Stockholms universitet

The Immigrant Wage Gap
A Panel Data Forecasting Analysis

Alice Pipping

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2020:17
Matematisk statistik
September 2020

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2020:17**
http://www.math.su.se

The Immigrant Wage Gap
A Panel Data Forecasting Analysis

Alice Pipping*

September 2020

Abstract

Media reports state that there is a significant wage gap between native borns and non-native borns in Sweden. Using societal data from all Swedish municipalities and a simple linear regression, this study concludes that not only is this true, but that the gap grew between 2002 and 2017. In order to change this trend, it is helpful to build a predictive model so we can find the factors that affect the gap the most. To do this, the study builds three different panel data models: a pooled OLS regression, a fixed effects model and a random effects model. This in order to answer two questions: *can we predict the future of the wage gap using a panel data model*, and *which panel data model makes the best predictions?* The models are then evaluated and compared, in order to determine which one is best suited for the data. The study concludes that it is possible for a panel data model to make predictions of this wage gap based on the data we have, and that the fixed effects model is best suited for this analysis. It also concludes that some kind of panel data model could help in minimizing the wage gap. However, a more extensive analysis would be needed.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: alcp@telia.com. Supervisor: Mathias Millberg Lindholm, Kristofer Lindensjö, Felix Wahl.

Contents

1 Background **5**

2 Purpose and research questions **5**

3 Data **6**
3.1 Variables 7

4 Theory and method **7**
4.1 Regression 7
4.2 Residuals and model evaluation 8
4.2.1 Uncorrelated residuals 9
4.2.2 Zero mean 10
4.2.3 Constant variance (homoscedasticity) 10
4.2.4 Normal distribution 10
4.2.5 Linearity 10
4.3 Building the models 11
4.3.1 Pooled OLS regression 12
4.3.2 The fixed effects model 12
4.3.3 The random effects model 13
4.3.4 Choosing variables 14
4.4 Comparing the models 15
4.4.1 Standard error 15
4.4.2 R-squared 15
4.5 Evaluating the predictions - Cross-validation 16
4.5.1 Mean Squared Error of Prediction 17

5 Results **18**
5.1 National average income difference 18
5.1.1 A simple regression model 19
5.2 The contending models 20
5.3 Residuals - Are the assumptions met? 22
5.3.1 Uncorrelated residuals and Zero mean 22
5.3.2 Residuals vs fitted 23
5.3.3 Q-Q plot 25
5.4 Which model is better? 26
5.4.1 R squared and standard error 26
5.5 Accuracy of predictions 27
5.5.1 MSEP 27
5.5.2 Plotting the predictions 27

6 Discussion **29**

7 References **30**

A Appendix A **33**
A.1 Simple regression model 33
A.2 Pooled model 33
A.3 Fixed effects model 34
A.4 Random effects model 36
A.5 The pooled model's residuals 36
A.6 List of municipalities 38

1 Background

You have probably heard of the gender wage gap, as many studies have been published on the subject (see Arulampalam et al. and Blau & Kahn). Something you do not hear about as often - but probably does not surprise you - is the wage gap between native borns and non-native borns. This wage gap has sometimes been the focus of media attention, and many things have been discussed as possible factors in deciding your salary. The media discussion has mentioned factors such as how far from the host country[1] you are born (Olsson 2017), what your gender is (Katz & Österberg 2013), how long you have been in the host country as well as the age you arrived, and your skills in the host country's language (Expressen 2017). It has also been suggested that having a name that sounds local, as well as looking and "seeming" local, helps (Expressen 2017). Political organizations have deemed this issue worthy of scientific reasearch. One example is an analysis from Saco (Edström et al. 2017) for the interested reader.

In order to change something in society, we first have to understand it. In the case of this wage gap, it would most easily be changed if we could find the factors that affect it the most. This would include trying to understand which factors to change, how much to change them and when to change them in order to eliminate the gap. One way to do that is to build a model that predicts the gap in future years. By analysing this model, we can understand what needs to change.

This study contributes to the understanding of how society affects this wage gap not by focusing on individual people and their backgrounds but instead on the society in the host country, or more specifically, the host municipality. It is limited to municipalities in Sweden between the years 2002 and 2017 and discusses how different societal variables affect the mean salary for an immigrant compared to the mean salary of a native born. It also analyses the differences between certain types of panel data models.

How much people make of course depends not only on what profession they work in, but also if they work full time. We have decided to ignore both of these aspects to accomodate to the frame of this study.

2 Purpose and research questions

The purpose of this analysis is to find the best model to predict the average income difference between native borns and non-native borns in different Swedish

[1] The European Commission's definition of 'host country' is "The EU State in which a non-EU national takes up legal residence" (European Commission DG Migration and Home Affairs 2020). In this study, it is not necissarily a non-EU national, but anyone born outside of the host country.

municipalities. The research questions are:

> How has the national average wage gap between immigrants and native borns changed in Sweden over the years 2002-2017?
>
> Can we make a predictive model for the average wage gap in each Swedish municipality based on the data we have?
>
> Which panel data model makes the best predictions for this data?

The idea is that the model should be able to predict the average wage gap in every municipality for future years. The model will be examined to see if it can predict the gap with a result close enough to the actual value to be deemed reliable. Where to draw the line for a reliable model is, however, not obvious, so we will use different well-known statistical analyses to determine how well the model makes predictions.

3 Data

The data for this analysis was downloaded from Statistics Sweden in March 2020. Statistics Sweden is responsible for all offical and government statistics in Sweden. They have a database of statistics that is updated every day (Statistics Sweden, *Statistical Database*). The data chosen for this analysis covers the years 2002-2017, because those were the years that had data for all of the sought after variables. This means that the analysis will have balanced data, which in turn means that the number of observations is the same in each cross-section unit, and no data is missing from the data frame. If this is not the case, and data is missing in some part of the data frame, the data is unbalanced.

From the years 2002-2017, data from all 290 Swedish municipalities on the variables average age, population, number of non-native born people (later changed to be proportion of non-native born people), average income and number of people with 'eftergymnasial utbildning', i.e. number of people with tertiary education. Beyond these, one variable was chosen that was Sweden-specific (rather than municipality-specific): Number of asylum seekers in Sweden. We have six explanatory variables, two variables that give us the panel data structure (year and municipality), and a response variable. The response variable is difference in average income between Swedish borns and non-native borns in a certain municipality in a certain year. This is calculated as:

$$\text{AVG(salary for Swedish borns)} - \text{AVG(salary for non-native borns)}, \quad (1)$$

(where AVG stands for average), and will thefeore be positive when the average salary for the Swedish born is higher than the avarage salary for the non-native born.

Both parts of the response variable were found on Statistics Sweden in the

format of average disposable income in the municipality, in a value called price base amount, a number yearly determined by the government (SFS 2010:110, 2 ch 6 & 7 §). The point of this value is to adjust for the inflation, so you can compare numbers between different years (Nationalencyklopedin, *Basbelopp*). To be able to compare the salary for native borns with the salary for non-native borns, the subtraction in equation (1) was made.

3.1 Variables

Table 1 lists the variables used in this study.

Type	Variable	Meaning	Size
Response	Y_{it}	AVG(salary for native borns) minus AVG(salary for non-native borns)	$-1 < Y_{it} < 6$
Explanatory	$X_{it}^{(1)}$	Municipal avg age	$36 < X_{it}^{(1)} < 50$
Explanatory	$X_{it}^{(2)}$	Municipal population	$2400 < X_{it}^{(2)} < 940000$
Explanatory	$X_{it}^{(3)}$	Municipal proportion non-natives	$0.02 < X_{it}^{(3)} < 0.05$
Explanatory	$X_{it}^{(4)}$	Number of asylumseekers in Sweden	$17\,000 < X_{it}^{(4)} < 170000$
Explanatory	$X_{it}^{(5)}$	Municipal avg income (thousand kronor)	$480 < X_{it}^{(5)} < 2800$
Explanatory	$X_{it}^{(6)}$	Municipal number of people with tertiary education	$200 < X_{it}^{(6)} < 360000$
Index	i	The municipality	$1 < i < 290$
Index	t	The year	$2002 < t < 2017$

Table 1: List of variables used in this study.

4 Theory and method

4.1 Regression

The panel data models, which will be used in this study and which we will get to in section 4.3, are, like many other models, based on basic linear regression. For this reason, we will first lay a basic ground of the linear regression model to help the reader understand the further theory. A lot of this is well described in Koop (2008) chapter 8, so if nothing else is stated, that is the source for section 4.1.

