Introduction to R: Part VI

An Economics Data Session

Alexandre Perera i Lluna [1,2]

[1]Centre de Recerca en Enginyeria Biomèdica (CREB)
Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial (ESAII)
Universitat Politècnica de Catalunya
mailto:Alexandre.Perera@upc.edu

[2]Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN)

Jan 2011 / Introduction to R
Universitat Rovira i Virgili

Contents I

Penn World Table Session

PWT

- The Penn World Table (PWT) displays a set of national accounts economic time series covering many countries.
- It also provides information about relative prices within and between countries, as well as demographic data and capital stock estimates.
- The Table contains data on about 30 variables for about 167 countries over some or all the years 1950-98.

The Penn World Tables are described in **Robert Summers** and **Alan Heston** "The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950-1988", Quarterly Journal of Economics, May 1991, pp.327–368

Penn World Table Session

R has PWT on the package *pwt*, in order to retrieve we first should install the package, then:

```
> install.packages(pwt)
> library(pwt)
```

Penn World Table Session

```
> help(package=pwt)
  Package:              pwt
Version:              6.3-0
Date:                 2009-10-15
Title:                Penn World Table
Author:               Achim Zeileis, Guan Yang
Maintainer:           Achim Zeileis <Achim.Zeileis@R-project.org>
Description:          The Penn World Table provides purchasing power
                      parity and national income accounts converted
                      to international prices for 189 countries for
                      some or all of the years 1950-2007.
License:              GPL-2
Repository:           CRAN
Date/Publication:     2009-10-14 22:32:00
Built:                R 2.11.0; ; 2010-04-23 15:24:46 UTC; unix

Index:

pwt5.6                    Penn World Table 5.6
pwt6.1                    Penn World Table 6.1
pwt6.2                    Penn World Table 6.2
pwt6.3                    Penn World Table 6.3
```

Getting to know PWT6.3

First, let's be familiar with our dataset.

```
> d <- pwt6.3
> dim(d)
```

```
[1] 11020    36
```

```
> names(d)
```

```
 [1] "country"   "isocode"   "year"      "pop"
 [5] "xrat"      "currency"  "ppp"       "cgdp"
 [9] "cc"        "cg"        "ci"        "p"
[13] "pc"        "pg"        "pi"        "openc"
[17] "cgnp"      "y"         "yeks"      "ycpdw"
[21] "rgdpl"     "rgdpl2"    "rgdpch"    "rgdptt"
[25] "openk"     "kc"        "kg"        "ki"
[29] "rgdpeqa"   "rgdpwok"   "rgdpl2wok" "rgdpl2pe"
[33] "rgdpl2te"  "rgdpl2th"  "grgdpch"   "grgdpl2"
```

PWT Main Variable Names I

- pop: population
- xrat: Exchange rate
- ppp: Purchasing power partity
- cgdp: Real Gross domestic product per capita
  - cc: Consumption Share of CGDP
  - ci: Investment Share of CGDP
  - cg: Government Share of CGDP
  - cnfb: Net Foreign Balance
- year
- country
- pg: Price Level of Government
- p: Price Level of Gross domestic product
- pc:. Price Level of Consumption
- pi: Price Level of Investment
- openc: Openness
- cgnp: Gross National Produc
- csave: Current Savings
- y: GNP Relative to the United States (US=100)
- rgdpl: Real GDP per capita (Laspeyres)

PWT Main Variable Names II

- rgdpch: Real GDP per capita (Chain)
- rgdpeqa: Real GDP chain per equivalent adult
- rgdpwok: Real GDP chain per worker
- rgdptt: Adjustment ofr changes in the Terms of Trade
- openk: Openness
- kc: Consumption Share of RGDPL
- kg: Government Share of RGDPL
- ki: Investment Share of RGDPL
- knfb: Net Foreign Balance: KNFB
- KapW Capital Stock per Worker
- KapD: Producers Durables
- KapNR: Non Residential Construction (% of Capital Stock)
- KapO: Other Construction (% of Capital Stock)
- KapR: Residential Construction (% of Capital Stock)
- KapT: Transport Equipment (% of Capital Stock)
- STLIV: Standard of Living
- KNDP: Net Domestic Product
- GRGDPCH: Growth Rate of RGDPCH

