Research in Applied Econometrics
Chapter 1. R

Pr. Philippe Polomé, Université Lumière Lyon 2

M1 APE Analyse des Politiques Économiques
M1 RISE Gouvernance des Risques Environnementaux

2018 – 2019

Outline

SWIRL

- Do Course 1 : R programming, Lessons 1-9 + 14 by yourself
  - To quit a lesson : esc
  - Answer "no" to any proposition to "register"
  - Following ...
    - press $\hookleftarrow$
    - Sometimes, much text is to be read – that is a good exercice
- Follow the commands in the RAE2017.R
  - They follow the slides
- We do just Lesson 1
  - To make sure you can start the other lessons by yourself

SWIRL R programming overview

1 : Basic Building Blocks	2 : Workspace and Files
3 : Sequences of Numbers	4 : Vectors
5 : Missing Values	6 : Subsetting Vectors
7 : Matrices and Data Frames	8 : Logic
9 : Functions	10 : lapply and sapply
11 : vapply and tapply	12 : Looking at Data
13 : Simulation	14 : Dates and Times
15 : Base Graphics	

Use RAE.r : A few commands outside of SWIRL

- In R-Studio, you should have created a new project (upper right button)
  - and called it "RAE" for example
  - Stored it where you can find it back
  - If not, do it now
- Execute the commands on RAE2017.R to see the output
- Usual math functions : log, exp, sign, sqrt, abs, min, max
  - log(exp(sin(pi/4)^2)*exp(cos(pi/4)^2)) Type in Console ↩
- Special vectors
  - ones <- rep(1, 10)
  - even <- seq(from = 2, to = 20, by =2)
  - trend <- 1981 :2005
- diag(4) Identity mtx of size 4

Mtx Operations

- A<-matrix(1 :6, nrow = 2)
  - A look what it looks like & how R gives the position of the elements
  - Look @ your environment window : A is now there
    - It remains in you project until erased (the brush)
- t(A) = transpose of A (*not A'*)
- dim(A) = dimensions of A (R then C)
- nrow(A) ; ncol(A) nbr R ; C
- A[i,j] extract element (i,j)
  - Does not remove it from the mtx
- A[,j] extract C j (all the R) into one vector
  - A[i,] same for R i
- A1<-A[1 :2, c(1, 3)] A1 has 2 R containing the elts in R 1 to 2 and C 1 & 3 from A
  - For this particular mtx, same result w/ A[,-2]

Mtx Operations

- det(A1) determinant
- solve(A1) inverse
- A %*% B mtx product
  - A*A element-by-element product
- crossprod(A, B) efficient calculation of A'B
- diag(A1) extract diag
- cbind(1, A1) "combine" one C of ones and A1

$$\begin{matrix} \cdot & & \cdot & \cdot \\ \cdot & \rightarrow & \cdot & \cdot \end{matrix}$$

- rbind(A1, diag(4, 2)) "stack" A1 & a diag mtx of size 2 w/ 4 on the diag

$$\begin{matrix} \cdot & \cdot \\ \cdot & \cdot \\ \uparrow & \\ \cdot & \cdot \end{matrix}$$

Outline

Dataframe

- "Frame" = "context"
  - In R, a "Dataframe" is a data mtx
    - a collection of vectors of same length
    - Stacked together horizontaly
- Each vector = 1 C = "variable"
  - Possibly of different natures
    - quantitative, numeric but qualitative, characters, dates...
  - it may further contain meta-data
    - e.g. variable type or categories name
- Each row = 1 obs in the sample

Dataframe Creation

- Several ways
  - keyboard (cfr Swirl programming lesson 7)
  - read R file
  - import
- keyboard example
  - alternative 1
    - mydata <- data.frame(one = 1 :10, two = 11 :20, three = 21 :30)
  - alternative 2
    - mydata <- as.data.frame(matrix(1 :30, ncol=3)) AND names(mydata) <- c("one", "two", "three")
- R is not very good for encoding data manually
  - But we use this example to explain attachment (below)

ATTACH

- A dataframe is "attached"
  - w/ command ATTACH
  - then variables' names in the dataframe maybe used directly in commands
- For example
  - mean(two) produce an error message
  - attach(mydata) and then mean(two) produces the average of variable "two"
- detach(mydata) is self-explanatory
  - Why detach ? e.g. to avoid confusions

Subset Selection

- As seen in SWIRL a subset of a Dataframe can be accessed by [or $
  - $ extract a single variable
- The command SUBSET sometimes work better (e.g. conditional selection)
  - e.g. mydata.sub<-subset(mydata, two<=16, select = -two)
  - selects all the obs. of variables one & three
    - fow which the obs of variable 2 are ≤ 16

Export (WRITE) a dataframe

- write.table(mydata, file="mydata.txt", col.names=TRUE)
  - create a txt file mydata.txt in the working directory
    - normally where your project is
  - Meta-data are not passed
  - The text file format is

"one"	"two"	"three"	
"1"	1	11	21
"2"	2	12	22
...			

