Applied Econometrics Homework
M2 FE

Khalil Janbek, Romain Jouhameau

01/01/2022

PART A

Option 1 Master thesis dataset

Khalil's Master thesis explores the topic of the performance of securitized mortgage loans in the US in the context of the coronavirus crisis. Based on issuance and performance data for more than 346,724 mortgage loans purchased by Freddie Mac in 2020Q1, combined with neighborhood-level coronavirus data, the thesis will examine, at the US local level, whether the coronavirus outbreak affected these loans' (i) delinquency status and (ii) prepayment. In parallel, the Master thesis also enquires whether the status of the lender that originated the loan - commercial bank or shadow bank, Fintech or not - has an influence on those two loan performance metrics.

In those datasets, data are available at the geographical level of the Metropolitan Statistical Area (or MSA). MSAs are defined as "a core area containing a substantial population nucleus (50,000+ inhabitants), together with adjacent communities having a high degree of economic and social integration with that core". The US counts 384 MSA across the country.

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?

For each loan, the database records:

- (i) time-invariant issuance metrics. Those variables vary only with respect to individuals (in other words by loan) and do not vary over. Those time-invariant variables regroup for instance borrower credit score at issuance, the original loan to value, Original borrower debt-to-income and the lender type.
- (ii) time-varying performance metrics. Those variables vary both with respect to time and with respect to individuals. For instance, prepayment (Zero Balance Code), Delinquency status, and unpaid principal.

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

The dataset counts 17 time periods, spanning from a month in 2020Q1 to June 2021. The sample consists of 346,724 loans, originated at various dates but purchased by Freddie Mac in 2020Q1.

```
#> # 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
```

Lenders originating and selling these mortgage loans are classified as either commercial banks or shadow banks. The latter category also includes Fintech lenders.

```
#> # 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)?

As shown in the previous question, our dataset has an unbalanced panel data structure, with time periods per individual loan varying from 1 to 17.

Starting by exploring missing observations, the table shows that 6.75% of monthly loan performance observations relate to mortgage loans originated in locations that are not in an MSA - amounting to 22,351 loans, or 6.4% of the total of mortgages in our database. Those loans are likely to have been originated in sparsely populated areas.

```
#> # A tibble: 17 x 5
#>    Number_Periods_avai~ Number_of_loans percent cumulative_sum cumulative_perce~
#>                   <dbl>           <dbl>   <dbl>          <dbl>             <dbl>
#>  1                    1             123   0.035            123             0.035
#>  2                    2             467   0.135            590             0.17
#>  3                    3             963   0.278           1553             0.448
#>  4                    4            1655   0.477           3208             0.925
#>  5                    5            3058   0.882           6266             1.81
#>  6                    6            6189   1.78           12455             3.59
#>  7                    7           10496   3.03           22951             6.62
#>  8                    8           11092   3.20           34043             9.82
#>  9                    9           11599   3.34           45642            13.2
#> 10                   10           11310   3.26           56952            16.4
#> 11                   11           11439   3.30           68391            19.7
#> 12                   12           12144   3.50           80535            23.2
#> 13                   13           10956   3.16           91491            26.4
#> 14                   14           10829   3.12          102320            29.5
#> 15                   15          108907  31.4           211227            60.9
#> 16                   16           77004  22.2           288231            83.1
#> 17                   17           58493  16.9           346724           100
```

Monthly developments in COVID-19 confirmed cases by MSA are also missing in 8.45% of our monthly loan performance observations. This percentage includes the 6.75% of loan performance observations that are not located in an MSA. As our loan performance data start in February 2020, the remaining missing COVID-19 observations are related to monthly loan performance data that date prior to March 2020 - date at which COVID-19 cases started to be tracked.

