Introduction to Handling Data

ECON20222 - Lecture 1

Ralf Becker and Martyn Andrews

January 2012

What is this course unit about?

- Help you implement and interpret the main estimation and inference techniques used in Economics
- Focus on:
  - causal inference
  - the main pitfalls of time-series analysis

At the end of this unit ...

You will be able to:

- Do intermediate data work in R
- Confidently apply regression analysis in R
- Apply more advanced causal inference techniques in R
- Find coding help for any new challenges in R
- Discuss strengths and weaknesses of particular empirical applications
- Perform inference appropriate for the model being estimated
- Interpret empirical results (with due caution!)

What you need to do

To learn in this unit you need to:

coding, cleaning data, struggling, self-learning, amazement at what you can do

answering real questions, that there is not always a clear answer

Assessment Structure and feedback

- Online test (on the use of R) - 10%
- End-of-Term exam (MC and short answer questions) - 50%
- Group coursework - 40% (see extra info)

Aim for today

Statistics/Econometrics

- Summary Statistics
- Difference between population and sample
- Hypothesis testing
- Graphical Data Representations
- Diff-in-Diff Analysis
- Simple regression analysis

R Coding

- Introduce you to R and RStudio
- How do I learn R
- Import data into R
- Perform some basic data manipulation
- Perform hypothesis tests
- Estimate a regression

Today's Empirical Question

Card, David ; Krueger, Alan B. (1994) Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania, The American Economic Review, 84, 772-793.

Do higher minimum wages decrease employment (as predicted by a simplistic labour market model)?

The Research Question

"This paper presents new evidence on the effect of minimum wages on establishmentlevel employment outcomes. We analyze the experiences of 410 fast-food restaurants in New Jersey and Pennsylvania following the increase in New Jersey's minimum wage from $ 4.25 to $ 5.05 per hour. Comparisons of employment, wages, and prices at stores in New Jersey and Pennsylvania before and after the rise offer a simple method for evaluating the effects of the minimum wage."

Card, David ; Krueger, Alan B. (1994, p.772)

Why Data Matter

The debate is still alive:

- Overall negative effect on employment, IZA.
  "Research findings are not unanimous, but especially for the US, evidence suggests that minimum wages reduce the jobs available to low-skill workers."
- An overview of the mpirical evidence is provided in this report by Arindrajit Dube for the UK Government.
  "Especially for the set of studies that consider broad groups of workers, the overall evidence base suggests an employment impact of close to zero."

The Plan for today

- Replicate the above summary statistics in R
- Why one can get two very different results based on the same data
- Use this example to
  - introduce you to R
  - review some summary statistics
  - review simple regression and its implementation
  - introduce some basic visualisations

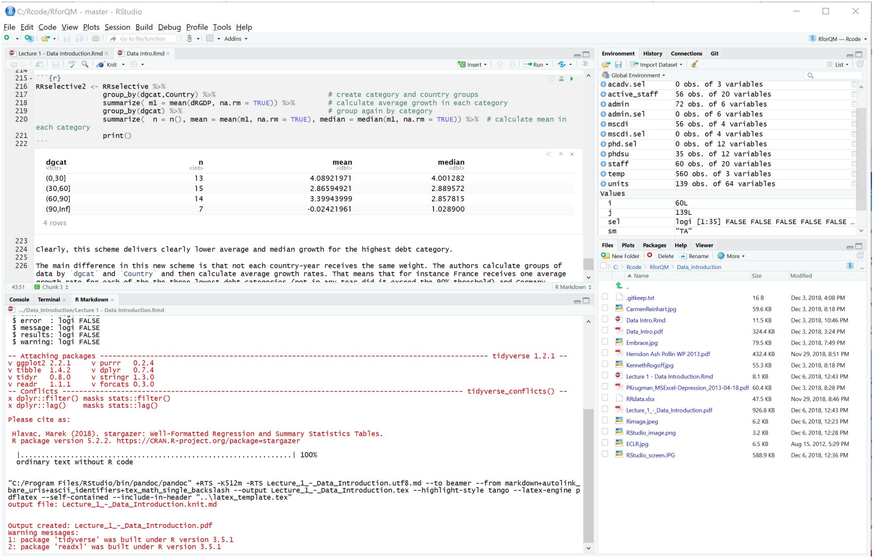
Introduce R/R-Studio

- R is a statistical software package, it is open source and free
- a lot of useful functionality is added by independent researchers via packages (also for free)

- RStudio is a user interface which makes working with R easier. You need to install R before you install RStudio.

