ETH
Swiss Federal Institute of Technology Zurich

Department of Mathematics

Semester Paper — Fall 2015

David Pham

Public Employment: Data Analysis with OECD Economic Outlook Quarterly Data

Submission Date: March 15th 2016

Co-Adviser: Prof. Norbert Hungerbühler
Adviser: Prof. Jan-Egbert Sturm

To Myriam for her support during the long nights of work and to our future children.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Since the start of the 2008 financial crisis, countries and governments heavily have relied on their respective central banks to boost their economies and support growth. Thanks to low interest rates, governments and their leaders could probably invest in the economy and provide employment to the population. They can provide it by several means: keeping a low corruption, creating an employment friendly environment and naturally employ professionals directly, the so-called public servants.

The share of civil servants with respect to total labor force varies over time, and several hypothesis exist to explain why. One of the recurring idea is that incumbent governments have incentives to increase the number of civil servants before elections to raise their chances of reelection and this would be easier in environment where fiscal transparency is low.

The purpose of this semester paper is to synthesize the current literature of the subject, and most importantly to gather and to prepare data for a replication of the results of the literature by using quarterly data from the OECD Economic Outlook.

The added value of this work is that a quarterly data set is used as input to study the problem. This would be for the first time, to the best of our knowledge. The main challenge to complete the work has been to gather the data, to tidy them and to merge them. Moreover, as the data set contains hundreds of observational variables, statistical analysis should be executed cautiously.

With modern technology, studies should be as transparent and reproducible as possible.

Hopefully, the Github repository[1] of the project offers full transparency over the data and the code used to handle them. The analysis is performed with the statistical environment R (R Core Team (2015)) and the main scripts to perform the analysis is supplemented in the appendix.

The structure of the semester paper is following: a first part is devoted to summarize the research of recent papers, then it focuses on the statistical analysis of the data, before concluding.

[1] https://github.com/davidpham87/public_employment_analysis

1.1 Theory

This section synthesizes three papers used as theoretical background to run our analysis. These are:

i.) Alesina, Baqir, and Easterly (2000), which studies American states about their election and in order to find if any inequality measures could predict the share of public servants.

ii.) Alt, Lassen, and Wehner (2014), that describes the effect of accounting practices of EU countries over public employment.

iii.) Aaskoven (2015), which analyzes the effect fiscal transparency on civil servants.

1.2 Theoretical models

Alesina et al. (2000) supposes that American cities use public employment as a discreet mean for wealth redistributing: It channels resources from middle class voters to disadvantaged citizens when an explicit tax-transfer scheme would not find political support. In order to justify their idea, they set the following theoretical framework: define a two periods time-frame, with an election after the first period. There are two classes of voters (*middle* class and the *poor*) and two contestants for the government. Each candidate need the support of the middle class in order to win the election. The challenge of the incumbent government is to know whether it should start a public project to employ people from the poor class, at the risk of losing political support from the middle class.

Define B as the benefit of a public project. This benefit can be thought of the employment of the *poor* to complete the project. Then restrict B to a discrete random variable with

$$B = \begin{cases} B_L & \text{with probability } 1 - \theta \\ B_H & \text{with probability } \theta \end{cases}$$

where $0 < B_L < B_H$ and θ is random variable taking either θ_L or θ_H with $0 < \theta_L < \theta_H < 1$. When $\theta = \theta_L$, it is more efficient to make a cash transfer than implementing the project. Intuitively, θ is the risk of the project. Moreover, the incumbent government observes the realisation of θ before deciding to implement a public project or not.

As there are two contestants for the government they also possess different preferences: one supports the middle class and conduct the public project only if $\theta = \theta_H$, whereas the other favors the poor, leading the public project for any value of θ as long as the latter action does not prevent the candidate from winning the next election. Voters do not know which type are the politicians, but they have prior believes: they are not completely certain whom the candidates support. Moreover, being reelected is always favored by the incumbent candidate as it maximizes her utility.

Under these conditions, Alesina et al. (2000) shows there exists an optimal decision for the incumbent government given θ and its preferences. With the appropriate data, the paper observes that several inequality measures are correlated with public employment shares, hence supporting their view.

Alt et al. (2014) and Aaskoven (2015) have a similar hypothesis: the incumbent government will boost the share of public employment as election dates are nearing in order to stay elected. They assert that the magnitude of these changes can be explained by the degree of fiscal transparency of the governments. Fiscal transparency is often measured measured by evaluating the quality and frequency of financial reports from a country. They assume that fiscal opacity allows governments to use unorthodox accounting methodology to please political partners, financial markets and voters. It also allows them to use use windfall revenues to employ more civil servants without voters noticing, even though if these would prefer to have a tax cut or a cash transfer. Furthermore, increasing the number public employees is a fast and easy process for the incumbent government, because it usually does not require a modification in the laws or the approval of the parliament.

1.3 Statistical Analysis

In practice, two challenges were encountered during the analysis: data interpolation and statistical analysis.

Data interpolation The Economic Outlook dataset from the OECD exists in annual and quarterly frequencies. Nevertheless, numerous variables are not measured quarterly and quarterly interpolation of annual data is a reasonable choice in order to retain information. However, economists often require for level variables (i.e. measure with units) that the sum of quarterly data to be equal to the annual value. Sax and Steiner (2014) proposes the Denton(-Chollette) method as a decent choice for performing such a task. Heuristically, this method (so-called temporal distribution or *disaggregation*) solves the problem by dividing and spreading the annual value into quarterly values, such that the interpolation looks smooth in the end. This process can be thought as a smart spline interpolation adjusted by a scaling factor. More sophisticated methodology might use correlated quarterly variables in order to adjust the annual ones, but this usually leads to over-fitting.

Data analysis The standard statistical methods in the literature are the regular multiple linear regression. The cited references use the a common model. Assume that Y_{it} denotes the public employment rate (number of civil servants over total labor force) for the i-th country and time t. Then Y_{it} is fitted as

$$Y_{it} = X_{it}\beta + \eta_i + \tau_t + \varepsilon_{it}, \quad i \in \{1, \dots, n\}, \; t \in \{1, \dots, T\}, \tag{1.1}$$

where $X_{it} \in \mathbb{R}^{n\times p}$ is a matrix of explanatory and control variables, $\beta \in \mathbb{R}^p$ is the regression coefficient, η_i is a country fixed-effect and τ_t is a time-fixed effect, ε_{it} are non-correlated centered gaussian random variables, n, respectively T, is the number of countries, respectively, period of observations. One weakness in the analysis of the literature is that assumptions of the multiple regression are seldom checked. This is problematic as for our data set the fitted values of ε_{it} and $\varepsilon_{i(t+1)}$ are highly correlated, contradicting the assumption. Additionally, statistical significance is often reported with raw p-values. Nonetheless, using these unprocessed p-values leads to a higher rate of false positive. Best practice recommend to adjust these by controlling the false discovery rate (see Benjamini and Hochberg (1995)).

Chapter 2

Empirical Analysis

This chapter is structured as such: a first section describes the data and its variables, their source and a reason to incorporate them in the analysis. This is followed by a presentation of the results and their analysis.

2.1 Data

The main data comes from the OECD Economic Outlook (98^{th} edition) (cf. OECD (2015)) with annual and quarterly frequency. We only use a subset of the data set with the following variables.

- Public employment rate of OECD countries between 1990 to 2012. Public employment rate is defined as the number of civil servant divided by the total labor force. Figure 2.1 shows the public employment rate used for the data analysis.
- Unemployment rate in percent. Everything else being equal, the correlation with public employment rate should be negative: the total labor force is composed of employed plus unemployed people, hence if the number of workers in the private industry remain stable, a diminishing unemployment rate should increase the public employment rate.
- Government revenues in percent of GDP: the share of GDP that the government receive from taxes and others sources of income. This variable captures the size of the government.

- Net lending in percent of GDP: the difference between revenues and expenditures scaled by GDP. The net lending captures how well the incumbent government manages its budget.

- GDP Growth in percent. It is believed that this variable represent the effect economic cycle and momentum in an economy.

- GDP per capita in 2010 USD Purchasing power parity, in order to control for the Wagner's Law, stating that richer populations care more about common goods (often provided by government). The effect of wealth over the public employment share should be non-linear, reason for which its log is used in the models.

Moreover, the following time series have been collected to test the assertion of the literature.

- The 105 election quarters for the relevant OECD countries were recorded from Wikipedia (Wikipedia (2015)); From the election dates, one can deduce the number of years until the normal end of the term.

- Fiscal transparency score, Wang, Irwin, and Murara (2015) from the IMF. A low fiscal transparency allows governments to use windfall revenues to boost the number of public employees or to adjust their national accounts.

- Political direction of the executive government (left or right) from the World Bank (Beck, Clarke, Groff, Keefer, and Walsh (2001)). One desires to capture the effect or correlation of the political partisanship over public employment rate. It should be increased when a left-wing governments is elected.

- Gini coefficient, before and after tax from the Standardizing the world income inequality database (Solt (2009)). The Gini coefficient is a standard measure to assess the income inequality within a country. A high level of inequality should lead to a bigger rate of public employees, as a mean for the government to redistribute wealth.

Plots of these variable are showed in Appendix A. We also assumed that structural breaks could have affected the public employment rate in the data set. James and Matteson (2014) provides a non-parametric algorithm to detect such point. As Figure 2.2 depicts, although there are some individual changes, there are no overall break at given time point. Hence the idea has been abandoned.

2.2 Analysis

2.2.1 Results

For our analysis, we extend the model from Equation (1.1) with the following equation:

$$Y_{it} = \alpha Y_{it-1} + X_{it}\beta + \eta_i + \tau_t + \varepsilon_{it} \quad i \in \{1, \ldots, n\},\ t \in \{1, \ldots, T\},$$

where X_{it} is the independent variable and contains the unemployment rate, government revenues and net lending for i-th country at time t. The other variables are similar to Equation (1.1). This model, albeit counter-intuitive, is stable and robust. Additionally, the auto-regressive aspect reduces the correlation of the error terms ε_{it}. Table 2.1 shows the coefficient of the regression[1]. Moreover, Figure 2.2.2 provides some support about the soundness of our statistical model: the residuals of the regression seem to be uncorrelated, although there is evidence they do not follow a Gaussian distribution. The regression tables for the additional variables are in the Appendix B. Note the following supplementary observations.

i.) The number of fitted parameter is much bigger than what is displayed in the tables: one degree of freedom is allocated for each quarter to fit τ_t (88 for the 22 years of observations) and for the country parameter η_i (17 parameters).

ii.) The main variables (unemployment rate, government revenue and net lending) keep the same regression coefficient and their statistical significance.

iii.) The absolute size of these coefficients are quite small.

iv.) GDP growth and the Gini coefficient after taxes have significant regression coefficient before adjustment of the p-values, but become insignificant afterwards.

v.) The effect of the wealth of a nation, measured by the GDP per capita, is unexpectedly not correlated. This observation does not support the Wagner's law.

vi.) Unfortunately, the number of years before the next official election, the left-wing partisanship, the fiscal transparency do not seem to influence the public employment rate, contradicting literature.

