Applied Econometrics Homework
M2 FE

Khalil Janbek, Romain Jouhameau

01/01/2022

```
library(data.table)
library(dtplyr)
library(dplyr, warn.conflicts = FALSE)
library(readxl)
library(janitor)
library(modelsummary)
library(kableExtra)
library(ggrepel) # for spacing text inside plots
library(tidyverse)
library(sf)
library(usmap)
library(tsibble)
library(DBI)
```

```
con <- dbConnect(RSQLite::SQLite(), "Data/DB.sqlite")
DB = tbl(con, "DB")

DB %>%
  head() %>%
  collect()
```

```
#> # A tibble: 6 x 43
#>   Loan_Seq_Number Reporting_Period Current_UPB Delinquency_Status Loan_Age
#>   <chr>                      <dbl>       <dbl> <chr>                 <int>
#> 1 F20Q10000703               18293      342000 0                         0
#> 2 F20Q10000703               18322      342000 0                         1
#> 3 F20Q10000703               18353      341000 0                         2
#> 4 F20Q10000703               18383      341000 0                         3
#> 5 F20Q10000703               18414      340000 0                         4
#> 6 F20Q10000703               18444      340000 0                         5
#> # ... with 38 more variables: Time_to_Maturity <int>, Zero_Balance_Code <int>,
#> #   Current_Interest_Rate <dbl>, E_LoanToValue <int>,
#> #   Delinquency_Due_To_Disaster <chr>, Borrower_Assistance_Status_Code <chr>,
#> #   Interest_Bearing_UPB <dbl>, Credit_score <int>, First_Payment_Date <int>,
#> #   FstTime_HB_Flag <chr>, Maturity_Date <int>, MSA <int>,
#> #   Mortgage_Insurance_pct <int>, Number_of_Units <int>,
#> #   Occupancy_Status <chr>, O_CombinedLoanToValue <int>, ...
```

PART A:

1. In your data set, which are the variables which are varying with respect to two indices (or more if you consider inflows and outflows from one individual or country to another individual or countries? Which are the variables which are varying only with respect to time? Which are the variables which are varying only with respect to individuals?

```
DB %>%
  count(Loan_Seq_Number) %>%
  summarise(Number_Loans = n(),
            Max_T = max(n, na.rm = TRUE),
            Min_T = min(n, na.rm = TRUE),
            Mean_T = mean(n, na.rm = TRUE),
            Median_T = median(n, na.rm = TRUE),
            ) %>%
  collect()
```

2. What is the largest number of period T for individuals? What is the number of individuals?

```
#> # A tibble: 1 x 5
#>   Number_Loans Max_T Min_T Mean_T Median_T
#>          <int> <int> <int>  <dbl>    <int>
#> 1       346724    17     1   14.0       15
```

```
DB %>%
  mutate(Type = case_when(
           Lender_Type == 1 ~ 'Bank',
           Lender_Type == 0 & Fintech == 0 ~ 'Shadow Bank',
           Lender_Type == 0 & Fintech == 1 ~ 'Fintech'
         )) %>%
  group_by(Type) %>%
  count(Loan_Seq_Number) %>%
  summarise(Number_Loans = n(),
                   Max_T = max(n, na.rm = TRUE),
                   Min_T = min(n, na.rm = TRUE),
                   Mean_T = mean(n, na.rm = TRUE),
                   Median_T = median(n, na.rm = TRUE)) %>%
  collect()
```

```
#> # A tibble: 3 x 6
#>   Type        Number_Loans Max_T Min_T Mean_T Median_T
#>   <chr>              <int> <int> <int>  <dbl>    <int>
#> 1 Bank              147506    17     1   14.3       15
#> 2 Fintech            52273    17     1   13.7       15
#> 3 Shadow Bank       146945    17     1   13.9       15
```

3. Comment on the structure of the unbalanced panel (how many (and which) countries have a single observation, discontinuities between observations, how many have at least 2 consecutive observations (which is useful to compute lags, autocorrelations, first difference and within estimators)?

4. Compute between transformed and within transformed variables for all variables. Present a table with the within, between and pooled variance for each variable. Compute the share of between and within variance in the total variance for each variable. Comment these results.

5. Plot the distribution of the within and between transformed dependent variable and of you key (preferred) explanatory variable (not all the explanatory variable) [in Burnside and Dollar: GDP growth and foreign aid EDA/GDP], using on the same graph an histogram, a normal law with same empirical mean and standard error and a kernel continuous approximation. Comment the between and within difference for each variable, and compare within/within for dependent and explanatory variable, and between/between for dependent and explanatory variable: kurtosis, skewness, non-normality, high leverage observation (far from the mean), several modes (mixture of distribution)?

