Week 1 - AYU - Pod

Table of Contents

*This document will help you install R, the most popular statistical programming language, and Rstudio, the most popular programming editor for R, into your computer. You will also learn about how R can be used as a powerful calculator and about how to import data and implement linear regression. Follow the section 1 to 3 to do the AYU Questions in section 4.*

###### Question 1

Install R and Rstudio

*How to do it?*

* Download R at [this link](https://cran.r-project.org/bin/windows/base/)
* Download Rstudio at [this link](https://posit.co/download/rstudio-desktop/)
* Install R
* Install Rstudio

###### Question 2

Run linear model on the following data

| x | 2 | 3 | 5 | 6 | 1 | 9 | 10 | 15 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| y | 1 | 4 | 6 | 4 | 4 | 3 | 20 | 25 |

*How to do it?*

* Open Rstudio
* In the console, create two vectors and by typing in

x = c(2, 3, 5, 6, 1, 9, 10, 15)  
y = c(1, 4, 6, 4, 4, 3, 20, 25)

* To regress on use the lm function

my\_linear\_model = lm(y~x)

* To show the result use

summary(my\_linear\_model)

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.5325 -1.2797 0.3865 3.2580 5.8837   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.7217 3.2754 -0.526 0.61797   
## x 1.5838 0.4224 3.749 0.00952 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.274 on 6 degrees of freedom  
## Multiple R-squared: 0.7009, Adjusted R-squared: 0.651   
## F-statistic: 14.06 on 1 and 6 DF, p-value: 0.009516

## 1. Setup the enviroment

1. Download and Install R at: <https://cran.r-project.org/bin/windows/base/R-4.2.3-win.exe>
2. Download and Install R-Studio at: <https://download1.rstudio.org/electron/windows/RStudio-2023.03.0-386.exe>

## 2. Using R as a calculator

### 2.1 Operate on one vector

R can be used as a powerful calculator. In this example, we will calculate summarized statistics for the following and inputs.

|  | 1 | 2 | 3 | 4 |
| --- | --- | --- | --- | --- |
|  | 1 | 5 | 2 | 6 |

We first define these variables. Open Rstudio, in the console type:

x = c(1,2,3,4)  
y = c(3,5,2,6)

Two variables and can be seen as one dimensional vectors. We now can calculate by sum(x). Type sum(x) in the console.

sum(x)

## [1] 10

can be calculated using mean(x). Type mean(x) in the console.

mean(x)

## [1] 2.5

Similarly, we have median(x) to calculate the median of , sd(x) for the standard deviation and var(x) for the variance of .

median(x)

## [1] 2.5

sd(x)

## [1] 1.290994

var(x)

## [1] 1.666667

R operates vectors on a element-wise manner. For example will add 3 to all the element of

x + 3

## [1] 4 5 6 7

Or will square all element of

x^2

## [1] 1 4 9 16

### 2.2 Operate on two vectors

As seen before, will add elements to elements of and similarly for multiplication (x\*y) and division (x/y)

x + y

## [1] 4 7 5 10

x\*y

## [1] 3 10 6 24

x/y

## [1] 0.3333333 0.4000000 1.5000000 0.6666667

We can compute by simply sum(x\*y)

sum(x\*y)

## [1] 43

The function cor(x,y) calculates the correlation of and .

cor(x,y)

## [1] 0.4242641

Apply your understanding by doing Question 3.

## 3. Simple Linear Regression

### 3.1 Manually input data

To run SLR in R we use the function lm as follows.

lm(y~x)

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Coefficients:  
## (Intercept) x   
## 2.5 0.6

To obtain all the important information of the regression, use

summary(lm(y~x))

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Residuals:  
## 1 2 3 4   
## -0.1 1.3 -2.3 1.1   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.5000 2.4799 1.008 0.420  
## x 0.6000 0.9055 0.663 0.576  
##   
## Residual standard error: 2.025 on 2 degrees of freedom  
## Multiple R-squared: 0.18, Adjusted R-squared: -0.23   
## F-statistic: 0.439 on 1 and 2 DF, p-value: 0.5757

Apply your understanding by doing Question 3.

### 3.2 Outside Data

To import a csv dataset into R, we do the follows, we will use the function read\_csv. This function does not come with R but belongs to the tidyverse package. We first need to install this package. Type the following into Rstudio console (You only need to install it one time)

install.packages('tidyverse')

We are now ready to use `read\_csv`` to import a dataset file. There are two Possibilities.

* If the dataset file is on your computer say the data file data.csv at C:\SRM\Data\data.csv. We use

library(tidyverse)  
d = read\_csv('C:\\SRM\\Data\\data.csv')

Notice the double \\ instead of \. There is a better practice to import data that we will use in later AYU assignments.

* If the dataset is on cloud, for example, the Automobile Insurance Claims dataset at the link: <https://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/CSVData/NAICExpense.csv>

library(tidyverse)  
d = read\_csv('https://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/CSVData/AutoClaims.csv')

You can use the function View(d) to view the data.

The dataset is about claims experience from a large midwestern (US) property and casualty insurer for private passenger automobile insurance. Let’s run regression of PAID on AGE.

model <- lm(PAID~AGE, data = d)  
summary(model)

##   
## Call:  
## lm(formula = PAID ~ AGE, data = d)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1851 -1329 -848 280 58142   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1786.738 195.031 9.161 <2e-16 \*\*\*  
## AGE 1.039 3.015 0.345 0.73   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2647 on 6771 degrees of freedom  
## Multiple R-squared: 1.754e-05, Adjusted R-squared: -0.0001301   
## F-statistic: 0.1188 on 1 and 6771 DF, p-value: 0.7304

From the above result, we can write the equation of the best fitted line: . This model has a very low R-squared (almost zero). Let’s use the model to predict the value of PAID when AGe are 20 and 40.

predict(model, list(AGE = c(20, 40)))

## 1 2   
## 1807.517 1828.296

As can be seen, when the predictions of PAID are 1807.517 and 1828.296 respectively.

Apply your understanding by doing Question 3.

## 4. AYU - Question

**Question 1.**

Calculate the mean, standard deviation of and the correlation of and . Given that

|  | 1 | 2 | -3 | -4 |
| --- | --- | --- | --- | --- |
|  | 1 | 5 | -2 | -6 |

**Question 2.**

Using the data in Question 1, find the slope, intercepts and the p-value of the F-test on the linear regression of on .

**Question 3.**

Import the Automobile UK Collision Claims dataset at <https://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/CSVData/AutoCollision.csv>

This dataset considered collision losses from private passenger United Kingdom (UK) automobile insurance policies. Run regression of Claim\_Count on Severity and

1. Write the equation of the best fitted line
2. What is the p-value of the test against
3. Use the model to predict claim count when the Severity are 200, 250 and 300.

## 5. Submission

* Address all the questions in a word document. Copy and paste the codes and the results of the code to the same word document. (Notice: You will learn that this is not the best way to present your statistical analysis. We will learn a more professional way to present it in the next Pod-AYU)