Efficient Inference with Panel Data

On the pass-through of the Dutch 2001 and 2012 VAT increases to consumer prices.

by

George J.C. van Hooft

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday August 28, 2018 at 15:00.

Student number:	4432215	
Thesis committee:	Dr. ir. G.N.J.C. Bierkens,	TU Delft, supervisor
	Dr. ir. F. H. van der Meulen,	TU Delft, chair
	Dr. D. Kurowicka,	TU Delft

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Abstract

This thesis evaluates the pass-through of the 2001 and 2012 Dutch Value Added Tax (VAT) increases to customer prices using a difference-in-differences model. To this end, the first difference and feasible generalised least squares estimators are introduced. Contrary to the conventional pooled OLS estimator, these estimators always show significant causal effects for both VAT hikes. These also dramatically improve the accuracy of the estimates compared to earlier research on the incidence of VAT. For the 2012 tax increase, the null hypothesis of full pass through is even rejected. This result is a novelty in the econometric literature. Even in more general settings, the estimators used in this thesis prove far superior over conventional causal estimation techniques of difference-in-differences models.

Keywords: efficient linear regression, panel data, first difference, difference-in-differences, value added tax

Contents

Chapter 1

Introduction

Standard macroeconomic models suggest that taxation has a major disruptive effect on the economy (Stiglitz, 1988, p. 390). The Value Added Tax (hereinafter referred to as: VAT) is progressively winning importance in public finance. The VAT is best described as a tax on consumption. Economists stress the fact that consumption taxes tend to have a less disruptive effect on the economy than income taxes. Hence, a world wide shift from income taxes to the VAT can be observed (Sanger and Thomas, 2018). E.g., the Netherlands increased its standard rate from 17.5% to 19% in 2001 and to 21% in 2012. Furthermore, the Dutch government considers the increase of the reduced rate from 6% to 9%. These increases lead to all kind of economic impact questions, among which the question to what extent the VAT increases are passed through on consumer prices remains unclear. Naturally, research has been done on this matter but the conclusions are ambiguous. The Bureau of Economic Policy Analysis (CPB) for example suggests that we cannot reject the hypothesis that the 2001 and 2012 VAT increases were completely passed through to consumer prices (Vrijburg, Mellens, and Dijkstra, 2014). Benedek, de Mooij, Keen, and Wingender (2015) came to the same conclusion in a more general setting.

The statistical models used in these papers are often based on the Difference-in-Differences (DD) model. These models aim to capture a causal treatment effect of a policy change over a certain time period - in this case the VAT pass through. DD models are commonly used in econometric literature, yet Bertrand, Duflo, and Mullainathan (2004) suggest that the standard errors reported by most estimators are incorrect due to the presence of strong autocorrelation in the error term. In particular, this could lead to misleading inference. They recommend using robust standard errors that take the of serial correlation (also referred to as autocorrelation) into account. However, these can be substantially bigger than their non-robust counterparts. The non-rejection of the full pass through hypothesis in the paper of Vrijburg et al. (2014) and Benedek et al. (2015) was largely due to the fact that the estimates were accompanied with very high standard errors. This is likely a sign that the estimators were inefficient and consequently

produced inaccurate estimates.

1.1 Research question

This thesis deals with linear panel data models. It is largely known fact that the OLS estimator is only efficient provided the error term is serially uncorrelated and the variances are equal among the cross sectional units (homoscedasticity). In all other cases, OLS is not guaranteed to be efficient and other linear estimator may perform better. As noted in the introduction, a solution that is often used in practice is the use of OLS accompanied with serial correlation robust standard errors. Robust standard errors may only be applied when semi-parametric linear regression methods are used.[1] This thesis further investigates other semi-parametric linear regression estimators that are more efficient than the OLS estimator in presence of serial correlation and/or unequal variances among the cross sectional units (heteroscedasticity). These estimators will be used to see if they increase the accuracy of the estimates for the Dutch 2001 and 2012 VAT increases. In particular, we want to see if these estimators lead to the rejection of the null hypothesis that the VAT was fully passed to consumer prices. This leads to the formulation of the following research question:

Do semi-parametric linear regression estimators that are efficient in presence of serial correlation lead to the rejection of the hypothesis that the Dutch 2001 and 2012 VAT-increases were fully passed through to consumer prices?

Throughout this thesis, we restrict the analysis to the class of linear Generalised Method of Moments (GMM) estimators. This broad class of estimators only require appropriate moment conditions rather than the full specification of distributions. This is often considered a semi-parametric approach (e.g. Wooldridge, 2001), but Robert (2007) argues that such statistical methods qualify as non-parametric. However, non-parametric regression often refers to kernel density regression techniques. That is, qualifying the class of GMM estimators as non-parametric may cause confusion and so the class of GMM estimators is qualified as semi-parametric.

1.2 Scientific significance

The Difference-in-Differences model is among the most commonly used causal estimation techniques in econometrics. The DD is easy to implement and circumvents many endogeneity problems which would otherwise cause OLS to be inconsistent. Given the fact that these models are often heavily affected by serial correlation, most researchers apply the OLS estimator and accompany them with serial correlation robust standard error. This most commonly applied solution proposed by Bertrand et al. (2004) is far from satisfactory.[2] While it leads to a decrease of type I errors (rejecting a true null hypothesis), it also leads to a dramatic increase of type II errors (failing to reject a false

[1] In a Maximum Likelihood context, this is an indication of a model misspecification.

[2] They proposed more solutions, but these are not commonly implemented.

null hypothesis). This thesis investigates estimators that are efficient in presence of serial correlation. Such estimators simultaneously reduce type I errors without increasing type II errors. This is a far superior solution compared to OLS accompanied with serial correlation robust standard errors.

The estimators that are derived in this thesis will be used to capture the causal effect of the 2001 and 2012 VAT increases on consumer prices. Due to the fact that earlier research used inefficient estimators, this thesis will derive a more decisive answer on whether these VAT hikes were fully passed through to consumer prices. This gives policymakers extra tools for designing an efficient system of consumption taxes.

1.3 Structure and main findings

This thesis is roughly subdevided in two parts. The first part discusses the theoretical linear regression methods for panel data models. In the second part, these methods are applied on the incidence of VAT. The following topics are discussed to derive an answer on the research question:

Part I

Chapter 2 introduces the general framework of this thesis. The generalised method of moments estimator and its properties will be introduced and compared with maximum likelihood estimation (MLE). Furthermore, some general conventions that will be used throughout this thesis are introduced and the classical linear regression assumptions are revisited.

Extending these results to a regression framework, chapter 3 discusses non-parametric transformations of the OLS estimator that are efficient under alternative formulations of the classical assumptions. While the finite sample properties under the Gauss-Markov assumptions are discussed, special emphasis is placed on the asymptotic properties of these estimators. It appears that many desirable properties of these estimators can be derived asymptotically under assumptions that are considerably weaker than the ones needed for the finite sample properties. This includes an asymptotic efficiency result. These estimators are compared with their fully parametric Bayesian or maximum likelihood analogue.

These results are extended in chapter 4 for parametric transformations of the OLS estimator. While these estimators allow a more accurate specification of dependence structures, it also appears that these estimators in general violate the Gauss-Markov assumptions. However, they do preserve some asymptotic efficiency results. Because efficiency strongly relies on the correct specification of the model, chapter 5 introduces some specification tests that are useful in determining which model is most appropriate to use. This chapter also gives a brief introduction in causal estimation, where special emphasis is placed the difference-in-differences model. It appears that these models are usually affected by heavy serial dependence, reducing the efficiency of the usual OLS estimator. Given the theory and simulations made in the previous chapters, estimators

are proposed that are likely to perform well for causal estimation using difference-in-differences.

Part II

Chapter 6 introduces the legal and economic concepts. We will need those results to fully justify the choices made in the causal regression framework. This chapter starts by discussing the Dutch VAT act after which the economic theory behind the incidence of VAT is discussed. In particular, a theoretic economic model is derived that is used to formulate some hypotheses. This model provides for example a relative theoretical upper bound for the VAT pass through.

This theory is then used to make causal inference on the tax incidence of the 2001 and 2012 VAT hikes, which is done in chapter 7. These results are compared with the outcome of the CPB anaysis (Vrijburg et al., 2014). It appears that most of their results are only significant at a 10% level. Even though this thesis adheres a 5% bound, we can reproduce most of their findings at a 1% level. Furthermore, some of their conclusions are rejected. This includes the rejection of the null hypothesis that the 2012 VAT was fully passed through to consumer prices.

1.4 Basic statistics and probability

While the reader is assumed to be familiar with basic probablistic and statistics, some of the necessary preliminaries will be reconsidered. It is especially recommended to read paragraph 1.4.1 for the short review of semi-parametric statistical methods.

Definition 1.4.1 (Conditional Expected Value)**.** Let $X : \Omega \to \mathbb{R}$ be an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{A} \subset \mathcal{F}$ be a sigma-algebra. Then the **conditional expectation** of X w.r.t. $\mathcal{A}$, denoted as $\mathbb{E}[X|\mathcal{A}]$ is the unique $\mathcal{A}$-measurable random variable satisfying

$$\int_A X d\mathbb{P} = \int_A \mathbb{E}[X|\mathcal{A}] d\mathbb{P}, \quad A \in \mathcal{A} \tag{1.1}$$

In this thesis the law of iterated expectations will often be used - the theorem and proof will be stated below:

Theorem 1.4.1 (Law of Iterated Expectations, LIE)**.** *Let X be an integrable random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{A} \subset \mathcal{F}$ be a σ-algebra. Then*

$$\mathbb{E}[\mathbb{E}[X|\mathcal{A}]] = \mathbb{E}[X] \tag{1.2}$$

Proof. Note that $\Omega \in \mathcal{A}$ and simply pick $A = \Omega$, i.e.

$$\begin{aligned}\mathbb{E}[X] &= \int_{\Omega} X d\mathbb{P} \\ &= \int_{\Omega} \mathbb{E}[X|\mathcal{A}] d\mathbb{P} \\ &= \mathbb{E}[\mathbb{E}[X|\mathcal{A}]]\end{aligned}$$

□

This thesis heavily relies on asymptotic theory, hence the following convergence notions are important.

Definition 1.4.2. Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

1. The sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is said to converge **almost surely** to a random variable X iff
$$\mathbb{P}[\lim_{n\to\infty} X_n = X] = 1 \tag{1.3}$$
Notation: $X_n \xrightarrow{a.s.} X$ or as-$\lim_{n\to\infty} X_n = X$.

2. The sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is said to converge **in probability** to a random variable X iff
$$\forall \varepsilon > 0 \ : \ \lim_{n\to\infty} \mathbb{P}[|X_n - X| > \varepsilon] = 0 \tag{1.4}$$
Notation: $X_n \xrightarrow{p} X$ or p-$\lim_{n\to\infty} X_n = X$.

3. The sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is said to converge **in distribution** to a random variable X iff
$$\forall x \in \mathbb{R} \text{ at which } F \text{ is continuous } \lim_{n\to\infty} F_{X_n}(x) = F_X(x) \tag{1.5}$$
Where F_{X_n} and F_X represent the cumulative distribution functions of X_n and X respectively. Notation: $X_n \xrightarrow{d} X$ or d-$\lim_{n\to\infty} X_n = X$.

Theorem 1.4.2 (Law of Large Numbers, LLN)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of iid random variables. Then*
$$\bar{X}_n := \frac{1}{n}\left(\sum_{i=1}^{N} X_i\right) \xrightarrow{a.s.} \mathbb{E}[X_i] \tag{1.6}$$

Theorem 1.4.3 (Central Limit Theorem, CLT)**.** *Let $\{X_N\}_{N\in\mathbb{N}}$, $X_N \in \mathbb{R}^T$ be a sequence of iid random variables with $\mathbb{E}[X_N] = 0$ and $\mathbb{E}[X_N X_N^T] < \infty$. Then,*
$$N^{-1/2} \sum_{i=1}^{N} X_i \xrightarrow{d} N(0, \mathbf{D}) \tag{1.7}$$

where $\underset{T\times T}{\mathbf{D}} = \mathbb{E}[X_N X_N^T]$.

Theorem 1.4.4 (Continuous Mapping Theorem, CMT - Mann and Wald (1943))**.** *Let* $\{X_n\}_{n\in\mathbb{N}}$ *be a* K*-dimensional sequence of random vectors. If* $g : \mathbb{R}^K \to \mathbb{R}^L$ *is a continuous function, then:*

$$X_n \xrightarrow{a.s.} X \implies g(X_n) \xrightarrow{a.s.} g(X) \tag{1.8}$$

1.4.1 Basic statistical inference

Definition 1.4.3. Let $\hat{\theta}$ be an estimator for θ, then:

1. $\hat{\theta}$ is said to be unbiased when $\mathbb{E}[\hat{\theta}] = \theta$;
2. $\hat{\theta}$ is said to be weakly consistent when p-lim$_{n\to\infty}\hat{\theta} = \theta$;
3. $\hat{\theta}$ is said to be strongly consistent when as-lim$_{n\to\infty}\hat{\theta} = \theta$.

This thesis focusses on strongly consistent estimators. They will simply be referred to as consistent estimators.

Definition 1.4.4 (Asymptotic normality of an estimator)**.** Let $\{\hat{\theta}_N\}_{N\in\mathbb{N}}$, $\hat{\theta}_N \in \mathbb{R}^K$ be a sequence of estimators for $\theta \in \mathbb{R}^K$.

(i) If

$$\sqrt{N}\left(\hat{\theta}_N - \theta\right) \xrightarrow{d} N(0, \mathbf{V}) \tag{1.9}$$

then $\hat{\theta}_N$ is said to be $\sqrt{\mathbf{N}}$**-asymptotically normally distributed**.

(ii) The matrix $\mathbf{V}$ is called the **asymptotic variance-covariance matrix**, which is a positive semi-definite matrix.

Notation: $\mathbf{V} = \text{Avar}\left(\sqrt{N}\left(\hat{\theta}_N - \theta\right)\right)$.

(iii) Assume we have a consistent estimator $\hat{\mathbf{V}}_N$ for $\mathbf{V}$. Then the **asymptotic standard error** of $\hat{\theta}_{N_i},\ i = 1, \ldots, K$ is defined as

$$se\left(\hat{\theta}_{N_i}\right) = \left(\hat{\mathbf{V}}_{N_{i,i}}/N\right)^{1/2} \tag{1.10}$$

(iv) The asymptotic variance of the estimator $\hat{\theta}_N$ is defined as $\text{Avar}\left(\hat{\theta}_N\right) = \frac{\mathbf{V}}{N}$.

The expression for the asymptotic variance of $\hat{\theta}_N$ may be confusion due to devision by N. The name 'asymptotic variance' should in particular not be taken literally. Assume for the moment that $\hat{\theta}_N \xrightarrow{a.s.} \theta$. Since θ is non-stochastic, $\text{Avar}\left(\hat{\theta}_N\right) \xrightarrow{a.s.} 0$. The definition of asymptotic variance is better seen as the large sample behaviour of the variance.

Efficiency

The question often arrises whether an estimator is efficient compared to other estimators. For this thesis, the concept of relative efficiency will play a central role in this thesis.

Definition 1.4.5 (Efficient estimators)**.** Let $\hat{\theta}$ be a consistent and a $\sqrt{N}$-asymptotically normally distributed estimator for θ with $\text{Avar}\left(\hat{\theta}\right) = \frac{\mathbf{V}}{N}$.

(i) Let $\bar{\theta}$ be another $\sqrt{N}$-asymptotically normally distributed estimator for θ with $\text{Avar}\left(\bar{\theta}\right) = \frac{\mathbf{D}}{N}$. $\hat{\theta}$ is said to be **asymptotically efficient relative to** $\bar{\theta}$ if

$$\mathbf{D} - \mathbf{V} \tag{1.11}$$

is a positive semi-definite (PSD) matrix (Wooldridge, 2001, p. 42).

(ii) An estimator $\hat{\theta}$ is said to be **efficient for a finite sample** if it is unbiased and, for any other estimator $\bar{\theta}$,

$$\text{Var}\left(\bar{\theta}\right) - \text{Var}\left(\hat{\theta}\right) \tag{1.12}$$

is a PSD matrix.

1.4.2 Hypothesis Testing

This section introduces the asymptotic theory for hypothesis testing for a certain parameter $\underset{K\times 1}{\theta}$. Let $\underset{K\times 1}{\hat{\theta}}$ be a consistent estimator for θ s.t.

$$\sqrt{N}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N\left(0, \mathbf{V}\right) \tag{1.13}$$

where $\mathbf{V}$ denotes a positive definite asymptotic variance-covariance matrix.

Wald Statistic

Wald (1939) introduced a statistic that is useful to test hypotheses in a semi-parametric setting where no distributions have been prespecified (a likelihood ratio test is infeasible). As a preliminary remark, the Wald test greatly relies on asymptotics and may be unreliable for small samples.

Consider testing J hypotheses on the K-parameters in $\hat{\theta}$, which are denoted by a (non-stochastic) matrix $\underset{J\times K}{\mathbf{R}}$. That is, we wish to test

$$H_0 = \mathbf{R}\theta = r$$
$$H_1 = \mathbf{R}\theta \neq r$$

For which the test statistic is given by

$$W_N = \left(\sqrt{N}\mathbf{R}\hat{\theta} - r\right)^T \left(\mathbf{R}\mathbf{V}\mathbf{R}^T\right)^{-1} \left(\sqrt{N}\mathbf{R}\hat{\theta} - r\right) \underset{H_0}{\xrightarrow{d}} \chi^2_J. \tag{1.14}$$

If the variance-covariance matrix $\mathbf{V}$ is unknown, it can be replaced by a consistent estimator of $\mathbf{V}$, which will be denoted as $\hat{\mathbf{V}}$, in which case it is useful to rewrite the Wald statistic in equation 1.14 as

$$W_N = \left(\mathbf{R}\hat{\theta} - r\right)^T \left(\mathbf{R}\frac{\hat{\mathbf{V}}}{N}\mathbf{R}^T\right)^{-1} \left(\mathbf{R}\hat{\theta} - r\right) \underset{H_0}{\overset{d}{\rightarrow}} \chi^2_J \tag{1.15}$$

Let α be the significance level (in this thesis $\alpha = 0.05$), then the H_0 hypothesis is rejected if for a critical region C_α

$$\alpha = \mathbb{P}_{H_0}[W_N > C_\alpha] = 1 - \mathbb{P}_{H_0}[W_N \leq c_\alpha] \underset{H_0}{\overset{d}{\rightarrow}} 1 - F\left(C_\alpha\right)$$

where F denotes the CDF of a χ^2_J distribution. This result suggests

$$C_\alpha = F^{-1}\left(1 - \alpha\right) \tag{1.16}$$

And so, H_0 is rejected if

$$W_N > C_\alpha \tag{1.17}$$

Theorem 1.4.5. *The Wald-statistic is asymptotically equivalent to the Likelihood Ratio test.*

Proof. See Atkison and Lawrance, 1983. □

Example 1.4.1 (Wald F-test)**.** Assume we are considering the regression model

$$\underset{T\times 1}{y_i} = \underset{T\times K}{\mathbf{X}_i}\ \underset{K\times 1}{\beta} + \underset{T\times 1}{\varepsilon_i}\ , \tag{1.18}$$

where $\beta = [\alpha, \theta]$ is the parameter to be estimated, α the intercept and $\mathbf{X}_i$ is a design matrix. The F-test tests the hypothesis

$$H_0 : \theta = 0$$
$$H_1 : \theta \neq 0$$

i.e. under the null hypothesis, the true model merely consists of the intercept α while under the alternative hypothesis, θ is significantly different from 0. The Wald F-test uses

$$\mathbf{R} = \mathbf{I}_{K-1}$$
$$\underset{K-1\times 1}{r} = [0,\ 0, \cdots,\ 0]^T$$

Example 1.4.2 (Wald t-test)**.** We consider again the regression model described in equation 1.18. Let $\beta_i,\ i = 1, \ldots, k$ be i-th component of β. The t-test tests the hypothesis

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

i.e. under the null hypothesis, β_i is not significantly different from 0. The Wald t-test uses

$$\underset{K\times 1}{\mathbf{R}} = e_i^T$$
$$\underset{K\times 1}{r} = 0$$

where e_i is a vector with a 1 on the i-th row and 0 everywhere else.

1.4.3 Bayesian Inference

Bayesian inference strongly relies on Bayes' theorem, which will be stated here for completeness.

Theorem 1.4.6. ***(Bayes)***

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A, B \in \mathcal{F}$ be two events. Then

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]} \tag{1.19}$$

The probability $\mathbb{P}[A]$ can be interpreted as the initial degree of belief for the event A.

In the previous chapters, the idea was to infer the fixed parameter θ based on observations of the random variable X. This is based on the classical or *frequentist* approach to statistics. In Bayesian statistics, the parameter θ is assumed to be a priori random. More specifically, the Bayesian setup is given by

$$\begin{aligned} \theta \quad &\sim \pi(\theta), && \text{Prior} \\ \mathbf{X}|\theta &\sim \pi\left(X|\theta\right), && \text{Likelihood} \\ \theta|\mathbf{X} &\sim \pi\left(\theta|X\right) && \text{Posterior} \end{aligned}$$

Using the Bayes' theorem, we get

$$\pi\left(\theta|x\right) = \frac{\pi(\theta)\,\pi\left(x|\theta\right)}{\int_\Theta \pi(\theta')\,\pi\left(x|\theta'\right)d\theta'} \propto \pi(\theta)\,\pi\left(x|\theta\right) \tag{1.20}$$

As per the Bernstein-von Mises theorem, Bayesian statistics is mainly justified for finite samples. This theorem suggests that as the sample size increases, the influence of the prior parameter reduces. That is, the posterior becomes independent of the of the prior distribution as $N \to \infty$ (van der Vaart, 1998, p.141-142).

Prior selection

Bayesian inference strongly relies on the selection of the priors. In this paragraph we will provide some commonly used priors.

Non-informative priors If little to nothing is known about the prior distribution of θ, then the prior distribution should be as uninformative as possible. That is. the prior should affect inference as little as possible. Such priors are called non-informative, although the this name is misleading as no prior is truely non-informative (Robert, 2007). Many non-informative priors have been proposed over the years, but Jeffreys prior remains among the most commonly used non-informative prior in pratice.

Definition 1.4.6. (Jeffreys, 1961)

Jeffreys prior for a parameter $\theta \in \Theta$ is given by

$$\pi(\theta) \propto \sqrt{\mathcal{I}(\theta)} \tag{1.21}$$

Where $\mathcal{I}(\theta)$ represents the fisher information of θ contained in the likelihood function $\pi\left(y|\theta\right)$.

Jeffreys prior is considered to be non-informative because it is invariant under reparametrisation.

Conjugate Priors The posterior distribution depends on the choice of prior distribution. However, in most cases the former is intracable (van der Meulen, 2017). However, if the prior has the same distribution as the posterior, then there exists a closed form expression. This loose statement is formalised in the concept of conjugacy.

Definition 1.4.7. (A formalisation of the definition given in Lavine, 2014)

Let $\mathcal{F}$ and $\mathcal{G}$ denote classes of density functions for the likelihood function and the prior, respectively. Pick $\pi\left(y|\theta\right) \in \mathcal{F}$ and $\pi(\theta) \in \mathcal{G}$ arbitrarily. The class $\mathcal{G}$ of prior density functions is said to be conjugate for the class $\mathcal{F}$ of likelihood functions if the posterior density $\pi\left(\theta|y\right) \in \mathcal{G}$.

In words, a prior for θ is said to be conjugate if the posterior has the same distribution as the prior. A conjugate prior has the property that inference on θ can be conducted analytically.

Part I

Theoretical Analysis of Panel Data

Chapter 2

Assumption in Linear Regression models

All linear panel data model estimators derived in this thesis are based on a common set of assumptions. Because this thesis will use regression methods that may not be familiar to the reader, these assumptions will be analysed. Shortly summarised, the aim is to make reduce the amount of assumptions made compared to conventional maximum likelihood estimation OLS. This allows for a more flexible regression design. In this chapter, the following model problem will be considered

$$\underset{\mathbb{R}}{y_{i,t}} = \underset{1\times K}{x_{i,t}} \underset{K\times 1}{\beta} + \underset{\mathbb{R}}{\varepsilon_{i,t}},\ i = 1,\dots,N,\ t = 1,\dots,T \tag{2.1}$$

where N is the amount of cross-sectional units and T the amount of time periods. In T-dimensions, the model is written as

$$\underset{T\times 1}{y_i} = \underset{T\times K}{\mathbf{X}_i} \underset{K\times 1}{\beta} + \underset{T\times 1}{\varepsilon_i},\ i = 1,\dots,N \tag{2.2}$$

The parameter of interest is β. Note that this is a special case of multivariate regression, where we assume that the parameter β is *common* among the cross-sectional units and time periods. The reader is assumed to be familiar with basic properties of the OLS estimator. If this is not the case, then it may be better to read section 3.2 first and return to this chapter afterwards.

2.1 Conventions

2.1.1 Random Sampling

Throughout this thesis, a random sampling assumption is imposed. This means that $\{\mathbf{X}_i, y_i\}_{i=1}^N$ is an i.i.d. sequence. In particular, this implies that $\{\varepsilon_i\}_{i=1}^N$ is an i.i.d. sequence. Given that a population model is used, the random sampling assumption can

be described as the fact that an i.i.d. sample can be drawn from the population. This assumption is only imposed in the cross sectional dimension of the model, no assumptions are yet imposed on the dependence for the time dimension (also referred to as 'serial dependence'). In particular, $\{x_{i,t}, y_{i,t}\}_{t=1}^{T}$ may be be an arbitrarily dependent sequence.

2.1.2 Stochastic Regressors

A general convention in this thesis is that the matrix of regressors (also referred to as 'design matrix') is assumed to be stochastic rather than fixed. In regression models, the data used for linear regression is more often than not an outcome of a random variable. Assuming that this data is obtained from a deterministic process would immediately lead to a misspecification of the model. The fixed regressor assumption is usually defended with the argument that most OLS properties do not depend on assumptions on the regressors (Cf. Greene, 2011, p. 56 and Wooldridge, 2001, p. 9-10). While it is quite correct that OLS results like unbiasedness, consistency and the Gauss-Markov theorem can be derived with fixed regressors, it does pose some possible serious threats. This thesis also discusses estimators that only have a relevant interpretation under a random regressor assumption.

Example 2.1.1. Consider the maximum likelihood estimation of the model problem (2.2) under the assumption $\varepsilon_i|\mathbf{X}_i \overset{iid}{\sim} N\left(0, \sigma_\varepsilon^2 \mathbf{I}_T\right)$. Assume now, counterfactually, that the regressors are fixed. This means that it is automatically independent of ε_i and so

$$y_i \overset{iid}{\sim} N\left(\mathbf{X}_i\beta, \sigma_\varepsilon^2 \mathbf{I}_T\right),\ i = 1, \ldots, N \tag{2.3}$$

i.e. apperently the normality assumption can be checked by plotting the values of y_i (and possibly rejected) while in reality only

$$y_i|\mathbf{X}_i \overset{iid}{\sim} N\left(\mathbf{X}_i\beta, \sigma_\varepsilon^2 \mathbf{I}_T\right) \tag{2.4}$$

holds.

2.1.3 Further conventions

Conditional expectation

The conditional expectation as in definition 1.4.1 was defined conditional on a σ-algebra. And so, as in most literature, when $\mathbb{E}[\varepsilon_i|\mathbf{X}_i]$ is written, it is meant to say $\mathbb{E}[\varepsilon_i|\sigma(\mathbf{X}_i)]$, where $\sigma(\mathbf{X}_i)$ is the smallest σ-algebra containing $\mathbf{X}_i$.

Existence of the moment conditions

Throughout this thesis, no specific distributional assumption on $\mathbf{X}_i$ and ε_i are required. However, it is assumed that $\mathbf{X}_i$ and ε_i are distributed in such a way that the appropriate moments actually exist. This excludes for example that $\varepsilon_i|\mathbf{X}_i \sim \text{Cauchy}(\epsilon_0, \gamma)$.

2.2 Generalised Method of Moments

The class Generalised Method of Moments estimators proposed by Hansen (1982) is a class of semi-parametric estimators that only require moment conditions to hold. That is, no distributions need to be predetermined. This estimator is based on the analogy principle rather than the likelihood principle.

Suppose the parameters $\theta \in \mathbb{R}^K$ is to be estimated from a matrix $\mathbf{Z}_i \in T \times L$, where $L \geq K$ and the true value of this parameter is $\theta_0 \in \mathbb{R}^K$. The GMM requires, for some function $g(\mathbf{Z}_i, \theta)$,

$$\mathbb{E}[g(\mathbf{Z}_i, \theta_0)] = 0 \tag{2.5}$$

to hold.

The GMM solves for a quadratic form of the sample analogue of 2.5, i.e.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left[\sum_{i=1}^{N} g(\mathbf{Z}_i, \theta) \right]^T \hat{\mathbf{W}} \left[\sum_{i=1}^{N} g(\mathbf{Z}_i, \theta) \right], \tag{2.6}$$

where as-$\lim_{N\to\infty} \hat{\mathbf{W}} = \mathbf{W}$ and $\mathbf{W}$ is positive definite.

The GMM-estimated parameters have, under some relatively weak assumptions, some desirable properties like consistency and and asymptotic normality. In this thesis however, these properties will be proved separately for each estimator as they give a good example on how the specific assumptions are used. It is also important to note that for most estimators, the GMM estimator reduces to the more familiar Method of Moments estimator. More specifically, if θ is just-identified (necessarily $K = L$), then for linear regression, the Method of Moments estimator is appropriate.[1]

The GMM estimator described above is very general and also allows for non-linear estimation. Since only linear models will be considered, we write equation (2.6) as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \mathbf{Z}_i (y_i - \mathbf{X}_i\theta) \right]^T \hat{\mathbf{W}} \left[\sum_{i=1}^{N} \mathbf{Z}_i (y_i - \mathbf{X}_i\theta) \right]. \tag{2.7}$$

The GMM estimator for θ is given by (Wooldridge, 2001, p. 190)

$$\hat{\theta}_{GMM} = \left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{Z}_i \hat{\mathbf{W}} \mathbf{Z}_i^T \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{Z}_i \hat{\mathbf{W}} \mathbf{Z}_i^T y_i \right) \tag{2.8}$$

In most cases, $\mathbf{Z}_i = \mathbf{X}_i$, but this is formulation is slightly more general and useful in some cases of model misspecification.[2] Assume for the moment that $\mathbf{Z}_i = \mathbf{X}_i$ and assume

[1] GMM estimation is only useful in case of Generalised Instrument Variable regression, a topic which will be shortly touched upon.

[2] In case of endogeneity, we will briefly discuss this later on in this thesis.

that all the inverses $\left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{X}_i\right)^{-1}$ and $\hat{\mathbf{W}}^{-1}$ exist a.s. (we will make a more careful analysis later). In full matrix notation, the GMM estimator equals

$$\begin{aligned}\hat{\theta}_{GMM} &= \left(\mathbf{X}^T\mathbf{X}\hat{\mathbf{W}}\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\hat{\mathbf{W}}\mathbf{X}_i y\right) \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\hat{\mathbf{W}}^{-1}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\hat{\mathbf{W}}\left(\mathbf{X}y\right) \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T y\right)\end{aligned}$$

which is the famous OLS estimator. In particular, it does not matter which weighting matrix is chosen as long as $\mathbf{Z}_i = \mathbf{X}_i$ (in fact, $\left(\sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{X}_i\right)$ being invertable is enough). We will continue writing this as

$$\left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{N} \mathbf{X}_i^T y_i\right) \tag{2.9}$$

2.2.1 Exogeneity

Exogeneity is not unambiguously defined in linear regression literature. Some authors say that the matrix of regressors $\mathbf{X}_i$ is said to be exogenous if $\mathbb{E}[\mathbf{X}_i^T \varepsilon_i] = 0$, i.e. $\mathbf{X}_i$ and the error term ε_i are orthogonal (cf. Wooldridge, 2001). Furthermore, we assume that $\mathbb{E}[\varepsilon_i] = 0$. Note this is not a major assumption, since the inclusion of an intercept in the model already guarantees that $\mathbb{E}[\varepsilon_i] = 0$. The orthogonality definition of exogeneity can be defended from the perspective that a *consistent* OLS estimator can be derived under this rather weak assumption. However, for many practical purposes a stronger assumption commonly called 'strict exogeneity' is preferred over uncorrelatedness, as is also the case in this thesis. Hence, the following definition of exogeneity shall be used in this thesis

Definition 2.2.1 (Exogeneity and endogeneity, cf. Greene (2011))**.** Let $\mathbf{X}_i$ be a matrix of regressors (also commonly referred to as design matrix).

1. $\mathbf{X}_i$ is said to be **strictly exogenous** if

$$\mathbb{E}[\varepsilon_i | \mathbf{X}_i] = 0 \tag{2.10}$$

2. $\mathbf{X}_i$ is said to be **endogenous** if

$$\mathbb{E}[\mathbf{X}_i^T \varepsilon_i] \neq 0 \tag{2.11}$$

This assumption is one of the two key assumptions of linear regression (e.g. required for consistency), but is also the most difficult one to test. If endogeneity is present, then OLS-estimation is inconsistent and biased. In this case, (generalised) instrumental

variables estimation (discussed in chapter 5) may be preferred, but this is generally speaking very difficult. One way to see this is to note that endogeneity basically means that we got the causation wrong. This *could* be the result of not omitting relevant variables in the model (Cf. Greene, 2011, p. 89-93). That is, we either need to find these variables or change the causality assumption. Since the former may be impossible due to the fact that these variables are unobservable, many researchers tend to choose for the latter which is what the method of instrumental variables aims to do.

2.2.2 Full rank condition

An assumption that may look trivial at first but nonetheless one of the two core assumption of linear regression is the rank condition.

Definition 2.2.2 (Rank condition)**.** The matrix of regressors $\mathbf{X}_i$ is said to be of **full rank** if

$$\mathbb{P}[\text{rank}\left(\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{X}_i\right) = K] = 1 \tag{2.12}$$

i.e. $\text{rank}\left(\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{X}_i\right) \overset{a.s.}{=} K$.

In some literature (cf. Wooldridge, 2001), the slightly weaker assumption,

$$\text{rank}\left(\mathbb{E}[\mathbf{X}_i^T\mathbf{X}_i]\right) = K, \tag{2.13}$$

is imposed which suffices to derive asymptotic properties of the OLS estimator. The rank condition defined in equation (2.12) is more natural and this assumption in practice almost never fails for the OLS estimator anyway. This condition will prove to be challenging for some other linear estimators, but this does not depend on the formulation of the rank condition.

2.2.3 Homoscedasticity and serial correlation

The serial uncorrelatedness and homoscedasticity of the error term variance is, contrary to popular belief, an auxiliary assumption. That is, it is not required for the consistency and unbiasedness of the OLS estimator. They are required to prove the efficiency properties of the linear regression estimator.

Definition 2.2.3 (Efficiency assumptions)**.** Let $\mathbf{X}_i$ denote the matrix of regressors and ε_i the error term.

1. The variance of the error term is said to be homoscedastic if

$$\text{diag}\left(\mathbb{E}[\varepsilon_i\varepsilon_i^T|\mathbf{X}_i]\right) = \sigma_\varepsilon^2\mathbf{I}_T,\ i = 1,\dots,N \tag{2.14}$$

2. The variance of the error term is said to be heteroscedastic if

$$\text{diag}\left(\mathbb{E}[\varepsilon_i \varepsilon_i^T | \mathbf{X}_i]\right) = \text{diag}[\sigma_1^2,\ \sigma_2^2, \cdots, \sigma_T^2],\ i = 1, \ldots, N \tag{2.15}$$

and for at least one t and s, $t \neq s$, $\sigma_t^2 \neq \sigma_s^2$.

3. The variance of the error term is said to be serially uncorrelated if

$$\mathbb{E}[\varepsilon_{i,t} \varepsilon_{i,s} | \mathbf{X}_i] = 0,\ t \neq s,\ t, s = 1, \ldots, T \tag{2.16}$$

If the variance of the error term is heteroscedastic or serially correlated, then the estimates for β are still consistent (and possibly unbiased). However, the usual expression for the standard errors is wrong as these are based on the homoscedasticity assumption. To get around this problem, heteroscedastic and serial correlation robust estimators can be used. These are usually, *but not necessarily*, larger than the non-robust standard errors (Auld, 2012). Note that in presence of heteroscedasticity or serial correlation in the error term, OLS is usually no longer efficient.

2.3 Maximum Likelihood Estimation

The class of maximum likelihood estimators (hereinafter also referred to as 'MLE') is among the most common methods to estimate linear regression models (or, more generally, statistical models).

Let $\pi\left(y_i | \mathbf{X}_i; \theta\right)$[3] be a likelihood function and let $\underset{K \times 1}{\theta}$ be the estimand. Then the maximum likelihood estimator for θ equals

$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}}\, \pi\left(y_i | \mathbf{X}_i; \theta\right). \tag{2.17}$$

In addition to being consistent, maximum likelihood estimation in exponential family models are asymptotically efficient (van der Meulen, 2017, p. 38)[4]. This is a considerably stronger result than relative efficiency, which is usually best efficiency property that can be proven for estimators in the class of the generalised method of moments. At the other hand, MLE methods strongly rely on the correct specification of the model and the distributions. That is, if a model is misspecified then MLE is often inconsistent. In this section, some properties of the MLE estimator will be discussed that are needed to derive the usual OLS estimator.

2.3.1 Normality of the error term

The usual OLS estimator can only be obtained provided the error $\varepsilon_i | \mathbf{X}_i \overset{iid}{\sim} N\left(0, \mathbf{\Sigma}\right)$, where $\mathbf{\Sigma}$ is a variance-covariance matrix. That is, if the errors are not normally dis-

[3] This notation is motivated by the Bayesian notation for the likelihood function.

[4] This is a sufficient condition, not necessary. See Wooldridge, 2001, Ch. 13

tributed, then is some cases generalised linear models may be used provided the distribution of the error term is supported. Otherwise, MLE estimators are often non-linear and may not have a closed form expression.

2.3.2 Homoscedasticity

In the introduction about GMM-estimation for linear models it was shown that the homoscedasticity assumption was not actually needed to derive a consistent estimator for β. This is different in MLE estimation. That is, if the variance of the error term is heteroscedastic, then assuming $\varepsilon_i|\mathbf{X}_i \overset{iid}{\sim} N\left(0, \sigma_\varepsilon^2 I_T\right)$ is a misspecification of the model. In this particular case, it will not cause inconsistency. At the other hand, MLE is (i) not efficient and (ii) the reported standard errors are wrong, potentially leading to improper inference.

2.4 Generalised Method of Moments vs. Maximum Likelihood Estimation

The focus on this thesis will lie on GMM-based estimators because they are more robust to model misspecifications (more precisely formulated, the class of GMM estimators require less model specifictions). That, however, does not mean that GMM estimators are always more useful than MLE methods. While under appropriate assumptions the GMM-OLS estimator can be proven to be the Best Linear Unbiased Estimator[5], it does not necessarily mean it is the *Best Consistent Estimator*. As noted before, that property remains to the MLE estimator.

[5] Also known as the Gauss-Markov theorem.

Chapter 3

Non-parametric transformations

So far, the basic assumptions that are made in linear regression models have been revisited. It was argued that the generalised method of moments provides a much more flexible regression framework compared to maximum likelihood methods at the cost of efficiency properties. This chapter continues developing the linear regression framework based on the generalised method of moments. First, the pooled OLS estimator is introduced. This estimator is essentially the OLS extension for panel data models. It is also the core of this thesis since all other estimators derived in this thesis to some extent relate to pooled OLS. This is also the reason that the properties of this estimator are treated thorougly.

After deriving the pooled OLS estimator, the analysis turns to non-parametric transformations of this estimator. The term non-parametric transformation refers to the fact that the transformation itself does not induce the estimation of more *variance* parameters. While pooled OLS produces unbiased and consistent estimates under relatively weak assumptions, the transformed estimators prove to have some desirable finite sample efficiency properties in some specific cases. That is, using such estimators may dramatically improve the accuracy of the estimates compared to the ones produced by the conventional pooled OLS estimator.

3.1 Model Problem

Throughout this chapter we consider the following linear regression model:

$$\underset{\mathbb{R}}{y_{i,t}} = \underset{1\times K}{x_{i,t}} \underset{K\times 1}{\beta} + \underset{\mathbb{R}}{\varepsilon_{i,t}}, \quad i = 1, \dots, N;\ t = 1, \dots, T \tag{3.1}$$

where $y_{i,t}$ is the dependent variable, $x_{i,t}$ the explanatory variables, β a vector of regressors and $\varepsilon_{i,t}$ the error term with

$$\varepsilon_{i,t} = \mu_i + v_{i,t}. \tag{3.2}$$

μ_i is said to be the unobserved individual specific effect and $v_{i,t}$ is a remainder term. In particular, μ_i is time-invariant. As a T-dimensional vector, equation (3.1) can be

rewritten as

$$\underset{T\times 1}{y_i} = \underset{T\times K}{\mathbf{X}_i}\,\beta + \overbrace{\underset{T\times 1}{\vec{1}_T\mu_i} + \underset{T\times 1}{v_i}}^{\varepsilon_i}, \tag{3.3}$$

where $\vec{1}_T$ represents a T-dimensional vector of ones.

3.2 Pooled Ordinary Least Squares Estimator

As mentioned in the introduction of this chapter, the pooled OLS estimator is essentially the OLS extension for panel data. While the OLS estimator and its properties are well known, the GMM formulation of OLS is not. The pooled OLS forms the basis of all other estimators derived in this thesis and it is consequently important that all properties are understood.

3.2.1 Finite Sample properties

This paragraph establishes the finite sample properties of the pooled OLS estimator. Caution must be paid to the fact that the assumptions under which these properties can be derived are considerably stronger than under which similar asymptotic properties can be derived.

- Assumption F.POLS.1: $\mathbb{E}[\varepsilon_i|\{\mathbf{X}_j\}_{j=1}^N] = 0$
- Assumption F.POLS.2: $\mathbb{P}[\text{rank}\left(\sum_{i=1}^N \mathbf{X}_i^T\mathbf{X}_i\right) = K] = 1$
- Assumption F.POLS.3: $\mathbb{E}[\varepsilon_i\varepsilon_i^T|\{\mathbf{X}_j\}_{j=1}^N] = \sigma_\varepsilon^2\mathbf{I}_T$

In this finite sample analysis of the pooled OLS estimator, we want to derive the results *conditional* on $\{\mathbf{X}_i\}_{i=1}^N$ (e.g. conditional unbiasedness). These properties are due to the law of iterated expectations stronger than their unconditional counterparts. It is important to note that assumption F.POLS.2 can be attenuated to rank $\left(\mathbb{E}[\mathbf{X}_i^T\mathbf{X}_i]\right) = K$ if we are satisfied with unconditional results. The goal of this paragraph is to derive the famous Gauss-Markov theorem. This theorem basically states that the pooled OLS estimator is the Best Linear Unbiased Estimator for $\hat{\beta}$.