- ECLR is a web-resource we have set up to support you in your R work.

Welcome to RStudio



Write Code Files or the Basic Workflow

- keep an original data file (usually '.xlsx' or '.csv') and do not overwrite this file
- any manipulation we make to the data (data cleaning, statistical analysis etc.) is command based and we collect all these commands in a script file. R will then interpret and execute these commands. It is hence like a recepie which you present to a chef. These script files have extension '.r'
- you can also learn to write Rmarkdown files ('.rmd'). They combine code with normal text and output.
- When you write code you should ensure that you add comments to your code. Comments are bit of text which is ignored by R (everything after an '#') but helps you or someone else to decipher what the code does.

By following the above examples you make it easy for yourself and others to replicate your work. At the core of the Reinhard/Rogoff controversy was the ability of independent researchers to replicate their work!

Prepare your code

We start by uploading the extra packages we need in our code.

The first time you need these packages at a computer you may need to install these. Use the following code to do this

```
install.packages(c("readxl","tidyverse","ggplot2","stargazer")
```

This only needs to be done once on a particular computer. However, every time you want to use any of these packages in a code you need to make them available to your code (load them):

```
library(tidyverse)     # for almost all data handling tasks
library(readxl)        # to import Excel data
library(ggplot2)       # to produce nice graphiscs
library(stargazer)     # to produce nice results tables
```

The data

Then we load the data from excel

```
RRData <- read_excel("RRdata.xlsx")
RRData <- as.data.frame(RRData) # forces data.frame structure
```

and check some characteristics of the data which are now stored in `RRData`:

```
str(RRData)  # prints some basic info on variables
```

```
## 'data.frame':    1171 obs. of  4 variables:
##  $ Year   : num  1946 1947 1948 1949 1950 ...
##  $ Country: chr  "Australia" "Australia" "Australia" "Austr
##  $ debtgdp: num  190 177 149 126 110 ...
##  $ dRGDP  : num  -3.56 2.46 6.44 6.61 6.92 ...
```

Discuss data.frame, number of obs and number of variables, their names and variable types

Variable types

These are four basic data types.

- `character: "a", "swc"`
- `numeric: 2, 15.5`
- `integer: 2L (the L tells R to store this as an integer)`
- `logical: TRUE, FALSE`

It is important that you know and understand differences between data types. Each variable has has a particular type and some operations only work for particular datatypes. For instance, we need `num` or `int` for any mathematical operations.

In our data.frame three variables are `num` and one is of `chr` type.

We will encounter `logical` variables frequently. they are very powerful

factor variables

It is necessary to change categorical variables (here `Country`) to `factor` variables.

```
RRData$Country <- as.factor(RRData$Country)
str(RRData)
```

```
## 'data.frame':    1171 obs. of  4 variables:
##  $ Year   : num  1946 1947 1948 1949 1950 ...
##  $ Country: Factor w/ 20 levels "Australia","Austria",..: 1
##  $ debtgdp: num  190 177 149 126 110 ...
##  $ dRGDP  : num  -3.56 2.46 6.44 6.61 6.92 ...
```

- `RRData$Country` calls variable `Country` in dataframe `RRData`
- `<-` assigns what is on the right `as.factor(RRData$Country)` to the variable on the left `RRData$Country`
- `as.factor(RRData$Country)` calls a function `as.factor` and applies it to `RRData$Country`

factor variables

factor variables are variables with discrete categories. Which ones they are you can find out with the levels() function:

```
levels(RRData$Country)
```

```
##  [1] "Australia"   "Austria"     "Belgium"     "Canada"
##  [6] "Finland"     "France"      "Germany"     "Greece"
## [11] "Italy"       "Japan"       "Netherlands" "New Zealand
## [16] "Portugal"    "Spain"       "Sweden"      "UK"
```

