Learning R

Causal Inference and Impact Evaluation

Diego A. Martin

February 21, 2019

What is R?

R is a programming environment:

- **R** itself was initially developed by Robert Gentleman and Ross Ihaka at the University of Auckland, New Zealand in 1996.
- They designed the language to combine the strengths of two existing languages, S and Scheme.
- Tools are distributed as packages, which any user can download to customize the **R** environment.
- Free system.

What is R?

- http://CRAN.R-project.org/
- RStudio is a better view (Similar to stata). Problematic with big database.
- For more information:
  - A Modern Approach to Regression with R.
  - An Introduction for R for Quantitative Economics.
  - R for STATA users.
  - Applied Econometric with R.

Introduction

- Installing Packages. $install.packages(\text{“}AER\text{”})$
- Vector and Matrix.
  $A < -matrix(c(a1, a2, a3, a4, a5), columns, rows)$
  - $byrow = TRUE$ indicates that the matrix should be filled by rows.
  - $byrow = FALSE$ indicates that the matrix should be filled by columns (the default).
  - dimnames provides optional labels for the columns and rows.
- $write.table$ and $read.table$

Data types

1. Vector.
2. Matrices.
3. Arrays: Similar to matrices but can have more than two dimensions.
   $Z < -array(data_{vector}, dim_{vector})$
4. Data Frames: A data frame is more general than a matrix, in that different columns can have different modes (numeric, character, factor, etc.). This is similar to SAS and SPSS datasets.

Exercise 1

- Generate a data frame with five type of individuals.
- Assign the ID, Age, Country, Education and a variable X1 with the salary.

5 List: An ordered collection of objects (components). A list allows you to gather a variety of (possibly unrelated) objects under one name.

6 Factor (**encode**): Tell R that a variable is nominal by making it a factor. The factor stores the nominal values as a vector of integers in the range [1... k] (where k is the number of unique values in the nominal variable), and an internal vector of character strings (the original values) mapped to these integers.
R will treat, in statistical procedures and graphical analyses:

- Factors as nominal variables
- Ordered factors as ordinal variables.
- Class of files class().

Some useful functions

- $length(object)$ number of elements or components
- $str(object)$ structure of an object
- $class(object)$ class or type of an object
- $names(object)$ names
- $c(object, object, ...)$ combine objects into a vector
- $cbind(object, object, ...)$ combine objects as columns
- $rbind(object, object, ...)$ combine objects as rows

Some useful functions

- $object$ prints the object
- $ls()$ list current objects
- $rm(object)$ delete an object
- $newobject < -edit(object)$ edit copy and save as newobject
- $fix(object)$ edit in place
- $merge()$

Where can we find help?

- Help ?command.
- Using Internet: UCLA, Berkeley, R etc.

Regressions

The demand for economics journals

We begin with a small data set taken from Stock and Watson (2007) that provides information on the number of library subscriptions to economic journals in the United States of America in the year 2000. The data set, originally collected by Bergstrom (2001), is available in package AER under the name Journals.

- Graph date.
- Writing OLS as a funcition $lm()$ $abline()$
- Store results in a list.

Quantile Regressions

Quantile Regressions: More complete view of the entire conditional distribution, not just the mean as OLS.

R 's fitting functions for regression models typically possess virtually identical syntax. In fact, in the case of quantile regression models, all we need to specify in addition to the already familiar formula and data arguments is tau, the set of quantiles that are to be modeled; For the next example set to 0.2, 0.35, 0.5, 0.65, 0.8.

Determinate of wages

Exercise 2

Considering a estimation of a wage equation in semi-logarithmic form based on data taken from Berndt (1991).

1. Loading the data set CPS1985 from the package AER.
2. Generate a data cps equal to CPS1985.
3. Estimate with OLS a Mincer equation (1958) with experience squared and education.
4. Graph Wage with experience, experience squared and education.
5. Install the package "quantreg".
6. Estimate with quantile regressions the Mincer equation (1956). **Hint**: use the help.

Relationship between wages and years of experience.

- Generate data cps2, where education is held constant and experience vary over a range.
- Add predicted values ($\hat{\beta}$), lower and upper bund.
- Add predicted values for percentiles.
- Visualize a result from wage to experience.

It can be seen that wages are highest for individuals with around 30 years of experience. The curvature of the regression lines is more marked at lower quartiles, whereas the relationship is much flatter for higher quantile.

Graph

Other ways to see results:

- Graph the results from quantile regressions $plot(summary(cps_rq))$.
- Heatmap: Bivariate kernel density.

Compared with the scatterplot in Figure 1, this brings out more clearly the empirical relationship between log(wage) and experience.

What about R-Studio

- It looks great.
- It can be seen the data.
- The graphs do not need a click.
- It can be selected the packages.