Regression is a very important econometric tool used to understand the relationship between variables, and is especially useful for an analysis of many variables with complex interactions. On the most basic level, a linear regression

model describes the linear relationship between two variables, X and Y:

$$Y = \alpha + \beta X + \varepsilon, \tag{2}$$

where Y is our response variable, X is our explanatory variable, α is the intercept and β the slope. ε, in turn, is the error, and is added because the regression line $Y = \alpha + \beta X$ is very likely missing some important variables that may affect Y (Koop 2008, pp. 31-35). This regression model is chosen to be the best-fitting line to our observations, or our 'individuals', as we will refer to them. In order to explain what the best-fitting line is, we must start by introducing the residual. A residual is the difference between the actual value and its estimate. If we rearrange equation (2) and adjust it to be just one individual i we find that

$$\varepsilon_i = Y_i - \alpha - \beta X_i.$$

If we now replace α and β with their estimates (denoted by $\hat{\alpha}$ and $\hat{\beta}$) we get

$$\hat{\varepsilon}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i,$$

where $\hat{\varepsilon}$ is the residual. The values from this equation give a straight line that deviates slightly from the true values. What we call the best-fitting regression line is the one that deviates the least from the actual values, or the one that has the smallest residuals according to some criteria. There are different versions of this criteria, but the most common is the sum of squared residuals (SSR):

$$SSR = \sum_{i=1}^{N} \hat{\varepsilon}_i^2,$$

and the best-fitting line is the one with the $\hat{\alpha}$ and $\hat{\beta}$ that give the smallest SSR. These are referred to as the *ordinary least squares* (OLS) estimates, which are the ones we will use in this analysis. There are certain assumptions for the residuals that need to be fulfilled in order for OLS to be the best estimation for a certain model. These assumptions - and how to test them - will be adressed in section 4.2. If we conclude OLS to be the best estimator, we say that OLS is BLUE (best, linear, unbiased estimator) (Koop 2008, p. 72). If the assumptions are not fulfilled, OLS is not BLUE, which means that another estimation theory could possibly give better results. (The inclined reader can dive deeper into OLS estimation for multiple regression in Koop 2008, ch. 2.3.1.)

4.2 Residuals and model evaluation

As mentioned in 4.1, we have to make different assumptions about our model's residuals. Testing residuals is an essential part of the model-building process. As mentioned above, if the assumption 1-4 below are not fulfilled, OLS is not BLUE, and we cannot really assume that any tests (including the ones that we will use) are accurate. Therefore, one must know these possible specification errors, and know how to test a model for them (Andersson et al. 2019, p. 63).

Hyndman (2018, ch 3.3) and Koop (2008, p. 66) write as follows:

1. The residuals are uncorrelated. If there are correlations between residuals, then there is information left in the residuals which should be used in computing forecasts ($\text{cov}(\hat{\varepsilon}_{it}, \hat{\varepsilon}_{jt}) = 0$ for $i \neq j$)

2. The residuals have zero mean. If the residuals have a mean other than zero, then the forecasts are biased ($\text{E}(\hat{\varepsilon}_{it}) = 0$)

3. The residuals have constant variance ($\text{var}(\hat{\varepsilon}_{it}) = \text{E}(\hat{\varepsilon}_{it}^2) = \sigma^2$) (homoscedasticity)

4. The residuals are normally distributed

In Koop, these assumptions will be found without index t. We have adapted them to panel data by adding the time aspect. More on why panel data needs this in section 4.3.

The assumptions 1-4 are built for a normal linear regression model. However, we will work with panel data models. As we will see in section 4.3, the pooled model and the fixed effects model are really just normal regression models, except that the fixed effects model has a new α for every individual. Because of this, the classical assumptions 1-4 are still the ones to take into account for both of these models (Koop 2008, p. 260). What goes for the random effects model will be discussed in section 4.3.3.

As mentioned in section 4.1, assumptions 1-4 need to be tested in order to determine if OLS is BLUE for a certain model. This is based on the *Gauss-Markov Theorem* (Koop 2008, p. 72). Accoring to Koop (2008), the proof of this theorem uses the first 3 assumptions, and a fifth one (X_i is not a random variable), but not assumption number 4. So if 1-3 and 5 are true, OLS is BLUE even if the residuals are not normally distributed. However, we will still test assumption 4 as well. As for the fifth assumption, we cannot test whether out $X : s$ are random variables. We know that they are observable and therefore we can estimate them and will consider the fifth assumption to be true.

In the following sections, we will go through how to test assumptions 1-4, and these will be checked for the fixed effects model and the pooled model. Assumptions 1-3 are not easily checked for panel data models (panel data models will be covered in section 4.3). Theory of how to check them for panel data models exists, but it is mostly very complicated. Therefore, we will analyse these graphically, and draw conclusions from that.

4.2.1 Uncorrelated residuals

Correlation in residuals is not easily checked in a plot. One possibility is plotting the residuals against time to check for patterns. This is not a statistical test, but since correlation means that the residual for a year t is affected by the residual for the year before, $t - 1$, this plot gives us an indication of possible

correlation. If there is no pattern, we have no indication of correlation. If there is a pattern, there is a possibility of some type of correlation.

There are ways to determine correlation via tests. However, these tests are not usually made for panel data, but rather for time series data. One example is the Breusch-Godfrey test (see Koop 2008, ch. 5.4.3). Some of these tests can be modified to fit panel data, but the theory is complicated. Therefore, we will analyse this using a residuals vs time plot.

4.2.2 Zero mean

For panel data, there is no statistical test to use in order to analyse whether the residuals have zero mean. We will therefore calculate a moving average over time. This will be calculated so that we have one residual average per year, over all municipalities. The moving average will be plotted in the same residuals vs time plot that is used in section 4.2.1, and analysed. If the values are close enough to 0, we will consider our residuals to have a zero mean. This will give us an indication of whether our model fulfills this assumption, but it is important to note that it is not actually a proper test.

4.2.3 Constant variance (homoscedasticity)

Homoscedasticity, or the constant variance of the residuals, is often checked using a residuals vs fitted plot (Andersson et al. 2019, p. 68). This is exactly what it sounds like, residuals are plotted against fitted values (estimates). If the plot is an even band centered around 0, with no particular pattern, constant variance can be assumed. If not, the variance can not be assumed to be constant. This plot can also come in other forms, for example residuals vs observed values, and is then analysed in the same way as the residuals vs fitted plot.

4.2.4 Normal distribution

Whether the residuals are normally distributed is easiest checked through a Normal Quantile plot, or Q-Q plot. In a Q-Q plot, you plot the ordered residuals against $\gamma_i = \Phi^{-1}[(i - \frac{3}{8})/(N + \frac{1}{4})]$, where Φ is the standard normal cumulative distribution function (Andersson et al. 2019, p. 69). This analyses the residuals against the theoretical normal quantiles. If the plot looks more or less like a straight line, the residuals can be assumed to be normally distributed. If it looks like a straight line with 'tails' (or more like an 's'-shape) it is not completely normally distributed.

4.2.5 Linearity

Something else that is often tested in multiple linear regressions is linearity. According to Andersson et al. (2019, p. 64), if the relationship between response and explanatory variables is not linear, using OLS is "unsatisfactory".

According to Andersson et al. (2019, p. 64), if the relationship is not linear "parameter estimates are not only biased but also without any meaning". James et al. (2013, p. 92) suggests that a useful tool for checking linearity graphically is using the same residuals vs fitted plots we are already using to check homoscedasticity. We will add a smooth fit to these plots, which will help us identify trends. If these smooth fits are more or less straight lines, linearity can be assumed. If not, steps can be taken to improve linearity. For example, non-linear transformations of the predictive variables (such as logX and X^2) can be used.

4.3 Building the models

Panel data is gaining popularity in empirical research (Woolridge 2013, p. 448). Panel data is data that has both cross-sectional and time series aspects (Koop 2008, p. 255). Cross-sectional data is data collected over many units (for example a number of different companies), and time series data is data collected from the same individual many times (for example a specific company at different times) (Koop 2008, pp. 2-3). The unit in cross-sectional data is indicated using an index i, and the time unit in time series data is indicated using an index t. In panel data, both indices are used and the notation is Y_{it}.

For a panel data analysis, the data set is supposed to be collected by first randomly selecting the groups from which to collect it, and then collecting data on the same variables a number of times over a certain period (Woolridge 2013, p. 448). The data set used here was not collected specifically for this analysis, and therefore was not necessarily collected according to this method. However, it seems to be collected in an appropriate manner. The data is collected from a certain group (people living in Sweden), and the same variables have been collected from this group once a year. The collection of data also follows the same 'individuals', or municipalities in our case. Within these individuals, things probably change over time as people move, die and are born, but that is all encased in the time aspect of this investigation. From what we can gather, the data set was collected in line with panel data methods, and should therefore be possible to use for a panel data investigation.