6. Plot boxplot of within distribution and between distribution for the dependent variable and the key explanatory variables. Comment that you find the same insights from question 5.

7. Compute univariate descriptive statistics (min, Q1, median, Q3, max, mean, standard error) for Within and Between transformed variables. Is the mean different from the median and why? How many standard errors from the mean are the min and max extremes (report (MAX-average)/standard error and (MIN-average)/standard error in the tables)?

8. Plot the boxplot of within transformed dependent variable and the key explanatory variable by a few individual (all of them if N around 50) and only the first 20 of them for larger data set. Comment on their differences of standard errors and means for each individuals

9. Compare and comment the within and between transformed bivariate correlation matrix for all variables (include a time trend 1,2,.,T). Check poor simple correlation with the dependent variables and high correlation between explanatory variables.

10. Comment the bivariate auto-correlation and trend-correlations (check the number of observations).

11. Comment the bivariate graphs with linear, quadratic and Lowess fit for dependent and key explanatory variable (aid/gdp and growth of gdp): Within transformed, Between transformed.

12. Comment the results of estimations of Between, Within (fixed effects, (fe)) and Mundlak (random effects (re) including all X(i.) as regressors), two-way fixed effects (add year dummies in fe regression) and First differences, including all explanatory variables except the ones with high near-multicollinearity in their respective between or within space.

13. If one of your variable is time-invariant z(i) (Institutional quality ICRG for Burnside Dollar), run a baseline Hausman Taylor estimation including all X(i.) as instruments. Comment the results.

14. If one of your variable is time-invariant z(i) (Institutional quality ICRG for Burnside Dollar), run a between regression on z(i) explained by X(i.) and other time invariant variable (only with N observations). If the R2 is low, this may signal X(i.) are weak instruments poorly correlated with the variable z(i) to be instrumented. Comment.

15. Optional: mention or propose improvements to the Python, STATA, SAS or R code (copy it here). Optional: propose improvements, additional insights, and you do not know how to code them.

PART B (update results)

```
# Import Data
# Inside the Data folder, get all the .RDS files except MSA_Large
panel_data <- list.files(path =  'Data/', pattern="*[^(MSA_Large)].RDS") %>%
  map(., ~read_rds(paste0('Data/', .))) %>%
  reduce(inner_join, by = c('iso2c', 'country', 'year'))

panel_data
```

1. Download 5 panel data variables from World Bank and/or IMF and/or FRED databases for the recent period (1990-2020) and for the largest coverage of emerging economies: GDP/head, GDP/head PPP-adjusted (very last update), Log(population), Foreign aid/GDP (ODA), of log an index of corruption (or good public sector governance) from the World Bank. From now on, consider as your sample only country-year observations which are available for ALL the 5 variables for at least TWO CONSECUTIVE years for a given country. The full class may coordinate for this updated database. In all the following questions except perhaps the last one, the PPP adjusted GDP is not used. So we consider 4 variables excluding GDP/head PPP adjusted.

```
#> # A tibble: 1,848 x 9
#>    iso2c country corruption  year gdp_ppp gdp_per_cap oda_gni oda_net population
#>    <chr> <chr>        <dbl> <int>   <dbl>       <dbl>   <dbl>   <dbl>      <dbl>
#>  1 AL    Albania     -0.533  2019  13657.       4549.   0.187    9.95    2854191
#>  2 AL    Albania     -0.479  2018  13317.       4434.   2.27   120.      2866376
#>  3 AL    Albania     -0.421  2017  12771.       4250.   1.29    58.6     2873457
#>  4 AL    Albania     -0.405  2016  12292.       4090.   1.42    59.5     2876101
#>  5 AL    Albania     -0.479  2015  11878.       3953.   2.91   116.      2880703
#>  6 AL    Albania     -0.548  2014  11587.       3856.   2.11    97.3     2889104
#>  7 AL    Albania     -0.698  2013  11361.       3781.   2.08    93.3     2895092
#>  8 AL    Albania     -0.726  2012  11228.       3736.   2.86   120.      2900401
#>  9 AL    Albania     -0.683  2011  11053.       3678.   2.94   131.      2905195
#> 10 AL    Albania     -0.525  2010  10749.       3577.   3.09   125.      2913021
#> # ... with 1,838 more rows
```

```
panel_data %>%
  count(year)
```

```
#> # A tibble: 21 x 2
#>     year     n
#>    <int> <int>
#>  1  1996    88
#>  2  1998    88
#>  3  2000    88
#>  4  2002    88
#>  5  2003    88
#>  6  2004    88
#>  7  2005    88
#>  8  2006    88
#>  9  2007    88
#> 10  2008    88
#> # ... with 11 more rows
```