[1] The output of linear models has produced with the Stargazer package (Hlavac (2015)) used with the statistical programming language R.

2.2.2 Robustness analysis

The coefficients of the base model remain stable and significant when additional indepedent variables are added. The same holds when the frequency is lowered to annual data as well. Furthermore, interaction terms have also been introduced in the robustness tests, without creating major modifications to our previous finding.

Note that for the robustness analysis, missing data lead to a different number of observations as input for the model, making the comparison more difficult.

One of the additional difficulty in this analysis is that country means are sufficient to predict the values of the public employment rate: Variations are small with a high auto-correlation coefficient. In order to tackle this problem, we tried to fit the time-difference of the public employment rate, but the explained variance (commonly known as the R^2) is negligible (about 9% with an over-fitting model).

In short, unemployment rate, government revenues and net lending compose a statistically sound and robust model to explain the public employment rate. Unfortunately, according to the data, other variables, such as fiscal transparency, the remaining years until the end of the government mandate or measure of inequality do not offer additional information.

Table 2.1: Result of linear regression of the main model. One observes that government revenue and net lending are not significant without unemployment rate.

	Dependent variable:			
	Public employment rate			
	(1)	(2)	(3)	(4)
Unemployment rate	-0.009^{***}	-0.006^{***}		
	(0.002)	(0.002)		
Government revenue	0.008^{***}		0.002	
	(0.003)		(0.002)	
Net Lending in percent of GDP	-0.005^{***}			-0.002
	(0.001)			(0.001)
Constant	0.585^{***}	0.771^{***}	0.658^{***}	0.728^{***}
	(0.130)	(0.103)	(0.129)	(0.102)
Auto-correlation effect	Yes	Yes	Yes	Yes
Time effect	Yes	Yes	Yes	Yes
Country effect	Yes	Yes	Yes	Yes
Observations	1,579	1,592	1,579	1,579
R^2	0.999	0.999	0.999	0.999
Adjusted R^2	0.999	0.999	0.999	0.999
Residual Std. Error	0.145 (df = 1466)	0.146 (df = 1481)	0.146 (df = 1468)	0.146 (df = 1468)

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

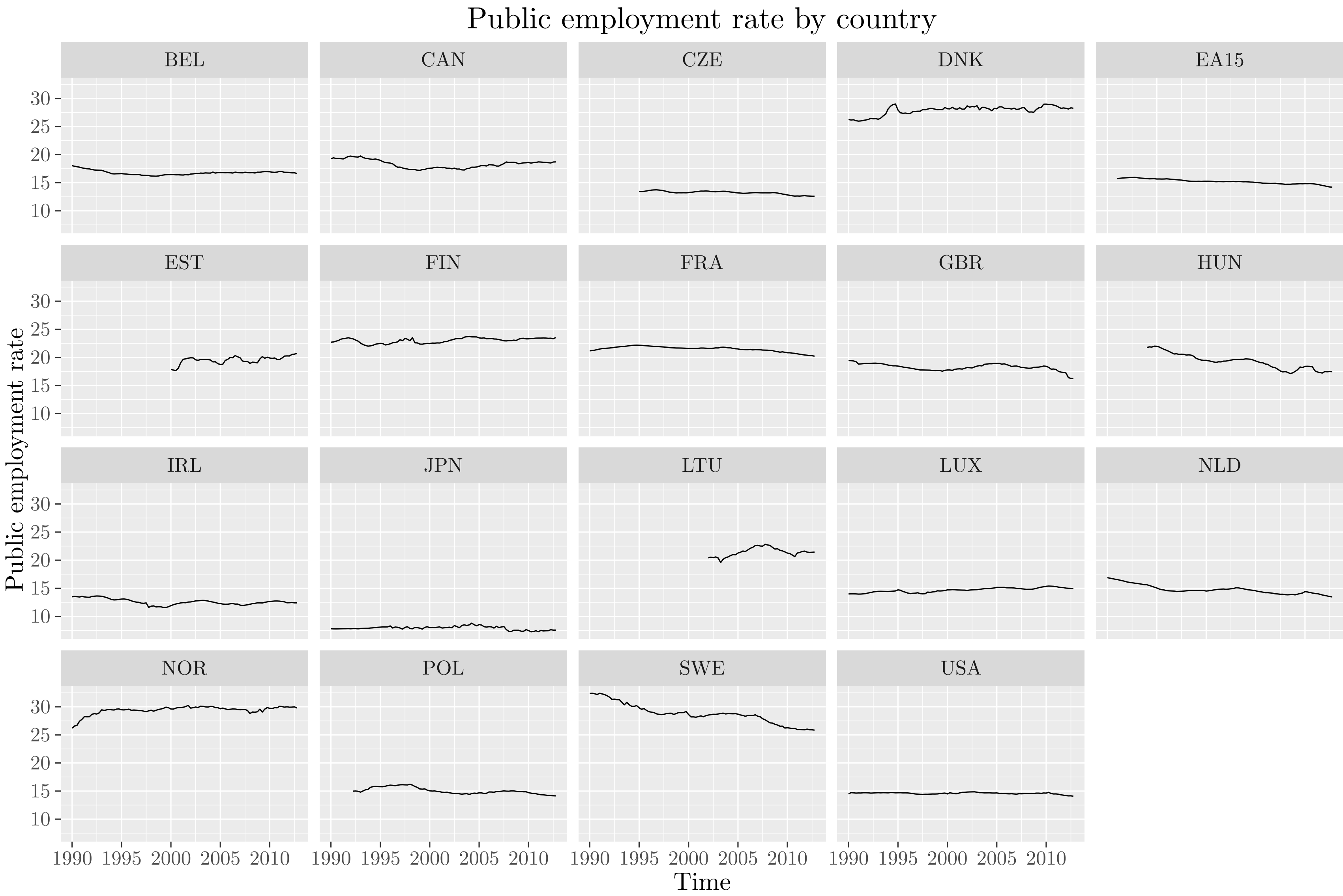

Figure 2.1: Ratio of civil servants over total labor force, in the OECD data set. Note that the curves are quite stable in the last 20 years.

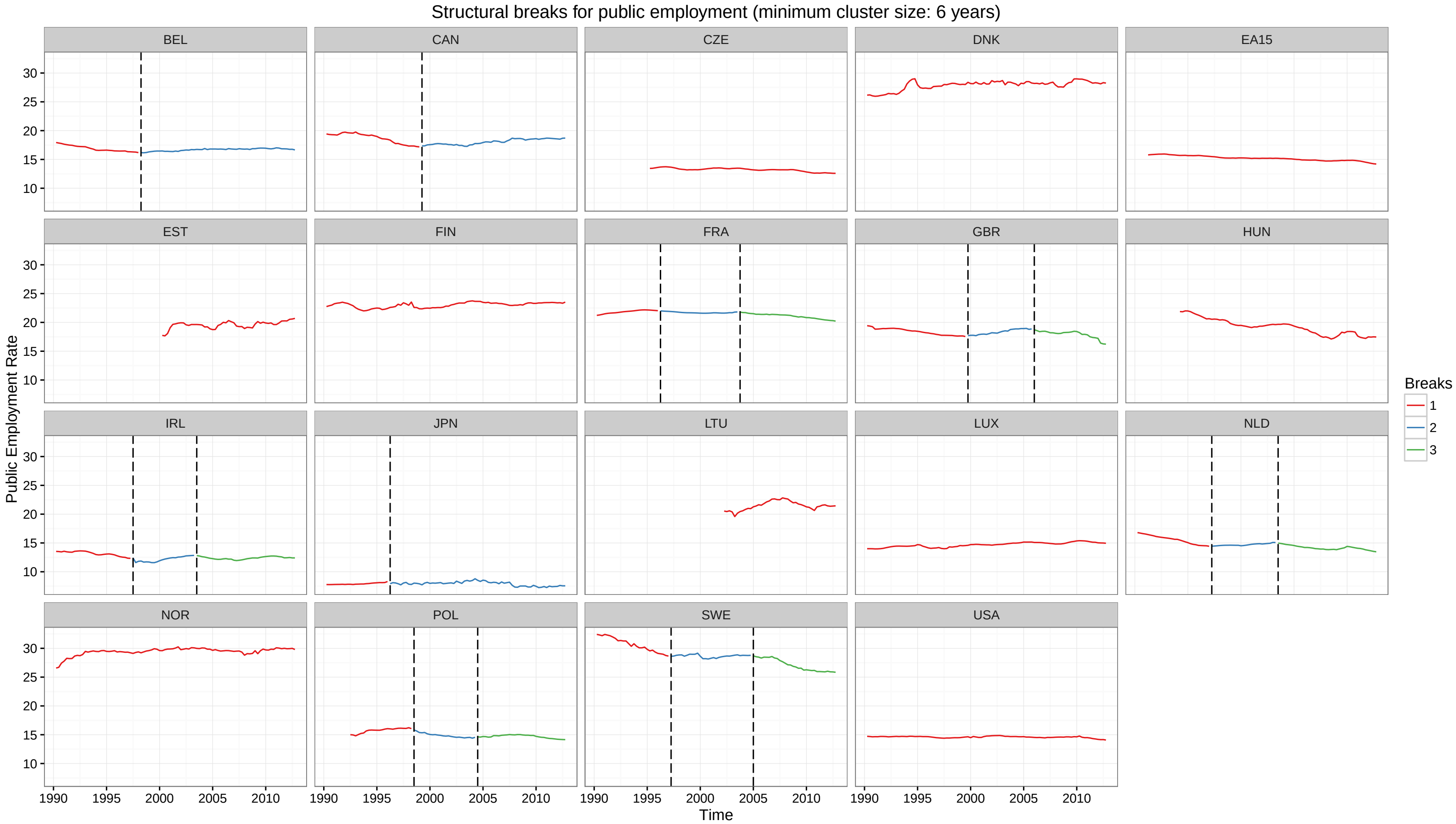

Figure 2.2: Structural breaks of the public employment rate using a non-parametric estimation of the breaks. Although there seems to have some breaks in individual countries, these do not justify a split in the analysis of the whole data set.