  - So that it looks like the C headers are shifted left
    - Take that into account accordingly w/ the software you use to open it

Import (READ) a dataframe

- The Environment window has a button that makes it easy
  - a preview is generated
  - Use the "import dataset" button in the Environment window to read "mydata.txt" back into R
- Import from another software : excel, stata, sas...
  - Easiest : if you have access to the software, export the data file in txt or csv
    - loss of meta-data
  - R-Studio proposes several formats
    - It does not work often as these softwares change their formats often
  - Use Google
    - e.g. “R import Stata 17 data”
  - Also www.statmethods.net/input/importingdata.html
    - for a few formats

Outline

Plot

- First SWIRL
  - course R-programming, lesson 15 Base graphics
- A few additional graphic elements using package **plot**
  - Packages **lattice ggplot2** are better
    - varianceexplained.org/RData/code/code_lesson2/
  - R has many publication-quality graphics
    - But they are not very intuitive
- plot( ) is the default graphic command for many objects :
  - dataframes, time series, fitted linear models
    - it is also an old, crude, command

Examples with data("CPS1988")

- Data file is CPS1988 preloaded in the AER package
  - Pop. survey March 1988, US Census Bureau
  - 28 155 obs., cross-section
  - Men, 18-70 y-o
  - Income $>$ US$ 50 in 1988
  - Not self-employed, not working w/o salary
- summary(CPS1988)
- Quantitative data
  - wage $/week
  - education & experience (=age-education-6) in years

"Scatterplots" – dispersion – XY

- Probably the + common in stat, with histograms
  - We use CPS1988 : a census data file on wage and its determinants
  - From the AER package
- attach(CPS1988)
  - plot(education, log(wage))
    - First is on arg in x-axis, 2nd in y-axis
  - To export a plot : "Export" button in Plots window
    - There are several formats
    - png is easiest to use in word processing
- detach(CPS1988)
- plot(log(wage)~education, data=CPS1988)
  - alternative to avoid attaching the dataframe

R Graphic Parameters

- A plot results may be modified in many ways
  - E.g. argument type controls if the plot is made points (type = p), lines (type = l), both (type = b), steps (type = s) or others
- Several dozens parameters may be modified
  - See ?par
  - They may be modified *after* the plot w/ command par( )
  - Or they can be supplied in the plot( ) command e.g. plot(log(wage)~education, data=CPS1988, pch=20, col="blue", ylim=c(4,10), xlim=c(0,20), main="Wage by education years")

R Graphic Parameters

- Add layer(s) to a plot : lines( ), points( ), text( ), legend( )
  - Add a straight line abline(a, b)
    - a intercept, b slope
- 1 plot over another
  - x <- rnorm(50)
  - x2 <- rnorm(50, -1)
  - plot(ecdf(x), xlim = range(c(x, x2)))
    - ecdf empirical cumulative density function
  - plot(ecdf(x2), add = TRUE, lty = "dashed")
- Barplots, pie charts, boxplots, QQ plots & histograms
  - barplot( ), pie( ), boxplot( ), qqplot( ), hist( )
  - We'll see later

Histograms & boxplots

- Continue w/ CPS1988 data base on wage & its determinants
  - summary(CPS1988) reveals that some variables are categorical
  - Categorical : called *factors* in R
- **Factors** are vectors of categories
  - sometimes w/ **metadata**
    - e.g. categories names
  - g <- rep(0 :1, c(2,4))
  - g <- factor(g, levels=0 :1, labels=c("male", "female"))
    - Name categories (0,1) of g into "Male"(=0) & "Female"
    - so g is [1] male male female female female female

Factors in CPS1988

- In CPS1988, the factors are
  - ethnicity is caucasian “cauc” & african-american “afam”
  - smsa residence in urban area
  - region
  - parttime works part-time
- Plots according to data type
  - Numerical/Quantitative or categorical
  - Single variable or 2 in relation

