W271 Live Session 12: Analysis of Panel Data 2

Devesh Tiwari

8/1/2017

Main topics covered in Week 12 (Async Unit 12)

```
- Fixed Effect Model
- A Digression: differencing when there are more than 2 time periods
- Random effect model
- Fixed effect vs. random effect models
```

Readings:

W2016: Jeffrey Wooldridge. *Introductory Econometrics: A Modern Approach.* 6th edition. Cengage Learning

```
- Ch. 14.1 - 14.2
- [package plm](https://cran.r-project.org/web/packages/plm/plm.pdf)
- [plm vignettes](https://cran.r-project.org/web/packages/plm/vignettes/plm.pdf)
```

Breakout Session 1

Imagine you are on a data science team working for a company that is interested in expanding its market in a developing country, Tiwaristan. The company's main source of revenue is its mobile application, which is why they are keen on understanding why users are happy or unhappy. By some miracle, they were able to conduct 4 rounds of surveys (once every month). They have survey data on 30 users from 6 different provinces. They have collected the following data:

(a) Customer satisfaction score (0 - 100)

(b) Number of hours the phone was not able to connect to the data network due to bandwidth related issues (conectivity).

(c) Gender

(d) Annual income

(e) Education level

(f) Province

(g) Date / round of survey

1. Describe this dataset. How many observations are there in this dataset? What is the unit of analysis?
2. What estimation challenges would you face if you examined the relationship between customer satisfication and connectivity? How would you overcome them?
3. Assume that you no longer have access to which province a user lives in. Can you use the same strategy you used to estimate #2 above to estimate the relationship with the other covariates and the DV? What are the benefits of using this approach? Costs?

Breakout Session 2

In this exercise, we will analyze a simple panel dataset from the World Bank. This dataset has country level data from the years 2005 and 2010. In particular, it has data on a country's GDP per capita, the amount of money received from remittances abroad, and population. We are interested in estimating the relationship between GDP per capita, reminttances, and population. BE SURE TO CONDUCT AN EDA AND MAKE ANY NECCESSARY TRANSFORMATIONS TO THE DATA!!

1. Load this dataset as a *plm* package.
2. Estimate the pooling estimator using the plm function in R, and then examine the residuals. What do you notice?
3. Estimate the same data using a "within" or "fixed effect" estimator. Why do you think that this estimator is called a "within" estimator? What are we measuring? Examine these residuals as well.
4. Using the lm function in r, estimate a model that contains a dummy variable for each country. Compare these results to the fixed effects model you just estimated. What do you notice?
5. Compare your within estimate to the pooled-OLS estimate. Do they seem different? If so, what does that tell you?

Group Discusssion: Random effects

- Fixed effects estimator is strict in the sense that we can only reliably estimate time-variant variables that have a decent amount of within-unit variation.
- We might want to estimate a time-variant varible using more information than the fixed effects model, or we might want to estimate time-invarient variables.
- Doing so means that we have to assume that our variables of interest are exogenous with respect to the error term. If they are not exogenous, we have to interpret the results with a lot of caution because the coefficients will be biased.
- So, even though we are "OK" with generating biased estimates, we still should not use the pooled-OLS estimator because the errors within cross-sectional units are likely correlated.
- The random effects model helps us deal with this by "removing"" this correlation from the observations.

Breakout Session 3: Random effects model

1. Run a random effects model using the plm package. What do the "effects" portion of the output refer to? What is theta? Can you describe theta in plain English?
2. Examine and compare the regression outcome of the pooled-OLS, within estimator, and the random effects model. Pay attention to both the coefficients and standard errors. What do you notice? Which model produces the smallest standard errors? Largest? If your coefficients are different across models, what does that tell you about the relationship between the covariate and the error?
3. Run a Hausman test to determine whether or not a fixed effects model or a random effects model is appropriate.

Open questions

Remainder of time

Fixed effects code

```
rm(list = ls())
library(plm)
```

```
## Loading required package: Formula
```

```
setwd("~/Documents/Projects/MIDS/Summer 2017")
wb <- read.csv("wb.csv", stringsAsFactors = FALSE)

head(wb)
```

```
##   country year remittance gdp_capita keep population
## 1 Albania 2005   6.51e+06       2710    1    2960000
## 2 Albania 2010   2.42e+07       4090    1    2960000
## 3 Algeria 2005   2.70e+07       3100    1   34650000
## 4 Algeria 2010   2.80e+07       4470    1   34650000
## 5  Angola 2005   2.15e+08       1580    1   19550000
## 6  Angola 2010   7.14e+08       3890    1   19550000
```