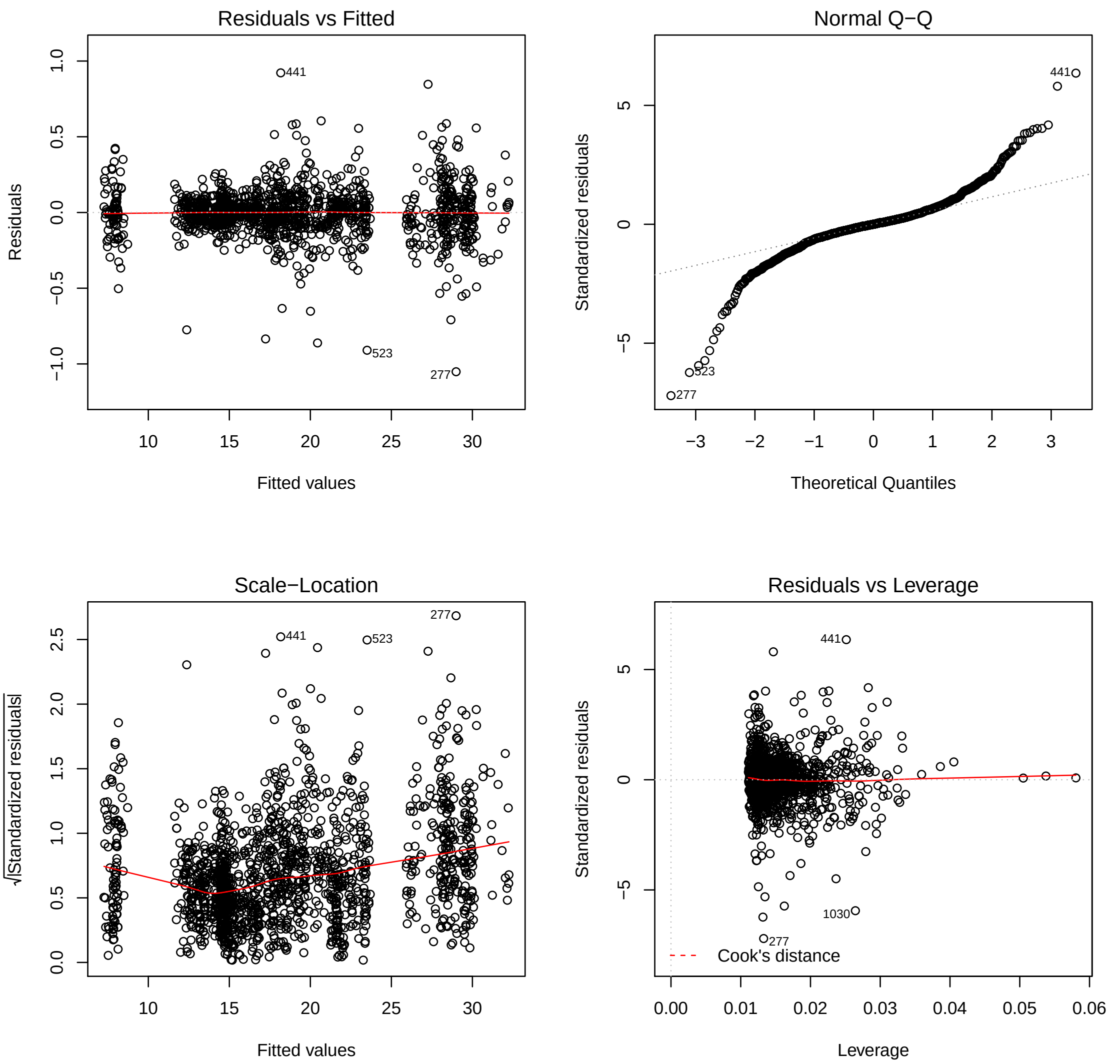

Figure 2.3: Diagnostic plots for the base line model. On the upper-left plot, one observes that the residuals are not uniformly distributed on the fitted values. The upper-right plot shows that the residuals are probably not distributed as a Gaussian variable as well. Hence some care should be taken with the assumption of the model and its output.

Chapter 3

Conclusion

This semester paper quickly summarizes the current literature on the theory of civil servants and tries to find a reasonable model which could explain quarterly variation of the share of public employees in the labor force. With analysis on annual data, it is believed that the level of wealth, the fiscal transparency, the government political partisanship, the proximity of election terms or the degree of inequality would correlate with the public employment rate. With quarterly frequency, data suggest that the influence of these variables on the public employment rate might be weak. Even though our analysis showed some significant unadjusted p-values with these variables, these become statistically insignificant after correction. These results do not support the conclusions of the existing literature. Nevertheless, one can not exclude that the adjustment might have been too sharp or that some observations of the data were imprecise.

In order to test these ideas with quarterly frequency, the OECD Economic Outlook quarterly data set has been retrieved and many missing variables were interpolated from annual data from the same data set or alternative sources, leading to potential mistakes.

In order to confirm or invalidate this conclusion, one could replicate the study in order to check the consistency of the data, restrict the number of variables at the beginning of the study in order to increase the power of the statistical test, and maybe use more precise methods for the interpolation in order to capture undetected signals. The challenge of over-fitting and overstating statistical significance remains and one should probably stick to simple statistical models.

In order to ease the results, the data and programming scripts are shared on Github[1]. References should provide enough indication concerning the technical and the economical background.

Finally, the base model with unemployment rate, government revenues and net lending seems to be surprisingly consistent to explain public employment share with quarterly data from the OECD. This is a deceptive result as it is difficult to see why these variables would be correlated with the share of civil servants according to the literature.

[1] https://github.com/davidpham87/public_employment_analysis

Bibliography

Aaskoven, L. (2015). Fiscal transparency and public employment: Evidence from the oecd. *Journal of Political Economy, submitted*.

Alesina, A., R. Baqir, and W. Easterly (2000). Redistributive public employment. *Journal of Urban Economics 48*(2), 219–241.

Alt, J., D. D. Lassen, and J. Wehner (2014). It isn't just about greece: domestic politics, transparency and fiscal gimmickry in europe. *British Journal of Political Science 44*(04), 707–716.

Beck, T., G. Clarke, A. Groff, P. Keefer, and P. Walsh (2001). New tools in comparative political economy: The database of political institutions. *the world bank economic review 15*(1), 165–176.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Hlavac, M. (2015). *stargazer: Well-Formatted Regression and Summary Statistics Tables.* Cambridge, USA: Harvard University. R package version 5.2.

James, N. A. and D. S. Matteson (2014). ecp: An R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software 62*(7), 1–25.

OECD (2015). Economic outlook.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Sax, C. and P. Steiner (2014). *tempdisagg: Methods for Temporal Disaggregation and Interpolation of Time Series.* R package version 0.24.0.

Solt, F. (2009). Standardizing the world income inequality database. *Social Science Quarterly 90*(2), 231–242.

Wang, R. F., M. T. Irwin, and L. K. Murara (2015). *Trends in Fiscal Transparency: Evidence from a New Database of the Coverage of Fiscal Reporting.* Number 15-188. International Monetary Fund.

Wikipedia (2015). Wikipedia, the free encyclopedia. [Online; accessed 2015-12-31].

Appendix A

Data visualizations

The purpose of the following graphics is to offer some visual checks in order to detect any anomalies in the data set.