Learn more about your data

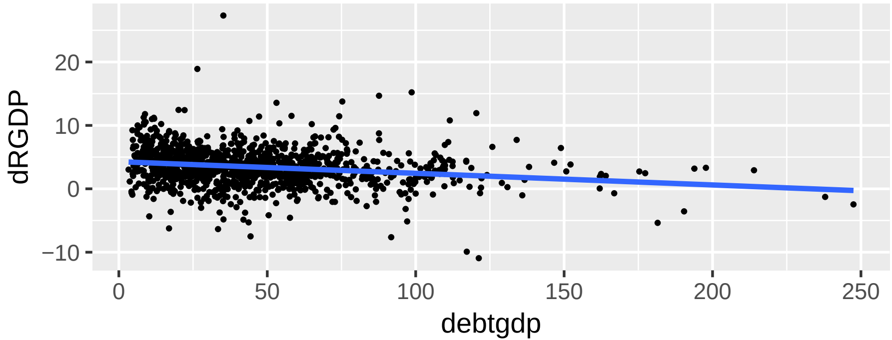
Use the `stargazer` function for some initial summary stats for `num` or `int` variables

```r
stargazer(RRData, type = "text")
```

```
##
## ====================================================================
## Statistic   N     Mean    St. Dev.   Min   Pctl(25) Pctl(75)   Max
## --------------------------------------------------------------------
## Year      1,171 1,979.155  17.944   1,946   1,964    1,995    2,009
## debtgdp   1,171  46.283    32.395   3.279   22.148   61.474  247.482
## dRGDP     1,171   3.430    2.981   -10.942  1.911    5.100   27.329
## --------------------------------------------------------------------
```

Scatter plot of the data

```
p1 <- ggplot(RRData,aes(debtgdp,dRGDP)) +
  geom_point(size=0.5) +     # this produces the scatter plot
  geom_smooth(method = "lm", se = FALSE)  # adds the line
p1
```



Point out that each dot represents one country/year's data, e.g. France in 1991. Point out line of best fit

Regression Line

The line in the previous plot is the line of best fit coming for a linear regression

$$rGDP = \alpha + \beta debtgdp + u \text{ (Population Model)}$$

- The population model is defined by unknown parameters α and β and the unknown error terms u. We will use sample data to obtain sample estimates of these parameters.
- The error terms u contain the effects of any omitted variables and reflect that any modelled relationship will only be an approximation. The u are random variables

$$rGDP_{it} = \widehat{\alpha} + \widehat{\beta}\ debtgdp_{it} + \widehat{u}_{it} \text{ (Estimated Sample Model)}$$

Here we have two subscripts as the data have a cross-section (`i`) and a time-series dimension (`t`).

The regression line in the previous figure is represented by

$$\widehat{rGDP}_{it} = \widehat{\alpha} + \widehat{\beta} debtgdp_{it} \quad \text{(Regression Line)}$$

Simple Regression Model and OLS

Regression analysis is the core technique used in Econometrics. It is based on certain assumptions about the *Population Model* and the error terms u (more on this in the next few weeks).

How to estimate parameters (get $\widehat{\alpha}$ and $\widehat{\beta}$) using the available sample of data? This is typically done by Ordinary Least Squares (OLS).

Simple Regression Model and OLS

```
mod1 <- lm(dRGDP~debtgdp, data= RRData)
summary(mod1)
```

```
##
## Call:
## lm(formula = dRGDP ~ debtgdp, data = RRData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9958  -1.5200  -0.0774   1.5707  23.6960
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.279290   0.148970   28.73  < 2e-16 ***
## debtgdp     -0.018355   0.002637   -6.96 5.67e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.922 on 1169 degrees of freedom
## Multiple R-squared:  0.03979,    Adjusted R-squared:  0.03897
## F-statistic: 48.44 on 1 and 1169 DF,  p-value: 5.666e-12
```

OLS - nice output

```
stargazer(mod1,type="text")
```

```
##
## ================================================
##                          Dependent variable:
##                     ---------------------------
##                                 dRGDP
## ------------------------------------------------
## debtgdp                       -0.018***
##                                (0.003)
##
## Constant                       4.279***
##                                (0.149)
##
## ------------------------------------------------
## Observations                    1,171
## R2                              0.040
## Adjusted R2                     0.039
## Residual Std. Error       2.922 (df = 1169)
## F Statistic           48.439*** (df = 1; 1169)
## ================================================
```

OLS - calculation and interpretation

How were $\widehat{\beta}$ and $\widehat{\alpha}$ calculated?

$$\begin{aligned}
\widehat{\beta} &= \frac{\widehat{Cov}(dRGDP_{it}, debtgdp_{it})}{\widehat{Var}(debtgdp_{it})} \\
\widehat{\alpha} &= \overline{dRGDP}_{it} - \widehat{\beta} * \overline{debtgdp}_{it}
\end{aligned}$$

How to interpret $\widehat{\beta} = -0.018$?