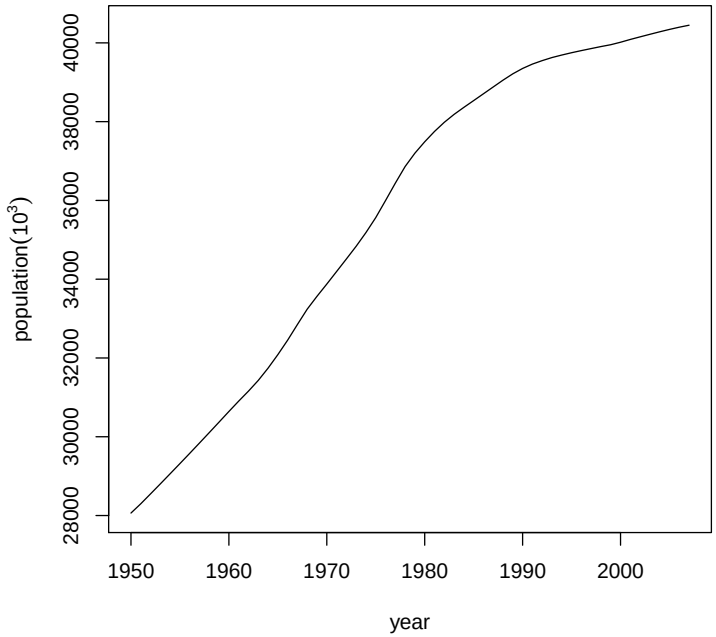
Dataset description: Countries

```
> class(d$country)
[1] "factor"
> nlevels(d$country)
[1] 190
> head(levels(d$country))
[1] "Afghanistan"          "Albania"
[3] "Algeria"              "Angola"
[5] "Antigua and Barbuda" "Argentina"
> tail(levels(d$country))
[1] "Vanuatu"   "Venezuela" "Vietnam"   "Yemen"
[5] "Zambia"    "Zimbabwe"
> (spain <- which(d$country == "Spain"))
 [1] 8991 8992 8993 8994 8995 8996 8997 8998 8999 9000 9001
[12] 9002 9003 9004 9005 9006 9007 9008 9009 9010 9011 9012
[23] 9013 9014 9015 9016 9017 9018 9019 9020 9021 9022 9023
[34] 9024 9025 9026 9027 9028 9029 9030 9031 9032 9033 9034
[45] 9035 9036 9037 9038 9039 9040 9041 9042 9043 9044 9045
[56] 9046 9047 9048
```

Years

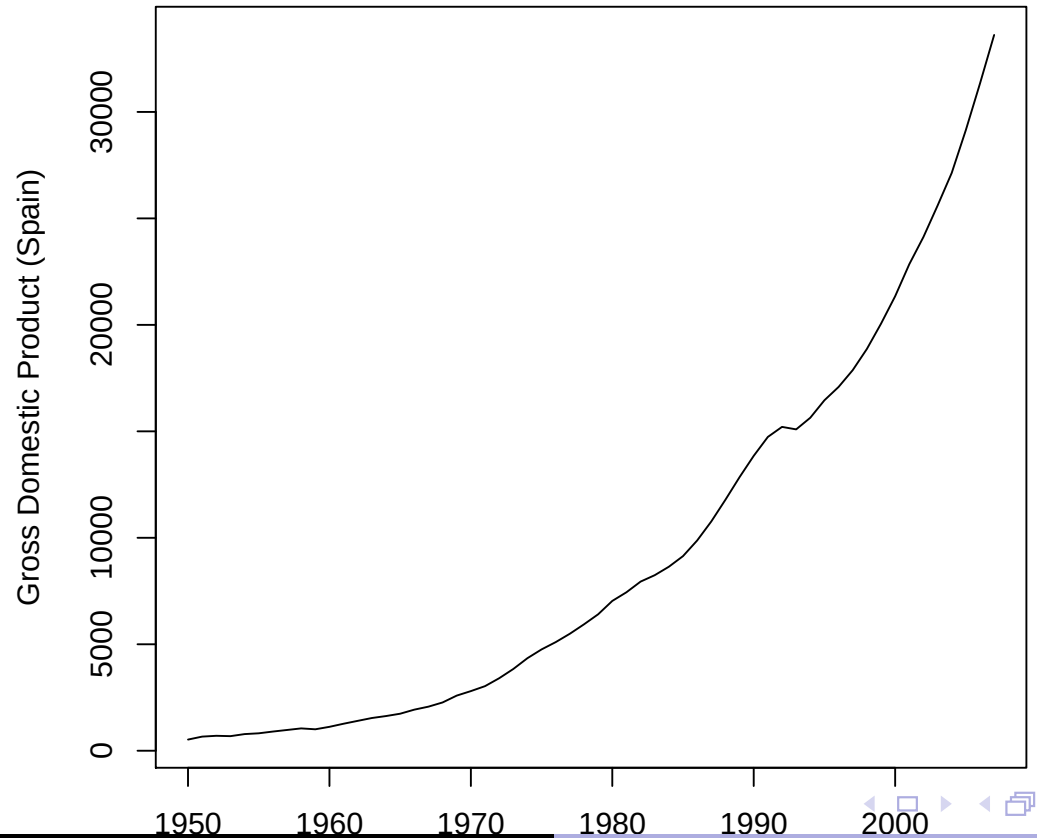
```
> unique(d$year)

 [1] 1950 1951 1952 1953 1954
 [6] 1955 1956 1957 1958 1959
[11] 1960 1961 1962 1963 1964
[16] 1965 1966 1967 1968 1969
[21] 1970 1971 1972 1973 1974
[26] 1975 1976 1977 1978 1979
[31] 1980 1981 1982 1983 1984
[36] 1985 1986 1987 1988 1989
[41] 1990 1991 1992 1993 1994
[46] 1995 1996 1997 1998 1999
[51] 2000 2001 2002 2003 2004
[56] 2005 2006 2007
```

```
> plot(d$year[spain],
            d$pop[spain],
            xlab="year",
            ylab=expression(
              plain(population )*(10^3)),
            type='l')
```