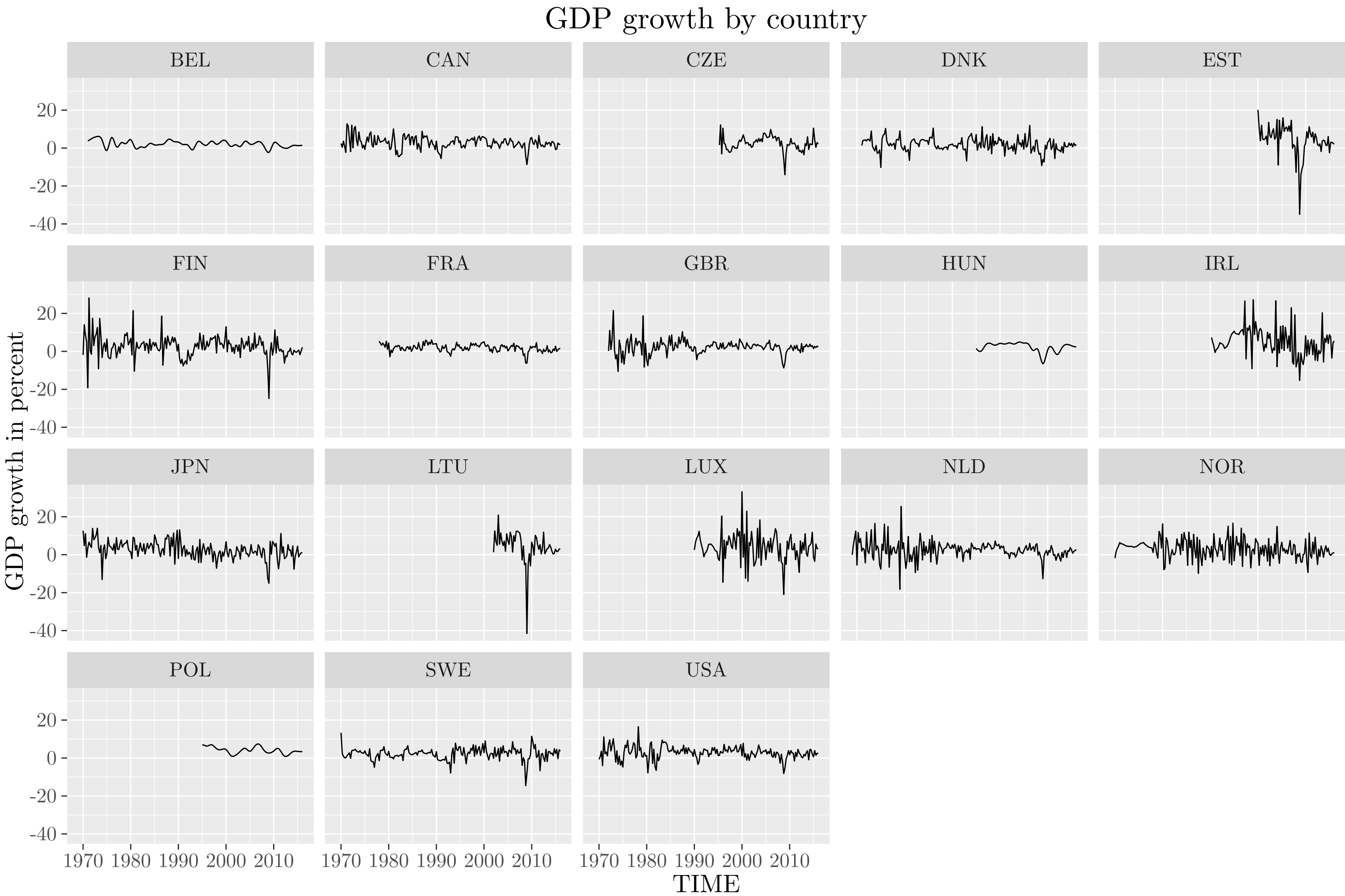

Figure A.1: GDP growth, volume

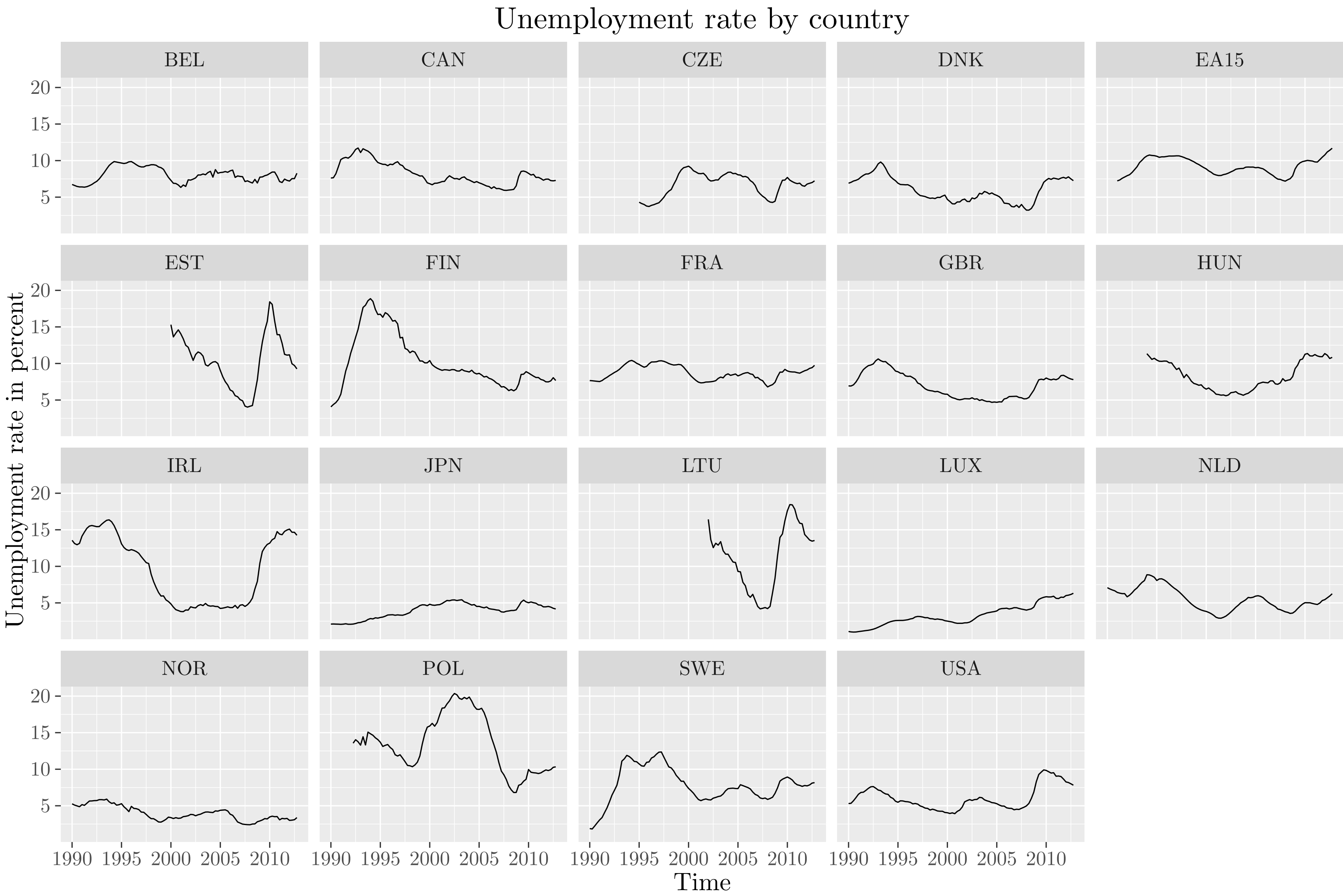

Figure A.2: Unemployment rate

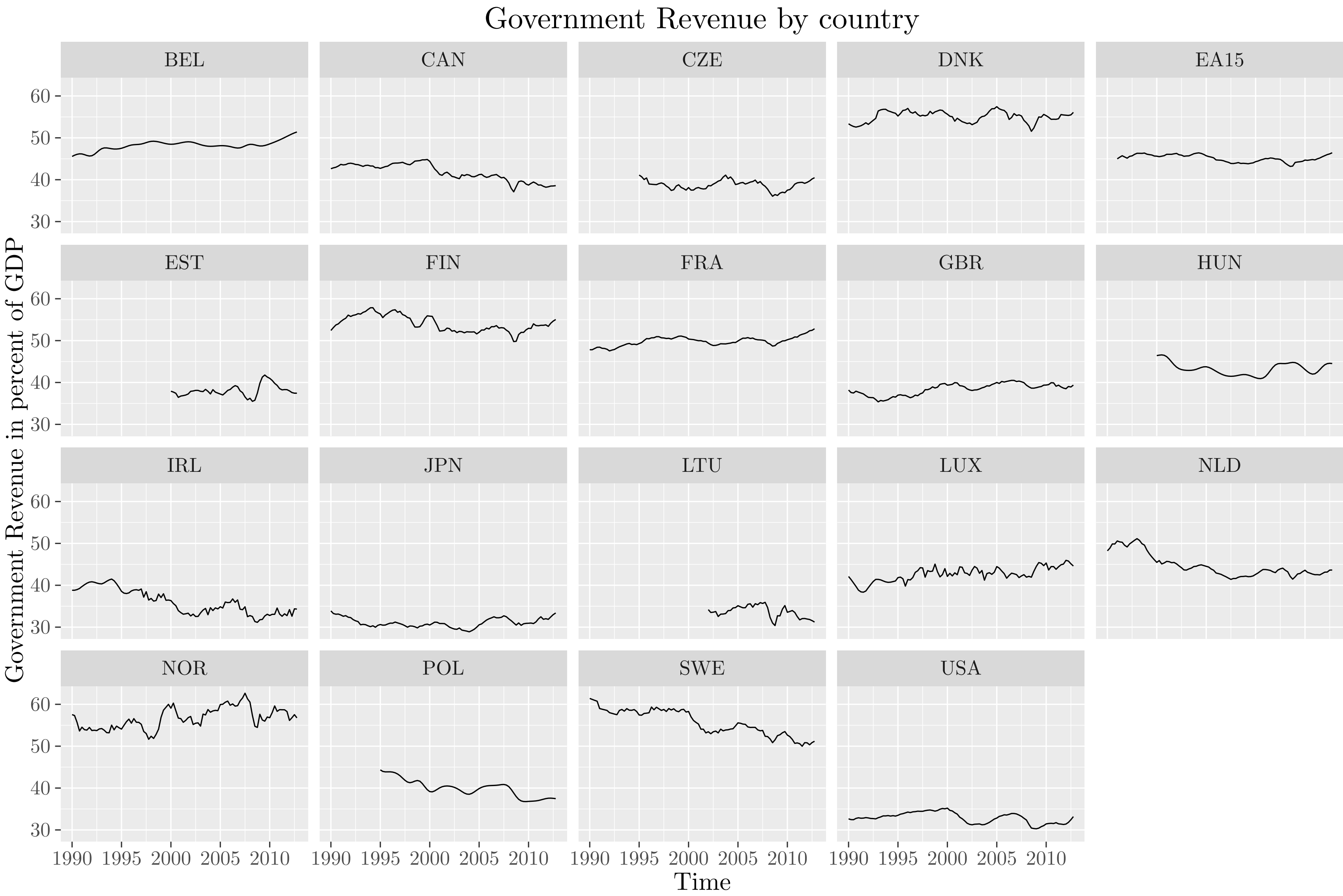

Figure A.3: Government revenue in percent of GDP

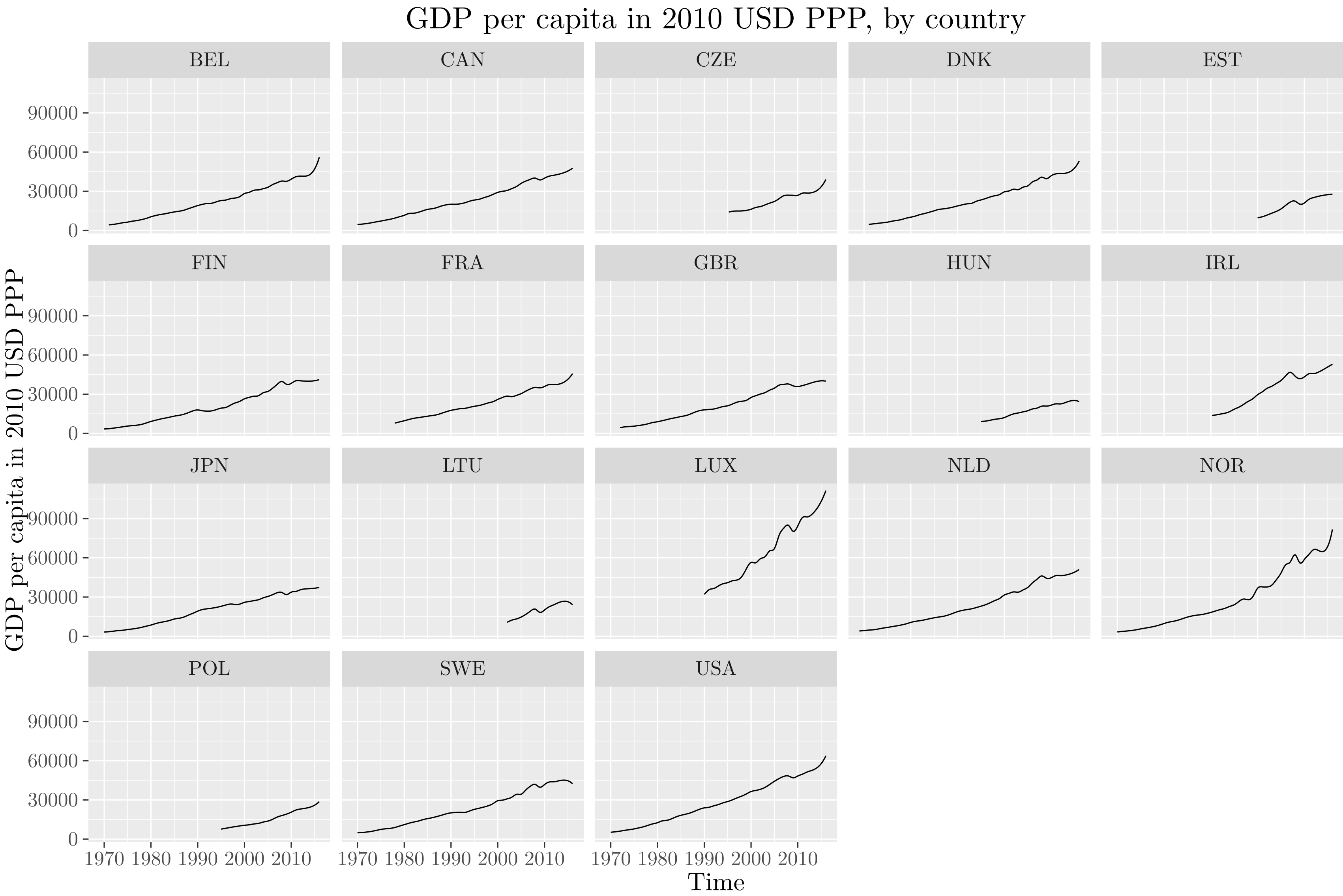

Figure A.4: GDP per Capita in USD millions

Figure A.5: Net lending in percent of GDP

Appendix B

Regression tables

This part provides the regression tables of the many regressions used to test the different variables and to check the robustness of the base model, which includes the unemployment rate, government revenues and net lending as main variables.