```
#> # A tibble: 405 x 3
#>      MSA LoansPerMSA Percent
#>    <int>       <int>   <dbl>
#>  1    NA       22351    6.45
#>  2 38060       11947    3.45
#>  3 31084       11555    3.33
#>  4 16984       10082    2.91
#>  5 12060        8319    2.4
#>  6 40140        7655    2.21
#>  7 19740        7428    2.14
#>  8 47894        7176    2.07
#>  9 19124        6679    1.93
#> 10 26420        6170    1.78
#> # ... with 395 more rows
```

As briefly shown in question 2, we have between 1 and 17 consecutive observation for each loan contract. Nonetheless, the number of loans for which we have only 1 period of observation is very small compared to the total number of loans of the dataset - 123 out 346,724, or 0.3%. All the other loans therefore have at least 2 consecutive observations.

None of our loan performance data display discontinuities between observations.

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.

The pooled, within and between are defined as below:

- Pooled deviation = $x_{i,t} - \overline{x}$
- Within deviation = $x_{i,t} - \overline{x}_i$
- Between deviation = $\overline{x}_i - \overline{x}$

Then we take the variance for each variable and each transformation.

The below table decomposes the total (pooled) deviation into a Within and Between part.

variable	Pooled	Within	Between
Delinquency_Status	2.547150e-02	1.144690e-02	1.402550e-02
Zero_Balance_Code	2.254630e-02	2.057160e-02	1.974700e-03
confirmed	1.037464e+14	4.955898e+13	5.408003e+13
Credit_score	1.860216e+03	0.000000e+00	1.860216e+03
Current_Interest_Rate	2.290216e-01	2.960000e-05	2.289920e-01
Current_UPB	1.952840e+10	2.764186e+09	1.676421e+10
E_LoanToValue	2.879469e+02	3.328202e+01	2.587328e+02
O_DebtToIncome	9.073039e+01	0.000000e+00	9.073039e+01
O_LoanToValue	2.748828e+02	0.000000e+00	2.748828e+02
Time_to_Maturity	4.824118e+03	1.985388e+01	4.804264e+03
Lender_Type	2.456347e-01	0.000000e+00	2.456347e-01
Fintech	1.253963e-01	0.000000e+00	1.253963e-01

We can firstly single out the variances of the variables “Confirmed” (which captures the confirmed COVID-19 cases by MSA) and Current_UPB”, as they have a significantly greater magnitude. On the other extreme of the spectrum, variables such as the “Delinquency status”, the “Zero Balance Code” display among the lowest variances. Those variables have both non-zero within and between variance.

On the other hand, the variables “Credit Score”, “Current Interest Rate”, “Original Debt to Income” and “Original Loan to Value”, “Lender type” and “Fintech” do not vary over time, as they are characteristics defined at the origination of the loan. The therefore have no Within variance.

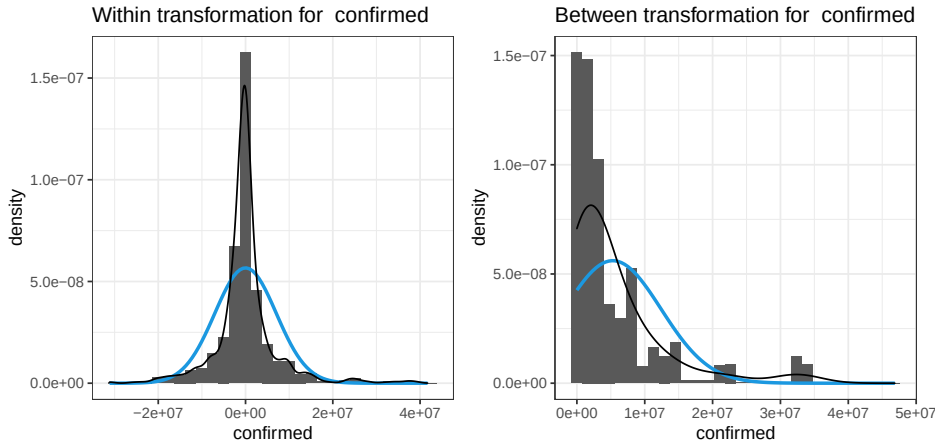
Confirmed COVID cases therefore supply most of the variance, followed by the current UPB, time to maturity and credit score.

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)?

We carried out the calculation of within and between transformations for all variables except the binary variables - namely the lender type - as it would not deliver any useful information. We therefore plot the distribution of the Within and Between transformed variables for the "Delinquency status" dependent variable - which is a categorical variable - and one chosen explanatory variable, "Confirmed COVID cases" - which is continuous.

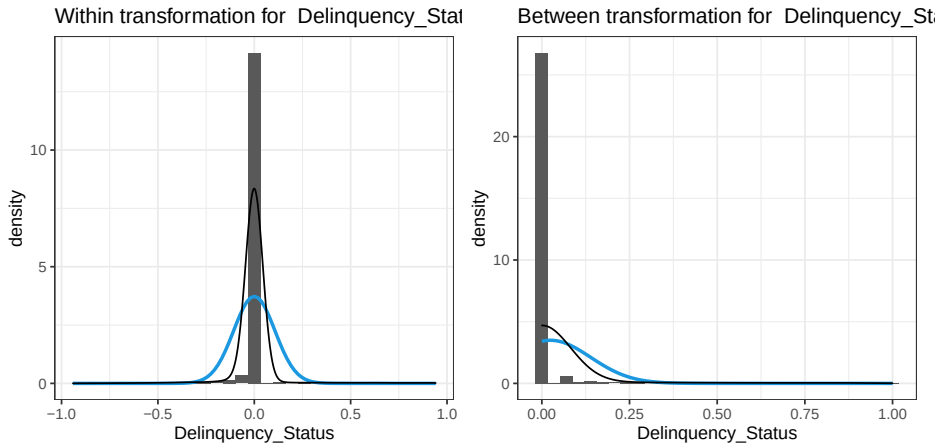
We chose to plot them on separate charts for better visual representation, to facilitate comparison.

5.1. Between and Within transformation of the independent variable (confirmed COVID cases)



The most striking feature is the non-normality of the "Confirmed COVID cases" variable in the Between transformation case. A particularly striking point in the "Between" case is that the histogram and line are clearly skewed towards the right, owing to the presence of outliers (extremely large values at x = 30,000,000). We note that the greatest density is around x=0 in both the Within and Between cases. The striking point of the "Within" case is the leptokurtic normal distribution, with a relatively thicker right tail.

5.2. Between and Within transformation of the dependent variable (Delinquency Status)



The Within and Between histograms of the "Delinquency status" variable are notably more contrasted, owing to the categorical nature of this variable. Both charts are strongly centered around x=0. The Within transformed variable displays a

normal and strongly leptokurtic distribution. The kernel tails, while flattening, are still visible, owing to the presence of outliers. Those outliers are clearly visible in the Between transformation case, which is skewed to the right.

	Confirmed Within	Confirmed Between	Delinquency_Status Within	Delinquency_Status Between
skewness	1.105	2.359	2.008	5.633
kurtosis	7.505	5.698	37.609	33.380

We compare the above skewness and kurtosis coefficients to those of a Normal distribution, which are respectively of 0 and 3.

While the skewness coefficients are somewhat close to those of a normal distribution for both within-transformed variable, it is not the case for the between-transformed variables, which are markedly above 0 - as expected visually.

The Kurtosis coefficients confirm that the within and between distributions for both variables are strongly leptokurtic, especially in the case of the Delinquency status variable.

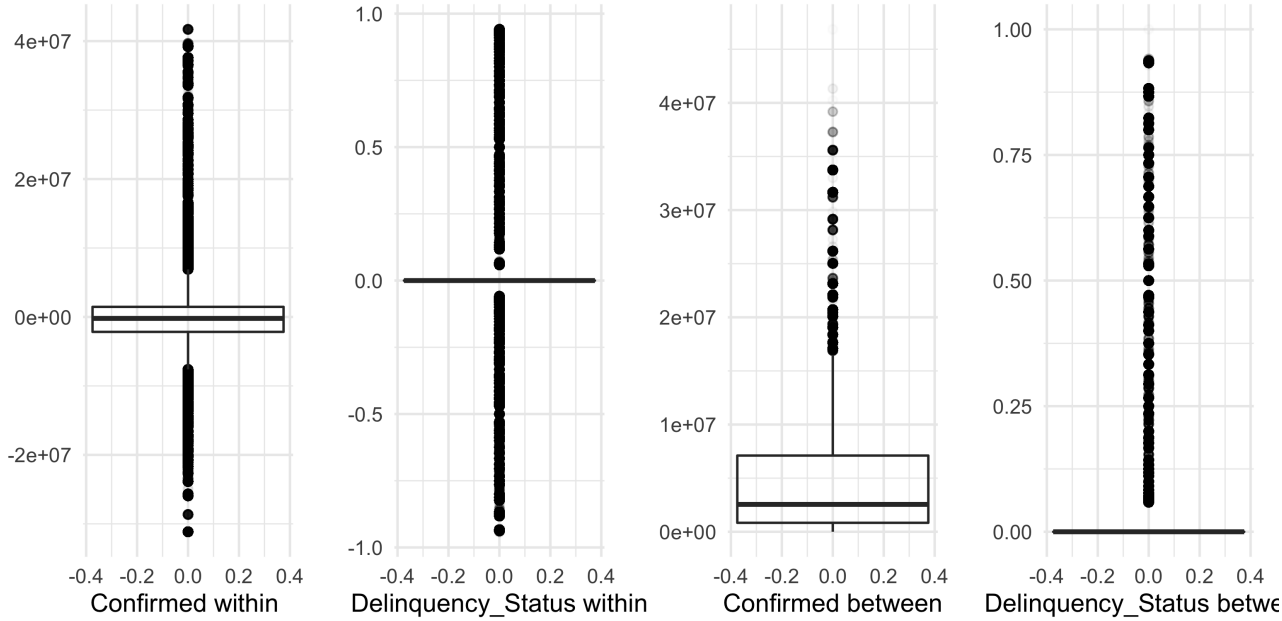


Figure 1: Boxplot for Confirmed Covid cases and Delinquency Status considering within and between transformation

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.

The result of the above boxplot confirms the intuitions derived from the graphs of question 5. Firstly, both variables - both in the within and between transformation case - are centered around 0. Secondly, in the Within-transformed case of both variables, we notice the low interquartile range, which matches their leptokurtic shapes in question 5. The above chart also confirms the presence of tails on both sides on the mean, with relatively more outliers above the mean in the case of the "Confirmed COVID cases" variable.

Finally, turning to the Between-transformed case, the above chart confirms the significantly greater interquartile range for the "Confirmed COVID cases" variable and the presence extreme outliers. It also confirms the very small dispersion of the Between-transformed "Delinquency status" variable.

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)?

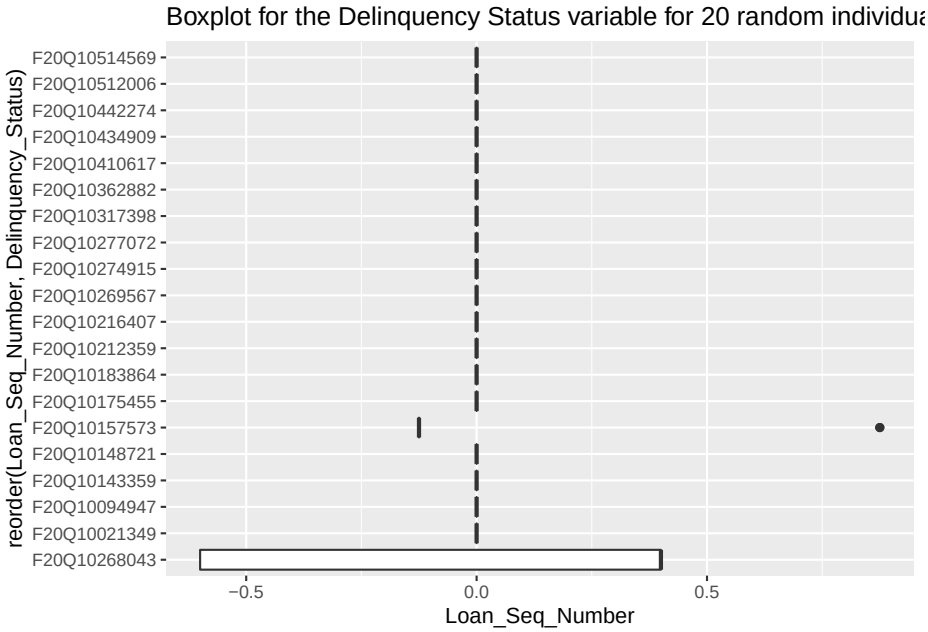
	Confirmed Within	Confirmed Between	Delinquency_Status Within	Delinquency_Status Between
mean	0.0	5282374.0	0.000	0.025
median	-212441.4	2552574.5	0.000	0.000
std	7039814.1	7112223.8	0.107	0.114
min	-31151504.8	39.0	-0.941	0.000
max	41706533.5	46854386.1	0.941	1.000
q25	-2164924.5	835707.9	0.000	0.000
q75	1463448.3	7107209.3	0.000	0.000

The above table confirms the intuitions we derived from the charts and tables in question 5.

We firstly notice a very significant difference between the mean and the median of the "Confirmed COVID cases" variables, both in the Within and Between transformed case. This result matches the large kurtosis of this distribution that we identified in question 5. In the "Within" case, while the distribution is centered around 0, we note that the median is negative: this result is in part driven by the presence of outliers that we highlighted in question 6.

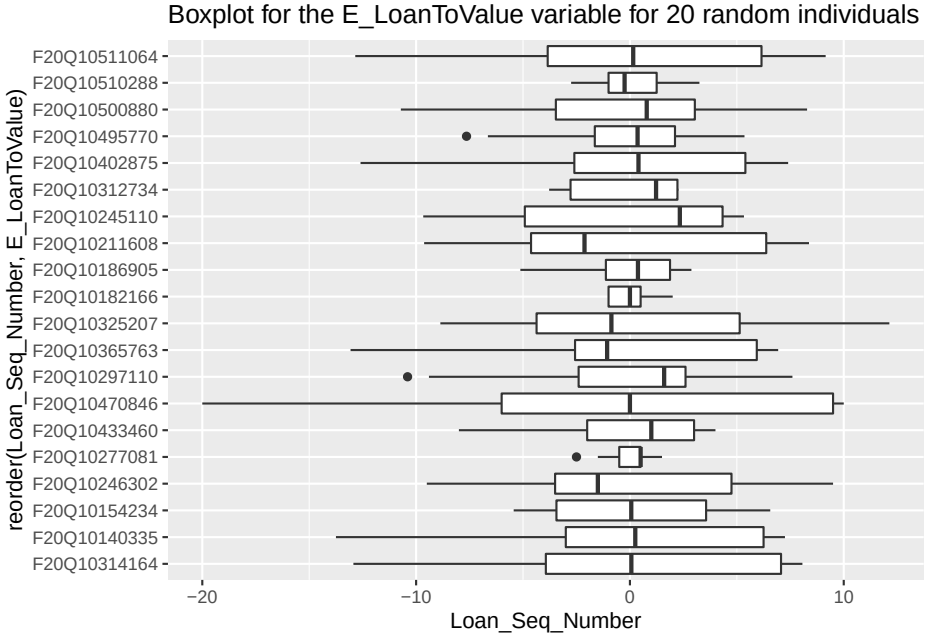
Secondly, turning to the variable "Delinquency status", the very high kurtosis that we identified in question 5 (of 19 and 25 for respectively the Within and the Between cases) is not reflected in the median (of 0.0). This statement is especially true in the "within" case. This is because the negative and positive outliers offset eachother, thereby resulting in a median of 0.

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

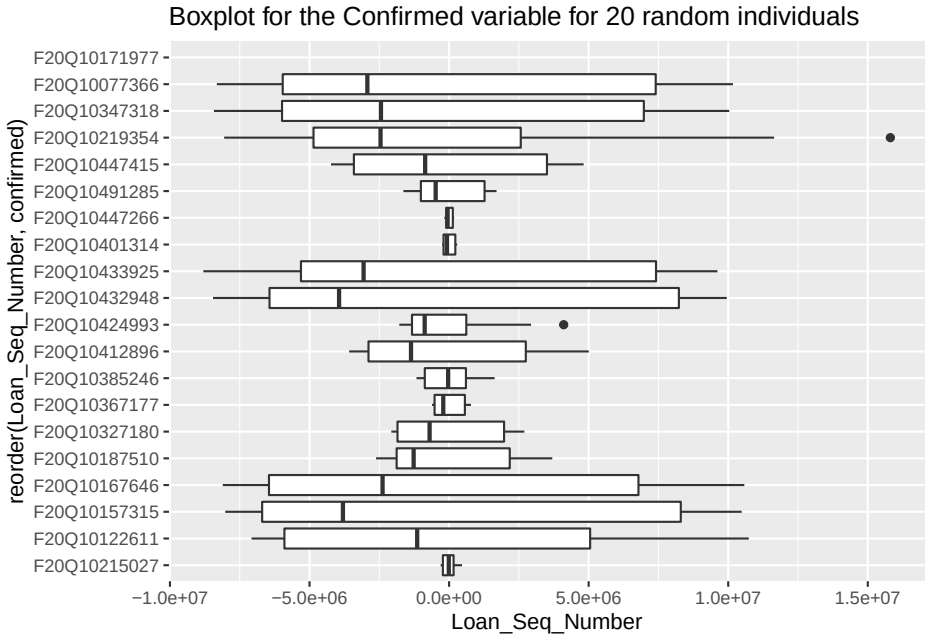


This graphical representation unfortunately cannot provide useful insights for our dependent variable, as it is a categorical variable which takes on the value of 0 for most loans.

Nonetheless, we can infer more conclusions from the dependent variables' charts.



The means are in overall centered around 0. Standard errors and interquartile ranges are very heterogeneous across different loans.



Means precisely centered around 0. Standard errors and interquartile ranges are very heterogeneous across loans.

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.

We exclude from the within variables bivariate correlation matrix the variables which are time-invariant (which therefore have 0 standard deviation).

Table 1: Correlation matrix considering a within transformation

	Delinquency_Status	Zero_Balance_Code	confirmed	Current_Interest_Rate	Current_UPB	E_LoanToValue	Time_to_Maturity
Delinquency_Status	1	.	.	.	.	.	.
Zero_Balance_Code	−0.017	1	.	.	.	.	.
confirmed	−0.004	0.118	1	.	.	.	.
Current_Interest_Rate	0.035	0.000	−0.008	1	.	.	.
Current_UPB	0.018	−0.901	−0.151	−0.001	1	.	.
E_LoanToValue	−0.003	−0.027	−0.447	0.008	0.185	1	.
Time_to_Maturity	−0.030	−0.171	−0.597	−0.147	0.206	0.722	1

As our dependent variable of interest is the "Delinquency status", we focus on the first column of this table.

Those results are surprising: the number of confirmed cases, the estimated loan to value ratio and the loan's time to maturity are inversely related to the delinquency status. As increases in those variables indicate a degradation in the loans' credit quality, we would have expected those variables to be positively correlated with the delinquency status. Only interest rate has an expected positive sign - as a higher cost of debt can reasonably be associated with a more important loan delinquency.

Table 2: Correlation matrix considering a between transformation

	Delinquency_Status	Zero_Balance_Code	confirmed	Credit_score	Current_Interest_Rate	Current_UPB	E_LoanToValue	O_DebtToIncome	O_LoanToValue	Time_to_Maturity	Lender_Type	Fintech
Delinquency_Status	1	.	.	.	.	.	.	.	.	.	.	.
Zero_Balance_Code	−0.073	1	.	.	.	.	.	.	.	.	.	.
confirmed	0.067	−0.207	1	.	.	.	.	.	.	.	.	.
Credit_score	−0.120	0.067	−0.024	1	.	.	.	.	.	.	.	.
Current_Interest_Rate	0.087	0.019	−0.028	−0.329	1	.	.	.	.	.	.	.
Current_UPB	0.057	0.118	0.190	0.080	−0.173	1	.	.	.	.	.	.
E_LoanToValue	0.062	0.074	−0.057	−0.092	0.186	0.177	1	.	.	.	.	.
O_DebtToIncome	0.090	−0.004	0.075	−0.171	0.162	0.051	0.103	1	.	.	.	.
O_LoanToValue	0.053	−0.026	−0.067	−0.092	0.182	0.095	0.928	0.097	1	.	.	.
Time_to_Maturity	0.040	0.064	0.002	−0.039	0.392	0.183	0.328	0.109	0.301	1	.	.
Lender_Type	0.024	−0.075	0.012	0.008	0.038	−0.078	0.031	−0.016	0.055	−0.018	1	.
Fintech	0.009	0.052	−0.003	−0.055	−0.005	−0.016	−0.027	0.036	−0.045	−0.064	−0.363	1

Thanks to the presence of variance in more explanatory variables, we can identify many more interesting relationships with the "Delinquency status" variable.

Focusing again on the first column, the first striking point is that the correlation coefficients for nearly all variables take on an expected positive sign. Those above results indeed verify our working hypothesis that (i) confirmed COVID cases are positively related to the loan delinquency, similarly as (ii) interest rate, loan size, loan-to-value and debt-to-income. Nonetheless, it remains surprising that the "Credit score" correlation coefficient is negative. Similarly, the positive coefficient for the "Lender type" variable is surprising, as it seems to indicate that the fact that the lender is a commercial bank is positively related to loan delinquency, compared to if it was a shadow bank.

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

We now turn to exploring autocorrelation and trend-correlation of the within-transformed variable and pooled variables (as the between-transformed variable has no time dimension).

Starting with autocorrelation, we compute the autocorrelation coefficients for the following variables: delinquency status, confirmed COVID cases, the current interest rate, the current unpaid principal balance and estimated loan to value.