Gross Domestic Product per capita (Spain)

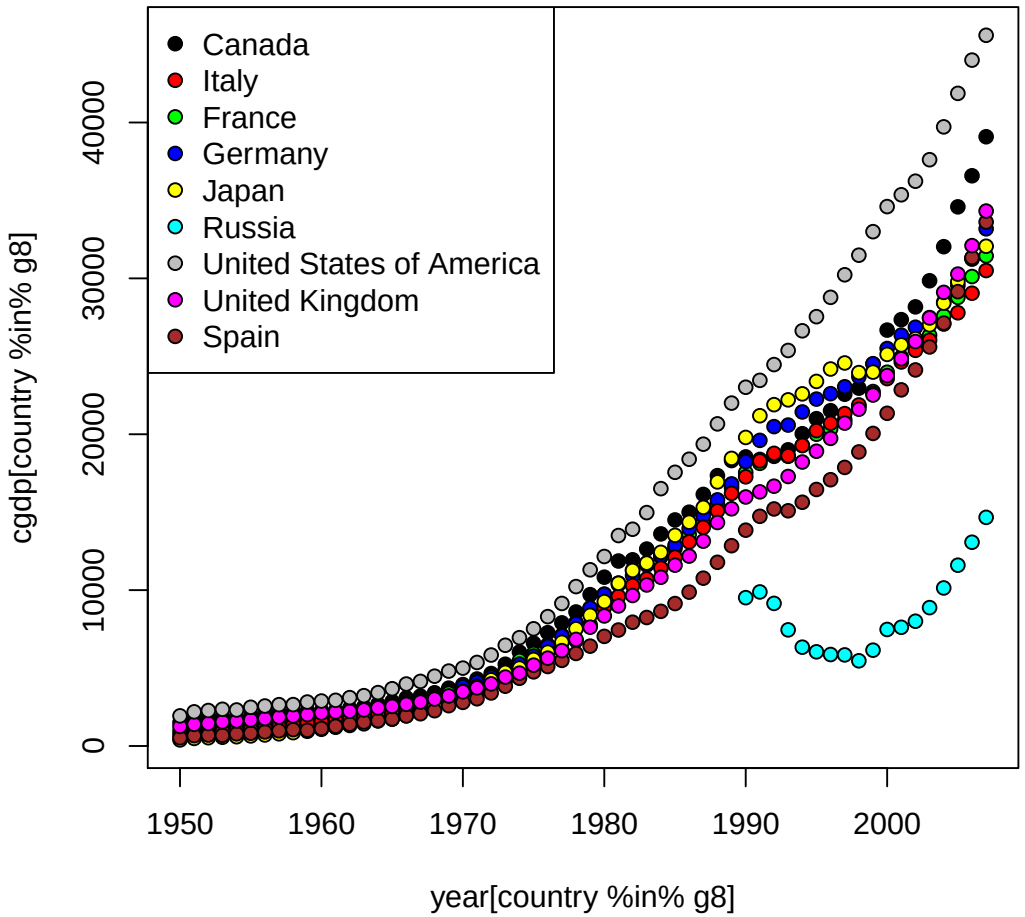
```
> plot(d$year[spain], d$cgdp[spain], xlab = "year",
+      ylab = "Gross Domestic Product (Spain)",
+      type = "l")
```