There are a number of different panel data models, and they are all variants of the classic linear regression model (Andersson et al. 2019, p. 121):

$$Y_i = \beta X_i + \varepsilon_i,$$

where

$$Y_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{iT} \end{bmatrix}, \quad X_i = \begin{bmatrix} 1 & X_{i1}^{(2)} & \cdots & X_{i1}^{(k)} \\ 1 & X_{i2}^{(2)} & \cdots & X_{i2}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{iT}^{(2)} & \cdots & X_{iT}^{(k)} \end{bmatrix}, \quad \varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix},$$

β is a vector of size $k \times 1$, and i is the ith individual. This means that all individuals i are observed T times. The three main models for panel data - pooled OLS regression, fixed effects and random effects - all fall in to the category of variants of the classic linear regression model. These are the ones we will have a closer look at. We will now discuss them one at a time.

4.3.1 Pooled OLS regression

Pooled OLS regression is the most basic method of panel data, but it more or less ignores the panel structure of the data. This means that it does not distinguish between different individuals, and therefore does not give any indication of the difference between them (Andersson et al. 2019, pp. 121-122). As we are interested in the difference between individuals (one of the reasons we are using panel data), pooled OLS regression might not be the appropriate choice of model.

To explain why pooled regression is unsuitable for certain types of analyses, the following example is used in Koop (2008, pp. 256-257). Imagine we have two variables: Y for income and X for years of education. A pooled regression in this form

$$Y_{it} = \alpha + \beta X_{it} + \varepsilon_{it},$$

(also recognized as a standard regression model), only works if all individuals have the same relationship between our two variables, in this case income and years of education. For example, the intercept (α) describes the starting point for an individual's earnings. This will, more often than not, be different for different individuals. Because of this, the pooled regression might not be the best choice. However, this study will still build a pooled model, to use as a comparison to the other models we build.

As indicated by the name, in pooled OLS regression OLS is BLUE given that $\hat{\varepsilon}_{it}$ satisfies the classical assumptions (see section 4.2) (Koop 2008, p. 256). In other words, it is assumed that $\hat{\varepsilon}_{it}$ are independent, identically distributed, have a zero expected value and a variance σ^2.

Now, we will look at fixed effects and random effects models. Both of these are so called 'individual effect models', which are models with an intercept that varies across the individuals and is denoted α_i. This gives different individuals the possibility to have different starting points (Koop 2008, ch. 8.3).

4.3.2 The fixed effects model

In the fixed effects model, dummy variables are used to model the individual effect. A dummy variable is either 0 or 1, and is often used to represent an individual's gender, having 1 represent females and 0 represent males. When used in the fixed effects model, the dummy variables are not explanatory variabels, but are used to decide which intercept should be used. These dummies have the

quality that $D_{ij} = 1$ for the jth individual (and $j = 2, ..., N$, where N is the number of individuals), and 0 for all other individuals (Koop 2008, p. 260). In other words, if $i = j$, $D_{ij} = 1$, and if $i \neq j$, $D_{ij} = 0$. The fixed effects regression looks like

$$Y_{it} = \alpha_1 + \alpha_2 D_{i2} + \cdots + \alpha_N D_{iN} + \beta X_{it} + \varepsilon_{it},$$

where the vector β is the same for all individuals. To avoid multicollinearity, we use one dummy variable less than the number of individuals (Andersson et al. 2019, p. 123), and individual 1 is here our reference unit. The reference unit was chosen simply because it is the first one. However, the reference unit could have been any of our individuals. The individuals are ordered after Statistics Sweden's numbering of the municipalities, where each municipality has a four-digit-number. These (as they are at the time of writing) can be found in the appendix, in A.6.

As with the pooled model, if $\hat{\varepsilon}_{it}$ satisfies the classical assumptions, OLS is BLUE for the fixed effects model (see section 4.2) (Koop 2008, p. 260). That means the same assumptions as in the pooled model are applied to $\hat{\varepsilon}_{it}$.

The fixed effects model often leads to a regression with many explanatory variables. If the dataset has N individuals, T years and k explanatory variables, the fixed effects model will have $N + k$ coeffeicents to estimate, something that might be a problem if N is very large, which it often is (Koop 2008, ch. 8.3.1). Whether this is a problem in our analysis will be discussed in section 6.

4.3.3 The random effects model

We have now covered a very simple panel data model (pooled OLS regression) and a slightly more complicated one (fixed effects model). On the even more complicated side of the spectrum, we have the random effects model. The random effects model does not use dummy variables to model the individual effect, but rather a random variable. This means that the many estimations needed in order to build a fixed effects model disappear in use of a random effects model, and if many estimations are necessary, a random effects model might be less work.

The random effects model explains the individual effect with a random variable, and is written

$$Y_{it} = \alpha_i + \beta X_{it} + u_{it}, \tag{3}$$

where

$$\alpha_i = \alpha + v_i.$$

v_i is a random variable, which means that α_i, the individual effect, is a random variable (Koop 2008, p. 263), and α is a constant. In equation (3), the error is denoted u_{it} rather than the usual ε_{it}. If one wants to write the random effects model in the form of a general regression model, it is possible to rewrite it as

$$Y_{it} = \alpha + \beta X_{it} + \varepsilon_{it},$$

where

$$\varepsilon_{it} = v_i + u_{it}.$$

This second set of equations is an alternative way of writing the first set of equations. Koop (2008, p. 263) assumes that u_{it} and v_i both satisfy the classical assumptions. This means that we assume that they "are independent random variables each with a $\mathrm{N}(0, \sigma_u^2)$ distribution", and the same for v_i, with $\mathrm{var}(u_{it}) = \sigma_u^2$ being the variance for u (and corresponding for v_i). We also assume that u_{it} and v_i are uncorrelated with each other. Other than that, the assumptions for the residuals are slightly different from the pooled model and the fixed effects model. Koop (2008, p. 263) lists the following properties:

$$\begin{aligned}
\mathrm{E}(\hat{\varepsilon}_{it}) &= 0 \\
\mathrm{var}(\hat{\varepsilon}_{it}) &= \sigma_u^2 + \sigma_v^2 \\
\mathrm{cov}(\hat{\varepsilon}_{it}, \hat{\varepsilon}_{jt}) &= 0, \quad \text{for} \quad i \neq j \\
\mathrm{cov}(\hat{\varepsilon}_{it}, \hat{\varepsilon}_{js}) &= 0, \quad \text{for} \quad i \neq j \quad \text{and} \quad s \neq t \\
\mathrm{cov}(\hat{\varepsilon}_{it}, \hat{\varepsilon}_{is}) &= \sigma_v^2 \quad \text{for} \quad s \neq t
\end{aligned}$$

As discussed above, beacuse this model does not use dummy variables it only has one intercept and k explanatory variables - therefore the problem discussed with the fixed effects model (that there are many coefficients to estimate) does not apply here (Koop 2008, ch. 8.3.2). Because of this, the random effects model might be attractive for a data set with many individuals. However, the random effects model's residuals are a little different from what we are used to, and this adds a more complicated side of this model. In the last of the properties above, it is assumed that two errors for the same individual at two different times are correlated with each other (i.e. $\mathrm{cov}(\varepsilon_{it}, \varepsilon_{is}) \neq 0$) (Koop 2008, p. 264). This means that these errors do not satisfy the classical assumptions and OLS is not BLUE. Instead, GLS (*generalized least squares*) should be used. For this reason, we will stick to comparing the fixed effects model to the pooled model. As for the random effects model, we will build it for the sake of seeing if this more complicated version gives better estimations, but stick with the other models for the bigger analysis. Since we will not be using GLS we will not go into detail about it. The inclined reader can read about it in Koop (2008), ch. 5.

4.3.4 Choosing variables

Usually, when working with regressions, some form of stepwise variable selection is used in order to narrow down the number of variables in the model. However, when working with as few variables as we are, this step is not needed. The reason is that with so few variables, we can easily check them one by one to decide which ones are relevant. When checking if the variables are relevant, we use a Wald significance test, based on the Wald statistic (Held & Bové 2014,

ch. 4.2.4). The general version of the Wald statistic is $H_0 : \beta = \beta_0$ and

$$\frac{(\hat{\beta} - \beta_0)}{\text{se}(\hat{\beta})} \overset{a}{\sim} N(0, 1),$$

where se is the standard error.

A significance test is a way to test a hypothesis H_0 against a hypothesis H_1 (Held & Bové 2014, ch. 3.3). When using the Wald test to choose variables, H_0 is $\beta_0 = 0$ (and should therefore not be used), and H_1 is $\beta_0 \neq 0$ (and should therefore be used). A low p-value (usually determined to be under 0.05) means a rejection of the null hypothesis, and in this case means we should keep the variable. The lower the p-value, the more significant the variable is in the regression. All variables significant according to the Wald test will be used in the respective models. We will include the random effects model in this part of the analysis, even though we can not know if the results are accurate.

4.4 Comparing the models

After having checked, and hopefully confirmed, that the fixed effects and pooled models fulfill the assumptions from section 4.2, we can compare them. We will compare our models in two ways: with the standard error, and with the adjusted R-squared. This section will cover how to compare models in these two ways.