Table B.1: Effect of GDP

	Dependent variable:	
	Public employment rate	
	(1)	(2)
Government Revenue	0.008***	0.008***
	(0.003)	(0.003)
Net Lending in percent of GDP	−0.006***	−0.005***
	(0.001)	(0.001)
Unemployment rate	−0.009***	−0.009***
	(0.002)	(0.002)
GDP growth, YoY in percent	0.004**	
	(0.002)	
Constant	0.517***	0.584***
	(0.133)	(0.131)
Auto-correlation effect	Yes	Yes
Time effect	Yes	Yes
Country effect	Yes	Yes
Observations	1,570	1,570
R^2	0.999	0.999
Adjusted R^2	0.999	0.999
Residual Std. Error	0.146 (df = 1456)	0.146 (df = 1457)
F Statistic	22,388.310*** (df = 113; 1456)	22,512.150*** (df = 112; 1457)

Note: *p<0.1; **p<0.05; ***p<0.01

Table B.2: Effect of GDP per Capita

	Dependent variable:	
	Public employment rate	
	(1)	(2)
Government Revenue	0.006**	0.006**
	(0.003)	(0.003)
Net Lending in percent of GDP	−0.004***	−0.004***
	(0.001)	(0.001)
Unemployment rate	−0.009***	−0.007***
	(0.002)	(0.002)
Log of GDP per capita, in USD Millions	−0.069	
	(0.070)	
Constant	1.330*	0.586***
	(0.765)	(0.131)
Auto-correlation effect	Yes	Yes
Time effect	Yes	Yes
Country effect	Yes	Yes
Observations	1,449	1,449
R^2	0.999	0.999
Adjusted R^2	0.999	0.999
Residual Std. Error	0.145 (df = 1337)	0.145 (df = 1338)
F Statistic	22,321.440*** (df = 111; 1337)	22,524.810*** (df = 110; 1338)

Note: *p<0.1; **p<0.05; ***p<0.01

Table B.3: Effect of Years until next Election

	Dependent variable:	
	Public employment rate	
	(1)	(2)
Unemployment rate	−0.009***	−0.009***
	(0.002)	(0.002)
Government Revenue	0.008***	0.008***
	(0.003)	(0.003)
Net Lending in percent of GDP	−0.005***	−0.005***
	(0.002)	(0.002)
Years until next election	−0.004	
	(0.003)	
Constant	0.595***	0.583***
	(0.134)	(0.134)
Auto-correlation effect	Yes	Yes
Time effect	Yes	Yes
Country effect	Yes	Yes
Observations	1,492	1,492
R^2	0.999	0.999
Adjusted R^2	0.999	0.999
Residual Std. Error	0.149 (df = 1379)	0.149 (df = 1380)
F Statistic	21,095.540*** (df = 112; 1379)	21,281.620*** (df = 111; 1380)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table B.4: Effect of IMF GFS Score

	Dependent variable:	
	Public employment rate	
	(1)	(2)
Unemployment rate	−0.004	−0.004
	(0.003)	(0.003)
Government Revenue	0.011**	0.011**
	(0.005)	(0.005)
Net Lending in percent of GDP	−0.004*	−0.004*
	(0.002)	(0.002)
IMF GFS Index	−0.0001	
	(0.0004)	
Constant	0.221	0.230
	(0.328)	(0.325)
Auto-correlation effect	Yes	Yes
Time effect	Yes	Yes
Country effect	Yes	Yes
Observations	680	680
R^2	0.999	0.999
Adjusted R^2	0.999	0.999
Residual Std. Error	0.144 (df = 619)	0.144 (df = 620)
F Statistic	18,739.620*** (df = 60; 619)	19,085.990*** (df = 59; 620)

Note: *p<0.1; **p<0.05; ***p<0.01

Table B.5: Effect of Government Political Orientation

	Dependent variable:	
	Public employment rate	
	(1)	(2)
Unemployment rate	-0.009^{***}	-0.009^{***}
	(0.002)	(0.002)
Government Revenue	0.007^{***}	0.008^{***}
	(0.003)	(0.003)
Net Lending in percent of GDP	-0.005^{***}	-0.005^{***}
	(0.002)	(0.002)
Government with a left-wing partisanship	0.014	
	(0.011)	
Constant	0.601^{***}	0.583^{***}
	(0.134)	(0.134)
Auto-correlation effect	Yes	Yes
Time effect	Yes	Yes
Country effect	Yes	Yes
Observations	1,492	1,492
R^2	0.999	0.999
Adjusted R^2	0.999	0.999
Residual Std. Error	0.149 (df = 1379)	0.149 (df = 1380)
F Statistic	$21{,}103.400^{***}$ (df = 112; 1379)	$21{,}281.620^{***}$ (df = 111; 1380)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table B.6: Effect of Gini coefficient, data up to 2010 (included)

	Dependent variable:	
	Public employment rate	
	(1)	(2)
Unemployment rate	−0.012***	−0.012***
	(0.003)	(0.002)
Government Revenue	0.013***	0.012***
	(0.003)	(0.003)
Net Lending in percent of GDP	−0.008***	−0.007***
	(0.002)	(0.002)
Gini Coefficient, Market Income	−0.005	
	(0.003)	
Gini Coefficient, Net Income	0.010**	
	(0.005)	
Constant	0.539***	0.617***
	(0.178)	(0.152)
Auto-correlation effect	Yes	Yes
Time effect	Yes	Yes
Country effect	Yes	Yes
Observations	1,276	1,276
R^2	0.999	0.999
Adjusted R^2	0.999	0.999
Residual Std. Error	0.151 (df = 1174)	0.151 (df = 1176)
F Statistic	19,893.780*** (df = 101; 1174)	20,255.620*** (df = 99; 1176)
Note:		*p<0.1; **p<0.05; ***p<0.01

Appendix C

R Implementation Details

The complete R code to generate the results is stored on https://github.com/davidpham87/public_employment_analysis where the procedures are more documented. The first script provides some functions useful for the analysis, the second displays how the data were manipulated to create the data matrix, whereas the third is used to produce the results.

Initialization script

```
PCKGS <- c('DescTools',
           'data.table',
           'readstata13',
           'magrittr',
           'ggplot2',
           'tikzDevice',
           'stargazer',
           'rsdmx',
           'softImpute',
           'rpart',
           'glmnet',
           'parallel',
           'plm',
           'plotly',
           'tempdisagg')

loadPackages <- function(){
  lapply(PCKGS, require, character.only=TRUE)
}

###########################################################################
### Utilities

joinDataTable <- function(lDT, kx=c('LOCATION', 'TIME')){
  Reduce(function(x, y) merge(x, y, all=TRUE), lDT)
}


unscale <- function(s) {
```

```
  s * attr(s, 'scaled:scale') + attr(s, 'scaled:center')
}

VIF ← function(data){
  vif.unique ← . %>% {reformulate('.', .)} %>% lm(data) %>% summary %$%
  adj.r.squared %>% {1/(1-.)}
  cn ← colnames(data)
  f ← function(s) tryCatch(vif.unique(s), error=function(error) NaN)
  res ← mclapply(cn, f, mc.cores=8)
  names(res) ← cn
  res
}

missingRatePerColumn ← function(x, p=0){
  missing.rate ← vapply(x, function(y) mean(is.na(y)), 0.0) %>%
    sort(TRUE) %>% round(4) %>% {Filter(function(x) x ≥ p, .)}
  missing.rate
}

unselect ← function(data, cols){
  new.cols ← Filter(function(x) !x %in% cols, colnames(data))
  subset(data, select=new.cols)
}

unselectVector ← function(x, kx) x[!x %in% kx]

butlast ← function(x, k=1) x[1:(length(x)-k)]

fctr2num ← function(x) as.numeric(levels(x)[x])

robustnessAnalysis ← function(data, cols, to.drop, formula=egr ~ .){
  cols.extended ← unselectVector(cols, to.drop)
  x.lm ← lm(formula, data[, cols.extended, with=FALSE])
  print(summary(x.lm))
  x.lm
}


###########################################################################
### Robustness Analysis

##' Useful this when doing robustness analysis (including excluding variable)
completeLmData ← function(lm.model, DT, new.cols){
  new.DT ← as.data.table(lm.model$model)
  setkey(new.DT, TIME, country)
  new.DT ← merge(new.DT, x.new[, c('country', 'TIME', new.cols), with=FALSE],
                 by=c('TIME', 'country'), all.x=TRUE)
}


###########################################################################
### Imputation functions

scaleNumeric ← function(x){
  if (mode(x) == 'numeric'){
```

```
    scale(x)
  } else {
    x
  }
}


str2fctrs ← function(dataset){
  fctrs ← sapply(1:ncol(dataset), function(jdx)
    any(c("string", "character") %in% class(dataset[1, jdx])))
  dataset[, fctrs] ← lapply(dataset[, fctrs, drop=FALSE], as.factor)
  dataset
}

imputeDataMi ← function(dataset, n, column.type.mi=NULL, ...){
  args ← list(...)
  valid.column.type ← c("unordered-categorical", "ordered-categorical",
                        "binary", "interval", "continuous", "count",
                        "irrelevant")

  ## check that the modification are valid
  if (!is.null(column.type.mi)){
    stopifnot(all(vapply(column.type.mi, is.element, TRUE,
                        set=valid.column.type)))
  } else {
    column.type.mi ← list()
  }

  mdf ← missing_data.frame(dataset) # missing data.frame

  for (k in names(column.type.mi)){
    mdf ← change(mdf, y=k, what="type", to=column.type.mi[[k]])
  }

  imputations ← do.call(mi, c(list(mdf, n.iter=30, n.chains=4), args))
  data.mi ←  mi::complete(imputations, n) # creates 20 different versions of imputations

  ## mi append columns providing the stating the missingnes, so we have to delete them
  data.mi ← lapply(data.mi,  function(df) df[, 1:ncol(dataset)]) # restrict the number of columns
  return(data.mi)
}

imputeDataSoftImpute ← function(dataset, ...){

  args ← list(...)
  is.null.args ← length(args) == 1 & is.null(args[[1]])

  ## boolean vectors stating factors columns
  fctrs ← sapply(1:ncol(dataset), function(jdx)
    any(c("factor", "string") %in% class(dataset[1, jdx])))
  lvls ← lapply(dataset[, fctrs], levels)

