**Computer lab 1 in MMS075, Jan 23, 2020**

1. The R and RStudio software are installed on all computers in the computer lab. If you also want to have them on your home computer, you can download them for free.

R can be downloaded from here:

<https://cran.r-project.org/>

RStudio Desktop can be downloaded from here: <https://rstudio.com/products/rstudio/download/#download>

1. Start the R or RStudio software and type the code in Sections 2.3.1, 2.3.2 and 2.3.3 of the ISL book. These codes are also available online at <http://faculty.marshall.usc.edu/gareth-james/ISL/code.html>.
2. Consider again the context of Exercise 1 in the first exercise class: a (hypothetical) very large company called Maintain-IT is responsible for a project task that needs to be repeated every year. They want to determine how the number of employees assigned to the project affects the completion time. An analyst at Maintain-IT decides to use simple linear regression to model this dependence, based on the following observations:

|  |  |  |
| --- | --- | --- |
| Year | Employees assigned to the project | Completion time (days) |
| 1 | 70 | 20 |
| 2 | 30 | 60 |
| 3 | 10 | 100 |
| 4 | 90 | 20 |

Do the following steps using the above data:

1. Input the data in the table above in R;
2. Create a scatter plot using the “plot” command;
3. Compute the least squares coefficient estimates and corresponding 95% confidence intervals for the linear model:
	* Specify the model using the “lm” command; example: lm(time~employees), where “time” and “employees” should be replaced by the name of the variables you defined in part a).
	* Get a summary output using “summary”;
	* Get the confidence intervals using “confint”.
4. Find the quantity in your output that specifies the proportion of variability in completion time explained by the number of employees assigned to the project. Confirm that this indeed equals the squared correlation of the two variables.
5. Draw the least squares line in your scatter plot using “abline”. Hint: use …$coefficients[1] and …$coefficients[2] as arguments, where “…” should be replaced by the name of your model defined in part c.

Check whether the output you got agrees with the solution provided in “Exercise class 1 – solution to Exercise 1, parts c and e.pdf”.

**Code example:**

Employees=c(70,30,10,90)

CompTime=c(20,60,100,20)

plot(Employees,CompTime,xlab="Number of employees on project",ylab="Completion time",xlim=c(0,90),ylim=c(0,100),col="red",pch=20,cex=2)

LinMod=lm(CompTime~Employees)

summary(LinMod)

cor(Employees,CompTime)^2

abline(LinMod$coefficients[1],LinMod$coefficients[2],col="blue",lwd=2)

1. Download the advertising example in ISL from <http://faculty.marshall.usc.edu/gareth-james/ISL/data.html>. Import the data in R using the read.table command, e.g. AdData=read.table("[FileName]", header=TRUE, sep=","), where [FileName] should be replaced by the name of the downloaded file, including the path on the computer – e.g. in the format C:/Users/andrasb/Documents/Advertising.csv.

**Code example (after downloading Advertising.csv in my Documents folder):**

AdData=read.table("C:/Users/andrasb/Documents/Advertising.csv", header=TRUE, sep=",")

**Note**: in the computer lab, the document had to be saved on the Desktop (i.e. in a computer-specific folder), otherwise it did not work to import the data. Also, it worked better to use the menu in RStudio choosing File > Import Dataset > From Text (base)…



1. Create the three plots in Exercise 2 of the first exercise class, showing sales in 1000 units as a function of 1000 dollars invested. Hint: use the “plot(AdData$radio, AdData$sales, …)” command with further arguments xlab, ylab specifying axis names and col specifying the color of plotted points. Hint2: experiment with the pch and cex arguments to modify the type and size of plotted points; e.g. try pch=20, cex=2.

**Code example:**

plot(AdData$TV,AdData$sales, xlab="TV advertisement (1000$)",ylab="Sales",col="red",pch=20,cex=1)

plot(AdData$radio,AdData$sales, xlab="Radio advertisement (1000$)",ylab="Sales",col="red",pch=20,cex=1)

plot(AdData$newspaper,AdData$sales, xlab="Newspaper advertisement (1000$)",ylab="Sales",col="red",pch=20,cex=1)

1. Construct three separate linear models for each of the advertisement forms and display the regression lines on the plots above. Do also a summary of the regression models and compare the intercept and slope values. Do you see anything surprising?

**Code example:**

par(mfrow=c(1,3)) # This code line allows placing 3 plots side-by-side in 1 row.

TVModel=lm(AdData$sales~AdData$TV)

plot(AdData$TV,AdData$sales, xlab="TV advertisement (1000$)",ylab="Sales",col="red",pch=20,cex=1)

abline(TVModel$coefficients[1],TVModel$coefficients[2],col="blue")

RadioModel=lm(AdData$sales~AdData$radio)

plot(AdData$radio,AdData$sales, xlab="Radio advertisement (1000$)",ylab="Sales",col="red",pch=20,cex=1)

abline(RadioModel$coefficients[1],RadioModel$coefficients[2],col="blue")

NewspaperModel=lm(AdData$sales~AdData$newspaper)

plot(AdData$newspaper,AdData$sales, xlab="Newspaper advertisement (1000$)",ylab="Sales",col="red",pch=20,cex=1)

abline(NewspaperModel$coefficients[1],NewspaperModel$coefficients[2],col="blue")

summary(TVModel)

summary(RadioModel)

summary(NewspaperModel)

1. Display again the plots for the different advertisement forms, adding an argument “xlim=c(0,300)” in each case, and again display the regression lines. What was the effect of this change? Discuss whether and how your understanding of the results has changed compared to the previous point.

**Code example:**

par(mfrow=c(1,3)) # This code line allows placing 3 plots side-by-side in 1 row.

TVModel=lm(AdData$sales~AdData$TV)

plot(AdData$TV,AdData$sales, xlab="TV advertisement (1000$)",ylab="Sales",col="red",pch=20,cex=1,xlim=c(0,300))

abline(TVModel$coefficients[1],TVModel$coefficients[2],col="blue")

RadioModel=lm(AdData$sales~AdData$radio)

plot(AdData$radio,AdData$sales, xlab="Radio advertisement (1000$)",ylab="Sales",col="red",pch=20,cex=1,xlim=c(0,300))

abline(RadioModel$coefficients[1],RadioModel$coefficients[2],col="blue")

NewspaperModel=lm(AdData$sales~AdData$newspaper)

plot(AdData$newspaper,AdData$sales, xlab="Newspaper advertisement (1000$)",ylab="Sales",col="red",pch=20,cex=1,xlim=c(0,300))

abline(NewspaperModel$coefficients[1],NewspaperModel$coefficients[2],col="blue")

1. Feedback quiz (optional): Go to [www.menti.com](http://www.menti.com) and use the code 30 42 05.