Had we decided to use the random effects model in this study, this section would have contained a test that compared the fixed effects and random effects models. One test that could have been used is the Hausman test. The inclined reader can read about the Hausman test in Koop (2008), pp. 154-156 and pp. 264-266.

4.4.1 Standard error

The standard error represents the average distance between the regression line and the actual values. There is one standard error per individual, and it is calculated as the square root of the variance (Held & Bové 2014, p. 56), and we want it to be as small as possible. The standard error gives a quick overview of the difference between two models. A model that has consistently higher standard errors than another model, most probably makes inferior predictions.

4.4.2 R-squared

The R-squared is a way to measure model fit and is defined as

$$R^2 = 1 - \frac{\Sigma_{it}\hat{\varepsilon}_{it}^2}{\Sigma_{it}(Y_{it} - \overline{Y}_t)^2},$$

(where $\overline{Y}_t$ is the mean of Y_{it} for a certain year) and can, according to Koop (2008, p. 95) "be interpreted as the proportion of the variability of the response variable that can be explained by the explanatory variables". R^2 is, however, not perfect when dealing with multiple regression as adding new explanatory variables always increases the R^2. Because of this, something called the adjusted R^2 (or R^2_{adj}) was introduced. R^2_{adj} is defined as

$$R^2_{adj} = 1 - \frac{s^2}{\frac{1}{N-1}\Sigma_{it}(Y_{it} - \overline{Y}_t)^2},$$

where $s^2 = \frac{\Sigma_{it}\hat{\varepsilon}^2_{it}}{N-k-1}$ is an estimate of the variance σ^2 and $\frac{1}{N-1}\Sigma_{it}(Y_{it} - \overline{Y}_t)^2$ is the sample variance. As we are working with models with multiple explanatory variables, R^2_{adj} will be used in this analysis. The formulas for R^2 and R^2_{adj} above are adjusted to work or panel data. In original form, R^2 and R^2_{adj} only contain the index i and not it. However, we will still calculate these using the sum of squared residuals ($\Sigma\hat{\varepsilon}$), so we replaced the i with it, and added index t to $\overline{Y}$.

For R^2 and R^2_{adj}, 1 is a perfect fit (Koop 2008, p. 37) and 0 is a terrible fit. The closer to 1, the better. Koop gives the following example: if you have Y = cost of production on X = output for 123 electric utility companies, an R^2 of 0.92 means that "92% of the variation in costs across companies can be explained by the variation in output".

4.5 Evaluating the predictions - Cross-validation

Within forecasting with regression methods, there are multiple ways to do what is called cross-validation, a way to evaluate a predictive model (Sundberg 2016, p. 70). Cross-validation is usually done by dividing the data into a traning and a test sample. The former is used to estimate parameters, and the latter is a sample to compare with. This division is often proposed to follow the 80-20 rule (Hastie et al. 2017, ch. 7.10): using 80% of the data for the training sample, and the remaining 20% for the test sample. For simplicity reasons, this study will divide the data into 15 years for the training sample and the last year for the test sample. The reason is that it will be quicker and easier to only look at one time unit when analysing the test data.

Cross-validation for panel data forecasting is not covered by the panel data literature at our disposal. Because of this, the cross-validation theory in section 4.5.1 is based in general regression theory. We have made an attempt at adapting it to panel data models, in order to have something to base our analysis on.

4.5.1 Mean Squared Error of Prediction

MSEP, or Mean Squared Error of Prediction, is a form of cross-validation. It determines the size of the prediction error with the following equation:

$$\text{MSEP} = \frac{1}{N}\sum_{1}^{N}(y_i - \hat{y}_{i,-i})^2,$$

where y_i is the response variable for i, $\hat{y}_{i,-i}$ is the predicted y_i value from data without observation i and N is the number of observations (Sundberg 2016, p. 70). Here, N is used because that is the total number of observations in a 'normal' regression. In panel data, N is the number of individuals, and the total number of observations is NT, where T is the number of years. However, we will only make predictions for the year 2017, and will therefore keep only N. Also, our response variable is not Y_i, but Y_{it}, where $t = 2017$, since that is the year we are predicting. Besides all of this, our prediction is made not by taking out individual values but an entire year. This means that in our case, we use $\hat{y}_{it,-t}$ instead of $\hat{y}_{i,-i}$. We now have:

$$MSEP = \frac{1}{N}\sum_{1}^{N}(y_{it} - \hat{y}_{it,-t})^2, \quad \text{where} \quad t = 2017.$$

The MSEP value is often used to compare prediction models, and the model is better the lower the value. According to Sundberg (2016, p. 70), taking the square root of MSEP, what is called RMSEP, is equivalent to using MSEP. Since they are equivalent, we will be using MSEP to compare our models.

The prediction model will be based on the first 15 years (2002-2016). Using this model, we will then make predictions for the year 2017. Using the predictions for 2017 ($\hat{y}_{it,-t}$) and the actual values for 2017 (y_{it}), we will calculate an MSEP value. This value will tell us if the model makes acceptable predictions. However, no source really informs us what constitutes an acceptable MSEP value. Since the value is based on the difference between the actual value and the predicted value, we want it to be as small as possible. However, how large of a difference is acceptable is up to us to determine, and might generally depend on the size of the actual values, since the MSEP value is based on the difference between the actual value and the predicted value. This means that the size of the MSEP value is connected to how far off the predictions are, and we want them to be as close to the actual values as possible. We can say that the size of the MSEP value is relative. As an example, consider a model built to predict the population in Stockholm. To evaluate the model, you build it based on earlier years and use it to predict the population in June 2020 (which we know is around 1.5 million). In this case, an MSEP of a couple of thousand would be pretty good, and an MSEP of a couple of hundred would be really great. This is because that means the MSEP value is very low compared to the actual value of around 1.5 million.

In our case however, the actual values are between -1 and 6 (see table 1), and an MSEP of 100, 10, or even 1, would be bad. However, around 0.1 and 0.01, we might start viewing it as an MSEP value indicating a reliable model. This shows us that an MSEP value can be 1000 and only be around 1% of the actual value, and indicate a reliable model. However, the MSEP value can also be 1 and be 100% of the actual value, and indicate a non-reliable model.

5 Results

5.1 National average income difference

To describe how the national average wage gap has changed between 2002 and 2017 (and answer the first research question), we decided to plot the national average income difference in the years 2002-2017 (figure 1). Since the income is in price base amount, it is adjusted for the inflation. It shows a clear increase in difference over these years. Therefore, the answer to our first research question is that the average wage gap has increased between 2002 and 2017.

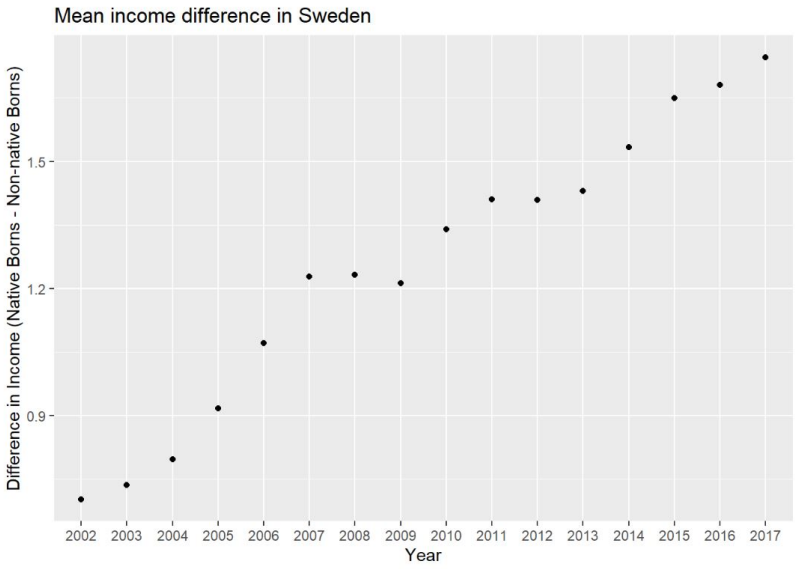

Figure 1: The mean income difference between native borns and non-native borns 2002-2017, in price base amount.

This increase can at least in part, according to Statistics Sweden, be explained by the strong economic growth in Sweden between 2006 and 2008 (Statistics Sweden, *Vanligare med låg ekonomisk standard bland utrikes födda*). A similar

trend can be found between 1996 and 1998, a period that was characterized by strong economic growth. During booms, incomes tend to grow in all social strata, but the incomes grow more in higher strata than lower ones, increasing the income differences. As you can see in figure 1, the income difference grew before the economic boom that started in 2006, as well as long after its end in 2008. This indicates a general increase in income difference in Sweden.