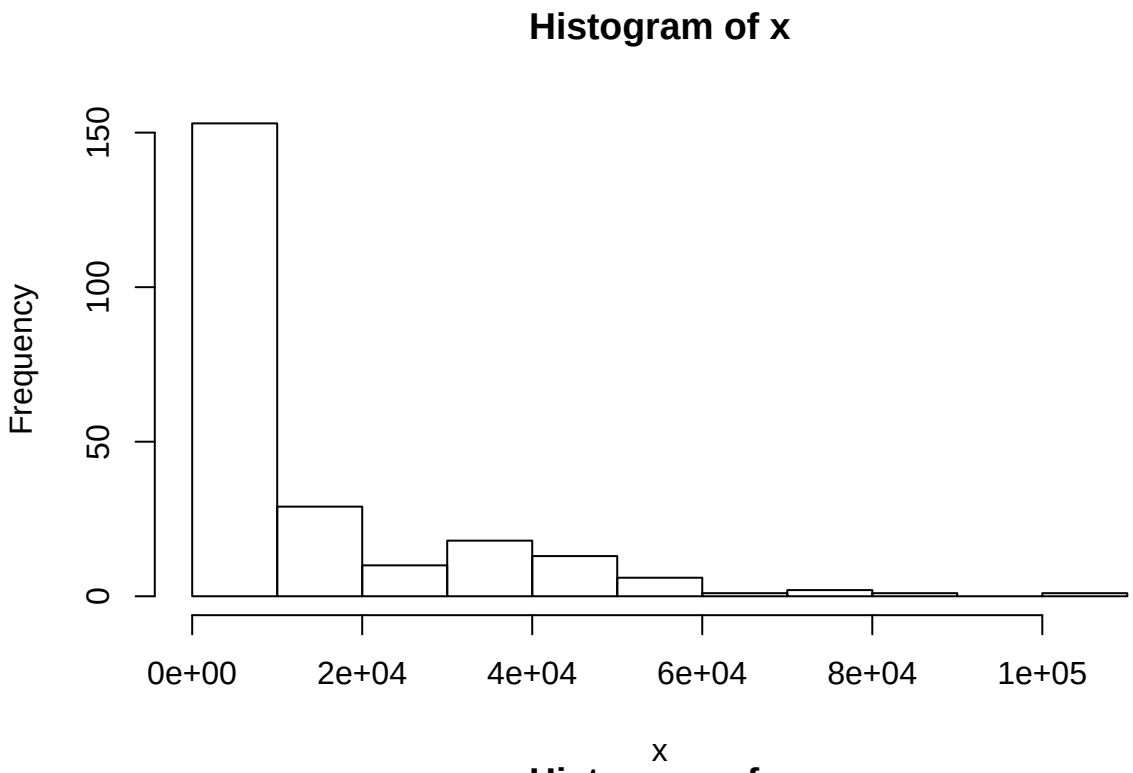
```
summary(wb)
```

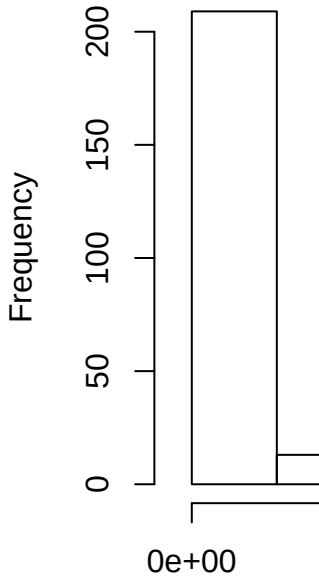
```
##    country               year        remittance          gdp_capita
##  Length:234         Min.   :2005   Min.   :0.000e+00   Min.   :   141
##  Class :character   1st Qu.:2005   1st Qu.:4.772e+07   1st Qu.:  1272
##  Mode  :character   Median :2008   Median :1.955e+08   Median :  4690
##                     Mean   :2008   Mean   :2.069e+09   Mean   : 12919
##                     3rd Qu.:2010   3rd Qu.:1.340e+09   3rd Qu.: 17800
##                     Max.   :2010   Max.   :5.080e+10   Max.   :103000
##       keep     population
##  Min.   :1   Min.   :4.860e+05
##  1st Qu.:1   1st Qu.:4.430e+06
##  Median :1   Median :1.055e+07
##  Mean   :1   Mean   :5.038e+07
##  3rd Qu.:1   3rd Qu.:3.315e+07
##  Max.   :1   Max.   :1.320e+09
```

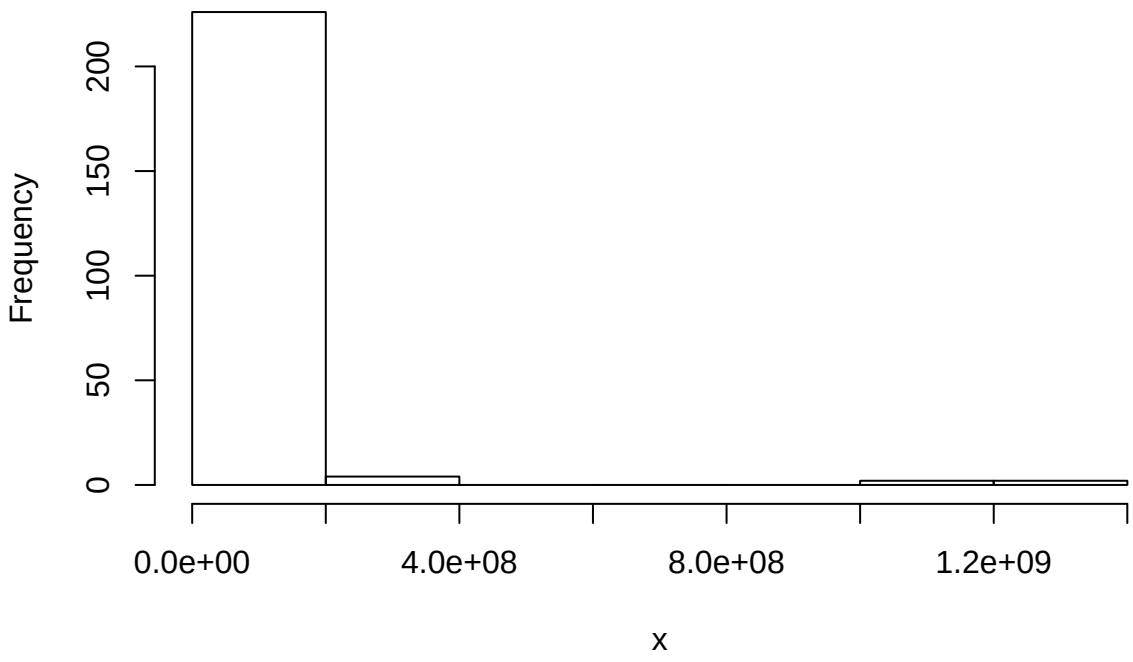
```
all(table(wb$country) == 2) # Every country has two observations
```

```
## [1] TRUE
```

```
apply(wb[, c("gdp_capita", "remittance", "population")], 2,
      function(x) hist(x))
```








```
## $gdp_capita
## $breaks
##  [1]      0  10000  20000  30000  40000  50000  60000  70000  80000  90000
## [11] 100000 110000
##
## $counts
```