Gross Domestic Product per capita (G8+Spain)

```
> g8 <- c("Canada", "Italy", "France", "Germany",
+     "Japan", "Russia", "United States of America",
+     "United Kingdom", "Spain")
> cols <- c("black", "red", "green", "blue",
+     "yellow", "cyan", "gray", "magenta",
+     "brown")
> attach(d, warn.conflicts = FALSE)
> plot(year[country %in% g8], cgdp[country %in%
+     g8], pch = 21, bg = cols[match(d$country[d$country %in%
+     g8], g8)])
> legend("topleft", legend = g8, pch = 21,
+     pt.bg = cols)
```

Gross Domestic Product per capita (G8+Spain)



Mean Gross Domestic Product in past 50 years

Question

Which has been the mean gross domestic product per capita in all years in the G8+Spain?

```
> by(d$cgdp, d$country, mean)[g8]

d$country
                   Canada                     Italy
                12339.937                 10659.572
                   France                   Germany
                11063.889                        NA
                    Japan                    Russia
                11501.853                        NA
United States of America            United Kingdom
                15433.164                 10761.944
                    Spain
                 9388.628
```

Mean Gross Domestic Product in past 50 years

Question

Which has been the mean gross domestic product per capita in all years in the G8+Spain?

```
> by(d$cgdp, d$country, mean)[g8]

d$country
                  Canada                   Italy
               12339.937               10659.572
                  France                 Germany
               11063.889                      NA
                   Japan                  Russia
               11501.853                      NA
United States of America          United Kingdom
               15433.164               10761.944
                   Spain
                9388.628
```

Dealing with NA's

```
> cgdp[country == "Germany"]
 [1]        NA        NA        NA        NA
 [5]        NA        NA        NA        NA
 [9]        NA        NA        NA        NA
[13]        NA        NA        NA        NA
[17]        NA        NA        NA        NA
[21]  3808.967  4105.509  4426.771  4831.593
[25]  5256.147  5768.995  6392.438  7058.109
[29]  7860.196  8859.355  9727.715 10354.658
[33] 10932.733 11609.992 12215.551 12872.955
[37] 13953.899 14763.932 15792.964 16818.171
[41] 18246.479 19609.792 20491.730 20591.520
[45] 21432.337 22256.316 22603.238 23049.579
[49] 23689.091 24529.103 25502.508 26361.604
[53] 26867.506 27486.528 28475.380 29547.739
[57] 31291.090 33181.091
> all(!is.na(cgdp[country == "Germany"]))
[1] FALSE
> all(!is.na(cgdp[country == "Spain"]))
[1] TRUE
```

Dealing with NA's

```
> by(cgdp, country, function(x) mean(x,
+     na.rm = TRUE))[g8]

country
                Canada                  Italy
             12339.937              10659.572
                France                Germany
             11063.889              16647.981
                 Japan                 Russia
             11501.853               8510.398
United States of America         United Kingdom
             15433.164              10761.944
                 Spain
              9388.628
```

Past 10 years?

Which has been the mean gross domestic product per capita in the past 10 years in the G8+Spain?

```
> by(cgdp[year > 1990], country[year > 1990],
+     function(x) mean(x, na.rm = TRUE))[g8]

country[year > 1990]
                   Canada                     Italy
                26053.526                  23282.679
                   France                   Germany
                23493.737                  25115.656
                    Japan                    Russia
                25499.364                   8451.076
United States of America            United Kingdom
                33294.899                  23517.539
                    Spain
                21539.701
```

Past 10 years?

Which has been the mean gross domestic product per capita in the past 10 years in the G8+Spain?

```
> by(cgdp[year > 1990], country[year > 1990],
+      function(x) mean(x, na.rm = TRUE))[g8]

country[year > 1990]
                   Canada                     Italy
                26053.526                  23282.679
                   France                   Germany
                23493.737                  25115.656
                    Japan                    Russia
                25499.364                   8451.076
United States of America            United Kingdom
                33294.899                  23517.539
                    Spain
                21539.701
```

Regression

Could we predict the exchangerate (xrat) given:

- population (pop)
- Gross Domestic Product per capita (cgdp)
- Puchasing Power Parity (ppp)
- Openness (openc)

```
> mod <- lm(xrat ~ pop + cgdp + ppp + openc,
+     subset = spain)
```

Regression

Could we predict the exchangerate (xrat) given:

- population (pop)
- Gross Domestic Product per capita (cgdp)
- Puchasing Power Parity (ppp)
- Openness (openc)