An increase of one unit in `debtgdp` (=1% point) is related to a decrease of GDP growth of 0.018 units (=0.018% points - pp)

Have we established that higher debt levels **cause** lower GDP growth?

NO

Regression Analysis - Underneath the hood

Need to recognise that in a sample $\hat{\beta}$ and $\hat{\alpha}$ are really random variables.

$$\begin{aligned}
\hat{\beta} &= \frac{\widehat{Cov}(dRGDP, debtgdp)}{\widehat{Var}(debtgdp)} \\
&= \frac{\widehat{Cov}(\alpha + \beta\ debtgdp + u, debtgdp)}{\widehat{Var}(debtgdp)} \\
&= \frac{\widehat{Cov}(\alpha, debtgdp) + \beta\widehat{Cov}(debtgdp, debtgdp) + \widehat{Cov}(u, debtgdp)}{\widehat{Var}(debtgdp)} \\
&= \beta\ \frac{\widehat{Var}(debtgdp)}{\widehat{Var}(debtgdp)} + \frac{\widehat{Cov}(u, debtgdp)}{\widehat{Var}(debtgdp)} = \beta + \frac{\widehat{Cov}(u, debtgdp)}{\widehat{Var}(debtgdp)}
\end{aligned}$$

So $\hat{\beta}$ is a function of the random term u and hence is itself a random variable. Once $\widehat{Cov}(dRGDP, debtgdp)$ and $\widehat{Var}(debtgdp)$ are replaced by sample estimates we get ONE value which is draw from a random distribution.

OLS - estimator properties

What can we learn from this?

- If u_{it} is a random variable, so is $\widehat{\beta}$
- Any particular value we get is a draw from a random distribution
- An estimator is unbiased if, on average, the estimates would be equal to the unknown β
  at this stage the concept of unbiasedness may still be a little hazy and that is fine
- For this to happen we need to assume that $Cov(u, x) = 0$ as then $E(\widehat{\beta}) = \beta$
  Why do we need to assume this? Because while we do have values for x_{it} we do not have values for the unobserved error terms u_{it}. Hence we cannot test this. As you will find out this is mainly a thinking exercise and one at the core of much of what we do.

OLS - the exogeneity assumption

For $\widehat{\beta}$ in $y_{it} = \alpha + \beta x_{it} + u_{it}$ to be unbiased (i.e. on average correct) we needed

$$Cov(u_{it}, x_{it}) = 0$$

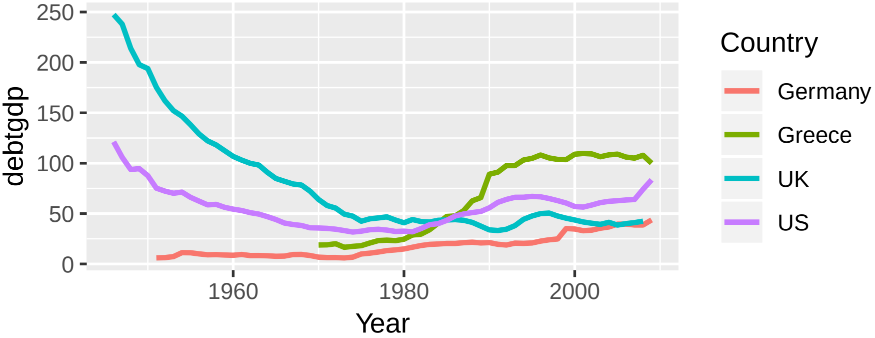
This is sometimes called the Exogeneity assumption. The error term has to be uncorrelated to the explanatory variable x_{it}

There are a lot of reasons why this assumption may be breached.