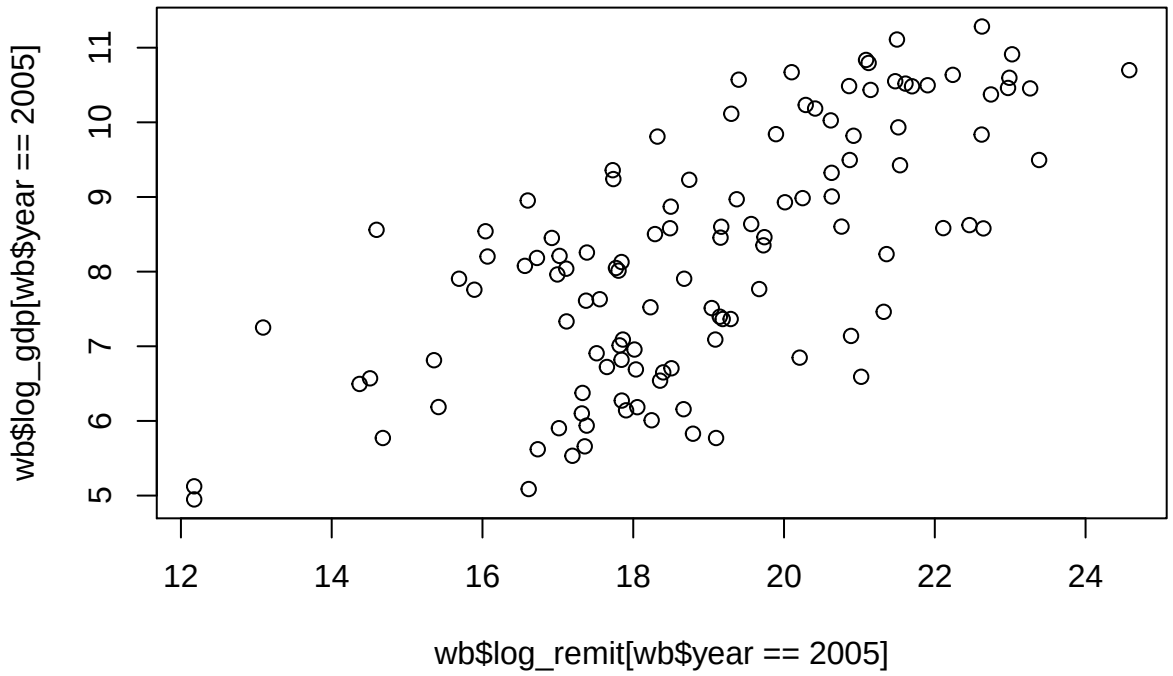
```
##  [1] 153  29  10  18  13   6   1   2   1   0   1
##
## $density
##  [1] 6.538462e-05 1.239316e-05 4.273504e-06 7.692308e-06 5.555556e-06
##  [6] 2.564103e-06 4.273504e-07 8.547009e-07 4.273504e-07 0.000000e+00
## [11] 4.273504e-07
##
## $mids
##  [1]   5000  15000  25000  35000  45000  55000  65000  75000  85000  95000
## [11] 105000
##
## $xname
## [1] "x"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $remittance
## $breaks
##  [1] 0.0e+00 5.0e+09 1.0e+10 1.5e+10 2.0e+10 2.5e+10 3.0e+10 3.5e+10
##  [9] 4.0e+10 4.5e+10 5.0e+10 5.5e+10
##
## $counts
##  [1] 209  13   7   1   1   1   0   0   0   1   1
##
## $density
##  [1] 1.786325e-10 1.111111e-11 5.982906e-12 8.547009e-13 8.547009e-13
##  [6] 8.547009e-13 0.000000e+00 0.000000e+00 0.000000e+00 8.547009e-13
## [11] 8.547009e-13
##
## $mids
##  [1] 2.50e+09 7.50e+09 1.25e+10 1.75e+10 2.25e+10 2.75e+10 3.25e+10
##  [8] 3.75e+10 4.25e+10 4.75e+10 5.25e+10
##
## $xname
## [1] "x"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $population
## $breaks
## [1] 0.0e+00 2.0e+08 4.0e+08 6.0e+08 8.0e+08 1.0e+09 1.2e+09 1.4e+09
##
## $counts
## [1] 226   4   0   0   0   2   2
##
## $density
```

```
## [1] 4.829060e-09 8.547009e-11 0.000000e+00 0.000000e+00 0.000000e+00
## [6] 4.273504e-11 4.273504e-11
##
## $mids
## [1] 1.0e+08 3.0e+08 5.0e+08 7.0e+08 9.0e+08 1.1e+09 1.3e+09
##
## $xname
## [1] "x"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

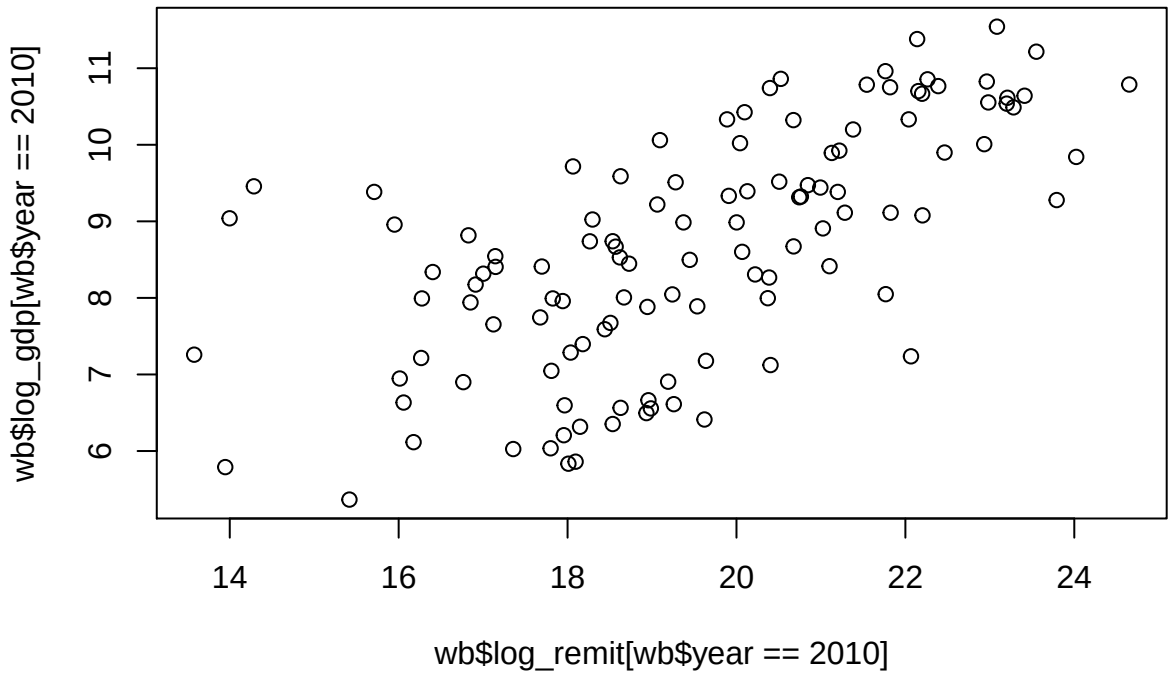
```
# Distributions are all skewed.
# Be sure to examine for each year too
# Will take log, note that remittances are zero for Paraguay.
wb[wb$remittance == 0, ]
```

```
##      country year remittance gdp_capita keep population
## 169 Paraguay 2005          0       1510    1    6005000
## 170 Paraguay 2010          0       3230    1    6005000
```

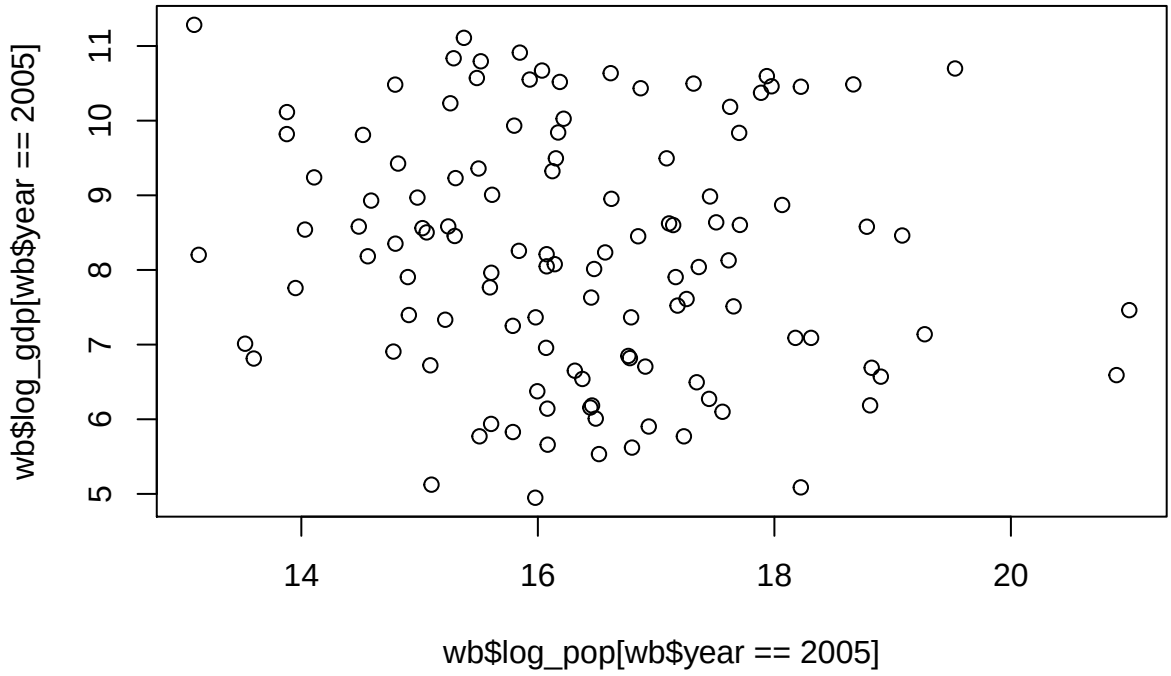
```
wb <- wb[wb$remittance > 0, ]