Pooled OLS Estimator as a GMM estimator

The assumptions formulated above suggest the use of a Generalised Method of Moments estimator. We have seen that under assumption F.POLS.1, the GMM estimator is invariant under the choice of the weighting matrix. F.POLS.1 implies that ε_i and $\mathbf{X}_i$ are orthogonal,

$$\begin{aligned}\mathbb{E}[\mathbf{X}_i^T\varepsilon_i] &= \mathbb{E}[\mathbf{X}_i^T\mathbb{E}[\varepsilon_i|\{\mathbf{X}_i\}_{i=1}^N]] \\ &= \mathbb{E}[\mathbf{X}_i\mathbb{E}[0]] \\ &= 0\end{aligned}$$

so in particular,

$$\mathbb{E}[\mathbf{X}_i^T (y_i - \mathbf{X}_i\beta)] = 0 \tag{3.4}$$

the analogy principle suggests using the sample average to estimate β

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i \left(y_i - \mathbf{X}_i\hat{\beta}\right) = 0. \tag{3.5}$$

Solving for $\hat{\beta}$ results in the **pooled OLS estimator**,

$$\hat{\beta}_{POLS} = \left(\sum_{i=1}^{N} \mathbf{X}_i^T\mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^T y_i\right), \tag{3.6}$$

where $\left(\sum_{i=1}^{N} \mathbf{X}_i^T\mathbf{X}_i\right)^{-1}$ exists a.s. due to assumption F.POLS.2.

Unbiasedness

The pooled OLS estimator is conditionally unbiased under assumptions F.POLS.1 and F.POLS.2. Indeed,

$$\mathbb{E}\left[\left(\sum_{i=1}^{N} \mathbf{X}_i^T\mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^T y_i\right) |\{\mathbf{X}_i\}_{i=1}^N\right] \overset{LIE}{=} \beta + \left(\sum_{i=1}^{N} \mathbf{X}_i^T\mathbf{X}_i\right)^{-1} \times \left(\sum_{i=1}^{N} \mathbf{X}_i^T \underbrace{\mathbb{E}[\varepsilon_i|\{\mathbf{X}_i\}_{i=1}^N]}_{0}\right)$$

which implies that, conditional on $\{\mathbf{X}_j\}_{j=1}^N$, $\hat{\beta}_{POLS}$ is unbiased. This property is also preserved unconditionally due to the law of iterated expectations:

$$\begin{aligned}\mathbb{E}[\hat{\beta}_{POLS}] = \mathbb{E}[\mathbb{E}[\hat{\beta}_{POLS}|\{\mathbf{X}_i\}_{i=1}^N]] &= \mathbb{E}[\beta] \\ &= \beta\end{aligned}$$

so under F.POLS.1-F.POLS.2, $\hat{\beta}$ is also unconditionally unbiased.

Variance

Under assumptions F.POLS.1 and F.POLS.2, the conditional variance-covariance matrix of $\hat{\beta}_{POLS}$ is given by

$$\begin{aligned}
&\text{Var}\left(\hat{\beta}|\{\mathbf{X}_i\}_{i=1}^N\right) \\
&= \text{Var}\left(\beta + \left(\sum_{i=1}^N \mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^N \mathbf{X}_i^T\varepsilon_i\right)\Bigg|\{\mathbf{X}_i\}_{i=1}^N\right) \\
&= \mathbb{E}\left[\left(\sum_{i=1}^N \mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^N \mathbf{X}_i^T\varepsilon_i\right)\left(\sum_{i=1}^N \varepsilon_i^T\mathbf{X}_i\right)\left(\sum_{i=1}^N \mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\Bigg|\{\mathbf{X}_i\}_{i=1}^N\right] \\
&\overset{RS}{=} \left(\sum_{i=1}^N \mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^N \mathbf{X}_i^T\mathbb{E}\left[\varepsilon_i\varepsilon_i^T|\{\mathbf{X}_i\}_{i=1}^N\right]\mathbf{X}_i\right)\left(\sum_{i=1}^N \mathbf{X}_i^T\mathbf{X}_i\right)^{-1}
\end{aligned}$$

This expression of the finite sample variance matrix is valid in absence of assumption F.POLS.3. That is, no restrictions have been imposed on the variance matrix. It is consequently **robust** to arbitrary serial correlation and heteroscedasticity (cf. Arellano, 1987). Under assumption F.POLS.3, this matrix reduces to the familiar variance-covariance matrix of the OLS estimator,

$$\sigma_\varepsilon^2\left(\sum_{i=1}^N \mathbf{X}_i^T\mathbf{X}_i\right)^{-1} \tag{3.7}$$

The unconditional variance-covariance matrix of the POLS-estimator is obtained by the law of total variance.

$$\text{Var}\left(\hat{\beta}\right) = \sigma_\varepsilon^2\mathbb{E}\left[\left(\sum_{i=1}^N \mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\right] \tag{3.8}$$

where we used that $\text{Var}\left(\mathbb{E}[\hat{\beta}|\{\mathbf{X}\}_{i=1}^N]\right) = 0$ due to the unbiasedness of the pooled OLS esimator.

The Gaus-Markov Theorem

The Gauss-Markov theorem proves an efficiency property of the pooled OLS estimator. This theorem states that no other linear unbiased estimator achieves a lower variance. A more precise statement is given in theorem 3.2.1. The Gauss-Markov theorem is usually proved under a fixed regressor assumption. If stochastic regressors are assumed, then the Gauss-Markov theorem is commonly called the *conditional* Gauss-Markov theorem. It is important to note that the Gauss-Markov theorem strongly relies on assumption F.POLS.3, which was not needed to derive the unbiasedness of the POLS estimator.

Theorem 3.2.1 (Gauss-Markov)**.** *Assume assumptions F.POLS.1-F.POLS.3 hold. if $\tilde{\beta}$ is another linear unbiased estimator of β, then, conditional on $\{\mathbf{X}_i\}_{i=1}^N$,*

$$\text{Var}\left(\tilde{\beta}|\{\mathbf{X}_i\}_{i=1}^N\right) - \text{Var}\left(\hat{\beta}|\{\mathbf{X}_i\}_{i=1}^N\right) \tag{3.9}$$

is a positive semidefinite matrix. That is to say, the pooled OLS estimator is the Best Linear Unbiased Estimator (BLUE) conditionally on $\{\mathbf{X}_i\}_{i=1}^N$.

Proof. Let $\bar{\beta}$ be another estimator for β given by

$$\bar{\beta} = \sum_{i=1}^{N} \underset{K\times T}{\mathbf{C}_i}\, y_i. \tag{3.10}$$

This estimator is not necessarily unbiased. To be unbiased, it must satisfy

$$\sum_{i=1}^{N} \mathbf{C}_i\mathbf{X}_i = \mathbf{I}_K.$$

To see this, note

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{N} \mathbf{C}_i y_i|\{\mathbf{X}_i\}_{i=1}^N\right] =& \mathbb{E}\left[\sum_{i=1}^{N} \mathbf{C}_i\mathbf{X}_i + \mathbf{C}_i\varepsilon_i|\{\mathbf{X}_i\}_{i=1}^N\right] \\ \overset{F.POLS.1}{=}& \left(\sum_{i=1}^{N} \mathbf{C}_i\mathbf{X}_i\right)\beta\end{aligned}$$

and so indeed, $\sum_{i=1}^N \mathbf{C}_i\mathbf{X}_i = \mathbf{I}_K$ for $\bar{\beta}$ to be unbiased.

Continuing the analysis, consider the (finite) variance of $\bar{\beta}$,

$$\begin{aligned}\text{Var}\left(\sum_{i=1}^{N} \mathbf{C}_i y_i|\{\mathbf{X}_i\}_{i=1}^N\right) &= \sum_{i=1}^{N} \mathbf{C}_i\, \text{Var}\left(\varepsilon_i|\{\mathbf{X}_i\}_{i=1}^N\right)\mathbf{C}_i^T \\ &= \sigma_\varepsilon^2 \sum_{i=1}^{N} \mathbf{C}_i\mathbf{C}_i^T \end{aligned} \tag{3.11}$$

Define $\mathbf{D}_i = \mathbf{C}_i - \left(\sum_{i=1}^N \mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\mathbf{X}_i$, or, equivalently,

$$\mathbf{C}_i = \mathbf{D}_i + \left(\sum_{i=1}^{N} \mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\mathbf{X}_i. \tag{3.12}$$

Equation 3.11 becomes

$$\begin{aligned}
\sigma_\varepsilon^2 \sum_{i=1}^{N} \mathbf{C}_i \mathbf{C}_i^T &= \sigma_\varepsilon^2 \sum_{i=1}^{N} \left(\mathbf{D}_i + \left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{X}_i \right)^{-1} \mathbf{X}_i \right) \left(\mathbf{D}_i + \left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{X}_i \right)^{-1} \mathbf{X}_i^T \right)^T \\
&= \sigma_\varepsilon^2 \sum_{i=1}^{N} \left(\mathbf{D}_i \mathbf{D}_i^T + \mathbf{D}_i \mathbf{X}_i^T \left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{X}_i^T \right)^{-1} \right) \\
&\overset{(*)}{=} \sigma_\varepsilon^2 \sum_{i=1}^{N} \mathbf{D}_i \mathbf{D}_i^T + \cancel{2 \mathbf{D}_i \mathbf{X}_i \left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{X}_i \right)^{-1}} + \left(\sum_{i=1} \mathbf{X}_i^T \mathbf{X}_i \right)^{-1} \\
&= \mathrm{Var}\left(\hat{\beta} | \{\mathbf{X}_i\}_{i=1}^N \right) + \sigma_\varepsilon^2 \sum_{i=1}^{N} \mathbf{D}_i \mathbf{D}_i^T
\end{aligned}$$

(*) follows from the fact that $\sum_{i=1}^{N} \mathbf{D}_i \mathbf{X}_i = 0$. Noting that $\sum_{i=1}^{N} \mathbf{D}_i \mathbf{D}_i^T$ is a positive semi-definite matrix, it also follows that

$$\mathrm{Var}\left(\bar{\beta} | \{\mathbf{X}_i\}_{i=1}^N \right) - \mathrm{Var}\left(\hat{\beta} | \{\mathbf{X}_i\}_{i=1}^N \right)$$

is a positive semi-definite matrix. As required. □

As noted in the introduction on the Gauss-Markov theorem, this proof strongly relied on assumption F.POLS.3. If F.POLS.3 fails, then the OLS estimator is not necessarily BLUE. There does exist an estimator that preserves the Gauss-Markov property even when assumption F.POLS.3 fails, which will be discussed later on.

The Gauss-Markov has some strengths and weaknesses. A strength of the Gauss-Markov theorem is that it is the Best Linear Unbiased Estimator for *finite samples*. Usually, comparable results for general estimators can only be derived asymptotically. For example, MLE estimation only guarantees, under some weak assumptions, that the estimator is asymptotically efficient. A weakness of the Gauss-Markov theorem is that it only proves that the POLS estimator is the best *linear* unbiased estimator. That is, it does not rule out that there may exist non-linear unbiased estimators with smaller finite sample variances.

3.2.2 Asymptotic properties

This paragraph follows chapter 7 of Wooldridge (2001). To derive a consistent OLS estimator for β, the following assumptions are imposed (these are slightly reformulated from Wooldridge (2001)),

- **Assumption A.POLS.1:** $\mathbb{E}[\mathbf{X}_i^T \varepsilon_i] = 0, \ i = 1, \ldots, N$

- **Assumption A.POLS.2:** $\operatorname{rank}\left(\mathbb{E}[\mathbf{X}_i^T\mathbf{X}_i]\right) = K \;\; i = 1, \ldots, N$
- **Assumption A.POLS.3:** $\mathbb{E}[\varepsilon_i\varepsilon_i^T|\mathbf{X}_i] = \mathbf{I}_T\sigma_\varepsilon^2, \;\; i = 1, \ldots, N$

Assumption A.POLS.1 is the weakest assumption possible to obtain a consistent estimator for β. Assumption A.POLS.2 may seem trivial, because it can only fail if $\mathbf{X}_i$ has linearly dependent rows (which is usually not the case). In fact, this assumption is often omitted in text books. This assumption is included in this thesis because it will prove to be restrictive for other estimators. It should be noted that assumption A.POLS.1 is insufficient to derive the unbiasedness of the pooled OLS estimator - the stronger assumption F.POLS.1 is required for that.

Consistency

The pooled OLS estimator consistently estimates $\hat{\beta}$ under assumption A.POLS.1 and A.POLS.2. In particular, we do not need assumption A.POLS.3 to derive the consistency of this estimator.

Theorem 3.2.2. *Under A.POLS.1 and A.POLS.2, pooled OLS is consistent. That is,*

$$\underset{n\to\infty}{\text{as-lim}}\, \hat{\beta}_{POLS} = \beta \tag{3.13}$$

Proof.

$$\begin{aligned}
&\underset{n\to\infty}{\text{as-lim}} \left(\sum_{i=1}^{n} \mathbf{X}_i^T\mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{X}_i^T y_i\right) \\
=&\beta + \underset{n\to\infty}{\text{as-lim}} \left(\sum_{i=1}^{n} \mathbf{X}_i^T\mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{X}_i^T\varepsilon_i\right) \\
\overset{CMT}{=}&\beta + \left(\underset{n\to\infty}{\text{as-lim}} \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^T\mathbf{X}_i\right)^{-1} \left(\underset{n\to\infty}{\text{as-lim}} \frac{1}{n}\sum_{i=1}^{n} X_i^T\varepsilon_i\right) \\
=&\beta + \underbrace{\mathbb{E}[\mathbf{X}_i^T\mathbf{X}_i]^{-1}}_{\text{A.POLS.2}} \underbrace{\mathbb{E}[\mathbf{X}_i^T\varepsilon_i]}_{\text{A.POLS.1}} \to 0 \\
=&\beta
\end{aligned}$$

□

Consistency is generally considered a minimum property an estimator should satisfy for asymptotic inference. Consistency for example does not imply unbiasedness ('asymptotic' pooled OLS is an example). The converse is neither true but we will not see any example of that in this thesis.

Asymptotic variance and normality

Even though no distributions were specified so far, the Central Limit Theorem in combination with the random sampling assumption does provide the $\sqrt{N}$-asymptotic normality of the pooled OLS estimator. As noted before, assumption A.POLS.3 is an auxiliary assumption. It was not needed to prove consistency and will not be needed to prove the $\sqrt{N}$-normality of the Pooled OLS estimator. This is a useful property, as assumption A.POLS.3 is rarely met in practice (i.e. heteroscedasticity and/or serial correlation in the error term). Failure of this assumption implies that the regular standard errors, which rely on assumption A.POLS.3, are incorrect. This affects reliable inference in the linear model. White (1984) provided an expression for the asymptotic variance that is robust to general heteroscedasticity. This result was extended by Arellano (1987) to take arbitrary serial correlation into account. The latter result will be derived here.

Theorem 3.2.3. *Assume assumptions A.POLS.1 and A.POLS.2 hold, then*

1. *The pooled OLS estimator is $\sqrt{N}$ asymptotically normally distributed, i.e.*

$$\sqrt{N}\left(\hat{\beta}_{POLS} - \beta\right) \xrightarrow{d} N\left(0, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right) \tag{3.14}$$

2. *The asymptotic variance of $\sqrt{N}\left(\hat{\beta}_{POLS} - \beta\right)$ is given by*

$$\text{Avar}\left(\sqrt{N}\left(\hat{\beta}_{POLS} - \beta\right)\right) = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} \tag{3.15}$$

where $\mathbf{B} = \text{Var}[\mathbf{X}_i^T\varepsilon_i]$ and $\mathbf{A} = \mathbb{E}[\mathbf{X}_i^T\mathbf{X}_i]$

Proof. (For an alternative proof, see Wooldridge, 2001).

Recall

$$\sqrt{N}\left(\hat{\beta}_{POLS} - \beta\right) = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{X}_i^T\varepsilon_i\right) \tag{3.16}$$

1. The convergence in distribution follows from the fact that

$$\begin{aligned}\underset{N\to\infty}{\text{d-lim}}\sqrt{N}\left(\hat{\beta}_{POLS} - \beta\right) &= \underset{N\to\infty}{\text{d-lim}}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{X}_i^T\varepsilon_i\right)\\ &\overset{CMT}{=} \underset{N\to\infty}{\text{d-lim}}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{A}^{-1}\mathbf{X}_i^T\varepsilon_i\\ &\overset{CLT}{=} N\left(0, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right)\end{aligned}$$

where the Central Limit Theorem can be applied due to the random sampling assumption (see paragraph 2.1.1), in particular $\{\mathbf{X}_i^T \varepsilon_i\}_{i=1}^N$ is an i.i.d. sequence.

The asymptotic variance equals

$$\begin{aligned}
&\text{Avar}\left[\sqrt{N}\left(\hat{\beta}_{POLS} - \beta\right)\right] \\
=&\underset{n\to\infty}{\text{as-lim}} \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{X}_i\right)^{-1} \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \mathbf{X}_i^T \varepsilon_i\right) \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \mathbf{X}_i^T \varepsilon_i\right)^{T} \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{X}_i\right)^{-1} \\
\overset{RS}{=}&\underset{n\to\infty}{\text{as-lim}} \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{X}_i\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i^T \varepsilon_i \varepsilon_i^T \mathbf{X}_i\right) \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{X}_i\right)^{-1} \\
\overset{CMT}{=}&\underbrace{\mathbb{E}[\mathbf{X}_i^T \mathbf{X}_i]^{-1}}_{\mathbf{A}^{-1}} \underbrace{\mathbb{V}[\mathbf{X}_i^T \varepsilon_i]}_{\mathbf{B}} \underbrace{\mathbb{E}[\mathbf{X}_i^T \mathbf{X}_i]^{-1}}_{\mathbf{A}^{-1}}
\end{aligned}$$

which proves the asymptotic variance part. □

This result suggests that the usual standard errors are incorrect in absence of assumption A.POLS.3. However, this theorem suggests there exists an estimable expression for the asymptotic variance-covariance matrix $\text{Avar}\left(\hat{\beta}_{POLS}\right)$. Note that the asymptotic variance-covariance matrix of $\hat{\beta}_{POLS}$ is given by

$$\text{Avar}\left(\hat{\beta}_{POLS}\right) = \frac{\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}}{N}. \tag{3.17}$$

Taking the sample analogue of $\mathbf{A}$, the estimator

$$\hat{\mathbf{A}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{X}_i \tag{3.18}$$

consistently estimates $\mathbf{A}$. To obtain a consistent estimator for $\mathbf{B}$, we define the pooled OLS residuals as $\hat{\varepsilon}_{i,t} = y_i - \mathbf{X}_i \hat{\beta}_{POLS}$. A consistent estimator for $\mathbf{B}$ is given by (Cf. Wooldridge, 2001, p. 160)

$$\hat{\mathbf{B}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i \hat{\varepsilon}_i \hat{\varepsilon}_i^T \mathbf{X}_i. \tag{3.19}$$

A consistent estimator for $\text{Avar}\left(\hat{\beta}_{POLS}\right)$ is obtained by substituting $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ in equation (3.17), i.e.

$$\hat{\text{Avar}}\left(\hat{\beta}_{POLS}\right) = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}. \tag{3.20}$$

This estimator of the asymptotic variance-covariance matrix is called the estimator for the **robust asymptotic variance-covariance matrix**.

If assumption A.POLS.3 is met, the expression of the variance reduces to a more familiar expression.

Corollary 3.2.1. *Assume assumptions A.POLS.1 - A.POLS.3 hold, then*

$$\text{Avar}\left(\sqrt{N}\left(\hat{\beta}_{POLS}-\beta\right)\right)=\sigma^2\mathbb{E}\left[\mathbf{X}_i^T\mathbf{X}_i\right]^{-1} \tag{3.21}$$

Proof. In this case, matrix $\mathbf{B}=\sigma_\varepsilon^2\mathbb{E}[\mathbf{X}_i^T\mathbf{X}_i]$. Substituting this in the expression for the robust variance-covariance matrix proves the required. □

Asymptotic Efficiency

The Gauss-Markov theorem suggests that, under appropriate assumptions, the Pooled OLS estimator is the Best Linear Unbiased Estimator for finite samples. Asymptotic efficiency is considerably more difficult, because it must attain the Cramer-Rao lower bound. The class of GMM estimators in general cannot be shown to satisfy this property since no distributions are specified.

In this thesis, the model problem will often be transformed by pre-multiplying equation (3.3) with a transformation matrix $\underset{S\times T}{\mathbf{Q}}$, after which Pooled OLS is applied on the transformed system. This results in a certain estimator $\bar{\beta}$. That is, $\bar{\beta}$ is the pooled OLS estimator applied on

$$\mathbf{Q}y_i=\mathbf{Q}\mathbf{X}_i\beta+\mathbf{Q}\varepsilon_i \tag{3.22}$$

Throughout this thesis, $\bar{\beta}$ is (i) consistent and (ii) $\sqrt{N}\left(\bar{\beta}-\beta\right)\to N\left(0,\mathbf{D}\right)$. These can be proven to hold provided the following assumptions are imposed.

- **Assumption A.POLS.1(b):** $\mathbb{E}[\varepsilon_i|\mathbf{X}_i]=0$
- **Assumption A.POLS.2(b):** $\text{rank}\left(\mathbb{E}[\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i]\right)=K$.

Assumption A.POLS.1(b) implies assumption A.POLS.1. The central asymptotic efficiency question in this thesis is whether the estimator $\hat{\beta}$ outperforms $\bar{\beta}$. That is,

$$\text{Avar}\left(\hat{\beta}\right)-\text{Avar}\left(\bar{\beta}\right) \tag{3.23}$$

is a positive semidefinite matrix. In that case, $\hat{\beta}$ is said to be efficient relative to $\bar{\beta}$.

Theorem 3.2.4 (Efficiency of the pooled OLS estimator — Mostly own work)**.** *Assume assumptions A.POLS.1(b), A.POLS.2, A.POLS.2(b) and A.POLS.3 hold. Let $\bar{\beta}$ be the estimator obtained after applying Pooled OLS on the transformed system given in equation* (3.22)*. Then,*

$$\text{Avar}\left(\bar{\beta}\right)-\text{Avar}\left(\hat{\beta}_{POLS}\right) \tag{3.24}$$

is a positive semi-definite matrix.

Proof. Under assumption A.POLS.1 - A.POLS.3, $\text{Avar}\left(\hat{\beta}_{POLS}\right) = \sigma_\varepsilon^2 \mathbb{E}[\mathbf{X}_i^T \mathbf{X}_i]^{-1}/N$, where we will denote $\mathbf{V} = \sigma_\varepsilon^2 \mathbb{E}[\mathbf{X}_i^T \mathbf{X}_i]^{-1}$. The asymptotic variance of $\bar{\beta}$ is given by $\text{Avar}\left(\bar{\beta}\right) = \mathbf{D}/N$. Matrix $\mathbf{D}$ is given by

$$\begin{aligned}
\mathbf{D} &= \text{Avar}\left(\sqrt{N}\left(\bar{\beta} - \beta\right)\right) \\
&= \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{Q}^T \mathbf{Q} \mathbf{X}_i\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{Q}^T \varepsilon_i \varepsilon_i^T \mathbf{Q} \mathbf{X}_i\right) \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{Q}^T \mathbf{Q} \mathbf{X}_i\right)^{-1} \\
&\overset{LLN}{\to} \mathbb{E}[\mathbf{X}_i^T \mathbf{Q}^T \mathbf{Q} \mathbf{X}_i]^{-1} \,\text{Var}\left(\mathbf{X}_i^T \mathbf{Q}^T \mathbf{Q} \varepsilon_i\right) \mathbb{E}[\mathbf{X}_i^T \mathbf{Q}^T \mathbf{Q} \mathbf{X}_i]^{-1} \\
&= \left(\sigma_\varepsilon^2\right)^2 \text{cov}\left(\mathbf{X}_i^T \mathbf{Q}^T \mathbf{Q} \varepsilon_i, \mathbf{X}_i^T \varepsilon_i\right)^{-1} \text{Var}\left(\mathbf{X}_i^T \mathbf{Q}^T \mathbf{Q} \varepsilon_i\right) \text{cov}\left(\mathbf{X}_i^T \mathbf{Q}^T \mathbf{Q} \varepsilon_i, \mathbf{X}_i^T \varepsilon_i\right)^{-1}
\end{aligned}$$

Analogously, the asymptotic variance of $\mathbf{V}$ can be written ass

$$\mathbf{V} = \left(\sigma_\varepsilon^2\right)^2 \text{Var}\left(\mathbf{X}_i^T \varepsilon_i\right)^{-1}$$

We have now written the asymptotic variance of $\bar{\beta}$ in such a way that it can easily be compared with $\hat{\beta}$. In fact, we can now follow the method used by Newey and McFadden (1983/2007, vol. 4, p. 2166) to prove $\mathbf{D} - \mathbf{V}$ is positive semi-definite.[1] Let $\mathbf{A}_1 = \mathbf{X}_i^T \varepsilon_i$ and $\mathbf{A}_2 = \mathbf{X}_i^T \mathbf{Q}^T \mathbf{Q} \varepsilon_i$, then

$$\begin{aligned}
\mathbf{D} - \mathbf{V} &\propto \text{cov}\left(\mathbf{A}_2, \mathbf{A}_1\right)^{-1} \left(\text{Var}\left(\mathbf{A}_2\right) - \text{cov}\left(\mathbf{A}_2, \mathbf{A}_1\right) \text{Var}\left(\mathbf{A}_1\right)^{-1} \text{cov}\left(\mathbf{A}_1, \mathbf{A}_2\right)\right) \text{cov}\left(\mathbf{A}_1, \mathbf{A}_2\right)^{-1} \\
&= \text{cov}\left(\mathbf{A}_2, \mathbf{A}_1\right)^{-1} \text{Var}\left(\mathbf{A}_2 - \text{cov}\left(\mathbf{A}_1, \mathbf{A}_2\right) \text{Var}\left(\mathbf{A}_1\right)^{-1} \mathbf{A}_1\right) \text{cov}\left(\mathbf{A}_1, \mathbf{A}_2\right)^{-1},
\end{aligned}$$

so $\mathbf{D} - \mathbf{V}$ is proportional to a PSD matrix. Since $\sigma_\varepsilon^2 > 0$, we have established that $\mathbf{D} - \mathbf{V}$ is in fact a PSD matrix. That is, no transformed system outperforms $\hat{\beta}_{POLS}$. As required. □

3.2.3 Simulations

In this paragraph, we will consider some simulations of the Pooled OLS estimator. In particular, we want to show that the normality assumption is not actually needed for the theoretical results to hold. Both a small sample and large sample simulation are considered.

Small sample simulation

Consider a panel data model with $N = 50$ cross sectional units and $T = 20$ time periods. Two models are simulated in this paragraph. Throughout this thesis, the abbriviation SS stands for Small Sample.

[1] They prove that the MLE estimator is asymptotically more efficient than minimum distance estimators, yet their proof can be applied in a similar fashion.

Case 1: Gauss-Markov assumptions hold The model is simulated under the following assumption.

- Assumption SS.POLS.1: $\varepsilon_i|\mathbf{X}_i \overset{iid}{\sim} \tau_3(0, \mathbf{I}_T)$, where $\tau_\nu(\mu, \mathbf{\Sigma})$ denotes a student t distribution with ν degrees of freedom, μ a location vector and $\mathbf{\Sigma}$ a shape matrix.

and consider the model

$$y_i = \alpha + \mathbf{X}_i\beta \tag{3.25}$$

where $\mathbf{X}_i$ is some arbitrary generated data based following a normal distribution. The true values of the parameters are $\alpha = 1$ and $\beta = 2$.

The Pooled OLS regression is performed $R = 3000$ times. Table 3.1 presents the results for the simulation, where we especially wish to compare the type I and type II errors made in this simulation. A type I error refers to the fact that the estimate was significantly different from the true value of the parameter, while a type II error refers to the fact that the estimate was not significantly different from 0 (both at a 5% level). In statistics, type I errors are generally considered worse than type II errors (although it is not as black and white as text books tell and the consequences should be determined case-by-case). The amount of type I errors are expected to be around 5% in both cases.

The amount of type I errors are, as expected, around 5%. Furthermore, no type II errors were observed. We note that the maximum likelihood standard errors on average are lower than the pooled OLS standard errors. This is not an indication that the Gauss-Markov assumptions are not satisfied, since the MLE estimator for the parameter β is non-linear in this case (Liu and Rubin, 1995). This illustrates that Gauss-Markov does not have power against non-linear estimators. While the MLE estimator would theoretically be preferred as the estimator of choice, this simulation clearly shows that pooled OLS in absence of normality may still produce very accurate estimates.

Table 3.1: Pooled OLS summary (R=3000 simulations)

	Estimate	Std. Error	t value	Pr(>|t|)	Type I	Type II
(Intercept)	1.0018	0.0542	19.2859	0	4.3%	0%
x	1.9997	0.0207	101.1183	0	5.33%	0%

Table 3.2: MLE estimation summary (R=3000 simulations)

	Estimate	Std.Error	Z value	p-value	Type I	Type II
(Intercept)	1.0016	0.0396	25.4376	0	4.83%	0%
beta	1.9996	0.0151	133.41	0	5.5%	0%

Case 2: Heteroscedasticity In this example we compare the MLE estimator with the Pooled OLS estimator under heteroscedasticity.

- Assumption SS.POLS.1: $\varepsilon_i \overset{iid}{\sim} \tau_3\left(0, \text{diag}\left(\sigma_t^2\right)\right)$, where $\sigma_t^2 = t^2$

Under assumption SS.POLS.1, the pooled OLS estimator is still supposed to produce unbiased estimates. The results are based on heteroscedastic robust standard errors. Note that pooled OLS no longer satisfies the Gauss-Markov theorem, i.e. there may exist another linear estimator that outperforms pooled OLS.[2] A summary of the simulation is presented in table 3.3. Note that pooled OLS produces about the same standard errors and as the maximum likelihood estimator, although the amount of type I errors produced by pooled OLS is considerably larger than the MLE type I errors. This is likely a result of overestimation of the actual standard errors in the MLE estimates, meaning that these estimates in reality are more accurate than displayed here.

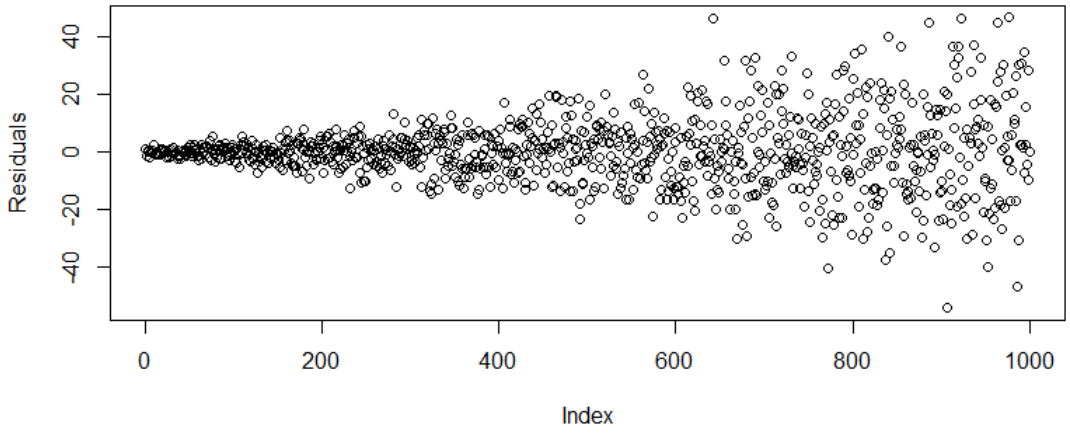
Table 3.3: Summary Pooled OLS (R=3000 simulations)

	Estimate	Std. Error	t value	Pr(>|t|)	Type I	Type II
(Intercept)	1.001	0.3854	2.6002	0.0658	5.03%	26.2%
x	2.0018	0.1464	13.741	0	4.83%	0%

Table 3.4: Summary MLE (R=3000 simulations)

	Estimate	Std.Error	Z value	p-value	Type I	Type II
(Intercept)	0.9961	0.3073	3.2466	0.02	4.13%	
x	2.0039	0.117	17.1742	0	4%	

Figure 3.1: Residuals pooled OLS estimator



[2]We will find out later that such estimator indeed exists.

Large sample simulations

In this paragraph we will consider some large sample simulations for the Pooled OLS estimator. Given the nature of this research, this means that the amount of cross sectional units N is large while the amount of periods T remains fixed. In particular, we consider a panel of $N = 1000$ and $T = 20$. Only the estimates of the pooled OLS estimator will be presented as maximum likelihood is computationally too expensive.

Case: Efficiency assumptions holds

Again we assume

- Assumption LS.POLS.1: $\varepsilon_i|\mathbf{X}_i \overset{iid}{\sim} \tau_3(0, \mathbf{I}_T)$.

and so the efficiency assumptions hold. Table 3.5 presents the results for the Pooled OLS estimator. From this table we can observe that the estimates are very accurate (i.e. low standard errors) compared to its small sample counterpart.

Table 3.5: Summary pooled OLS (R=3000 simulations)

	Estimate	Std. Error	t value	Pr(>|t|)	Type I	Type II
(Intercept)	1.0018	0.0542	19.2859	0	4.3%	0%
beta	1.9997	0.0207	101.1183	0	5.33%	0%

3.2.4 Bayesian Inference

The pooled OLS estimator can also be described in a Bayesian context. In this paragraph, we will show that the maximum a posteriori (MAP) estimator may in fact be equal to the pooled OLS estimator. MAP estimation is not very common in Bayesian statistics. The reason why we opt for MAP estimation is the fact that it also makes a link with the MLE estimator. That is, we will use a prior such that the MAP estimator is in fact also the MLE estimator.

- **Assumption B.POLS.1:** $\varepsilon_i|\mathbf{X}_i \overset{iid}{\sim} N\left(0, \sigma_\varepsilon^2 \mathbf{I}_T\right)$
- **Assumption B.POLS.2:** $\mathbb{P}[\text{rank}\left(\sum_{i=1}^N \mathbf{X}_i^T \mathbf{X}_i\right) = K] = 1$

These assumptions imply that

$$y_i|\mathbf{X}_i, \beta, \sigma_\varepsilon^2 \overset{iid}{\sim} N\left(\mathbf{X}_i\beta, \mathbf{I}_T\sigma_\varepsilon^2\right) \tag{3.26}$$

and so the likelihood function is given by

$$\pi\left(y_{1:N}|\mathbf{X}_{1:N},\beta,\sigma_\varepsilon^2\right) \propto \prod_{i=1}^{N}\left(\frac{1}{\sigma^2}\right)^{T/2}\exp\left(-\frac{1}{2\sigma_\varepsilon^2}\left(y_i-\mathbf{X}_i\beta\right)^T\left(y_i-\mathbf{X}_i\beta\right)\right)$$
$$= \left(\frac{1}{\sigma_\varepsilon^2}\right)^{NT/2}\exp\left(-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{N}\left(y_i-\mathbf{X}_i\beta\right)^T\left(y_i-\mathbf{X}_i\beta\right)\right). \quad (3.27)$$

We impose a non-informative Jeffreys prior. The prior distribution is given by (Box and Tiao, 1992, p. 426)

$$\pi\left(\mu,\sigma_\varepsilon^2\right) \propto \frac{1}{\sigma_\varepsilon^2}. \quad (3.28)$$

In that case, the posterior is given by

$$\pi\left(\theta|y_{1:N},\mathbf{X}_{1:N}\right) \propto \left(\frac{1}{\sigma_\varepsilon^2}\right)^{NT/2+1}\exp\left(-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{N}\left(y_i-\mathbf{X}_i\beta\right)^T\left(y_i-\mathbf{X}_i\beta\right)\right) \quad (3.29)$$

Under this particular prior, the maximum a-posteriori estimator for β is in fact the pooled OLS estimator.

3.3 Fixed Effects Estimator

In the introduction of this chapter, the error term was written as $\varepsilon_i = \mu_i\vec{1}_T + v_i$. The pooled OLS estimator "ignored" the unobserved individual specific fixed effect μ_i. The estimates provided by the pooled OLS estimator are hence only consistent provided $\mathbb{E}[\mu_i|\mathbf{X}_i] = 0$. But if $\text{cov}\left(\mu_i,\mathbf{X}_i\right) \neq 0 \implies \mathbb{E}[\mu_i|\mathbf{X}_i] \neq 0$, then pooled OLS is inconsistent due to endogeneity with the unobserved individual specific fixed effect. The fixed effects estimator provides a first relaxation of the strict exogeneity assumption as formulated in assumption A.POLS.1. In fixed effect models, μ_i is assumed to be a parameter in the model and $v_{i,t}$ the (random) error term. In particular, the fixed effects estimator allows $\mathbb{E}[\mu_i|\mathbf{X}_i] \neq 0$. The results for this estimator are only derived asymptotically. This section follows chapter 10 of Wooldridge (2001). Under the following assumptions, a consistent estimator can be derived for a model with unobserved individial specific fixed effects.

3.3.1 Derivation of the estimator

In order to derive the fixed effect estimator, the model problem given in equation (3.1) is first averaged over time. That is,

$$\bar{y}_i = \bar{x}_i + \mu_i + \bar{\varepsilon}_i, \quad (3.30)$$

where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{i,t}$, $\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{i,t}$ and $\bar{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{i,t}$. Substracting equation (3.30) from the model problem results in

$$\begin{aligned} y_{i,t} - \bar{y}_i &= \left(x_{i,t} - \bar{x}_i\right)\beta + \left(\varepsilon_{i,t} - \bar{\varepsilon}_i\right) \\ &:= \dot{y}_{i,t} = \dot{x}_{i,t}\beta + \dot{\varepsilon}_{i,t} \end{aligned} \tag{3.31}$$

The individual specific effect μ_i disappeared due to its time invariance. The fixed effects estimator is obtained by applying the Pooled OLS estimator on (3.31), i.e.

$$\hat{\beta}_{FE} = \left(\sum_{i=1}^{N} \dot{\mathbf{X}}_i^T \dot{\mathbf{X}}_i\right)^{-1} \left(\sum_{i=1}^{N} \dot{\mathbf{X}}_i^T \dot{y}_i\right) \tag{3.32}$$

In vector form, equation (3.31) is written as

$$\underset{T\times 1}{\dot{y}_i} = \underset{T\times K}{\dot{\mathbf{X}}_i} \underset{K\times 1}{\beta} + \underset{T\times 1}{\dot{\varepsilon}_i} \tag{3.33}$$

To derive the fixed effects estimator with a transformation matrix, the time-demeaning matrix can be written as (Baltagi, 2005, p. 12)

$$\underset{T\times T}{\mathbf{Q}} = \mathbf{I}_T - \vec{1}_T \left(\vec{1}_T^T \vec{1}_T\right)^{-1} \vec{1}_T^T. \tag{3.34}$$

This is an idempotent matrix with rank $T-1$. Pre-multiplying $\mathbf{Q}$ with equation (3.31) results in the set of equations (3.33).

3.3.2 Asymptotic properties

- **Assumption A.FE.1**: $\mathbb{E}\left[v_i|\mathbf{X}_i, \mu_i\right] = 0$.
- **Assumption A.FE.2**: $\text{rank}\left(\sum_{t=1}^{T} \mathbb{E}\left[\left(x_{i,t} - \bar{x}_i\right)^T \left(x_{i,t} - \bar{x}_i\right)\right]\right) = K$
- **Assumption A.FE.3**: $\mathbb{E}[\varepsilon_i \varepsilon_i^T|\mathbf{X}_i, \mu_i] = \sigma_\varepsilon^2 \mathbf{I}_T$

Again, note that $\mathbb{E}[\mu_i|\mathbf{X}_i]$ may be unequal to 0. This property comes at a price, as per assumption A.FE.2 time-invariant data cannot be included in the regression.

Consistency

A natural question that may arrise is whether the fixed effects transformation under assumptions A.FE.1 - A.FE.3 still yields a consistent estimator for β. The answer is given in corollary 3.3.1.

Corollary 3.3.1. *Assume assumption A.FE.1 - A.FE.2 hold, then fixed effects estimator consistently estimates β. That is,*

$$\underset{N\to\infty}{\text{as-lim}}\, \hat{\beta}_{FE} = \beta$$

Proof. Because the Fixed Effects estimator is a POLS-estimator on equation (3.33), it suffices to show that assumption POLS.1 holds under A.FE.1. for the transformed systsem, i.e. $\mathbb{E}[\dot{\mathbf{X}}_{i,t}^T \dot{\varepsilon}_{i,t}] = 0$. Indeed,

$$\begin{aligned}\mathbb{E}[\dot{x}_{i,t}^T \dot{v}_{i,t}] &= \mathbb{E}[\left(x_{i,t} - \bar{x}_i\right)^T \left(v_{i,t} - \bar{\varepsilon}_i\right)] \\ &= \mathbb{E}[x_{i,t}^T v_{i,t} - x_{i,t}^T \bar{v}_i - \bar{x}_i v_{i,t} + \bar{x}_i \bar{\varepsilon}_i]\end{aligned}$$

All these terms are, due to A.FE.1, again uncorrelated (this is easily proved using the Law of Iterated Expectations). So indeed,

$$\mathbb{E}[\dot{\mathbf{X}}_{i,t}^T \dot{\varepsilon}_{i,t}] = 0$$

□

Serial correlation

The Fixed Effects transformation was obtained by substracting the time series average. This induces serial correlation in the system, as shown in corollary 3.3.2. Note that under assumption A.FE.3 (required for efficiency) there are no obvious restrictions for the transformed errors.