We don't have any observations for the Year 1997, 1999 and 2001. Thus, we will restrict ourselves to the time period between 2002 to 2019 for 88 countries.

```
panel_data <- panel_data %>%
  filter(!year %in% c(1996, 1998, 2000))
```

```
panel_data %>%
  count(country)
```

```
#> # A tibble: 88 x 2
#>    country                 n
#>    <chr>               <int>
#>  1 Albania                18
#>  2 Algeria                18
#>  3 Antigua and Barbuda    18
#>  4 Argentina              18
#>  5 Bangladesh             18
#>  6 Belize                 18
#>  7 Benin                  18
#>  8 Bhutan                 18
#>  9 Bolivia                18
#> 10 Botswana               18
#> # ... with 78 more rows
```

```
# We remove them because they have negative ODA for at least one year
# We would need to find Gross ODA in order to have only positives values
country_to_remove <- panel_data %>%
  group_by(country) %>%
  filter(oda_net < 0) %>%
  distinct(country) %>%
  pull()

country_to_remove
```

```
#>  [1] "Argentina"   "China"       "Gabon"       "Indonesia"   "Malaysia"
#>  [6] "Mauritius"   "Panama"      "Peru"        "Philippines" "Sri Lanka"
#> [11] "Thailand"
```

```
panel_data <- panel_data %>%
  filter(!country %in% country_to_remove)

panel_data %>%
  count(country)
```

```
#> # A tibble: 77 x 2
#>    country                 n
#>    <chr>               <int>
#>  1 Albania                18
#>  2 Algeria                18
#>  3 Antigua and Barbuda    18
#>  4 Bangladesh             18
#>  5 Belize                 18
#>  6 Benin                  18
#>  7 Bhutan                 18
#>  8 Bolivia                18
#>  9 Botswana               18
#> 10 Brazil                 18
#> # ... with 67 more rows
```

```
panel_data <- panel_data %>%
  arrange(country, year) %>%
  group_by(country) %>%
  mutate(g_gdp_per_cap = log(gdp_per_cap) - lag(log(gdp_per_cap)),
         g_population = log(population) - lag(log(population)),
         g_oda_net = log(oda_net) - lag(log(oda_net))) %>%
  ungroup()
```

2. Compute 2 growth rates using the difference of log: the growth of GDP/head (difference of log, denoted GDPg), the growth of foreign aid ODAg (but NOT the growth for foreign aid/GDP: remove the difference of log of GDP from the difference of log of foreign aid/GDP).

```
panel_data %>%
  group_by(country) %>%
  mutate(oda_net_gdp_cap = oda_net / gdp_per_cap) %>%
  # Summarise is used  to transform our dataframe and calculate the mean for each country
  summarise(across(where(is.double), ~mean(., na.rm = T))) %>% # across apply a function (here the mean) g
  arrange(desc(oda_net_gdp_cap)) %>%
  relocate(country, oda_net_gdp_cap) %>%
  slice_max(oda_net_gdp_cap, n = 10)
```

3. Compute the between average over time for the first period and for the second period for the 6 variables. Provide the top 10 of countries for ODA/GDP with average over time for each period.

```
#> # A tibble: 10 x 11
#>    country        oda_net_gdp_cap corruption gdp_ppp gdp_per_cap oda_gni oda_net
#>    <chr>                    <dbl>      <dbl>   <dbl>       <dbl>   <dbl>   <dbl>
#>  1 Solomon Islan~          0.182     -0.291    2382.       2050.   24.9    377.
#>  2 Burundi                 0.174     -1.18      821.        304.   26.5     53.0
#>  3 Malawi                  0.174     -0.640    1316.        344.   16.0     60.1
#>  4 Central Afric~          0.164     -1.15     1025.        454.   15.9     69.8
#>  5 Sierra Leone            0.139     -0.837    1483.        565.   18.8     77.8
#>  6 Rwanda                  0.139      0.208    1555.        618.   15.9     85.2
#>  7 Vanuatu                 0.130      0.0793   2981.       2755.   14.1    363.
#>  8 Cabo Verde              0.126      0.792    5946.       2887.   12.7    359.
#>  9 Guinea-Bissau           0.126     -1.29     1761.        590.   13.7     73.9
#> 10 Burkina Faso            0.0986    -0.255    1753.        595.    9.95    58.0
#> # ... with 4 more variables: population <dbl>, g_gdp_per_cap <dbl>,
#> #   g_population <dbl>, g_oda_net <dbl>
```

```
panel_data %>%
  group_by(country) %>%
  mutate(oda_net_gdp_cap = oda_net / gdp_per_cap) %>%
  filter(oda_net_gdp_cap >= 0,
         oda_net_gdp_cap < 0.05) %>%
  count(country) %>%
  mutate(prop = round(n / 18, 2)) %>%
  arrange(desc(n))
```