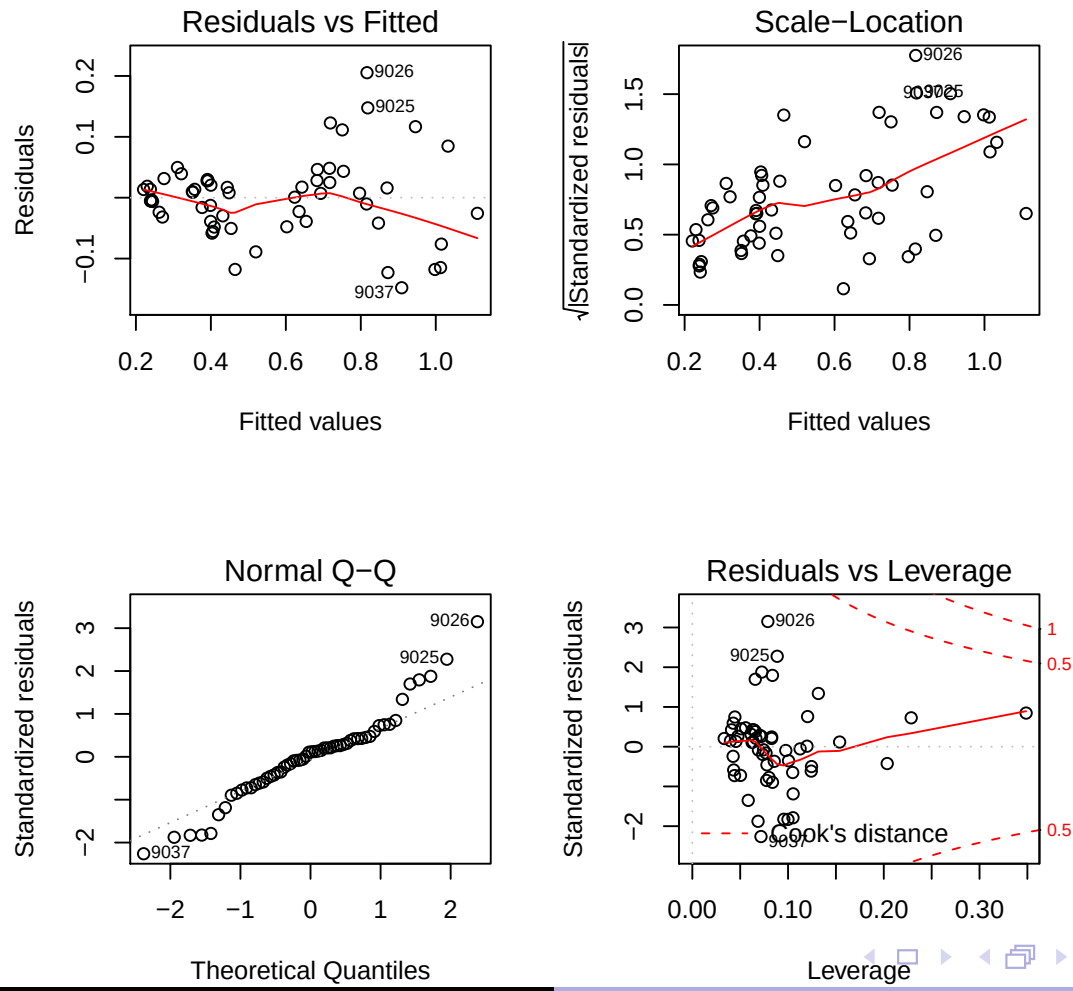
```
> mod <- lm(xrat ~ pop + cgdp + ppp + openc,
+     subset = spain)
```

Regression

```
> summary(mod)

Call:
lm(formula = xrat ~ pop + cgdp + ppp + openc, subset = spain)

Residuals:
      Min        1Q    Median        3Q       Max
-0.147923 -0.037132  0.007242  0.028070  0.205417

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.213e+00  2.279e-01   5.323 2.11e-06 ***
pop         -4.488e-05  8.410e-06  -5.337 2.01e-06 ***
cgdp        -3.403e-05  4.108e-06  -8.285 3.93e-11 ***
ppp          1.620e+00  1.809e-01   8.954 3.46e-12 ***
openc        2.001e-02  2.342e-03   8.545 1.53e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06791 on 53 degrees of freedom
Multiple R-squared: 0.9384, Adjusted R-squared: 0.9337
F-statistic: 201.8 on 4 and 53 DF,  p-value: < 2.2e-16
```

Linear Regression Residuals

```
> layout(matrix(c(1, 2, 3, 4), 2, 2))
> plot(mod)
```