Corollary 3.3.2. *Assume assumptions A.FE.1 - A.FE.3 hold. Then, the correlation coefficient for the transformed error is given by*

$$\rho\left(\dot{\varepsilon}_{i,t}, \dot{\varepsilon}_{i,s}\right) = -\frac{1}{T-1} \tag{3.35}$$

Proof. (Clarification of the proof given by Wooldridge (2001, p. 270))

Recall that the correlation coefficient of equation (3.35) is defined as

$$\rho\left(\dot{\varepsilon}_{i,t}, \dot{\varepsilon}_{i,s}\right) = \frac{\operatorname{cov}\left(\dot{\varepsilon}_{i,t}, \dot{\varepsilon}_{i,s}\right)}{SD\left(\dot{\varepsilon}_{i,t}\right) SD\left(\dot{\varepsilon}_{i,s}\right)} \tag{3.36}$$

where $SD\left(\dot{\varepsilon}_{i,t}\right) = \sqrt{\operatorname{Var}\left(\dot{\varepsilon}_{i,t}\right)}$. First, an expression for the covariance will be derived.

$$\begin{aligned}\operatorname{cov}\left(\dot{\varepsilon}_{i,t}, \dot{\varepsilon}_{i,s}\right) &= \mathbb{E}[\left(\varepsilon_{i,t} - \bar{\varepsilon}_i\right)\left(\varepsilon_{i,s} - \bar{\varepsilon}_i\right)] \\ &= \mathbb{E}[\varepsilon_{i,t}\varepsilon_{i,s}] - \mathbb{E}[\varepsilon_{i,t}\bar{\varepsilon}_i] - \mathbb{E}[\varepsilon_{i,s}\bar{\varepsilon}_i] + \mathbb{E}[\bar{\varepsilon}_i^2]\end{aligned} \tag{3.37}$$

The first term equals zero due to assumption A.FE.3, since the original errors are serially uncorrelated. As for the second term,

$$\begin{aligned}\mathbb{E}\left[\varepsilon_{i,t}\bar{\varepsilon}_i\right] &= \frac{1}{T}\mathbb{E}\left[\varepsilon_{i,t}\sum_{j=1}^{T}\varepsilon_{i,j}\right] \\ &= \frac{1}{T}\sigma_\varepsilon^2\end{aligned} \tag{3.38}$$

since $\varepsilon_{i,t}$ and $\varepsilon_{i,s}$ are serially uncorrelated for $s \neq t$. The expression for the third term is the same second term. The expression for the fourth term is given by

$$\begin{aligned}\mathbb{E}[\bar{\varepsilon}_i^2] &= \frac{1}{T^2}\mathbb{E}\left[\sum_{j=1}^{T}\varepsilon_{i,j}\sum_{k=1}^{T}\varepsilon_{i,k}\right] \\ &= \frac{1}{T^2}\sum_{j=1}^{T}\mathbb{E}[\varepsilon_{i,l}^2] \\ &= \frac{1}{T}\sigma_\varepsilon^2 \end{aligned} \tag{3.39}$$

Substituting these results in equation (3.38) results in

$$\text{cov}\left(\dot{\varepsilon}_{i,t}, \dot{\varepsilon}_{i,s}\right) = -\frac{1}{T}\sigma_\varepsilon^2. \tag{3.40}$$

An expression for the standard deviation is also required in order to derive the serial correlation of the transformed errors. Noting that $\text{Var}\left(\dot{\varepsilon}_{i,t}\right) = \mathbb{E}[\dot{\varepsilon}_{i,t}^2]$ due to A.FE.1,

$$\begin{aligned}\mathbb{E}[\dot{\varepsilon}_{i,t}] &= \mathbb{E}[\varepsilon_{i,t}^2] - 2\mathbb{E}[\varepsilon_{i,t}\bar{\varepsilon}_i] + \mathbb{E}[\bar{\varepsilon}_i^2] \\ &= \sigma_\varepsilon^2 - \frac{2}{T}\sigma_\varepsilon^2 + \frac{1}{T}\sigma_\varepsilon^2 \\ &= \sigma_\varepsilon^2\left(1 - \frac{1}{T}\right) \end{aligned} \tag{3.41}$$

i.e.

$$SD\left(\dot{\varepsilon}_{i,t}\right) = \sqrt{\sigma_\varepsilon^2\left(1 - \frac{1}{T}\right)}. \tag{3.42}$$

Combining equations (3.38) and (3.42) results in

$$\rho\left(\dot{\varepsilon}_{i,t}, \dot{\varepsilon}_{i,s}\right) = -\frac{1}{T-1}, \tag{3.43}$$

as required. □

These results have some implications. The transformed system induces negative serial correlation which tends to 0 as $T \to \infty$. Also, the expression for the variance given in equation (3.41) suggests that it does not depend on t, i.e. it is heteroscedastic across t.

Asymptotic normality and variance

Given the proof for corollary 3.3.1, it should not come as a surprise that $\sqrt{N}\left(\hat{\beta} - \beta\right)$ is asymptotically normally distributed.

Corollary 3.3.3. *Assume assumptions A.FE.1 and A.FE.2 hold. Then,*

$$\sqrt{N}\left(\hat{\beta}_{FE}-\beta\right) \xrightarrow{d} N\left(0, \mathbf{A}^{-1}\operatorname{Var}\left(\dot{\mathbf{X}}_i\dot{\varepsilon}_i\right)\mathbf{A}^{-1}\right) \tag{3.44}$$

where $\mathbf{A}=\mathbb{E}[\dot{\mathbf{X}}_i^T\dot{\mathbf{X}}_i]$.

If, in addition, A.FE.3 holds, then $\mathbf{A}^{-1}\operatorname{Var}\left(\dot{\mathbf{X}}_i\dot{\varepsilon}_i\right)\mathbf{A}^{-1}=\mathbf{A}^{-1}$ *and so*

$$\operatorname{Avar}\left(\hat{\beta}_{FE}\right)=\frac{\mathbf{A}^{-1}}{N} \tag{3.45}$$

We will not prove the asymptotic $\sqrt{N}$-normality of the fixed effects estimator, because it analogous to the proof of the asymptotic $\sqrt{N}$-normality of the pooled OLS estimator. The structure of the asymptotic variance is a special property of the fixed effects estimator which we will investige further.[3]

Proof. The asymptotic variance of the fixed effects estimator is given by

$$\begin{aligned}\operatorname{Avar}\left(\hat{\beta}_{FE}\right) =&\mathbb{E}[\dot{\mathbf{X}}_i^T\dot{\mathbf{X}}_i]^{-1}\mathbb{E}[\dot{\mathbf{X}}_i^T\dot{\varepsilon}_i\dot{\varepsilon}_i^T\dot{\mathbf{X}}_i]\mathbb{E}[\dot{\mathbf{X}}_i^T\dot{\mathbf{X}}_i]^{-1}\\ =&\mathbb{E}[\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i]^{-1}\mathbb{E}[\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\varepsilon_i\varepsilon_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i][\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i]^{-1}\end{aligned}$$

Noting that $\mathbf{Q}$ is idempotent, this equation reduces to (applying the LIE),

$$\begin{aligned}=&\sigma_\varepsilon^2\mathbb{E}[\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i]^{-1}\mathbb{E}[\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i][\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i]^{-1}\\ =&\sigma_\varepsilon^2\mathbb{E}[\dot{\mathbf{X}}_i^T\dot{\mathbf{X}}_i]^{-1}\end{aligned}$$

As required. □

Asymptotic Efficiency

This paragraph rejects a claim of Wooldridge (2001, p. 269) about the fixed effects estimator. He suggests this estimator is the most efficient estimator under assumption A.FE.1 - A.FE.3 (without proof). This claim is false. Consider the following theoretical model;

- **Assumption AM.FE.1:** Assumption A.POLS.1 holds
- **Assumption AM.FE.2:** Assumption A.FE.2 holds
- **Assumption AM.FE.3:** Assumption A.POLS.3 holds

[3] It is special in the sense that it does **not** have the structure of a robust variance-covariance matrix, i.e. $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ reduces to $\mathbf{A}^{-1}$ which is not at all obvious under assumption A.FE.3.

Assumption AM.FE.1 and AM.FE.3 imply there are no unobserved fixed effects. Assumption AM.FE.2 in particular means there are no time-invariant regressors. Under this theoretical model, both A.POLS.1 - A.POLS.3 and A.FE.1 - A.FE.3 hold.

Theorem 3.2.4 (asymptotic efficiency of the pooled OLS) suggests that under POLS.1 - POLS.3, no transformation outperforms the pooled OLS estimator. So at best, the fixed effects estimator is as efficient as the pooled OLS estimator. Given the expression of the fixed effects asymptotic variance-covariance matrix, this is not an unreasonable conjecture. However, another hypothesis is that the fixed effects estimator is less efficient than pooled OLS since the fixed effects transformation induces serial correlation in the system. Theorem (3.3.1) confirms the latter hypothesis. Before turning to this theorem, the following linear algebra fact is useful.

Lemma 3.3.1. *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N \times N}$ *be non-singular matrices. If*

$$\mathbf{A}^{-1} - \mathbf{B}^{-1} \tag{3.46}$$

is a positive semi-definite matrix, then

$$\mathbf{B} - \mathbf{A} \tag{3.47}$$

is a positive semi-definite matrix.

Theorem 3.3.1 (Own work)**.** *Under Assumption AM.FE.1 - AM.FE.3, the Pooled OLS estimator is asymptotically more efficient than the fixed effects estimator, i.e.*

$$\text{Avar}\left(\hat{\beta}_{FE}\right) - \text{Avar}\left(\hat{\beta}_{POLS}\right) \tag{3.48}$$

is a positive semi-definite matrix.

Proof. To prove this result, we use the time demeaning matrix of equation (3.34). So the asymptotic variance of the fixed effects estimator is given by

$$\begin{aligned} \text{Avar}\left(\hat{\beta}_{FE}\right) &= \mathbb{E}[\mathbf{X}_i^T \mathbf{Q}^T \mathbf{Q} \mathbf{X}_i] \\ &= \mathbb{E}[\mathbf{X}_i^T \mathbf{Q} \mathbf{X}_i]^{-1}, \end{aligned}$$

where the last equality follows due to the fact that $\mathbf{Q}$ is idempotent. Consider now

$$\begin{aligned} \text{Avar}\left(\hat{\beta}_{POLS}\right)^{-1} - \text{Avar}\left(\hat{\beta}_{FE}\right)^{-1} &= \sigma_\varepsilon^2 \mathbb{E}[\mathbf{X}_i^T \mathbf{X}_i - \mathbf{X}_i^T \mathbf{Q} \mathbf{X}_i] \\ &= \mathbb{E}[\mathbf{X}_i^T \left(\mathbf{I}_T - \mathbf{Q}\right) \mathbf{X}_i] \end{aligned} \tag{3.49}$$

so

$$\mathbf{I}_T - \mathbf{Q} = \frac{1}{T}\left(\vec{1}_T \vec{1}_T^T\right), \tag{3.50}$$

which is proportional to a matrix consisting of ones. This matrix is obviously PSD, suggesting that matrix (3.49) is a positive semi-definite. To finish the proof, note

$$\begin{aligned} &\text{Avar}\left(\hat{\beta}_{POLS}\right)^{-1} - \text{Avar}\left(\bar{\beta}_{FE}\right)^{-1} \text{ is positive semi-definite} \\ \implies &\text{Avar}\left(\hat{\beta}_{FE}\right) - \text{Avar}\left(\hat{\beta}_{POLS}\right) \text{ is positive semi-definite,} \end{aligned}$$

as required. □

The point of this example is that it is unadvisable to just use the fixed effects estimator in all cases in which no time-invariant regressors are present. After all, if Wooldridge's claim were true and assumption A.FE.2 holds (no time-invariant regressors), the fixed effect estimator would always be preferable over the pooled OLS estimator because it can never be ruled out that no unobserved fixed effect exists. This point is especially important if T is small.

3.3.3 Simulation

In this paragraph we consider a simulation of a model in which the use of the fixed effects estimator leads us to make a type II error (it fails to reject a false null hypothesis) while the pooled OLS leads to correct inference. We consider a slightly adjusted model than the ones used in previous simulations:

$$\underset{T\times 1}{y_i} = \underset{T\times 1}{X_i}\ \underset{\mathbb{R}}{\beta} + \underset{T\times 1}{\varepsilon_i}\ . \tag{3.51}$$

where the true value of $\beta = 2$. The intercept α has been removed from the model because it is time invariant, i.e. inclusion of an intercept would be a violation of the rank condition for the fixed effects estimator. In this simulation, we take a relatively small sample of $N = 50$ and $T = 3$. The key assumption in this model is:

- **Assumption SS.FE.1:** $\varepsilon_i|\mathbf{X}_i \overset{iid}{\sim} N\left(0, \sigma_\varepsilon^2 \mathbf{I}_T\right)$, where the true value of $\sigma_\varepsilon^2 = 25$.

Case 1: Inefficiency

Under Assumption SS.FE.1, pooled OLS is the Best Linear Unbiased Estimator and asymptotically more efficient than the fixed effects estimator. The results of the pooled OLS estimator applied on this model are presented in table 3.6. It is for now important to notice that, in this simulation, the Pooled OLS estimator only produces type II errors in 2.2% of the cases.

The results for the fixed effects estimator are given in table 3.7. The fixed effects estimator has clearly a much higher type I error compared to the Pooled OLS estimator, suggesting that this estimator is much less accurate that the pooled OLS estimator. Serial correlation robust standard errors have been used, suggesting that they apparently underestimate the actual standard errors. The point of this simulation is to show that,

especially in short panels (small T), the fixed effects estimator may easily be inefficient. As noted before, the fixed effect transformation induces serial correlation in the model under assumption SS.FE.1. Consequently, the fixed effects estimator is less efficient than the pooled OLS model. Given $T = 3$, the autocorrelation coefficient of the transformed model equals

$$\rho\left(\dot{\varepsilon}_{i,t}, \dot{\varepsilon}_{i,s}\right) = -\frac{1}{T-1} = -\frac{1}{2},\ s \neq t.$$

Table 3.6: Pooled OLS (R=1000 simulations)

	Estimate	Std. Error	t value	Pr(>|t|)	Type I	Type II
beta	1.9991	0.4627	4.4917	0.0048	5.3%	2.2%

Table 3.7: Fixed effects (R=1000 simulations)

	Estimate	Std. Error	t value	Pr(>|t|)	Type I	Type II
beta	2.0213	0.6899	3.2273	0.0498	7.3%	19.2%

Case 2: Presence of fixed effects

The pooled OLS estimator was argued to be inconsistent in presence of an unobserved individual specific fixed effect that caused assumption A.POLS.1 to fail. In this paragraph we consider a panel that has $N = 1000$ cross-sectional units and $T = 20$ time periods. This model will be formulated such that it includes an individual specific fixed effect. The model problem considered here is the standard fixed effects model, i.e.

$$y_i = \mathbf{X}_i\beta + \mu_i\vec{1}_T + v_i \tag{3.52}$$

The following assumptions are imposed.

- **Assumption LS.FE.1:** $v_i|\mathbf{X}_i, \mu_i \overset{iid}{\sim} N\left(0, \mathbf{I}_T\right)$
- **Assumption LS.FE.2:** $\mu_i \overset{ind}{\sim} N\left(\bar{\mu}_i, \sigma^2_{\mu_i}\right)$, where both $\sigma^2_{\mu_i}$, $\bar{\mu}_i \in [1, 10]$ are arbitrary integers.

The results for the pooled OLS estimator are presented in table 3.8. The excessive amount of type I errors produced by this estimator indeed suggests that it produces inconsistent estimates due to the presence of an individual specific fixed effect. The POLS estimator should clearly be rejected as a reliable estimator.

The outcomes for the fixed effects estimator are given in table 3.9. The dramatic reduction in type I errors produced by the fixed effects estimator obviously suggests that this estimator is to be preferred over the POLS estimator. At the other hand, the amount

of type I errors did not converge to 5%, again an indication that the robust standard errors underestimate the actual standard errors.

Table 3.8: Pooled OLS summary (R=1000 simulations)

	Estimate	Std. Error	t value	Pr(>|t|)	Type I	Type II
beta	2.3960	0.0256	110.2407	0	98.30%	0.00%

Table 3.9: Fixed effects summary (R=1000 simulations)

	Estimate	Std. Error	t value	Pr(>|t|)	Type I	Type II
beta	2.000	0.0028	712.919	0	6.00%	0.00%

3.3.4 Bayesian inference

The Fixed Effect transformation can also be justified from a Maximum Likelihood Assumption under the usual assumptions for panel data models.

- **Assumption B.FE.1:** $y_i|\mathbf{X}_i, \mu_i \overset{iid}{\sim} N\left(\mathbf{X}_i\beta + \mu_i\vec{1}_T, \sigma_v^2\mathbf{I}_T\right)$
- **Assumption B.FE.2:** $\mathbb{P}[\text{rank}\left(\sum_{i=1}^N \dot{\mathbf{X}}_i^T\dot{\mathbf{X}}_i\right) = K] = 1$

Under assumption B.FE.1 and assumption B.FE.2, the likelihood ratio for $y_{1:N}|\mathbf{X}_{1:N}$ is given by

$$\pi\left(y_{1:N}|\mathbf{X}_{i:N}\right) = \left(\frac{1}{2\pi\sigma_v^2}\right)^{NT/2} \times \exp\left(-\frac{1}{2\sigma_v^2}\sum_{i=1}^N\left(y_i - \mathbf{X}_i\beta - \vec{1}_T\mu_i\right)^T\left(y_i - \mathbf{X}_i\beta - \vec{1}_T\mu_i\right)\right)$$

A non-informative Jeffreys prior for $\theta_i = \left(\sigma_v^2, \beta, \mu_i\right)$ is given by $\pi\left(\theta\right) \propto \frac{1}{\sigma_v^2}$. In this case, the posterior is given by

$$\pi\left(\theta_{1:N}|y_{1:N}, \mathbf{X}_{1:N}\right) \propto \left(\frac{1}{\sigma_v^2}\right)^{NT/2+1} \times \exp\left(-\frac{1}{2\sigma_v^2}\sum_{i=1}^N\left(y_i - \mathbf{X}_i\beta - \vec{1}_T\mu_i\right)^T\left(y_i - \mathbf{X}_i\beta - \vec{1}_T\mu_i\right)\right) \tag{3.53}$$

Theorem 3.3.2. *The MAP-estimator for the Bayesian Fixed Effect model with Jeffreys prior is given by the frequentist fixed effect estimator, i.e.*

$$\hat{\beta} = \left(\sum_{i=1}^{N} \dot{\mathbf{X}}_i^T \dot{\mathbf{X}}_i \right)^{-1} \left(\sum_{i=1}^{N} \dot{\mathbf{X}}_i^T \dot{y}_i \right) \tag{3.54}$$

Proof. The log-posterior is given by

$$\begin{aligned} \log\left(\pi\left(\theta_{1:N}|y_{1:N}, X_{1:N}\right)\right) \propto & -\left(\frac{NT}{2}+1\right)\log\left(\sigma_v^2\right) \\ & -\frac{1}{2\sigma_v^2}\sum_{i=1}^{N}\left(y_i - \mathbf{X}_i\beta - \vec{1}_T\mu_i\right)^T\left(y_i - \mathbf{X}_i\beta - \vec{1}_T\mu_i\right) \end{aligned} \tag{3.55}$$

The MAP-estimator for μ_i is derived by taking the derivative of (3.55) with respect to μ_i, setting this 0 and solve for μ_i.

$$\frac{\partial \log(\pi)}{\partial \mu_i} \propto \frac{1}{\sigma_v^2}\sum_{i=1}^{N} y_i^T \vec{1}_T + 2\beta^T \mathbf{X}_i^T \vec{1}_T + 2\mu_i T = 0 \tag{3.56}$$

Rewriting this results in

$$\hat{\mu_i} = \bar{y}_i - \bar{X}_i\beta \tag{3.57}$$

Substituting estimator 3.57 in equation (3.55) results in

$$\begin{aligned} \log\left(\pi\left(\theta_{1:N}|y_{1:N}, X_{1:N}\right)\right) \propto & -\left(\frac{NT}{2}+1\right)\log\left(\sigma_v^2\right) \\ & -\frac{1}{2\sigma_v^2}\sum_{i=1}^{N}\left(\dot{y}_i - \dot{x}_i\beta\right)^T\left(\dot{y}_i - \dot{x}_i\beta\right) \end{aligned} \tag{3.58}$$

Where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{i,t}$ and $\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{i,t}$.

This is again a transformed pooled Bayesian regression model with a Jeffreys prior, which MAP estimation is known and given by

$$\hat{\beta}_{BFE} = \left(\sum_{i=1}^{N} \dot{\mathbf{X}}_i^T \dot{\mathbf{X}}_i \right)^{-1} \left(\sum_{i=1}^{N} \dot{\mathbf{X}}_i \dot{y}_i \right) \tag{3.59}$$

as required. □

3.4 First Difference Estimator

Another method to eliminate inconsistencies caused by individual specific fixed effects is to take the first differences of the model problems and apply Pooled OLS on that transformation. This results in the first difference estimator. This estimator has received little attention in the panel data literature compared to the fixed effects estimator. However, the first difference estimator has very desirable properties in presence of heavy serial correlation in the error term. That is, it may even be preferred over pooled OLS in absence of unobserved individual specific fixed effects.

Derivation of the Estimator

The First Difference estimator is derived after taking the first difference transformation of the model problem formulated in equation section 3.1. That is, the first differenced model becomes

$$\Delta y_{i,t} = \Delta x_{i,t}\beta + \Delta\varepsilon_{i,t},\ t = 2, \ldots, T \tag{3.60}$$

where $\Delta y_{i,t} = y_{i,t} - yi, t-1$ and $\Delta x_{i,t},\ \Delta\varepsilon_{i,t}$ are defined analogously. As noted in the introduction, this transformation removes the unobserved individual specific fixed effect μ_i. Indeed,

$$\Delta\varepsilon_{i,t} = \mu_i + v_{i,t} - \mu_i - v_{i,t-1} = v_{i,t} - v_{i,t-1}$$

and so the individual specific fixed effect is removed.

In vector form, the first differenced transformed model will be written as

$$\underset{T-1\times 1}{\Delta y_i} = \underset{T-1\times K}{\Delta \mathbf{X}_i}\ \underset{K\times 1}{\beta} + \underset{T-1\times 1}{\Delta\varepsilon_i}\ . \tag{3.61}$$

The **first difference estimator** is the pooled OLS estimator applied on the transformed model in equation (3.61), i.e.

$$\hat{\beta}_{FD} = \left(\sum_{i=1}^{N} \Delta\mathbf{X}_i\Delta\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{N} \Delta\mathbf{X}_i\Delta y_i\right) \tag{3.62}$$

In matrix form, define the following transformation matrix

$$\underset{T-1\times T}{\mathbf{Q}} = \begin{bmatrix} -1 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & & 0 \\ 0 & \cdots & \cdots & \cdots & -1 & 1 \end{bmatrix} \tag{3.63}$$

Then the set of equations (3.61) is obtained by premultiplying the model problem with $\mathbf{Q}$.

3.4.1 Serial correlation

As noted in the introduction of this section, the FD estimator has some desirable properties in presence of heavy serial correlation. A distinction will be made between the *differenced errors* and the *original errors*. The former refers to $\Delta\varepsilon_i$, while the latter refers to ε_i. To see why the FD estimator has special properties, consider the covariance of $\Delta\varepsilon_{i,t}$ and $\Delta\varepsilon_{i,t-1}$. That is,

$$\begin{aligned}\operatorname{cov}\left(\Delta\varepsilon_{i,t}, \Delta\varepsilon_{i,t-1}\right) &= \mathbb{E}[\Delta\varepsilon_{i,t}\Delta\varepsilon_{i,t-1}] \\ &= \operatorname{cov}[\varepsilon_{i,t}, \varepsilon_{i,t-1}] - \operatorname{cov}[\varepsilon_{i,t}, \varepsilon_{i,t-2}] - \operatorname{Var}[\varepsilon_{i,t-1}] + \operatorname{cov}[\varepsilon_{i,t-1}, \varepsilon_{i,t-2}].\end{aligned}$$

This term is already an indication that the $\rho\left(\Delta\varepsilon_{i,t}, \Delta\varepsilon_{i,t-1}\right) \leq 0$ if the original errors are positively serially correlated. Continuing the analysis, extra assumptions are made on the serial correlation of the original error term.

Original errors follow a random walk

Let $\varepsilon_{i,t}$ follow a random walk, i.e. $\varepsilon_{i,t} = \varepsilon_{i,t-1} + \eta_{i,t}$ where $\mathbb{E}[\eta_{i,t}|\mathbf{X}_i] = 0$. Then

$$\begin{aligned}\operatorname{cov}[\varepsilon_{i,t}, \varepsilon_{i,t-1}] &= \operatorname{cov}[\varepsilon_{i,t-1} + \eta_{i,t}, \varepsilon_{i,t}] \\ &= \operatorname{cov}[\varepsilon_{i,t-1}\varepsilon_{i,t-1}] + \cancel{\mathbb{E}[\eta_{i,t}\varepsilon_{i,t-1}]} \\ &= \operatorname{Var}[\varepsilon_{i,t-1}\varepsilon_{i,t-1}]\end{aligned}$$

and in a similar fashion

$$\operatorname{cov}[\varepsilon_{i,t}, \varepsilon_{i,t-2}] = \operatorname{cov}[\varepsilon_{i,t-1}, \varepsilon_{i,t-2}]$$

resulting in

$$\operatorname{cov}\left(\Delta\varepsilon_{i,t}, \Delta\varepsilon_{i,t-1}\right) = 0 \tag{3.64}$$

$$\implies \rho\left(\Delta\varepsilon_{i,t}, \Delta\varepsilon_{i,t-1}\right) = 0 \tag{3.65}$$

So if the original errors follow a random walk, the differenced errors are serially uncorrelated.

Original error is serially uncorrelated

Assume now that the original error is serially uncorrelated, as is the case under Assumption A.FE.3. In that case, equation (3.64) reduces to

$$\begin{aligned}\operatorname{cov}[\Delta\varepsilon_{i,t}, \Delta\varepsilon_{i,t-1}] &= -\operatorname{Var}[\varepsilon_{i,t-1}] \\ &= -\sigma^2_\Delta \\ \rho\left(\Delta\varepsilon_{i,t}, \Delta\varepsilon_{i,t-1}\right) &= -\frac{\sigma^2_\Delta}{2\sigma^2_\Delta} \\ &= -1/2\end{aligned}$$

ergo, if the original errors are serially uncorrelated, then the first order lag of the differenced errors have an autocorrelation equalling $-1/2$.

3.4.2 Finite sample properties

For the finite sample properties to hold, the following assumptions are imposed.

- **Assumption F.FD.1:** $\mathbb{E}[v_i|\{\mathbf{X}_j, \mu_j\}_{j=1}^N] = 0$
- **Assumption F.FD.2:** $\mathbb{P}[\text{rank}\left(\Delta\mathbf{X}_i \left(\Delta\mathbf{X}_i\right)^T\right) = K] = 1$
- **Assumption F.FD.3:** $\mathbb{E}[\Delta\varepsilon_i \left(\Delta\varepsilon_i\right)^T |\{\mathbf{X}_j, \mu_j\}_{j=1}^N] = \sigma_\Delta^2 \mathbf{I}_{T-1}$

Note that the FD estimator is indeed identified due to assumptions F.FD.1 and F.FD.2.

Unbiased

The FD estimator is an unbiased estimator for β.

Corollary 3.4.1. *The FD estimator $\hat{\beta}_{FD}$ is, conditional on $\{\mathbf{X}_j, \mu_j\}_{j=1}^N$, an unbiased estimator for β.*

Proof. This is a shortened proof.

$$\begin{aligned}\mathbb{E}[\hat{\beta}_{FD}|\{\mathbf{X}_j, \mu_j\}_{j=1}^N] &= \beta + \left(\sum_{i=1}^N \left(\Delta\mathbf{X}_i\right)^T \Delta\mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^N \sum_{t=1}^T x_{i,t}^T \mathbb{E}[v_{i,t} - v_{i,t-1}|\{\mathbf{X}_j, \mu_j\}_{i=1}^N]\right) \\ &= \beta\end{aligned}$$

as required. □

The unconditional unbiasedness follows from the Law of Iterated Expectations.

Variance matrix

The variance-covariance matrix of the FD estimator, conditional on $\{\mathbf{X}_j, \mu_j\}_{j=1}^N$, equals

$$\text{Var}\left(\hat{\beta}_{FD}|\{\mathbf{X}_j, \mu_j\}_{j=1}^N\right) = \sigma_\Delta^2 \left(\sum_{i=1}^N \left(\Delta\mathbf{X}_i\right)^T \Delta\mathbf{X}_i\right)^{-1}, \tag{3.66}$$

implying that the unconditional variance-covariance matrix of $\hat{\beta}_{FD}$ equals

$$\text{Var}\left(\hat{\beta}_{FD}\right) = \sigma_\Delta^2 \mathbb{E}\left[\left(\sum_{i=1}^N \left(\Delta\mathbf{X}_i\right)^T \Delta\mathbf{X}_i\right)^{-1}\right]. \tag{3.67}$$

Finite sample efficiency

The FD estimator under assumptions F.FD.1-F.FD.3 satisfies the Gauss-Markov properties and hence the fact that the FD estimator is the Best Linear Unbiased Estimator should not come as a surprise. Rather than proving this result here, it may be useful to compare the pooled OLS estimator with the first difference estimator in the following theoretical model:

- **Assumption FM.FD.1**: Assumption F.POLS.1 holds;
- **Assumption FM.FD.2**: Assumption F.FD.2 holds;
- **Assumption FM.FD.3**: Assumption F.FD.3 holds;

Assumption FM.FD.1 implies that there are no unobserved individual specific fixed effects and so the Pooled OLS estimator is unbiased. Assumption FM.FD.2 suggests, among others, that there are no a.s. time-invariant regressors. Assumption FM.FD.4 implies that the pooled OLS does not satisfy the Gauss-Markov assumption as per the analysis made in paragraph 3.4.1. It is hence not guaranteed to satisfy to be the most efficient estimator. To prove the efficiency of the First Difference estimator, a lemma is needed. Recall that $\mathbf{A}^g$ is called the **generalised inverse** if

$$\mathbf{A}\mathbf{A}^g\mathbf{A} = \mathbf{A} \tag{3.68}$$

Lemma 3.4.1 (Own work)**.** *Assume assumptions FM.FD.1 - FM.FD.3 hold, then the variance-covariance of the original errors equals*

$$\mathbb{E}[\varepsilon_i\varepsilon_i^T|\{\mathbf{X}_i\}_{i=1}^N] = \sigma_\Delta^2\left(\mathbf{Q}^T\mathbf{Q}\right)^g \tag{3.69}$$

where $\left(\mathbf{Q}^T\mathbf{Q}\right)^g$ *denotes the generalised inverse of* $\left(\mathbf{Q}^T\mathbf{Q}\right)$.[4]

Proof. From assumption FM.FD.3, it follows

$$\begin{aligned}\sigma_\Delta^2\mathbf{I}_{T-1} &= \mathbb{E}[\Delta\varepsilon_i\Delta\varepsilon_i^T|\{\mathbf{X}_j\}_{j=1}^N] \\ &= \mathbf{Q}\mathbb{E}[\varepsilon_i\varepsilon_i^T|\{\mathbf{X}_j\}_{j=1}^N]\mathbf{Q}^T\end{aligned} \tag{3.70}$$

Premultiply equation (3.70) with $\mathbf{Q}^T$ and post multiply this equation with $\mathbf{Q}^T$. Then,

$$\left(\mathbf{Q}^T\mathbf{Q}\right)\mathbb{E}[\varepsilon_i\varepsilon_i^T|\{\mathbf{X}_j\}_{j=1}^N]\left(\mathbf{Q}^T\mathbf{Q}\right) = \sigma_\Delta^2\mathbf{Q}^T\mathbf{Q} \tag{3.71}$$

from the definition of pseudo-inverse,

$$\mathbb{E}[\varepsilon_i\varepsilon_i^T|\{\mathbf{X}_j\}_{j=1}^N] = \sigma_\Delta^2\left(\mathbf{Q}^T\mathbf{Q}\right)^g \tag{3.72}$$

as required. □

[4]It does not necessarily have to be the reflective generalised inverse.

Lemma 3.4.2 makes it possible to prove the following theorem. Let $\mathbf{\Sigma} = \sigma_\Delta^2 \left(\mathbf{Q}^T\mathbf{Q}\right)^g$.

Corollary 3.4.2 (Own work)**.** *Under assumptions FM.FD.1 - FM.FD.3, the Pooled OLS estimator is less efficient for finite samples than the First Difference estimator.*

Proof.

$$\underset{K\times K}{\text{Var}}\left(\hat{\beta}_{POLS}|\{\mathbf{X}_j\}_{j=1}^N\right) = \left(\sum_{i=1}^N \mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^N \mathbf{X}_i\mathbf{\Sigma}\mathbf{X}_i\right)\left(\sum_{i=1}^N \mathbf{X}_i^T\mathbf{X}_i\right)^{-1}, \tag{3.73}$$

whereas

$$\underset{K\times K}{\text{Var}}\left(\hat{\beta}_{FD}|\{\mathbf{X}_j\}_{j=1}^N\right) = \sigma_\Delta^2\left(\sum_{i=1}^N \mathbf{X}_i\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i\right)^{-1}. \tag{3.74}$$

To see that $\text{Var}\left(\hat{\beta}_{POLS}|\{\mathbf{X}_j\}_{j=1}^N\right) - \text{Var}\left(\hat{\beta}_{FD}|\{\mathbf{X}_j\}_{j=1}^N\right)$ is a positive semi-definite matrix, a slightly changed version of the Gauss-Markov tric is needed. To do so, first define

$$\mathbf{C}_i = \left(\sum_{j=1}^N \mathbf{X}_j^T\mathbf{X}_j\right)\mathbf{X}_i.$$

Hence, the variance-covariance matrix of the Pooled OLS estimator can be rewritten as

$$\text{Var}\left(\hat{\beta}_{POLS}|\{\mathbf{X}_j\}_{j=1}^N\right) = \sum_{i=1}^N \mathbf{C}_i\mathbf{\Sigma}\mathbf{C}_i^T \tag{3.75}$$

This expression does not have a Gauss-Markov structure. We consequently take a slightly adjusted difference matrix,

$$\mathbf{D}_i = \mathbf{C}_i - \left(\sum_{i=1}^N \mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q} \tag{3.76}$$

This expression has the usual properties, i.e.

1. $\sum_{i=1}^N \mathbf{D}_i\mathbf{X}_i = 0$
2. $\sum_{i=1}^N \mathbf{D}_i y_i = \hat{\beta}_{POLS} - \hat{\beta}_{FD}$

Rewrite $\mathbf{C}_i$ as

$$\mathbf{C}_i = \mathbf{D}_i + \left(\sum_{i=1}^N \mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q} \tag{3.77}$$

and so expression (3.75) can be rewritten as

$$
\begin{aligned}
&\sum_{i=1}^{N}\left(\mathbf{D}_i+\left(\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\right)\left(\mathbf{Q}^T\mathbf{Q}\right)^g \\
&\qquad\qquad\times\left(\mathbf{D}_i+\left(\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\right)^T \\
=&\sum_{i=1}^{N}\mathbf{D}_i\boldsymbol{\Sigma}\mathbf{D}_i^T+\sigma_\Delta^2\left(\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i\right)^{-1} \\
&\qquad\qquad\times\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\left(\mathbf{Q}^T\mathbf{Q}\right)^g\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i \\
&\qquad\qquad\times\left(\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i\right)^{-1} \\
\underset{3.4.2}{\overset{lemma}{=}}&\sum_{i=1}^{N}\mathbf{D}_i\boldsymbol{\Sigma}\mathbf{D}_i^T+\sigma_\Delta^2\left(\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i\left(\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i\right)^{-1} \\
=&\sum_{i=1}^{N}\mathbf{D}_i\boldsymbol{\Sigma}\mathbf{D}_i^T+\sigma_\Delta^2\left(\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{Q}^T\mathbf{Q}\mathbf{X}_i\right)^{-1} \\
=&\sum_{i=1}^{N}\mathbf{D}_i\boldsymbol{\Sigma}\mathbf{D}_i^T+\mathrm{Var}\left(\hat{\beta}_{FD}|\{\mathbf{X}_i\}_{i=1}^N\right)
\end{aligned}
$$

In particular, this means that

$$
\mathrm{Var}\left(\hat{\beta}_{POLS}|\{\mathbf{X}_j\}_{j=1}^N\right)-\mathrm{Var}\left(\hat{\beta}_{FD}|\{\mathbf{X}_j\}_{j=1}^N\right) \tag{3.78}
$$

is a positive semi-definite matrix and so the FD estimator is indeed more efficient in finite samples than the Pooled OLS estimator under assumptions M.FD.1 - M.FD.3. As required. □

The unconditional results are preserved by the law of iterated expectations.

3.4.3 Asymptotic properties

The assumptions required to derive the appropriate asymptotic results are weaker than the ones needed to derive the finite sample results. That is, F.FD.1 - F.FD.3 imply the following assumptions.

- **Assumption A.FD.1:** $\mathbb{E}[v_{i,t}|\mathbf{X}_i, \mu_i] = 0,\ t = 1, \ldots, T$
- **Assumption A.FD.2:** $\text{rank}\left(\mathbb{E}[\sum_{t=2}^{T} \Delta x_{i,t}^T \Delta x_{i,t}]\right) = K$
- **Assumption A.FD.3:** $\mathbb{E}[\Delta\varepsilon_i \Delta\varepsilon_i^T|\mathbf{X}_i, \mu_i] = \sigma_\Delta^2 \mathbf{I}_{T-1}$ with $\Delta\varepsilon_i \in \mathbb{R}^{T-1}$.

Consistency

The first difference estimator consistently estimates β under assumptions A.FD.1 and A.FD.2.

Corollary 3.4.3. *If A.FD.1 and A.FD.2 hold, then the first difference estimator consistently estimates β, i.e.*

$$\underset{N\to\infty}{\text{as-lim}}\, \hat{\beta}_{FD} = \beta \tag{3.79}$$

Proof. One way to prove this is to show that assumptions A.FD.1 imply A.POLS.1. We take that for granted.

$$\begin{aligned}\underset{N\to\infty}{\text{as-lim}}\, \hat{\beta}_{FD} &= \beta + \underset{N\to\infty}{\text{as-lim}} \left(\sum_{i=1}^{N} (\Delta\mathbf{X}_i)^T \Delta\mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} (\Delta\mathbf{X}_i)^T \Delta\varepsilon_i\right) \\ &= \beta + \mathbb{E}[(\Delta\mathbf{X}_i)^T \Delta\mathbf{X}_i]^{-1} \cancel{\mathbb{E}[(\Delta\mathbf{X}_i)^T \Delta\varepsilon_i]}\end{aligned}$$

proving the required. □

Asymptotic normality and asymtotic variance

Just as in the pooled OLS case, the first difference estimator is asymptotically normally distributed under A.FD.1 and A.FD.2, i.e.

$$\underset{N\to\infty}{\text{d-lim}}\, \sqrt{N}\left(\hat{\beta}_{FD} - \beta\right) = N\left(0, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right) \tag{3.80}$$

where $\mathbf{A} = \mathbb{E}[(\Delta\mathbf{X}_i)^T \Delta\mathbf{X}_i]$ and $\mathbf{B} = \text{Var}\left((\Delta\mathbf{X}_i)^T \Delta\varepsilon_i\right)$ implying that

$$\text{Avar}\left(\hat{\beta}_{FD}\right) = \frac{1}{N}\left(\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right). \tag{3.81}$$

Under Assumption A.FD.3

$$\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} = \sigma_\Delta^2 \mathbf{A}^{-1}, \tag{3.82}$$

suggesting that

$$\text{Avar}\left(\hat{\beta}_{FD}\right) = \frac{1}{N}\left(\sigma_\Delta^2 \mathbf{A}^{-1}\right) \tag{3.83}$$

3.4.4 Simulation

We compare the Pooled OLS estimator with the first difference estimator in a panel of $N = 50$ cross sectional units and $T = 20$ time units. We assume that

- **Assumption SS.FD.1:** Assumption F.POLS.1 holds and $\{\mu_i\}_{i=1}^N$ is independent in mean of $\{\varepsilon_i|\mathbf{X}_i\}_{i=1}^N$
- **Assumption SS.FD.2:** Assumption F.FD.2 holds
- **Assumption SS.FD.3:** Assumption F.FD.3 holds

The model we consider is again

$$y_i = X_i\beta + \varepsilon_i \tag{3.84}$$

where the true value of β equals 2. Furthermore, $\varepsilon_{i,t} = \rho\varepsilon_{i,t-1} + \eta_{i,t}$, where $\eta_i \overset{iid}{\sim} N\left(0, \sigma_\eta^2\mathbf{I}_T\right)$. The true value of $\rho = 0.99$ and $\sigma_\eta^2 = 100$.

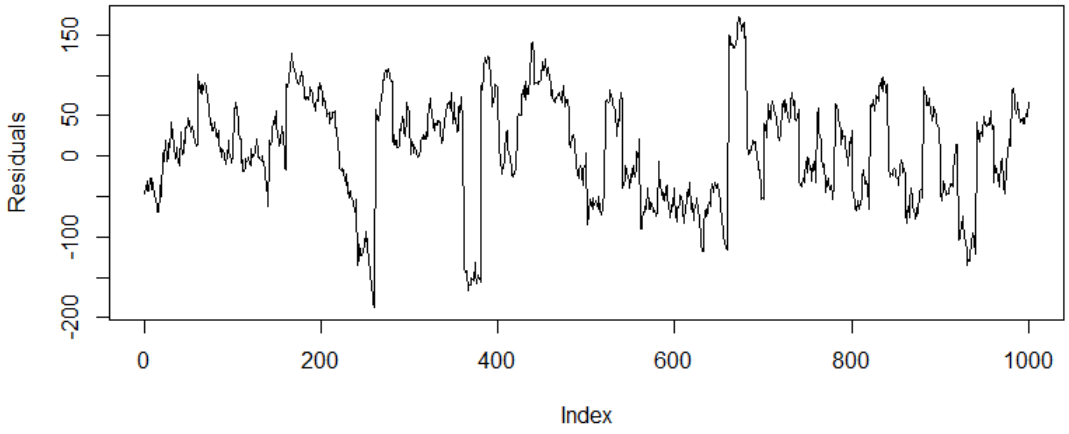
Pooled OLS estimates

The pooled OLS estimator is supposed to be inefficient because it does not satisfy the Gauss-Markov properties. The pooled OLS estimates are given in table 3.10. Obviously, the pooled OLS estimator preserves the property of being an unbiased estimator, so the average of the estimates should not come as a surprise. However, in over 50% of the cases we did make a type II error, underlining the poor performance of the pooled OLS estimator in case of heavy serial correlation. From figure 3.3 we can clearly observe that the pooled OLS residuals are strongly serially correlated.

Table 3.10: Pooled OLS estimates (R = 3000 simulations)

	Estimate	Std. Error	t value	Pr(>|t|)	Type I	Type II
1	2.0034	1.0603	1.9622	0.1698	5.83%	51.4%

Figure 3.2: POLS Residuals



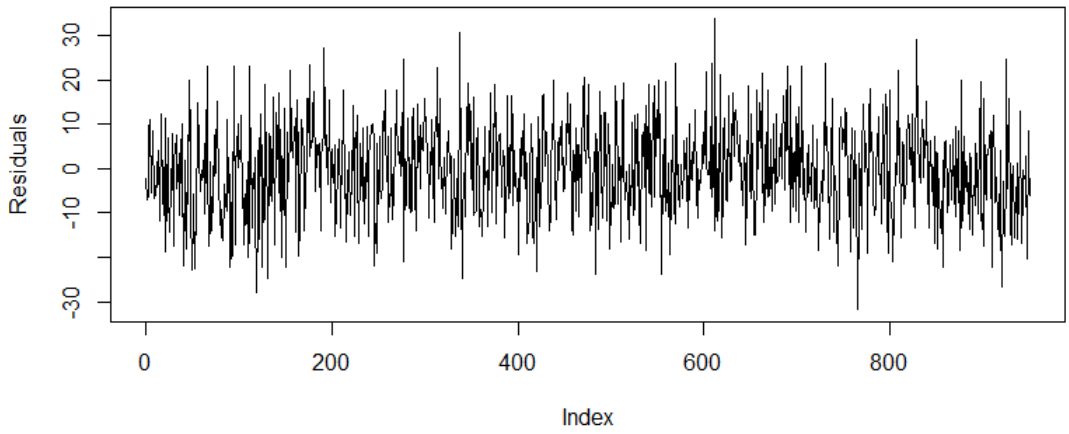
First difference estimates

Under assumptions SS.FD.1 - SS.FD.3, the first difference estimator is the Best Linear Unbiased Estimator. The results of the simulations are given in table 3.11. The first difference estimator clearly outperforms the pooled OLS estimator under this model formulation. No type II errors were observed for the the first difference estimator, while the percentage of type I errors made by the first difference estimator remains constant. From figure 3.3 we can also observe that the first difference residuals seem to follow a white noise process, suggesting that serial correlation has been removed.