One numerical variable : histogram & density

- hist(wage, freq=FALSE)
  - option freq=FALSE
    - relative frequencies, else absolute (counting)
  - option binwidth=zzz
    - "bin" = container : chose the length of the base of the rectangles
- hist(log(wage), freq=FALSE)
- lines(density(log(wage)), col=4)
  - Command density is actually a non-parametric estimate of the density function (next year)
- Remarks
  - log distribution is less asymetrical than the raw data
  - data in log are often closer to a normal
    - That is often the case w/ econ. data & a rationale for the normal hypothesis

One categorical

- W/ categorical data
  - Mean & variance have no meaning
  - But frequencies do
- summary(region) : absolute frequencies (counts)
- tab <- table(region) : stores these freq. in a table called tab
- prop.table(tab) computes the proportions (relative freq.)
- Barplots & pie visualise often quite well cat. data
  - barplot(tab)
  - pie(tab)
  - These plots can be modified using parameters

2 categorical

- Usually presented in a Contingency Table
  - xtabs( ) w/ a **formula** interface :
    - e.g. xtabs(~ ethnicity + region, data = CPS1988)
    - data is optional si it is still *attached*
  - table(ethnicity, region) mêmes résultats
- A plot of that is a "spine plot"
  - plot(ethnicity ~ region) Formula
  - plot(ethnicity, region) What differences ?

2 numerical

- The Correlation Coefficient r is typical
  - For positive & asymetrical variables : Spearman's ρ
    - *ranks* correlation, instead of values, is often prefered because r is not robust to asymetry
- cor(log(wage), education)
- cor(log(wage), education, method="spearman")
  - Results differ a bit
- plot(log(wage)~education)
  - scatterplot shows little correlation
  - but log makes it difficult to see graphically

1 numerical & 1 categorical

- Often, conditionnal moments are calculated
  - e.g. average wage by ethnicity
  - tapply(log(wage), ethnicity, mean)
    - "Applies" the cmd "mean" on the 2 variables ethnicity & log(wage)
    - Mean maybe replaced by any valid cmd, e.g quantile
- The Box plots & QQ (quantile-quantile) plots are often used

1 numerical & 1 categorical : Box plot

- A box plot is a crude representation of an empirical distribution
  - The box is limited by "hinges" (1º & 3º quartiles) and show the median
  - Outside of the box, 2 lines indicate the smallest & largest obs.
    - within 1.5 $\times$ size of the box from the closest hinge
  - Any obs. outside is represented by separate points
- boxplot(log(wage)~ethnicity)

1 numerical & 1 categorical : QQ plot

- A QQ plot **matches** the quantiles of 2 (empirical) distributions
  - Recall that quantiles are quantities
    - e.g. the 1º quartile of afam wage is the wage s.t. 25% of afam make less & 75% +
  - If the 2 distributions are identical : QQ plot = diagonal
  - Otherwise, if e.g. cauc make more than afam, then
    - with cauc on the x-axis, the QQ plot will be below the diag.
    - A bit like the plot of income inequality, but w/ 2 var.
  - awage <- subset(CPS1988, ethnicity == "afam")$wage
  - cwage <- subset(CPS1988, ethnicity == "cauc")$wage
  - qqplot(awage, cwage)
  - abline(0,1) overlay the diag (intercept 0, slope 1)
- detach(CPS1988) to close CPS1988

Outline

Basic Regression Commands in R

- Linear Regression Model LRM

$$y_i = x_i' \beta + \epsilon_i$$

w/ $i = 1...n$

  - In mtx form $y = X\beta + \epsilon$
- Typical Hyp. in cross-sections
  - $\mathtt{E}(\epsilon|X) = 0$ (exogeneity)
  - $\mathtt{Var}(\epsilon|X) = \sigma^2 I$ ("sphericity" : homoscedasticity & no autoc.)
- In R, models are usually fitted by calling a cmd
  - For the LRM in cross-section : fm <- lm(formula, data,...)
  - Argument ... replace a series of arguments
    - describing the model
    - or choosing the computation mode (algorithm)
    - or options

Basic Regression Commands in R

- The lm cmd returns an ***object***
  - Here : the fitted model under the name fm
    - Maybe visualised in many ways or summarized
- The lm object can be used to compute :
  - Predictions & fitted values, residuals, ... by means of fm$... see RAE2017
  - Tests & several postestimations diagnostics
- Most estimation commands work the same way

SWIRL

- Do Lessons 1-6, course « Regression Models » in Swirl
  - The others : later
  - Concentrate on code, you know the econometrics
  - Think of closing files that may have remained opened from the previous session

1. "Introduction"
   - To remember "A coefficient will be within 2 standard errors of its estimate about 95% of the time"
2. "Residuals" is + difficult (reading + programming +concepts)