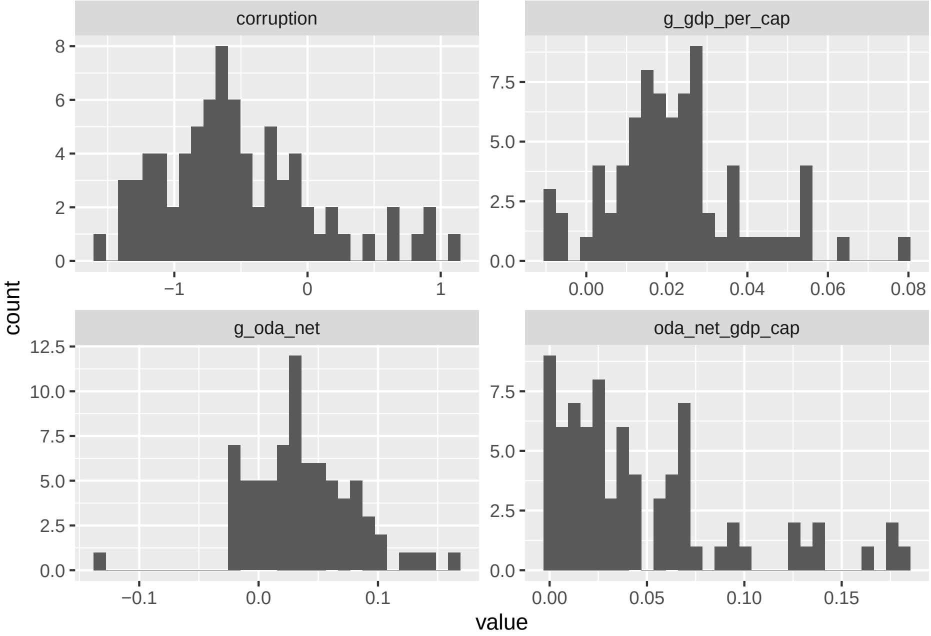
4. Compute the proportion of country-years observations in your database such that 0<=ODA/GDP<0.5%

```
#> # A tibble: 65 x 3
#> # Groups:   country [65]
#>    country                 n  prop
#>    <chr>               <int> <dbl>
#>  1 Albania                18     1
#>  2 Algeria                18     1
#>  3 Antigua and Barbuda    18     1
#>  4 Bangladesh             18     1
#>  5 Belize                 18     1
#>  6 Bolivia                18     1
#>  7 Brazil                 18     1
#>  8 Colombia               18     1
#>  9 Costa Rica             18     1
#> 10 Dominican Republic     18     1
#> # ... with 55 more rows
```

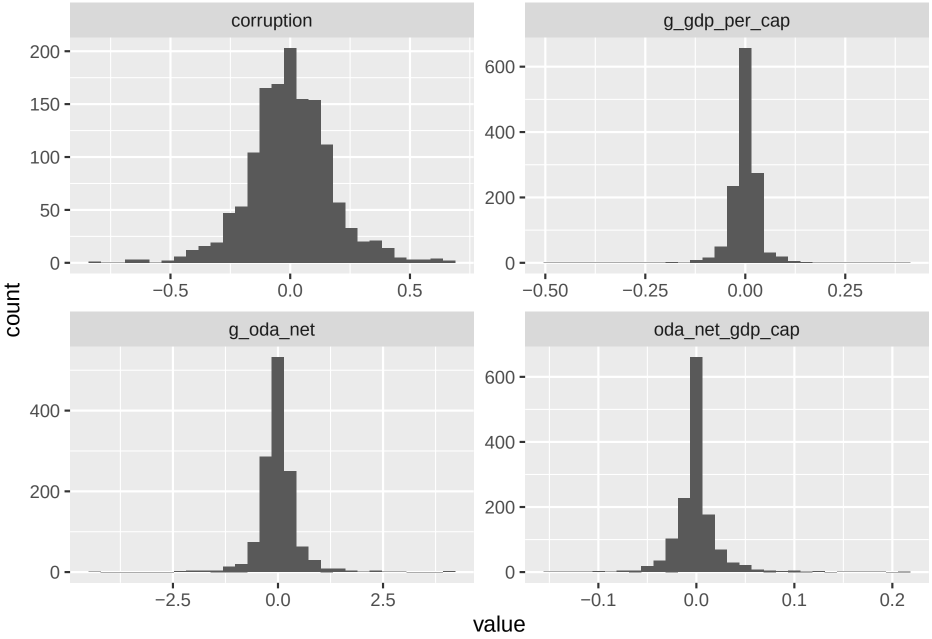
```
# Between transformation
between_transformation <- panel_data %>%
  group_by(country) %>%
  mutate(oda_net_gdp_cap = oda_net / gdp_per_cap) %>%
  summarise(across(where(is.double), mean, na.rm = T))
```

```
between_transformation %>%
  select(country, oda_net_gdp_cap, g_oda_net, g_gdp_per_cap, corruption) %>%
  pivot_longer(-country) %>%
  ggplot(aes(x = value)) +
  geom_histogram() +
  facet_wrap(~name, scales = 'free')
```

5. Compute the between and within transformations of the 6 variables over the full period. Provide the 4 histograms for ODA/GPD, growth of ODA, growth of GDP/head, corruption index for both between and within transformed variables (hence 8 histograms). Comment.