```
#> # A tibble: 7 x 2
#>   name                 value
#>   <chr>                <dbl>
#> 1 Delinquency_Status   0.846
#> 2 Zero_Balance_Code    0.97
#> 3 confirmed            0.984
#> 4 Current_Interest_Rate 0.987
#> 5 Current_UPB          0.981
#> 6 E_LoanToValue        0.974
#> 7 Time_to_Maturity     1
```

The striking feature of the above table is that the within-transformed variables display a strongly auto-regressive pattern, with the delinquency status variables displaying the lowest coefficient as expected.

Those results come as expected:

- (i) confirmed COVID-19 cases display a gradual rise over time
- (ii) the current interest rate exhibits little change over time
- (iii) the current unpaid principal balance displays a gradual decrease over time.

We verify, using the below code, the p_value significance of those autocorrelation coefficients.

```
#> # A tibble: 7 x 2
#>   name                    value
#>   <chr>                   <dbl>
#> 1 Delinquency_Status    3.69e- 5
#> 2 Zero_Balance_Code     5.81e-10
#> 3 confirmed             4.14e-11
#> 4 Current_Interest_Rate 1.27e-12
#> 5 Current_UPB           2.38e-11
#> 6 E_LoanToValue         1.82e-10
#> 7 Time_to_Maturity      4.23e-24
```

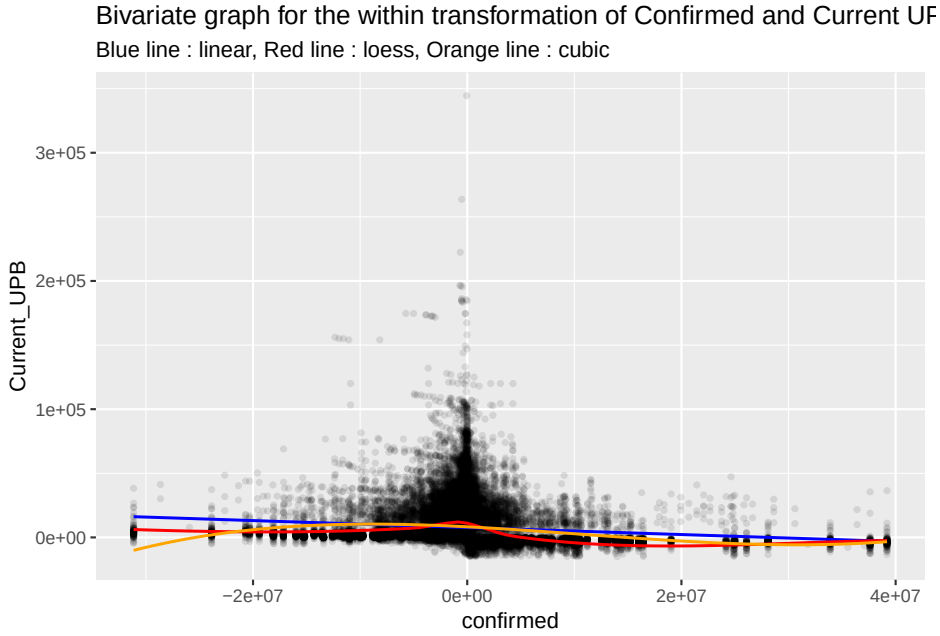
The extremely small p-values reported in the above table shows the significance of the strong auto-regressive pattern that we identified for those variables.

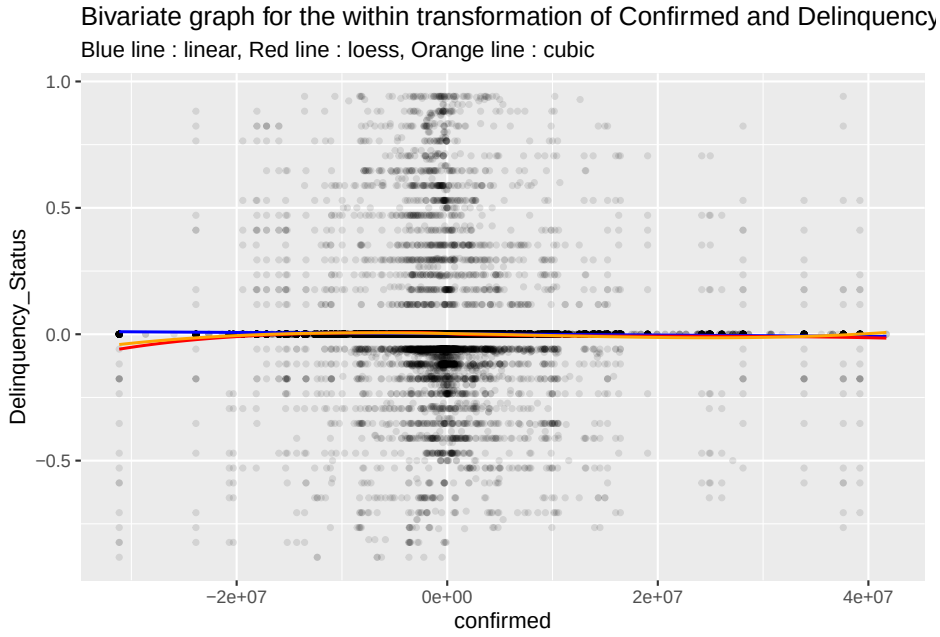
We now examine the time-trend of those within-transformed variable.