   - Explains loops
   - Forces to re-read previous cmds
   - Make sure to execute program res_eqn.r when it shows up

SWIRL

3. "Least Squares Estimation" – nothing in particular
4. Introduction to Multivariable Regression
   - Install *manipulate* previously
     - I am not sure of the stability of this lesson
     - Do not edit the function myplot which will show up
     - ! cor(gpa_nor, gch_nor) will be $\neq \hat{\beta}$, SWIRL expects $=$, so a bug
5. "Residual Variation"
   - "Gaussian elimination" shows that a k-regressors regression
     - may be seen as a succession of k 1-regressor regressions
     - DO NOT interpret this as model building or presentation or a way to select results
6. "MultiVar Examples" – nothing in particular

Multivariare Linear Regression w/ Factors

- The purpose of this example is to demonstrate various R tools
  - that are used to transform & combine regressors
- Dataframe : CPS1988 as before
- SWIRL Course « Regression Models »
  - lesson 7 : "MultiVar Examples2"
    - Plots window for BoxPlot
    - sapply : use help in help window

Wage Equation

- Wage Equation

$$\log(wage) = \beta_1 + \beta_2 exp + \beta_3 exp^2 + \beta_4 education + \beta_5 ethnicity + \epsilon$$

```
cps_lm<-lm(log(wage)~experience+I(experience^2)
+education+ethnicity, data=CPS1988)
```

- "Insulation function" I( )
  - indicates to R that ^2 be understood as the square of exp
    - otherwise, R is unsure of the meaning and withdraws experience^2
  - This might be clearer w/ a formula y ~ a + (b+c)
    - Are there 2 variables on the RHS of the formula : a et (b+c), or are there 3 ?
    - To clarify, write y ~ a + I(b+c)

Results & Testing

- summary(cps_lm)
  - The return of education (to the wage) is 8.57%/year
    - % interpretation because wage is in log model
  - Categorical variables are managed by R
    - that selects the reference cat.
- Compare Nested Models : Anova (Analysis of Variance) Table
  - Regression + constraint
    - cps_noeth<-lm(log(wage)~experience+ I(experience^2)+education, data=CPS1988)
    - Usually, the test is on + than one variable
  - anova(cps_noeth,cps_lm)

Interactions : effects of combined regressors

- e.g. in labor econ : the combined effect of education & ethnicity
  - Does one year of Education have the same return for different ethnicities ?
- This is modeled w/ **multiplicative** terms
  - Consider

    $$\log(wage) = \beta_1 + \beta_2 ethnicity + \beta_3 ethnicity \times education + \beta_4 education + \epsilon$$

  - Then $\partial \log(wage) / \partial education = \beta_3 ethinicity + \beta_4$
    - If $ethinicity = 0$, then the effect of 1 year of education is β_4
    - If $ethinicity = 1$, then the effect of 1 year of education is $\beta_3 + \beta_4$
- Let a, b, c three factors
  - so that each has several discrete levels
- and x, y two continuous variables (quantitative)

Several Models/Formulas with Interactions

- y~a+x : no interaction
  - A single slope (of x) but one intercept for each level of factor a
- y~a*x : same as previous model +
  - one interaction term for each level of a with x (different slopes)
  - In a more formal notation, let $d_{ai} = I(a = i)$:

$$[y \sim a * x] \equiv \left[y = \beta_{ai} \sum_i d_{ai} + \gamma_{ai} x \sum_i d_{ai} \right]$$

Formulas with Interactions

- y~(a+b+c)^2
  - models all the interactions at 2 variables
    - but not at 3
  - So this is like as many dichotomous var. as the nbr of levels $d_{ai-bj} = I(a = i \wedge b = j)$ for a & b
    - and similarly for a & c and for c & b
- SWIRL course Regression Models
  - Lesson 8 : MultiVar Examples3

Interactions Wage eq. : ethnicity & education

- cps_int<-lm(log(wage)~experience+I(experience^2) +education*ethnicity, data=CPS1988)
  - Only one of the "+" from cps_lm has been replaced by *
- coeftest(cps_int)
  - A + compact version of summary( )
  - That can also be used on some other regression cmds
- The regression outputs the effects of education & ethnicity
  - called "main effects"
  - and the product of education & an indicator for the level "afam" of ethnicity
    - Why afam ? Probably because it is less numerous than cauc

Interactions Wage eq. : ethnicity & education

- afam has a neg. effect on the intercept
  - lower average wage for african-american
  - AND on the slope of education
    - lower return of education for african-american
- The effect is not much significant though
  - since a 5% significance with a sample of nearly 30 000 individuals is not much convincing