Table 3.11: First difference (R=3000 simulations)

	Estimate	Std. Error	t value	Pr(>|t|)	Type I	Type II
1	1.9964	0.088	23.0416	0	5.5%	0%

Figure 3.3: FD Residuals



3.4.5 Bayesian Inference

The variance-covariance matrix $\mathbf{\Sigma}$ is only positive semi-definite, so a multivariate normal distribution will not have a density w.r.t. a Lebesgue measure. On the contrary, we can obtain a Bayesian approach to first differencing under the following assumptions.

- **Assumption B.FD.1:** $\Delta\varepsilon_i|\mathbf{X}_i \overset{iid}{\sim} N\left(0, \sigma_\Delta^2 \mathbf{I}_{T-1}\right)$
- **Assumption B.FD.2:** $\mathbb{P}[\text{rank}\left(\sum_{i=1}^N \mathbf{X}_i^T \mathbf{Q}^T \mathbf{Q} \mathbf{X}_i\right) = K] = 1$

Under assumptions B.FD.1 and B.FD.2,

$$y_{i:N}|\mathbf{X}_{i:N}, \beta, \sigma_v^2 \overset{iid}{\sim} N\left(0, \sigma_\Delta^2 \mathbf{I}_{T-1}\right). \tag{3.85}$$

The maximum a posteriori estimator under Jeffreys' prior is equal to the first difference estimator. This point will not be shown again.

3.5 Summary

This paragraph discussed several estimators that could be obtained with a non-parametric transformation. First, the pooled OLS estimator was derived as a (generalised) method of moments estimator. It appeared that most finite sample properties relied on stronger assumptions than comparable asymptotic properties. In fact, asymptotic efficiency properties could be derived under relatively weak assumptions.

We also saw that the conventional standard errors of the OLS estimator are often invalid due to serial correlation or heteroscedasticity. To circumvent this problem, the notion of robust standard errors was introduced. These errors remain valid in presence of

serial correlation and heteroscedasticity. They are usually bigger than their conventional counterparts, but there are some cases in which they are smaller (i.e. the normal OLS standard errors are biased upwards).

The second estimator discussed in this chapter was obtained by a fixed effects transformation. The resulting fixed effects estimator consistently estimates the parameter β in presence of an endogeneous unobserved individual specific fixed effect. On the other hand, if such unobserved individual specific effect is absent, this estimator may easily be inefficient compared to pooled OLS. It was argued that this is especially a problem when T is small since the fixed effects transformation induces negative serial correlation in the model. This serial correlation increases monotonically to zero as $T \to \infty$.

This chapter was concluded with the first difference estimator, which is obtained by subtracting the first differences from the model. This estimator proved to be efficient for finite samples in case the original error term follows a random walk. Also for weaker types of serial dependence, this estimator may prove far superior over the conventional pooled OLS estimator in terms of accuracy.

Chapter 4

Parametric Transformations

In the previous chapters it was shown that the assumptions of homoscedasticity and serial uncorrelatedness of the error term were not actually needed to derive consistent and unbiased estimates for β. However, these estimators are not guaranteed to be efficient in presence of serial correlation or heteroscedasticity. This chapter starts by reviewing the generalised least squares (GLS) estimator proposed by Aitken (1934). This estimator has both finite sample and asymptotic efficiency properties. GLS is in pratice infeasible, because it requires knowing the variance-covariance matrix of the error term up to a constant. Therefore, the analysis turns to parametric transformations of the pooled OLS estimator. That is, the transformation induces extra variance parameters. Those methods are commonly referred to as feasible GLS because of their close relation with the actual GLS estimator. After deriving some general results of feasible GLS estimation, the analysis considers how this estimation technique can be implemented to take serial correlation into account. Furthermore, the conventional random effects estimator will be derived as a feasible GLS estimator.

The model problem in this chapter is again the standard Panel Data regression model given by equation (4.1)

$$y_i = \mathbf{X}_i\beta + \varepsilon_i \tag{4.1}$$

4.1 Generalised Least Squares

As mentioned in the introduction in this chapter, GLS is efficient even if in presence of heteroscedasticity and serial correlation in the error term. However, the basic GLS method requires that the variance-covariance matrix of the error term is known up to a constant. This is more often than not the case. Hence, the GLS estimator discussed in this chapter is often considered infeasible. The GLS estimator is included in this thesis because the feasible GLS estimator strongly relies on it.

4.1.1 Derivation of the GLS estimator

The GLS estimator is justified by a transformation of the model problem described in (4.1). That is, both sides are multiplied by $\mathbf{\Sigma}^{-1/2}$, i.e.

$$\begin{aligned} \mathbf{\Sigma}^{-1/2} y_i &= \mathbf{\Sigma}^{-1/2} \mathbf{X}_i + \mathbf{\Sigma}^{-1/2} \varepsilon_i \\ y_i^* &= \mathbf{X}_i^* \qquad\quad + \varepsilon_i^* \end{aligned} \tag{4.2}$$

The **generalised least squares estimator** is obtained by applying pooled OLS on equation (4.2). That is,

$$\hat{\beta}_{GLS} = \left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{\Sigma}^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i \mathbf{\Sigma}^{-1} y_i \right) \tag{4.3}$$

4.1.2 Finite Sample properties

The finite sample properties of the GLS estimator can be derived under the following set of assumptions.

- Assumption F.GLS.1: $\mathbb{E}[\varepsilon_i | \{\mathbf{X}_j\}_{j=1}^N] = 0$
- Assumption F.GLS.2: $\mathbb{P}[\text{rank}\left(\sum_{i=1}^N \mathbf{X}_i^T \mathbf{\Sigma}^{-1/2} \mathbf{X}_i^T\right) = K] = 1$, where $\mathbf{\Sigma} = \mathbb{E}[\varepsilon_i \varepsilon_i^T]$ is assumed to be *known*.
- Assumption F.GLS.3: $\mathbb{E}[\varepsilon_i \varepsilon_i^T | \{\mathbf{X}_j\}_{j=1}^N] = \mathbb{E}[\varepsilon_i \varepsilon_i^T]$

Unbiased

The unbiasedness of the GLS estimator may sound trivial, but in fact it is not. The only reason why the GLS estimator can be proven to be unbiased without more assumptions is due to the fact that the variance-coviarance matrix $\mathbf{\Sigma}$ is assumed to be known.

Corollary 4.1.1. *Assume assumptions F.GLS.1 and F.GLS.2 hold. Then the GLS estimator is unbiased conditionally on* $\{\mathbf{X}_j\}_{j=1}^N$.

Proof.

$$\begin{aligned} &\mathbb{E}\left[\left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{\Sigma}^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i \mathbf{\Sigma}^{-1} y_i \right) \middle| \{\mathbf{X}_i\}_{i=1}^N \right] \\ =& \beta + \left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{\Sigma}^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{\Sigma}^{-1} \mathbb{E}[\varepsilon_i | \{\mathbf{X}_j\}_{j=1}^N] \right) \\ \overset{F.GLS.1}{=}& \beta. \end{aligned}$$

□

The unconditional unbiasedness follows from the law of iterated expectations.

Variance

The variance of the GLS estimator is important, because its expression will make the Gauss-Markov property of the GLS estimator obvious.

$$\begin{aligned}
\operatorname{Var}\left(\hat{\beta}_{GLS}|\{\mathbf{X}_i\}_{i=1}^N\right) &= \operatorname{Var}\left(\beta + \left(\sum_{i=1}^N \mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^N \mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\varepsilon_i\right)|\{\mathbf{X}_i\}_{i=1}^N\right) \\
&= \mathbb{E}\left[\left(\sum_{i=1}^N \mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^N \mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\varepsilon_i\right)\left(\sum_{i=1}^N \varepsilon_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right)\right. \\
&\qquad \left.\times\left(\sum_{i=1}^N \mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right)\Big|\{\mathbf{X}_i\}_{i=1}^N\right] \\
&= \left(\sum_{i=1}^N \mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^N \mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\mathbb{E}[\varepsilon_i\varepsilon_i^T|\{\mathbf{X}_i\}_{i=1}^N]\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right) \\
&\qquad \times\left(\sum_{i=1}^N \mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right) \\
&\overset{F.GLS.3}{=} \left(\sum_{i=1}^N \mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right)^{-1}
\end{aligned}$$

The unconditional variance follows from the law of total variance, i.e.

$$\operatorname{Var}\left(\hat{\beta}\right) = \mathbb{E}\left[\left(\sum_{i=1}^N \mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right)^{-1}\right] \tag{4.4}$$

Gauss-Markov theorem

As noted in the introduction of this chapter, the conditional Gauss-Markov theorem also holds under assumptions F.GLS.1 - F.GLS.3.

Theorem 4.1.1 (Conditional Gauss-Markov)**.** *Assume F.GLS.1 and F.GLS.3 hold. Then, if $\bar{\beta}$ is another linear unbiased estimator,*

$$\operatorname{Var}\left(\bar{\beta}\right) - \operatorname{Var}\left(\hat{\beta}_{GLS}\right)$$

is a semi-positive definite matrix. That is to say, the GLS estimator is BLUE.

Proof. The proof may be obvious from the expression of the finite sample variance-covariance matrix of the GLS-estimator as given in equation (4.4). It is, however, easier to note that the GLS transformed system implies assumption F.POLS.3. Recall F.POLS.3 states that

$$\mathbb{E}[\varepsilon_i \varepsilon_i^T | \{\mathbf{X}_j\}_{j=1}^N] = \sigma_\varepsilon^2 \mathbf{I}_T. \tag{4.5}$$

For the transformed system,

$$\begin{aligned}\mathbb{E}[\varepsilon_i^* \varepsilon_i^{*T} | \{\mathbf{X}_j\}_{j=1}^N] &= \mathbf{\Sigma}^{-1/2} \mathbb{E}[\varepsilon_i \varepsilon_i^T | \{\mathbf{X}_j\}_{j=1}^N] \mathbf{\Sigma}^{-1/2} \\ &= \mathbf{\Sigma}^{-1/2} \mathbf{\Sigma} \mathbf{\Sigma}^{-1/2} \\ &= \mathbf{I}_T\end{aligned}$$

implying the transformed system given by (4.2) satisfies the Gauss-Markov assumptions and hence the GLS-estimator is BLUE. □

4.1.3 Asymptotic properties

To derive asymptotic properties, the following assumptions are imposed. These assumptions are sufficient for consistency, but they do not guarantee the estimator to be unbiased.

- **Assumption A.GLS.1:** $\mathbb{E}[\varepsilon_i | \mathbf{X}_i] = 0$
- **Assumption A.GLS.2:** $\operatorname{rank}\left(\mathbb{E}[\mathbf{X}_i^T \mathbf{\Sigma}^{-1} \mathbf{X}_i]\right) = K$, where $\mathbf{\Sigma} = \mathbb{E}[\varepsilon_i \varepsilon_i^T]$ denotes a *known* variance-covariance matrix.
- **Assumption A.GLS.3:** $\mathbb{E}[\varepsilon_i \varepsilon_i^T | \mathbf{X}_i]$

Assumption A.GLS.1 is the strict exogeneity assumption. Imposing the equivalent of assumption A.POLS.1 is insufficient for consistency.[1]

Consistency

A.GLS.1 and A.GLS.2 do not imply that the GLS estimator is unbiased. However, they do imply that the GLS estimator is consistent.

Theorem 4.1.2. *Under A.GLS.1 and A.GLS.2, $\hat{\beta}_{GLS}$ consistently estimates β. That is,*

$$\operatorname*{as-lim}_{N \to \infty} \hat{\beta}_{GLS} = \beta \tag{4.6}$$

[1] For consistency, the weaker assumption $\mathbb{E}[\mathbf{X}_i \otimes \varepsilon_i] = 0$ is sufficient.

Proof. Note that

$$\begin{aligned}
\underset{N\to\infty}{\text{as-lim}}\,\hat{\beta}_{GLS} =& \beta + \underset{N\to\infty}{\text{as-lim}} \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\varepsilon_i\right)\\
=& \beta + \mathbb{E}[\mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i]^{-1}\mathbb{E}[\mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\varepsilon_i]\\
\overset{LIE}{=}& \beta + \mathbb{E}[\mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i]^{-1}\mathbb{E}[\mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\cancelto{0}{\mathbb{E}[\varepsilon_i|\mathbf{X}_i]}]\quad {}^{(GLS.1)}\\
=& \beta
\end{aligned}$$

□

Asymptotic normality and variance

The GLS estimator is asymptotic normally distributed.

Theorem 4.1.3. *Under assumption A.GLS.1 and A.GLS.2, the GLS estimator is asymptotically normally distributed. That is,*

$$\sqrt{N}\left(\hat{\beta}_{GLS} - \beta\right) \xrightarrow{d} N\left(0, \mathbf{V}\right) \tag{4.7}$$

where $\mathbf{V} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1},\ \mathbf{A} = \mathbb{E}[\mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i],\ \mathbf{B} = \text{Var}\left(\mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\varepsilon_i\right)$

Proof.

$$\begin{aligned}
\underset{N\to\infty}{\text{d-lim}}\,\sqrt{N}\left(\hat{\beta}_{GLS} - \beta\right) &= \underset{N\to\infty}{\text{d-lim}} \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right)\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{X}_i\boldsymbol{\Sigma}^{-1}\varepsilon_i\right)\\
&= \underset{N\to\infty}{\text{d-lim}}\,\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{A}^{-1}\mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\varepsilon_i\\
&= N\left(0, \mathbf{A}^{-1}\,\text{Var}\left(\mathbf{X}_i^T\boldsymbol{\Sigma}^{-1}\varepsilon_i\right)\mathbf{A}^{-1}\right).
\end{aligned}$$

□

Consequently,

$$\text{Avar}\left(\hat{\beta}_{GLS}\right) = \frac{1}{N}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} \tag{4.8}$$

The need for assumption GLS.3 may now become apparent, as expression (4.8) does not say anything about the asymptotic efficiency of the GLS estimator. However, with assumption GLS.3, (4.8) reduces to (by the Law of Iterated Expectations)

$$\text{Avar}\left(\hat{\beta}_{GLS}\right) = \frac{\mathbf{A}^{-1}}{N} \tag{4.9}$$

suggesting asymptotic efficiency.

Asymptotic efficiency

In this paragraph, both the asymptotic efficiency of the GLS estimator will be proven and a comparison will be made with the pooled OLS estimator. As in the Pooled OLS case, the GLS estimator is relatively efficient among all other consistent linear estimators that are asymptotically normally distributed.

Theorem 4.1.4. *Assume assumptions A.GLS.1 - A.GLS.3 hold. Assume* $\mathbf{Q}_T$ *is a positive definite transformation matrix pre-multiplied on the system* (4.2)*, after which Pooled OLS is applied on the transformed system, resulting in the estimator* $\bar{\beta}$*. Then,*

$$\text{Avar}\left(\bar{\beta}\right) - \text{Avar}\left(\hat{\beta}\right) \tag{4.10}$$

is positive semi-definite.

Proof. This proof is analogous to the proof given for the pooled OLS estimator. That is, the result follows from taking $\mathbf{A}_1 = \mathbf{X}_i^T \boldsymbol{\Sigma}^{-1} \varepsilon_i$ and $\mathbf{A}_2 = \mathbf{X}_i^T \mathbf{Q}^T \mathbf{Q} \boldsymbol{\Sigma}^{-1} \varepsilon_i$ in the proof of theorem 3.2.4.

□

It is easy to remark that the pooled OLS estimator is at best as efficient as the GLS estimator. To obtain the pooled OLS estimator, pick transformation matrix $\mathbf{Q} = \boldsymbol{\Sigma}^{1/2}$. It is as efficient as GLS provided $\boldsymbol{\Sigma} = \mathbf{I}_T \sigma_\varepsilon^2$ and less efficient otherwise. This result is in line with 3.2.4, since assumption A.POLS.3 was required to hold, which is the case iff $\boldsymbol{\Sigma} = \mathbf{I}_T \sigma_\varepsilon^2$.

4.1.4 Maximum Likelihood Estimation

The GLS estimator as derived in equation (4.3) is also the MLE estimator under the following conditions

- **Assumption MLE.GLS.1:** $\varepsilon_i \overset{iid}{\sim} N\left(0, \boldsymbol{\Sigma}\right)$
- **Assumption MLE.GLS.2:** $\mathbb{P}[\text{rank}\left(\mathbf{X}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{X}_i\right) = K] = 1$

and so

$$y_i | \mathbf{X}_i \overset{iid}{\sim} N\left(\mathbf{X}_i \beta, \Sigma\right) \tag{4.11}$$

for which the log likelihood function equals

$$\begin{aligned} \log\left(\pi\right) = &- \frac{NT}{2} \log\left(2\pi\right) + \frac{1}{2} \log\left(\det\left(\boldsymbol{\Sigma}\right)\right) \\ &- \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \mathbf{X}_i \beta\right)^T \boldsymbol{\Sigma}^{-1} \left(y_i - \mathbf{X}_i \beta\right). \end{aligned}$$

The estimator for β is obtained in the usual fashion (first order condition, assuming $\mathbf{\Sigma}$ is known)

$$\frac{\partial \log(\pi)}{\partial \beta} = \frac{1}{2}\sum_{i=1}^{N}(y_i - \mathbf{X}_i\beta)^T \mathbf{\Sigma}^{-1}\mathbf{X}_i = 0$$
$$\iff \beta = \left(\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{\Sigma}^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{\Sigma}^{-1}y_i\right)$$

4.1.5 Simulation

We consider a panel data model with $N = 50$ cross-sectional units and $T = 20$ time periods. The model is constructed such that the heavy heteroscedasticity and serial dependence is present in the error term. More specifically,

- **Assumption SS.GLS.1:** $\varepsilon_{i,t} = \rho\varepsilon_{i,t-1} + \eta_{i,t}$ where the true value of $\rho = 0.9$ and $\eta_i \overset{iid}{\sim} N\left(0, \sigma_i^2\mathbf{I}_T\right)$. The true value of $\sigma_t^2 = t^2$.

Pooled OLS

Under these assumptions, the pooled OLS estimator is not BLUE. This can obviously be seen from the rejection rates of the pooled OLS estimator given in table 4.1. The null hypothesis for the intercept was falsely not rejected in about 92% of the cases. At the other hand, no type II errors were observed for the parameter β. The plot of the residuals is given in figure 4.1. This plot clearly shows how the residuals are affected by serial dependence. Figure 4.2, however, better captures how the residuals are affected by heteroscedasticity.

Table 4.1: Pooled OLS (based on R=1000 simulations)

	Estimate	Std. Error	t value	Pr(>|t|)	Type I	Type II
(Intercept)	1.011	2.6138	0.3938	0.4665	6.8%	91.5%
beta	1.995	0.3249	6.3519	1e-04	5.3%	0%

Figure 4.1: Plot of residuals (showing serial dependence)

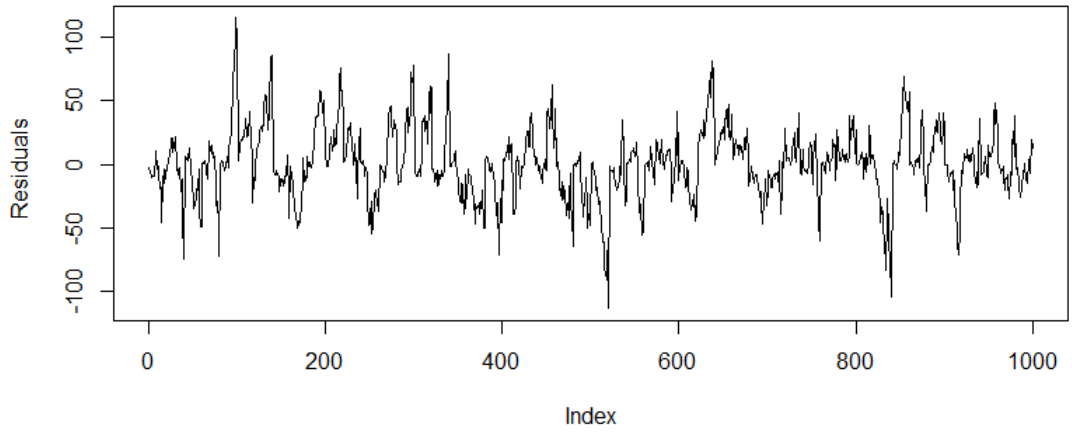
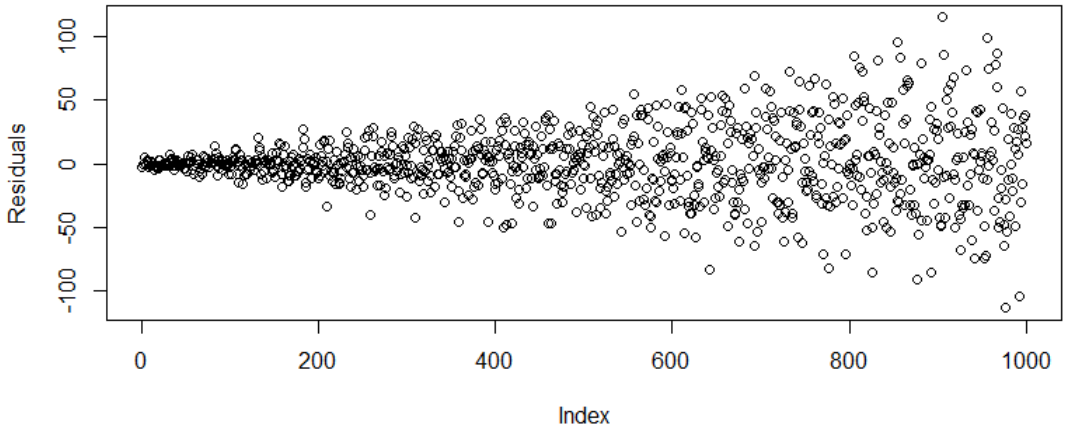


Figure 4.2: Plot of residuals (showing heteroscedasticity)



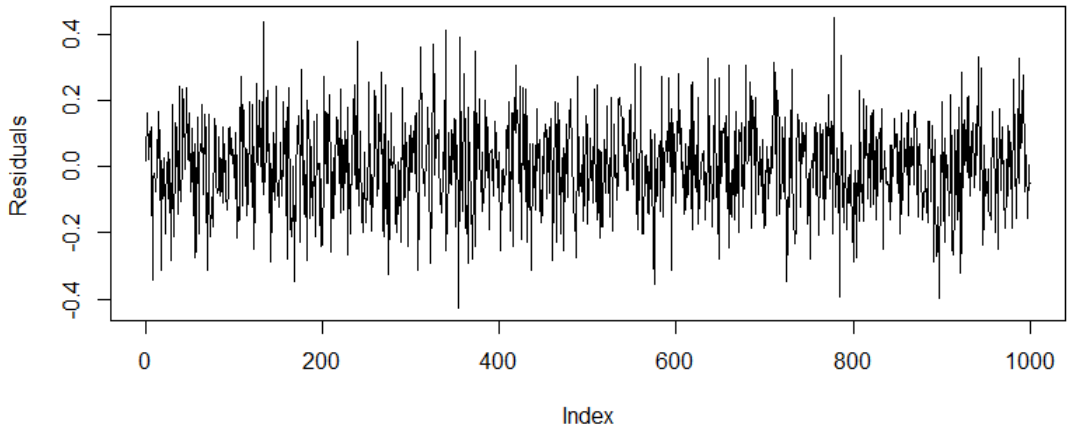
Generalised Least Squares

In this particular case, we have the luxury of knowing the variance-covariance matrix. As expected, GLS performs considerably better than the pooled OLS estimator. The null hypothesis was always rejected out of the 1000 simulations made. From figure 4.3 it can be observed that the residuals now follow some white noise process. They do in particular not seem to be affected by serial correlation anymore.

Table 4.2: GLS simulations (R=1000 simulations)

	Estimate	Std. Error	t value	Pr(>|t|)	Type I	Type II
(Intercept)	0.9999	0.0044	226.0202	0	7.2%	0%
beta	1.9985	0.0289	71.2367	0	4.6%	0%

Figure 4.3: Plot of GLS residuals



4.1.6 Application: Weighted Least Squares

An important application of the GLS method is Weighted Least Squares (WLS). Weighted Least Squares takes heteroscedasticity into account, but fails to repair the estimates for serial correlation. In particular, the variance matrix in a WLS-regression for the model problem (4.1) is given by

$$\mathbf{\Sigma} = \text{diag}\left(w_i^2\right) \tag{4.12}$$

where $w_i^2 = \frac{1}{\sigma_i^2}$ is the weight of the particular observation.

4.2 Feasible Generalised Least Squares

The GLS estimator is often infeasible because the variance-covariance matrix $\mathbf{\Sigma}$ is usually unknown. If $\mathbf{\Sigma}$ in the GLS analysis is replaced by a consistent estimator, denoted by $\hat{\mathbf{\Sigma}}$, then we obtain the **feasible GLS estimator** (FGLS). That is, the feasible GLS estimator is then given by

$$\hat{\beta}_{FGLS} = \left(\sum_{i=1}^{N} \mathbf{X}_i \hat{\mathbf{\Sigma}}^{-1} \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i \hat{\mathbf{\Sigma}}^{-1} \mathbf{X}_i\right) \tag{4.13}$$

Example 4.2.1 (Unrestricted feasible GLS)**.** To get a consistent estimator for $\mathbf{\Sigma}$, Wooldridge (2001) suggests fitting the model in two stages. First, pooled OLS is used

to fit the model from which estimator $\hat{\beta}_{POLS}$ is obtained. Under assumptions GLS.1 and GLS.2, this estimator consistently estimates β. Let now $\hat{\varepsilon}_i$ denote the pooled OLS residuals. The estimator

$$\hat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_i \hat{\varepsilon}_i^T \tag{4.14}$$

is consistent for $\mathbf{\Sigma}$. This may seem like a great choice, but in reality it is not. This unrestricted variance estimator requires the estimation of $O\left(T^2\right)$ variance parameters. This in particular requires N to be very large compared to T to obtain reliable estimates, which is often not the case.

The feasible GLS estimator at the other hand is extremely useful to model more complex serial dependence structures. Also, the classical random effects estimator is a special case of a restricted feasible GLS estimator. We will first discuss the general properties of the feasible GLS estimator, after which we propose some restricted estimators for the variance-covariance matrix $\mathbf{\Sigma}$.

4.2.1 Finite sample properties

Remarkably little is known about the finite sample properties of the FGLS estimator. Assume we have a consistent estimator for $\hat{\mathbf{\Sigma}}$. The usual finite sample assumptions are imposed, i.e.

- Assumption F.FGLS.1: $\mathbb{E}[\varepsilon_i | \{\mathbf{X}_j\}_{j=1}^N] = 0$
- Assumption F.FGLS.2: $\mathbb{P}[\text{rank}\left(\mathbf{X}_i^T \hat{\mathbf{\Sigma}} \mathbf{X}_i\right) = K] = 1$, where $\mathbb{E}[\varepsilon_i \varepsilon_i^T] = \mathbf{\Sigma}$
- Assumption F.FGLS.3: $\mathbb{E}[\varepsilon_i \varepsilon_i^T | \{\mathbf{X}_j\}_{j=1}^N] = \mathbb{E}[\varepsilon_i \varepsilon_i^T]$

Unbiased

In general, the feasible GLS estimator cannot be proven to be an unbiased estimator of β. Assumption F.FGLS.1 is insufficient because the estimator $\hat{\mathbf{\Sigma}}$ is a random variable itself. In some special cases, like $\hat{\mathbf{\Sigma}} = \mathbf{\Sigma}$, the feasible GLS estimator is unbiased under assumption F.FGLS.1. This also shows that the GLS is a special case of feasible GLS. So far, we did not impose any assumption on the distribution of the error term (other than moment conditions). Kakwani (1967) proves, under some mild assumptions on the distribution of the error term, that the feasible GLS is an unbiased estimator of β.

Theorem 4.2.1 (Kakwani, 1967)**.** *In addition to assumptions F.FGLS.1 and F.FGLS.2, assume that (i)* $\mathbb{E}[\hat{\beta}_{FGLS}] < \infty$ *and (ii) the distribution of* ε_i *follows a symmetric distribution. Then*

$$\mathbb{E}[\hat{\beta}_{FGLS}] = \beta \tag{4.15}$$

Note that these assumptions in particular hold when $\varepsilon_i|\mathbf{X}_i \sim N(0, \mathbf{\Sigma})$. We will not further eleborate on the finite sample properties of the feasible GLS estimator. There is no general Gauss-Markov extension of the feasible GLS estimator. Even under the normality of the error term, $\hat{\mathbf{\Sigma}}$ may show heavy non-linear behaviour. As a result, the FGLS estimator is not guaranteed to be efficient in finite samples. If the feasible GLS estimator reduces to the GLS estimator, then it does satisfy the Gauss-Markov theorem.

4.2.2 Asymptotic properties

As noted in the introduction, the feasible GLS estimator is mainly justified for its asymptotic properties. The assumptions imposed for the asymptotic results are exactely the same as A.GLS.1 - A.GLS.3. The feasible GLS estimator can be proven to be $\sqrt{N}$-equivalent to the GLS estimator. This means that it also satisfies the same asymptotic properties as the GLS estimator. In particular, it is asymptotically efficient provided A.GLS.3 also holds.

Consistency and efficiency

It is clear that the FGLS estimator is consistent, as the variance-covariance matrix is also consistent. A more interesting result, also the result that justifies the use of the FGLS estimator, is the fact that $\hat{\beta}_{FGLS}$ and $\hat{\beta}_{GLS}$ are $\sqrt{N}$ equivalent.

Theorem 4.2.2. *Assume assumptions A.FGLS.1 and A.FGLS.2 hold, then*

$$\underset{N\to\infty}{\text{as-lim}} \left(\hat{\beta}_{FGLS} - \hat{\beta}_{GLS}\right) = 0 \tag{4.16}$$

This statement is much stronger than just saying that both estimators are consistent estimators for β. As we have seen, a large class of estimators satisfies this consistency property but these estimators are in general not $\sqrt{N}$ equivalent with the GLS estimator.

Proof. (Obtained from Wooldridge, 2001, p. 159) We start by noting that

$$\sqrt{N}\left(\hat{\beta} - \beta\right) = \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i \hat{\mathbf{\Sigma}} \mathbf{X}_i\right)^{-1} \left(N^{-1/2} \sum_{i=1}^{N} \mathbf{X}_i \hat{\mathbf{\Sigma}} \varepsilon_i\right) \tag{4.17}$$

Then the second term is given by (where $\text{vec}(\mathbf{A})$ denotes the vectorisation of matrix $\mathbf{A}$):

$$N^{-1/2}\left(\sum_{i=1}^{N} \mathbf{X}_i^T \hat{\mathbf{\Sigma}} \varepsilon_i\right) - N^{-1/2}\left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{\Sigma} \varepsilon_i\right) = N^{-1/2}\left(\sum_{i=1}^{N} \varepsilon_i \otimes \mathbf{X}_i\right)^T \text{vec}\left(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right)$$

The central limit theorem implies (where $\otimes$ denotes the Kronecker product)

$$\underset{N\to\infty}{\text{as-lim}}\, N^{-1/2} \sum_{i=1}^{N} (\varepsilon_i \otimes \mathbf{X}_i)^T = \mathbf{Z} \tag{4.18}$$

where $\mathbf{Z}$ is some normally distributed random variable. This follows from assumption A.FGLS.1. At the other hand,

$$\underset{N\to\infty}{\text{as-lim}} \operatorname{vec}\left(\hat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\right) = 0 \tag{4.19}$$

due to the consistency of $\hat{\mathbf{\Sigma}}$. So we can say that

$$N^{-1/2}\sum_{i=1}^{N}\mathbf{X}_i^T\hat{\mathbf{\Sigma}}\varepsilon_i = N^{-1/2}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{\Sigma}^{-1}\varepsilon_i + d_N \tag{4.20}$$

where $d_N \xrightarrow{a.s.} 0$. Similary,

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{X}_i = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{\Sigma}^{-1}\mathbf{X}_i + k_n \tag{4.21}$$

where $k_N \xrightarrow{a.s.} 0$. We have now shown that

$$\sqrt{N}\left(\hat{\beta}_{FGLS} - \beta\right) = \sqrt{N}\beta_{GLS} + p_N \tag{4.22}$$

where $p_N \xrightarrow{a.s.} 0$. This in particular implies

$$\underset{N\to\infty}{\text{as-lim}}\sqrt{N}\left(\hat{\beta}_{FGLS} - \hat{\beta}_{GLS}\right) = 0 \tag{4.23}$$

as required. □

To conclude this paragraph,

Theorem 4.2.3. *The feasible GLS estimator is efficient relatively to any other transformed pooled OLS estimator.*

Proof. Direct consequence of the $\sqrt{N}$-equivalence with the $\hat{\beta}_{GLS}$ estimator. That is, $\sqrt{N}$ equivalence implies that these estimators have the same limiting distribution including the same asymptotic variance-covariance matrix. □

4.2.3 Autocorrelation in error term

The GLS and unrestricted FGLS explicitely take serial dependence into account, yet both methods are in practice never applied in their pure forms. However, as T grows larger, there is also a bigger need to model serial dependence in the error term. The first difference transformation provided a non-parametric method that explicitely takes serial correlation into account, but this method only satisfied some efficiency properties provided the error term follows a random walk. In this paragraph we will extend these results to the case where the error term follows an $ARMA\,(1,1)$ process.

Definition 4.2.1. $\varepsilon_{i,t}$ is said to follow an $ARMA\,(p,q)$ process when

$$\varepsilon_{i,t} = \sum_{s=1}^{p} \rho_s \varepsilon_{i,t-s} + \left(\sum_{j=1}^{q} \phi_j \eta_{i,t-j} \right) + \eta_{i,t} \tag{4.24}$$

where $\eta_{i,t}$ is some white noise process.

In that case, the autocovariance function (ACF) is given by (Shumway and Stoffer, 2016)

$$\operatorname{cov}\left(\varepsilon_{i,t}, \varepsilon_{i,t-s}\right) = \frac{(1+\theta\rho)\,(\rho+\theta)}{1-\rho^2}\rho^{s-1}. \tag{4.25}$$

We will not derive the transformation matrix $\hat{\mathbf{\Sigma}}^{-1/2}$. MaCurdy (1982) derive the results for general $ARMA\,(p,q)$ matrices.

First order autoregressive errors (ARMA(1,0))

Assume that $\varepsilon_{i,t}$ follows an $AR\,(1)$ proces, i.e.

$$\varepsilon_{i,t} = \rho\varepsilon_{i,t-1} + \eta_{i,t},\ |\rho < 1|. \tag{4.26}$$

- **Assumption AR1.FGLS.1:** $\mathbb{E}[\eta_i|\mathbf{X}_i] = 0$
- **Assumption AR1.FGLS.2:** $\operatorname{rank}\left([\mathbf{X}_i^T\mathbf{\Sigma}^{-1}\mathbf{X}_i]\right) = K$
- **Assumption AR1.FGLS.3:** $\mathbb{E}[\eta_i\eta_i^T|\mathbf{X}_i] = \sigma_\eta^2\mathbf{I}_T$

The variance-covariance matrix $\mathbf{\Sigma}$ is obtained by substituting $\theta = 0$ in equation (4.25).

$$\operatorname{cov}\left(\varepsilon_{i,t}, \varepsilon_{i,t+s}\right) = \frac{\rho^s}{1-\rho^2} \tag{4.27}$$

Prais and Winsten (1954) show that in this case the GLS transformation matrix yields

$$\mathbf{\Sigma}^{-1/2} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\rho & 1 \end{bmatrix} \tag{4.28}$$

That is, we have an expression of the GLS estimator. To obtain a feasible GLS estimator, a consistent estimator for ρ is required. Wooldridge's two stage procedure is also useful to obtain a consistent estimator for ρ. We derive it using the method proposed by

Cochrane and Orcutt (1949). In the first stage, the pooled OLS model is fitted and we apply pooled OLS again on (where $\hat{\varepsilon}_{i,t}$ are the pooled OLS residuals)

$$\hat{\varepsilon}_{i,t} \text{ on } \hat{\varepsilon}_{i,t-1}. \tag{4.29}$$

so that the estimator for ρ is given by

$$\hat{\rho} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\varepsilon}_{i,t-1}^2 \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\varepsilon}_{i,t-1} \hat{\varepsilon}_{i,t} \right). \tag{4.30}$$

$\hat{\rho}$ consistently estimates ρ as $N \to \infty$. Furthermore, Wooldridge (1991) shows that the usual asymptotic properties hold for this estimator.

It is theoretically possible to extend this procedure to general $AR(p)$, $p \leq T$ processes, but this gets increasingly complicated as p gets higher since we need to compute $\mathbf{\Sigma}^{-1/2}$. Note that this is not very restrictive from a practical point of view, since in most econometric applications $\rho_1 > \rho_2 > \cdots > \rho_p$. For stationarity, $|\sum_{j=1}^{P} \rho_j| < 1$ and so the influence of the higher order terms dies out quickly anyway. If we expect high order autocorrelation, it may be more useful to use transformation matrix (4.28) and robustify the standard errors for the remainder serial correlation rather than computing $\mathbf{\Sigma}^{-1/2}$ for the small efficiency gain. Baltagi (2005, p. 87) derives a transformation matrix $\mathbf{\Sigma}^{-1/2}$ for a second order autoregressive process, which may be useful if ρ_2 is expected to be large. The model given in (4.29) can easily be extended for an estimate of the second order correlation coefficient.

While correcting for autoregressive errors provides a asymptotic solution to the problem, the estimates for ρ are biased. To see this,

$$\mathbb{E}[\hat{\rho}] = \rho + \mathbb{E}\left[\left(\sum_{i=1}^{N} \sum_{i=1}^{T} \hat{\varepsilon}_{i,t-1}^2 \right)^{-1} \left(\sum_{i=1}^{N} \sum_{i=1}^{T} \hat{\varepsilon}_{i,t-1} \eta_{i,t} \right) \right] \tag{4.31}$$

For ρ to unbiased, $\eta_{i,t}$ is mean independent of $\{\hat{\varepsilon}_{i,t}\}_{i=1,t=1}^{N,T}$. But this fails, as

$$\hat{\varepsilon}_{i,t+1} = \rho \hat{\varepsilon}_{i,t} + \eta_{i,t+1}. \tag{4.32}$$

4.2.4 Simulations

This paragraph considers a few simulations for the feasible GLS estimator. In particular, the performance between this estimator and the first difference estimator is compared. We consider a panel with $N = 50$ cross-sectional units and $T = 20$ time periods. The model is given by

$$y_i = \beta \mathbf{X}_i + \varepsilon_i \tag{4.33}$$

The true value of $\beta = 2$. The error term follows an $ARMA(2,1)$ process, i.e.

$$\varepsilon_{i,t} = \rho_1 \varepsilon_{i,t-1} + \rho_2 \varepsilon_{i,t-2} + \theta \eta_{i,t-1} + \eta_{i,t} \tag{4.34}$$

where $\eta_i \overset{iid}{\sim} N(0, \mathbf{I}_T)$. The true value of $\rho_1 = 0.8$, $\rho_2 = 0.1$, $\theta = 0.2$. These numbers are chosen such that they roughly correspond with the serial correlation dependence structure in the 2001 and 2012 VAT studies. The aim of this simulation is to show that first differencing performs very well in practical applications, even if the error term does not follow a random walk. The first difference estimates are given in 4.3 and the feasible GLS restricted to $ARMA(1,1)$ errors in table 4.4. $ARMA(2,2)$ posed convergence problems due to ρ_2 being small.

From the tables we can observe that the first difference estimator indeed does not underform for the feasible GLS estimator. Only the type I errors produced by the FGLS estimator is somehwat higher than the type I errors produced by the feasible GLS error. This may be an indication that the robust standard errors underestimate the actual standard errors. However, there is no reason to worry about this. The autocorrelation function (ACF) plots of both the first difference and feasible GLS estimator indeed show that serial correlation disappeared.