```
#>      Reporting_Period    Delinquency_Status     Zero_Balance_Code
#>             1.0000000             0.1953637             0.9470408
#>             confirmed Current_Interest_Rate           Current_UPB
#>             0.9668491            -0.7896294            -0.9570242
#>         E_LoanToValue      Time_to_Maturity
#>            -0.9321785            -0.9998331
```

The above table suggests that the the within-transformed COVID confirmed cases, current loan interest rate, unpaid principal balance and estimated loan-to-value ratio display strong time trends.

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.





The relationship between confirmed COVID cases and our dependent variable (namely the loans' "Delinquency status") can unfortunately not be correctly captured through a linear correlation setup.

To estimate the strength of those relationship, we turn to a Linear Probability Model (LPM) framework.

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.

12.1. Linear Probability Models (LPM) on between-transformed variables.

We run a between Linear Probability Model (LPM) regression of the loans' delinquency status on :

- (i) the COVID-19 confirmed cases
- (ii) the borrower's estimate loan-to-value
- (iii) the loan's current interest rate.

```
#> Oneway (individual) effect Between Model
#>
#> Call:
#> plm(formula = Delinquency_Status ~ confirmed + E_LoanToValue +
#>     Current_Interest_Rate, data = DB_KJ, model = "between", index = c("Loan_Seq_Number",
#>     "Reporting_Period"))
#>
#> Unbalanced Panel: n = 314445, T = 1-16, N = 4125176
#> Observations used in estimation: 314445
#>
#> Residuals:
#>      Min.   1st Qu.    Median   3rd Qu.      Max.
#> -0.220030 -0.033148 -0.023961 -0.013500  0.996021
#>
#> Coefficients:
#>                          Estimate  Std. Error t-value  Pr(>|t|)
#> (Intercept)           -8.6004e-02  1.8253e-03 -47.119 < 2.2e-16 ***
#> confirmed              1.1932e-09  3.0322e-11  39.351 < 2.2e-16 ***
#> E_LoanToValue          4.0235e-04  1.3770e-05  29.219 < 2.2e-16 ***
#> Current_Interest_Rate  2.1532e-02  4.6711e-04  46.095 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Total Sum of Squares:    4713.6
#> Residual Sum of Squares: 4641
#> R-Squared:      0.015401
#> Adj. R-Squared: 0.015392
#> F-statistic: 1639.51 on 3 and 314441 DF, p-value: < 2.22e-16
```

Although all the regressors are statistically significant, the extremely small size of the regressors suggests the weak power of those variables to account for variability in the Delinquency status variable. We suspect that this is due to the significant imbalance in the dependent variable's distribution, as we have only 127,000 observations of significant loan delinquency against more than 4,700,000 observations of non-delinquency. This explains the very low R-squared of this regression.

As this homework exercise is an opportunity for us to explore this Freddie Mac loan performance dataset, we run a Linear Probability Model (LPM) with other variables, in order to answer a different question: "During the first periods of the COVID-19 crisis, did shadow banks, especially fintechs, serve riskier borrowers?"

Based on Wang (2020), we run the following regression:

```
#> Oneway (individual) effect Between Model
#>
#> Call:
#> plm(formula = Fintech ~ E_LoanToValue + Credit_score + E_LoanToValue *
#>     Credit_score + Current_Interest_Rate, data = DB_KJ, model = "between",
```

```
#>     index = c("Loan_Seq_Number", "Reporting_Period"))
#>
#> Unbalanced Panel: n = 335404, T = 1-17, N = 4469982
#> Observations used in estimation: 335404
#>
#> Residuals:
#>     Min.  1st Qu.   Median  3rd Qu.     Max.
#> -0.29018 -0.16023 -0.14284 -0.12642  1.16249
#>
#> Coefficients:
#>                               Estimate  Std. Error  t-value  Pr(>|t|)
#> (Intercept)                  8.8355e-01  4.6011e-02  19.2030 < 2.2e-16 ***
#> E_LoanToValue               -4.2158e-03  6.8337e-04  -6.1692 6.873e-10 ***
#> Credit_score                -8.4310e-04  5.9897e-05 -14.0759 < 2.2e-16 ***
#> Current_Interest_Rate       -1.3950e-02  1.4048e-03  -9.9304 < 2.2e-16 ***
#> E_LoanToValue:Credit_score   4.7008e-06  9.0085e-07   5.2182 1.808e-07 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Total Sum of Squares:    43249
#> Residual Sum of Squares: 43054
#> R-Squared:      0.0045018
#> Adj. R-Squared: 0.0044899
#> F-statistic: 379.18 on 4 and 335399 DF, p-value: < 2.22e-16
```

As in Wang (2020), the R-squared is very low. Consistently with the intuition from the literature and to Wang (2020), the lower the credit score at origination, the more likely the loan is to have been originated by a Fintech lender. The above coefficients can be interpreted the following way: a 1 unit decrease in borrower's FICO score raises the probability of the loan being originated by a Fintech of 0.084%. Turning to the interpretation of the loan-to-value ratio coefficient: a 1% decrease in the borrower LTV ratio raises the probability that the loan was originated by a Fintech lender. Those results suggest that Fintech lenders, while targeting borrowers with lower credit scores, might compensate the risk by requiring a lower LTV ratio (in other words, higher collateral).

12.2. Linear Probability Models (LPM) on within-transformed variables.

For the coming regressions, we will focus on the first relationship, on the drivers of mortgage loan delinquency, as unfortunately, running the Linear Probability Models for predicting the probability of Fintech lending on the different variables transformations is computationally very heavy.

We now run within fixed effects, random fixed effects and first-difference estimations, starting by our first relationship of interest, between COVID cases at the MSA level and loan delinquency status.

The fixed effect estimation controls for MSA-level characteristics thanks to the inclusion of MSA fixed effects. While the MSA fixed effects avoid the risk of omitted variable bias - as it controls for time-invariant factors that vary across MSA.

```
#> Oneway (individual) effect Within Model
#>
#> Call:
#> plm(formula = Delinquency_Status ~ confirmed + E_LoanToValue,
#>     data = DB_KJ, model = "within")
#>
#> Unbalanced Panel: n = 314445, T = 1-16, N = 4125176
#>
#> Residuals:
#>        Min.     1st Qu.      Median     3rd Qu.        Max.
```

```
#> -9.4122e-01 -2.3589e-04 -1.8488e-05  1.3719e-04  9.4188e-01
#>
#> Coefficients:
#>                  Estimate  Std. Error  t-value  Pr(>|t|)
#> confirmed     -1.1234e-10  8.8682e-12 -12.6677 < 2.2e-16 ***
#> E_LoanToValue -3.0422e-05  1.0782e-05  -2.8214  0.004781 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Total Sum of Squares:    46885
#> Residual Sum of Squares: 46882
#> R-Squared:      4.48e-05
#> Adj. R-Squared: -0.082467
#> F-statistic: 85.3638 on 2 and 3810729 DF, p-value: < 2.22e-16
```

12.3. Random effect estimation