```

```
  dataset[, fctrs] ← lapply(dataset[, fctrs], as.numeric)
  x ← as.matrix(dataset)

  fit ← if (is.null.args){
    do.call(softImpute::softImpute, c(list(x)))
  } else {
    do.call(softImpute::softImpute, c(list(x), args))
  }

  dataset ← as.data.frame(softImpute::complete(x, fit))

  ## Correct the factors
  f ← function(s) {
    cut(round(dataset[, s]), c(0, seq_along(lvls[[s]])),
        labels=lvls[[s]], include.lowest=TRUE)
  }

  dataset[, fctrs] ← lapply(names(lvls), f)
  list(dataset)
}

############################################################################
### Quarterly Functions

##' Expects annual regular data with t being the year
interpolateQuarter ← function(t, y, max.time=2016,
                              method.interpolation=NULL){
  data ← na.omit(data.table(t, y))
  tout ← t
  t.idx ← data[, tout ≥ min(t) & tout ≤ max.time]
  data ← data[t ≤ max.time]
  tout ← tout[t.idx]
  yout.index ← data$t
  method.interpolation ← if (is.null(method.interpolation)) 'spline' else
      method.interpolation
  tryCatch({
    if (method.interpolation == 'spline'){
      yout ← spline(data$t, data$y, xout=tout)$y
    }

    if (method.interpolation == 'denton-cholette'){
      y.ts ← ts(data$y, start=min(data$t))
      y.td ← td(y.ts ~ 1, conversion='sum', to='quarterly',
                method=method.interpolation)
      y.predict ← predict(y.td)
      yout ← as.numeric(y.predict) # max time not considered here. FIXME
      yout.index ← index(y.predict)
    }

    if (method.interpolation %in% c('locf', 'fill-forward')){
      yout ← zoo::na.locf(y)
    }

    na.size.before ← length(t[t<min(yout.index)])
    na.size.after ← length(t) - na.size.before - length(yout)
```

```r
    res <- c(rep(NA, times=na.size.before),
             yout,
             rep(NA, times=na.size.after))

    return(res)
  }, error = function(e) y)
}


interpolateQuarterColumn <- function(eo.q, eo.a, col, max.time,
                                     method.interpolation=NULL){
  setkey(eo.a, country, TIME)
  setkey(eo.q, country, TIME)

  col.new <- paste0(col, '_annual_data')
  col.q <- paste0(col, '_interpolated')
  tryCatch(setnames(eo.a, col, col.new), error=function(e) NA)
  eo.q <- merge(eo.q,
                eo.a[, c('country', 'TIME', col.new), with=FALSE],
                all.x=TRUE)
  eo.q[, (col.q):=interpolateQuarter(TIME, get(col.new), get('max.time'),
                                     get('method.interpolation')),
       by='country']
  eo.q[, (col.new):=NULL]
  eo.q
}


################################################################################
### Plots

##' Shortcut to compare to variable in data.table x
compareValue <- function(x, ...){
  require(plotly)
  argx <- unlist(list(...))
  plot.data <- melt(x[, c('TIME', 'country', argx), with=FALSE],
                    id.vars=c('TIME', 'country'))
  gg <- ggplot(plot.data, aes(TIME, value)) +
    geom_line(aes(color=variable)) +
    facet_wrap(~country)
  print(ggplotly(gg))
  gg
}

showdiag <- function(lm.obj){
  par(mfrow = c(2, 2))
  plot(lm.obj)
}
```

Data manipulation script

```r
source('init.R')
pckgs <- loadPackages()
```

```
library(magrittr)
library(lattice)

MAKE_PLOTS ← TRUE
MAX_YEAR_EXTRAPOLATION ← 2014

cols ← c(# 'egr_diff', # public employement rate
  'gdpvd',
  'gdp_per_capita',
  'gdpv_yoy_annpct', # gdp growth
  'QUARTER',
  'unr',
  'population_interpolated',
  'government_revenue',# yrg over gdpvd
  'openness',
  'wage_share',
  'nlg_to_gdp' # net landing in % of gdp
)

cols.to.save ← c(
  cols,
  'egr',
  'country',
  'YEAR',
  'fiscal_transparency_interpolated', # 'imf_gfs_scores'
  'gini_market_interpolated',
  'gini_net_interpolated',
  'gini_red_abs',
  'gini_red_rel',
  'gap_interpolated',
  'gaplfp_interpolated',
  'left',
  'govfrac',
  'yrcurnt', # year until next election
  'is_election_date', # If the quarter is an election quarter
  "natural_ressource_rent",
  "revenueindex_interpolated",
  "employmentindex_interpolated",
  "regulationindex_interpolated",
  "subsidisationindex_interpolated",
  'muni_interpolated',
  'state_interpolated',
  'author_interpolated',
  "auton_interpolated",
  "self_employment_rate"
  ) %>% sort %>% {c('TIME', .)}


###########################################################################
## Load data

eos ← readRDS('../data/eo-data.rds')
eo.desc ← readRDS('../data/eo-colnames-dt.rds')
setkey(eo.desc, VARIABLE) # enamble eo.desc['bsii'] => Balance of income,
    value, BOP basis
```

```r
eos[[2]][ , list(country, eg)] %>% na.omit %>% {unique(.$country)} -> country.q # get non.missing country

##########################################################################
## Splines for interpoalting between years

## cols.interpolation.denton.cholette ←
##   c('yrg', 'nlg', 'xgs', 'mgs', 'gdp', 'gdpv', 'gdpvd', 'wage')
## cols.interpolation.splines ←
##   c('unr', 'gap', 'gaplfp', 'es', 'lf')


cols.interpolation.denton.cholette ← c()
cols.interpolation.splines ←
  c(c('yrg', 'nlg', 'xgs', 'mgs', 'gdp', 'gdpv', 'gdpvd', 'wage'),
    c('unr', 'gap', 'gaplfp', 'es'))

eo.a ← copy(eos[[1]])
eo.q ← copy(eos[[2]])

## pdf('plot/quarterly_vs_annual_levels', 12, 8)
## xyplot(yrg ~ TIME | country, eo.a[country=='USA'], type='l', main='YRG quarterly')
## xyplot(yrg ~ TIME | country, eo.q[country=='USA'], type='l', main='YRG annual')
## dev.off()

## NOTE: the function add a _interpolated at the end of the variables in
## cols.interpolation

eo.q ← Reduce(
  function(x, y) interpolateQuarterColumn(x, eo.a, y, MAX_YEAR_EXTRAPOLATION, 'denton-cholette'),
  cols.interpolation.denton.cholette, init=eo.q)

## TODO debug denton chollette

eo.q ← Reduce(
  function(x, y) interpolateQuarterColumn(x, eo.a, y, MAX_YEAR_EXTRAPOLATION),
  cols.interpolation.splines, init=eo.q)

##########################################################################
## Patching quarterly data

for (col in c(cols.interpolation.denton.cholette, cols.interpolation.splines)) {
  try({
    eo.q[is.na(get(col)), (col):= get(paste0(col, '_interpolated'))]
  })
}

eo.q[, gdpv_yoy_annpct:=c(NA, NA, NA, NA,
                          100*gdpv[-(1:4)]/butlast(gdpv, 4)-100), by='country']

##########################################################################
```

```r
## New Data
## Gathering of of additional data in order to make robustness analysis.
## The data are not a part of the oecd economic outlook data set.
## SWIID provides measures of gini

new.data.names <- new.data <-
  c('population', 'imf_gfs_scores', 'SWIID', 'wdi_rest_federalism')

# SWIID

new.data %<>% {paste0('../data/', ., '_cleaned.csv')} %>% lapply(fread) %>%
  lapply(function(dt) {
    dt[, V1:=NULL]
    setnames(dt, colnames(dt), tolower(colnames(dt)))
    setnames(dt, 'time', 'TIME')
    dt[, TIME:=as.numeric(TIME)]
    setkeyv(dt, c('location', 'TIME'))}) %>% joinDataTable

setnames(new.data, 'location', 'country')
setkeyv(eo.a, c('country', 'TIME'))

cols.to.add.chollette <-
  c("ny_gdp_totl_rt_zs", "revenueindex", "employmentindex", "regulationindex",
    "subsidisationindex")

# TODO add columns left/execl, authon, muni.
cols.to.add <-
  c('pop', 'gini_net', 'gini_market', 'fiscal_transparency', "stconst", "
      parlsys")

cols.to.add.locf <- c('muni', 'state', 'author', "auton")

new.data[, author:=as.double(author)]
new.data[, auton:=as.double(auton)]
new.data[, stconst:=as.double(stconst)]
new.data[, parlsys:=as.double(parlsys)]

for (col in cols.to.add.chollette){
  eo.q <- interpolateQuarterColumn(eo.q, new.data, col, MAX_YEAR_EXTRAPOLATION
      , 'denton-cholette')
}

for (col in cols.to.add){
  eo.q <- interpolateQuarterColumn(eo.q, new.data, col, MAX_YEAR_EXTRAPOLATION
      )
}

for (col in cols.to.add.locf){
  eo.q <- interpolateQuarterColumn(eo.q, new.data, col, MAX_YEAR_EXTRAPOLATION
      , 'locf')
}

## Fill forwards for muni and state

eo.q[, stconst_interpolated:=as.integer(stconst_interpolated)]
```

```
eo.q[, lpop_interpolated:=log(pop_interpolated)]
setnames(eo.q, 'pop_interpolated', 'population_interpolated')
eo.q[, gini_red_abs:=(gini_market_interpolated - gini_net_interpolated)]
eo.q[, gini_red_rel:=100*gini_red_abs/gini_net_interpolated]



DT ← fread('../data/execrlc_govfrac_yrcurnt_quartery_cleaned.csv')
DT[, V1:=NULL]
setnames(DT, 'location', 'country')
eo.q ← merge(eo.q, DT, by=c('country', 'TIME'), all=TRUE)

##########################################################################
### Transformation of the data to create the data matrix

### x is the data set with annual observation for eg
x ← copy(eo.q)
setkey(x, 'country')
x ← x[country.q]
# x ← x[TIME< 2013 & TIME > 1984.75]
x[, country:=as.factor(country)]
time.numeric ← x$TIME
x[, TIME.NUMERIC:=time.numeric]
x[, TIME:=as.factor(TIME)]
x[, QUARTER:=as.factor(QUARTER)]
x[, YEAR:= as.factor(YEAR)]

x[, egr := 100*eg/lf] # et: General Government employment, lf: Total labor
    force
x[, self_employment_rate := 100*es/lf]

x[, government_revenue:=100*yrg_interpolated/gdp, by='country'] # TODO gdp and
     not gdpv
x[, nlg_to_gdp:=100*nlg_interpolated/gdp, by='country'] # TODO change to gdp
    and not gdpv

x[, wage_share:=100*wage/gdp]
x[, openness:=100*(xgs+mgs)/gdp]

x[, gdp_per_capita:=gdpvd/population_interpolated/1e6]
x[, gdp_per_capita_log:=log(gdp_per_capita)]

setnames(x, 'ny_gdp_totl_rt_zs_interpolated', 'natural_ressource_rent')

##########################################################################
## LAGs might be useful in the future

## x[, ypgtq_interpolated_diff:=c(NA, diff(ypgtq_interpolated)), by='country']

## x[, gdp_per_capita_diff:=c(NA, diff(gdp_per_capita)), by='country']

## x[, egr_level_lagged:= c(NA, butlast(egr)), by='country'] # et: General
    Government employment, et: Total employment
## x[, egr_diff:= c(NA, diff(egr)), by='country'] # et: General Government
    employment, et: Total employment
```

```
## x[, egr_diff:= 100*egr/shift(egr, 1), by='country'] # percent change
## x[, egr_lagged:= c(NA, butlast(egr_diff)), by='country']
## x[, egr_lagged_2:= c(NA, butlast(egr_lagged)), by='country']

## x[, ydrh_to_gdpv_diff:=c(NA, diff(ydrh_to_gdpv)), by='country']

## x[, unr_lagged:=c(NA, butlast(unr)), by='country']
## x[, unr_diff:=c(NA, diff(unr)), by='country']
## x[, unr_diff_lagged:=c(NA, butlast(unr_diff)), by='country']

# x[, gdpv_annpct_quarterly_lagged:=c(NA, butlast(gdpv_annpct_quarterly)), by
    ='country']
# x[, gdpv_annpct_quarterly_lagged_2:=c(NA, NA, butlast(gdpv_annpct_quarterly,
     2)), by='country']

##############################################################################
## Remove NAs
x <- x[!is.na(egr)] # Non na observation

##############################################################################
## Diagnostic plots

if (MAKE_PLOTS){

  pdf('plot/variable_validation_check.pdf', 12, 7, onefile=TRUE)

  xyplot(egr ~ TIME.NUMERIC | country,
         na.omit(x[, c(cols, 'egr', 'country', 'TIME.NUMERIC'), with=F]),
         type='l', main='Public Employment in percent of labor force')

  xyplot(gdpv_yoy_annpct ~ TIME.NUMERIC | country,
       x[, c('gdpv_yoy_annpct', 'country', 'TIME.NUMERIC'), with=F],
       type='l', main='GDP Growth, YoY in %')

  xyplot(gap_interpolated ~ TIME.NUMERIC | country,
       x[, c(cols, 'gap_interpolated', 'country', 'TIME.NUMERIC'), with=F],
       type='l', main='GDP Output GAP in %' )

  xyplot(gaplfp_interpolated ~ TIME.NUMERIC | country,
         x[, c(cols, 'gaplfp_interpolated', 'country', 'TIME.NUMERIC'), with=F
             ],
         type='l', main='Labor gap')

  xyplot(gdpvd ~ TIME.NUMERIC | country,
         x[, c('gdpvd', 'country', 'TIME.NUMERIC'), with=F],
         type='l', main='GDPVD Quartery')

  xyplot(log(gdpvd) ~ TIME.NUMERIC | country,
         x[, c('gdpvd', 'country', 'TIME.NUMERIC'), with=F],
         type='l', main='GDPVD Quarterly (log)')

  xyplot(government_revenue ~ TIME.NUMERIC | country,
         x[, c('government_revenue', 'country', 'TIME.NUMERIC'), with=F],
         type='l', main='Government Revenue')