Data Standardization

Standardization

All variable k to have $\mu_k = 0$ and $\sigma_k = 1$.

```
> s <- d[spain, ]
> data <- as.data.frame(scale(s[, c("xrat", "pop", "cgdp",
+     "ppp", "openc")]))
> names(data) <- c("sxrat", "spop", "scgdp", "sppp", "sopenc")
> s <- cbind(s, data)
> names(s)

 [1] "country"   "isocode"   "year"      "pop"       "xrat"
 [6] "currency"  "ppp"       "cgdp"      "cc"        "cg"
[11] "ci"        "p"         "pc"        "pg"        "pi"
[16] "openc"     "cgnp"      "y"         "yeks"      "ycpdw"
[21] "rgdpl"     "rgdpl2"    "rgdpch"    "rgdptt"    "openk"
[26] "kc"        "kg"        "ki"        "rgdpeqa"   "rgdpwok"
[31] "rgdpl2wok" "rgdpl2pe"  "rgdpl2te"  "rgdpl2th"  "grgdpch"
[36] "grgdpl2"   "sxrat"      "spop"      "scgdp"     "sppp"
[41] "sopenc"
```

Regression (after std)

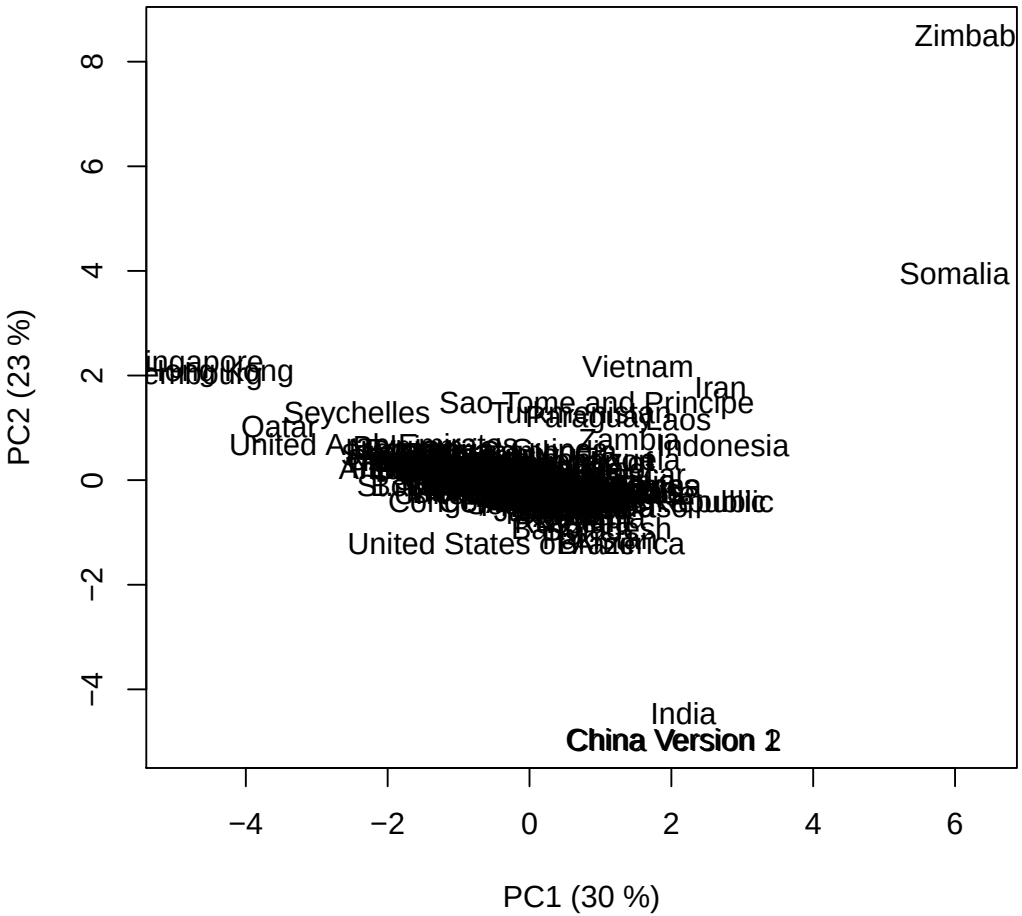
```
> mod2 <- lm(sxrat ~ spop + scgdp + sppp + sopenc, data = s)
> summary(mod2)

Call:
lm(formula = sxrat ~ spop + scgdp + sppp + sopenc, data = s)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56074 -0.14076  0.02745  0.10640  0.77868

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  3.255e-16  3.380e-02 9.63e-15        1
spop        -7.081e-01  1.327e-01   -5.337 2.01e-06 ***
scgdp       -1.174e+00  1.417e-01   -8.285 3.93e-11 ***
sppp         1.545e+00  1.726e-01    8.954 3.46e-12 ***
sopenc       1.207e+00  1.413e-01    8.545 1.53e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2574 on 53 degrees of freedom
Multiple R-squared: 0.9384, Adjusted R-squared: 0.9337
F-statistic: 201.8 on 4 and 53 DF,  p-value: < 2.2e-16

> detach(d)
```