5.1.1 A simple regression model

To demonstrate a simple regression model, we will make one out of the data for figure 1. It looks like this:

$$Y_i = -137.4 + 0.069X_i + \varepsilon_i,$$

where X_i is the year, and Y_i is the income difference said year. The full model summary can be found in A.1, which also tells us that the R^2_{adj} is over 0.96, which is very good. This means that our model only has a 4% margin of error, and shows how much simpler a simple regression model is compared to a multiple regression model.

This model only has one β, since we are only working with one explanatory variable. This β is 0.069, meaning that when one year goes by, the mean income difference increases by 0.069 price base amount. The model also has one α, at -137.4. Looking at A.1, we can see that for the β, the standard error was 3.448×10^{-03}, and for α, the standard error was 6.928. We have not tested our residuals for this model, and will not go in to detail about it. Because of that, we do not really know if our conclusion about the wage gap from figure 1 actually holds any truth. We are mostly using this to compare a simple regression model to our multiple regression models. In order to actually base anything on a model like this, more extensive testing would be needed.

		Pooled	Fixed effects	Random effects
	Intercept	-2.6116	See figure 6	-3.4306
1	Municipal avg age	4.8851×10^{-02}	7.4729×10^{-02}	6.8818×10^{-02}
2	Municipal population	3.9174×10^{-06}	-4.1219×10^{-05}	3.0767×10^{-06}
3	Municipal proportion non-natives	9.2100×10^{-01}	6.7290	4.0766
4	Number of asylum-seekers in Sweden	4.5293×10^{-08}	-2.5388×10^{-07}	-1.1590×10^{-07}
5	Municipal avg income (thousand kronor)	2.0773×10^{-03}	1.2472×10^{-03}	1.6285×10^{-03}
6	Municipal number of people with tertiary education	-9.6378×10^{-06}	5.2800×10^{-05}	-9.1137×10^{-06}

Table 2: Summary of the estimates of our three models

5.2 The contending models

As mentioned in 4.4, we started the process by dividing our data into a training sample (the data for 2002-2016) and a test sample (the data for 2017). Based on the training sample, three panel data models were built. The models' estimates can be found in table 2, and the models look like follows:
Pooled:

$$Y_{i2017} = \alpha + \beta_1 X^{(1)}_{i2017} + \beta_2 X^{(2)}_{i2017} + \beta_3 X^{(3)}_{i2017} + \beta_5 X^{(5)}_{i2017} + \beta_6 X^{(6)}_{i2017} + \varepsilon_{i2017}$$

Fixed effects:

$$Y_{i2017} = \alpha_i + \beta_1 X^{(1)}_{i2017} + \beta_2 X^{(2)}_{i2017} + \beta_3 X^{(3)}_{i2017} + \beta_4 X^{(4)}_{i2017} + \beta_5 X^{(5)}_{i2017} + \beta_6 X^{(6)}_{i2017} + \varepsilon_{i2017}$$

Random effects:

$$Y_{i2017} = \alpha + \beta_1 X^{(1)}_{i2017} + \beta_2 X^{(2)}_{i2017} + \beta_3 X^{(3)}_{i2017} + \beta_5 X^{(5)}_{i2017} + \beta_6 X^{(6)}_{i2017} + \varepsilon_{i2017}$$

Table 2 contains the calculated estimates for all three of our models. A.2, A.3 and A.4 contain the full model summaries, including the results of the Wald significance test for each estimate. The α values for the fixed effects model can be found in A.3, figure 6. The X:s are different for each municipality and year and are too many to cover here. The intervals can be found in table 1, and the exact values can be found in the original data at Statistics Sweden.

Table 2 shows that in almost all cases, the estimates for the same variable k are of similar size, and of the same sign (positive or negative), in all three models. Exceptions are *Municipal population* (and *Municipal number of people with tertiary education*), where the fixed effects model has a negative (positive) estimate

whereas the other two are positive (negative), and *Number of asylumseekers in Sweden*, where the pooled model has a positive estimate whereas the other two models have negative ones. The one variable that is not municipality-specific (*Number of asylumseekers in Sweden*) is the only variable deemed irrelevant (according to the Wald significance test, see section 4.3.4), and this in the pooled model (with a p-value of around 0.76) and the random effects model (with a p-value of around 0.26). This variable was concluded to be relevant (though less so than the others) in the fixed effects model, with a p-value of about 0.01. All other variables were concluded to be relevant in all three models, with the biggest p-value being *Municipal population* in the random effects model, on the size of 10^{-05}. In the other two models, no p-value (except for the ones mentioned) was over the size of 10^{-11}, making all variables very relevant. It might not be very surprising that the only non-municipality-specific variable is the only variable deemed irrelevant in two of the models. This is the least specific variable and the only one that is the same over all municipalities. Important to note is that while we are analysing these p-values like we know they are reliable, they too are effected by whether or not the residual assumptions are true, something we will cover in section 5.3.

Since the difference over a single variable is very little between the models, let us look at the pooled model as an example analysis of the estimates for the β values. Looking at the pooled model's column, we can see that the biggest estimate is the one for municipal proportion non-natives, and the smallest one is for number of asylumseekers in Sweden. We will use these two to discuss estimate sizes. It is easy to assume that the biggest β makes the biggest impact on the response variable. However, the impact is also affected by the size of the explanatory variable itself. Looking at table 1, we can see that the size of municipal proportion non-natives will be $0.02 < X_{it}^{(3)} < 0.05$, while the size number of asylumseekers in Sweden will be $17000 < X_{it}^{(4)} < 170000$. If both of these were to be in some kind of middle (say for example 0.035 and 93 500 respectively), βX would be around 0.032 for $X_{it}^{(3)}$ and around 0.0042 for $X_{it}^{(4)}$. The impact of municipal proportion non-natives is still bigger, but if $X_{it}^{(3)}$ would be at its lowest (0.02) and $X_{it}^{(4)}$ would be at its highest (170000), this might change. This shows that we cannot simply analyse the β-values, they need to be analysed with their corresponding X-values.

Moving on from the β values, we will look at the α values for the fixed effects model in figure 6. Most of them are between 0 and -5, with a couple of exceptions. The most extreme outliers are the highest ones: Stockholm at 15.72394129, Göteborg at 8.95646865 and Malmö at 3.58801080. A high α means that the starting point for difference in salary is high. Interestingly, the three highest α:s are found in Sweden's three largest cities. The lowest α is found in Haparanda at 5.882704, meaning that there, the starting point for difference in salary is the furthest away from Stockholm.

5.3 Residuals - Are the assumptions met?

In this section, we will analyse whether the assumptions discussed in section 4.2 are met by our fixed effects and pooled models. As mentioned in 4.2, we will do this purely visually, by plotting our residuals in different ways. This is not ideal and not as exact as using tests, but will give us a starting point for an analysis.

As was also mentioned in section 4.3.3, the random effects model's residuals will not be analysed. The reason is that we know that the residuals for the random effects model are not uncorrelated, and therefore cannot assume that the tests are accurate. As the reader may have noticed, the Wald significance test and the estimates of the models have already been analysed (see section 5.2), including for the random effects model. These analyses might not be accurate if this section concludes that the assumptions are not met. This also goes for the tests that can be found in appendix A.1, A.2, A.3 and A.4.

Plots for the two models we are analysing (the fixed effects model and the pooled model) are similar in their distribution. Since that is the case, we will show the plots for the fixed effects model in the text, and show the pooled model's plots in the appendix (A.5).

5.3.1 Uncorrelated residuals and Zero mean

This section will analyse the residuals as residuals vs time plots with moving averages. Since we have 290 municipalities and a plot with every one would be very chaotic, we chose to only show a plot containing five of our municipalities. In order to show as accurate of a representation of all 290 municipalities as possible, we chose two municipalities with some extreme residuals, and three municipalities with residuals closely resembling the average residual in our models. The plot for the fixed effects model, with the moving average in it, can be found in figure 2, and the plot for the pooled model can be found in A.5, in figure 7.

Our two residuals vs time plots are very similar. We can see a clear cyclical trend in Danderyd, and an outlier in Bromölla. The rest of the residuals show a slight cyclical trend that was more clear in a more zoomed-out version of the plots than the ones discussed in this essay. When a plot of all 290 municipalities was analysed, it was concluded that this general trend (most lightly cyclical, and some more cyclical) was the same over all municipalities. Most residuals can be found between -0.5 and 0.5, with some exeptions and outliers. Two municipalities with some of the most extreme trends were Danderyd and Lidingö, which both happen to have had among the top 5 highest mean incomes of Swedish municipalities in 2018 (Ekonomifakta, *Din kommun i siffror*). These two do not simply have some outliers, like Bromölla, but have extreme values for all years. Over all, the mostly slight and sometimes more extreme trend we see indicates some form of autocorrelation. This because, as described in section

4.2.1, a pattern indicates that any one year's value is affected by the value of the year before. Therefore, we will assume slight autcorrelation and accept that assumption number 1 (uncorrelated residuals) is not fulfilled.