```

```
xyplot(gdp_per_capita_log ~ TIME.NUMERIC | country,
       x[, c(cols, 'gdp_per_capita_log', 'country', 'TIME.NUMERIC'), with=F],
       type='l', main='GDP per capita (log)')

xyplot(gdp_per_capita ~ TIME.NUMERIC | country,
       x[, c(cols, 'gdp_per_capita', 'country', 'TIME.NUMERIC'), with=F],
       type='l', main='GDP per capita')

## For Gini: JPN and CAD -> Data stops in 2007. Hence the number afterwards
## are not trustable.

xyplot(gini_net_interpolated ~ TIME.NUMERIC | country,
       x[, c('gini_net_interpolated', 'country', 'TIME.NUMERIC'), with=F],
       type='l', main='Gini Net (post-tax and post subsidies)')

xyplot(gini_market_interpolated ~ TIME.NUMERIC | country,
       x[, c('gini_market_interpolated', 'country', 'TIME.NUMERIC'), with=F],
       type='l', main='Gini Market (pre-tax and pre-subsidies)')

xyplot(gini_red_abs ~ TIME.NUMERIC | country,
       x[, c('gini_red_abs', 'country', 'TIME.NUMERIC'), with=F], type='l',
       main='Gini Reduction, Difference between pre and post tax/subsidies, Absolute')

xyplot(gini_red_rel ~ TIME.NUMERIC | country,
       x[, c('gini_red_rel', 'country', 'TIME.NUMERIC'), with=F],
       type='l', main='Gini Reduction (Relative)')

xyplot(fiscal_transparency_interpolated ~ TIME.NUMERIC | country,
       x[, c('fiscal_transparency_interpolated', 'country', 'TIME.NUMERIC'), with=F],
       type='l', main='IMF Fiscal Transparency Score')

xyplot(openness ~ TIME.NUMERIC | country,
       na.omit(x[, c(cols, 'openness', 'country', 'TIME.NUMERIC'), with=F]),
       type='l', main='Openness')

xyplot(wage_share ~ TIME.NUMERIC | country,
       na.omit(x[, c(cols, 'wage_share', 'country', 'TIME.NUMERIC'), with=F]),
       type='l', main='Wage share')

xyplot(yrcurnt ~ TIME.NUMERIC | country,
       na.omit(x[, c(cols, 'yrcurnt', 'country', 'TIME.NUMERIC'), with=F]),
       main='Years until next election')

xyplot(parlsys_interpolated ~ TIME.NUMERIC | country,
       na.omit(x[, c(cols, 'parlsys_interpolated', 'country', 'TIME.NUMERIC'), with=F]),
       main='Parlsys')

xyplot(auton_interpolated ~ TIME.NUMERIC | country,
       na.omit(x[, c(cols, 'auton_interpolated', 'country', 'TIME.NUMERIC'),
```

```
                 with=F]),
        main='Auton')

  xyplot(natural_ressource_rent ~ TIME.NUMERIC | country,
        na.omit(x[, c(cols, 'natural_ressource_rent', 'country', 'TIME.NUMERIC'), with=F]),
        main='Natural Ressource Rent', type='l')

  xyplot(left ~ TIME.NUMERIC | country,
        na.omit(x[, c(cols, 'left', 'country', 'TIME.NUMERIC'), with=F]),
        main='Left', type='l')

  xyplot(self_employment_rate ~ TIME.NUMERIC | country,
        na.omit(x[, c(cols, 'self_employment_rate', 'country', 'TIME.NUMERIC'), with=F]),
        main='Self Employment Rate', type='l')

  dev.off()
}


############################################################################
## Save into CSV

write.csv(x, '../data/public_employment_data_all.csv')
write.csv(x[, cols.to.save, with=F], '../data/public_employment_design_matrix.csv')
```

Analysis script

```
source('init.R')
pckgs ← loadPackages()
library(mice)
library(impute)
library(magrittr)
library(parallel)
library(lattice)

MAKE_PLOTS ← TRUE
MAX_YEAR_EXTRAPOLATION ← 2014

cols ← c(
  'egr',
  'egr_lagged', # lagged public employment rate
  'unr',
  'government_revenue',# yrg over gdpvd
  'nlg_to_gdp', # net landing in % of gdp
  'TIME',
  ## 'QUARTER',
  'country'
)

cols.to.save ← c(
```

```
  cols,
  'egr',
  'country',
  'YEAR',
  'nlg_to_gdpv',
  'fiscal_transparency_interpolated', # 'imf_gfs_scores'
  'gini_market_interpolated',
  'gini_net_interpolated',
  'gini_red_abs',
  'gini_red_rel',
  'gap_interpolated',
  'gaplfp_interpolated',
  'execrlc',
  'govfrac',
  'yrcurnt', # year until next election
  'is_election_date' # If the quarter is an election quarter
  ) %>% sort %>% {c('TIME', .)}


# fread('../data/public_employment_data_all.csv')
x ← fread('../data/public_employment_design_matrix.csv')
head(x)
x[, V1:=NULL]
setkey(x, 'country')

x[, gdp_per_capita_log:= log(gdp_per_capita)]
x[, gdpv_yoy_annpct_lagged:= c(NA, butlast(gdpv_yoy_annpct)), by='country']
x[, gdpv_yoy_annpct_lagged_2:=c(NA, NA, butlast(gdpv_yoy_annpct, 2)), by='
    country']
x[, lpop_interpolated:=log(population_interpolated)]

time.numeric ← x$TIME
x[, TIME:=as.numeric(TIME)]
x ← x[TIME < 2013 & TIME > 1989.75]

x[, TIME:=as.factor(TIME)]
x[, QUARTER:=as.factor(QUARTER)]
x[, YEAR:= as.numeric(YEAR)]

x[, egr_lagged:= c(NA, butlast(egr)), by='country']
x[, country:=as.factor(country)]


x[, unr_lagged := c(NA, butlast(unr)), by='country']
x[, gdp_per_capita_log_lagged := c(NA, butlast(gdp_per_capita_log)), by='
    country']

## Keep data for plotting
lvl2num ← function(x) as.numeric(levels(x)[x])
x.model.lm ← na.omit(x[, c(cols, 'country', 'TIME'), with=FALSE])


x[, egr_country := scale(egr, center=TRUE, scale=FALSE), by='country']


