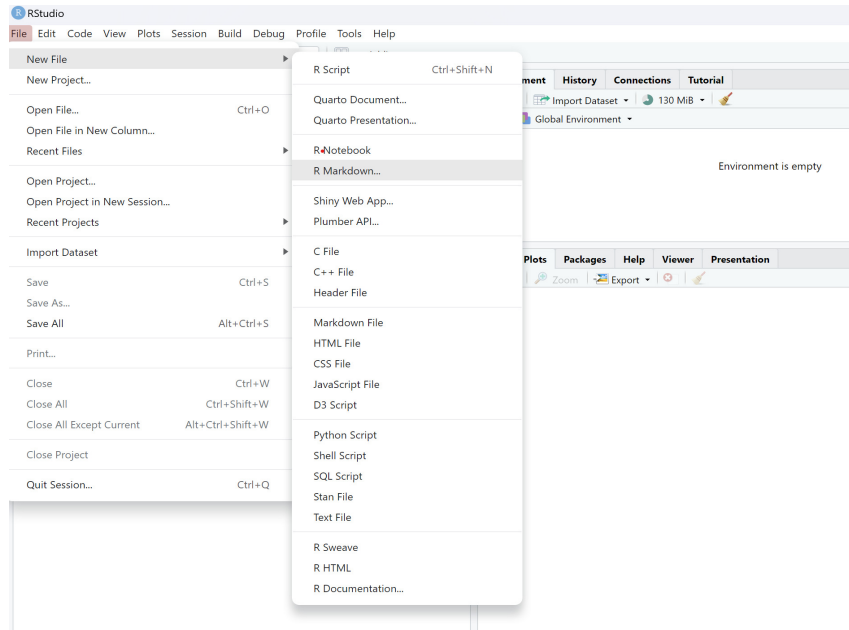


Figure 3: Go to your RStudio, click on the “File” option, then go to “New File” and choose “R Markdown”.

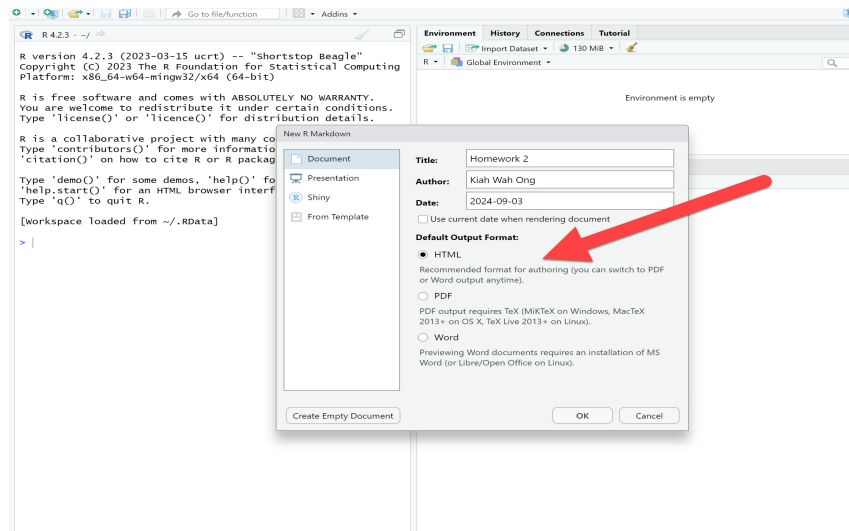


Figure 4: You can choose to output your document as either HTML or PDF. Select HTML as your output format, run the document, and you will see the result as shown in Figure 5.

## Homework 2

Kiah Wah Ong  
2024-09-03

### R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

	speed	dist
## Min.	: 4.0	Min. : 2.00
## 1st Qu.	:12.0	1st Qu.: 26.00
## Median	:15.0	Median : 36.00
## Mean	:15.4	Mean : 42.98
## 3rd Qu.	:19.0	3rd Qu.: 56.00
## Max.	:25.0	Max. :120.00

### Including Plots

You can also embed plots, for example:

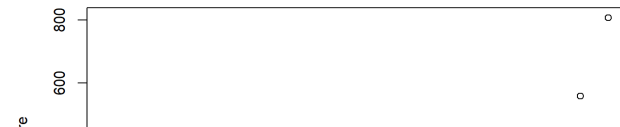


Figure 5: After compilation, you will see the default HTML conversion of your first R Markdown document.