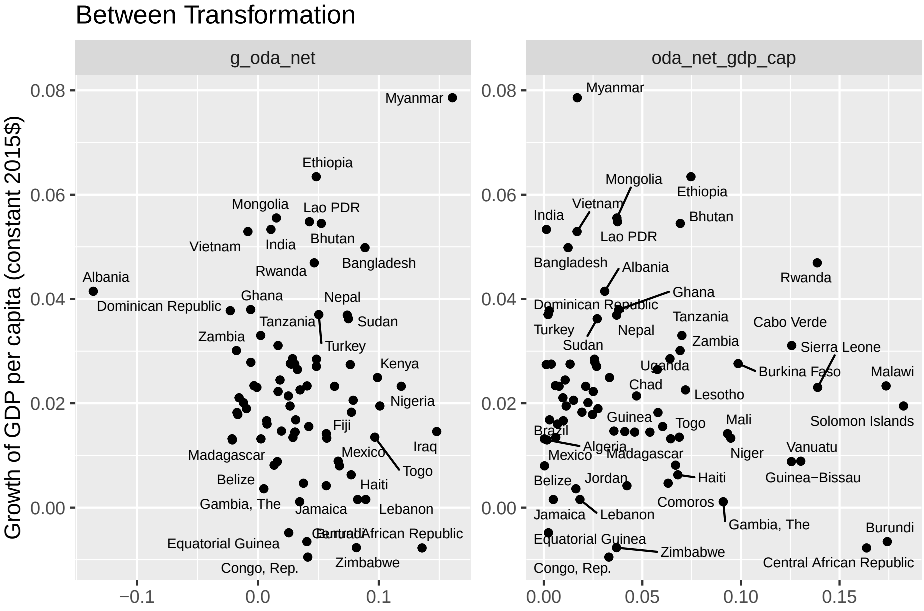


```
# Pas sur pour celui la
# Within transformation
panel_data %>%
  group_by(country) %>%
  mutate(oda_net_gdp_cap = oda_net / gdp_per_cap) %>%
  mutate(across(where(is.double),~ . - mean(., na.rm = T))) %>%
  select(country, oda_net_gdp_cap, g_oda_net, g_gdp_per_cap, corruption) %>%
  pivot_longer(-country) %>%
  ggplot(aes(x = value)) +
  geom_histogram() +
  facet_wrap(~name, scales = 'free')
```