Basis

Exercise 3: Remembering Vectors

1. Generate the next vector: x= (1.8 3.14 4 88.169 13)
2. Multiple the vector x by 2 and add 3.
3. Generate a vector x1 equal to multiple the first component of x by 5, the second by 4 until the fifth by 1.
4. Generate a vector x2 equal to add the first component of x 1, the second by 2 until the fifth by 5.
5. Replace x equal to x=x1*x+x2.
6. How can do the last three steps in on line?

Some useful functions

Vectors

- $x[c(posituin1, position)]$ subsetting vectors
- $rep(a, repetitions)$
- $seq(from =, to =, by =)$
- $a : b$ patterned Vectors

Matrices

- $matrix(i : j, nrow = 2)$
- $t(A)$ transposed
- $dim(a)$ Dimensions
- $ncol(A)$ number of columns
- $A[a : b, c(takerow1, takerow2)]$ subtracting a square matrix

- $det(A1)$ Determinant
- $solve(A1)$ Inverse of A1
- $A1\% * \%solve(A1)\ \% * \%$ matrix multiplication
- $diag(4)$ Matrix diagonal
- $cbind(1, A1)$ combining matrices columns
- $rbind(A1, diag(4, 2))$ combining matrices diagonal

Logical Comparisons

- Logical proof with vectors and equations.
- $is.numeric(x)$

Random Number

- $rnorm()$
- $sample()$

Loops

$$if\ (cond)\ expr1\ else\ expr2$$

- $if(rnorm(1) > 0)\ sum(x)\ else\ mean(x)$
- $for(i\ in\ i:j)$
- $while()$ is a condition that may change in every run of the loop so that it finally can become FALSE.

Exercise 4:

1. Generating a function f with is equal to y as a function of x.
2. Replace x with a sequence from 0 to 10 with intervals of 0.5. (**Hint:** use function $seq()$ and see de help).
3. Replace y with the next equation: $3 * x + 2 +$ a random number between 1 and 2.
4. Graph y as function of x
5. Graph OLS line.
6. Creating a frame with three columns, first column numbers from 1 to 10, second column numbers from 11 to 20 and third column numbers from 21 to 30. Finally take the mean from column 2.

Export and import date

- $setwd()$ direction to export.
- $write.table()$ Export frame.
- $read.table()$ Import frame.

Import different kind of database

1. csv. $read.table("c:/mydata.csv", sep = ",")$
2. xlsx. $library(xlsx)$ and then $read.xlsx("c:/myexcel.xlsx", 1)$
3. dta. $library(foreign)$ and then $read.dta(import - stata.dta)$
4. rda. $save(mydata, file = "mydata.rda")$ it makes all objects stored in $mydata.rda$ directly available within the current environment. The advantage of using .rda files is that several arbitrary R objects, can be stored, including functions or fitted models,without loss of information.
5. Similarly functions to export to export.

Factors

Typical econometric examples of categorical variables include gender, union membership, or ethnicity. Encoding (e.g., 0 for males and 1 for females)

- $g <- factor(g, levels = 0:1, labels = c("male","female"))$

Missing

In R, such missing values are coded as NA (for "not available").

$newdata <- read.table("mydata.txt", na.strings = "-999")$

Object Orientation

Creating objects of a certain "class" and the creating function "generic functions" to this functions.

- Vector
- Factor
- Generic Function : $summary()$
- $methods(summary)$ will return a list of methods for all sorts of different classes.
- $summary.default()$

Defining a class

- Object "normsample" than contains a sample from a normal distribution.
- Creating function $normsample$.
- This function takes a required argument n (the sample size) and further arguments ..., which are passed on to $rnorm()$, the function for generating normal random numbers. In addition to the sample size, it takes further arguments—the mean and the standard deviation.
- After generation of the vector of normal random numbers, it is assigned the class "normsample" and then returned.

- Defining a $summary()$ method that reports the empirical mean and standard deviation for this sample.
- Function $sumary.norsample$.

Graphics

- Using again the data Journals from AER.
- Functions $plot()$.
- Function $rug()$ add ticks to visualize the marginal distribution of the variables.
- Function $par()$ show graphical parameters.
- Export Graph $pdf(\text{"}myfile.pdf\text{"}, height = 5, width = 6)$.
- Other kind of graphs $barplot()$, $pie()$, $boxplot()$, $qqplot()$ and $hist()$
- $demo(\text{“}demographics\text{''})$.

Pattern of the Graphs

- $plot(1:20, pch = 1:20, col = 1:20, cex = 2)$
- pch first 20 plotting symbol.
- $colors$ The first eight colors, then they are repeated.
- cex size of the point symbol.
- $dev.off()$ close the graph.

Mathematical annotation of plots.