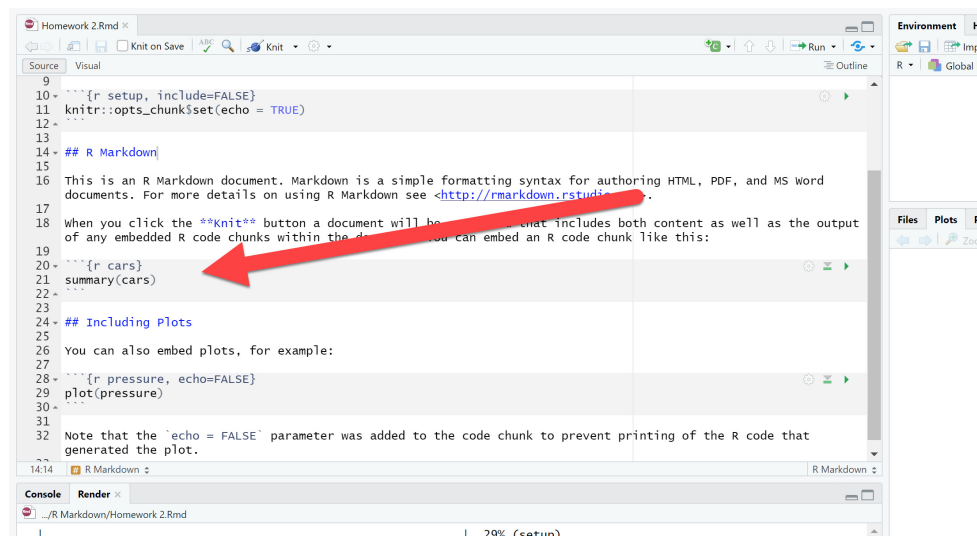
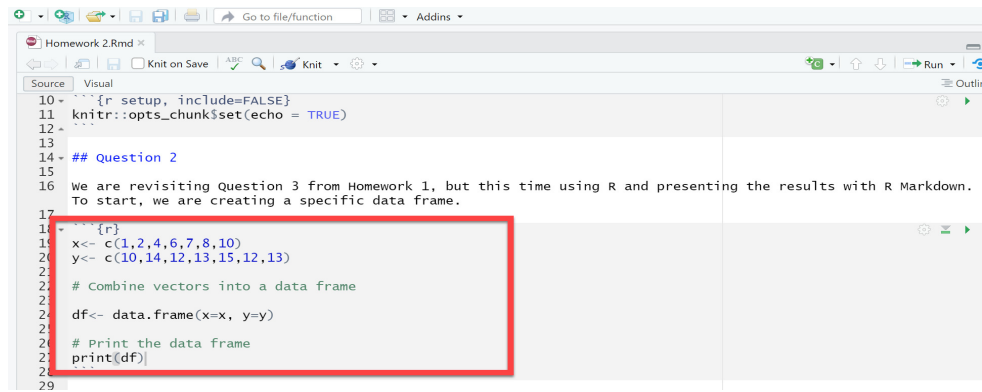


Figure 6: Return to your R Markdown document, and you'll find the section marked by the red arrow. This is where the R code is compiled to generate the summary that appears in your HTML document.