```
#> Oneway (individual) effect Random Effect Model
#>    (Swamy-Arora's transformation)
#>
#> Call:
#> plm(formula = Delinquency_Status ~ confirmed + E_LoanToValue,
#>     data = DB_KJ, model = "random")
#>
#> Unbalanced Panel: n = 314445, T = 1-16, N = 4125176
#>
#> Effects:
#>                   var std.dev share
#> idiosyncratic 0.01230 0.11092 0.456
#> individual    0.01467 0.12112 0.544
#> theta:
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>  0.3246  0.7623  0.7699  0.7583  0.7768  0.7768
#>
#> Residuals:
#>     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
#> -0.74195 -0.00747 -0.00643  0.00017 -0.00540  0.94985
#>
#> Coefficients:
#>                 Estimate Std. Error z-value  Pr(>|z|)
#> (Intercept)   1.3460e-02 6.2263e-04 21.6178 < 2.2e-16 ***
#> confirmed     4.0783e-11 8.2387e-12  4.9501 7.417e-07 ***
#> E_LoanToValue 1.9484e-04 8.5106e-06 22.8942 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Total Sum of Squares:    50570
#> Residual Sum of Squares: 50569
#> R-Squared:      6.5365e-05
#> Adj. R-Squared: 6.488e-05
#> Chisq: 537.741 on 2 DF, p-value: < 2.22e-16
```

12.4. First difference estimation

```
#> Oneway (individual) effect First-Difference Model
#>
#> Call:
#> plm(formula = Delinquency_Status ~ confirmed + E_LoanToValue,
#>     data = DB_KJ, model = "fd")
#>
#> Unbalanced Panel: n = 314445, T = 1-16, N = 4125176
#> Observations used in estimation: 3810731
#>
#> Residuals:
#>       Min.    1st Qu.     Median    3rd Qu.       Max.
#> -1.0197066 -0.0017979 -0.0015259 -0.0012574  1.0861467
#>
#> Coefficients:
#>                  Estimate  Std. Error t-value  Pr(>|t|)
#> (Intercept)    1.3890e-03  6.6475e-05  20.894 < 2.2e-16 ***
#> confirmed     -1.3557e-10  3.1301e-11  -4.331 1.484e-05 ***
#> E_LoanToValue -2.2669e-04  1.9164e-05 -11.829 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Total Sum of Squares:    46013
#> Residual Sum of Squares: 46011
#> R-Squared:      4.2126e-05
#> Adj. R-Squared: 4.1601e-05
#> F-statistic: 80.2681 on 2 and 3810728 DF, p-value: < 2.22e-16
```

One conclusion can be drawn of the above results: while the variable of confirmed COVID cases and for the estimated loan-to-value are significant at the 1% level, the R-squared and the correlation coefficients are extremely low.
Those results corroborate the findings of the between estimation: those two variables therefore do not seem to capture much variability in the Delinquency status data, due to the strong imbalance between delinquent and non-delinquent loans.

Finally, to check the existence of individual fixed effects, we run a Hausman test, comparing the results of the Between and Within Estimators.

Table 3: Hausman endogeneity test for the random effects wage model

statistic	p.value	parameter	method	alternative
2201.251	0	2	Hausman Test	one model is inconsistent

12.5. Hausman endogeneity test

The above output shows a low p -value of the test, which indicates that the null hypothesis - stating that the individual random effects are exogenous - is rejected. The random effects estimation is therefore inconsistent. The fixed effects model is therefore the correct specification.

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.

The effect of any time-invariant variable (such as the credit score) is eliminated through the "de-meaning transformation" of the variables. We therefore need to find a way to estimate the coefficients of "Confirmed" and "Estimated Loan-to-Value" without including fixed effects, while dealing with correlations with the error term, owing to ommitted variable bias. The Hausmann-Taylor approach will therefore be useful for this purpose: it allows us to account for panel data biases and still estimate our panel data coefficients - which cannot be done in a standard fixed effect approach.

For conducting the Hausman-Taylor estimation, the R software does all the calculations for deciding which variable will be instrumented by a de-meaned instrumental variable and which variables will be employed in random effects.

The results are reported in the below table.

```
#> Oneway (individual) effect Random Effect Model
#>    (Hausman-Taylor's transformation)
#> Instrumental variable estimation
#>    (Baltagi's transformation)
#>
#> Call:
#> plm(formula = Delinquency_Status ~ confirmed + E_LoanToValue +
#>     Credit_score | E_LoanToValue + confirmed, data = DB_KJ, model = "random",
#>     random.method = "ht", inst.method = "baltagi")
#>
#> Unbalanced Panel: n = 314421, T = 1-16, N = 4124879
#>
#> Effects:
#>                   var std.dev share
#> idiosyncratic 0.01230 0.11092 0.383
#> individual    0.01984 0.14087 0.617
#> theta:
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>  0.8069  0.8069  0.8069  0.8069  0.8069  0.8069
#>
#> Residuals:
#>       Min.    1st Qu.     Median    3rd Qu.       Max.
#> -7.4930828 -0.2040439 -0.0092638  0.1249319  8.8673496
#>
#> Coefficients:
#>                   Estimate  Std. Error  z-value  Pr(>|z|)
#> (Intercept)    2.0767e+00  6.2122e-02  33.4299 < 2.2e-16 ***
#> confirmed     -7.9239e-11  8.7659e-12  -9.0395 < 2.2e-16 ***
#> E_LoanToValue -4.2250e-05  1.0782e-05  -3.9185 8.912e-05 ***
#> Credit_score  -2.7201e-03  8.2089e-05 -33.1358 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Total Sum of Squares:    49306
#> Residual Sum of Squares: 4133100
#> R-Squared:      0.0007482
#> Adj. R-Squared: 0.00074747
#> Chisq: 1388.06 on 3 DF, p-value: < 2.22e-16
```

We now end up with theoretically unbiased estimations of those panel marginal effects. The results are quite comparable to the other specifications: all variables are significant to the 1% level. Nonetheless, the R-squared is extremely low.

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.

To conclude, we verify the strength of our instruments by examining their correlation with our time-invariant variable "Credit score".

```
#> Oneway (individual) effect Between Model
#>
#> Call:
#> plm(formula = Credit_score ~ E_LoanToValue + confirmed + Current_Interest_Rate,
#>     data = DB_KJ, model = "between", index = c("Loan_Seq_Number",
#>         "Reporting_Period"))
#>
#> Unbalanced Panel: n = 314421, T = 1-16, N = 4124887
#> Observations used in estimation: 314421
#>
#> Residuals:
#>      Min.   1st Qu.    Median   3rd Qu.      Max.
#> -185.2864  -26.5679    5.9014   30.5716  133.4394
#>
#> Coefficients:
#>                          Estimate  Std. Error  t-value  Pr(>|t|)
#> (Intercept)            8.6861e+02  6.0370e-01 1438.815 < 2.2e-16 ***
#> E_LoanToValue         -8.7056e-02  4.5544e-03  -19.115 < 2.2e-16 ***
#> confirmed             -2.1977e-07  1.0028e-08  -21.914 < 2.2e-16 ***
#> Current_Interest_Rate -2.9414e+01  1.5449e-01 -190.391 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Total Sum of Squares:    571430000
#> Residual Sum of Squares: 507600000
#> R-Squared:      0.1117
#> Adj. R-Squared: 0.11169
#> F-statistic: 13179.1 on 3 and 314417 DF, p-value: < 2.22e-16
```

Although each explanatory variable is significant at the 1% level, the reported R-squared is low (11%). Those results suggest the relative weakness of our the estimated Loan-to-Value ratio, the Confirmed COVID cases and Current interest rates variables for instrumenting the time-invariant variable Credit score.

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)

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.

We download yearly data for 88 countries over 23 periods (between 1996 and 2019) for the following variables:

- (i) the World Bank corruption index
- (ii) GDP per capita in euros
- (iii) PPP-adjusted GDP per capita
- (iv) foreign aid - as % of Gross National Income and per capita
- (v) population.

```
#> # 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
```

Based on the below table, as we lack observations for the years 1997, 1999 and 2001, we restrict our analysis to the 2002-2019 period for those 88 countries.

```
#> # 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 therefore remove the years 1996, 1998 and 2000 from our sample.

```
#> # 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
```

As we are concerned with the impact of foreign aid on economic growth, we also remove from our sample countries that have a negative net ODA.

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

We therefore arrive to the following final dataset, which comprises 77 countries over 18 years (2002-2019).

```
#> # 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
```

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).

Taking the log for GDP per capita, Population and ODA net, we then take the first difference in order to get the growth rate.

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

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.

We separate our dataset in two periods of approximately equal size in terms of available observations: 2002-2013 for Period 1 and 2014-2019 for Period 2.

As in Jia and Williamson (2018), we compare the countries that are the top recipients of foreign aid compared to GDP in those two subsets of the database.
First, Period 1 2002-2013 :

```
#> # 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 Malawi                   0.171     -0.589   1234.        322.   15.8     55.7
#>  2 Rwanda                   0.147     -0.006   1339.        533.   17.4     79.8
#>  3 Cabo Verde               0.146      0.757   5610.       2724.   14.9    404.
#>  4 Sierra Leone             0.134     -0.919   1381.        526.   19.3     69.1
#>  5 Burkina Faso             0.107     -0.272   1617.        549.   11.0     59.2
#>  6 Niger                    0.09      -0.746   1002.        429.    9.62    38.8
#>  7 Ethiopia                 0.084     -0.652   1076.        387.   13.2     32.4
#>  8 Central Afric~           0.084     -1.11    1091.        483.    9.59    40.7
#>  9 Lesotho                  0.081      0.043   2181.        925.    6.51    78.9
#> 10 Tanzania                 0.08      -0.561   1864.        691.    9.46    55.5
#> # ... with 4 more variables: population <dbl>, g_gdp_per_cap <dbl>,
#> #   g_population <dbl>, g_oda_net <dbl>
```

Secondly, Period 2 2014-2019 :

```
#> # 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 Central Afric~           0.324     -1.23     892.        395.   28.4    128.
#>  2 Malawi                   0.178     -0.743   1479.        386.   16.3     68.9
#>  3 Vanuatu                  0.172     -0.028   3031.       2799.   16.0    480.
#>  4 Burundi                  0.164     -1.32     799.        296.   18.9     48.5
#>  5 Sierra Leone             0.148     -0.675   1688.        643.   17.6     95.3
#>  6 Solomon Islan~           0.138     -0.144   2615.       2250.   13.8    311.
#>  7 Rwanda                   0.122      0.636   1985.        789.   12.9     96.0
#>  8 Guinea-Bissau            0.117     -1.52    1862.        624.   10.4     72.8
#>  9 Gambia, The              0.114     -0.61    2098.        673.   11.3     76.8
#> 10 Niger                    0.104     -0.632   1166.        499.    9.35    51.8
#> # ... with 4 more variables: population <dbl>, g_gdp_per_cap <dbl>,
#> #   g_population <dbl>, g_oda_net <dbl>
```

We notice that the 10 countries with the highest ODA/GDP average over the period differ between those two samples. While Malawi, Rwanda, Sierra Leone, Niger and the Central African Republic are top recipients in foreign aid / GDP on average for both periods 1 and 2, Burkina Faso, Cabo Verde and Tanzania have a lower relative to period 1 and are no more among the top 10 recipients in period 2.