wb$log_gdp <- log(wb$gdp_capita)
wb$log_remit <- log(wb$remittance)
wb$log_pop <- log(wb$population)

plot(wb$log_gdp[wb$year == 2005] ~ wb$log_remit[wb$year == 2005])
```



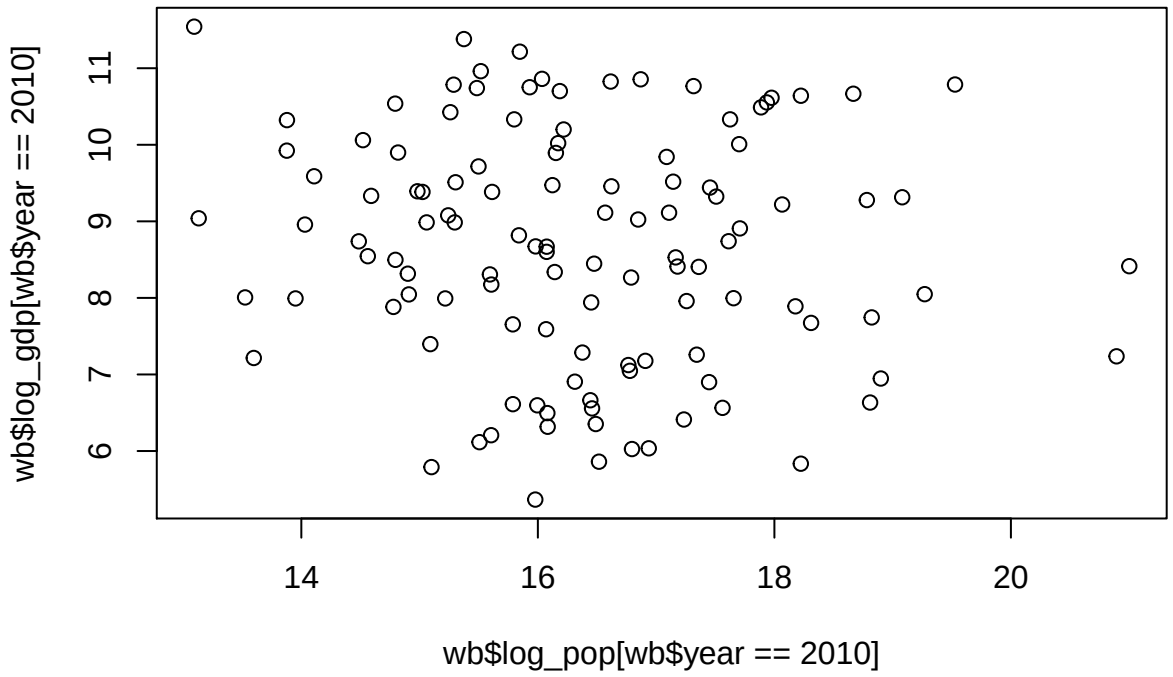

```
plot(wb$log_gdp[wb$year == 2010] ~ wb$log_remit[wb$year == 2010])
```



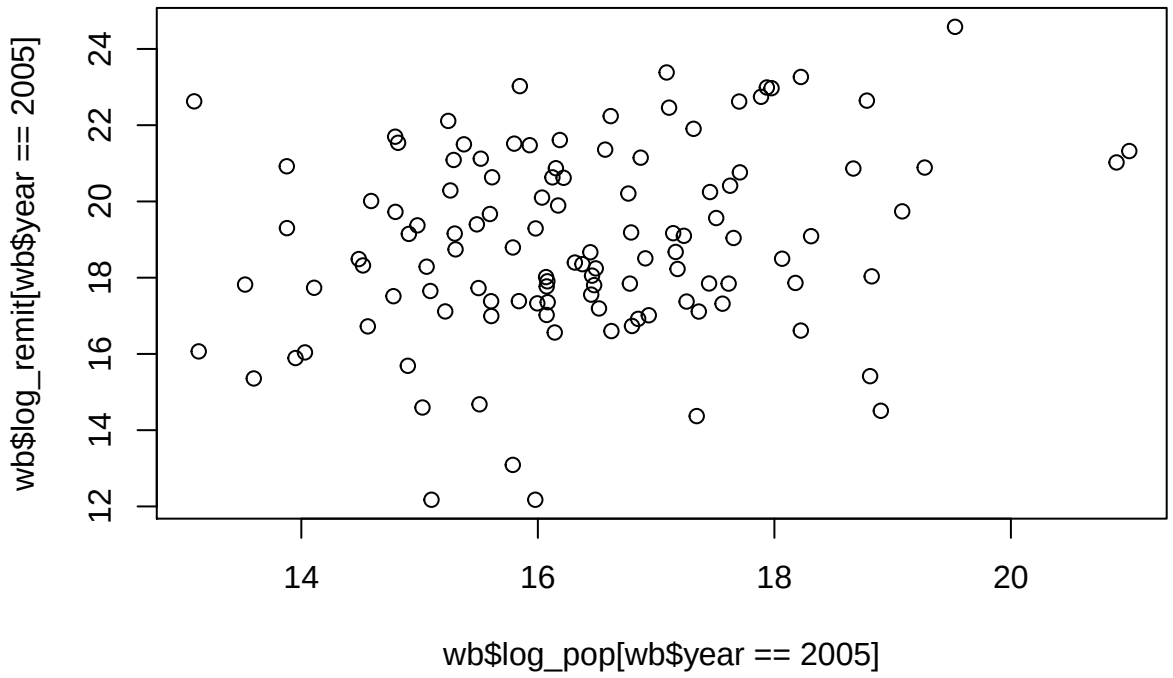