```

```

## Simple lm model
x.lm ← lm(egr  ~ ., x[, c(cols), with=FALSE])
summary(x.lm)
showdiag(x.lm)

y.fit.simple.lm ← fitted(x.lm)

base.cols ← c('egr_lagged', 'government_revenue', 'nlg_to_gdp',
               'unr', 'country', 'TIME', 'egr')

gov.lm ← robustnessAnalysis(
  x, c('egr_lagged', 'government_revenue', 'country', 'TIME', 'egr'), '')
nlg.lm ← robustnessAnalysis(
  x, c('egr_lagged', 'nlg_to_gdp', 'country', 'TIME', 'egr'), '')
unr.lm ← robustnessAnalysis(
  x, c('egr_lagged', 'unr', 'country', 'TIME', 'egr'), '')

baseline.lm ← robustnessAnalysis(
  x, c('egr_lagged', 'government_revenue', 'nlg_to_gdp',
       'unr', 'country', 'TIME', 'egr'), '')

gdp.lm ← robustnessAnalysis(x, c(base.cols, 'gdpv_yoy_annpct'), '')
gdp.wo.lm ← robustnessAnalysis(as.data.table(gdp.lm$model), base.cols, '', ff)
# gdp.lm ← robustnessAnalysis(x, c('gdpv_yoy_annpct_lagged', base.cols), '')

gdp.capita.lm ← robustnessAnalysis(x, c(base.cols, 'gdp_per_capita_log'), '')
gdp.capita.wo.lm ← robustnessAnalysis(
  as.data.table(gdp.capita.lm$model), base.cols, '', ff)

open.lm ← robustnessAnalysis(x, c('openness', base.cols), '')
self.lm ← robustnessAnalysis(x, c('self_employment_rate', base.cols), '')

## Fiscal Transparency
ff ← egr ~ .
imf.gfs.lm ← robustnessAnalysis(x, c(cols, 'fiscal_transparency_interpolated'), '', ff)
imf.gfs.wo.lm ← robustnessAnalysis(as.data.table(imf.gfs.lm$model), cols, '', ff)

## Lassen
x.new ← copy(x)
lassen ← fread('../data/lassen_fiscal_scores.csv')
setnames(lassen, 'Index Score', 'lassen_score')
x.lassen ← merge(x.new, lassen, by.x='country', by.y='ISO')
lassen.lm ← robustnessAnalysis(x.lassen, c(cols, 'lassen_score', 'left'), '',
                                egr ~ . + left*lassen_score + left - lassen_score)

## Left or right goverment
govrlc.lm ← robustnessAnalysis(x, c(cols, 'left'), '', ff)
govrlc.base.lm ← robustnessAnalysis(x, c(base.cols, 'left'), '', ff)
govrlc.wo.lm ← robustnessAnalysis(as.data.table(govrlc.lm$model), cols, '', ff)
```

```

is_election.base.lm ← robustnessAnalysis(x, c('is_election_date', base.cols),
    '', ff)

## Years until election
nrr.lm ← robustnessAnalysis(
  x, c('natural_ressource_rent', 'is_election_date', base.cols), '', ff)

nrr.mix.lm ← robustnessAnalysis(
  x, c('natural_ressource_rent', 'is_election_date', base.cols), '',
  egr ~ . + natural_ressource_rent*is_election_date)

lassen.lm ← robustnessAnalysis(
  x.lassen, c(base.cols, 'fiscal_transparency_interpolated', 'is_election_date
    '), '',
  egr ~ . + is_election_date*fiscal_transparency_interpolated)

stconst.lm ← robustnessAnalysis(
  x, c('state_interpolated', 'is_election_date', base.cols), '', ff)

yrcurnt.lm ← robustnessAnalysis(x, c(cols, 'yrcurnt'), '', ff)
yrcurnt.wo.lm ← robustnessAnalysis(as.data.table(yrcurnt.lm$model), cols, '',
    ff)

pop.lm ← robustnessAnalysis(x, c(cols, 'lpop_interpolated'), '', ff) # log
    population
pop.wo.lm ← robustnessAnalysis(as.data.table(pop.lm$model), cols, '', ff)

## Inequality
y ← copy(x)
y[, TIME:=fctr2num(TIME)]
y ← y[TIME < 2010] # restrict time because of interpolation
y[, TIME:=as.factor(TIME)]

gini.red.lm ← robustnessAnalysis(y, c(cols, 'gini_red_abs'), '', ff)
gini.lm ← robustnessAnalysis(
  y, c(cols, 'gini_market_interpolated', 'gini_net_interpolated'), '', ff)
## gini.lm ← robustnessAnalysis(
##   y, c(cols, 'gini_red_abs', 'gini_red_rel'), '', ff)
gini.wo.lm ← robustnessAnalysis(
  as.data.table(gini.lm$model), cols, '', ff)

## Net lending
nlg.lm ← robustnessAnalysis(x, c(cols, 'nlg_to_gdp'), '', ff)
nlg.wo.lm ← robustnessAnalysis(as.data.table(nlg.lm$model), cols, '', ff)

## Government fractionalization
govfrac.lm ← robustnessAnalysis(x, c(cols, 'govfrac'), '', ff)
govfrac.base.lm ← robustnessAnalysis(x, c('govfrac', base.cols), '', ff)
govfrac.wo.lm ← robustnessAnalysis(
  as.data.table(govfrac.lm$model), cols, '', ff)

## Federalism
x[, muni_interpolated:=gsub("-999", NA, muni_interpolated)]
x[, state_interpolated:=gsub("-999", NA, state_interpolated)]
```

```
x[, state_interpolated:=gsub("No local elections",
                              0, state_interpolated)]
x[, state_interpolated:=gsub("Legislature locally elected",
                              1, state_interpolated)]
x[, state_interpolated:=gsub("Legislature and executive locally elected",
                              1, state_interpolated)]

fed.lm <- robustnessAnalysis(
  x, c(cols, 'muni_interpolated', 'auton_interpolated',
       'state_interpolated'), '',
  egr ~ . + muni_interpolated*gdpv_yoy_annpct_lagged +
    auton_interpolated*gdpv_yoy_annpct_lagged +
    state_interpolated*gdpv_yoy_annpct_lagged - state_interpolated)

##############################################################################

descriptions <-
  list(`gdpv\\_yoy\\_annpct`='GDP growth, YoY in percent',
       unr='Unemployment rate',
       ypgtq='Total disbursements, general government, in percent of GDP',
       egr='Public employment rate',
       lpop='Log of population in million',
       `lpop\\_interpolated` ='Log of population in million',
       `ydrh\\_to\\_gdpv`='Household net income, in percent of GDP',
       `gdp\\_per\\_capita`='GDP per capita in USD Millions',
       `fiscal\\_transparency`='IMF GFS Index',
       incomeineq='Gini coefficient',
       lpoptot='Log of total population in million',
       'TIME'='Time',
       egr_diff='Change in Public Employment Rate (CPER)',
       `egr\\_lagged`='Lagged public employment rate',
       ## egr_lagged='Lagged change in Public Employment Rate',
       unr_lagged='Lagged unemployment rate',
       `government\\_revenue`='Government Revenue',
       `gdp\\_per\\_capita\\_log`='GDP per capita, 2010 PPP, in USD/person',
       'nlg\\_to\\_gdpv'='Net Lending, in percent of GDP',
       'gap_interpolated'='Output Gap in percent',
       'gdpv_yoy_annpct'='GDP growth, YoY in percent',
       'egr\\_lagged'='Public employment rate (1 Quarter Lag)'
       )


description <-
  c(list(gdpv_annpct='GDP growth',
         ydrh_to_gdpv='Household net income, in \\% of GDP',
         `gdp_per_capita_log`='Log of GDP per capita, in USD Millions',
         fiscal_transparency_interpolated='IMF GFS Index',
         `ypgtq_interpolated`='Government expenditure in \\% of GDP (
             interpolated)',
         country='Country',
         `gdpv_annpct:fiscal_transparency_score`='Effect of fiscal
             transparency on GDP growth',
         fiscal_transparency_score='Fiscal Transparency',
         'gini_toth'='Gini coefficient (Toth 2015)',
         egr_diff='Difference with previous public employment rate (CPER)',
```

```r
         egr_lagged='Lagged of difference in public employment rate',
         lpop_interpolated='Log of population in million',
         QUARTER='Quarter',
         YEAR='Year',
         TIME='Time',
         government_revenue='Government Revenue',
         execrlc='Left Orientated Government',
         yrcurnt='Years until next election',
         gini_market_interpolated='Gini Coefficient, Market Income',
         gini_net_interpolated='Gini Coefficient, Net Income',
         govfrac='Government Fractionalization',
         nlg_to_gdpv='Net Lending in percent of GDP',
         nlg_to_gdp='Net Lending in percent of GDP',
         gini_red_abs='Diff. of Gini Market and Net Income',
         left='Government with social orientation'),
    descriptions)


queryList ← function(l, kx){
  kx %>%
    lapply(function(s) if(is.null(d ← l[[s]])) NA else d) %>%
    unlist
}


x.lm$model$TIME ← lvl2num(x.lm$model$TIME)
queryList(description, colnames(x.lm$model)) %>% {
  stargazer(x.lm$model, out='model_output/simple_statistic_quarterly.tex',
            covariate.labels=.,
            font.size='footnotesize', title='Data statistics')
}

description[['egr_lagged']] ← NA
description[['YEAR']] ← NA

toTexModel ← function(li.lm, title, out, dep.name='Public employment rate'){
  cov.labs ← na.omit(queryList(description, names(coef(li.lm[[1]]))[-1]))

  argx ← c(li.lm, list(title=title, out=out, covariate.labels=cov.labs,
                       dep.var.labels=dep.name, omit=c('egr_lagged', 'TIME', 'country'),
                       omit.labels = c('Auto-correlation effect', 'Time effect', 'Country effect')))
  do.call(stargazer, argx)
}

dep.name ← 'Public employment rate'


toTexModel(list(x.lm, unr.lm, gov.lm, nlg.lm),
           'Main variable result',
           'model_output/simple_lm_quarterly.tex')

toTexModel(list(imf.gfs.lm, imf.gfs.wo.lm),
           'Effect of IMF GFS Score',
```

```r
           'model_output/simple_lm_imf_quarterly.tex')
toTexModel(list(govrlc.lm, govrlc.wo.lm),
           'Effect of Government Political Orientation',
           'model_output/simple_lm_govrlc_quarterly.tex')
toTexModel(list(govfrac.lm, govfrac.wo.lm),
           'Effect of Government Fractionalization',
           'model_output/simple_lm_govfrac_quarterly.tex')
toTexModel(list(yrcurnt.lm, yrcurnt.wo.lm),
           'Effect of Years until next Election',
           'model_output/simple_lm_yrcurnt_quarterly.tex')
toTexModel(list(gini.lm, gini.wo.lm),
           'Effect of Gini coefficient, data up to 2010 (included)',
           'model_output/simple_lm_gini_quarterly.tex')
toTexModel(list(gini.red.lm, gini.wo.lm),
           'Effect of Difference of Gini coefficient (Market and Net), data up to 2010 (included)',
           'model_output/simple_lm_gini_red_quarterly.tex')
toTexModel(list(gdp.lm, gdp.wo.lm),
           'Effect of GDP',
           'model_output/simple_lm_gdp_quarterly.tex')
toTexModel(list(gdp.capita.lm, gdp.capita.wo.lm),
           'Effect of GDP per Capita',
           'model_output/simple_lm_gdp_capita_quarterly.tex')



## Plots
if (FALSE) {
  cols.plot <- cols[cols!='TIME']
  data.plot <- melt(x[, c(cols.plot, 'YEAR', 'QUARTER'), with=FALSE],
                    id.vars=c('country', 'YEAR', 'QUARTER'))

  quarter.substitute <-
    lapply(1:4, function(i) list(paste0('-Q', i), paste0('.', 100*(i-1)/4)))

  quarters.time <- Reduce(function(x, l) gsub(l[[1]], l[[2]], x),  quarter.substitute,
                          data.plot[, paste0(YEAR, '-', QUARTER)]) %>% as.numeric
  data.plot[, Time:=quarters.time]


  data.plot[, variable:=gsub('_', '\\\\_', variable)]

  data.plot[, {
    options(tikzDefaultEngine = 'pdftex')
    s <- paste0('plot/simple_model_quarterly_', .BY[[1]], '.tex')
    s <- gsub('\\', '', s, fixed=TRUE)
    gg2 <- ggplot(.SD, aes(Time, value)) + geom_line() + facet_wrap(~ country) +
      ggtitle(paste0(descriptions[[.BY[[1]]]], ' by country'))
    tikz(s, height=6, width=9)
    print(gg2)
    dev.off()
  }, by='variable']
```

```

  pdf('plot/model_diagnostic_quarterly.pdf', width=9, height=9)
  par(mfrow=c(2,2))
  plot(x.lm)
  dev.off()

}
```

Declaration of Originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor .

Title of work (in block letters):

Public Employment: Data Analysis with OECD Economic Outlook Quartely Data

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Pham	David-Olivier

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the Citation etiquette information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work .
- I am aware that the work may be screened electronically for plagiarism.
- I have understood and followed the guidelines in the document *Scientific Works in Mathematics.*

Place, date:

Zürich, 15.03.2016

Signature(s):

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.