Average ODA/GDP are higher in period 2 - ranging from 10.4% to 32% compared to 7.9% and 17% in period 1 - explained by the presence of other countries compared to period 1, namely the Gambia, Guinea-Bissau, the Solomon Islands, Burundi and Vanuatu.

The magnitude of the change in the average ODA/GDP ratio for the Central African republic is the most striking, from 17% in period 1 to 32% on average over the period 2.

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

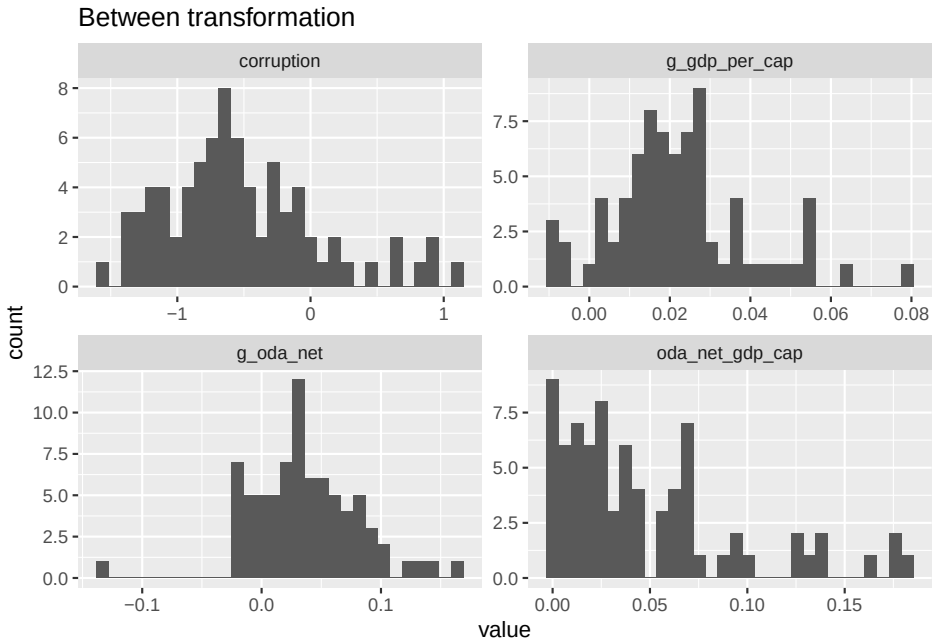
Now turning to the countries with the lowest ODA/GDP ratios, 23 countries of our dataset have ratios between 0% and 0.5%. Notably, Algeria, Brazil, Costa Rica, Dominican Republic, India, Mexico and Turkey have a ratio lower than 0.5% for all the 18 years of observations.

Albania, Côte d'Ivoire, Iraq and Vietnam, while having among the lowest ODA/GDP ratios, have an ODA/GDP lower than 0.5% for only 1 year of observation.

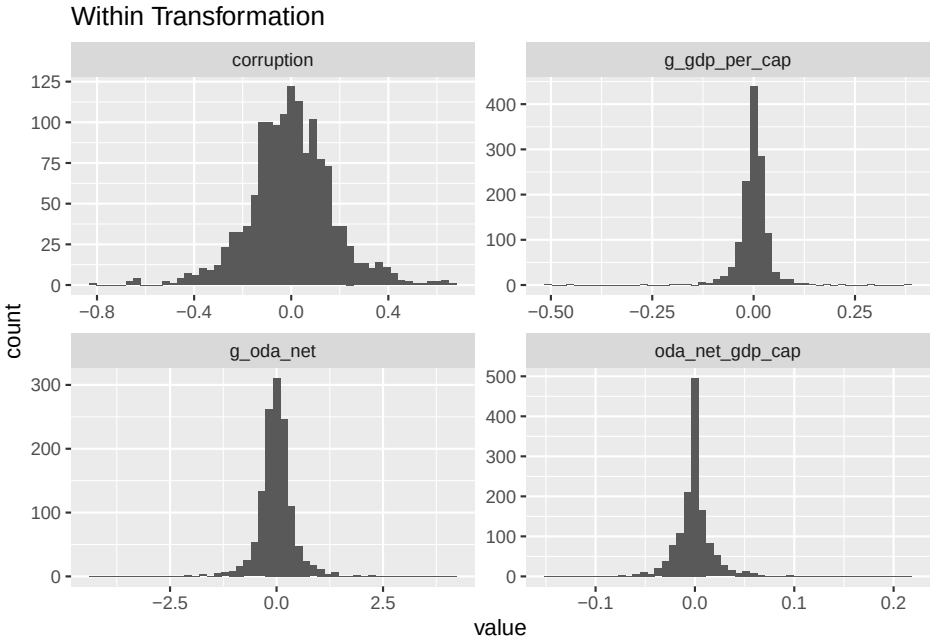
```
#> # A tibble: 23 x 3
#> # Groups:   country [23]
#>    country                n  prop
#>    <chr>              <int> <dbl>
#>  1 Algeria               18  1
#>  2 Brazil                18  1
#>  3 Costa Rica            18  1
#>  4 Dominican Republic    18  1
#>  5 India                 18  1
#>  6 Mexico                18  1
#>  7 Turkey                18  1
#>  8 Ecuador               17  0.94
#>  9 Colombia              16  0.89
#> 10 Equatorial Guinea     16  0.89
#> # ... with 13 more rows
```

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.

The important message of the below charts in the non-normality of the 'Between-transformed' variables. While we have no numeric information regarding the parameters of those distributions, we notice that the Between-transformed 'corruption' variable displays a positive skewness - it has a relatively fatter right tail. The Between-transformed 'ODA / GDP per capita' rather has the shape of a gamma distribution.

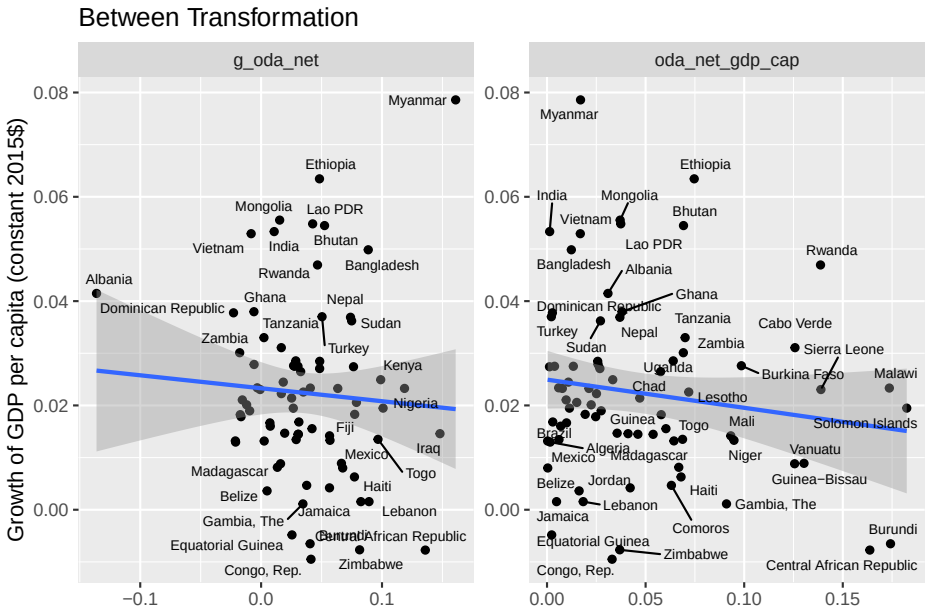


The most striking characteristic in the 'within-transformed' case is the apparent normality of the corruption variable and the 'ODA / GDP per capita' variable. Although those charts give no precisions about the kurtosis of those within-transformed variable, we can at least conclude that the distributions are leptokurtic and that those variables are evenly-distributed around the mean.


Within Transformation
corruption
g_gdp_per_cap
g_oda_net
oda_net_gdp_cap
count
value

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.

We start by examining the bivariate graphs in the between transformation case.



In the between case, a negative relationship between GDP per capita growth and (i) ODA/GDP and (ii) ODA/GDP growth seems to be captured by this linear model. Observations seems to be clustered around ODA/GDP = 0.05 and ODA/GDP growth a little above 0. We nonetheless need to examine whether to make sure whether the fitting of those 2 univariate models is not only due to noise, and therefore to examine the significance of the correlation coefficient between these two variables.