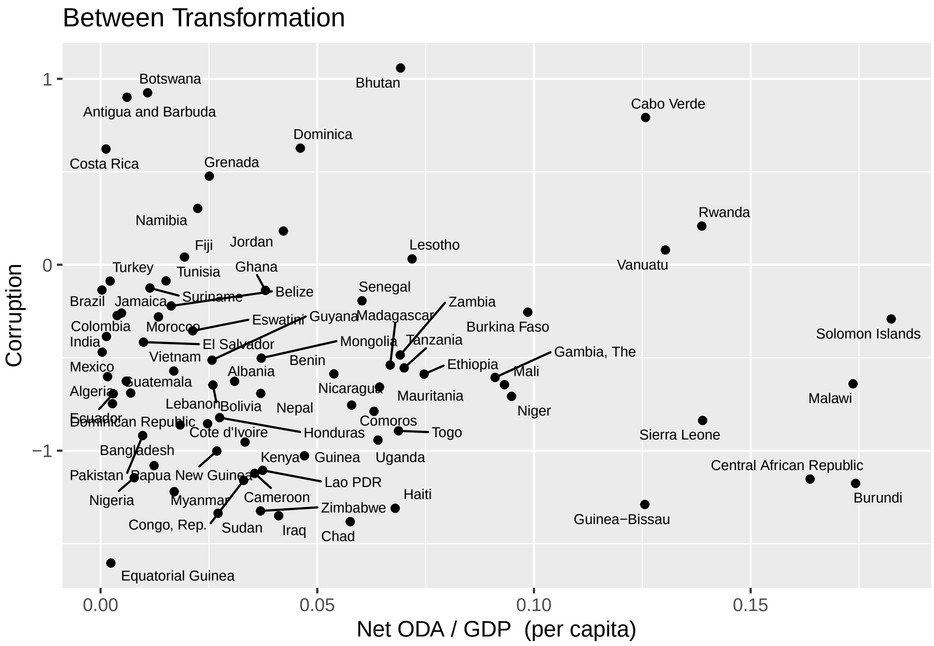

corruption
g_gdp_per_cap
g_oda_net
oda_net_gdp_cap
count
value
200
150
100
50
0
−0.5
0.0
0.5
600
400
200
0
−0.50
−0.25
0.00
0.25
400
200
0
−2.5
0.0
2.5
600
400
200
0
−0.1
0.0
0.1
0.2

```
between_transformation %>%
  pivot_longer(-c(country, g_gdp_per_cap)) %>%
  filter(name %in% c('g_oda_net', 'oda_net_gdp_cap')) %>%
  ggplot(aes(x = value, y = g_gdp_per_cap, label = country)) +
  geom_point() +
  facet_wrap(~name, scales = 'free') +
   geom_text_repel(size = 2.5) +
  labs(title = 'Between Transformation',
       x = '',
       y = 'Growth of GDP per capita (constant 2015$)')
```

6. Provide the 3 bivariate graphs (with acronyms for observations NIC12, for Nicaragua 2012) for between and within (hence 6 graphs) of growth of GDP/head (vertical axis) with (1) ODA/GDP, (2) the growth of ODA; of corruption index with ODA/GDP. Comment.



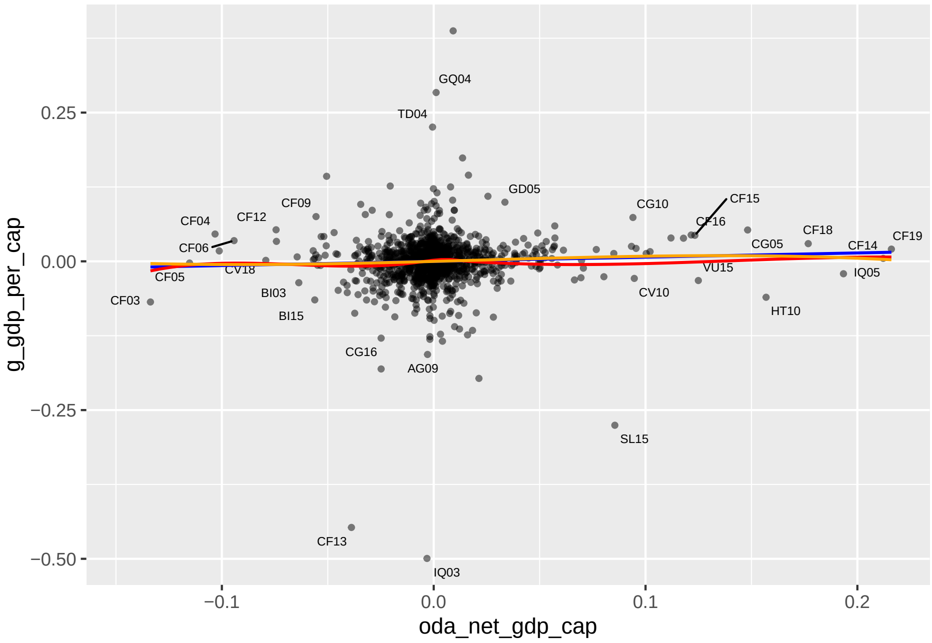
```
between_transformation %>%
  ggplot(aes(x = oda_net_gdp_cap, y = corruption, label = country)) +
  geom_point() +
  geom_text_repel(size = 2.5) +
  labs(title = 'Between Transformation',
       x = 'Net ODA / GDP  (per capita)',
       y = 'Corruption')
```