```
plot(wb$log_gdp[wb$year == 2005] ~ wb$log_pop[wb$year == 2005])
```



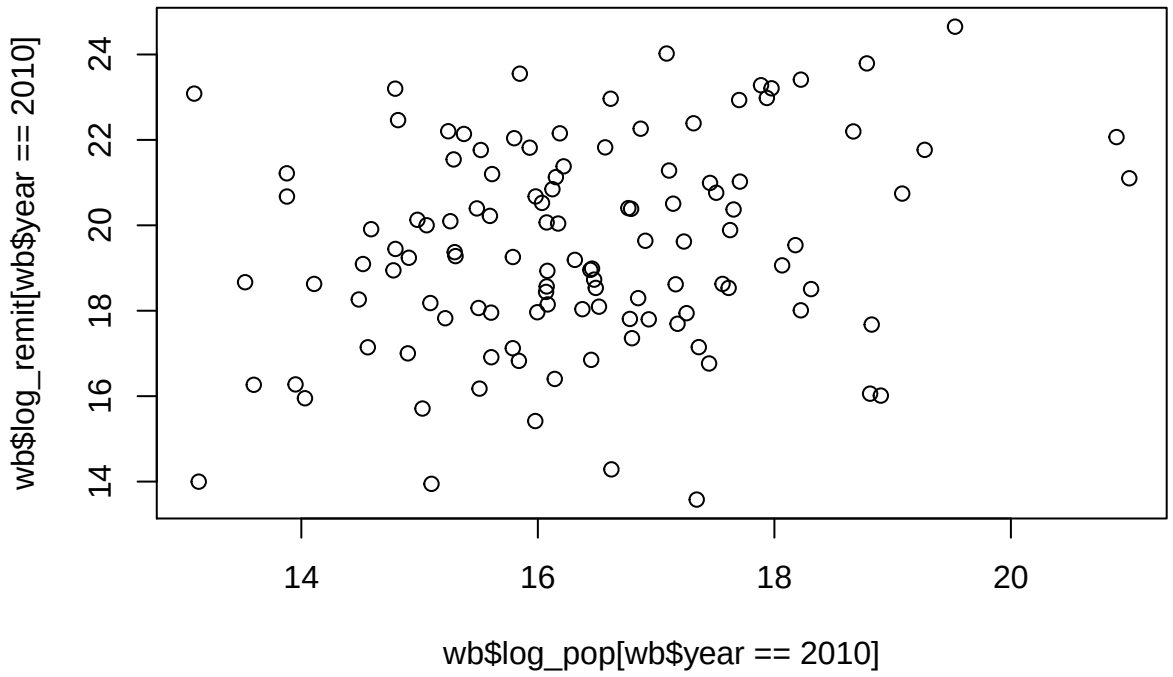

```
plot(wb$log_gdp[wb$year == 2010] ~ wb$log_pop[wb$year == 2010])
```




```
plot(wb$log_remit[wb$year == 2005] ~ wb$log_pop[wb$year == 2005])
```




```
plot(wb$log_remit[wb$year == 2010] ~ wb$log_pop[wb$year == 2010])
```



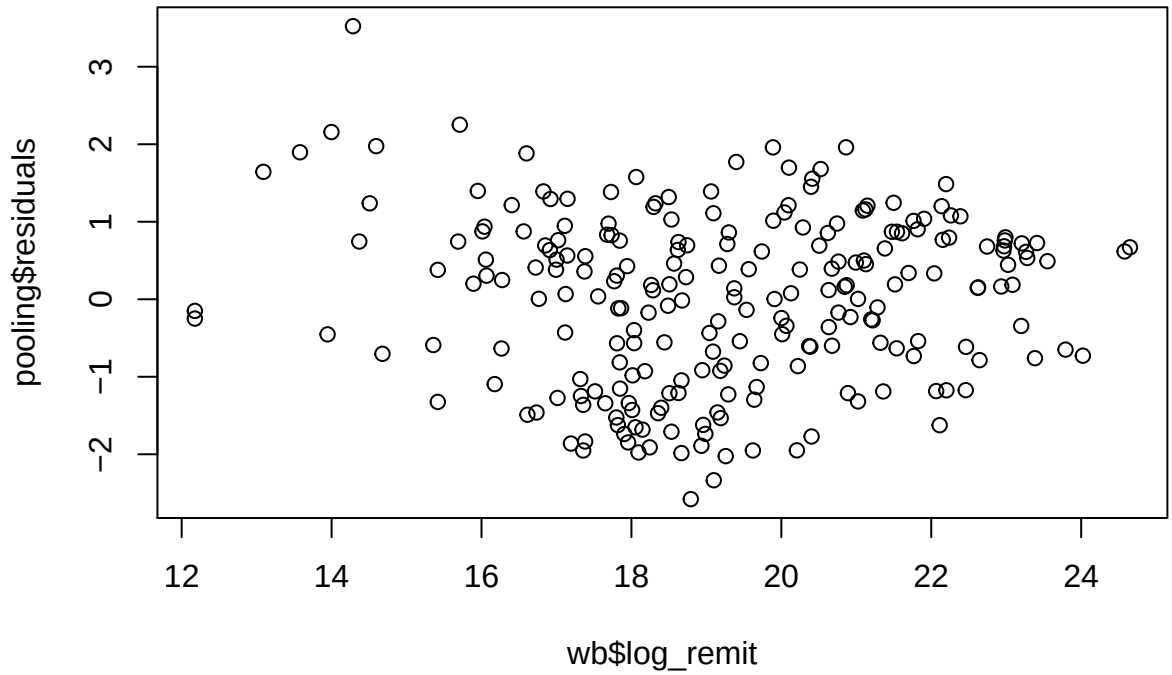

```
wb.plm <- pdata.frame(wb, index = c("country", "year"),
                      drop.index = TRUE)
```

```
## This series is constant and has been removed: keep
```

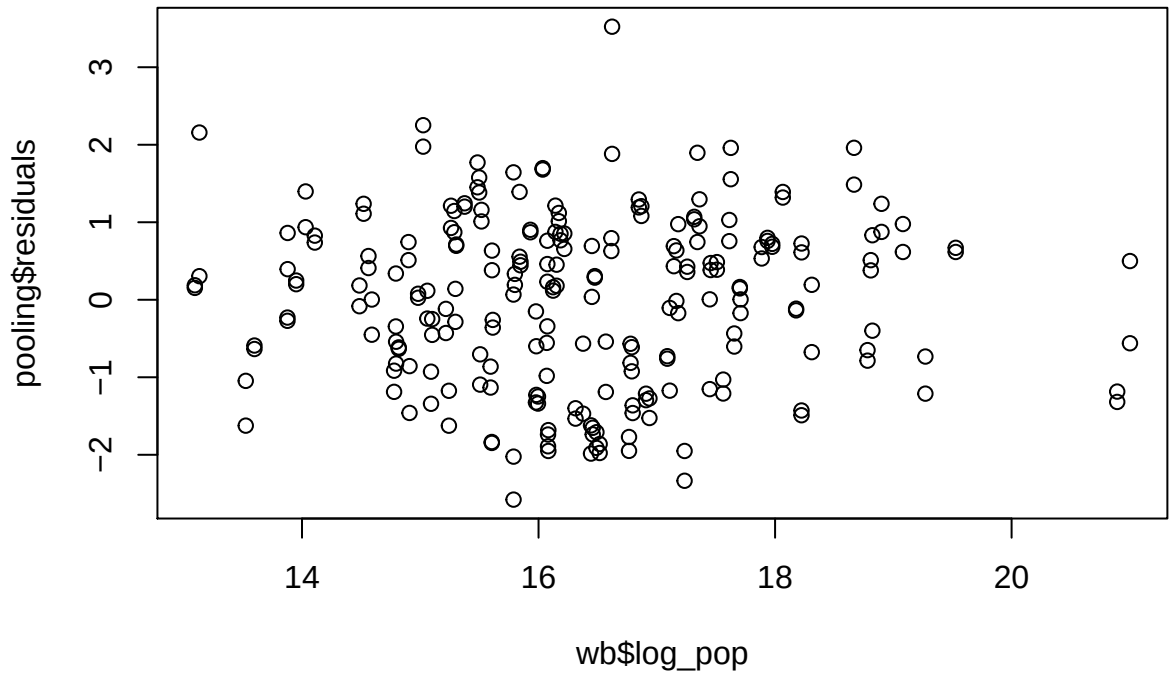
```
pooling <- plm(log_gdp ~ log_remit + log_pop, model = "pooling", data = wb.plm)
summary(pooling)
```