Predictions

- First define the values for which you want to predict.
  - We simplify the model to exp. & educ. for ease of presentation
  - Let's say we want to show the effect of Exp. at an average level of Educ.
- Create a new data frame w/ a C of average Educ & a C of all the possible values of Exp
  - Note that in the Census, some people have negative experience !
    - This is due to the way we compute Exp.
- Use a predict( ) cmd on
  - the lm object of interest : cps_lm here
  - the new data set for which we want prediction : cps2 here
    - predict( ) can not only gives a prediction but also bounds
    - Plot that on the data
- detach(CPS1988) when you are done to avoid confusion

Outline

When building a model, there are 2 contradictory forces

- If we omit a regressor, and it is in fact relevant
  - unobserved heterogeneity & inconsistency of LS estimators
    - if it is correlated to included regressors
  - we sometimes can deal w/ that using instruments or panel
- If we include irrelevant regressor that are correlated w/ relevant ones
  - we create multicollinearity w/ the csqce that both relevant & irrelevant regressors may appear non-signif.
  - That may even occur w/ 2 relevant regressors, e.g. in a Quantity-Price relation, the price of the subtitutes goods are relevant, but may be correlated w/ own price

Collinearity – Endogeneity Trade-off

- From a statistical point of view, 2 collinear variables carry the same information
  - Their separate influence on the dependant variable cannot be assessed in the present sample
    - Be pragmatic : reject one of the 2 or merge them in some way that makes sense in context
- It is not really possible to escape such a trade-off
  - Especially since in a particular sample, a relevant regressor may coincidentally appear non significant (if the sample is not large)
- Theory does not help by nature
  - since an empirical model is a trial of a model
  - theory helps interpreting results, not guide them

Progressive Inclusion

- is an old way of looking at model building

1. Among potential regressors x, take the one w/ highest correlation w/ y
2. Regress y on that single regressor
   - Is it significant ?
     - No : you don't have a model
     - Yes : estimate the one-regressor model & compute its residuals
3. Among the remaining regressors, take the one w/ highest correlation w/ the residuals
4. Repeat previous steps with progressively more regressors
   - Until one that is non-significant

Progressive Inclusion

- The issue w/ this approach is that if there are several relevant regressors
  - then at least the first step might be inconsistent
    - because at least one relevant regressor is missing
- This is a very serious issue that leads to non-sensical results

Progressive Elimination

- Instead, consider the "largest reasonnable set of regressors"
  - can be linked to the theory you want to test or to previous experience
- It is risky to just run this "encompassing" regression and report the results
  - because of multicollinearity

Progressive Elimination

- Gradually remove regressors one by one
  - Examine how the estimates of the remaining regressors evolve
  - If there is a noticeable increase in significance
    - but not so much change in estimates
    - collinearity was an issue
  - If estimated coefficients change wildly
    - omitted regressor endogeneity
- However
  - dropping collinear regressor could lead to jumps in coef estimates
    - after all, collinearity affects their variance
  - dropping a relevant regressor does not necessarily lead to major changes in the other coef
    - when that regressor is not much correlated to the others

Summing up

- Model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$ (no missing relevant regressor)
  - estimation by MCO when x_2 and x_1 are correlated
    - if they are not, there is NO serious consequences for $\hat{\beta}_1$
    - "not relevant but correlated to a relevant regressor" might not be empirically common

x_2		Consequences on $\hat{\beta}_1$	on $\hat{\beta}_2$
relevant	included	May appear insignificant	
	not incl.	Inconsistent	–
not relev.	included	May appear insignificant	should $\to 0$
	not incl.	??	–

Outline

Writing with R

- A few packages are designed to use R to write reports directly

1. The text is written directly in the script in the Editor window
   - Math formulas in latex may be included
   - Of course, R commands (graphics, regressions...)
2. If the data change, or the model, everything is adjusted automatically
3. LaTeX helps choose an appropriate format
   - report, paper, presentation

SWeave – Knitr – Markdown

- **SWeave** simply send the whole script to LaTeX
- **knitr** does the same but combine other packages and solve some issues in SWeave
- **Markdown** is the current standard
  - The script is directly printed using LaTeX or .doc (Word) or html (webpage)
  - Self-teach (I won't look into it)
    - http ://rmarkdown.rstudio.com/lesson-1.html
    - https ://www.r-bloggers.com/how-to-create-reports-with-r-markdown-in-rstudio/