Statistics for economics
R problem sets

University College Dublin

Spring 2017

For the R tutorials you will be working with socio-economic data mainly compiled by Eurostat at the NUTS2 region. The variables included in the data are briefly summarised in table 1.

Table 1: Description variables

Variable name	Description
GEO	NUTS2 indicator
ccode	Country code
mortality	General mortality rate out of 1000
infant.mortality	Infant mortality rate per 1000 live births
female.mortality	Female mortality rate out of 1000
middle.aged.mortality	Middle-aged (45-55y) mortality rate per 1000
mortality.hi	Binary indicator for regions with above average mortality rates
eur.pc	Per capita income in Euro
pps.pc	Per capita income in Euro adjusted for purchasing power
inh.per.doc	Number of inhabitants per doctor
income	Income group (1-low; 4-high)
structural.fund	Binary indicator for EU regions that qualify for EU funding
eastbloc	Binary indicator for countries that were part of the Warsaw Pact
candidate	Binary indicator for candidate countries for EU membership
efta	Binary indicator European Free Trade Association member countries
euro2102	Round reached at Euro 2012 football championship (0- not qualified; 4-final)

Note that the binary indicator for former Eastern Bloc countries excludes East Germany and for the candidate countries Iceland is excluded

Concerning the data, it can be the case that the data contains missing values or `NA` values. In that case you have to specify `na.rm=TRUE`, for instance when you want to get the mean of per capita income you use

```
mean(df$eur.pc,na.rm=TRUE)
```

You will also regularly have to subset the data, for instance if we want to know the average infant mortality rate for countries that are part of EFTA we use

```
mean(df$infant.mortality[df$efta==1])
```

Or we can use

```
mean(df[df$efta==1,]$infant.mortality)
```

Problem set 1: Frequencies

1. Which countries has the most regions in the dataset? How many does Ireland (`IE`) have?

2. Find the average general and infant mortality rate and compare these averages with those for regions in former Eastern Bloc countries.

3. Countries that have an income below 75% of the EU average qualify for EU funding from the Structural Funds and the Cohesion Plan. The dataset includes a variable (`structural.fund`) to indicate these regions. What is the average mortality rate for these regions? Is it higher or lower than the general average? How does it compare to countries from the former Eastern Bloc and countries that are candidate for EU membership?

4. What proportion of regions actually qualifies for the structural funding? And what proportion of the regions that qualify for EU funding are not in a former Eastern Bloc country?

5. What is the probability that a region has a general mortality rate that is above average, but does not qualify for EU funding?

6. What proportion of regions in the former East bloc is in the highest income category?

7. What proportion of regions in the middle income groups (2 & 3) have above average general mortality rates?

8. Is the Irish mortality rate above or below average? And which proportion of the regions has per capita income levels higher than the Irish average?

Here you have to use `df$ccode=="IE"`

Problem set 2: Exploratory data analysis

1. Plot a histogram for the general mortality rate. What type of distribution would best describe the data? Increase the number of bins by setting `breaks` equal to 10 and then to 100.
2. Plot a boxplot for the general mortality rate. Are there many outliers in the data?[1]
3. Let's examine the distribution of the general mortality rate across the different income groups using the boxplot. Are there any noticeable differences between the groups?[2]
4. Repeat the previous question but now for the
   (a) infant mortality rate
   (b) female mortality rate
   (c) middle aged mortality rate
5. Using a scatterplot, plot the the general mortality rate against the female mortality rate. Are there any patterns in the data? What does this tell use about the relation between the female mortality and general mortality rate.
6. Plot the general mortality rate against the log of income per capita (`eur.pc`). From the data, is there any descriptive evidence for a link between income and mortality? Proceed, by adding points for regions that qualify for EU funding.[3]
7. One could imagine that the number of inhabitants per doctor matters for the mortality rate. Check the relation between inhabitants per doctor and the general and infant mortality rate. What do you notice?

[1] Note that you can adjust the orientation of the boxplot by setting `horizontal=TRUE` in the `boxplot` command.