The steps provided should give you a clear understanding of how to use R Markdown. Now, let's return to the task at hand. You'll be revisiting Question 3 from Homework 1, but this time, you'll use R and present your results using R Markdown. Begin by creating a specific data frame. For guidance, refer to the example shown in Figure 7. You will receive the 6 points for this question by setting up the data frame.



```
10 {r setup, include=FALSE}
11 knitr::opts_chunk$set(echo = TRUE)
12
13
14 ## Question 2
15
16 We are revisiting Question 3 from Homework 1, but this time using R and presenting the results with R Markdown.
17 To start, we are creating a specific data frame.
18
19 {r}
20 x<- c(1,2,4,6,7,8,10)
21 y<- c(10,14,12,13,15,12,13)
22
23 # Combine vectors into a data frame
24
25 df<- data.frame(x=x, y=y)
26
27 # Print the data frame
28 print(df)
29
```

Figure 7: Use the code provided above to initialize the values of  $x$  and  $y$  that were used in Question 3 of Homework 1. Remember to “Knit” when compile using the R Markdown to convert the document to HTML.

3. (6 points) For this question, write your answer on paper, scan it, and submit the scanned document along with your HTML file (for Question 2, 4, and 5). Alternatively, you can use LaTeX, Word, or another word processing application to write your answer, but ensure that it is saved as a PDF before submission.

In statistics, we have the follow:

- Sum of squares of deviations of  $x$  values

$$s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

- Sum of squares of deviations of  $y$  values

$$s_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

- Sum of the product of deviations of  $x$  and  $y$  values

$$s_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Prove that these sum of squares can also be written as

(i)

$$s_{xx} = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

(ii)

$$s_{yy} = \sum_{i=1}^n y_i^2 - n\bar{y}^2$$

(iii)

$$s_{xy} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$$

4. (6 points) Back to your R Markdown.

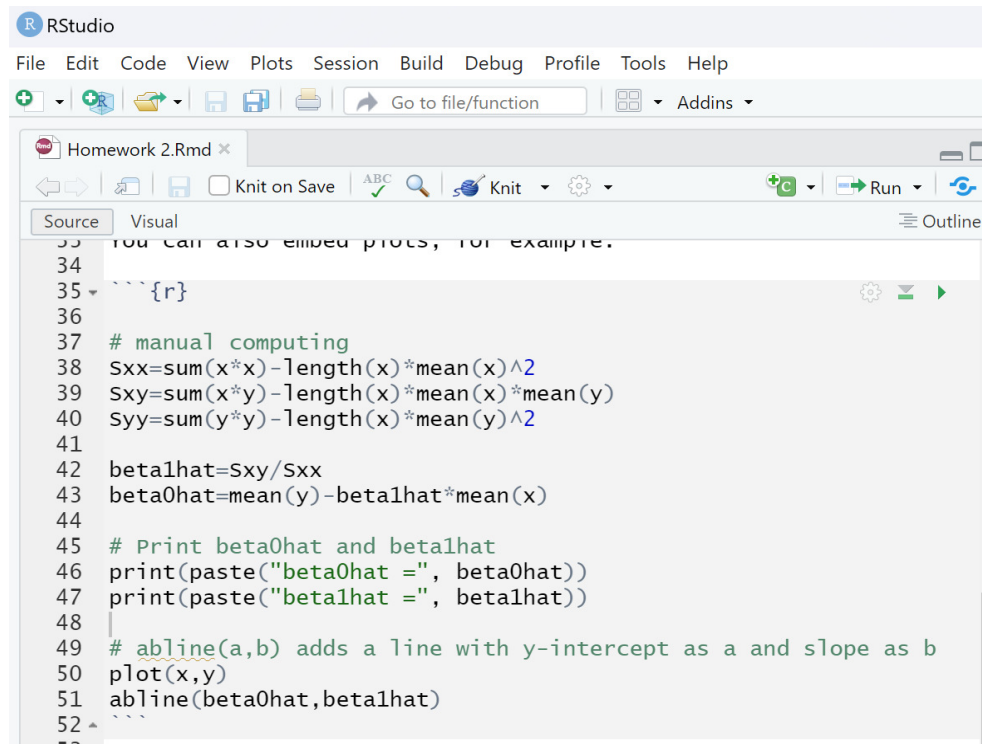
(a) Apply the knowledge you gained from Question 3 to compute the following using R:

(i)  $S_{xx}$ ,  $S_{xy}$  and  $S_{yy}$ .