```
## Pooling Model
##
## Call:
## plm(formula = log_gdp ~ log_remit + log_pop, data = wb.plm, model = "pooling")
##
## Balanced Panel: n=116, T=2, N=232
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
##  -2.580  -0.832   0.163   0.774   3.520
##
## Coefficients :
##              Estimate Std. Error t-value  Pr(>|t|)
## (Intercept)  4.107287   0.877070  4.6830 4.846e-06 ***
## log_remit    0.491001   0.030476 16.1111 < 2.2e-16 ***
## log_pop     -0.312064   0.048958 -6.3741 1.002e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    589.74
## Residual Sum of Squares: 272.47
## R-Squared:      0.53798
## Adj. R-Squared: 0.53394
## F-statistic: 133.324 on 2 and 229 DF, p-value: < 2.22e-16
```

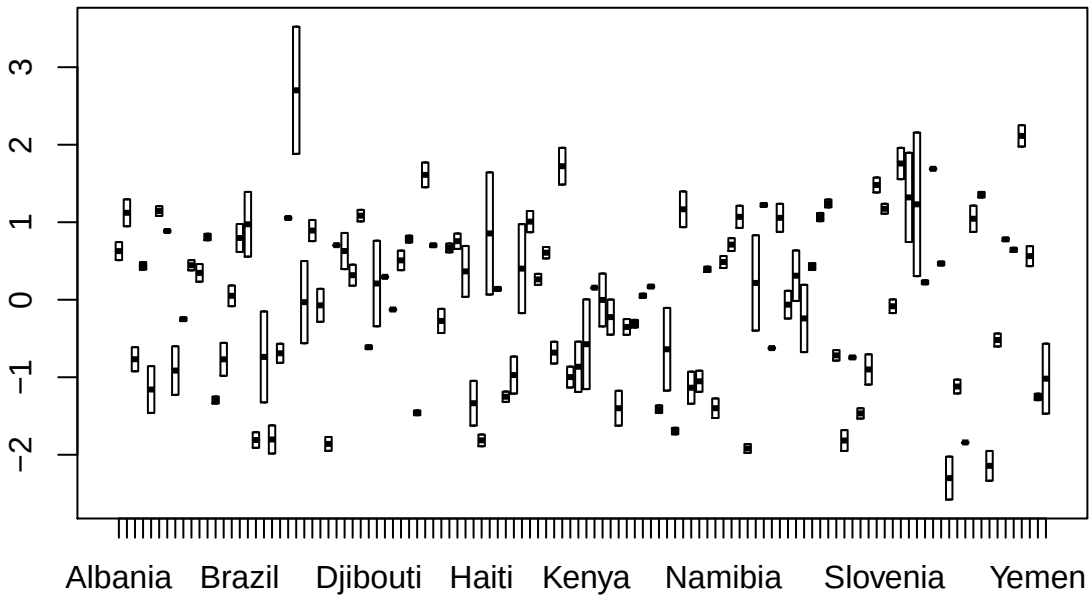
```
plot(pooling$residuals ~ wb$log_remit)
```




```
plot(pooling$residuals ~ wb$log_pop)
```



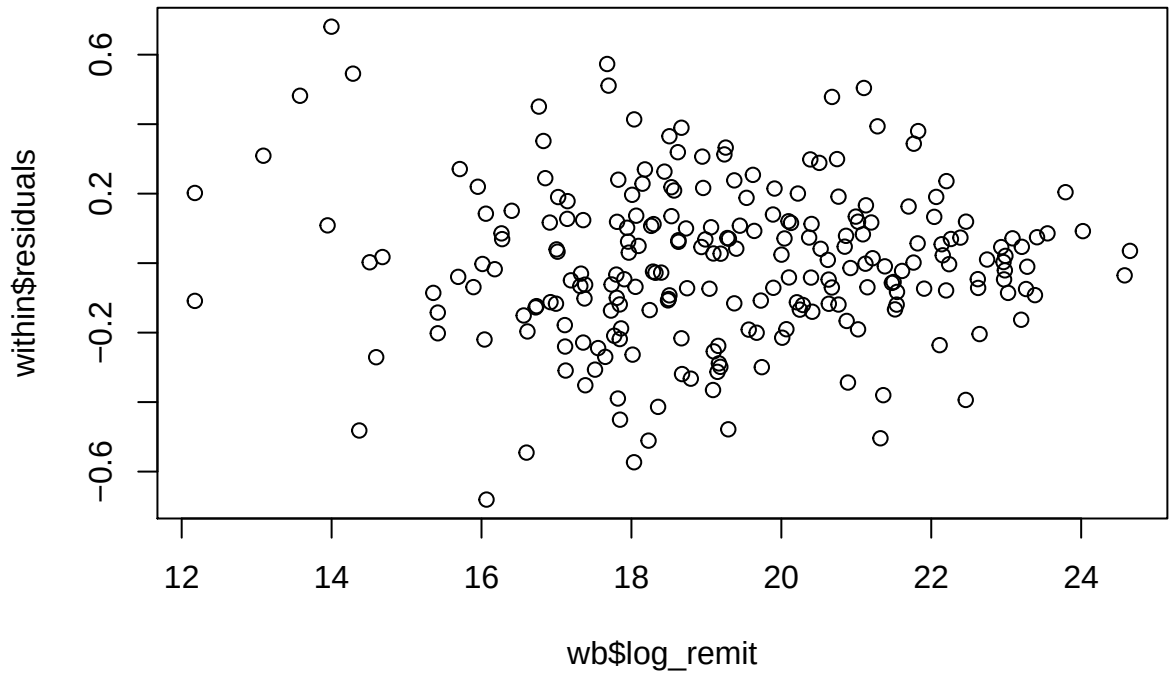