The moving average is the black line in each plot. This is the average municipal residual every year. We can see that while the average is not the same over the different years, it is always very close to 0. We will take this to be close enough to the zero mean we assume, and we consider the second assumption of a zero mean fulfilled.

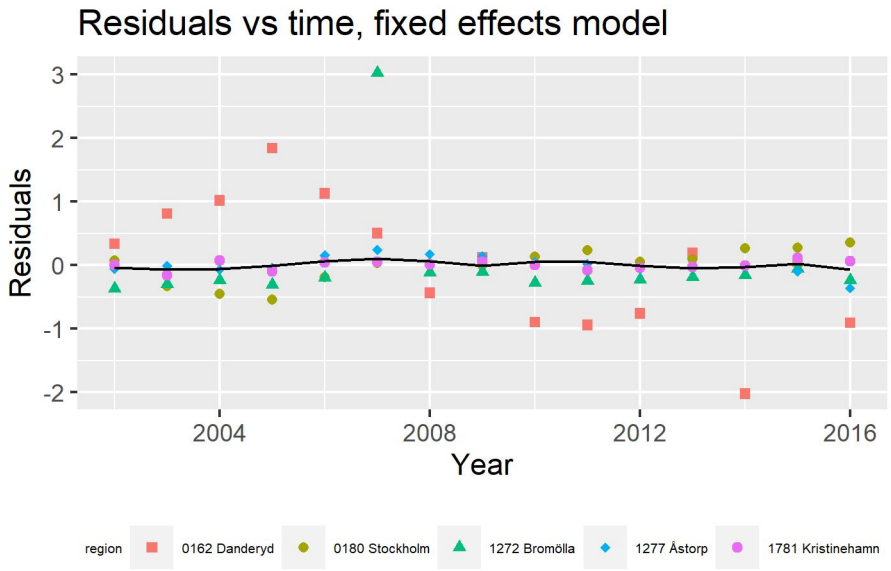

Figure 2: Residuals vs time for the fixed effects model

5.3.2 Residuals vs fitted

Figure 3 shows our residuals vs fitted plot for the fixed effects model, and figure 8 for the pooled model. A residuals vs observed values plot gives more or less the same result, and is therefore not discussed in this essay. Since the x-axis is ordered by fitted values, the municipalities with lower income difference will be at the far left, and the municipalities with higher income difference will be at the far right. We can see that most of the residuals are centered around 0 on the y-axis, which is desired for a plot of this kind. If we only look at the x-axis between 0 and 2.5, it looks like an even band around 0 (except for the one obvious outlier at $y = 3$). Looking beyond $x = 2.5$, we can see that almost all points belong to either Danderyd (orange squares) or Lidingö (blue circles), which we have already concluded have extreme residuals, explaining the more

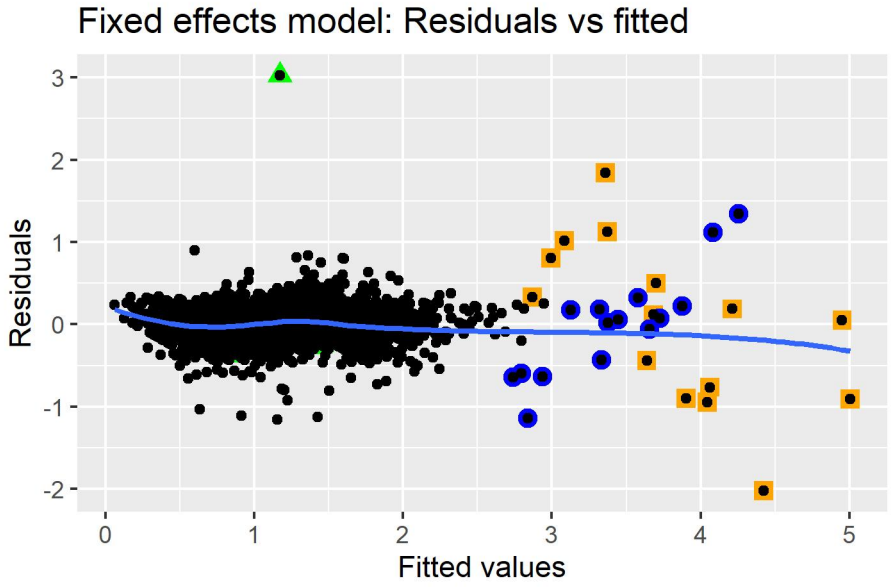

Figure 3: Residuals vs fitted plot for fixed effects model
Orange squares: Danderyd, Blue circles: Lidingö, Green triangles: Bromölla.

varied positioning along the y-axis. If we were to remove these two municipalities, constant variance could be assumed. However, as the plot includes both of these municipalities, a slight cone shape is suggested. This shape usually indicates that the variance increases with time (Andersson et al. 2019, p. 68). However, since our data not only includes time series aspects, but also cross-sectional aspects, it is not as simple for us. In this case, the plots suggest that the variance increases with certain cross-sectional aspects (or certain municipalities). Either way, we can not assume constant variance (and therefore do not consider the third assumption to be fulfilled), but note that granting the removal of two municipalities, constant variance could most likely be achieved.

In both figure 3 and figure 8 (in A.5), we have marked the most extreme outliers in color and shape depending on the municipality. We can note that there are three municipalities represented here: Danderyd, Lidingö and Bromölla. Going back to the residuals vs time plots, we remember that Danderyd had the most extreme cyclical trend, and Bromölla had one outlier, which exaplains their distribution along the y-axis in the residuals vs fitted plots as well. When analysed, the residuals vs time plots for Lidingö (both for fixed effects and the pooled model) show similar trends to the ones we see in Danderyd. This shows us a similarity in the residuals vs time plots compared to the residuals vs fitted plots. In figures 3 and 8, the fact that Danderyd and Lidingö have extreme

values also on the x-axis is explained by the high income difference in both of these municipalities. Bromölla is not as high up on this list, and is therefore more centered with the other points on the x-axis.

The blue line is the smooth fit that represents linearity. It is not a completely straight line, but has no apparent pattern or shape. We will therefore assume linearity in both models, and also note that the removal of Danderyd and Lidingö would most probably improve also the linearity of the models.

5.3.3 Q-Q plot

The Q-Q plot for our fixed effects model can be found in figure 4, and the Q-Q plot for the pooled model can be found in figure 9. They are similar, and both very close to a straight line except for some outliers. The outliers form tails, which means that the distribution is not completely normal. However, these tails do not include many values, and all points above the upper red line

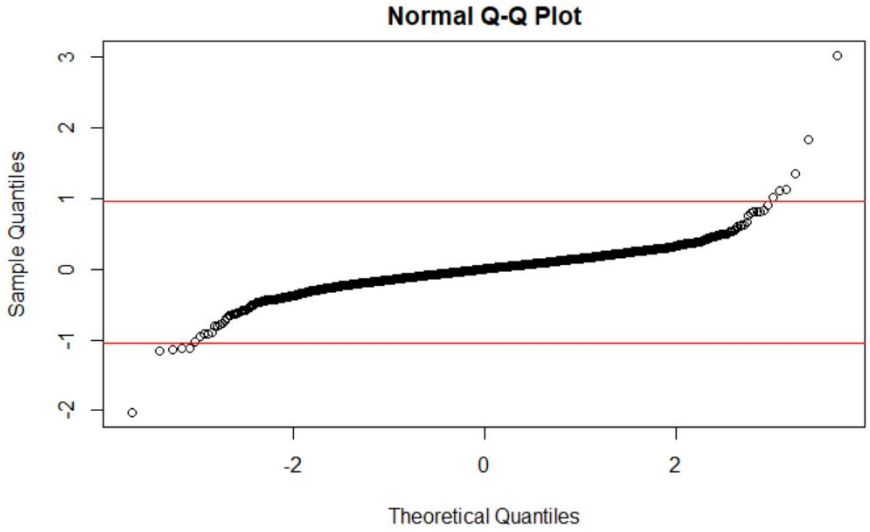

Figure 4: Q-Q plot for the fixed effects model.