Table 4.3: First difference summary (R=3000 simulations)

	Estimate	Std. Error	t value	Pr(>|t|)	Type I	Type II
beta	2.0001	0.0088	230.8955	0	5.73%	0%

Table 4.4: Feasible GLS summary (R=3000 simulations)

	Estimate	Std. Error	z value	Pr(>|z|)	Type I	Type II
beta	2.0001	0.0089	226.0238	0	5.13%	0%

Figure 4.4: ACF plot of First Difference estimator

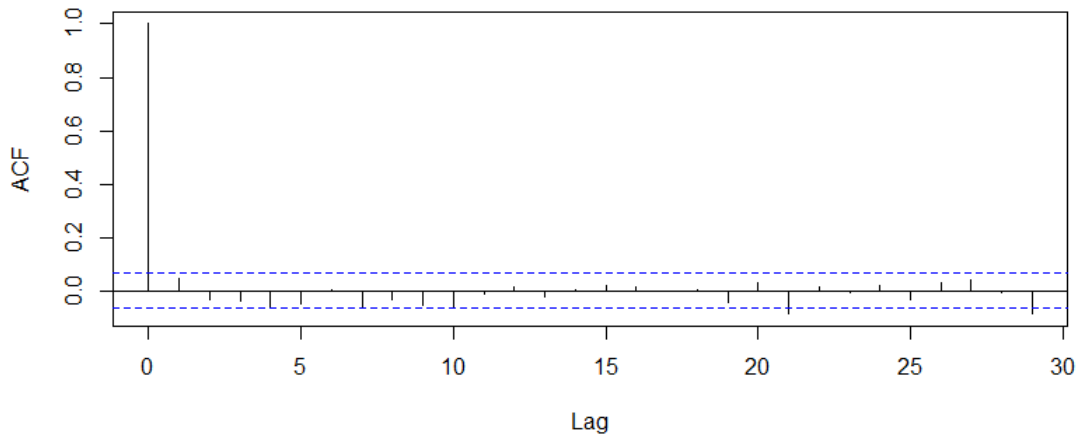
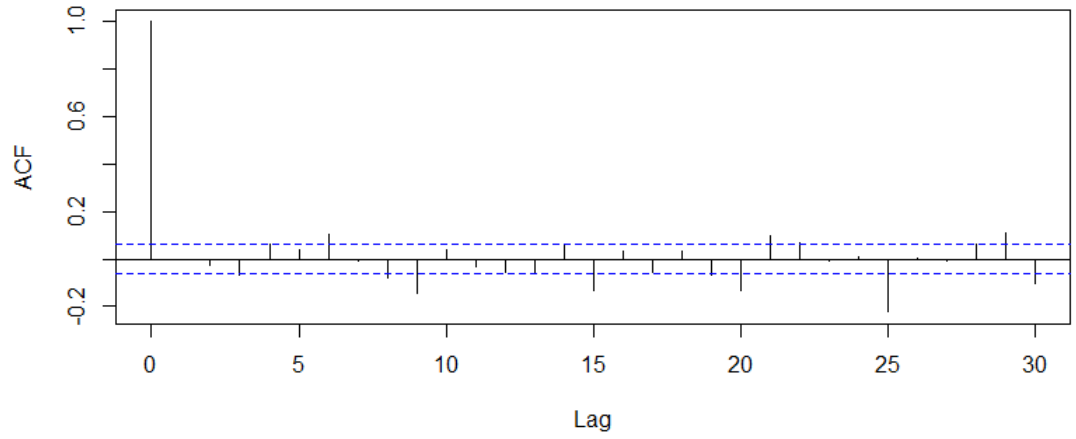


Figure 4.5: ACF plot of Feasible Generalised Least Squares estimator



4.2.5 Maximum Likelihood Estimation

The feasible GLS estimator is also a maximum likelihood estimator. We do not impose assumptions on the variance-covariance matrix $\mathbf{\Sigma}$, meaning that the maximum likelihood estimator in this case is also unrestricted. We impose the following assumptions:

- **Assumption MLE.FGLS.1:** $\varepsilon_i|\mathbf{X}_i \overset{iid}{\sim} N(0, \mathbf{\Sigma})$
- **Assumption MLE.FGLS.2:** $\mathbb{P}[\text{rank}\left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{\Sigma}^{-1} \mathbf{X}_i\right) = K] = 1$

This means

$$y_i|\mathbf{X}_i \overset{iid}{\sim} N(\mathbf{X}_i\beta, \mathbf{\Sigma}) \tag{4.35}$$

and so the MLE estimators can be derived as

$$\hat{\beta}_{MLE} = \left(\sum_{i=1}^{N} \mathbf{X}_i^T \hat{\mathbf{\Sigma}}_{MLE}^{-1} \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^T \hat{\mathbf{\Sigma}}_{MLE}^{-1} \mathbf{X}_i^T\right)$$

$$\hat{\mathbf{\Sigma}}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} \hat{\varepsilon}_i \hat{\varepsilon}_i^T$$

where $\hat{\varepsilon}_i = \left(y_i - \mathbf{X}_i \hat{\beta}_{MLE}\right)$.

The MLE estimator uses the residuals obtained by the FGLS estimator rather than the OLS estimator. This is, given the nature of the MLE estimator, a more efficient procedure. However, it is computationally more expensive than the GMM FGLS estimator derived above. The GMM FGLS estimator does not require simultaneous estimation of the parameters.

The estimator for the variance matrix easily shows why a fully correct specification of the variance matrix is also infeasible under a MLE setting: also the MLE variance-covariance matrix requires the estimation of $O\left(T^2\right)$ variance parameters. This usually only works well provided N is much bigger than T. Furthermore, in many practical applications such model formulation is overspecified.

4.3 Random Effects Estimator

The random effects estimator is a feasible GLS-based estimator, yet it shows much more similarities with the fixed effect estimator. Recall that in fixed effects estimation, the unobserved fixed effect μ_i was correlated with the design matrix $\mathbf{X}_i$. The fixed effects estimator made μ_i an estimand in the model. In the random effect models, μ_i is assumed to be *random*, just like in the Pooled OLS case. However, the random effects estimators allow μ_i to have a different variance, which is where the FGLS method kicks in. The random effects estimator may have some efficiency gains over pooled OLS.

4.3.1 Asymptotic Properties

The Random Effects estimator is best described as an example of a restricted FGLS method, since only two variance parameters need to be estimated. The Random Effects estimator is justified by the following set of assumptions.

- **Assumption A.RE.1:** $\mathbb{E}\left[v_{i,t}|\mathbf{X}_i, \mu_i\right] = 0,\ t = 1, \ldots, T.$
- **Assumption A.RE.2:** $\mathbb{E}\left[\mu_i|\mathbf{X}_i\right] = \mathbb{E}\left[\mu_i\right] = 0$, indepentent in mean.
- **Assumption A.RE.3:** $\text{Rank}\left(\mathbb{E}\left[\mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right]\right) = K$, with $\boldsymbol{\Sigma} = \mathbb{E}[\varepsilon_i\varepsilon_i^T]$.
- **Assumption A.RE.4:** $\mathbb{E}\left[v_i v_i^t|\mathbf{X}_i, \mu_i\right] = \sigma_v^2 I_t$.
- **Assumption A.RE.5:** $\mathbb{E}\left[\mu_i^2|\mathbf{X}_i\right] = \sigma_\mu^2$

As usual, assumptions A.RE.1 - A.RE.3 are the key identification assumptions while assumptions A.RE.4 - A.RE.5 are only needed for efficiency.

Lemma 4.3.1. *Under REE.1-REE.5, we can write*

$$\boldsymbol{\Sigma} = \sigma_v^2 I_T + \sigma_\mu^2 \vec{1}_T \vec{1}_T^T \tag{4.36}$$

Proof.

$$\begin{aligned}
\mathbb{E}[\varepsilon_i\varepsilon^T] &= \mathbb{E}[\left(\mu_i\vec{1}_T + v_i\right)\left(\mu_i\vec{1}_T + v_i\right)^T] \\
&= \underbrace{\mathbb{E}[\mu_i^2\vec{1}_T\vec{1}^T]}_{\substack{\sigma_\mu^2\vec{1}\vec{1}^T \\ \text{(REE.5)}}} + \underbrace{\mathbb{E}[\mu_i\vec{1}_T v_i^T] + \mathbb{E}[v_i\vec{1}_T^T\mu_i^T]}_{\substack{0 \\ \text{(LIE/A.RE.1)}}} + \underbrace{\mathbb{E}[v_i v_i^T]}_{\substack{\sigma_v^2\mathbf{I}_T \\ \text{(A.RE.4)}}}
\end{aligned}$$

□

Assume we have consistent estimators $\hat{\sigma}_\mu$ and $\hat{\sigma}_v$. Then $\mathbf{\Sigma}$ is consistently estimated by

$$\hat{\mathbf{\Sigma}} = \hat{\sigma}_\mu^2 \vec{1}_T \vec{1}_T^T + \hat{\sigma}_v^2 I_T \tag{4.37}$$

The random effects estimator is now defined as the feasible GLS estimator with the restricted variance-covariance matrix $\hat{\mathbf{\Sigma}}$, i.e.

$$\hat{\beta}_{RE} = \left(\sum_{i=1}^{N} \mathbf{X}_i^T \hat{\mathbf{\Sigma}}^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^T \hat{\mathbf{\Sigma}}^{-1} y_i \right) \tag{4.38}$$

Given the consistency of the feasible GLS estimator, we can also conclude that the random effects estimator is consistent.

Asymptotic Variance

The random effects estimator is obviously normally distributed and has an asymptotic variance given by (under A.RE.1 - A.RE.5)

$$\hat{\beta}_{RE} = \mathbb{E}[\mathbf{X}_i^T \mathbf{\Sigma}^{-1} \mathbf{X}_i]^{-1} \tag{4.39}$$

The main purpose of this paragraph is to derive consistent estimators for σ_μ^2

Lemma 4.3.2. *Assume A.RE.1 - A.RE.5 hold, then*

$$\mathbb{E}[\varepsilon_{i,t}^2] = \mathbb{E}[\mu_i^2] + \mathbb{E}[v_{i,t}^2] \tag{4.40}$$

Proof.

$$\begin{aligned} \mathbb{E}[\varepsilon_{i,t}^2] &\overset{RE.1}{=} \mathbb{E}[\left(\mu_i^2\right)] - \cancel{2\mathbb{E}[\mu_i v_{i,t}]} + \mathbb{E}[v_{i,t}^2] \\ &\overset{RE.4/5}{=} \sigma_\mu^2 + \sigma_v^2 \end{aligned}$$

□

In the same way as the FGLS estimator, we use the pooled OLS residuals to obtain consistent estimates for σ_μ^2 and σ_v^2. Recall $\mathbb{E}[v_{i,t} v_{i,s}] = \sigma_\mu^2,\ t \neq s$. Note that all possible combinations are given by:

$$\begin{aligned} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \mathbb{E}[v_{i,t} v_{i,s}] &= \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \sigma_\mu^2 \\ &= \frac{T(T-1)}{2} \sigma_\mu^2 \\ \iff \sigma_\mu^2 &= \frac{1}{T(T-1)/2} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \mathbb{E}[v_{i,t} v_{i,s}] \end{aligned} \tag{4.41}$$

Equation (4.41) justifies the following proposed estimator.

Corollary 4.3.1. *The estimator*

$$\hat{\sigma}_{\mu}^{2} = \frac{1}{NT\left(T-1\right)/2} \sum_{n=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{\varepsilon}_{i,t} \hat{\varepsilon}_{i,s} \tag{4.42}$$

consistently estimates σ_{μ}^{2}.

Proof. This is a direct consequence of equation (4.42) and the strong law of large numbers. □

That heteroscedastic and serial correlation robust asymptotic variance-covariance matrix is obtained the usual way and given by

$$\text{Avar}\left(\hat{\beta}_{RE}\right) = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} \tag{4.43}$$

where $\mathbf{A} = \mathbb{E}[\mathbf{X}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{X}_i]$ and $B = \mathbb{E}[\mathbf{X}_i^T \boldsymbol{\Sigma}^{-1} \varepsilon_i \varepsilon_i \boldsymbol{\Sigma}^{-1} \mathbf{X}_i]$.

Efficiency

Given the fact that the feasible GLS estimator is relatively efficient in the class of transformed pooled OLS estimators, it should not come as a surprise that the random effects estimator is asymptotically efficient too. After all, the Random Effects estimator is just a feasible GLS estimator. In particular, it is more efficient than the fixed effects estimator. Wooldridge (2001) argues this fact with a unnecessarily complicated proof, which we will not consider in detail. This proof, however, does show a very important fact: the random effects estimator is in fact the pooled OLS estimator applied on a quasi time demeaned system of equations. In particular, it is the pooled OLS estimator on

$$y_{i,t} - \lambda \bar{y}_i = \left(x_{i,t} - \lambda \bar{x}_i\right)\beta + \varepsilon_{i,t} - \lambda \varepsilon_{i,t}. \tag{4.44}$$

where $\lambda = 1 - \sqrt{1/[1 + T\left(\hat{\sigma}_{\mu}^{2}/\hat{\sigma}_{v}^{2}\right)]}$. Apart from the fact that this quasi-time demeaning shows an obvious relationship between the random effects and fixed effects estimator, it also shows that as $T \to \infty$, the estimates for the random effects and fixed effects are equal. This in particular suggests that for sufficiently big T, there is no real reason to prefer the random effects estimator over the fixed effects estimator.

4.4 Summary

This chapter discussed several estimators that aim to improve efficiency in presence of heteroscedasticity and serial correlation. We have seen that the generalised least squares estimator preserves Gauss-Markov and asymptotic efficiency in presence of arbitrary serial correlation and heteroscedasticity. At the other hand, it is almost always infeasible

in practical applications because it requires knowing the variance-covariance matrix of the error term up to a constant.

The GLS estimator was then extended to the feasible GLS estimator, which aims to solve the infeasibility of the GLS estimator by replacing the variance-covariance matrix of the error term by a consistent estimator. The feasible GLS estimator is an example of a so-defined parametric transformation of the pooled OLS estimator. While the feasible GLS estimator preserves the same asymptotic efficiency properties as the GLS estimator, it may have poor finite sample properties compared to other consistent estimators. The FGLS framework was then used to derive an estimator that is efficient in presence of $ARMA\,(p, q)$ serial correlation and in presence of exogenous individual specific fixed effects. The latter captures inefficiencies in presence of exogeneous unobserved individual specific fixed effects. The simulations showed that first differencing in presence of heavy serial dependence hardly underpeforms compared to modelling the serial correlation in the error term.

Chapter 5

Model specification and causality

This chapter discusses issues related to model specifications. It starts by considering solutions to endogeneity problems, where special emphasis is put on the method of instrumental variables regression. The analysis then continues to considering some specification tests to determine which model is the most useful to apply in practice. Topics include the determination of the presence of serial correlation in the error term and endogeneity caused by unobserved individual specific fixed effects. This chapter is concluded by discussing causal effects, where we especially consider the famous difference-in-differences model. It appears that these models are heavily affected by serial correlation, making the usual pooled OLS estimator inefficient.

5.1 Endogeneity

The previous chapters discussed panel data models where $\mathbf{X}_i$ was assumed to be exogenous. This appeared to be the key assumption for the consistency of such models. The fixed effects estimator provided a first relaxation. This estimator appeared to consistently estimate β in presencde of an endogenous unobserved fixed effect μ_i, but the fixed effects estimator still required $\mathbb{E}[v_i|\mathbf{X}_i] = 0$.

These exogeneity assumptions may be appropriate for experimental studies, but often are not for observational research. We will only very briefly discuss methods to circumvent inconsistency caused by endogenous regressors. Later on in this thesis we actually do observe endogenous regressors, but the methods in this paragraph will appear to be infeasible to implement.

Consider the T-dimensional panel data model.

$$\underset{T\times 1}{y_i} = \underset{T\times K}{\mathbf{X}_i}\ \underset{K\times 1}{\beta} + \underset{T\times 1}{\varepsilon_i} \tag{5.1}$$

Let $\underset{T\times L}{\mathbf{Z}_i}$ the matrix with L observable instrumental variables. IV-based estimators comply with the following set of assumptions.

- **Assumption IV.1:** $\mathbb{E}[\mathbf{Z}_i^T \varepsilon_i] = 0$.
- **Assumption IV.2:** $\text{rank}\left(\mathbb{E}[\mathbf{Z}_i^T \mathbf{X}_i]\right) = K$.

Since IV.1 does not imply that $\mathbb{E}[\mathbf{X}_i^T (y_i - \mathbf{X}_i\beta)] = 0$, it suggests replacing $\mathbf{X}_i^T$ with $\mathbf{Z}_i^T$ since $\mathbb{E}[\mathbf{Z}_i^T (y_i - \mathbf{X}_i\beta)] = 0$. A consistent estimator for β is then given by

$$\hat{\beta}_{IV} = \left(\sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{Z}_i^T y_i\right) \tag{5.2}$$

Theorem 5.1.1. *The estimator given by* (5.2) *consistently estimated* β.

Proof. Recall $\mathbb{E}[\mathbf{Z}_i^T (y_i - \mathbf{X}_i\beta)] \overset{IV.1}{=} 0$, hence

$$\begin{aligned}
\mathbb{E}[\mathbf{Z}_i^T (y_i - \mathbf{X}_i\beta)] &= \mathbb{E}[\mathbf{Z}_i^T y_i - \mathbf{Z}_i^T \mathbf{X}_i\beta] \\
&\overset{SLLN}{=} \text{as-lim}\, \frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}_i^T y_i - \mathbf{Z}_i^T \mathbf{X}_i\beta \\
\overset{IV.2}{\iff} \beta &= \text{as-lim} \underbrace{\left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{X}_i\right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}_i^T y_i\right)}_{\hat{\beta}_{IV}}
\end{aligned}$$

□

5.1.1 Pooled Two-Stage Least Squares Estimator

The method discussed above only works when $L = K$. If $L > K$, then the $\hat{\beta}_{IV}$ estimator does not usually have a solution (over identification). One way to proceed is to choose $\hat{\beta}$ such that it satisfies

$$\min_b \left(\sum_{i=1}^{N} \mathbf{Z}_i^T (y_i - \mathbf{X}_i b)\right)^T \left(\sum_{i=1}^{N} \mathbf{Z}_i^T (y_i - \mathbf{X}_i b)\right) \tag{5.3}$$

This fulfils the rank condition. More efficient, however, is the generalised method of moments estimator. In addition to IV.1-IV.2, let $\hat{W}$ be a weighting matrix that is positive, semi-definite. An important application of an GMM-estimator is the Two-Stage Least Squares (2SLS) estimator. The 2SLS estimator follows from choosing

$$\hat{\mathbf{W}} = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{Z}_i\right) \tag{5.4}$$

Obviously,

$$\underset{N\to\infty}{\text{as-lim}}\,\hat{\mathbf{W}} = \mathbb{E}[\mathbf{Z}_i^T\mathbf{Z}_i]$$

The **2SLS estimator** is given by

$$\begin{aligned}\hat{\beta}_{2SLS} = &\left[\left(\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{Z}_i\right)\left(\sum_{i=1}^{N}\mathbf{Z}_i^T\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{Z}_i^T\mathbf{X}_i\right)\right]^{-1}\\ &\times\left[\left(\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{Z}_i\right)\left(\sum_{i=1}^{N}\mathbf{Z}_i^T\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{Z}_i^T y_i\right)\right]\end{aligned} \tag{5.5}$$

To see why this estimator follows a two-stage procedure, we solve first the equation

$$\underset{T\times K}{\mathbf{X}_i} = \underset{T\times K}{\mathbf{Z}_i}\,\underset{K\times K}{\beta_{1S}} + \underset{T\times K}{\varepsilon_i}\,. \tag{5.6}$$

where estimator for β_{1S} is given by

$$\hat{\beta}_{1S} = \left(\sum_{i=1}^{N}\mathbf{Z}_i^T\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{Z}_i^T\mathbf{X}_i\right) \tag{5.7}$$

The predicted values become

$$\hat{\mathbf{X}}_i = \mathbf{Z}_i\left(\sum_{i=1}^{N}\mathbf{Z}_i^T\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{Z}_i\mathbf{X}_i\right) \tag{5.8}$$

The second step,

$$y_i = \hat{\mathbf{X}}_i\beta_{2S} + \varepsilon_i \tag{5.9}$$

Which leads to

Theorem 5.1.2 (Own work)**.** *Under the two-stage least squares procedure, we have*

$$\hat{\beta}_{2S} = \left(\sum_{i=1}^{N}\hat{\mathbf{X}}_i^T\hat{\mathbf{X}}_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{\mathbf{X}}_i^T y_i\right) = \hat{\beta}_{2SLS} \tag{5.10}$$

Proof.

$$
\begin{aligned}
\hat{\beta}_{2S} &= \left(\sum_{i=1}^{N} \hat{\mathbf{X}}_i^T \hat{\mathbf{X}}_i\right)^{-1} \left(\sum_{i=1}^{N} \hat{\mathbf{X}}_i^T y_i\right) \\
&= \left[\left(\sum_{i=1}^{N} \mathbf{Z}_i \left(\sum_{i=1} \mathbf{Z}_i^T \mathbf{Z}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{X}_i\right)\right)^T \right. \\
&\quad \left. \times \left(\mathbf{Z}_i \left(\sum_{i=1}^{T} \mathbf{Z}_i^T \mathbf{Z}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{X}_i\right)\right)\right]^{-1} \\
&\quad \times \left[\sum_{i=1}^{N} \mathbf{Z}_i \left(\sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{Z}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{X}_i\right)^T y_i\right] \\
&= \left[\left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{Z}_i\right) \left(\sum_{i=1} \mathbf{Z}_i^T \mathbf{Z}_i\right)^{-1} \right. \\
&\quad \left. \times \cancel{\left(\sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{Z}_i\right) \left(\sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{Z}_i\right)^{-1}} \left(\sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{X}_i\right)\right]^{-1} \\
&\quad \times \left[\left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{Z}_i\right) \left(\sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{Z}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{Z}_i^T \varepsilon_i\right)\right] \\
&= \left[\left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{Z}_i\right) \left(\sum_{i=1}^{T} \mathbf{Z}_i^T \mathbf{Z}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{X}_i\right)\right]^{-1} \\
&\quad \times \left[\left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{Z}_i\right) \left(\sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{Z}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{Z}_i^T \varepsilon_i\right)\right] \\
&= \hat{\beta}_{2SLS}
\end{aligned}
$$

As required. □

It is not recommended to actually perform 2SLS regression under a two-stage procedure, i.e. the standard errors are incorrect and may lead to improper inference (Wooldridge, 2001).

5.2 Specification tests

5.2.1 Presence of Fixed Effects

In presence of an individual specific unobserved fixed effect, the Pooled OLS estimator is either inefficient or inconsistent. Breusch and Pagan (1979) developed a test for the presence of such effects, which to date is among the most commonly used. However, this test relies on a full parametric specification of the model. Consequently, this test is infeasible since no distributions have been specified so far. Wooldridge (2001, p. 265) proposes a semi-parametric that tests

$$H_0 : \sigma_\mu^2 = 0$$

He proposes the following test statistic

$$H = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{\varepsilon}_{i,t} \hat{\varepsilon}_{i,s}}{\left[\sum_{i=1}^{N} \left(\sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{\varepsilon}_{i,t} \hat{\varepsilon}_{i,s} \right)^2 \right]^{1/2}} \underset{H_0}{\overset{d}{\rightarrow}} N(0,1) \tag{5.11}$$

I believe the H_0 of is poorly formulated, because this test in fact checks if the lower diagonal of the variance-covariance matrix (excluding the diagonal itself) has non-zero elements. While the variance-covariance matrix will have non-zero elements in presence of individual specific fixed effects, it will also have non-zero entries in presence of serial correlation. This means that H_0 should also be rejected in presence of serial correlation. Wooldridge (2001) acknowledges this fact and suggests that only a non-rejection of H_0 is evidence to the fact that $\sigma_\mu^2 = 0$. In all other cases, other methods should be consulted to get a decisive answer. In my opinion, a better formulation of the hypothesis test is given by

H_0 : No serial correlation and individual specific effects
H_1 : Assumption A.POLS.3 fails.

In practice, this test only really matters if a fixed effects transformation or first difference transformation is undesirable. For example when T is small, a variable of interest is time-invariant or the amount of cross-sectional units N is very large (the latter causing computational inefficiencies). If this is not a problem, fixed effects or first difference is a safe option as it takes care of the individual specific fixed effect while still producing consistent estimates.

5.2.2 Hausman Test

The Hausman test is a test constructed to two different estimators against eachother. The Hausman test in its purest form is hardly informative, because it tests

$$H_0: \operatorname*{as\text{-}lim}_{N\to\infty}\left(\hat{\beta}-\tilde{\beta}\right)=0, \text{ but } \hat{\beta} \text{ is more efficient than } \tilde{\beta}$$
$$H_1: \operatorname*{as\text{-}lim}_{N\to\infty}\left(\hat{\beta}-\tilde{\beta}\right)\neq 0$$

i.e., under the alternative hypothesis, at least one of the estimators is inconsistent. Under additional assumptions, the Hausman test may lead to more useful hypothesis testing. First, the most general form of the Hausman test will be derived and applied on specific cases.

Under H_0,

$$\sqrt{N}\left(\hat{\beta}-\tilde{\beta}\right)\underset{H_0}{\overset{d}{\to}} N\left(0,\boldsymbol{\Xi}\right) \tag{5.12}$$

where $\boldsymbol{\Xi}$ is a variance-covariance matrix. Hausman suggests using the test statistic

$$H=\left(\hat{\beta}-\tilde{\beta}\right)^T\left(\hat{\boldsymbol{\Xi}}/N\right)^{-1}\left(\hat{\beta}-\tilde{\beta}\right) \tag{5.13}$$

where $\hat{\boldsymbol{\Xi}}$ is a consistent estimator of the variance covariance matrix.

Also,

$$\boldsymbol{\Xi}=\operatorname{Avar}[\hat{\beta}-\tilde{\beta}]=\operatorname{Avar}\left(\hat{\beta}\right)+\operatorname{Avar}\left(\tilde{\beta}\right)-2\operatorname{Acov}\left(\hat{\beta},\tilde{\beta}\right) \tag{5.14}$$

Lemma 5.2.1. *If $\hat{\beta}$ is more efficient than $\tilde{\beta}$ (i.e. under H_0), then* $\operatorname{Acov}\left(\hat{\beta},\tilde{\beta}\right)=\operatorname{Avar}\left(\hat{\beta}\right)$.

Proof. Proof by contradiction.

Assume that

$$\operatorname{Acov}\left(\hat{\beta},\tilde{\beta}\right)\neq\operatorname{Avar}\left(\hat{\beta}\right) \tag{5.15}$$

a contradiction can now be derived. As suggested by Amemiya (1985, p. 146), consider the estimator

$$\bar{\beta}=\hat{\beta}+\left(\operatorname{Avar}\left(\hat{\beta}\right)-\operatorname{Acov}\left(\hat{\beta},\tilde{\beta}\right)\right)\left(\operatorname{Avar}\left(\hat{\beta}-\tilde{\beta}\right)\right)^{-1}\left(\tilde{\beta}-\hat{\beta}\right) \tag{5.16}$$

Note that $\mathbb{E}[\bar{\beta}]=\beta$. Its asymptotic variance is given by

$$\begin{aligned}\operatorname{Avar}\left(\bar{\beta}\right)=\operatorname{Avar}\left(\hat{\beta}\right)&-\left(\operatorname{Avar}\left(\hat{\beta}\right)-\operatorname{Acov}\left(\hat{\beta},\tilde{\beta}\right)\right)\\&\times\operatorname{Avar}\left(\hat{\beta}-\tilde{\beta}\right)^{-1}\left(\operatorname{Avar}\left(\hat{\beta}\right)-\operatorname{Acov}\left(\hat{\beta},\tilde{\beta}\right)\right)\end{aligned}$$

This is a contradiction, because $\hat{\beta}$ was more efficient. □

This lemma in particular implies that

$$\text{Avar}\left(\tilde{\beta}\right) - \text{Avar}\left(\hat{\beta}\right) \tag{5.17}$$

is positive definite and so

$$H \underset{H_0}{\overset{d}{\rightarrow}} \chi_k^2 \tag{5.18}$$

In a similar fashion as shown in paragraph 1.4.2, H_0 is rejected at signficance level α when

$$H > C_\alpha \tag{5.19}$$

where $C_\alpha = F^{-1}(1-\alpha)$, the inverse distribution function of χ_k^2.

Random effects vs. fixed Effects

Consider the random effects estimator $\hat{\beta}_{RE}$ and the fixed effects estimator $\hat{\beta}_{FE}$. Under A.RE.1-A.RE.5, the Fixed Effects estimator is consistent but the Random Effects estimator is consistent and BLUE. In particular, the Random Effects estimator will be more efficient. The Haussman hypothesis for the comparison of Random Effects and Fixed Effects estimators can so be formulated as

$$\begin{aligned} H_0: &\text{ Assumptions A.RE.1 - A.RE.5 hold and so } \underset{N\to\infty}{\text{as-lim}}\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) = 0 \\ H_1: &\text{ Assumption A.RE.2 fails and so } \beta = \underset{N\to\infty}{\text{as-lim}}\,\hat{\beta}_{FE} \neq \underset{N\to\infty}{\text{as-lim}}\,\hat{\beta}_{RE} \end{aligned}$$

Should A.RE.2 fail (endogeneity between the regressors and the individual specific effect), then the RE estimator is inconsistent while the FE estimator is consistent.

General endogeneity

Any OLS-based estimator can be compared with an Instrument Variable based estimator. IV-based estimators are less efficient than OLS-based estimators under the usual assumptions, so a Hauseman test can easily be formulated as

$$\begin{aligned} H_0: &\text{ Usual OLS assumptions hold and } \underset{N\to\infty}{\text{as-lim}}\left(\hat{\beta}_{IV} - \hat{\beta}_{OLS}\right) = 0 \\ H_1: &\text{ IV errors are uncorrelated } \beta = \underset{N\to\infty}{\text{as-lim}}\,\hat{\beta}_{IV} \neq \underset{N\to\infty}{\text{as-lim}}\,\hat{\beta}_{OLS} \end{aligned}$$

Are Hauseman Test result reliable?

The random effects estimator asymptotically efficient provided assumptions A.RE.1-A.RE.5 hold. However, if either A.RE.4 or A.RE.5 fail, then the random effects model loses this property. In particular, nothing guarantees that the difference between the

fixed effects and random effects asymptotic variances constitutes in a positive semidefinite matrix. Equation (5.14) provides an unrestricted form to perform Hauseman tests, but it requires knowing Acov $\left(\hat{\beta}_{RE}, \hat{\beta}_{FE}\right)$. As shown in lemma 5.2.1, this equals the asymptotic variance of the most efficient estimator and thus requires us to know (apriori) which estimator is most efficient. In case of heteroscedasticity or autocorrelation, this is something unknown and consequently the Hausman tests may lead to incorrect inference.

5.2.3 Test for Serial Correlation

Serial correlation in the error term may cause estimates to be inefficient. Hence, it is a useful to test for serial correlation. A proper test for serial correlation is usually tedious and often requires large T asymptotics. Wooldridge (2001), on the contrary, proposes a test for serial correlation that works without distributional assumptions. The test uses an AR(1) model for the residuals as described in equation (5.20).

$$\hat{\varepsilon}_{i,t} = \rho\hat{\varepsilon}_{i,t-1} + \eta_{i,t} \tag{5.20}$$

The test applies pooled OLS on equation (5.20) to estimate ρ. The appropriate hypothesis tests depend on the estimator that was used. For example, if a First Difference estimator is used and the interest lies in the question whether the original error term is serially uncorrelated, then

$$H_0 : \rho \neq -\frac{1}{2}$$
$$H_1 : \rho = -\frac{1}{2}$$

Under H_0, the original error term is serially correlated while under H_1 there is an *indication* that the original error term is serially uncorrelated (see paragraph 3.4.1).

This test is easy to implement, yet it has some disadvantages. For example, in a Pooled OLS setting it was required that $\mathbb{E}[\varepsilon_i\varepsilon_i^T|\mathbf{X}_i] = \sigma_\varepsilon^2\mathbf{I}_T$, i.e. the error term is serially uncorrelated for all time periods (future and past), not just for one time period back. While $AR(1)$ dependence is usually among the most important, it does not rule out other types of serial dependence. That is, this test captures potential efficiency problems caused by serial correlation, but rejection of the null hypothesis does not mean that the error term actually follows an $AR(p)$ process, it may easily be some arbitrary $ARIMA(p, r, q)$ process.

5.3 Causal Treatment Effects

The basic justification behind causal inference is represented by the standard statistical adagium that correlation does not imply causation. In particular, significant estimates by any of the previously discussed estimators do not necessarily imply causation. This

chapter will discuss methods to check for causal links. This chapter wil first discuss the theory behind Average Treatment Effects (ATE) and then continue with the difference-in-differences-estimator. The latter heavily relies on the ATE theory. In classical ATE theory, the treatment assignment is assumed to be ignorable. However, in many cases (in particular in this research), treatment assignment cannot be ignored.

5.3.1 Causal Effects

Let $y^0_{i,t}$ denote the outcome without treatment and $y^1_{i,t}$ the outcome with treatment. The interest lies in estimating the difference $y^0_{i,t} - y^1_{i,t}$. The fact that both $y^0_{i,t}$ and $y^1_{i,t}$ are random variables justifies the following definition of an average treatment effect.

Definition 5.3.1 (Average Treatment Effect - Rosenbaum and Rubin (1983))**.** The Average Treatment Effect is defined as

$$ATE = \mathbb{E}[y^0_{i,t} - y^0_{i,t}] \tag{5.21}$$

If w denotes a variable indicating treatment (with $w = 1$ equalling treatment and $w = 0$ otherwise), then the concept of the Average Treatment Effect can be extended with the following definition.

Definition 5.3.2 (Average Treatment Effect of the Treated — Wooldridge (2001, p. 605))**.** The Average Treatment Effect on the Treated, denoted as ATT, is defined as

$$ATT = \mathbb{E}[y^1_{i,t} - y^0_{i,t}|w = 1] \tag{5.22}$$

Equation (5.22) has the advantage that it only takes into account those groups actually eliagble for treatment.

Define the observable outcome as

$$y_{i,t} = \left(1 - w_{i,t}\right) y^0_{i,t} + w_{i,t} y^1_{i,t} \tag{5.23}$$

Following Wooldridge (2001), it is convinient to write

$$\begin{aligned} y^1_{i,t} &= \mu^1_i + v^1_{i,t} \\ y^0_{i,t} &= \mu^0_i + v^0_{i,t} \end{aligned}$$

Where $\mu_1 = \mathbb{E}[y^1_{i,t}]$ and $\mu^1_{i,t} = \mathbb{E}[y^1_{i,t}]$ and $v^1_{i,t}$ and $v^0_{i,t}$ the stochastic parts of $y^1_{i,t}$ and $y^0_{i,t}$, respectively. And so, equation (5.23) can be rewritten as

$$y_{i,t} = \mu^0_i + w\left(\mu^1_i - \mu^0_i\right) + v^0_{i,t} + w\left(v^1_{i,t} - v^0_{i,t}\right) \tag{5.24}$$

5.3.2 Estimator for Average Treatment Effect

In this section, Average Treatment Effect estimators will be proposed.

- **Assumption ATE.1:** $\mathbb{E}[\varepsilon_i^1|w_i] = 0$ and $\mathbb{E}[\varepsilon_i^0|w_i] = 0$.

To derive an estimator, note that under ATE.1 and ATE.2

$$\begin{aligned} \mathbb{E}[y_{i,t}^0] &= \mu_{i,t}^0 + \cancelto{0}{\mathbb{E}[v_{i,t}^0]} \\ \mathbb{E}[y_{i,t}^1] &= \mathbb{E}[y_{i,t}^0] + ATE \\ \implies ATE &= \mu_{i,t}^1 - \mu_{i,t}^0 \end{aligned}$$

The corresponding regression framework becomes[1]

$$\begin{aligned} \mathbb{E}[y_{i,t}|w] &= \mu_i^0 + ATE \cdot w_{i,t} + \mathbb{E}[v_{i,t}^0|w] + w_{i,t}\mathbb{E}[v_{i,t}^1 - v_{i,t}^0|w] \\ &\overset{ATE.1}{=} \mu_i^0 + ATE \cdot w_{i,t} \end{aligned} \tag{5.25}$$

that is to say, $\varepsilon_{i,t} = v_{i,t}^0 + w\left(v_{i,t}^1 - v_{i,t}^0\right)$ is a random variables with mean zero, which also justifies the use of $\varepsilon_{i,t}$

$$y_{i,t} = \mu_i^0 + ATE \cdot w_{i,t} + \varepsilon_{i,t} \tag{5.26}$$

This regression model may, with a superficial view, automatically suggest a fixed effect model. However, recall from the fixed effect analysis in paragraph 3.3 that the fixed effects estimate for equals $\mu_i^0 = \alpha^0 + \gamma_i^0$. And so, if $\mathbb{E}[\gamma_i^0|w_{i,t}] = \mathbb{E}[\gamma_i^0] = 0$, a random effect model may as well be appropriate.

Generalising this estimator to allow for exogenous control variables $x_{i,t}$ is trivial by changing the ATE assumptions to

- **Assumption ATE.1:** $\mathbb{E}[v_i^1|w, \mathbf{X}_i] = \mathbb{E}[v_i^1|\mathbf{X}_i] = 0$ and $\mathbb{E}[v_i^0|w, \mathbf{X}_i] = \mathbb{E}[v_i^0|\mathbf{X}_i] = 0$

The corresponding regression model becomes

$$y_{i,t} = \mu_i^0 + ATE \cdot w_{i,t} + x_{i,t}\beta + \varepsilon_{i,t} \tag{5.27}$$

5.3.3 Estimator for the Average Treatment Effect on the Treated

Before deriving an estimator for the ATT, it is important to note that the ATT and ATE are related through the following identity

$$ATT = ATE + \mathbb{E}[v_{i,t}^1 - v_{i,t}^0|w = 1] \tag{5.28}$$

[1] Own work

- **Assumption ATT.1:** $\mathbb{E}[v_i^1|w] = \mathbb{E}[v_i^1]$ and $\mathbb{E}[v_i^0|w] = \mathbb{E}[v_i^0]$

Assumption ATT.1 implies assumption ATE.1. Indeed, under ATT.1, $\mathbb{E}[v_{i,t}^1|w] = \mathbb{E}[v_{i,t}^0|w] = 0$. As a consequence, the same regression framework to estimate ATT becomes

$$y_{i,t} = \mu_{i,t}^0 + ATT \cdot w_{i,t} + \varepsilon_{i,t} \tag{5.29}$$

5.3.4 Application: Difference-in-differences estimation

The Difference-in-Differences estimator is one of the most popular causal estimators, but consequently also one of the most misunderstood estimators. Usually data is only available for (i) a treatment group before and after a treatment and (ii) for a group that does not receive treatment. This section deals with a slight generalisation and a formalisation of the Difference-in-Differences estimator proposed by Wooldridge, 2001, p. 130. Firstly, a model will be constructed for difference-in-differences estimation. Let $i = 1, \cdots, N$ denote the amount of cross sections, t indicates the treatment period, where $t = 0$ indicates pre-treatment and $t = 1$ treatment. Furthermore, let s_i indicate treatment status, where $s_i = 1$ means treatment and $s_i = 0$ means no treatment. Note that the treatment indicator is now defined as $w_{i,t} = t \cdot s_i$

The difference-in-differences model is given by

$$y_{i,t} = \alpha + \lambda t + \gamma s_i + \delta w_{i,t} + \varepsilon_{i,t} \tag{5.30}$$

The following assumption is imposed to get consistent estimates

- **Assumption DID.1:** $\mathbb{E}[\varepsilon_i|w_i] = 0$.

Let $y_{i,t}^0$ denote the outcome of the model without treatment and $y_{i,t}^1$ denote the outcome with treatment. Note that they do not need to be actually observed.

Corollary 5.3.1 (Own work)**.** *The difference-in-differences estimator estimates the Average Treatment Effect of the Treated, i.e.*

$$\delta = ATT \tag{5.31}$$

Proof. Observe that

$$\begin{aligned}\mathbb{E}[y_{i,t}^0] &= \alpha + \lambda t + \gamma s_i \\ \mathbb{E}[y_{i,t}^1] &= \alpha + \lambda t + \gamma s_i + \delta w_{i,t}\end{aligned}$$

Now, from the definition of ATT

$$\begin{aligned}ATT &= \mathbb{E}[y_{i,t}^1 - y_{i,t}^0|w_{i,t} = 1] \\ &= (\alpha + \lambda + \gamma + \delta) - (\alpha + \lambda + \gamma) = \delta\end{aligned}$$

As required. □

5.3.5 Reliability of Difference-in-Difference Estimates

The Difference-in-Differences estimator is very commonly used in econometric research. However, Bertrand et al. (2004) argued that its implementation in practice leaves much to be desired. They suggested that Difference-in-Differneces models are, apart from heteroscedasticity, often prone to serial correlation. Consequently, pooled OLS and the fixed effects estimator are inefficient and the standard errors reported in most statistical packages are wrong.

Bertrand et al. (2004) suggest that inference should at least be made using robust standard errors. However, estimators that are efficient in presence of serial correlation are not thoroughly discussed. They shortly consider modelling the serial correlation in the error term, but they advice against that for small samples due to estimation biases. They do not consider first differencing at all. It could be argued that the first difference estimator is not identified because of the inclusion of a time invariant variable, but this is obviously an incorrect interpretation.

To understand this point, recall that the first difference estimator is consistent in presence of an unobserved individual specific fixed effect. If no dummy variable for the control group was included, it would have been qualified as an omitted individual specific fixed effect. That is, the first difference estimator still consistently estimates the causal treatment effect.

5.4 Summary

This chapter started by discussing econometric methods to circumvent endogeneity. These methods provide consistent estimates from a mathematical perspective, but the application in practice is very difficult. The second paragraph dealt with specification tests. It appeared that the formulation of such tests is difficult due to the semi-parametric nature of the estimators discussed in this thesis. However, some specification tests were derived that did not require distributional assumptions, in particular tests that capture the presence of serial correlation.

The analysis was then turned to the estimation of causal treatment effects. The notion of Average Treatment Effect was introduced and some causal estimators were derived. An important application was the Difference-in-Differences model, a causal estimation technique that is usually robust to endogeneity. Special attention was paid to the paper of Bertrand et al. (2004). They argue that DD estimation is in many pratical applications heavily affected by serial correlation, reducing the accuracy of the conventional estimator. Given the simulations made in the previous chapters, first differencing is expected to give accurate and reliable results.

Part II

Application on VAT-increases

Chapter 6

Economics of Tax Incidence

This chapter focuses on the development of the theory that is required to properly understand the mechanisms of Value Added Taxation. We start by reviewing basic regulations of the European VAT system, which will prove to complicate capturing the incidence of VAT increases. This theory will be used to develop an economic model. This model is able to give theoretical predictions on how companies pass-through VAT to customers under some general equilibrium conditions. These results will be used to examine the reliability and performance of the estimators that are used to capture the causal treatment effect of the VAT-increases.

6.1 Summary of Value Added Taxation

The VAT aims to tax the consumption of goods and services of end consumers (business-to-customer transactions). This is achieved by taxing the company's revenue. More specifically, on every transaction between an entrepreneur and a customer, VAT is added on the transaction price and paid to the company. The company pays the collected tax to the tax authorities. Note that the tax collection is effectively outsourced to the enterprise, i.e. reducing the administrative burden of tax authorities.

Transactions made between two companies (business-to-business transactions) are *de facto* not taxed. While the supplying company needs to charge the price including VAT to the recipient company and pay this to the tax authorities, the recipient company is allowed to claim back the tax paid from the tax authorities. The following example gives some insight in how the system works in practice.

Consider a VAT-rate of 21% for a certain good A. Good A is produced by manufacturer X and sold to retailer Y (both companies):

Transaction X → Y	Price (€)
Good A (excl. VAT)	100.00
Value Added Tax	21.00
Total	121.00

This is a business-to-business transaction. Retailer Y pays the full amount (that is, €121) to manufacturer X. Manufacturer X pays the VAT he collects to the tax authorities. But, retailer Y is eligible to a refund of the €21 tax he paid. That is, he can claim back the €21 VAT he paid from the tax authorities. So in the end the VAT does not burden on retailer Y.

Retailer Y sells good A to customer Z:

Transaction Y → Z	Price (€)
Good A (excl. VAT)	200.00
Value Added Tax	42.00
Total	242.00

This is a business to customer transaction. The customer ends up paying €42 VAT and retailer Y has to transfer this to the tax authorities.

The VAT described above is applied in many countries around the world (notably not in the United States). However, this thesis shall only review the European VAT-system. The VAT is one of the few European harmonised taxes. All member states have to codify their VAT-systems in compliance with the EU VAT-directive. For example, the Netherlands has the *Wet op de Omzetbelasting*, Germany the *Umsatzsteuergesetz* and France the *Code de la Taxe sur la Valeur Ajoutée*. These laws are consequently essentially the same, because the EU VAT-directive basically dictates them.