- Density of the standard normal distribution.
- Include $expression()$
- $demo(\text{“}plotmath\text{”})$

Exploratory Data Analysis with R

Some Commands:

- $summary()$, $mean()$, $median()$, $var()$, $sd()$.
- $prop.table()$, $tapply()$ $corr()$.

Graphically

- $hist()$, $lines(density())$, $qqplot()$.

Linear Regressions

$$y = X\beta + \eta$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- $fm < -lm(formula, data...)$
- Parametric Test.
- No parametric test.

Exercise 5

1. Use data "Journals" from "AER".
2. Create a data "journals", with a data with two columns subscriptions ("subs") and Price "price", from "Journals".
3. Creating citeprice as price/citations and add to journals .
4. Graph the log of "subs" against the log of "citeprice".
5. Run the same regression as step 4 and add the fitted line to graph in that step.
6. And with Matrix

Some useful commands

- $jour_slm <- summary(jour_lm)$
- $jour_slm\$coefficients$
- $anova()$ analyse the variance.
- $coef()$ see the coefficients.
- $confint()$ see the confidence intervals.

Plotting "lm"

- $plot(jour_lm)$
  1. Residuals versus fitted values, perhaps the most familiar of all diagnostic plots for residual analysis.
  2. This curvature shows that the residuals more or less conform with normality.
  3. Scale-location versus leverages.
  4. Standardized residuals versus leverages.
- $par(mfrow = c(2, 2))$ Seeing four graph in one graph.

Testing a linear hypothesis

$$X\beta = r$$

Suppose we want to test, for the journals data, the hypothesis that the elasticity of the number of library subscriptions with respect to the price per citation equals to -0.5. Since this corresponds to the linear hypothesis $H_0 : \beta = -0.5$, we may proceed as follows:

- $linearHypothesis(jour_lm, "log(citeprice) = -0.5")$

Suggesting that the elasticity of interest is not substantially different from -0.5.

Exercise 6

1. Use data "CPS1988"
2. Run

$$log(wage) = experience + experience^2 + education + ethnicity$$

3. Summarize information.

Dummy variable

- factor are handled like this when they are included in regression models.
- $I()$ returns its argument "as is". This was used for computing experience squared in the regression above

Comparison of Models

With more than a single explanatory variable, it is interesting to test for the relevance of subsets of regressors. For any two nested models.

For example, it might be desired to test for the relevance of the variable ethnicity; i.e., whether there is a difference in the average log-wage (controlling for experience and education) between Caucasian and African-American men.

- $anova()$

- $anova(cps_noeth, cps_lm)$ Test that the RSS of both models are equal.
- $anova(cps_lm)$ the same test as before. When there is just one model, it adds one variable at time and does the same test.
- $update$ instead to write again the equation without ethnicity
- $waldtest()$ this produces the same F test as anova() but does not report the associated RSS.

Partial Linear Model

$$log(wage) = \beta_1 + g(experience) + \beta_2 education + \beta_3 ethnicity$$

Where g is an unknown function to be estimated from the data, and we use regression splines for this task

- $lm(\ bs(experience, df = 5))$ where df is the degree of freedom.

Factors, Interactions, and Weights

- $y\ a + x$ Model without interaction. identical slopes with respect to x but different intercepts with respect to a.
- $y\ a * x$ Model with interaction. This interaction included ethnicity, education and the interaction between two.
- $y\ a + x + a : x$ the term a:x gives the difference in slopes compared with the reference category, in other words just the interaction.

Separate Regressions for each level

As a further variation, it may be necessary to fit separate regressions for African-Americans and Caucasians.

- This model specifies that the terms within parentheses are nested within ethnicity.
- the term -1 removes the intercept of the nested model.
- A matrix to see results for both ethnicity
- $anova(model1, model2)$ the model where ethnicity interacts with every other regressor fits significantly better, at any reasonable level, than the model without any interaction term.

Change of the reference category

In CPS1988, "cauc" is the reference category for ethnicity, while "northeast" is the reference category for region.Bierens and Ginther (2001) employ "south" as the reference category for region.

- $relevel()$

Weighted least squares

Cross-section regressions are often plagued by heteroskedasticity. Weighted least squares (WLS) is one of the remedies.

- The problem is the variance, $V(\hat{\beta}) = S^2(X'X)^{-1}$.
- Going back to Journals data.

Feasible generalized least square (FGLS)

More often than not, we are not sure as to which form of the skedastic function to use and would prefer to estimate it from the data.

1. Estimating the residuals of the initial models with OLS and saving the results in an auxiliary regression.
2. Use the fited values of the auxiliary regression as the weights.

Hint: It may use a loop.

Exercise 7

1. Graph the logarithm of the subscription as a function of the logarithm of citeprice.
2. Add line from OLS estimation and FGLS estimation.
3. Add the respective legends.