Country variance analysis

```
> d2007 <- d[d$year == 2007, c("xrat", "pop", "cgdp", "ppp",
+     "openc")]
> countries <- d$country[d$year == 2007]
> mod <- prcomp(~., data = d2007, na.action = na.exclude,
+     center = TRUE, scale = TRUE)
> summary(mod)

Importance of components:
                          PC1   PC2   PC3   PC4   PC5
Standard deviation     1.234 1.061 0.960 0.873 0.815
Proportion of Variance 0.305 0.225 0.184 0.153 0.133
Cumulative Proportion  0.305 0.530 0.714 0.867 1.000
```

Score plot

```
> library(pls)
> scoreplot(mod, labels = countries)
```



Removing outliers

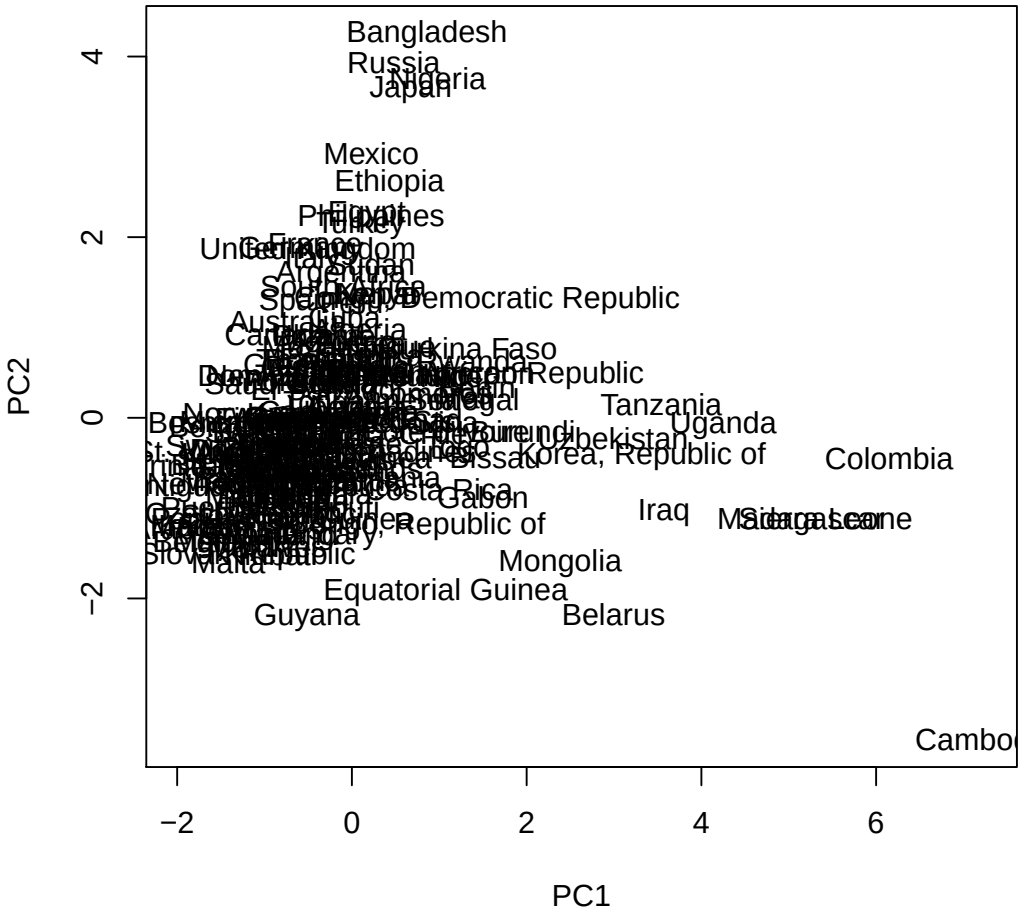
```
> coun <- which(is.na(rowMeans(d2007)))
> countries[coun]

[1] Bahrain Serbia
190 Levels: Afghanistan Albania Algeria Angola ... Zimbabwe

> d2007 <- d2007[-coun, ]
> countries <- countries[-coun]
> library(mvoutlier)
> outs <- aq.plot(d2007, quan = 1)

Projection to the first and second robust principal components.
Proportion of total variation (explained variance): 0.9682581

> table(outs$outliers)

FALSE  TRUE
  164    24

> d2007 <- d2007[!outs$outliers, ]
> countries <- countries[!outs$outliers]
```