[2] To plot the distribution across groups you have to use the ~ operator as in `dat$mortality ~dat$income`

[3] In `plot` you need to specify `log="x"`, if you put income on the x-axis. To add the points you need to use the `points` command and subset the data.

Problem set 3: Statistical tests

1. What is the average mortality rate and variance in former Eastern Bloc regions? How do these moments compare to the whole sample? [4]

[4] You can find the standard deviation using `var()`. Note that moments here is just a statistical term for the mean (first moment) and variance (second moment)

2. Let's look at the density of the mortality rate. You can do this by using the command

```
plot(density(df$mortality))
```

   Add the data for regions in the former Eastern Bloc using

```
lines(density(df$mortality[df$eastbloc==1]),lty=2,col="steelblue4")
```

   What do you notice about the data for the former Eastern Bloc?

3. Use a t-test to examine whether the general mortality rate in regions in the former Eastern bloc is the same as in other countries. What is the value of the t-statistic and the p-value? What is the average mortality rate of countries that are not in the former Eastern Bloc?

4. Let's have a look at the regions that qualify for EU funding. Sample 50 regions that qualify for EU funding and 50 that don't. Use a t-test on the two samples to examine whether the mortality rate is higher in regions that qualify for EU funding. What is an issue when using sampling to do this test?

5. For a number of reasons it could be the case that the mortality rate for a specific sex is different from the general mortality rate. Use the data to test whether the female mortality rate is lower than the general mortality rate. Does this also apply when focusing on regions that qualify for EU funding?

6. Use a z-test to determine whether the mortality rate in regions in the former East bloc is different from the average mortality rate. Repeat this test for regions that qualify for structural funds. Are the conclusions the same when you look at the infant mortality rate?[5]

[5] Note that due to missing values in the `structural.fund` variable you have to use `na.omit` to subset the data. Also, the z-test in R will only give the test statistic, so you have to look up the value in the table.

Problem set 4: Regression analysis

Regression analysis

1. Throughout the problem sets we have seen that there are some regional differences across Europe concerning mortality rates. Let's estimate a very simple model where we just regress the mortality rate on the indicator variable for regions that qualify for EU funding. What does the estimated effect tell us? Add the indicator variable for countries in former Eastern Bloc countries. Do the results change? If so, how?

2. Various studies have shown that there is a link between economic welfare and health. What do you think about how income and mortality are related? Let's check the relation using the data by regressing general mortality rate on the log of income per capite. What us the estimated effect of income on the mortality rate?[6]

[6] `log(eur.pc)`

3. Before we continue it is a good idea to check the fit of the model with the data. We can do this by looking at the relation between the fitted values and the residuals to be sure that there is no problem with the standard errors, and we can look at the relation between the observed and fitted values. So check the fit of the model by

   (a) plotting the residuals against the fitted values[7]

   [7] `plot(m1$residual,fitted(m1)`

   (b) plotting the fitted values against the observed values[8]

   [8] `plot(fitted(m1),m1$model$mortality)`

   Which conclusions would you draw here?

4. The variable used to capture income might not be adequate to account for cross-regional differences. Let's re-estimate the model but now using the variable that measures income accounting for purchasing power (`pps.pc`). Does this lead to any noticeable differences?

5. Specify a model with the log of income and add the indicator variable for regions that qualify for EU funding to the model. What is the estimated effect? Do you think that this is a good model specification?

6. Specify a model where you regress the mortality rate on the log of income per capita, the Eastern Bloc indicator, and the variable capturing the achievements of the country concerning the Euro 2012 football championships. Analyse the results and the fit of the model. Do you think that this is a good model?

7. Surely having plenty of doctors in an area is somehow related to the mortality rate. Estimate a model regressing the mortality rate on the log of income, the Eastern Bloc indicator, and the log of the

number of inhabitants per doctor (`inh.per.doc`). What are your conclusions? Are the results the same when you use the infant mortality rate?

8. Finally, specify a model where the outcome variable (`mortality`) is also log-transformed and regress it on the log of income per capita and the Eastern Bloc indicator.[9] How does this model compare to the one where the outcome variable is not log-transformed?

[9] To plot the fitted values against the observed values you have to use `plot(m9$model[,1],m9$fitted)`