(ii)  $\hat{\beta}_1$  and  $\hat{\beta}_0$ .

(b) Plot the points  $(x, y)$  and using the `abline()` function, add the least squares line to the plot.

See the figure below for reference.



```
33 # you can also embed plots, for example.
34
35 ```{r}
36
37 # manual computing
38 Sxx=sum(x*x)-length(x)*mean(x)^2
39 Sxy=sum(x*y)-length(x)*mean(x)*mean(y)
40 Syy=sum(y*y)-length(x)*mean(y)^2
41
42 beta1hat=Sxy/Sxx
43 beta0hat=mean(y)-beta1hat*mean(x)
44
45 # Print beta0hat and beta1hat
46 print(paste("beta0hat =", beta0hat))
47 print(paste("beta1hat =", beta1hat))
48
49 # abline(a,b) adds a line with y-intercept as a and slope as b
50 plot(x,y)
51 abline(beta0hat,beta1hat)
52 ```
```

Figure 8: Use the code provided above to answer Question 4.

5. (6 points) Mastering the use of R for regression analysis is a crucial skill for this course. We'll begin by working with a data set named **Advertising**, which contains information on media advertising expenditures (TV, radio, and newspapers) and corresponding product sales across 200 markets. The advertising budget is reported in thousands of dollars, and the sales figures are in thousands of units. In this exercise, you'll create least squares regression lines using R's built-in **lm()** function. Follow the steps below to complete this task.
- (i) Place the Excel file named **Advertising** into the same folder as your R Markdown document. This ensures that R can easily access the file from the correct directory.
  - (ii) Write the code below:

```
54 ## Question 5
55
56 ```{r}
57 df<-read.csv("Advertising.CSV", header=TRUE, sep=",")
58
59 x1<-df$TV
60 x2<-df$radio
61 x3<-df$newspaper
62 y<-df$sales
63
64 model1<-lm(y ~ x1)
65 options(digits = 9)
66 summary(model1)
67 plot(x1,y)
68 abline(model1)
```

Remarks:

Here the line of code in line 57 is used to load data from a CSV file into R.

In line 57, the `header=TRUE` is an argument that tells R that the first row of the CSV file contains the names of the columns. If this is set to `TRUE`, R will use the first row as the header row, which means the column names will be used as variable names in the data frame.

In line 57, `sep=","` specifies that the values in the CSV file are separated by commas. Since CSV stands for Comma-Separated Values, this is the default setting, but it's good practice to include it to make sure the data is read correctly.

Lines 59, 60, 61, and 62 renamed the variables TV, radio, newspaper and sales into  $x_1$ ,  $x_2$ ,  $x_3$  and  $y$  respectively.

In Line 64, we use the built-in regression function. Unlike in Problem 4 of this homework, the least squares regression line will be calculated automatically, so you won't need to compute values like  $S_{xx}$ ,  $S_{xy}$  and  $S_{yy}$ .

In line 65, we set the number of decimal points in the printout.

Line 66 give the summary of the regression analysis, while line 67 and 68 give the instruction to plot the points and then plot the least squares regression line.

- (iii) Regress  $y$  onto  $x_2$ , then regress  $y$  onto  $x_3$  using the steps in (ii). See the screen shot below for reference.

```
53
54 ## Question 5
55
56 ```{r}
57 df<-read.csv("Advertising.CSV", header=TRUE, sep=",")
58
59 x1<-df$TV
60 x2<-df$radio
61 x3<-df$newspaper
62 y<-df$sales
63
64 model1<-lm(y ~ x1)
65 options(digits = 9)
66 summary(model1)
67 plot(x1,y)
68 abline(model1)
69
70
71 model2<-lm(y ~ x2)
72 options(digits = 9)
73 summary(model2)
74 plot(x2,y)
75 abline(model2)
76
77
78 model3<-lm(y~x3)
79 summary(model3)
80 plot(x3,y)
81 abline(model3)
82
83 ```
```