Linear Regression with time series data

- In econometrics, time series regressions are often fitted by OLS.
- Hence, in principle, they can be fitted like any other linear regression model using $lm()$ if the data set is held in a $data.frame$.
- However, this is typically not the case for time series data, which are more conveniently stored in one of R's time series classes.
- An example is $ts()$, which holds its data in a vector or matrix plus some time series attributes (start, end, frequency).
- $dynlm()$ to run $y_t - y_{t-1} = \beta_1 + \beta_2(y_{t-1} - y_{t-2}) + \beta_3 x_{t-4}$

Exercise 8

1. Use data "USMacroG" and see the variables in this data.
2. Graph dpi as a function of consumption.
3. Try to use plot.type (**Hint** Use the help)
4. Install Packages $dynlm$.
5. Using $dynlm()$ run consumption as a function of dpi and a lagged of dpi ($L(dp1)$).
6. Run consumption as a function of dpi an a lagged of consumption.
7. Analyses the results of point 6 and 5.

An interesting way two see both results:

1. $merge()$ the original series with the fitted values from both models.
2. Plot a zero line.
3. Plot a residuals from both models.

Encompassing Test

Cox Test: The encompassing approach to comparing two nonnested models is to transform the problem into a situation we can already deal with comparing nested models.

Compare a two model where each one has one lagged variable with a model with two lags.

- $encomptest(cons_lm1, cons_lm2)$

In this case, both models perform significantly worse compared with the encompassing model, although the F statistic is much smaller for $cons_lm2$.

Linear Regression with Panel Data

- For methodological background, we refer to Baltagi (2005).
- Basic fixed- and random-effects methods. Using the well-known Grundfeld (1958).
- plm

Let use a subset of three firms for illustration and, utilizing $plm.data()$, tell R that the individuals are called "firm", whereas the time identifier is called "year".

$$y_{it} = X_{it}\beta + \eta_{it}$$

Whether there is omitted variable, then $cov(X_{it}, \eta_{it})! = 0$. The restriction of independence is not bidding, therefore OLS is biased.

- Fixed Effects

  $$y_{it} = \alpha_i + \beta X_{it} + u_{it}$$

  Where $\alpha_i = \alpha + v_i$. This means that can be decomposed in two, a fixed part, constant for each individual (v_i) and a random part, which has all the assumptions of OLS.
- Random Effects. It has the same specification as before but in this case v_i is a random variable with mean v_i and variance $Var(v_i! = 0)$

Exercise 9

1. Use data Grunfeld from package "AER".
2. Generate gr which is a subset of column firm with $(\%in\%)$ "Genral Electric", "General Motors", "IBM".
3. Generate pgr, using $plm.data()$ with arguments gr and an index composed of firm an year. (**Hint:** $plm.data()$).

$$invest_{it} = \beta_1 value_{it} + \beta_2 capital_{it} + \alpha_i + v_{it}$$

α_i denote the individual-specific effects

Some useful commands

- $plm(, model = "pooled")$
- $plm(, model = "within")$
- $pFtest()$ to see if fixed effects are necessary.
- $plm(, model = "random")$
- $plmtest(gr_pool)$ to see if random effects are necessary.

In both test the Ho: Effect are necessary

Random-effects methods are more efficient than the fixed-effects estimator under more restrictive assumptions, namely exogeneity of the individual effects. It is therefore important to test for endogeneity, and the standard approach employs a Hausman test.

- $phtest(gr_re, gr_fe)$

In line with the rather similar estimates presented above, endogeneity does not appear to be a problem here.

Dynamic linear models

To see in detail Arellano and Bond (1991).

- Autoregressive Models: Lags in endogenous variable of individual i.
- Distributed Lag Models: Lags in exogenous variable of individual i.

1. Set up a static formula containing the relevant variables average annual wage per employee (wage), the book value of gross fixed assets (capital), and an index of value-added output at constant factor cost (output).
2. The function providing the Arellano-Bond estimator is $pgmm()$, and it takes as its first argument a so-called dynformula, this being a static model equation, as given above, augmented with a list of integers containing the number of lags for each variable.

Systems of Linear Equations

Standard examples include seemingly unrelated regressions and various macroeconomic simultaneous equation models.

The package $systemfit$ can estimate a number of multiple-equation models.

Let see a seemingly unrelated regression (SUR) model.

Some advantage:

- Unlike the panel data models considered in the preceding section, which permit only individual-specific intercepts, the SUR model also allows for individual-specific slopes.
- The model assumes contemporaneous correlation across equations, and thus joint estimation of all parameters is, in general, more efficient than OLS on each equation.
- $systemfit()$

The output indicates again that there is substantial variation among the firms, and thus a single-equation model for the pooled data is not appropriate.

Please cite this presentation as: Martin, D. A. (2016). Learning R. Causal Inference and Impact Evaluation. Presentation. Universidad del Rosario.