```
boxplot(pooling$residuals ~ wb$country)
```



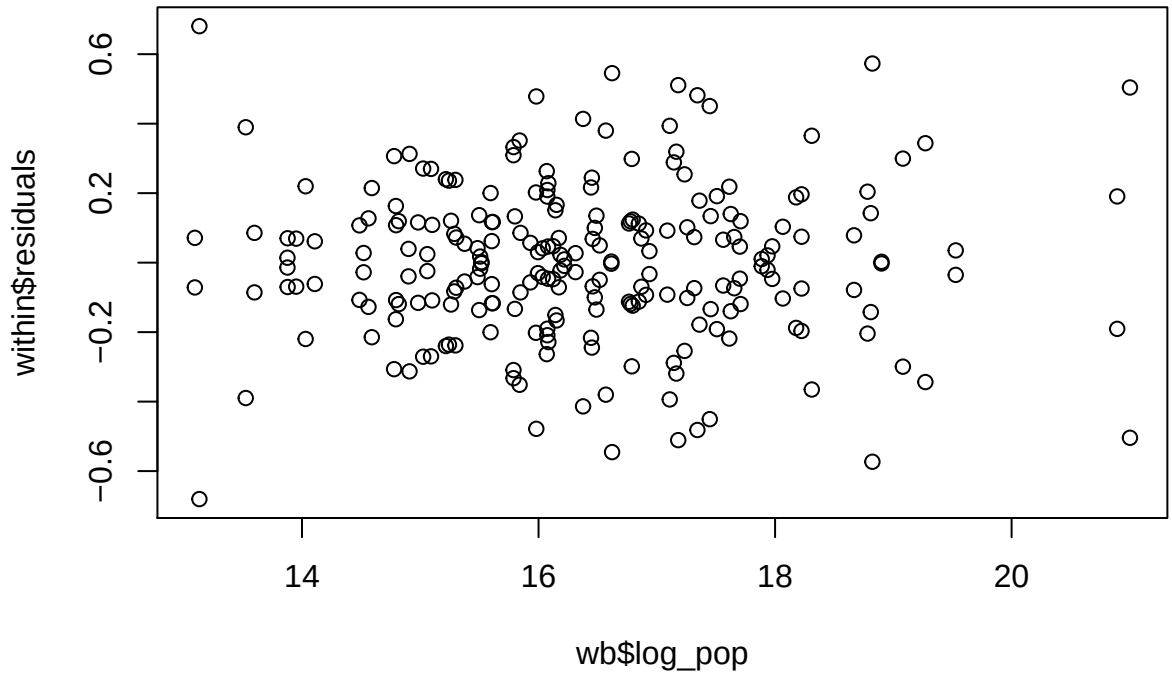

```
within <- plm(log_gdp ~ log_remit + log_pop, model = "within", data = wb.plm)
summary(within)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_gdp ~ log_remit + log_pop, data = wb.plm, model = "within")
##
## Balanced Panel: n=116, T=2, N=232
##
## Residuals :
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -6.81e-01 -1.20e-01  8.88e-16  1.20e-01  6.81e-01
##
## Coefficients :
##           Estimate Std. Error t-value  Pr(>|t|)
## log_remit 0.253308   0.040868  6.1982 9.187e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    15.228
## Residual Sum of Squares: 11.414
## R-Squared:      0.25041
## Adj. R-Squared: -0.5057
## F-statistic: 38.4175 on 1 and 115 DF, p-value: 9.1865e-09
```

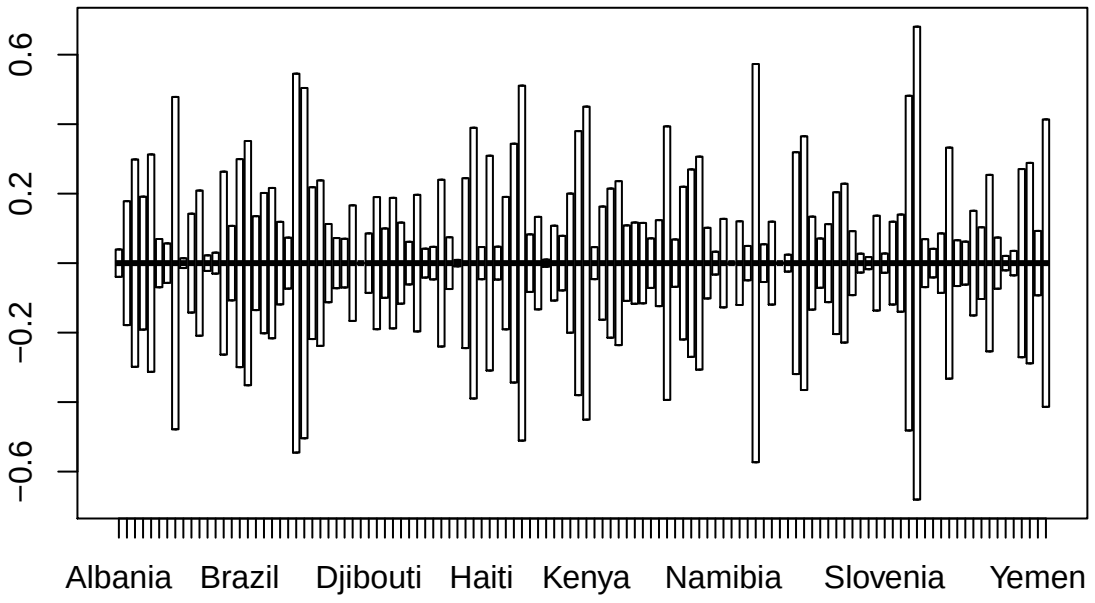
```
plot(within$residuals ~ wb$log_remit)
```



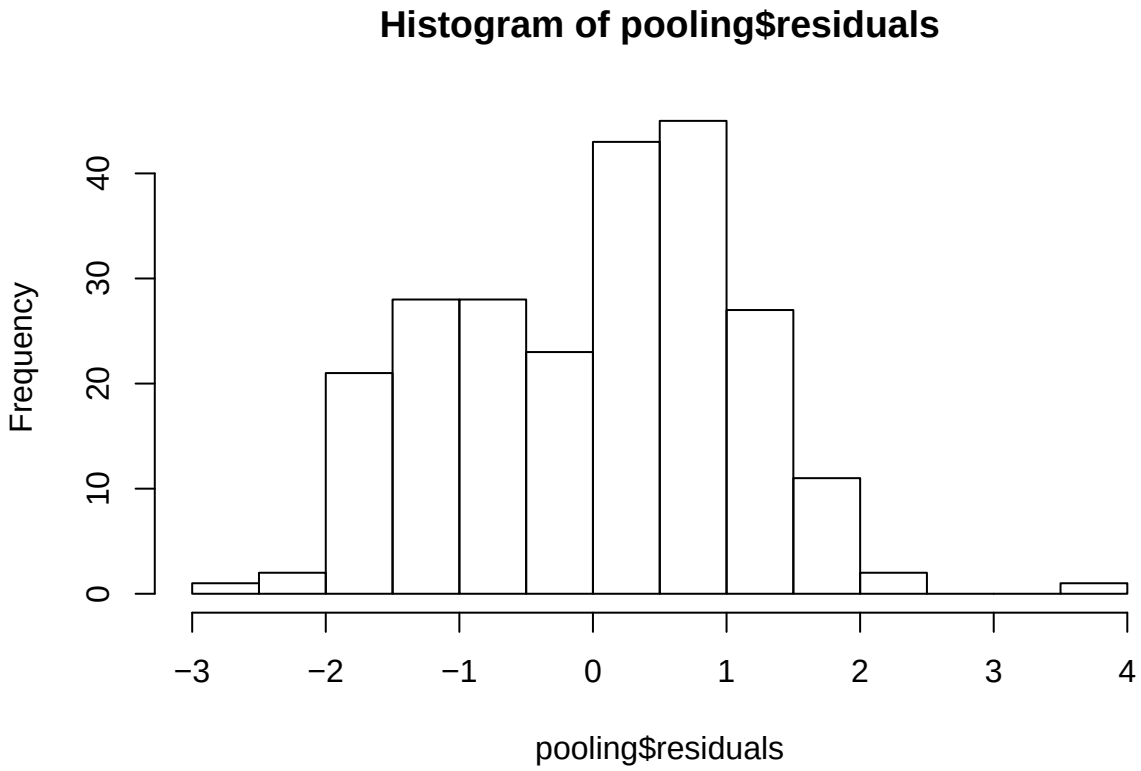