6.1.1 VAT-Rates

The VAT-directive outlines that every Member State has to choose a reference VAT-rate against which goods and services should be taxed. The Netherlands' reference rate is 21% and Germany's 19%. In addition to the reference rates, every Member State has the right to incorporate a maximum of two reduced tax rates for certain goods and services transactions named in the directive. The Netherlands has one reduced rate of 6%, Belgium two of 6% and 12% and Denmark has none at all. Goods and services that may be taxed against a lower rate include (but not limited to): food, non-alcoholic beverages, public transport, books, cultural activities and painting houses by a professional painter (the paint itself, however, is taxed against the reference rate). To latter leads to a

remarkable result, which shall be illustrated by an example.

Y is a professional house painter, residing in the Netherlands. In order to perform his activities, he buys the paint from manufacturer X. The following transaction was made:

Transaction X → Y	Price (€)
Paint(excl. VAT)	200.00
Value Added Tax (21%)	42.00
Total	242.00

Just like before, Y pays the full amount to entrepreneur X and claims the VAT paid to X from the tax authorities. Customer Z hires Y to paint his house. The Netherlands taxes painting activities against a reduced rate of 6%. The following transaction was made:

Transaction Y → Z	Price (€)
Painting house of Z (excl. VAT)	500.00
Value Added Tax (6%)	30.00
Total	530.00

Note that the amount of VAT Y can claim from the tax authorities (€42) is actually higher than the VAT he actually has to pay for the service he preformed (€30).

This previous example leads to the formulation of the first VAT-paradox: **Companies expecting to be liable to VAT, might in fact not be taxed at all.**

6.1.2 VAT-exemptions

Some companies have been awarded a VAT-exempt due to their specific characteristics. These entrepreneurs do not have to charge VAT for their goods and services, but they are neither allowed to deduct the VAT paid in business-to-business transactions. The VAT-exemptions are mandatory, that is, the EU VAT-directive dictates that VAT-exemptions should be incorporated in national law. VAT-exempt goods and services include (but not limited to) education, rent, hospitals, financial services and freelance journalism.

X is a Dutch freelance journalist and hence exempted from VAT (that is, he does not charge VAT for his services). He wrote a story for newspaper Z. He can assign the following costs to writing the story for Z.

Administration X → Z	Costs (incl. VAT)	VAT
Costs (reference tax rate)	250.00	52.50
Costs (reduced tax rate)	50.00	8.50
Revenue	100.00	**=5.5% tax burden**

The tax burden is still lower than the reduced VAT-rate, but this example clearly illustrates that his service still has a VAT-component. Examples can be extended to situations where the tax-burden actually increases to over 21% (the Dutch reference rate), but these are not meaningful for the purpose of this introduction.

The previous example leads to the formulation of the second VAT-paradox: **goods and services from companies that are exempted from this tax, still have a VAT-component.**

Little research has been done on how big the VAT-component of exempted goods and services actually is. The Dutch governmental institution Statistics Netherlands (herinafter referred to as: CBS) simply assume the tax burden to be 0. Also, while the Bureau for Economic Policy and Analysis (herinafter referred to as: CPB) often discusses the VAT-exemption, their economic reviews usually assumes this burden is 0 too. This is an unfortunate assumption, because about 40% to 30% of household's spending is on exempted goods and services, i.e. a very important factor.

Inefficiencies in the Production Process

A second problem that arrises for VAT-exempted companies is that they might be encouraged to make inefficient decisions. In particular, they might take activities in-company because in-company transactions are not subject to VAT, while outsourced activities are. The latter is illustrated by the following example.

Consider a university that wishes to outsource its cleaning activities. If outsourced, the costs (excl. VAT) would be €1.000 a month. If they would opt to keep these activities in-company, it would cost €1.100. The most optimal solution would be to outsource these activities. However, due to the fact that €210 VAT is added, it ends up opting to keep the cleaning in-company. That is, the company ends up selecting a suboptimal solution (i.e. wasting public money).

VAT-exempted companies might also have a distinct advantage over their taxed counterparts.

Waste treatment performed by public bodies like municipalities are exempted from VAT, while their private counterparts are subject to VAT.

VAT-Cumulation

Taxed companies also use services of their exempted counterparts. As mentioned before, such services have a VAT-component. The taxed company is unable to claim a refund for

this VAT-component and it will hence, so it will continue to burden in the production process. This also results in the fact that it will be included in the consumer prices, i.e. VAT will be charged on the VAT-component. This phenomenon is called VAT-cumulation.

6.2 Measures of Inflation

There are many measures of inflation, but the two main inflation measures are the Gross Domestic Product deflator (GDP-deflator) and the Consumer Price Index (CPI).

6.2.1 GDP-deflator

The GDP-deflator is among the easier to understand concepts of inflation. The GDP is a measure for the total size of the economy in a country. It equals the sum of all final goods and services produced within a certain year, i.e.

$$Y = C + I + G + (X - M) \tag{6.1}$$

Where

$$\begin{aligned} Y &= \text{GDP-value} \\ C &= \text{Household final expenditue on final goods} \\ I &= \text{Investments on final goods} \\ G &= \text{Government expenditure on final goods} \\ X &= \text{Export} \\ M &= \text{Import} \end{aligned}$$

The performance of the economy of a country can be measured in both real and nominal terms. The nominal GDP at time t (denoted as Y_t^{nom}) measures the current sales times *current prices*, whereas the real GDP at time t measures (denoted as Y_t^R) the current sales times *prices of a certain index year* t_0. The latter is hence robust to inflation and hence generally considered a better measure for the performance of an economy. These concepts justify the first measure of inflation.

Definition 6.2.1. 1. The **GDP-deflator** at time t is defined as

$$\text{GDP deflator}_t = \frac{Y_t^{nom}}{Y_t^R} \tag{6.2}$$

2. The **GDP-deflator inflation rate** on an annual basis at time t (in months) is defined as

$$\pi_t^D = \frac{\text{GDP deflator}_t - \text{GDP deflator}_{t-12}}{\text{GDP deflator}_t} \tag{6.3}$$

6.2.2 Consumer Price Index

The consumer price index is a measure of inflation that merely relies on changes in consumer prices. For a certain index year t_0, a basket of household goods and services is composed and kept fixed for a long time period (i.e. each product group i is assigned a weight w_i, t_0).

Definition 6.2.2. The CPI index at time t is defined as

$$CPI_t = \frac{\sum_{i=1}^{n} p_{i,t} \times w_{i,t_0}}{\sum_{i=1}^{n} w_{i,t_0}} \tag{6.4}$$

For the purpose of this thesis, the CPI is the preffered way to measure the influence of VAT-increases on inflation over the GDP-deflator. Note that the GDP-deflator, unlike the CPI, captures many factors that are not subject to VAT-increases[1]. While not every product group in the CPI gets the same VAT-treatment[2], it is easy to distinguish different VAT-product groups. In fact, they can even be used as control groups.

Within the European Union, the Harmonised Index of Consumer Prices (HICP) is used. That is, all EU countries have the same basket of consumer prices, but the CPI-weights may differ across the different member states. The HICP hence makes the use of other EU member states as control groups much easier, given they all measure the CPI in compliance with the HICP.

6.3 Theoretical Analysis

In this section, a micro-economic model will be proposed in order to determine the tax incidence. Firstly, the theoretical analysis will focus on partial equilibrium models. These are relatively straight-forward models that are able to capture most important characteristics of tax incidence. Partial equilibirum models can loosely be described as economic models that only take a few economic variables into account and hence produce clear and easily interpretable results (Jhingan (2014)). At the other hand, partial equilibrium models ignore the effect a certain policy change may have on the rest of the economy. In theory, better results are obtained using general equilibrium models. Such models take much more economic variables into account and hence provide more accurate results. However, general equilibrium models often do not provide easy to interpret results.

> Consider a Value Added Tax of 6% is imposed on train rides from Amsterdam to Paris, while flights are exempt from this tax.[a] In practice, this would lead to a shift from train travel to flights. Partial equilibrium models, however, ignore this

[1] Like investments by companies, which are de facto not subject to VAT

[2] Cf. differences in VAT-rates, exemptions.

substitution effect while general equilibrium models do take this into account.

[a] In pratice, international flights are taxed against 0% (this can be interpreted as a VAT-exemption with the right of deduction of input-VAT), while VAT-rates on train travel range from 6% to 19% (the latter in Germany).

General equilibrium models will not be considered, since the partial equilibrium model proves to provide the information needed to continue the practical analysis.

6.3.1 Competative Economy

Competitive markets are characterised by the fact that the producer price q equals the consumer price p, i.e. $q = p$. We use a slightly reformulation of the model derived in Myles (1995, p. 236) that only takes VAT increases into account. Assume a value added tax of τ is imposed on a certain good or service, so the consumer price now equals

$$p(\tau) = q(\tau)(1+\tau) \tag{6.5}$$

In the equilibrium condition (demand equals supply),

$$D\left((1+\tau)\,q(\tau)\right) = S\left(q(\tau)\right),$$

The goal is to determine $\frac{dq}{d\tau}$. Using implicit derivatives

$$\begin{aligned}\frac{dD}{d\tau} &= \frac{dD}{dp}\frac{dp}{d\tau} \\ &= \frac{dD}{dp}\left[(1+\tau)\frac{dq}{d\tau} + q(\tau)\right] \\ \frac{dS}{d\tau} &= \frac{dS}{dq}\frac{dq}{d\tau}\end{aligned}$$

equating both derivates and solving them for $\frac{dq}{d\tau}$ results in

$$\frac{dq}{d\tau} = -\frac{\frac{dD}{dp}q(\tau)}{(1+\tau)\frac{dD}{dp} - \frac{dS}{dq}} \tag{6.6}$$

This equation can be rewritten in terms of elasticity of both demand and supply.

Definition 6.3.1. (Elasticities)

1. The **price elasticity of demand** is defined as

$$\varepsilon_D = -\frac{p}{D}\frac{dD}{dp} \tag{6.7}$$

2. The **price elasticity of supply** is defined as

$$\varepsilon_S = \frac{q}{S}\frac{dS}{dq} \tag{6.8}$$

Some little algebra results in

$$\frac{dq}{d\tau} = \frac{\frac{q(\tau)}{1+\tau}\varepsilon_D}{-\varepsilon_D + \varepsilon_S} \tag{6.9}$$

or, equivalently,

$$\frac{1}{q(\tau)}\frac{dq}{d\tau} = \frac{\frac{1}{1+\tau}\varepsilon_D}{-\varepsilon_D + \varepsilon_S} \tag{6.10}$$

It is also possible to derive $\frac{dp}{d\tau}$ in a similar fashion by noting that $q(\tau) = \frac{p(\tau)}{1+\tau}$.

$$\begin{aligned}\frac{1}{p(\tau)}\frac{dp}{d\tau} &= \frac{\frac{dS}{dq}\frac{1}{1+\tau}}{\frac{dD}{dp} - \frac{1}{1+\tau}\frac{dS}{dq}} \\ &= \frac{\frac{1}{1+\tau}\varepsilon_S}{-\varepsilon_D + \varepsilon_S}\end{aligned}$$

Some points of interest include

1. The tax incidence is never more than the actual tax (note $\varepsilon_D < 0$).
2. If the price elasticity of demand is highly elastic, then

$$\lim_{\varepsilon_D \to \infty} \frac{\frac{1}{1+\tau}\varepsilon_S}{-\varepsilon_D + \varepsilon_S} = 0 \tag{6.11}$$

   i.e. the VAT will not be passed through on the consumers.
3. If the price elasticity of demand tends to be inelastic, then

$$\lim_{\varepsilon_D \downarrow 0} \frac{\frac{1}{1+\tau}\varepsilon_S}{-\varepsilon_D + \varepsilon_S} = \frac{1}{1+\tau} \tag{6.12}$$

   i.e. the tax is fully passed through.[3]

[3] This is easy to (albeit not necessary) see when $\tau \to 0$

6.3.2 Monopolistic Economy

This paragraph extends the tax incidence model for goods and services characterised by a monopoly developed by Fang (2004).

Let π denote the profit of a monopolistic firm. Then the monopolist maximises

$$\max_{p\in\mathbb{R}^+} \underbrace{\left\{ \frac{p}{1+\tau} D\left(p\right) - C\left(D\left(p\right)\right) \right\}}_{\pi} \tag{6.13}$$

In order to obtain the maximum, derivate π w.r.t. p, i.e.

$$\begin{aligned}
\frac{d\pi}{dp} &= \frac{1}{1+\tau}\left(D\left(p\right) + p\frac{dD}{dp}\right) - \frac{dC}{dD}\frac{dD}{dp} \\
&= \frac{p}{1+\tau}\left(\frac{1}{p}D\left(p\right) + \frac{dD}{dp}\right) - \frac{dC}{dD}\frac{dD}{dp} = 0 \\
\iff \frac{p}{1+\tau} &= \frac{\frac{dC}{dD}\frac{dD}{dp}}{\frac{1}{p}D\left(p\right) + \frac{dD}{dp}} \\
&= \frac{\frac{dC}{dp}}{-\frac{1}{\varepsilon_D} + 1}
\end{aligned}$$

To derive a useful expression for $\frac{dp}{d\tau}$, some extra assumptions must be imposed.

1. $C\left(D\left(p\right)\right) = cD\left(p\right) + a$ is a linear function of demand s.t. $\frac{dC}{dD} = c$
2. $D\left(p\right) = b - p\left(\tau\right)$

In that case,

$$\begin{aligned}
\frac{p\left(\tau\right)}{1+\tau} &= \frac{c}{2 - \frac{b}{p(\tau)}} \\
&= \frac{cp}{2p\left(\tau\right) - b} \\
\iff \frac{1}{1+\tau} &= \frac{c}{2p\left(\tau\right) - b} \\
\iff p\left(\tau\right) &= \frac{1}{2}\left[c\left(1+\tau\right) + b\right]
\end{aligned}$$

and so,

$$\frac{1}{p\left(\tau\right)}\frac{dp}{d\tau} = \frac{c}{c\left(1+\tau\right) + b}. \tag{6.14}$$

It is important to note that $\frac{1}{p(\tau)}\frac{dp}{d\tau} < 1$, i.e. in this framework a monopolist will never increase its prices by more than the tax imposed.

Full VAT-incidence analysis

In this section, an analysis will be made of the influence on the CPI if the VAT is fully passed through on consumers. Recall that, in the general model, $p(\tau) = (1+\tau)\, q(\tau)$. However, due to the fact that VAT is fully passed through, $q(\tau)$ is invariant under τ. So $dp = q \cdot d\tau$. The relative price change will then be given by

$$\frac{dp}{p(\tau)} = \frac{d\tau}{(1+\tau)} \tag{6.15}$$

The most important property of equation (6.15) is that an increase of the VAT by $\Delta\tau$ does not result in consumer prices to increase by $\Delta\tau$ (provided $\tau \neq 0$ initially).

Chapter 7

Causal Effects of the VAT increases

This chapter is the finale of this thesis. Before applying the statistical methods derived in this thesis on the causal effects of VAT increases on inflation, we will first consider two more theoretical sections. First, earlier statistical research on the causal effects of VAT changes on customer prices is discussed. Particular emphasis is placed on the paper of Vrijburg et al. (2014), which also investigates the causal effects of the 2001 and 2012 VAT increases on inflation. The resulting insights are used to develop a statistical model that improves existing estimates for the causal effects of the pass through of VAT increases, which are then applied on the causal effects of the 2001 and 2012 VAT increases.

7.1 Earlier research

A substantial amount of research has been done on the question whether VAT increases are fully passed through to consumer prices. This thesis focusses on the question whether the 2001 and 2012 VAT increases were fully passed through, so we will thoroughly consider the Dutch Bureau of Economic Policy Analysis (CPB) conducted by Vrijburg et al. (2014) since they also consider the 2001 and 2012 VAT increases. However, before turning to their analysis, we shortly summarise some of the results derived in other countries.

7.1.1 Evidence from other countries

The Deutsche Bundesbank (2008) considers the 2007 VAT-increase in Germany. In that year, the standard rate was increased by 3 percentage points from 16% to 19%. The Deutsche Bundesbank concludes that the VAT was largely passed through to consumer prices, but almost never by the full amount. In fact, some suppliers did not even make a change at all. Given the theoretical framework developed in this thesis, this is an indication that these are suppliers of commodities with a highly elastic demand function. Deutsche Bundesbank (2008) conducted the research on a micro level (i.e. per

commodity group), while we consider the VAT increases on a macro level (i.e. all commodity groups combined). Consequently, caution must be exercised while comparing these results.

Chirakijja, Crossley, and Lührmann (2009) consider the temporary United Kingdom 2008 VAT cut. They conclude that this VAT cut was only passed through for 75% in consumer prices, suggesting that the remainder was used to increase profit levels. While this analysis is on a related note, we must be aware that the passing through behaviour of VAT cuts may be different from VAT increases. This conjencture is especially justified by the fact that only labour intensive commodity groups were subject to the temporary VAT cut, while the 2001 and 2012 VAT increases affected a much broader range of commodity groups. Jongen, Lejour, and Massenz (2017) and Harju and Kosonen (2014) suggest that the VAT cuts for hairdressers and the restaurant industry were hardly passed through on consumer prices, further justifying this conjencture.

7.1.2 VAT increases 2001 and 2012

In this paragraph we will extensively consider the research done by Vrijburg et al. (2014) from both a legal and statistical perspective. Shortly summarised, they conclude that the hypothesis that the VAT increase was fully passed through on consumer prices cannot be rejected. Some mistakes were however made in their analysis, comprising the reliability of their findings.

Methodology

Vrijburg et al. (2014) consider a causal effect model that does not include dummies for the treatment period and control group specific effects. We have seen that the first difference estimator (and by extension also fixed effects estimators) consistently estimate a difference-in-differences model without including the control group specific effect. An alternative formulation of the fixed effects model is also able to consistently estimate the causal effect without the inclusion of treatment time period specific effects. We have not discussed this alternative formulation since it is not needed for this thesis. An indicator variable that controls for the treatment time period specific effect proves to suffice. They in particular consider the model

$$y_{i,s,t} = \mathbf{X}_{i,s}\beta + \mathbf{W}_{i,s}\gamma + \varepsilon_{i,s} \tag{7.1}$$

where $y_{i,s,t}$ denotes the inflation rate for commodity i at time t for country s, where $s = \{\text{Belgium, the Netherlands}\}$. Obviously, Belgium is the control group in the model. The matrix $\mathbf{X}_{i,s}$ is a vector of several treatment dummies, to capture the effect of the VAT-increase on consumer prices at several points in time. The most important treatment dummy is the one that equals 1 at the actual time of treatment - we will ignore the other two for the moment.[1]

[1]Vrijburg et al. (2014) also included a dummy for pre and post treatment, but these were almost never significantly different from 0.

Vrijburg et al. (2014) compare the fixed effects and the so-called 'Common Correlated Effects' (CCE) estimator for the treatment effects. The latter has not been discussed in this thesis. It is enough to remember that the CCE estimator only has a meaningful interpretation in presence of time specific unobserved fixed effects. Vrijburg et al. (2014) conclude that the fixed effects estimator does not produce significant treatment effects (which is likely an indication for a type II error). The CCE estimator produces a significant treatment effect, but it does not reject the hypothesis that the VAT was fully passed through on consumer prices.

Legal perspective

The list of included goods and services as high taxed gives reason to doubt the reliability of the study. Firstly, they classify CP0442 refuse collection as a high taxed commodity. In reality, refuse collection for households is usually exempted from VAT due to the fact that this is an exclusive power of the public sector (in particular municipalities). The European Court of Justice (supreme court of the EU) ruled in the Carpaneto Piacentino case that public bodies are exempted from VAT.[2] Given the fact that exempted goods and services still bear a VAT component, we may reasonably expect a treatment effect. At the other hand, the objective of the research of Vrijburg et al. (2014) was to capture the pass through of the 2001 and 2012 VAT increases on goods that are taxed against the standard rate. Inclusion of exempted goods and services is consequently unwanted noise and possibly resulting in lower causal treatment effect than they actually are.

Furthermore, they have included hairdressing salons and personal grooming establishments (CP1211) as a high taxed service. However, this is a mistake as these are mostly taxed against a reduced rate as per Table I item b.7 pertaining to the Dutch VAT act. Furthermore, they omitted alcohol and tabacco from the analysis. These goods are subject to additional excise taxes ('accijnzen'), but in the period of research, the excise tax on alcohol was never increased more than the actual inflation rate. At the other hand, the exclusion of tabacco is justified given the fact that during the treatment period, the increase in excise duty was higher than inflation, potentially causing upward biased estimates. While not a legal comment *per se*, it must be noted that they failed to include CP082_083 Telephone and Telefax Equipment and services as a high taxed service. Their paper does not mention why it has been excluded, but in my model it proved to be an outlier causing the estimates of the of the pooled OLS, fixed effects and random effects estimator to be much higher.

Statistical perspective

Vrijburg et al. (2014) note in a footnote that the presence serial correlation causes the non-robust standard errors to be invalid. Following Bertrand et al. (2004), they control for this using serial correlation robust standard errors. As noted before, this is a sufficient control for type I errors, but insufficient for type II errors. According to the

[2]ECJ 17 October 1989 (Carpaneto Piacentino), case 231/87

authors, the fixed effects estimator produces an insignificant treatment effect while the CCE does produce a significant treatment effect. Whether the CCE estimator really generates significant estimates is debatable, since the treatment effect captured by the CCE estimator is often only significant at a 10% level. Most academic literature consider 5% as upper bound. This upper bound is also adhered in this thesis. This in particular means that the CCE estimator in the paper of Vrijburg et al. (2014) did not actually produce a significant treatment effect.

Furthermore, they claim that the fixed effects estimator produces insignificant treatment effects due to the fact that it does not properly take the dependence between the unobserved individual fixed effect and the unobserved time specific effect into account, while the CCE estimator does. Firstly, they do not consider the argument that fixed effect is inefficient due to the presence serial correlation. They argue that serial correlation robust standard errors are sufficient control for serial correlation, but these only sufficiently control for type I but not for type II errors. This could also be observed from the DD simulation in this thesis. Moreover, it is not entirely trivial that there actually exists an unobserved time specific effect (other than the treatment time period specific dummy variable) that causes endogeneity. They do not further eleborate on that specific fact.

7.2 Methodology and data

The general statistical model is written as

$$\begin{aligned} y_{i,t,s} = \mathit{timed}\mathbf{1}[t_0 \leq t \leq t_1] + \mathit{countrd}\mathbf{1}[s = NL] \\ + \mathit{treat}\mathbf{1}[t_0 \leq t \leq t_1]\mathbf{1}[s = NL] \\ + infl_t + gdp_{s,t} + \varepsilon_{i,s,t} \end{aligned} \tag{7.2}$$

where $i = 1, \ldots, N$ and $t = 1, \ldots, \ t_0, \ldots, \ t_1, \ldots, T$ denote the amount of commodity groups and the amount of time periods, respectively. Let $s \in \{\text{Control group}, \text{Netherlands}\}$ denote the different countries. $y_{i,t,s}$ represents the inflation rate for commodity i at time t for country s. The parameters timed, $\mathit{countrd}$ and treat are analogously interpretable as their counterparts in the general DD model. The variable $infl_t$ controls for the euroarea wide inflation trend and $gdp_{s,t}$ controls for differences in economic growth rates between the countries. Benedek et al. (2015) argue that these variables provide sufficient control for potential endogneity issues. Following Bertrand et al. (2004), the standard errors accompanying the estimates are robustified for serial correlation and heteroscedasticity. We regress the model using pooled OLS, random and fixed effects, first difference and the feasible GLS estimator restricted to $ARMA\,(1,1)$ correlation. These estimators will be compared to see their performance.

Both Germany and Belgium are seperately used as control groups to infer the causal treatment effect parameter (treat). Using two seperate control groups allows for better comparison of the results, meaning that divergence between causal treatment results is an indication of potential endogeneity. This also gives more flexibility in deriving the

causal treatment effect, since we can exploit dependence structures in the data much better. This in particular proves to be useful when it comes to serial dependence. If such flexibility is used, it will always be argued why this is appropriate.

The data sources that are used for the causal inference are given in table 7.1. Shortly summarised, we use the EU Harmonised Index of Consumer Prices (HICP). This is an EU standardised method for measuring CPI inflation. This in particular suggests that the data used in this thesis is comparable. Eurostat also provides commodity weights at a country level, which can be interpreted as consumption weights. Weights are included the same manner as Weighted Least Squares, but its interpretation is different. If weights are included in the regression, the estimate for the *treat* parameter represents the weighted mean effect of the VAT increase on consumer prices. They do not necessarily reduce heteroscedasticity.

Most literature only document weighted estimates, which can be defended from the fact that unweighted estimates are affected by the level of subcategorisation in the HICP framework. To see this point, note that CP0451 Electricity does not have any subcatogory, while CP31 Clothing materials is further subcatagorised up to CP314 Cleaning, repair and hire of clothing. That is, the pass through of the VAT-increase of the former has a much bigger impact for consumers than the latter. Unweighted results do not capture this. Ideally though, both unweighted estimates and weighted estimates should not differ significantly. This thesis reports both weighted and unweighted estimates, although the final conclusions are based on the weighted results.

For each regression three central hypotheses are tested

1. Did the estimator produce a significant treatment effect?
2. Is autocorrelation present in the error term?
3. Is the estimator significantly different from the theoretical maximum pass through rate?

The first hypothesis translates to testing

$$H_0 : treat = 0$$
$$H_1 : treat > 0$$

in particular suggesting that we do not consider negative values.

The second hypothesis is tested with an $AR\,(1)$ fit of the residuals. This is a relatively easy procedure and usually sufficiently captures inefficiencies caused by autocorrelation. This in particular means we consider the regression

$$\hat{\varepsilon}_{i,t} = \rho\hat{\varepsilon}_{i,t-1} + \eta_{i,t} \tag{7.3}$$

where $\eta_{i,t}$ is some white noise term and $\hat{\varepsilon}_{i,t}$ the residuals produced by the appropriate estimator. The corresponding hypothesis test is then statistically equivalent to testing

$$H_0 : \rho = 0$$
$$H_0 : \rho \neq 0$$

Rejection of H_0 does not mean that the residuals actually follow an $AR(1)$ process. It is merely an easy test to capture inefficiencies caused by autocorrelation.

The third test is important for answering the research question. The theoretical bound is calculated by considering

$$\text{Percentual change} = \frac{p_{i,t_0} - p_{i,t_0-1}}{p_{i,t_0-1}} \tag{7.4}$$

if the VAT is fully passed through, then $p_{i,t_0} = (1 + \tau + \Delta\tau)\, q_{i,t_0-1}$ and $p_{i,t_0-1} = (1 + \tau)\, q_{i,t_0-1}$. And so equation (7.4) equals

$$\text{Percentual change} = \frac{\Delta\tau}{1 + \tau} \tag{7.5}$$

This means that the last hypothesis statistically turns into testing (cf. equation (6.15))

$$H_0 : \text{treat} \geq \frac{\Delta\tau}{1 + \tau}$$
$$H_1 : \text{treat} < \frac{\Delta\tau}{1 + \tau}$$

Note beforehand: Eurostat provides weights for each product group category. The `plm` package offers methods to implement these weights in the regression. I strongly recommend against the use of the built-in weighting function, as the standard errors reported with weights included are not invariant under multiplying them with a constant. Weighting should be done manually in the same way as the Weighted Least Squares procedure is implemented.

Table 7.1: Data sources

Parameters(s)	Data Source	Data Description
timed, countrd	Eurostat	Harmonised Index of Consumer Prices (HICP)
treat, infl		- monthly data (annual rate of change)
gdp	OECD	Quarterly National Accounts - GDP Growth

7.3 VAT Increase 2001

On January 1st 2001, VAT rate from 17.5% to 19%. The central hypothesis in this chapter is the question whether or not the VAT was fully passed through into consumer

prices. Given the theory developed in chapter 6, we can already derive a theoretical upper bound for the tax incidence. In particular, the maximum VAT pass through onto consumer prices equals

$$\frac{0.015}{1.175} = 0.01277 \tag{7.6}$$

or 1.277%. This means we expect

$$treat \in [0, 1.277] \tag{7.7}$$

Treatment starts at time period t_0 = 01-01-2001 and ends at t_1 = 31-01-2001. The panel consists of $T = 36$ time units, ranging from 01-01-1999 to 31-12-2001. $N = 39$ commodities are included in the regression.

Before beginning the regression analysis, we first discuss some 2001 specific policy changes that may cause endogeneity in estimating the treatment parameter.

7.3.1 Policy analysis

Treatment started on the first of january 2001, meaning that it coincided with the start of the new fiscal year. Apart from the change in VAT rate, some goods and services are subject to additional taxes. These taxes are commonly called 'excise taxes'. Usually, excise taxes are increased by the expected inflation every year. However, according to the *belastingplan 2001*, the following goods were subject to an increase of excise taxes that was higher than the expected inflation:

- Tabacco
- Fuel

We do not have proper instruments to control for the extra increase in excise taxes. These goods have been removed from the analysis, since including them would cause endogeneity in the treatment parameter. In particular, the treatment parameter would have had an upward bias.

7.3.2 Estimates

Pooled OLS

Table A.1 ('A' refers to appendix A) shows the estimates and $AR(1)$ fits for the unweighted and weighted pooled OLS estimator. There is no point deeply discussing the estimates, as the large standard errors in combination with the $AR(1)$ correlation coefficients suggest that the estimates are inefficient.

Belgium Only the unweighted pooled OLS estimator produced significant estimates for the treatment effect, i.e. the weighted treatment effect is not significantly different

from 0. The weighted estimator is considerably although not significantly different from the unweighted estimate. The unweighted estimate suggests a treatment effect that is higher than the theoretical maximum bound, but note that

$$CI_{95\%}\left(treat_{\text{Unw}}\right) = [0.683, 3.804] \tag{7.8}$$

which suggests that the estimate is certainly not significantly different from a wide range of values that lie within the theoretical bound.

On the contrary, the weighted pooled OLS estimator obviously produces a type II error. This estimator should consequently be rejected as a reliable estimator for the weighted model.

Turning to the question whether the pooled OLS is efficient, we test for

$$\begin{aligned} H_0 &: \ \sigma_\mu^2 = 0 \text{ and no serial correlation in the error term} \\ H_1 &: \ \text{Assumption A.POLS.3 fails} \end{aligned}$$

using the test formulated in paragraph 5.2.1. This was originally meant to only test for the presence of unobserved individual specific fixed effects, but this test unwillingly also checks for the presence of serial correlation. Both the unweighted and weighted estimates reject H_0. This also means that the null hypothesis that pooled OLS is efficient must be rejected. The $AR\,(1)$ fit of the residuals confirms this - the $AR\,(1)$ fits suggest autocorrelation coefficients around 0.9 for both the weighted and unweighted model, i.e. the efficiency of this estimator is highly affected by serial correlation.

Figure 7.1: ACF plot of weighted pooled OLS residuals (Belgium)

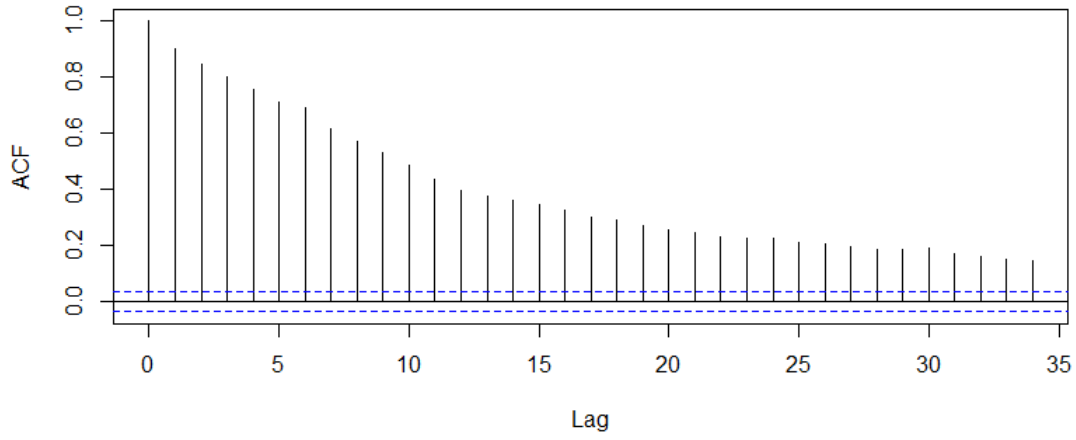
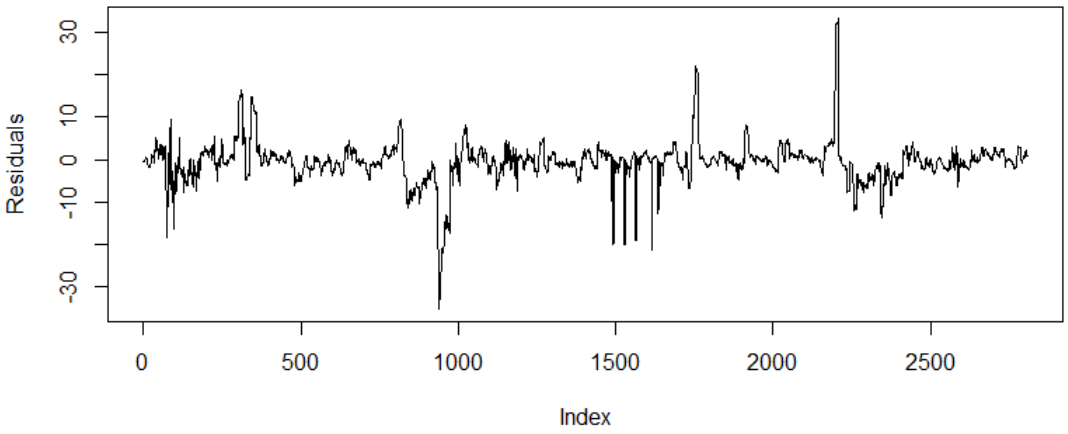


Figure 7.2: Plot of residuals



Diagnostic plots are usually very difficult to give for panel data analysis because of the multiple cross-sectional units. Plot 7.1 is the autocorrelation function (ACF) of the *first* cross sectional unit. While the plots are not at all representative for the full data set, they do provide useful insight on how the transformation changes the behaviour of serial correlation. For more deceicive results for the full data set, we would need to rely on the statistics reported in the tables. Figure 7.2 provides a plot for the residuals, that also shows strong serial dependence. The main results of pooled OLS regression are presented in table 7.2.

Table 7.2: Pooled OLS summary (Belgium)

Null hypothesis	Unweighted	Weighted
Insignificant treatment effect	Rejected (**)	Not rejected
Residuals AR(1) serially uncorrelated	Rejected (***)	Rejected (***)
Full pass through of VAT	Not rejected	Not rejected

[1] *p<0.05; **p<0.01; ***p<0.001

Germany Both the unweighted and weighted pooled OLS estimators produce significant treatment effects, yet they also do not reject the hypothesis that the VAT was fully passed through to consumer prices. While the pooled OLS estimtor does not produce a type II error when Germany is used as the control group, the large standard errors still suggest that this estimator is inefficient. Note that the unweighted estimator for the treatment effect suggests

$$CI_{95\%}\left(treat_{\text{Unw}}\right) = [1.488, 3.167] \tag{7.9}$$

and so rejects all values that are within the theoretical boundary. This *may* be an indication that the unweighted pooled OLS estimator is producing a type I error. While

the standard errors are robustified for serial correlation, Angrist and Pischke (2009) argues that robust standard errors can still underestimate actual standard errors. This potentially results in higher type I errors compared to efficient estimators. On the contrary, the weighted pooled OLS estimator does not reject a wide range of values that lie within the theoretical boundaries.

Figure 7.3: ACF plot of weighted pooled OLS residuals (Germany)

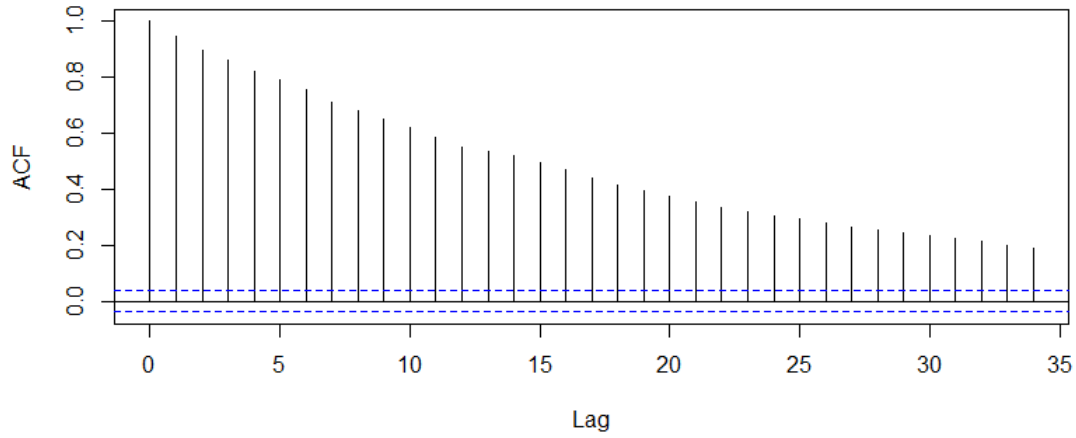
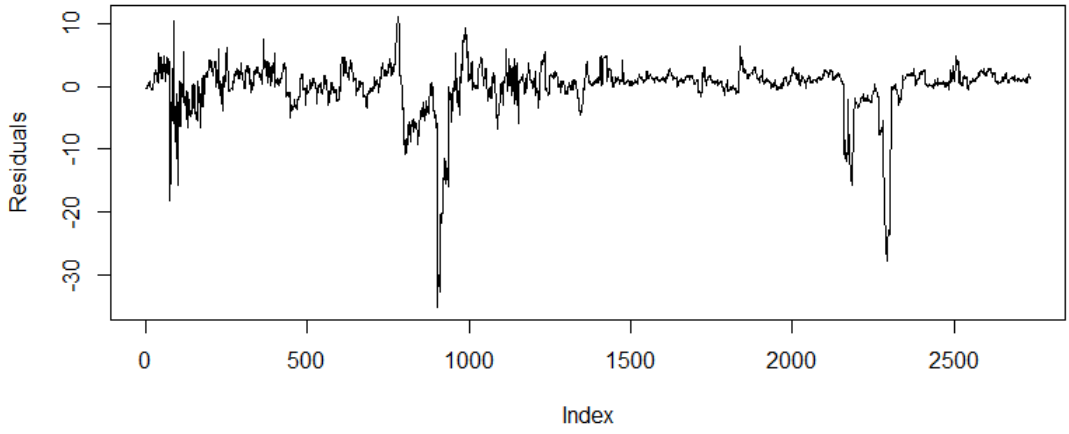


Figure 7.4: Plot of residuals



When Germany is used as the control group, the pooled OLS estimator seems to be inefficient. An $AR(1)$ fit of the residuals shows serial dependence that is close to a random walk. Both diagnostic plots suggest do not suggest differently, they both show very strong serial dependence. However, we do not yet exclude other types of serial dependence. The main results of the regression are summarised in table 7.3.

Table 7.3: Pooled OLS summary (Germany)

Null hypothesis	Unweighted	Weighted
Insignificant treatment effect	Rejected (***)	Rejected (**)
AR(1) serially uncorrelated	Rejected (***)	Rejected (***)
Full pass through of VAT	Not rejected (Type I?)	Not rejected

Random Effects and Fixed Effects

We will discuss the random effects and fixed effects together, as T is reasonably large and so the random effects estimates should approach the fixed effects estimates. The results of the random effects estimator are presented in table A.2 and the fixed effects in table A.3.

Belgium The analysis for the random and fixed effects estimator is roughly the same as the one made for the pooled OLS estimator. Only the unweighted pooled OLS estimator produces a significant treatment effect and it does not reject a large range of values within the theoretical bounds. While the point estimates are roughly on par with the pooled OLS estimator (certainly not significantly different), the AR(1) fit of the residuals reveals some odd behaviour in the serial correlation. The estimates are considerably (and significantly) lower than the estimates produced by the pooled OLS estimator. While this apparently did not lead to inconsistent estimates, it is an indication that there actually exists an unobserved individual specific effect.

The fixed effects estimates are almost equal to the random effects estimator. We are unable to make useful inference on the question whether the random effects estimator is preferred over the fixed effects estimator. Paragraph 5.2.2 discussed the Hausmann test, but it was argued that this test is only reliable provided assumptions A.RE.1 - A.RE.5 hold. Given the significant estimates for the first order autoregressive serial correlation in the residuals, this hypothesis is rejected. That is, assumption A.RE.5 is rejected. In this particular case this question is mostly theoretic anyway, since both the random and fixed effects estimator lead to the same inference and T is sufficiently large.

Figure 7.5: ACF plot of weighted fixed effects residuals

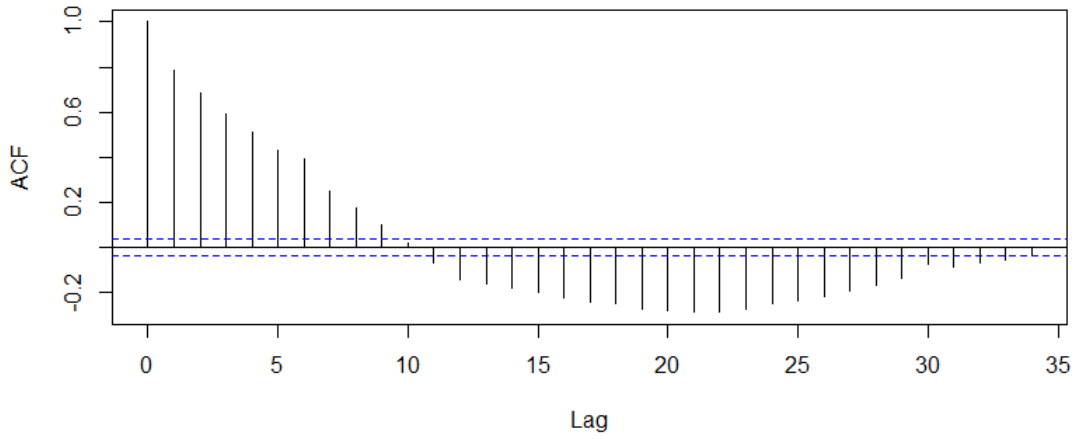
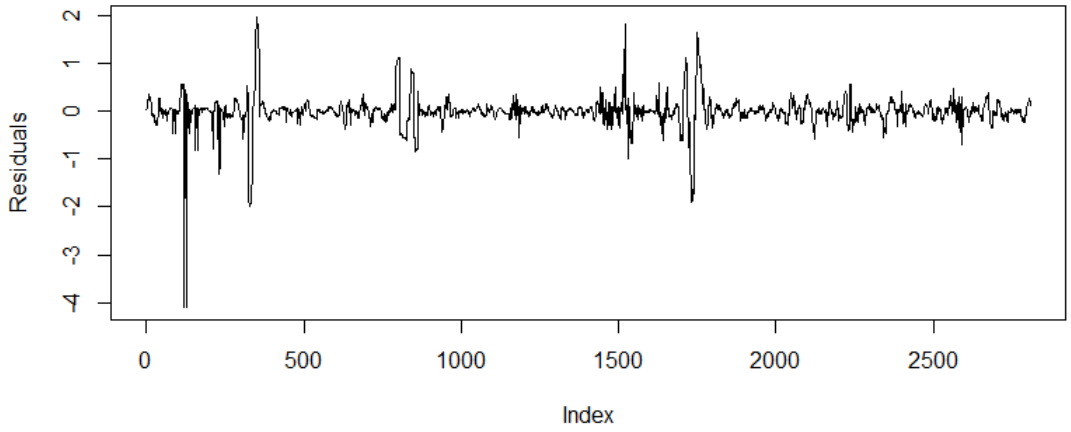


Figure 7.6: Plot of residuals



The diagnostic plots for the weighted fixed effects estimator have been included. The ACF plot indeed suggests somewhat lower serial dependence. It is more difficult to note the reduction in serial dependence from the residual plot. The results from regression are summarised in table 7.4. Note that the random effects and fixed effects estimators lead to exactely the same inference.