Removing outliers

```
> mod <- prcomp(d2007, center = TRUE, scale = TRUE)
> summary(mod)

Importance of components:
                          PC1   PC2   PC3   PC4    PC5
Standard deviation     1.434 1.155 0.945 0.772 0.3476
Proportion of Variance 0.411 0.267 0.179 0.119 0.0242
Cumulative Proportion  0.411 0.678 0.857 0.976 1.0000
```

Removing outliers

```
> scoreplot(mod$x[, 1:2], labels = countries)
```



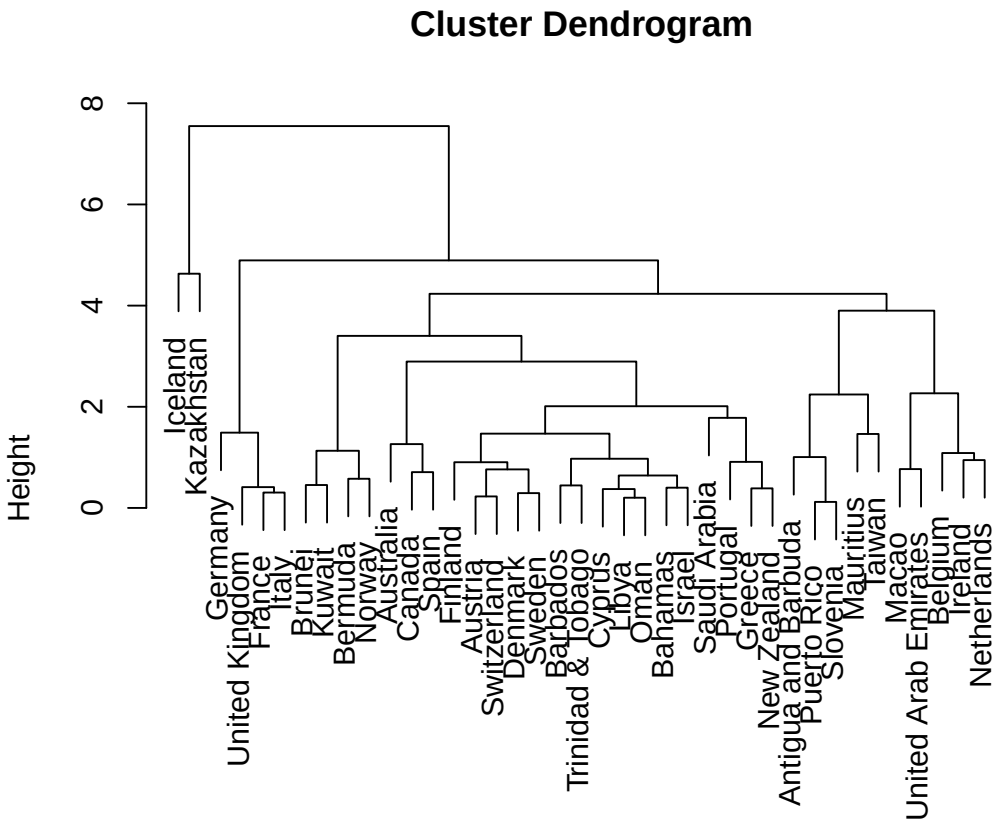
Clustering

```
> di <- dist(scale(d2007))
> h <- hclust(di)
> memb <- cutree(h, k = 6)
> sp <- which(countries == "Spain")
> idgrup <- which(memb == memb[sp])
> countries[idgrup]
```

```
 [1] Antigua and Barbuda  Australia            Austria
 [4] Bahamas              Barbados             Belgium
 [7] Bermuda              Brunei               Canada
[10] Cyprus               Denmark              Finland
[13] France               Germany              Greece
[16] Iceland              Ireland              Israel
[19] Italy                Kazakhstan           Kuwait
[22] Libya                Macao                Mauritius
[25] Netherlands          New Zealand          Norway
[28] Oman                 Portugal             Puerto Rico
[31] Saudi Arabia         Slovenia             Spain
[34] Sweden               Switzerland          Taiwan
[37] Trinidad & Tobago    United Arab Emirates United Kingdom
190 Levels: Afghanistan Albania Algeria Angola ... Zimbabwe
```

Clustering

```
> rownames(d2007) <- countries
> di <- dist(scale(d2007[idgrup, ]))
> h <- hclust(di)
> plot(h)
```



End Part VI