- Simultaneity ($debtgdp \rightarrow dRGDP$ and $dRGDP \rightarrow debtgdp$)
  Discuss the fact that we have to assume that causailty here goes in both directions. Hence we cannot attach one one-directional causal interpretation to the estimated coefficient. If you can estimate the model the other way round
- Measurement error in x_{it}
- Omitted relevant variables or unobserved heterogeneity

Reinhard/Rogoff Example - debtgdp

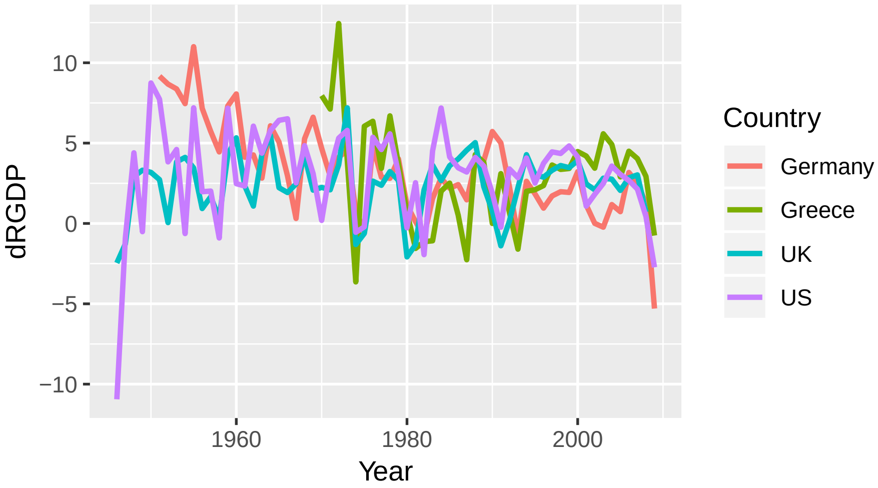
```
tempdata <- RRData %>% filter(Country %in% c("Germany","Greece","UK","US"))
ggplot(tempdata,aes(Year,debtgdp,color=Country)) +
  geom_line(size=1)      # this produces the line plot
```



Point out the piping operator and the filter function and then ggplot and geom_line

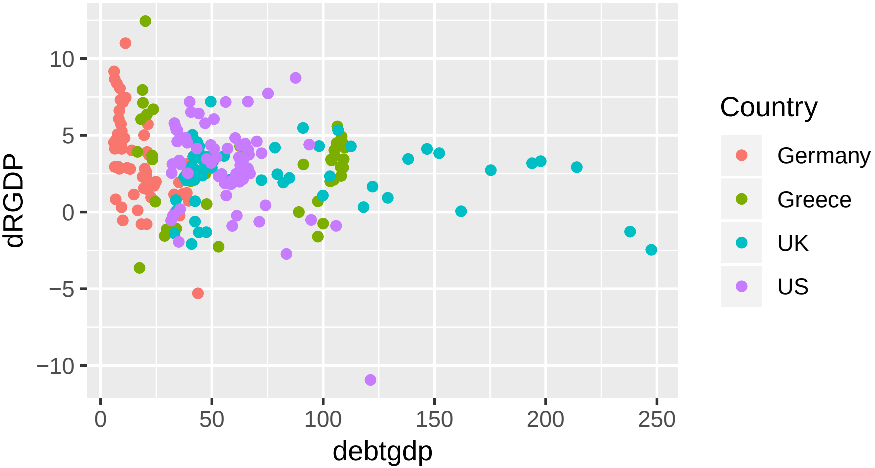
Reinhard/Rogoff Example - dRGDP

```
tempdata <- RRData %>% filter(Country %in% c("Germany","Greece","UK","US"))
ggplot(tempdata,aes(Year,dRGDP,color=Country)) +
  geom_line(size=1)      # this produces the line plot
```



Reinhard/Rogoff Example - dRGDP v debtgdp

```
tempdata <- RRData %>% filter(Country %in% c("Germany","Greece","UK","US"))
ggplot(tempdata,aes(debtgdp,dRGDP,color=Country)) +
  geom_point()      # this produces the scatter plot
```



Reinhard/Rogoff Example

What have we learned?