Table 7.4: FE and RE summary (Belgium)

Null hypothesis	Unweighted	Weighted
Insignificant treatment effect	Rejected (**)	Not rejected
Full pass through of VAT	Not rejected	Not rejected
AR(1) serially uncorrelated	Rejected (***)	Rejected (***)

Germany The same analysis approximately applies when Germany is used as a control group. It should be noted that the point estimates for the correlation coefficients are significantly different from their Belgian counterparts. For the weighted model we can even infer that the $AR(1)$ correlation of the residuals is significantly higher than the fixed effect residuals produced when Belgium is used as the control group.

Figure 7.7: ACF plot of weighted fixed effects residuals

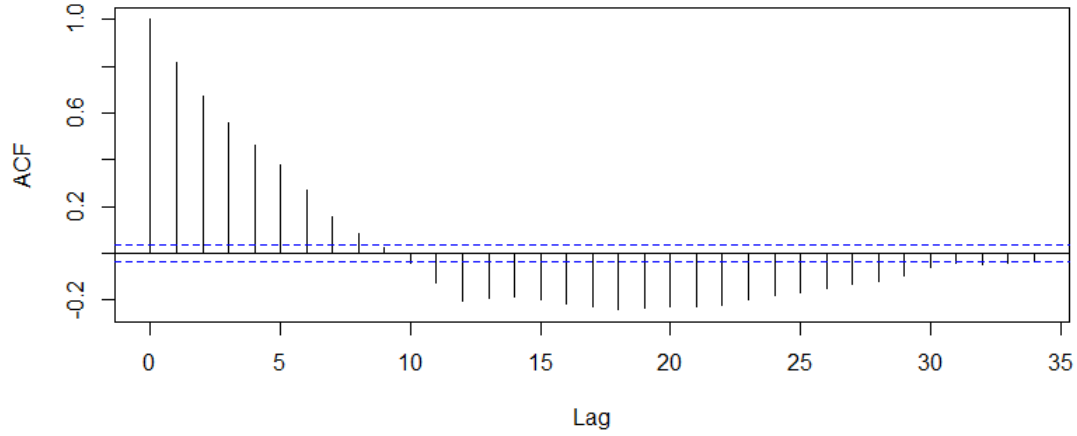
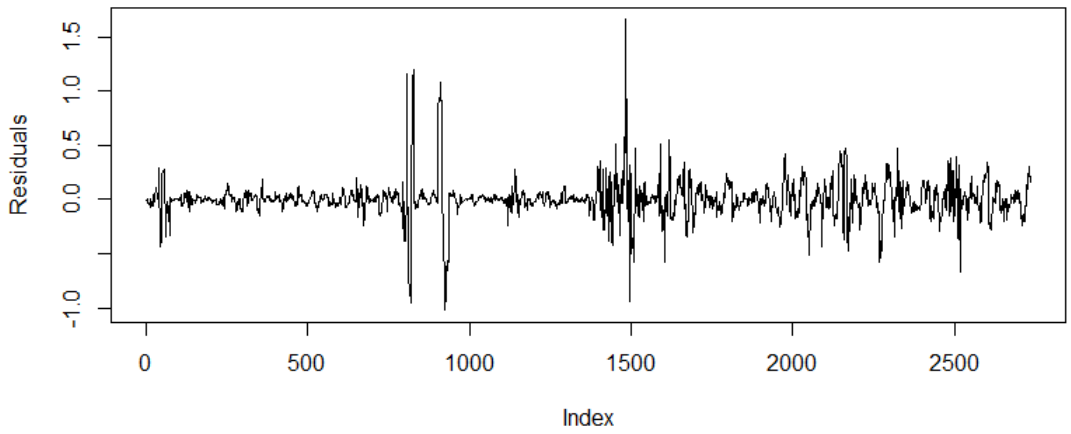


Figure 7.8: Plot of weighted fixed effects residuals



The plots again only show the results for the fixed effects estimator. The ACF plot also shows a slight reduction in serial correlation, although serial dependence is still high. This is also confirmed by a plot of the residuals. The main results from regression are presented in table 7.5

Table 7.5: FE and RE summary (Germany)

Null hypothesis	Unweighted	Weighted
Insignificant treatment effect	Rejected (***)	Rejected (***)
AR(1) serially uncorrelated	Rejected (***)	Rejected (***)
Full pass through of VAT	Not rejected (Type I?)	Not rejected

First Difference

Given the analysis made above, the first difference estimator is expected to produce much more reliable estimates than the other estimators discussed before. From a theoretical perspective this estimator should perform better provided the residuals follow a random walk, although the simulations also revealed that the FD estimator may outperform the other estimators in presence of other types of strong serial dependence in the error term. The results are presented in table A.4 and indeed suggest that all treatment effects are significant, i.e. the FD estimator does not seem to make type II errors while producing reasonably low standard errors.

Belgium The first difference estimator seems to increase the accuracy of the estimates considerably. Both the weighted and unweighted point estimates are within the theoretical bound, but they are not significantly different from values that are outside this bound. The unweighted first difference estimator has significant serial correlation in the residuals, while the weighted estimate is not significantly different from 0. The standard errors for the weighted and unweighted estimates are, however, about the same. Note that the standard errors are again robustified for arbitrary serial correlation and heteroscedasticity. From a theoretical perspective, the first difference estimator so far seems to outperform any other estimator discussed in this paragraph.

Figure 7.9: ACF plot of weighted first difference residuals (Belgium)

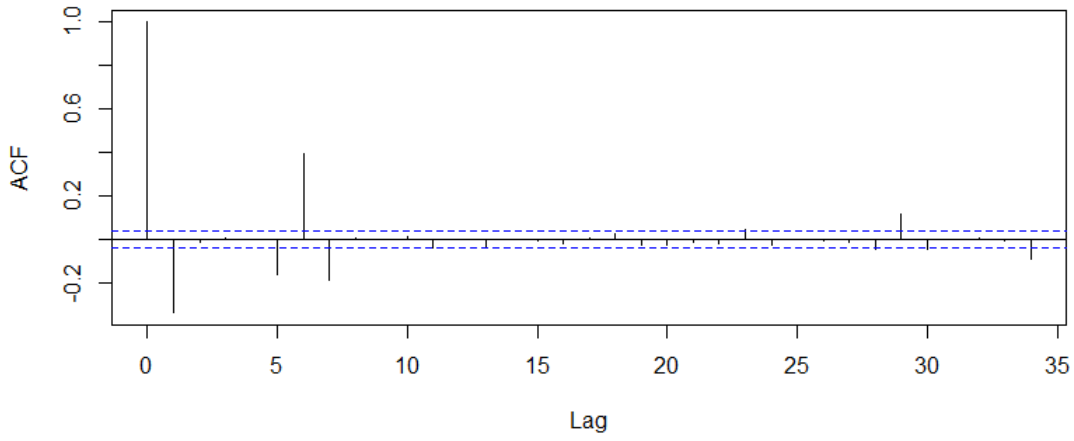
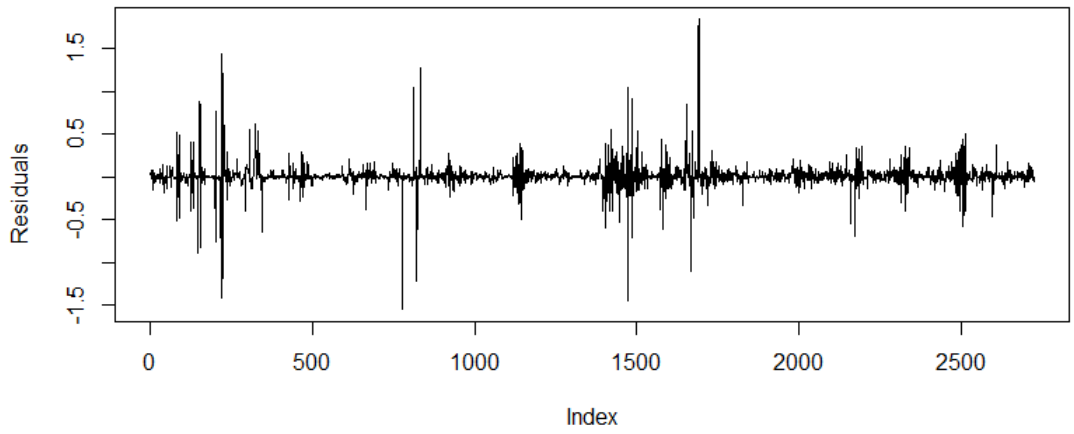


Figure 7.10: Plot of residuals



The ACF plot clearly shows a big change in serial dependence, i.e. the FD transformation considerably reduced serial dependence. The plots of the residuals also show more white noise behaviour, although it must be noted that this does not mean that the residuals actually follow a white noise process. The results from regression are sumarised in table 7.6.

Table 7.6: First difference summary (Belgium)

Null hypothesis	Unweighted	Weighted
Insignificant treatment effect	Rejected (***)	Rejected (***)
AR(1) serially uncorrelated	Rejected (***)	Not rejected
Full pass through of VAT	Not rejected	Not rejected

Germany The use of the first difference estimator seems to be fully justified when Germany is used as the control group. The $AR(1)$ correlation estimate of the residuals is not significantly different from 0, which is again an indication that the original errors follow a random walk. Also when Germany is used as the control group, the point estimates are never significantly different from full pass through. In fact, the weighted point estimate is slightly above the theoretical bound although it does not reject a wide range of values that are within the theoretical bound.

Figure 7.11: ACF plot of weighted first difference residuals (Germany)

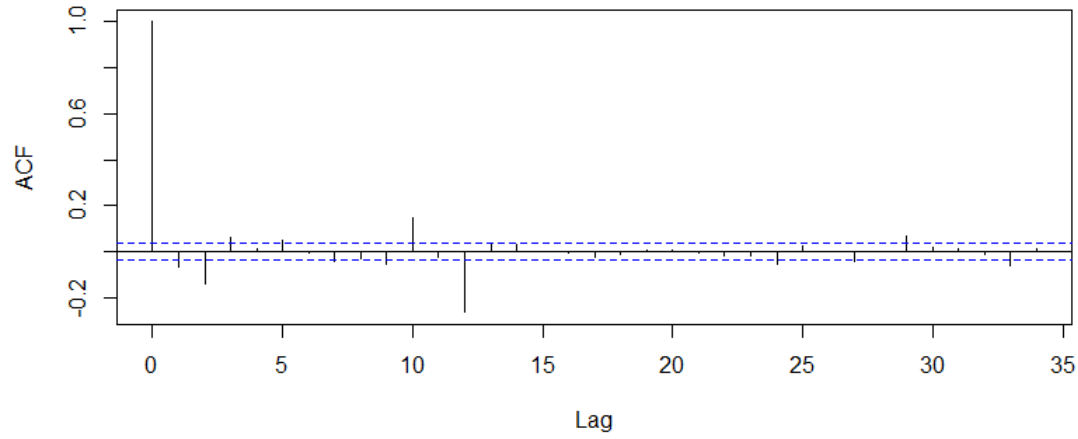
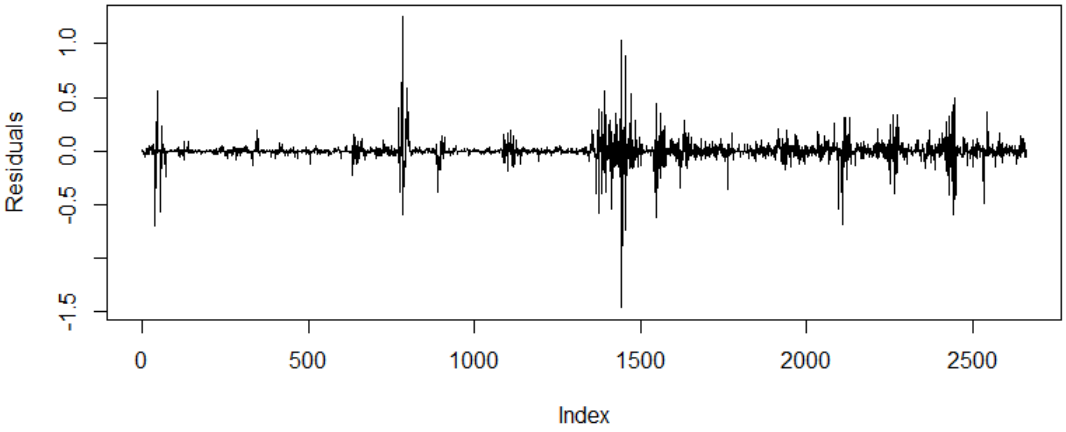


Figure 7.12: Plot of residuals



The ACF plot indeed suggests that the first difference transformation drastically reduced serial dependence. In fact, the result for the first order lag is lower than when Belgium is used as the control group. This can be explained from the fact that when Germany is used as the control group, the AR(1) correlation was not significantly different from 0. The main results from regression are presented in table 7.7.

Table 7.7: FD Summary (Germany)

Null hypothesis	Unweighted	Weighted
Insignificant treatment effect	Rejected (***)	Rejected (**)
AR(1) serially uncorrelated	Not rejected	Not rejected
Full pass through of VAT	Not rejected	Not rejected

Feasible Generalised Least Squares

This paragraph discusses the feasible GLS estimates for this model. We will only implement a restricted form of this estimator that captures serial dependence. The standard errors are robustified for the remainder heteroscedasticity. We consider an error term that follows an $ARMA\,(1,1)$ process, i.e.

$$\varepsilon_{i,t} = \rho\varepsilon_{i,t-1} + \theta\eta_{i,t-1} + \eta_{i,t} \tag{7.10}$$

where $\eta_{i,t}$ is some white noise process. Higher order ARMA processes are an option, but in this particular case that would have lead to an overfit. The estimates are given in A.5. Modelling the autocorrelation structure with an $ARMA\,(1,1)$ process does *not* increase the accuracy of the estimates in comparison with the first difference estimator.

Belgium The point estimates of the generalised least squares estimator are almost identical to the estimates for the first difference estimator. However, the standard errors produced by the GLS estimator are somewhat higher than the first difference estimates. This is a remarkable result, given the fact that the ACF plots reveal that the first order autocorrelation for the first difference estimator is considerably larger than the Generalised Least Squares estimator. This could be an indication that the robust standard errors for the first difference underestimate the actual standard errors. At the other hand, we should not worry about this too much since the differences are marginal. In particular, both the GLS and first difference estimator lead to the same inference. On the contrary, the generalised least squares produces significant results for most control variables. This is often not the case for the first difference estimator.

Figure 7.13: ACF plot of weighted first difference residuals (Belgium)

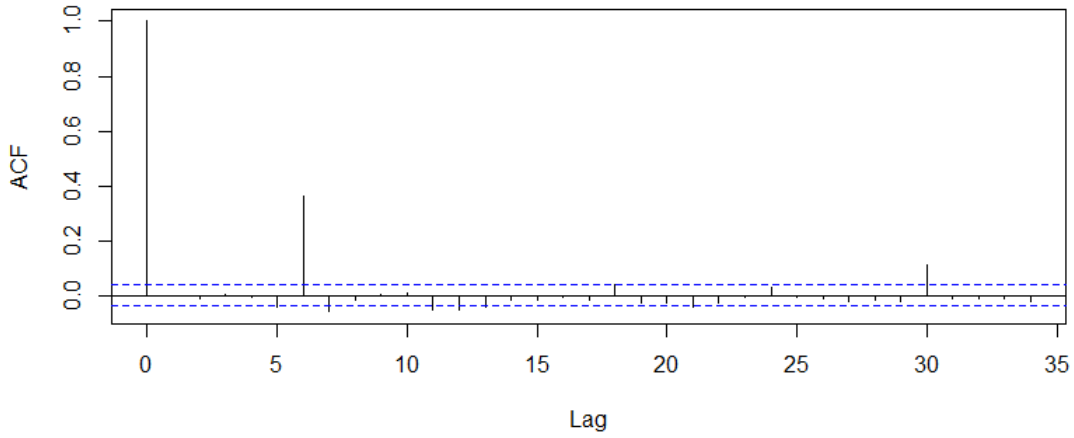
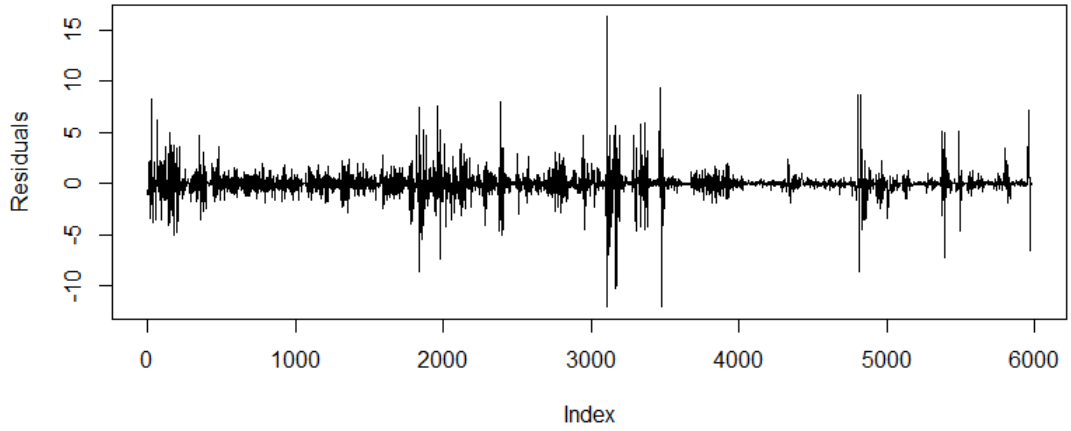


Figure 7.14: Plot of residuals



The ACF plot shows a big decrease in serial correlation. The white noise behaviour of the restricted feasible GLS estimator is also clearer from the plot of the residuals. Table 7.8 shows the main results of the regression.

Table 7.8: Feasible GLS summary (Belgium)

Null hypothesis	Unweighted	Weighted
Insignificant treatment effect	Rejected (***)	Rejected (***)
AR(1) serially uncorrelated	Not rejected	Not rejected
Full pass through of VAT	Not rejected	Not rejected

Germany The feasible GLS estimator again does not perform better than the first difference estimator. We already conjenctured that the error term may follow a random walk and the GLS estimator does not quite change that conjencture. This in particular means there is no real reason to prefer feasible GLS over the first differencing.

Figure 7.15: ACF plot of weighted feasible GLS residuals (Germany)

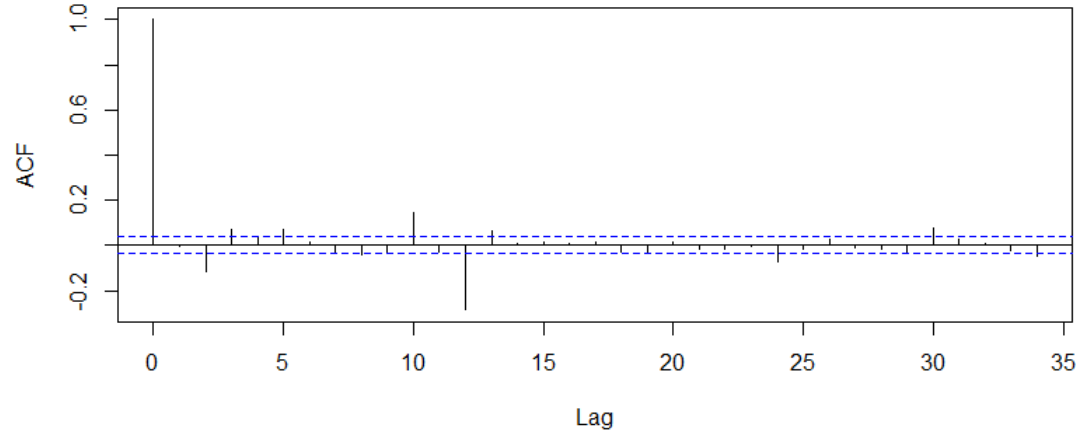
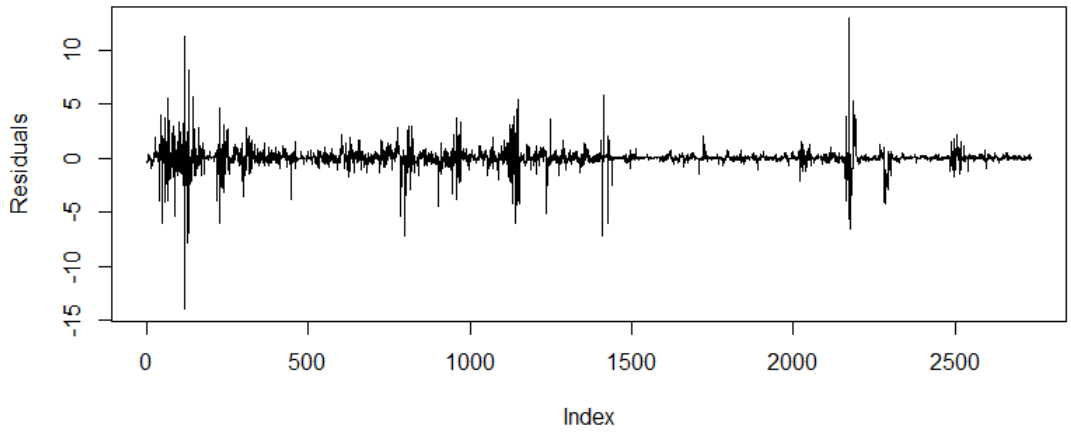


Figure 7.16: Plot of residuals



The ACF and residual plots do not give any interesting new insight compared to the restricted feasible GLS analysis made when Belgium is used as control group. The main results from regression are presented in 7.9.

Table 7.9: Feasible GLS summary (Germany)

Null hypothesis	Unweighted	Weighted
Insignificant treatment effect	Rejected (***)	Rejected (***)
AR(1) serially uncorrelated	Not rejected	Not rejected
Full pass through of VAT	Not rejected	Not rejected

7.3.3 Summary and conclusion

This paragraph considered the causal effects of the 2001 VAT increase on consumer prices. The final conclusions for the 2001 VAT increase are

1. **There has been a significant causal treatment effect as a result of the VAT increase.**
2. **The null hypothesis that the VAT increase was fully passed through is not rejected.**

The results are derived from the following observations.

Analysis of estimators Several estimators have been discussed, and we conclude that the pooled OLS, random effects and fixed effects estimators should be rejected as reliable estimators. Explicitely taking serial correlation into account leads to a dramatic improvement of the accuracy of the estimates. However, first differencing (non-parametric transformation) did not seem to underperform for feasible GLS restricted to $ARMA\,(1,1)$ errors. In fact, there was never real evidence that feasible GLS was to be preferred over first differencing.

Comparison with CPB analysis The CPB analysis of Vrijburg et al. (2014) never obtained significant treatment effects at a 5% level. At best they got treatment effects that were significant at a 10% level, which is insignificant to the standards used in this thesis. Recall that they only used Belgium as a control group, for which this thesis neither obtained significant treatment effects using the pooled OLS, fixed effects and random effects estimator. On the contrary, the first difference and generalised least squares estimators always produced significant treatment effects at a 0.1% significance level. This is a clear improvement from the CPB policy analysis of the 2001 VAT increase.

There is not enough evidence to reject the null hypothesis that the VAT increase was fully passed through on consumer prices. Vrijburg et al. (2014) did not reject this hypothesis either. At the other hand, this analysis obtained considerably more accurate estimates by explicitely taking serial correlation into account. The estimates produced by the efficient estimators in this thesis are arguably more accurate than the estimates of the CPB report. They are especially suitable for further tax policy analyses.

7.4 VAT increase 2012

In October 2012 the Dutch government increased the VAT rate from 19% to 21%. We again consider the theoretical maximum VAT-pass through rate. For the 2012 VAT-increase, this equals

$$\frac{1.21 - 1.19}{1.19} = 0.01681 \tag{7.11}$$

or 1.681%. This means that the theoretical models suggests

$$treat \in [0, 1.681]. \tag{7.12}$$

Before starting the estimation procedure, some 2012 specific policy changes that may cause endogeneity in the treatment parameter are discussed.

7.4.1 Policy analysis

There have been some policy changes that may cause endogeneity if we do not appropriate measures. Firstly, alcohol has been excluded from the analysis because Belgium considerably increased the excise tax on alcoholic drinks in 2013 (Federale Overheidsdienst Financiën, 2013). As per the *belastingplan 2013*, the Dutch government increased excise taxes on fuel and natural gas.

Furthermore, there is a potential endogeneity danger from the fact that the Netherlands followed a totally different budgetting pattern during the economic crisis than Belgium. That is, the Netherlands followed the nothern European approach of cutting spendings and increasing taxes while Belgium did the opposite. This also means that the economic growth in the Netherlands was lower compared to Belgium, which may have influence on inflation levels. To control for this endogeneity, we strongly rely on the *gdp* parameter. This parameter controls for the impact of the differences in economic growth on the inflation levels. We expect this parameter to be significant for most estimators (as opposed to the 2001 VAT increase).

7.4.2 Estimates

The estimates presented in this section are based on the analysis made above. When Belgium is used as the control group, the amount of commodity groups for which the HICP inflation was measured is considerably bigger than the amount of product groups considered in the 2001 VAT increase ($N = 84$). This may increase the accuracy of the estimators used in this thesis, leading to more reliable inference. Vrijburg et al. (2014) did not use the extra availability in data. There was no extra data available when Germany is used as the control group ($N = 36$). That is, the standard errors may be higher for the German estimates. Treatment start at time period $t_0 =$ 01-10-2012 and ends at $t_1 =$ 31-09-2013. The panel consists $T = 36$ time units, starting from 01-01-2011 up to 31-12-2013.

Pooled OLS

The estimates for the pooled OLS estimator are given in table B.1 ('B' refers to appendix B). The pooled OLS in general seems very inaccurate, which is on-par with the analysis of the 2001 VAT-increase.

Belgium We again observe that the unweighted estimates are significantly different from 0, while the weighted estimates do not reject that hypothesis. This was also observed for the 2012 VAT increase. There seems to be a rather big difference in unweighted and weighted estimates, but these differences are never significantly different from each other. The standard errors produced by both the weighted and unweighted estimators are reason to believe that the pooled OLS estimator is inefficient. We can immediately observe that from the $AR(1)$ fit of the residuals, but it is also confirmed by a specification test:

H_0 : No unobserved fixed effects or serial correlation in the error term

H_1 : Assumption A.POLS.3 fails

H_0 is rejected at a 0.1% significance level.

Figure 7.17: ACF plot of weighted pooled OLS residuals (Belgium)

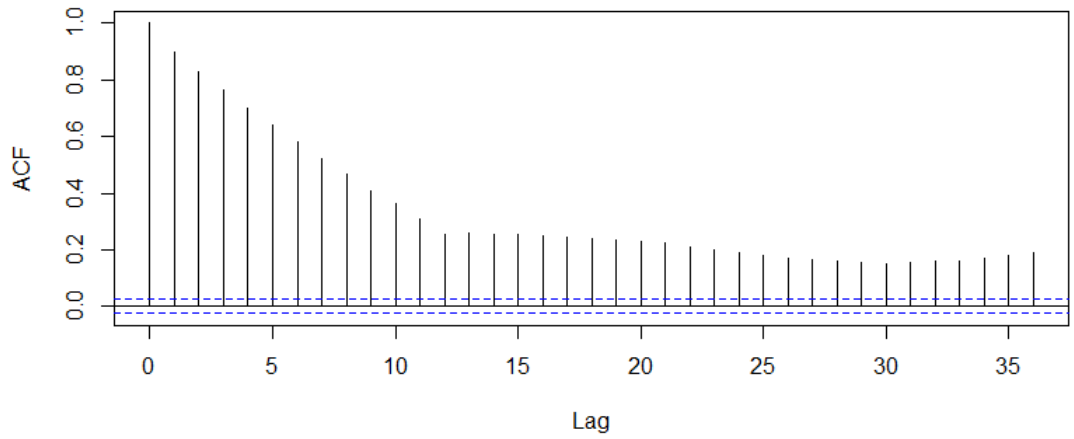
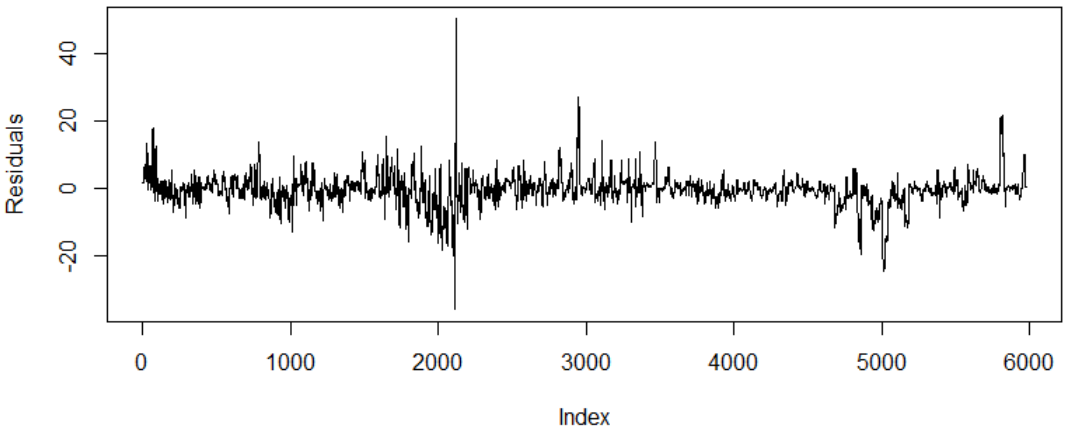


Figure 7.18: Plot of residuals



The rejection of H_0 can clearly be observed from the plots of the residuals and the ACF. The serial dependence seems very strong for the pooled OLS estimator, indeed suggesting that the estimates are inaccurate. A summary of the regression is given in 7.10.

Table 7.10: First difference summary (Belgium)

Null hypothesis	Unweighted	Weighted
Insignificant treatment effect	Rejected (**)	Not rejected
AR(1) serially uncorrelated	Not rejected	Not rejected
Full pass through of VAT	Not rejected	Not rejected

Germany Both the unweighted and weighted pooled OLS estimator give significant treatment effects at a 0.1% significance level. The point estimates for the weighted and unweighted pooled OLS estimator are very close and in particular not significantly different from each other. While the standard errors suggest that the estimates are slightly more accurate, the pooled OLS estimator is still an inefficient estimator. This is also confirmed by the same hypothesis test used for Belgium; H_0 is rejected at a 0.1% significance level.

Figure 7.19: ACF plot of weighted pooled OLS residuals (Germany)

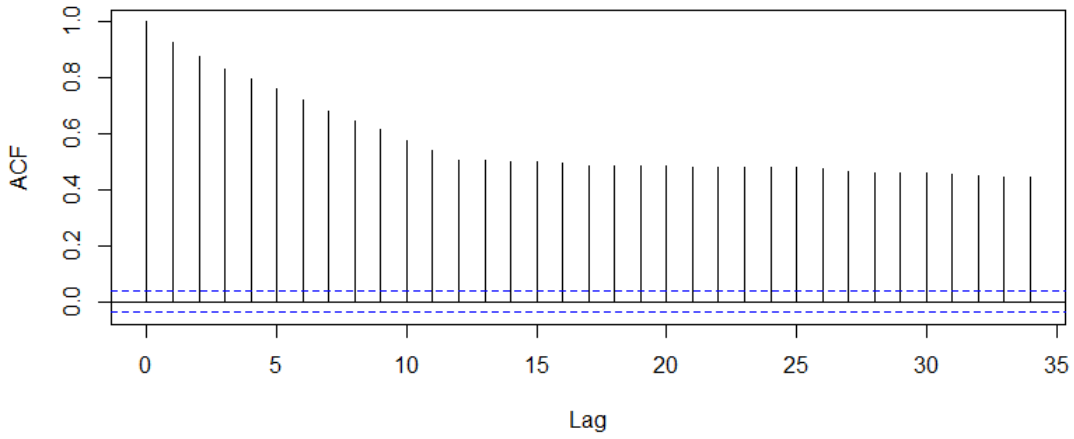
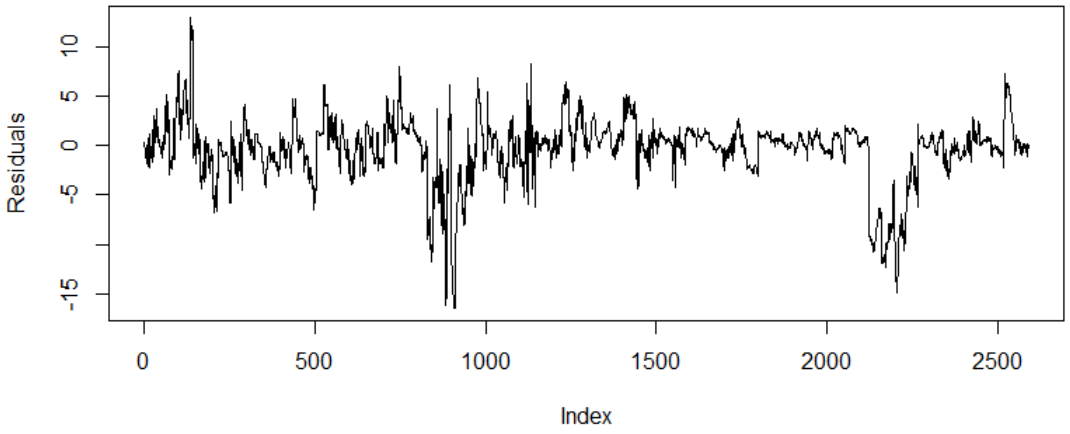


Figure 7.20: Plot of residuals



The plots do not give any new insights compared to the previous analysis. A summary of the regression results is given in table 7.11.

Table 7.11: Pooled OLS summary (Germany)

Null hypothesis	Unweighted	Weighted
Insignificant treatment effect	Rejected (***)	Rejected (***)
AR(1) serially uncorrelated	Not rejected	Not rejected
Full pass through of VAT	Not rejected	Not rejected

Random effects and fixed effects

The estimates of the random effects and fixed effects estimators are given in tables B.2 and B.3, respectively. The results for the random effects and fixed effects estimator are on par and do not provide any new insights. The results are consequently only shortly summarised.

It may be interesting to note that when Germany is used as the control group, the estimates for the $AR(1)$ correlation significantly decrease compared to their pooled OLS counterparts. This may be a potential sign of the existence of an unobserved individual specific fixed effect. No significant decrease in serial correlation is observed when Belgium is used as the control group. Table 7.12 presents a summary of the regression results.

Table 7.12: Fixed effects summary (Belgium and Germany)

Null hypothesis	Unweighted	Weighted
Insignificant treatment effect	Rejected (***)	Rejected (***)
AR(1) serially uncorrelated	N/A	N/A
Full pass through of VAT	Not rejected	Not rejected

First difference

The estimates are given in table B.4. The first difference estimates seem to show a dramatic improvement in the accuracy of the estimates. Due to some special results obtained for Belgium, we will first discuss the case where Germany is used as the control group.

Germany Both weighted and unweighted results show significant treatment effects, yet they both do not reject the null hypothesis that the VAT increase was fully passed through to consumer prices. The first difference transformation did not fully remove serial correlation, suggesting that the feasible GLS estimator may give more accurate results. The confidence interval for the first difference estimator is given by

$$CI_{95\%}\left(treat_{\text{Weighted}}\right) = [0.7725, 1.7676] \tag{7.13}$$

The upper bound of this interval is just a little above the full pass-through rate.

Figure 7.21: ACF plot of weighted first difference residuals (Belgium)

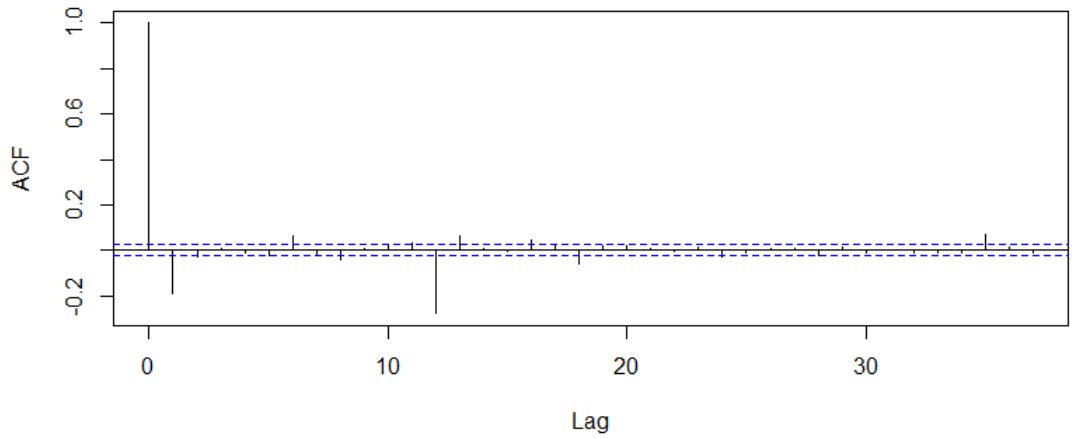


Figure 7.22: Plot of residuals

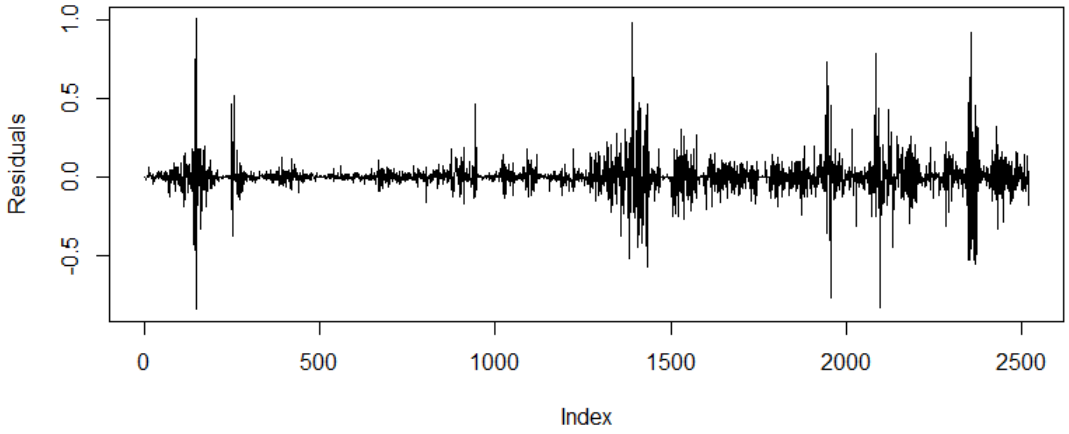


Table 7.13: First difference summary (Germany)

Null hypothesis	Unweighted	Weighted
Insignificant treatment effect	Rejected (***)	Rejected (***)
AR(1) serially uncorrelated	N/A	N/A
Full pass through of VAT	Not rejected	Not rejected

Belgium The weighted and unweighted results both show significant treatment effects at a 0.1% level. But there is more than that - the weighted results reject the null hypothesis that the 2012 VAT increase was fully passed through on consumer prices at a 1% significance level. That is,

$$CI_{99\%}\left(treat_{\text{Weighted}}\right) = [0.7354, 1.592] \tag{7.14}$$

which is the first time this hypothesis is rejected in this thesis. A fair question to ask is whether this justifies the conclusion that the 2012 VAT increase was not fully passed through. First note that the standard errors have been robustified for the remainder serial correlation that was not captured by the first difference estimator. Also, the estimates when Belgium is used as the control group rely on a much bigger sample than the German estimates. This may easily have resulted in an increase of accuracy, suggesting that we actually may rely on these estimates.

Figure 7.23: ACF plot of weighted first difference residuals (Germany)

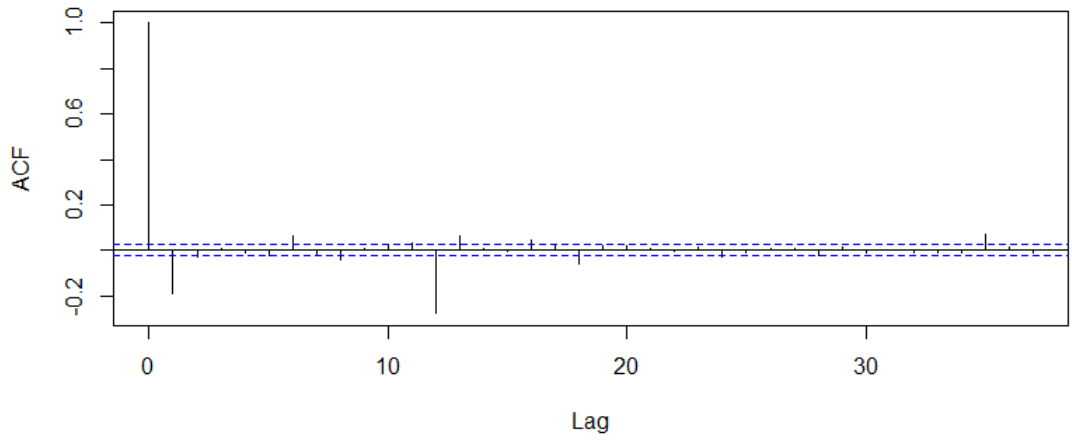
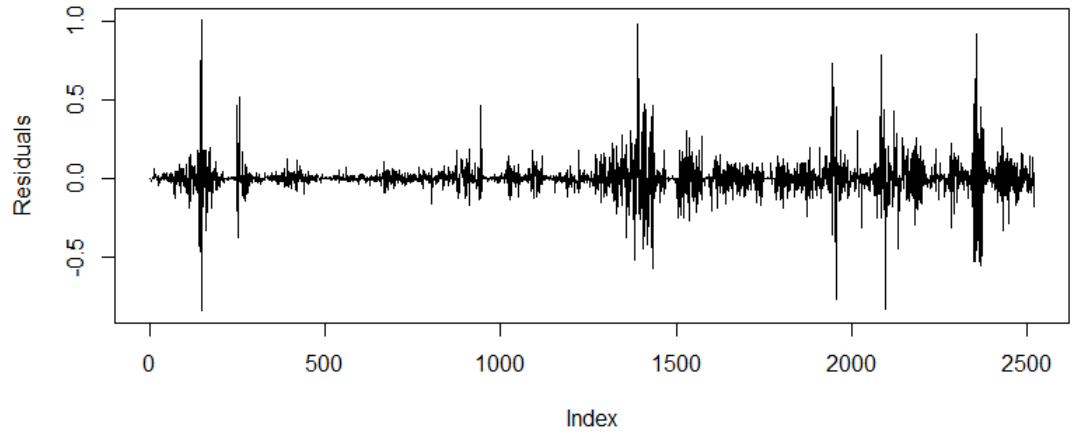


Figure 7.24: Plot of residuals



The plots of the residuals are almost identical to the first difference residual plot for Germany. Again, some white noise behaviour can be observed, suggesting that first differencing is an efficient method. We will first consider the estimates given by the feasible GLS estimator before making inference. The main results of this first difference regression are presented in table 7.14.