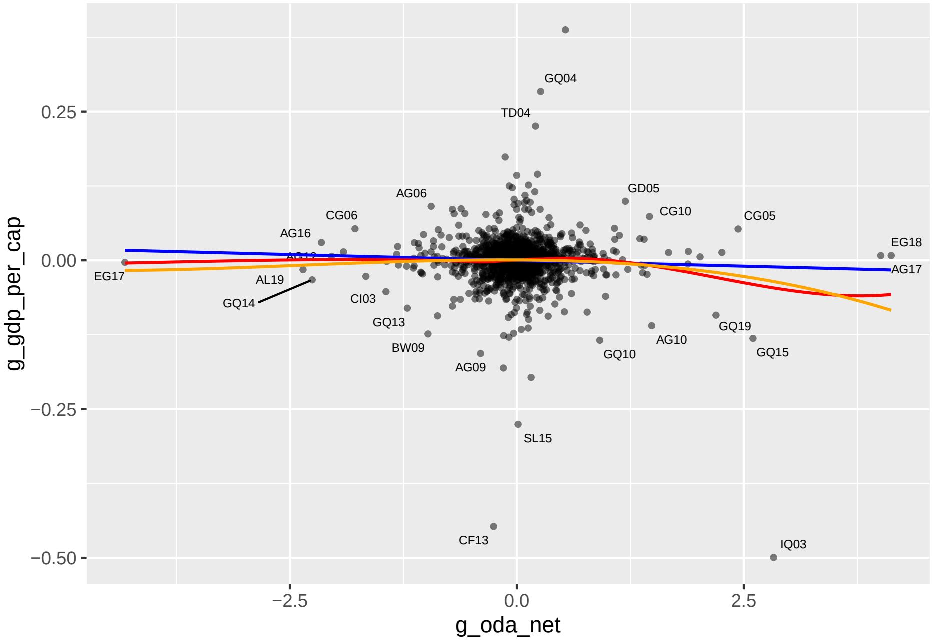
```
within_bivariate <- panel_data %>%
  group_by(country) %>%
  mutate(oda_net_gdp_cap = oda_net / gdp_per_cap) %>%
  mutate(across(where(is.double),~ . -  mean(., na.rm = T)),
         # '..$' : Regex pour sélectionner les deux (un point = n'importe quelle terme)
         # derniers ($) charactères du vecteur Year
         # '^..' : ^ pour selectionner les deux premiers charactères
         # '..$' : $ pour selectionner les deux derniers charactères
         iso_year = paste0(iso2c, str_extract_all(year, '..$'))) %>%
  ungroup() %>%
  select(iso_year, oda_net_gdp_cap, g_oda_net, g_gdp_per_cap, corruption)
```

```
plot_biv <- function(data, x, y) {

  data %>%
    ggplot(aes(x = {{ x }}, y = {{ y }}, label = iso_year)) +
    geom_point(size = 1, alpha = 0.5) +
    geom_text_repel(size=2) +
    geom_smooth(method = lm, se = FALSE, color = 'blue', size = 0.7) +
    geom_smooth(method = loess, se = FALSE, color = 'red', size = 0.7) +
    geom_smooth(method = lm, formula = y ~ splines::bs(x, 3), se = FALSE, color = 'orange', size = 0.7)

}
```

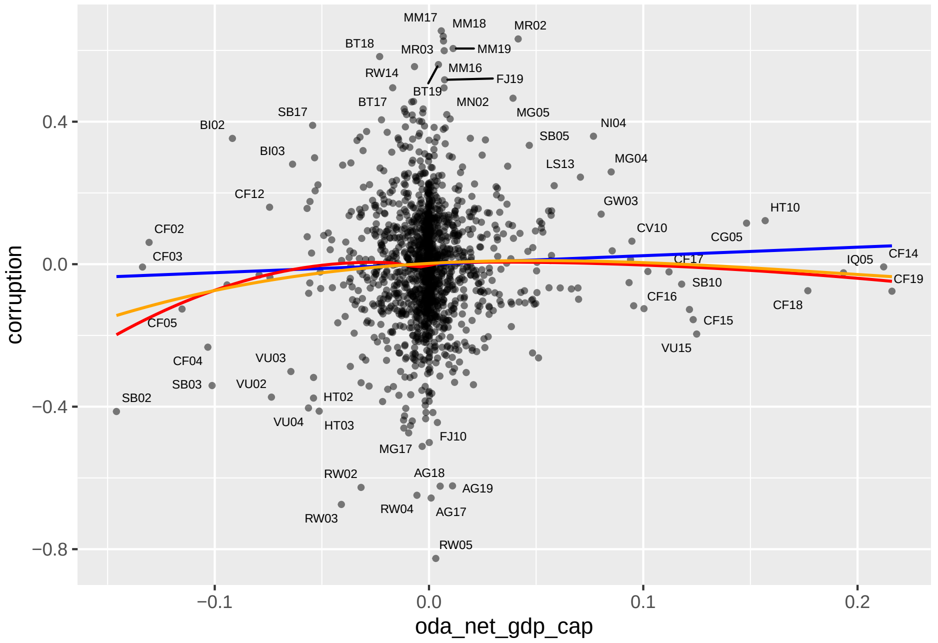
```
within_bivariate %>%
  plot_biv(x = oda_net_gdp_cap, y = g_gdp_per_cap)
```



```
within_bivariate %>%
  plot_biv(x = g_oda_net, y = g_gdp_per_cap)
```



```
within_bivariate %>%
  plot_biv(x = oda_net_gdp_cap, y = corruption)
```