- dept data are persistant
- growth data as well but less so
- data for different countries have different characteristics

Group the data into debt categories

Let's create the four debt categories: (0,30], (30,60], (60,90], (90,Inf].

```
RRData <- RRData %>% mutate(dgcat = cut(RRData$debtgdp, breaks=c(0,30,60,90,Inf)))

RRData %>% group_by(dgcat) %>%
            summarise_at("dRGDP", funs(mean, median)) %>%
            print()
```

```
## # A tibble: 4 x 3
##   dgcat      mean median
##   <fct>     <dbl>  <dbl>
## 1 (0,30]     4.17   4.15
## 2 (30,60]    3.12   3.11
## 3 (60,90]    3.22   2.9
## 4 (90,Inf]   2.17   2.34
```

Pint out the mutate function in combination with the cut function and the pipe, then the group_by and summarise_at function

These are the statistics we saw before (not the one reported by Reinhard and Rogoff)

Why did Reinhard and Rogoff report different results?

Debt Category	(0,30]	(30,60]	(60,90]	(90,Inf]
Avg Growth Rate (RR)	4.09%	2.87%	3.40%	-0.02%
Avg Growth Rate (HAP)	4.17%	3.12%	3.22%	2.17%

Why did RR get so much lower growth for the highest debt category, (90,Inf]?

Thomas Herndon, Michael Ash and Robert Pollin (2014) replicated the work. They identified the following differences to the above analysis.

1. They excluded early-postwar data for New Zealand, Australia and Canada, arguing that these data are atypical for later periods, essentially they are outliers
2. A spreadsheet error resulted in data for the five countries (Australia, Austria, Belgium, Canada and Denmark) to not be included.
3. Observations are not weighted equally.

Replicate Reinhard and Rogoff's results

```
## Selective treatment of early years
RRselective <- RRData %>%
  filter(!((Year<1950 & Country=="New Zealand") |
             (Year<1951 & Country=="Australia") |
             (Year<1951 & Country=="Canada") ))

## Spreadsheet error omitting five countries
RRselective <- RRselective %>%
  filter(!( Country %in%
              c("Australia","Austria","Belgium","Canada","Denmark") ))

RRselective %>% group_by(dgcat) %>%
            summarise_at("dRGDP", funs(mean, median)) %>%
            print()
```

```
## # A tibble: 4 x 3
##   dgcat     mean median
##   <fct>    <dbl>  <dbl>
## 1 (0,30]    4.24   4.4
## 2 (30,60]   2.98   3.06
## 3 (60,90]   3.16   2.85
## 4 (90,Inf]  1.69   2.33
```

Replicate Reinhard and Rogoff's results

So the first two differences explain some but not all of the differences. Let's implement the different weighting.

```
RRselective2 <- RRselective %>%
    group_by(dgcat,Country) %>%       # create category and country groups
    summarize( m1 = mean(dRGDP, na.rm = TRUE)) %>%  # calculate cat average
    group_by(dgcat) %>%                # group again by category
    summarize(  n = n(), mean = mean(m1, na.rm = TRUE),
                median = median(m1, na.rm = TRUE)) %>%
    print()
```

```
## # A tibble: 4 x 4
##   dgcat        n    mean median
##   <fct>    <int>   <dbl>  <dbl>
## 1 (0,30]      13  4.09     4.00
## 2 (30,60]     15  2.87     2.89
## 3 (60,90]     14  3.40     2.86
## 4 (90,Inf]     7 -0.0242   1.03
```

Don't worry too much about the full details of this weighting scheme.

The combination of the three changes make a massive difference

Why data (and their treatment) matter

The combination of these changes made a significant difference to the summary statistics.

Remember, the data were very persistent.

Reinhard and Rogoff
For each country all years which fall into one of the four categories are averaged and then treated as one observation

Herndon, Ash and Pollin
Each country/year observationis treated as one independent observation.

Perhaps it is right to not treat each observation as a new piece of information. But the RR weighting scheme seems to discard a lot of information.

These results made a significant difference in the political discourse.

Outlook

Over the next weeks you will learn

- to perform more advanced statistical analysis in R, such as:
  - Hypothesis testing
  - Multivariate regression analysis
  - specification testing
- to devise methods to draw causal inference
- to understand the main pitfalls of time-series modelling and forecasting