(in both fixed effects and pooled) are Danderyd, Lidingö or Bromölla. These are all familiar municipalities by now, as we have already concluded that their residuals seem to be outliers in general. Below the lower red line we have, for fixed effects: Danderyd, Örkelljunga, Lidingö, Höganäs and Strömsund, and for pooled: Danderyd, Ljusnarsberg, Höganäs, Orkelljunga and Haparanda. Some of these are new, but we still have the familiar outliers. The plots show relatively symmetric distributions. Except for the tails, these plots indicate normally distributed residuals, and we will consider the fourth assumption to

		Pooled	Fixed effects	Random effects
	Intercept	1.0426×10^{-01}	—	1.8789×10^{-01}
1	Municipal avg age	2.1878×10^{-03}	7.2259×10^{-03}	4.5653×10^{-03}
2	Municipal population	4.8195×10^{-07}	4.7778×10^{-06}	7.0156×10^{-07}
3	Municipal proportion non-natives	9.6377×10^{-02}	2.5507×10^{-01}	1.9913×10^{-01}
4	Number of asylumseekers in Sweden	1.5093×10^{-07}	9.8707×10^{-08}	1.0189×10^{-07}
5	Municipal avg income (thousand kronor)	3.3115×10^{-05}	5.1453×10^{-05}	4.0367×10^{-05}
6	Municipal number of people with tertiary education	1.4497×10^{-06}	6.8115×10^{-06}	1.8542×10^{-06}

Table 3: Summary of the standard errors of our three models

be true.

5.4 Which model is better?

5.4.1 R squared and standard error

We can see in A.2 and A.3 that the R^2_{adj} for the pooled model was 0.57798 and for the fixed effects model was 0.71234. This shows us that the simplicity of the pooled model does seem to give us a less accurate model than the fixed effects one. The fixed effects model's R^2_{adj} is not perfect, but it is higher than the pooled model. Following Koop's example (see section 4.4.2), the fixed effects model's variation in output explains around 71% of the variation in income difference across municipalities. The same number for the pooled model is around 58%. A model should be able to explain a high percentage of variation, and the fixed effects model is therefore more desirable here. However, we will also check the MSEP values for accuracy of predictions (see section 5.5.1).

In table 3, the standard erorrs for each model can be found. The standard error represents the average distance between the regression line and the actual values, and we want it to be as small as possible. A quick look at table 3 shows us that most rows keep standard errors of about the same size, and there is no systematic difference between the columns - which means that none of the models generally have bigger or smaller standard errors. This means that there is no indication in the standard errors that one model should give better or worse predictions than another.

5.5 Accuracy of predictions

According to R^2_{adj}, the fixed effects model gives more accurate predictions than the pooled model. Now, we will analyse the accuracy of the two models' predictions by comparing MSEP values (section 5.5.1) as well as plotting each models' difference in actual value and predicted value (section 5.5.2).

5.5.1 MSEP

Our MSEP values ended up being 0.1414581 for the pooled model and 0.07969323 for the fixed effects model. Since the model is better the lower the MSEP-value, this would contribute to the conclusion that the fixed effects model is better. Once again, the simplicity of the pooled model seems to make it less accurate. As discussed in section 4.5.1, what is an acceptable MSEP value is up for discussion. The MSEP values here are close to zero, but as concluded in 4.5.1, it is more important whether the value is small in comparison to the actual Y values. The $\overline{Y}_{2017} = 1.745862$, which means that our fixed effects model's MSEP value is around $4,5\%$ off, and our pooled model's MSEP value is around 8% off.

5.5.2 Plotting the predictions

In figure 5, we bring in the random effects model for the sake of comparing its predictions to the fixed effects model's and the pooled model's. The circles are the fixed effects model, the crosses the pooled model, and the dots the random effects model. Each point represent the difference in actual value and predicted value for each of the municipalities. The municipalities are here only referenced by a number, and the list to find the name of the municipality can be found in A.6. There seems to be no apparent difference between the three models when looking at it this way, as all dots are centered around 0, the way we want it. The closer to 0, the better the prediction. However, we do see a couple of outliers. The most extreme ones are found outside of the dashed helplines at 1.5 and -1.5. These are: the random effects model's value for Lidingö at 1.8436374, the fixed effects model's value for Danderyd at -1.7698345, the random effects model's value for Haparanda at -2.1398481, the pooled model's value for Lidingö at 1.5286365, and the pooled model's value for Danderyd at -1.9799257. Once again, Danderyd and Lidingö show up as outliers. Since there is no apparent difference between the three models here, our analysis of this plot is more or less that the models give very similar results. However, we can see that the fixed effects model only has one extreme outlier while both other models have two, and that the two most extreme outliers belong to the random effects model. This gives us slight indication that the fixed effects model may be better, and it seems that the random effects model might be unnecessarily complicated.

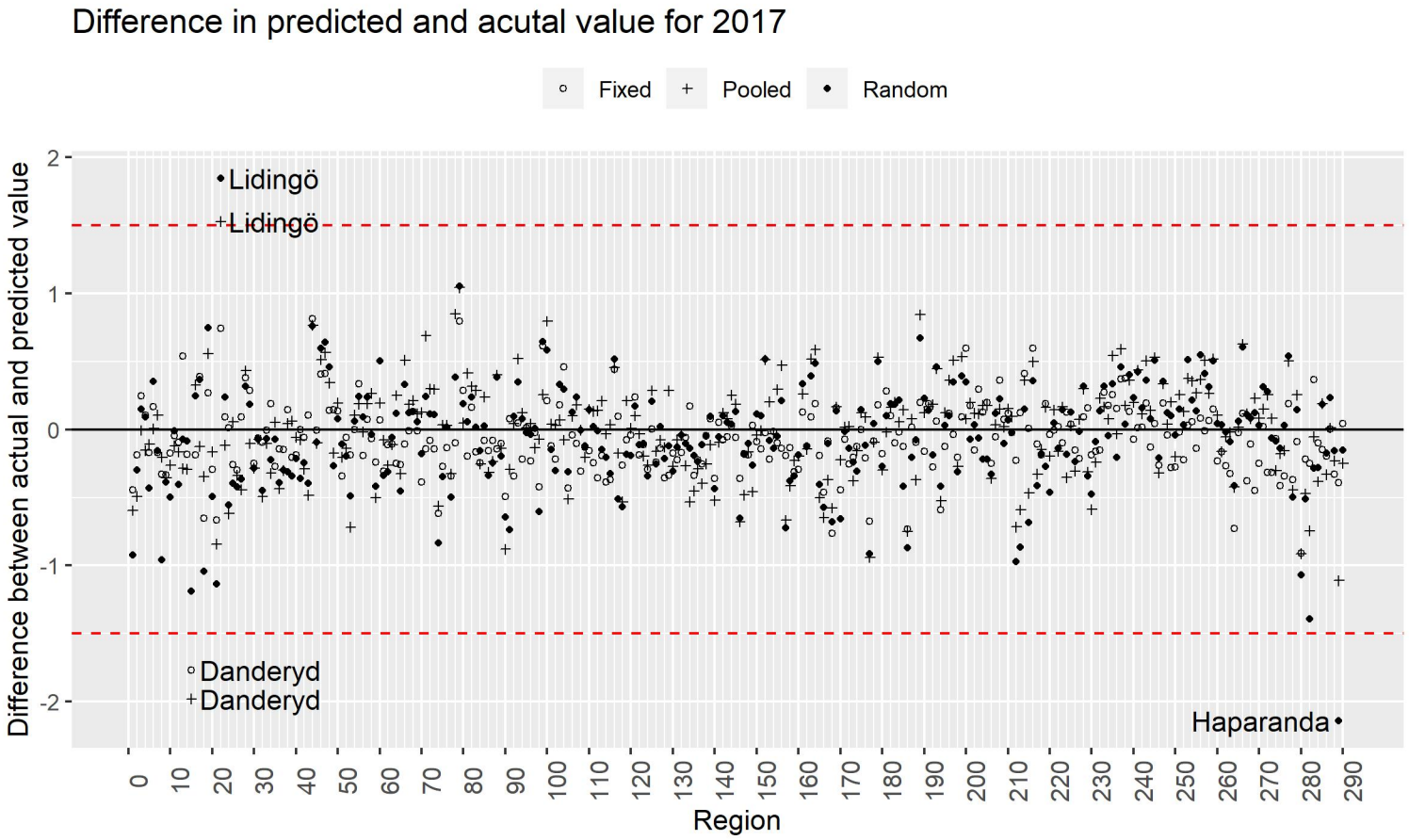

Figure 5: The predictions

6 Discussion

The purpose of this analysis was to investigate the income differences between native borns and non-native borns in different Swedish municipalities in the years 2002-2017, as well as to analyse differences between three types of panel data prediction models. The research questions were:

> How has the national average wage gap between immigrants and native borns changed in Sweden over the years 2002-2017?
>
> Can we make a predictive model for the average wage gap in each Swedish municipality based on the data we have?
>
> Which panel data model makes the best predictions for this data?

These questions were answered by processing and analysing data from Statistics Sweden using panel data methods. The first question has been answered using the mean of the income difference for every year, and for the second and third questions, three predictive models have been built. Now, the results will be discussed.

For the first reseach question, we built a simple regression model to represent the mean income difference in Sweden between the years of 2002-2017. As we plotted the line, it was shown that the income difference had increased, something that could be explained at least in part by the strong economic growth between 2006 and 2008 (see Statistics Sweden, *Vanligare med låg ekonomisk standard bland utrikes födda*). This model also acted as an example of the difference between a simple regression model and a multiple regression model.