MM17
MM18
MR02
BT18
MR03
MM19
RW14
MM16
FJ19
BT19
BT17
MN02
MG05
SB17
BI02
NI04
SB05
BI03
LS13
MG04
CF12
GW03
HT10
CF02
CV10
CG05
CF03
CF17
IQ05
CF14
SB10
CF19
CF16
CF18
CF05
CF15
VU15
CF04
VU03
SB03
VU02
SB02
HT02
VU04
HT03
FJ10
MG17
RW02
AG18
AG19
RW04
AG17
RW03
RW05
0.4
0.0
−0.4
−0.8
corruption
−0.1
0.0
0.1
0.2
oda_net_gdp_cap

```
between_transformation %>%
  select(c(g_oda_net, g_gdp_per_cap, gdp_per_cap, oda_net_gdp_cap, corruption, population)) %>%
  #cor()%>%
datasummary_correlation(title = 'Correlation matrix') %>%
kable_styling(latex_options = c("striped", "hold_position"))
```

Table 1: Correlation matrix

	g_oda_net	g_gdp_per_cap	gdp_per_cap	oda_net_gdp_cap	corruption	population
g_oda_net	1	.	.	.	.	.
g_gdp_per_cap	−0.07	1	.	.	.	.
gdp_per_cap	−0.04	−0.17	1	.	.	.
oda_net_gdp_cap	0.14	−0.15	−0.52	1	.	.
corruption	−0.16	0.14	0.43	−0.05	1	.
population	−0.04	0.26	−0.06	−0.20	−0.02	1

7. Comment the between versus within correlation matrix for the 6 variables in this order

8. Run a one-way fixed effect foreign aid regression on ODA/GDP function of Ln(Population) and Ln(GDP/head). Comment.

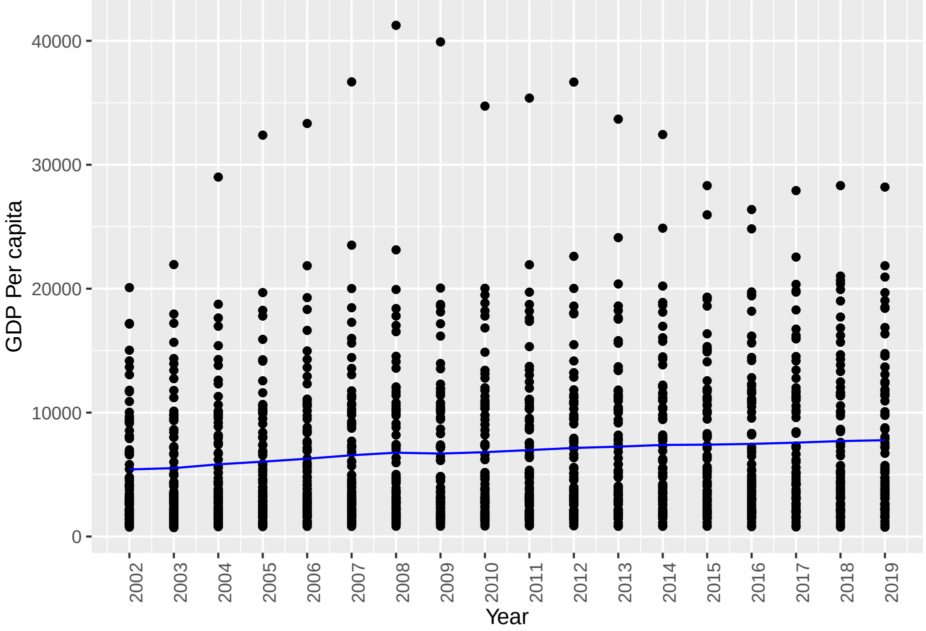
9. Run a one-way fixed effect of Corruption Index function of Ln(GDP/head), of ODA/GDP and the growth of ODA. Comment.

10. Run a one-way fixed effect with the growth of GDP/head function of Ln(GDP/head), ODA/GDP, the growth of ODA and the Corruption index.

11. Propose an additional interesting estimation using this database.

```
panel_data %>%
  ggplot(aes(x = year, y = gdp_ppp)) +
  geom_point() +
  geom_line(data = panel_data %>% group_by(year) %>% summarise(gdp_mean = mean(gdp_ppp)),
            aes(x = year, y = gdp_mean), col = "blue") +
  scale_x_continuous(labels = as.character(unique(panel_data$year)),
                     breaks = unique(panel_data$year)) +
  labs(x = "Year", y = "GDP Per capita") +
  theme(axis.text.x = element_text(angle = 90))
```

12. Compute the between and within transformations of the 11 variables over the full period. Provide histograms for ODA/GPD, growth of ODA, growth of GDP/head for both between and within transformed. Comment.


GDP Per capita
Year