Table 7.14: First difference summary (Belgium)

Null hypothesis	Unweighted	Weighted
Insignificant treatment effect	Rejected (***)	Rejected (***)
AR(1) serially uncorrelated	Rejected	Rejected
Full pass through of VAT	Not rejected	Rejected (**)

Generalised Least Squares

The feasible Generalised Least Squares estimator is restricted for errors that follow an $ARMA(1,1)$ process. An $ARMA(2,2)$ may be used (i.e. it is not an overfit), but the estimates are almost identical and certainly do not lead to different inference. More generally, modelling error structures does not seem to improve accuracy compared to first differencing at all.

Belgium It not surprising that the feasible GLS estimator produces significant treatment effects for both the unweighted and weighted estimates. In particular, the feasible GLS estimator rejects the null hypothesis of full pass-through at a 1% significance level, i.e.

$$CI_{99\%}\left(treat_{\text{Weighted}}\right) = [0.745, 1.647]. \tag{7.15}$$

It is insignificant at a 99.9% level. The estimates again do not reject this null hypothesis. Note that the inference is exactely the same as for the first difference estimator.

Figure 7.25: ACF plot of weighted feasible GLS residuals (Belgium)

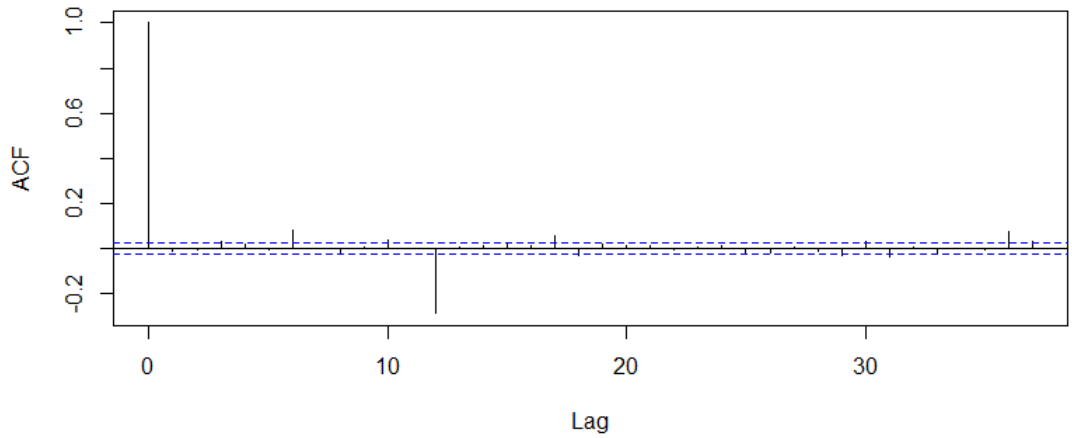


Figure 7.26: Plot of residuals

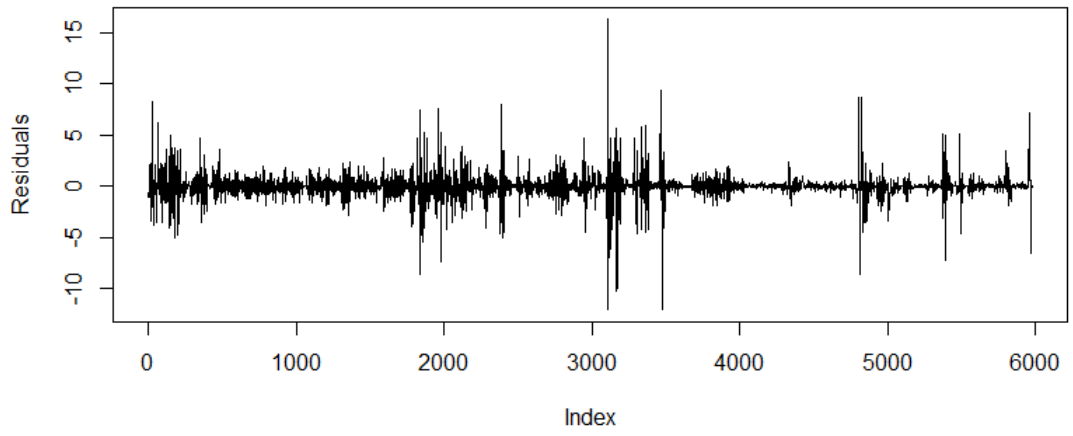


Table 7.15: Feasible GLS summary (Belgium)

Null hypothesis	Unweighted	Weighted
Insignificant treatment effect	Rejected (***)	Rejected (***)
AR(1) serially uncorrelated	N/A	N/A
Full pass through of VAT	Not rejected	Rejected (**)

Germany Both the weighted and unweighted estimates show significant treatment effects. However, they do not reject the null hypothesis of full pass-through. While the point estimates are roughly comparable with the Belgian estimates, it does not reject the null hypothesis. Yet they do not reject the hypothesis that the VAT was fully passed through to consumer prices. This is again due to the considerably higher standard errors when Germany is used as the control group. Again, this could easily be the result of the fact that the German estimates are based on fewer observations than their Belgian counterparts.

Figure 7.27: ACF plot of weighted feasible GLS residuals (Germany)

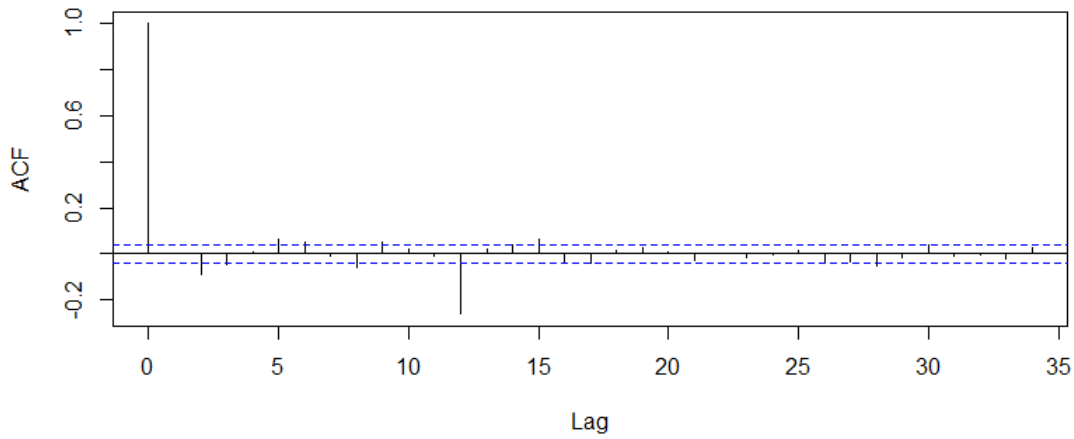


Figure 7.28: Plot of residuals

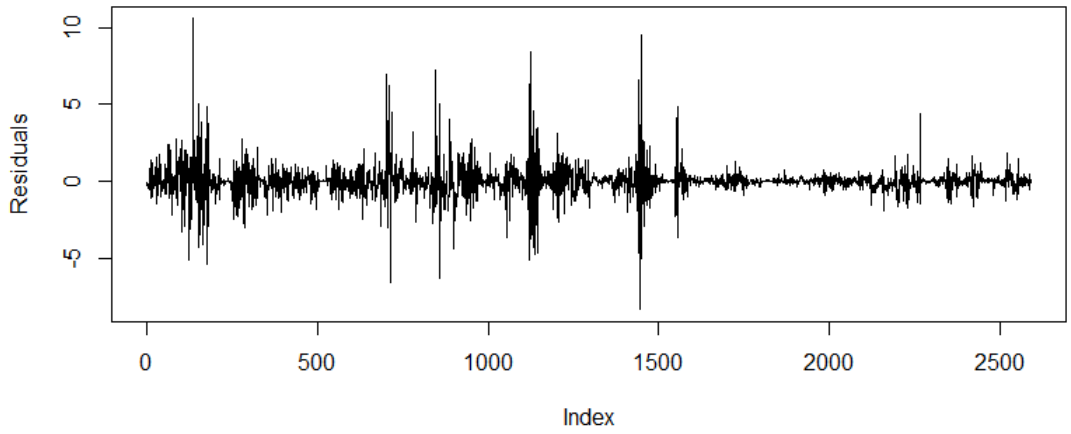


Table 7.16: Feasible GLS summary (Belgium)

Null hypothesis	Unweighted	Weighted
Insignificant treatment effect	Rejected (***)	Rejected (***)
AR(1) serially uncorrelated	N/A	N/A
Full pass through of VAT	Not rejected	Not rejected

Bayesian first differencing

When Belgium is used as the control group, a Bayesian weighted first differencing procedure proves to be suitable as an alternative for its frequentist counterpart. The robust standard errors are only marginally larger than the non-robust standard errors. We

assume that the likelihood ratio is given by

$$y_i|\mathbf{X}_i, \beta, \sigma \overset{iid}{\sim} N\left(\mathbf{X}_i\beta, \sigma_\varepsilon^2 \mathbf{I}_T\right). \tag{7.16}$$

The analysis will be performed using both uniform and informative priors.

Uniform prior The results are presented in table B.6. As expected, the inference does not change compared to the frequentist analysis. In particular, the credible interval for the parameter *treat* suggests that the null hypothesis of full incidence of VAT is rejected. Credible intervals have a different interpretation than confidence intervals. Since the true parameter is assumed to be a random variable, credible intervals capture the posterior uncertainty of the parameter location. A plot of the posterior density is given in figure 7.29. The treatment parameter seems a posteriori normally distributed, which is not surprising given the flat prior.

Figure 7.29: Plot of posterior density (treat)

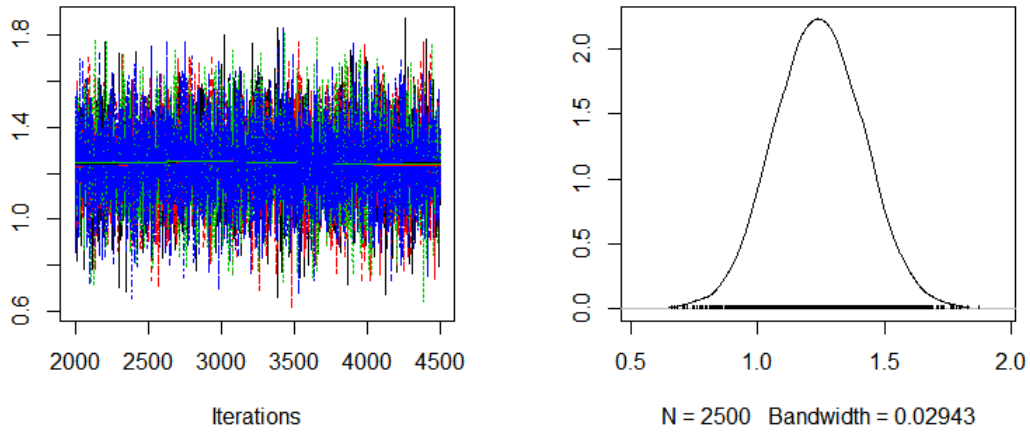


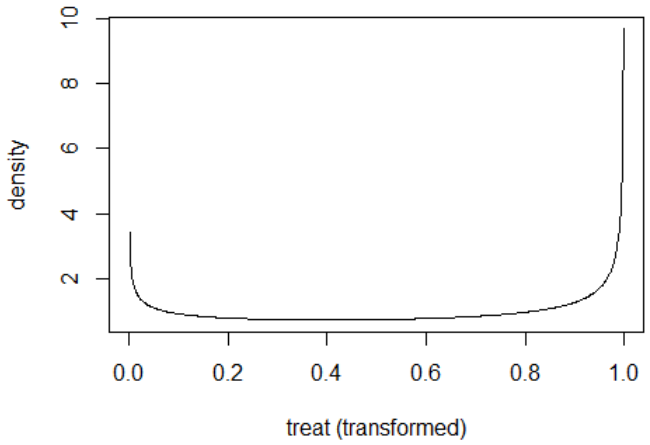
Table 7.17: FD Summary (Belgium)

Null hypothesis	Weighted
Insignificant treatment effect	Rejected (***)
AR(1) serially uncorrelated	N/A
Full pass through of VAT	Rejected (*)

Informative prior The results can be found in table B.7. The information for the informative prior is obtained from the analysis of the Deutsche Bundesbank (2008). As mentioned in the introduction, they suggest that most entrepreneurs fully pass-through the VAT to consumer prices. At the other hand, some entrepreneurs did not change prices

at all. This prior can loosely be modelled using a beta distribution, although the support of this distribution is $[0, 1]$ while $treat \in [0, 1.681]$. To circumvent this restriction, let $\tilde{treat} = \frac{treat}{1.681} \in [0, 1]$ denote the rescaled parameter. The prior distribution for the treatment parameter is $\tilde{treat} \sim \text{beta}(0.7, 0.45)$. This allows for some mass around no pass-through (i.e. 0), while placing most mass close to values indicating full pass-through. A plot of the prior distribution of the treatment parameter is given in figure 7.30.

Figure 7.30: Plot of prior distribution



Naturally, this is a very simplified model of the prior distribution, but it is able to capture the most important findings of Deutsche Bundesbank (2008). Figure 7.31 shows the posterior density of the treatment parameter. The prior distribution seems to have a slight influence on the posterior distribution, which is now somewhat skewed to the right. At the other hand, the influence of the prior seems mostly neglectable. This is also observable from the tables, the posterior mean hardly changes. The credible intervals neither show much change. The null hypothesis of full pass-through is still rejected under this prior.

Figure 7.31: Plot of posterior distribution (treat)

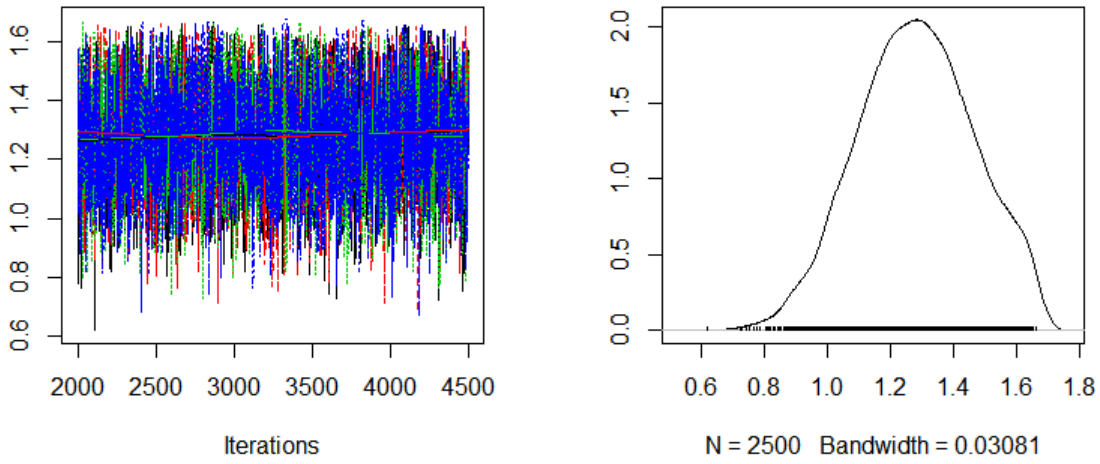


Figure 7.32: Plot of posterior distribution (σ_ε^2)

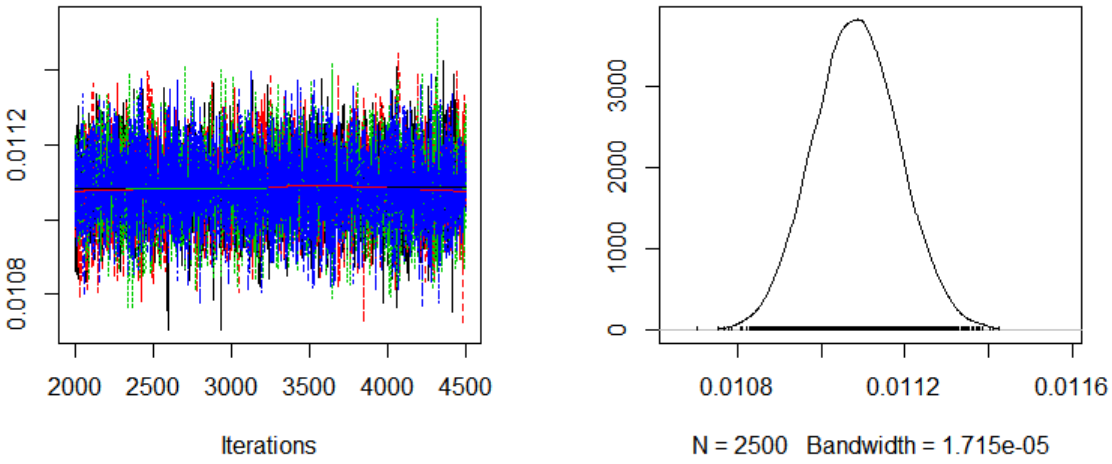


Table 7.18: FD Summary (Belgium)

Null hypothesis	Weighted
Insignificant treatment effect	Rejected (***)
AR(1) serially uncorrelated	N/A
Full pass through of VAT	Rejected (*)

7.4.3 Summary and conclusion

This paragraph discussed the causal treatment effects of the 2012 VAT increase. The following observations have been made

- **There has been a significant causal treatment effect**
- **The 2012 VAT increase was not fully passed through to consumer prices**

The results are based on the following observations:

Analysis of estimators

Several estimators were considered. Pooled OLS, fixed effects and random effects again provide too inaccurate estimates to be useful. Rather, we prefer a first differencing or a restricted feasible GLS procedure. The choice between either of them does not matter, they approximately have the same point estimates and standard errors. In particular, they do not lead to different inference. Both the first difference and restricted feasible GLS estimators lead to the rejection of the null hypothesis of full pass through when Belgium is used as the control group. When Germany is used as the control group, this hypothesis is not rejected. This is likely due to the fact that the Belgian estimates rely on much more observations than the German estimates. It is hence defendable to accept the first difference or the restricted feasible GLS estimator using Belgian as control group for the final inference for the 2012 VAT increase.

Comparison of the CPB results

Vrijburg et al. (2014) also concluded there had been a causal treatment effect at a 1% significance level. However, they did not reject the null hypothesis that the VAT was fully passed through in consumer prices. This means that the second conclusion in this thesis is to rejects this result from the CPB analysis. Again, the difference in inference is due to the fact that the estimators considered in this thesis provide much more accurate estimates than the estimators used by Vrijburg et al. (2014). Furthermore, they did not use the extra availability of data and sticked to the same commodity groups as in the 2001 regression.

Chapter 8

Conclusion

This thesis evualuated the tax incidence of the 2001 and 2012 Dutch Value Added Tax hikes. The research question of this thesis was whether the null hypothesis of full pass through could be rejected for both tax hikes. Based on the results of the weighted estimators, this question is answered as follows:

- **The hypothesis of full pass through is not rejected for the 2001 VAT increase**
- **The hypothesis of full pass through is rejected for the 2012 VAT increase**

Furthermore, for both VAT hikes a causal treatment effect was observed at a 0.1% significance level. This is a clear improvement over the findings in the CPB paper of Vrijburg et al. (2014). The results in this thesis also suggest that the first difference estimator should be included as one of the standard methods for estimating difference-in-differences models. These findings are based on the following observations

Theoretical analysis

This thesis considered linear regression models in which the error term is was heavily affected by serial correlation. While the classical panel data estimators may provide consistent and unbiased estimates, they are not efficient. For small and mid-sized samples, this may lead to inaccurate estimates and high standard errors. There is consequently a clear need to come up with estimators that are (more) efficient in presence of serial correlation.

The first difference estimator was the first that took serial correlation explicitely into account. This estimator is obtained by a non-parametric transformation of the model and is efficient provided the original errors follow a random walk. In that case, this estimator can even be proven to be the Best Unbiased Linear Estimator. However, this estimator also proved to perform well for other types of serial dependence.

Another estimator that took serial correlation into account was a restricted variant of the feasible GLS estimator. This is a parametric transformation and allows for the modelling of more complex serial dependence structures. While this estimator is not BLUE, it can be proven to be asymptotically efficient relatively to a large class of other linear estimators under appropriate assumptions. In many pratical applications however, there is no real reason to prefer this estimator over first differencing.

Applications on VAT increases

These insights were applied on the analysis of the 2001 and 2012 Dutch VAT increases. An economic model was derived that theoretically captures the influence of a VAT increase on customer prices. This model was used to derive some hypotheses for the statistical analysis of the 2001 and 2012 VAT increases. This analysis relied on a difference-in-differences model accompanied by two control variables. This model appeared to be heavily affected by serial correlation, resulting in the fact that estimators that do not properly capture serial correlation produce inaccurate estimates.

The CPB analysis of Vrijburg et al. (2014) used such inaccurate estimators, resulting in poor estimates. In particular, even the estimates provided by what they called an efficient estimator were often insignificant at a 5% level (i.e. not significantly different from 0). Using the theory gathered and developed in this thesis, a first differencing and a feasible GLS estimator restricted for $ARMA\,(1,1)$ errors were used. Both estimators dramatically improve existing estimates, with treatment effects that are significant at a 0.1% level. None of the estimators however rejected the null hypothesis of full VAT pass through to customer prices for the 2001 VAT increase.

For the 2012 VAT increase, the null hypothesis of full VAT pass-through were rejected at a 1% significance level when Belgium was used as a control group. When Germany was used as a control group, this hypothesis was not rejected. The point estimates were close to eachother and the non-rejection of the null hypothesis in the German case was mainly due to the fact that the standard errors were higher. The choice for appropriate control group was settled in favour for Belgium, since these relied on a considerably bigger sample than in the case Germany was used as control group. It is therefore in line with expectations that the Belgian estimates are more accurate, resulting in lower standard errors.

When Belgium was used as a control group for the 2012 VAT increase, a fully parametric Bayesian first differencing approach seemed plausible. Both a non-informative uniform and informative prior were used. The informative prior was based on earlier research of the Deutsche Bundesbank (2008). It did not matter for inference whether an uninformative or informative prior was used - both Bayesian methods rejected the null hypothesis of full pass-through.

Difference-in-differences analysis

Difference-in-differences models are prone to heavy serial correlation in the error term. Bertrand et al. (2004) proposed some solutions to take the inefficiencies caused by serial correlation into account, among which the use of robust standard errors is the most commonly applied. While robust standard errors provide sufficient control for type I errors, they may cause a big increase in type II errors. We propose the use of the feasible GLS estimator and first difference estimator. If the serial correlation is very high, inference for the first difference and feasible GLS estimators is rarely different. That is, complex modeling of error structures does not prove to perform better than first difference estimation. At the other hand, first differencing clearly underperforms compared to other estimators when the serial correlation is weak. As noted by Bertrand et al. (2004), this is rarely the case in DD estimation. Consequently, the first differencing estimator is in many pratical applications far superior over pooled OLS accompanied with robust standard errors.

Future research

Apart from a special case, Bayesian methods appeared to be very difficult to implement. This is due to the fact that fully parametric Bayesian regression frameworks do not offer the same flexibility as the class of frequentist generalised method of moments estimator. In particular, the lack of heteroscedastic and serial correlation robust standard errors in a Bayesian regression framework made it often infeasible to use it for this VAT pass-through study. While serial dependence structures were properly modelled, heteroscedasticity structures were not. That is, the inefficiency caused by heteroscedasticity was taken for granted and corrected for by using robust standard errors. Fully parametric Bayesian estimation would be feasible for this study provided heteroscedasticity is properly modelled. This is left for future research.

Appendix A

2001 VAT increase estimates

Table A.1: Pooled OLS

	Point Estimates			
	Belgium		Germany	
	Unweighted	Weighted	Unweighted	Weighted
	(1)	(2)	(3)	(4)
treat	2.243**	1.575	2.327***	1.846**
	(0.796)	(1.194)	(0.428)	(0.611)
timed	0.389	0.562	0.195	0.758
	(0.537)	(0.888)	(0.307)	(0.466)
countrd	0.397	1.677	1.643	2.224*
	(0.937)	(0.881)	(0.874)	(0.982)
gdp	−0.257***	−0.326**	−0.315***	−0.176*
	(0.069)	(0.107)	(0.063)	(0.074)
infl	0.345	0.589	0.038	−0.322
	(0.401)	(0.568)	(0.231)	(0.207)
Constant	1.203	0.931	0.427	0.362
	(1.063)	(0.874)	(0.448)	(0.773)
F Statistic	42.201***	122.240***	68.331***	111.616***
	Residuals $AR(1)$ fit			
	Belgium		Germany	
	Unweighted	Weighted	Weighted	Unweighted
AR1	0.906***	0.891***	0.961***	0.964***
	(0.022)	(0.044)	(0.009)	(0.012)

Note: *p<0.05; **p<0.01; ***p<0.001

Table A.2: Random Effects

	Point Estimates			
	Belgium		Germany	
	Unweighted	Weighted	Unweighted	Weighted
treat	2.243**	1.575	2.327***	1.846**
	(0.796)	(1.194)	(0.428)	(0.611)
timed	0.389	0.562	0.195	0.758
	(0.537)	(0.888)	(0.307)	(0.466)
countrd	0.397	1.677	1.643	2.224*
	(0.937)	(0.881)	(0.874)	(0.982)
gdp	−0.257***	−0.326**	−0.315***	−0.176*
	(0.069)	(0.107)	(0.063)	(0.074)
infl	0.345	0.589	0.038	−0.322
	(0.401)	(0.568)	(0.231)	(0.207)
Constant	1.203	0.931	0.427	0.362
	(1.063)	(0.874)	(0.448)	(0.773)
F Statistic	42.201***	73.318***	68.331***	104.001***
	Residuals $AR(1)$ **fit**			
	Belgium		Germany	
	Unweighted	Weighted	Unweighted	Weighted
AR1	0.793***	0.769***	0.845***	0.835***
	(0.042)	(0.075)	(0.031)	(0.034)

Note: *p<0.05; **p<0.01; ***p<0.001

Table A.3: Fixed Effects

	Point Estimates			
	Belgium		Germany	
	Unweighted	Weighted	Unweighted	Weighted
treat	2.243*	1.575	2.327***	1.846**
	(0.796)	(1.194)	(0.428)	(0.611)
timed	0.389	0.562	0.195	0.758
	(0.537)	(0.888)	(0.307)	(0.466)
gdp	−0.257	−0.326**	−0.315***	−0.176*
	(0.069)	(0.107)	(0.063)	(0.074)
infl	0.345	0.589	0.038	−0.322
	(0.401)	(0.568)	(0.231)	(0.207)
	Residuals $AR(1)$ **fit**			
	Belgium		Germany	
	Unweighted	Weighted	Unweighted	Weighted
AR1	0.790***	0.760***	0.842***	0.830***
	(0.042)	(0.076)	(0.031)	(0.036)

Note: *p<0.05; **p<0.01; ***p<0.001

Table A.4: First Difference

	Point Estimates			
	Belgium		Germany	
	Unweighted	Weighted	Unweighted	Weighted
treat	1.048^{***}	1.225^{***}	1.045^{***}	1.305^{***}
	(0.274)	(0.294)	(0.273)	(0.294)
timed	0.297^{*}	0.144	0.221^{***}	0.151^{*}
	(0.125)	(0.161)	(0.063)	(0.068)
gdp	0.255	0.272	−0.056	0.029
	(0.134)	(0.324)	(0.043)	(0.068)
infl	0.063	−0.026	0.206^{*}	0.287^{***}
	(0.163)	(0.295)	(0.081)	(0.083)
F Statistic	5.362^{**}	5.708^{***}	15.893^{***}	24.580^{***}
	Residuals $AR(1)$ **fit**			
	Belgium		Germany	
	Unweighted	Weighted	Unweighted	Weighted
AR1	-0.298^{***}	−0.339	−0.086	−0.067
	(0.086)	(0.186)	(0.048)	(0.085)

Note: *p<0.05; **p<0.01; ***p<0.001

Table A.5: Generalised Least Squares

	Point Estimates			
	Belgium		Germany	
	Unweighted	Weighted	Unweighted	Weighted
treat	1.103***	1.258***	1.126***	1.341***
	(0.270)	(0.347)	(0.264)	(0.309)
countrd	0.234	1.558	1.432	2.025**
	(0.909)	(1.112)	(0.897)	(0.767)
timed	0.529***	0.556*	0.292***	0.196**
	(0.150)	(0.237)	(0.064)	(0.075)
infl	0.277	0.284	0.241**	0.272***
	(0.169)	(0.268)	(0.089)	(0.072)
gdp	0.050	0.948	−0.095*	0.007
	(0.091)	(1.932)	(0.042)	(0.472)
Constant	0.786	0.022	−0.048	−0.053
	(0.641)	(0.046)	(0.545)	(0.051)

	OLS estimates for variance parameters			
	Belgium		Germany	
	Unweighted	Weighted	Unweighted	Weighted
AR1	0.885	0.896	0.885	0.896
MA1	0.061	0.039	0.061	0.039

Note: *p<0.05; **p<0.01; ***p<0.001

Appendix B

2012 VAT increase estimates

Table B.1: Pooled OLS Estimates

	Point Estimates			
	Belgium		Germany	
	Unweighted	Weighted	Unweighted	Weighted
treat	1.821**	1.212	1.936***	2.076***
	(0.564)	(0.779)	(0.505)	(0.541)
timed	−0.107	−0.148	−0.293	−0.465*
	(0.224)	(0.432)	(0.182)	(0.234)
countrd	−0.395	−0.643	−0.199	−0.397
	(0.555)	(0.652)	(0.726)	(0.633)
gdp	−0.114	−0.005	−0.150**	−0.159*
	(0.109)	(0.150)	(0.056)	(0.065)
infl	0.105	0.312	0.058	0.033
	(0.193)	(0.295)	(0.152)	(0.248)
Constant	0.859	1.103	0.918	1.529*
	(0.531)	(0.692)	(0.514)	(0.694)
F Statistic	23.408***	10.444***	24.374***	29.982***

	Residuals $AR(1)$ **fit**			
	Belgium		Germany	
AR1	0.900***	0.900***	0.942***	0.945***
	(0.015)	(0.013)	(0.010)	(0.010)
Constant	−0.006	−0.002	−0.009	−0.002
	(0.025)	(0.002)	(0.023)	(0.002)

Table B.2: Random Effects estimates

	Point Estimates			
	Belgium		Germany	
	Unweighted	Weighted	Unweighted	Weighted
treat	1.821^{**}	1.212	1.936^{***}	2.076^{***}
	(0.564)	(0.782)	(0.505)	(0.571)
timed	-0.107	-0.148	-0.293	-0.465^{*}
	(0.224)	(0.437)	(0.182)	(0.194)
countrd	-0.395	-0.183	-0.199	-0.269
	(0.555)	(0.724)	(0.726)	(0.753)
gdp	-0.114	-0.005	-0.150^{**}	-0.159^{***}
	(0.109)	(0.151)	(0.056)	(0.044)
infl	0.105	0.312	0.058	0.033
	(0.193)	(0.276)	(0.152)	(0.207)
Constant	0.859	0.409^{***}	0.918	1.048^{***}
	(0.531)	(0.028)	(0.514)	(0.039)

	Residuals $AR(1)$ **fit**			
	Belgium		Germany	
AR1	0.816^{***}	0.809^{***}	0.825^{***}	0.855^{***}
	(0.026)	(0.021)	(0.027)	(0.026)
Constant	-0.006	-0.001	-0.018	-0.0004
	(0.023)	(0.001)	(0.023)	(0.002)
Note:		*p<0.05; **p<0.01; ***p<0.001		

Table B.3: Fixed effects estimator

	Point Estimates			
	Belgium		Germany	
	Unweighted	Weighted	Unweighted	Weighted
treat	1.821**	1.212	1.936***	2.076***
	(0.564)	(0.782)	(0.505)	(0.576)
timed	−0.107	−0.148	−0.293	−0.465*
	(0.224)	(0.445)	(0.182)	(0.208)
gdp	−0.114	−0.005	−0.150**	−0.159***
	(0.109)	(0.156)	(0.056)	(0.044)
infl	0.105	0.312	0.058	0.033
	(0.193)	(0.294)	(0.152)	(0.216)
F Statistic	55.441***	23.924***	80.770***	99.261***

	Residuals $AR(1)$ **fit**			
	Residuals			
	Unweighted	Weighted	Unweighted	Weighted
AR1	0.901***	0.802***	0.943***	0.855***
	(0.015)	(0.022)	(0.010)	(0.027)
Constant	−0.002	−0.0002	−0.000	0.0001
	(0.025)	(0.001)	(0.023)	(0.002)

Note: *p<0.05; **p<0.01; ***p<0.001

Table B.4: First Differences

	Point Estimates			
	Belgium		Germany	
	Unweighted	Weighted	Unweighted	Weighted
treat	1.232^{***}	1.164^{***}	1.448^{***}	1.270^{***}
	(0.329)	(0.166)	(0.269)	(0.254)
timed	-0.019	0.030	-0.149	-0.076
	(0.069)	(0.085)	(0.098)	(0.109)
gdp	-0.100	0.042	-0.010	0.007
	(0.074)	(0.104)	(0.047)	(0.078)
infl	-0.082	0.002	0.047	0.252
	(0.137)	(0.141)	(0.157)	(0.237)
F Statistic	27.651^{***}	29.098^{***}	33.732^{***}	35.103^{***}

	Residuals $AR(1)$ **fit**			
	Residuals			
	Unweighted	Weighted	Weighted	Unweighted
AR1	-0.185^{***}	-0.194^{***}	-0.158^{***}	-0.145^{**}
	(0.053)	(0.048)	(0.038)	(0.046)

Note: *p<0.05; **p<0.01; ***p<0.001

Table B.5: Generalised Least Squares

	Point Estimates			
	Belgium		Germany	
	Unweighted	Weighted	Unweighted	Weighted
treat	1.330***	1.196***	1.559***	1.360***
	(0.327)	(0.159)	(0.272)	(0.247)
countrd	−0.406	−0.410	−0.007	0.026
	(0.492)	(0.548)	(0.660)	(0.655)
timed	−0.028	0.031	−0.188*	−0.115
	(0.073)	(0.092)	(0.094)	(0.103)
infl	−0.067	0.039	0.026	0.164
	(0.129)	(0.143)	(0.131)	(0.217)
gdp	−0.106	0.033	−0.050	−0.056
	(0.069)	(0.104)	(0.039)	(0.056)
Constant	1.186**	1.186*	0.749	0.548
	(0.425)	(0.529)	(0.498)	(0.678)

	First stage fit			
	Estimates for feasible GLS			
	Unweighted	Weighted	Unweighted	WeightedFEGLS
AR1	0.778	0.767	0.781	0.791
MA1	0.150	0.167	0.099	0.076

Note: *p<0.05; **p<0.01; ***p<0.001

Table B.6: Bayesian first difference with uninformative prior

	mean	se_mean	sd	2.5%	97.5%	n_eff	Rhat
treat	1.2416	0.0021	0.1763	0.9029	1.5835	7177.6814	1.0001
timed	0.0107	0.0016	0.1325	-0.2489	0.2723	7021.8111	1.0002
infl	0.1200	0.0011	0.1066	-0.0891	0.3283	10000.0000	1.0002
gdp	0.2314	0.0008	0.0830	0.0659	0.3957	10000.0000	1.0000
sigma	0.0111	0.0000	0.0001	0.0109	0.0113	10000.0000	0.9999

Table B.7: Bayesian first difference with informative prior

	mean	se_mean	sd	2.5%	97.5%	n_eff	Rhat
treat	1.2831	0.0031	0.1834	0.9230	1.6322	3502.3567	1.0003
timed	-0.0114	0.0021	0.1349	-0.2724	0.2546	3987.3857	1.0002
infl	0.1182	0.0013	0.1088	-0.0955	0.3298	7036.9671	0.9999
gdp	0.2333	0.0010	0.0828	0.0696	0.3970	6808.3819	1.0001
sigma	0.2300	0.0010	0.0841	0.0648	0.3960	7476.5947	0.9999

Appendix C

Commodities included in regression

Table C.1: Standard rate taxed goods (both Germany and Belgium)

HICP	Description
CP021	Alcoholic beverages
CP022	Tobacco
CP0311	Clothing materials
CP0312	Garments
CP0313	Other articles of clothing and clothing accessories
CP0314	Cleaning, repair and hire of clothing
CP032	Footwear
CP0431	Materials for the maintenance and repair of the dwelling
CP0451	Electricity
CP0452	Gas
CP0511	Furniture and furnishings
CP0512	Carpets and other floor coverings
CP052	Household textiles
CP0531_0532	Major household appliances whether electric (...)
CP0533	Repair of household appliances
CP054	Glassware, tableware and household utensils
CP055	Tools and equipment for house and garden
CP0561	Non-durable household goods
CP0562	Domestic services and household services
CP0711	Motor cars
CP0712-714	Motor cycles, bicycles and animal drawn vehicles
CP0721	Spare parts and accessories for personal transport equipment
CP0723	Maintenance and repair of personal transport equipment
CP0724	Other services in respect of personal transport equipment
CP082_083	Telephone and telefax equipment and services
CP0911	Equipment for the reception, recording and reproduction (...)
CP0912	Photographic and cinematographic equipment and (...)
CP0913	Information processing equipment
CP0914	Recording media
CP0915	Repair of audio-visual, photographic and information (...)
CP0921_0922	Major durables for indoor and outdoor recreation (...)
CP0931	Games, toys and hobbies
CP0932	Equipment for sport, camping and open-air recreation
CP0933	Gardens, plants and flowers
CP0934_0935	Pets and related products; veterinary and (...)
CP0953_0954	Miscellaneous printed matter; stationery (...)
CP1211	Hairdressing salons and personal grooming establishments
CP1212_1213	Electrical appliances for personal care; other appliances, (...)
CP1231	Jewellery, clocks and watches
CP1232	Other personal effects

Bibliography

Aitken, A. (1934). On the linear combination of observations and the general theory of least squares. *Proceedings of the Royal Society of Edinburgh*.

Amemiya, T. (1985). *Advanced econometrics*. Harvard University Press.

Angrist, J., & Pischke, J. (2009). *Mostlyy harlmess econometrics*. Princeton University Press.

Arellano, M. (1987). Computing robust standard errors for within-group estimators. *Oxford Bulletin of Economics and Statistics*.

Atkison, A., & Lawrance, A. (1983). A comparison of asymptotically equivalent test statistics for regression transformation. *Biometrika*.

Auld, C. (2012). The intuition of robust standard errors. Retrieved July 27, 2018, from http://chrisauld.com/2012/10/31/the-intuition-of-robust-standard-errors/

Baltagi, B. (2005). *Econometric analysis of panel data*. John Wiley & Sons Ltd.

Benedek, D., de Mooij, R., Keen, M., & Wingender, P. (2015). *Estimating vat pass through*. International Monutary Fund.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*.

Box, G., & Tiao, G. (1992). *Bayesian inference in statistical analysis*. John Wiley and Sons, inc.

Breusch, T., & Pagan, A. (1979). A simple test for heteroskedasticity and random coefficient variation. *Econometrica*.

Chirakijja, J., Crossley, T., & Lührmann, M. (2009). *The stimulus effect of the 2008 uk temporary vat cut*. Institute for Fiscal Studies.

Cochrane, D., & Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto- correlated error terms. *Journal of the American Statistical Association*, *44*(245), 32–61. Retrieved from http://www.jstor.org/stable/2280349

Deutsche Bundesbank. (2008). *Preis- und mengenwirkungender mehrwertsteueranhebung zum 1. januar 2007*. Deutsche Bundesbank.

Fang, H. (2004). *Lecture notes on economics of taxation*. Duke University.

Federale Overheidsdienst Financiën. (2013). Overzicht van de aangepaste accijnzen op alcohol vanaf 5 augustus 2013.

Greene, W. (2011). *Econometric analysis*. Pearson Education Ltd.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*.

Harju, J., & Kosonen, T. (2014). *The inefficiency of reduced vat rates: Evidence from restaurant industry*. Government Institute for Economic Research (Finland).

Jeffreys, H. (1961). *Theory of probability*. Oxford University Press.

Jhingan, M. (2014). *Microeconomic theory*. Vrinda Publications.

Jongen, E., Lejour, A., & Massenz, G. (2017). *Cheaper and more haircuts after vat cut? evidence from the netherlands*. Bureau of Economic Policy Analysis (Netherlands).

Kakwani, N. C. (1967). The unbiasedness of zellner's seemingly unrelated regression equations estimators. *Journal of the American Statistical Association*, *62*(317), 141–142. doi:10.1080/01621459.1967.10482895

Lavine, M. (2014). *Conjegate priors: Beta and normal*. Michigan University of Technology.

Liu, C., & Rubin, D. B. (1995). Ml estimation of the t distribution using em and its extensions, ecm and ecme. *Statistica Sinica*.

MaCurdy, T. E. (1982). The use of time series processes to model the error structure of earnings in a longitudinal data analysis. *Journal of Econometrics*, *18*(1), 83–114. doi:https://doi.org/10.1016/0304-4076(82)90096-3

Mann, H., & Wald, A. (1943). On stochastic limit and order relationships. *Annals of Mathematical Statistics*.

Myles, G. (1995). *Public economics*. Cambridge University Press.

Newey, W., & McFadden, D. (1983/2007). *Handbook of econometrics*.

Prais, S., & Winsten, C. (1954). *Trend estimators and serial correlation*. Cowles Foundation for Research in Economics.

Robert, C. (2007). *The bayesian choice*. Springer Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*.

Sanger, C., & Thomas, R. (2018). *The outlook for global tax policy in 2018*. EY.

Shumway, R., & Stoffer, D. (2016). *Time series analysis and its applications*. Springer Verlag.

Stiglitz, J. (1988). *Economics of the public sector*. Norton & Copmany.

van der Meulen, F. (2017). Lecture notes statistical inference. Technische Universiteit Delft.

van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge University Press.

Vrijburg, H., Mellens, M., & Dijkstra, J. (2014). *Robust estimation of the vat pass-through in the netherlands*. Centraal Planbureau.

Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*.

White, H. (1984). A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity. *Econometrica*.

Wooldridge, J. M. (1991). On the application of robust, regression- based diagnostics to models of conditional means and conditional variances. *Journal of Econometrics*, *47*(1), 5–46. doi:https://doi.org/10.1016/0304-4076(91)90076-P

Wooldridge, J. M. (2001). *Econometric analysis of cross section and panel data*. MIT Press Ltd.