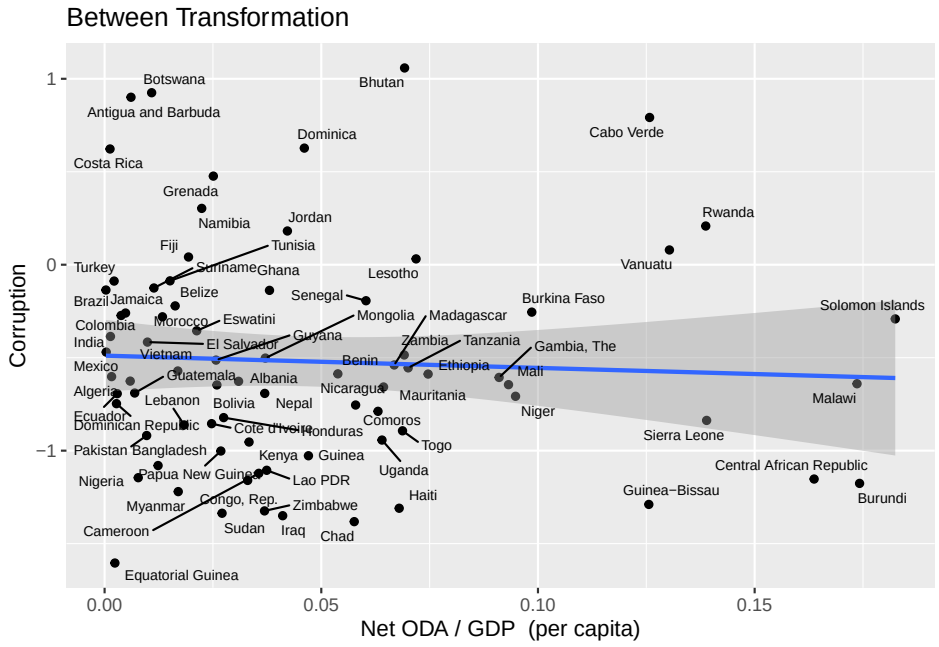
```
#>
#> Call:
#> lm(formula = g_gdp_per_cap ~ -1 + g_oda_net, data = .)
#>
#> Residuals:
#>       Min        1Q    Median        3Q       Max
#> -0.039045  0.001125  0.014976  0.023237  0.072861
#>
#> Coefficients:
#>           Estimate Std. Error t value Pr(>|t|)
#> g_oda_net  0.23051    0.04901   4.704 1.12e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.02486 on 76 degrees of freedom
#> Multiple R-squared:  0.2255, Adjusted R-squared:  0.2153
#> F-statistic: 22.12 on 1 and 76 DF,  p-value: 1.121e-05


#>
#> Call:
```

```
#> lm(formula = g_gdp_per_cap ~ -1 + oda_net_gdp_cap, data = .)
#>
#> Residuals:
#>       Min        1Q    Median        3Q       Max
#> -0.044436 -0.000812  0.013942  0.022106  0.074887
#>
#> Coefficients:
#>                 Estimate Std. Error t value Pr(>|t|)
#> oda_net_gdp_cap  0.21764    0.04188   5.196 1.66e-06 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.02426 on 76 degrees of freedom
#> Multiple R-squared:  0.2621, Adjusted R-squared:  0.2524
#> F-statistic:    27 on 1 and 76 DF,  p-value: 1.657e-06
```

As per the above results, the correlation coefficient between GDP per capita and ODA/GDP is significant at the 1% level in both cases, which suggests that those two univariate models are not fitting to noise or outliers.

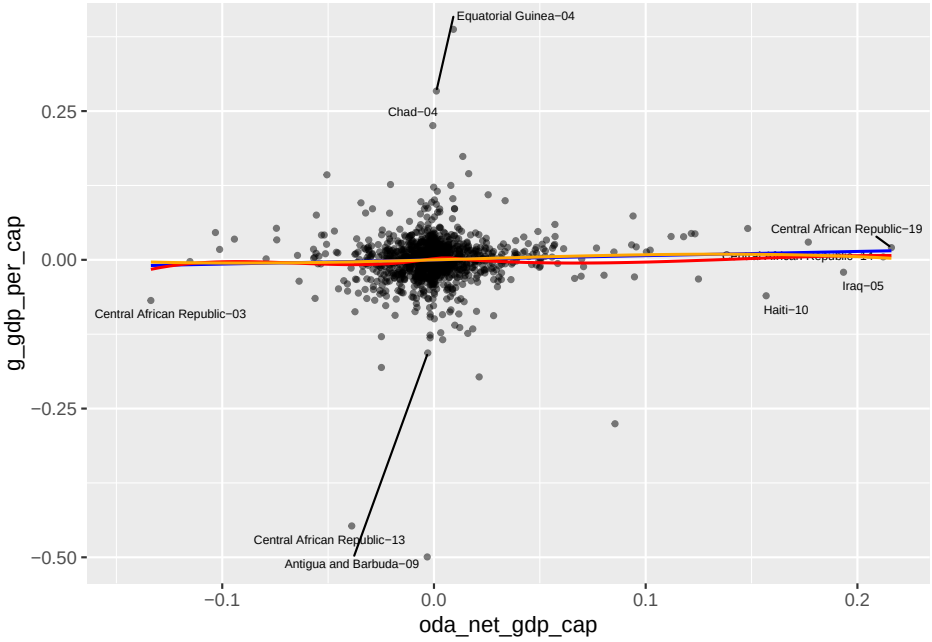
We continue by examining the bivariate graph of the corruption index and ODA/GDP in the between transformation case.



As opposed to the previous two cases, the bivariate linear specification does not manage to capture the variance of the corruption index. The regression line is close to y=0.

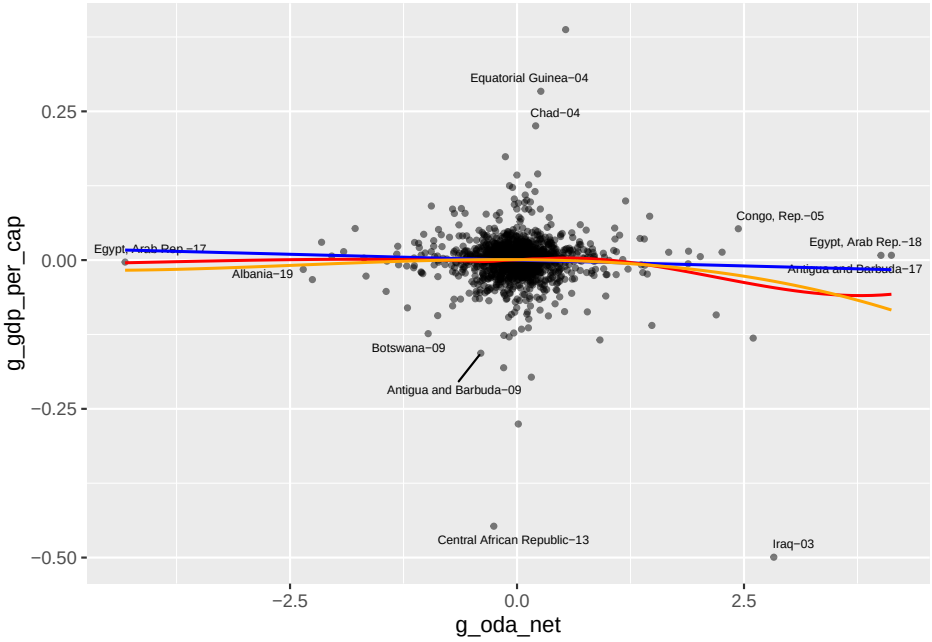
Turning now to the bivariate graphs of the within case:

We start by plotting the within bivariate graph of the growth of GDP/capita with ODA/GDP. We add 3 specifications of regression lines to those clouds of data: a linear model (in blue), a quadratic specification (in orange) and a LOWESS specification (in red).



In this first case, the 3 specifications fail to capture the variance of the growth of GDP/capita, as the fiting lines are nearly horizontal. Observations are markedly clustered around ODA/GDP = 0, which suggests orthogonality (or put simply, uncorrelation) between those two variables in the within-transformed case.

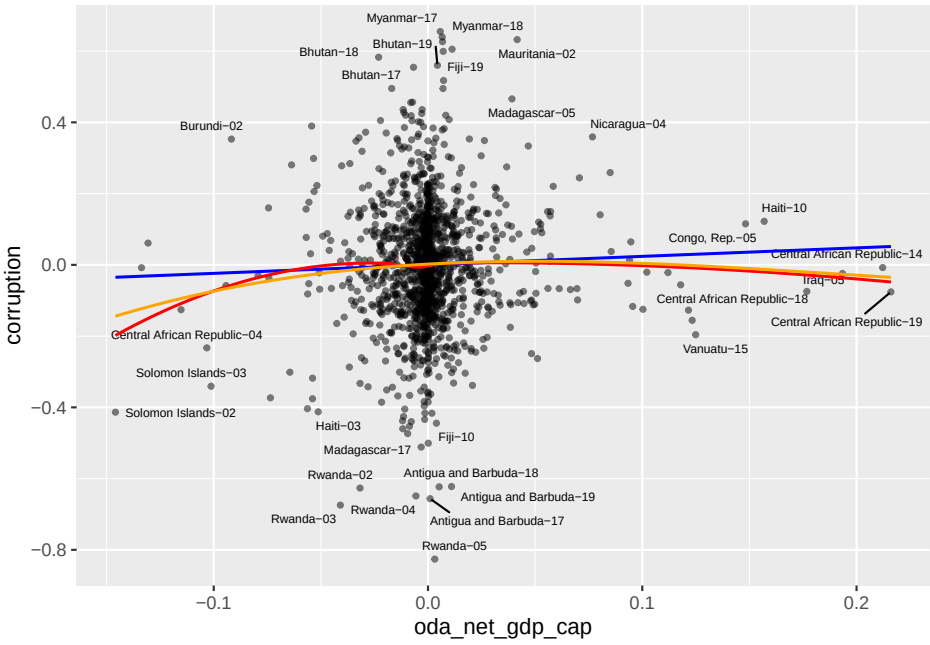
We then plot the within bivariate graph of the growth of GDP / capita with the growth rate of ODA.