```
plot(within$residuals ~ wb$log_pop)
```



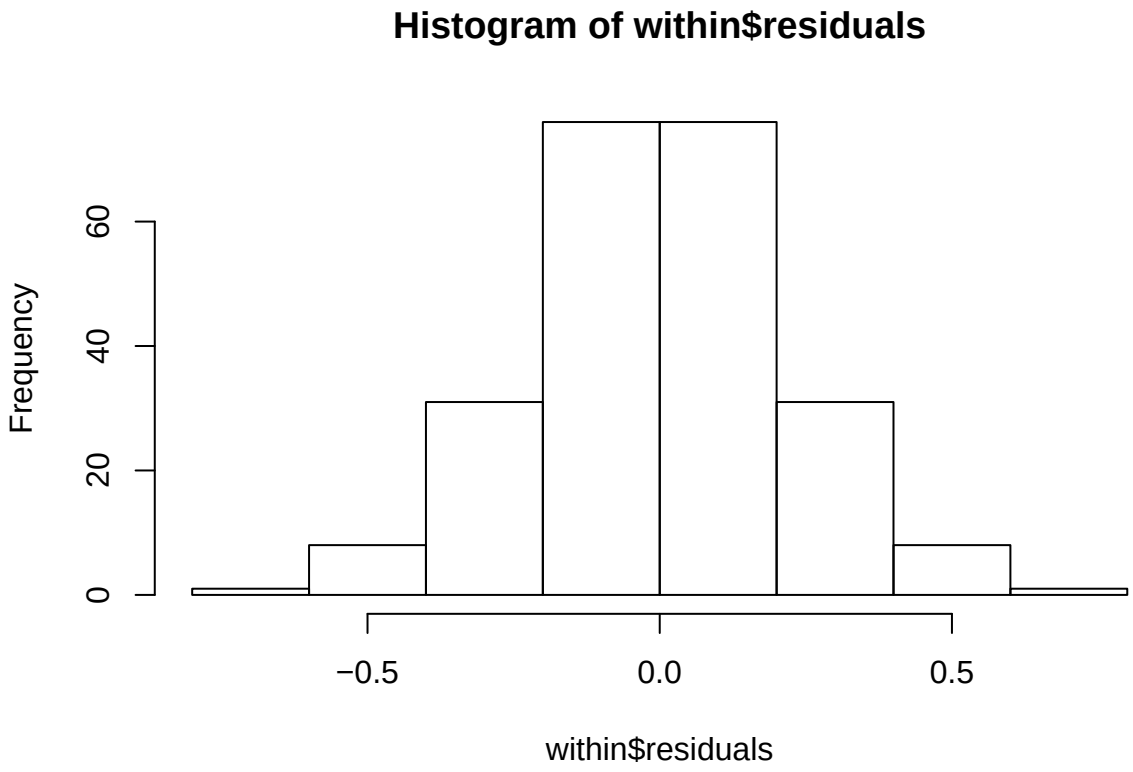

```
boxplot(within$residuals ~ wb$country)
```




```
hist(pooling$residuals)
```




```
hist(within$residuals)
```




```
random <- plm(log_gdp ~ log_remit + log_pop, model = "random", data = wb.plm)



summary(pooling) #coef = 0.49, se = 0.03
```

```
## Pooling Model
##
## Call:
## plm(formula = log_gdp ~ log_remit + log_pop, data = wb.plm, model = "pooling")
##
## Balanced Panel: n=116, T=2, N=232
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
##  -2.580  -0.832   0.163   0.774   3.520
##
## Coefficients :
##              Estimate Std. Error t-value  Pr(>|t|)
## (Intercept)  4.107287   0.877070  4.6830 4.846e-06 ***
## log_remit    0.491001   0.030476 16.1111 < 2.2e-16 ***
## log_pop     -0.312064   0.048958 -6.3741 1.002e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    589.74
## Residual Sum of Squares: 272.47
## R-Squared:      0.53798
## Adj. R-Squared: 0.53394
## F-statistic: 133.324 on 2 and 229 DF, p-value: < 2.22e-16
```

```r
summary(random) #      0.371592   0.030745
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = log_gdp ~ log_remit + log_pop, data = wb.plm, model = "random")
##
## Balanced Panel: n=116, T=2, N=232
##
## Effects:
##                  var std.dev share
## idiosyncratic 0.1001  0.3164 0.084
## individual    1.0895  1.0438 0.916
## theta:  0.7904
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -0.7860 -0.2480  0.0339  0.2290  1.1500
##
## Coefficients :
##              Estimate Std. Error t-value  Pr(>|t|)
## (Intercept)  5.662861   1.187965  4.7669 3.327e-06 ***
## log_remit    0.371592   0.030745 12.0864 < 2.2e-16 ***
## log_pop     -0.266966   0.069035 -3.8671 0.0001435 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    40.467
## Residual Sum of Squares: 24.481
## R-Squared:      0.39505
## Adj. R-Squared: 0.38977
## F-statistic: 74.7719 on 2 and 229 DF, p-value: < 2.22e-16
```

```r
summary(within) #     0.253308   0.040868
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_gdp ~ log_remit + log_pop, data = wb.plm, model = "within")
##
## Balanced Panel: n=116, T=2, N=232
##
## Residuals :
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -6.81e-01 -1.20e-01  8.88e-16  1.20e-01  6.81e-01
##
## Coefficients :
##           Estimate Std. Error t-value  Pr(>|t|)
## log_remit 0.253308   0.040868  6.1982 9.187e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    15.228
```

```
## Residual Sum of Squares: 11.414
## R-Squared:      0.25041
## Adj. R-Squared: -0.5057
## F-statistic: 38.4175 on 1 and 115 DF, p-value: 9.1865e-09
```

```
# Change in coefficients tell us that FX is required
# Note that SE is largest for within estimator
```