The same conclusion can be drawn for this bivariate relationship in the linear case, which seems to fail to capture the variance of GDP per capita growth. While the quadratic and LOWESS specifications seem to perform slightly better, this is driven by the presence of outliers.

Finally, We plot the within bivariate graphs of the corruption index with ODA/GDP. Similarly to the previous case, the strong

cluster around ODA/GDP = 0 signals the same orthogonality between those two variables, and the quadratic and LOWESS specifications seem to capture variance driven by the outliers.



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

Table 4: Correlation matrix

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

From the first column of the above table, we notice that as expected, corruption, population, and GDP per capita are negatively related to the level of foreign aid/GDP. This intuitively makes sense: higher GDP/capita calls for a lower need for foreign aid and higher corruption levels deters foreign aid. The correlation coefficient sign for the log of population and foreign aid/GDP is less intuitive to interpret. We also notice the interesting strong and positive sign of the correlation coefficient between GDP per capita and the corruption index in the second column.

We analyze the correlation matrix for the within case as well. We perform the two analyses separately: grouping the between and within transformed variables in one same correlation matrix would provide little information, as the between and within variables are orthogonal by essence.

Table 5: Correlation matrix

	oda_net_gdp_cap	corruption	population	gdp_per_cap	g_gdp_per_cap	g_oda_net
oda_net_gdp_cap	1	.	.	.	.	.
corruption	0.03	1	.	.	.	.
population	−0.02	0.04	1	.	.	.
gdp_per_cap	−0.06	0.05	0.17	1	.	.
g_gdp_per_cap	0.04	0.01	−0.04	0.06	1	.
g_oda_net	0.24	−0.03	−0.02	−0.13	−0.05	1

From the Within correlation matrix, we can observe that the correlation among our variables is much lower than in the case of the between correlation matrix. This confirms what we saw with the graphs in the previous question. Only three correlations are greater than 0,1. The highest correlation concerns the growth rate of ODA with ODA/GDP.

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

Table 6: Fixed effects model with ODA / GNI as dependent variable

	Model 1	Model 2
log(population)	−5.511*	−13.874**
	(2.417)	(4.124)
log(gdp_per_cap)	−5.104**	−8.000**
	(1.742)	(2.607)
Num.Obs.	1386	1386
R2	0.761	0.770
R2 Adj.	0.747	0.753
R2 Within	0.143	0.090
Std.Errors	Clustered (country)	Clustered (country)
FE: country	X	X
FE: year		X

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The above regression confirms the results of the above Between-correlation matrix: the two variables “population” and “GDP per capita” are negatively correlation with the foreign aid variable. The addition of year fixed effects improves only very slightly the high R-squared and adjusted R-squared (by less than 1 percentage point).

The coefficients are highly significant, at the 1% level for most. Those findings are therefore consistents with the results of the descriptive statistics.

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

Table 7: Fixed effects model with Corruption index as dependent variable

	Model 1	Model 2
oda_gni	0.007*	0.007*
	(0.003)	(0.003)
log(gdp_per_cap)	0.346***	0.614***
	(0.095)	(0.115)
g_oda_net	−0.011	−0.010
	(0.015)	(0.014)
Num.Obs.	1309	1309
R2	0.932	0.937
R2 Adj.	0.928	0.932
R2 Within	0.088	0.141
Std.Errors	Clustered (country)	Clustered (country)
FE: country	X	X
FE: year		X

+ p $<$ 0.1, * p $<$ 0.05, ** p $<$ 0.01, *** p $<$ 0.001

As we noted in the descriptive statistics part of question 7, GDP per capita is positively correlated with the corruption variable, with a coefficient relatively larger in magnitude compared to the other two coefficients. The two foreign aid variables (as % of GDP and % growth) make nearly no contribution to the variance of the corruption index: the foreign aid growth variable is in fact not significant (as per the below table). Those results are therefore very much in line with the findings of the bivariate correlation matrix in question 7. We finally note that the inclusion of year fixed effect in model 2 does not add explanatory power to the model.

Table 8: Fixed effects model with the growth rate of GDP per capita as dependent variable

	Model 1	Model 2
oda_gni	0.001	0.001
	(0.001)	(0.001)
log(gdp_per_cap)	0.016	0.089**
	(0.017)	(0.028)
g_oda_net	−0.005	−0.005
	(0.005)	(0.004)
corruption	−0.003	−0.015
	(0.011)	(0.013)
Num.Obs.	1309	1309
R2	0.155	0.234
R2 Adj.	0.099	0.174
R2 Within	0.008	0.048
Std.Errors	Clustered (country)	Clustered (country)
FE: country	X	X
FE: year		X

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

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.

The above regression results suggest no meaningful statistical relationship between GDP/capita growth and the Log(GDP/capita), foreign aid/GDP, foreign aid growth and the Corruption index: coefficients are of small magnitude, especially compared to standard errors, and the goodness-of-fit metrics (R-squared and Adjusted R-squared) are low.

The above table further proves that none of the regressors is statistically significant, to the exception of the log(GDP per capita) in the country and year fixed effect regression, quite understandably.

11. Propose an additional interesting estimation using this database.

Starting again from a Fixed-Effects model, considering Corruption as the dependent variable, ODA / GNI, log of GDP per cap and the growth rate of ODA as independent variable and finally, as Fixed Effects, country plus year.

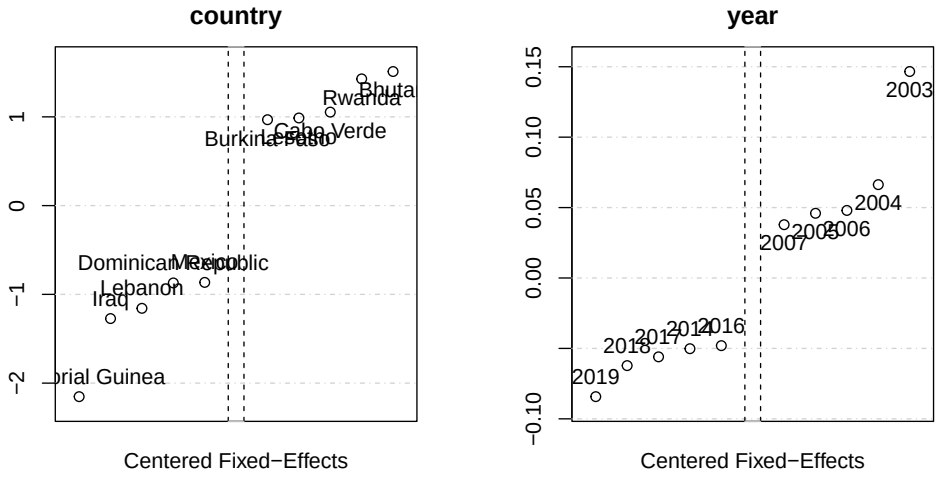
Table 9: Fixed effects model with Corruption index as dependent variable

	Model 1	Model 2
oda_gni	0.007*	0.007*
	(0.003)	(0.003)
log(gdp_per_cap)	0.346***	0.614***
	(0.095)	(0.115)
g_oda_net	−0.011	−0.010
	(0.015)	(0.014)
Num.Obs.	1309	1309
R2	0.932	0.937
R2 Adj.	0.928	0.932
R2 Within	0.088	0.141
Std.Errors	Clustered (country)	Clustered (country)
FE: country	X	X
FE: year		X

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

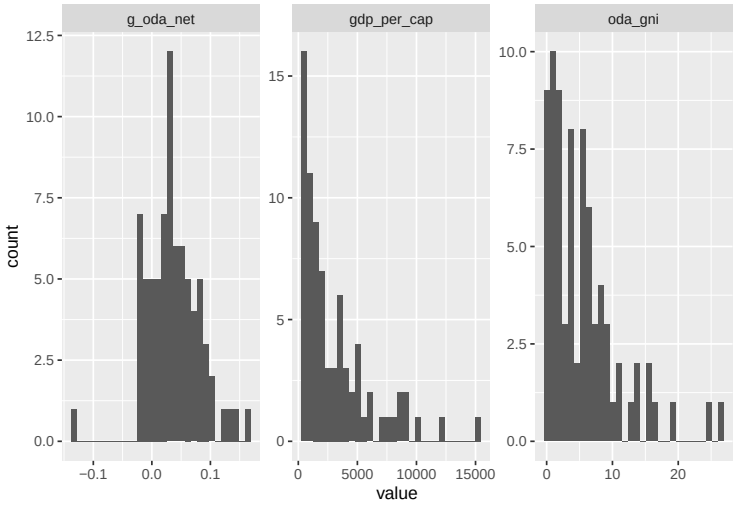
We can then analyse our fixed effects and see which countries and year are distributed around the mean for the corruption variable. A negative value means a higher level of corruption compared to the mean. Hence, Equatorial Guinea and Iraq have higher level of corruptions while Bhutan or the Rwanda have lower level of corruption.

Moreover, it seems that the level of corruptions is descreasing over time. The highest level of corruption concerns the year 2003, while 2019 is the year which have the lowest level of corruption.



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.
We begin by ploting the between transformation over the full period. The between growth rate of ODA can have negative values because we're considering Net ODA. Moreover, our between transformed variables are likely skewed.

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



And then, plotting the within transformation for *oda_gni, g_oda_net, gdp_per_cap* variables. Again, with the within transformation, it more likely that our variables are normally distributed around 0 even though, concerns about high kurtosis are there.