Moving on to the second research question, we can conclude that while a predictive panel data model for the average wage gap based on the data we have is not very accurate, all three versions do work and make predictions. We concluded in section 5.3 that for neither the pooled model, nor the fixed effects model, OLS was BLUE, according to Koops (2008) assumptions. However, we still used OLS as our estimator. In the random effects model's case, the residuals were not uncorrelated and we therefore did not go further with the analysis of that model's residuals. In the case of the pooled model and the fixed effects model, the residuals were very similar; slightly correlated but with a zero mean, and lack of constant variance but with a symmetric Q-Q plot with tails, indicating normally distributed residuals with some outliers. In other words: we considered assumptions 2 (zero mean) and 4 (normally distributed residuals) to be true, and assumptions 1 (uncorrelated residuals) and 3 (constant variance) not to be true. The conclusion is that both pooled OLS regression and fixed effects can be used for the predictive model we are seeking, based on the data we have. However, they might not be very accurate and some other type of estimation theory (other than OLS) might be needed. We can not say anything about the random effects model in this case. Yes, it makes predictions close

to the fixed effects model's predictions, but we have not tested it and can not draw these conclusions about it. As for the models we did test, it might have been possible to build a more accurate model. With more time and resources, it would have been easier to understand and use the theory in a more extensive way. The study might also have benefitted from more extensive data.

As for the third research question, the focus will be how complicated a model gives the best predictions, and if more complicated is always better. This because we have realized that the fixed effects model is a more complicated version of the pooled OLS regression, and the random effects model in turn is a more complicated version of the fixed effects model. Here, we did compare the random effects models estimates, p-values and standard errors to the other two models, for a more interesting analysis, even though we cannot know that they are accurate. We can conclude that while the added individual effects aspect made the fixed effects model much better than the pooled model, there is no sign of the added complexity of the random effects model being suitable for this particular analysis. While the many α:s of the fixed effects model give us a lot to calculate (the one downside we could find to the fixed effects model), that is not very hard for a computer to do. We also cannot find many positives about the random effects model that the fixed effects model does not have. Except for two outliers, there was no notable difference in figure 5, showing the subtraction of predicted value - actual value for the three models. However, since we did not analyse the random effects model closely, it is hard to conclude if it has any advantages over the fixed effects model in this case.

With more time, data and knowledge, it would probably be possible to build a better model. It is also possible that it would give a very different result regarding the difference between the fixed effects and random effects models. However, our conclusion is that given the frame we were working in, the fixed effects model gives the best predictions out of the three basic panel data models that we analysed. It did not give great predictions - probably because the variables were not extensive enough - but for this particular analysis, it was the better of the three models. Over all, the conclusion is that a panel data model could help in finding the factors that affect the gap the most. Given more extensive research, it could probably help in minimizing the wage gap between native and non-native borns.

7 References

Andersson, Patrik; Lindensjö, Kristoffer, & Tyrcha, Johanna. (2019). *Notes in Econometrics*. Stockholm University: Department of Mathematics.

Arulampalam, Wiji; Booth, Alison L. & Bryan, Mark L. (2005). Is there a glass ceiling over Europe? Exploring the gender pay gap across the wages distribution, *ISER Working Paper Series*, No. 2005-25, University of Essex, Institute for So-

cial and Economic Research (ISER), Colchester. http://hdl.handle.net/10419/92046 (collected 2020-04-27)

Blau, Francine & Kahn, Lawrence. (2003). Understanding International Differences in the Gender Pay Gap. *Journal of Labor Economics*, No. 21. 106-144. doi: 10.1086/344125. (collected 2020-04-27)

Edström, Josefin & Pokarzhevskaya, Galina. (2017). *Lönegap mellan akademiker med svensk och utländsk bakgrund En analys av Sacoförbundens medlemmar.* https://www.saco.se/opinion/rapporter/lonegap-mellan-akademiker-med-svensk-och-utlandsk-bakgrund/ (collected 2020-04-27)

Ekonomifakta. *Din kommun i siffror.* https://www.ekonomifakta.se/Fakta/Regional-statistik/Din-kommun-i-siffror/ (collected 2020-06-15)

European Commission DG Migration and Home Affairs (2020). Glossary. https://ec.europa.eu/home-affairs/e-library/glossary/host-country_en (collected 2020-07-02)

Expressen. (2017). Vi har inte råd med diskriminering på jobbet. *Expressen.* 13th July. https://www.expressen.se/ledare/vi-har-inte-rad-med-diskriminering-pa-jobbet/ (collected 2020-03-18)

Hastie, Trevor; Tibshirani, Robert, & Friedman, Jerome. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* 2nd edition. Springer Science+Business Media.

Held, Leonard, & Bové, Daniel Sabanés. (2014). *Applied Statistical Inference: Likelihood and Bayes.* Berlin, Heidelberg: Springer Berlin Heidelberg.

Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*, 2nd edition. Melbourne, Australia: OTexts. OTexts.com/fpp2. Accessed on 18th march.

James, Gareth; Witten, Daniela; Hastie, Trevor, & Tibshirani, Robert. (2013). *An Introduction to Statistical Learning with Applications in R.* Springer Science+Business Media.

Katz, Katarina & Österberg, Torun. (2013). Invandrade med svensk utbildning har lägre lön än infödda. *Institutet för Arbetsmarknads- och Utbildningspolistisk värdering.* 19th April. https://www.ifau.se/sv/Press/Meddelanden/Invandrade-med-svensk-utbildning-har-lagre-lon-an-infodda/ (collected 2020-03-18)

Koop, Gary. (2008). *Introduction to Econometrics.* Chichester: Wiley, cop.

Nationalencyklopedin. Basbelopp.
http://www.ne.se.ezp.sub.su.se/uppslagsverk/encyklopedi/lång/basbelopp

(collected 2020-06-04)

Olsson, Annika. (2017). Utländsk bakgrund ger akademiker lägre lön. *Arbetet.* 12th July. https://arbetet.se/2017/07/12/utlandsk-bakgrund-ger-akademiker-lagre-lon/ (collected 2020-03-18)

SFS 2010:110. *Prisbasbelopp.*

Statistics Sweden (SCB). *Statistical Database.* http://www.statistikdatabasen.scb.se/pxweb/en/ssd/ (collected 2020-03-03)

Statistics Sweden (SCB). (2017). *Vanligare med låg ekonomisk standard bland utrikes födda.* https://www.scb.se/hitta-statistik/artiklar/2017/Vanligare-med-lag-ekonomisk-standard-bland-utrikes-fodda/ (collected 2020-05-04)

Sundberg, Rolf. (2016). *Lineära Statistiska Modeller.* Stockholm University: Department of Mathematics.

Woolridge, Jeffrey M. (2013). *Introductory Econometrics: A Modern Approach,* 5th edition. South-Western: Cengage Learning.

A Appendix A

A.1 Simple regression model

	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	-1.374e+02	6.928	-19.83	1.20e-11
Year	6.900e-02	3.448e-03	20.02	1.06e-11

Multiple R-Squared: 0.9662
Adj. R-Squared: 0.9638

Table 4: Summary results for simple regression model.

A.2 Pooled model

	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	-2.6116	1.0426e-01	-25.04825	< 2.2e-16
Municipal avg age	4.8851e-02	2.1878e-03	22.3283	< 2.2e-16
Municipal population	3.9174e-06	4.8195e-07	8.1282	5.631e-16
Municipal proportion non-natives	9.2100e-01	9.6377e-02	9.5562	< 2.2e-16
Number of asylumseekers in Sweden	4.5293e-08	1.5093e-07	0.3001	0.7641
Municipal avg income (thousand kronor)	2.0773e-03	3.3115e-05	62.7288	< 2.2e-16
Municipal number of people with tertiary education	-9.6378e-06	1.4497e-06	-6.6482	3.334e-11

R-Squared: 0.57857
Adj. R-Squared: 0.57798

Table 5: Summary results for pooled model.

A.3 Fixed effects model

	Estimate	Std. Error	t-Value	Pr(>|t|)
Municipal avg age	7.4729e-02	7.2259e-03	10.3419	< 2.2e-16
Municipal population	-4.1219e-05	4.7778e-06	-8.6271	< 2.2e-16
Municipal proportion non-natives	6.7290	2.5507e-01	26.3806	< 2.2e-16
Number of asylumseekers in Sweden	-2.5388e-07	9.8707e-08	-2.5721	0.01014
Municipal avg income (thousand kronor)	1.2472e-03	5.1453e-05	24.2403	< 2.2e-16
Municipal number of people with tertiary education	5.2800e-05	6.8115e-06	7.7516	1.139e-14

R-Squared: 0.73185
Adj. R-Squared: 0.71234

Table 6: Summary results for fixed effects model.

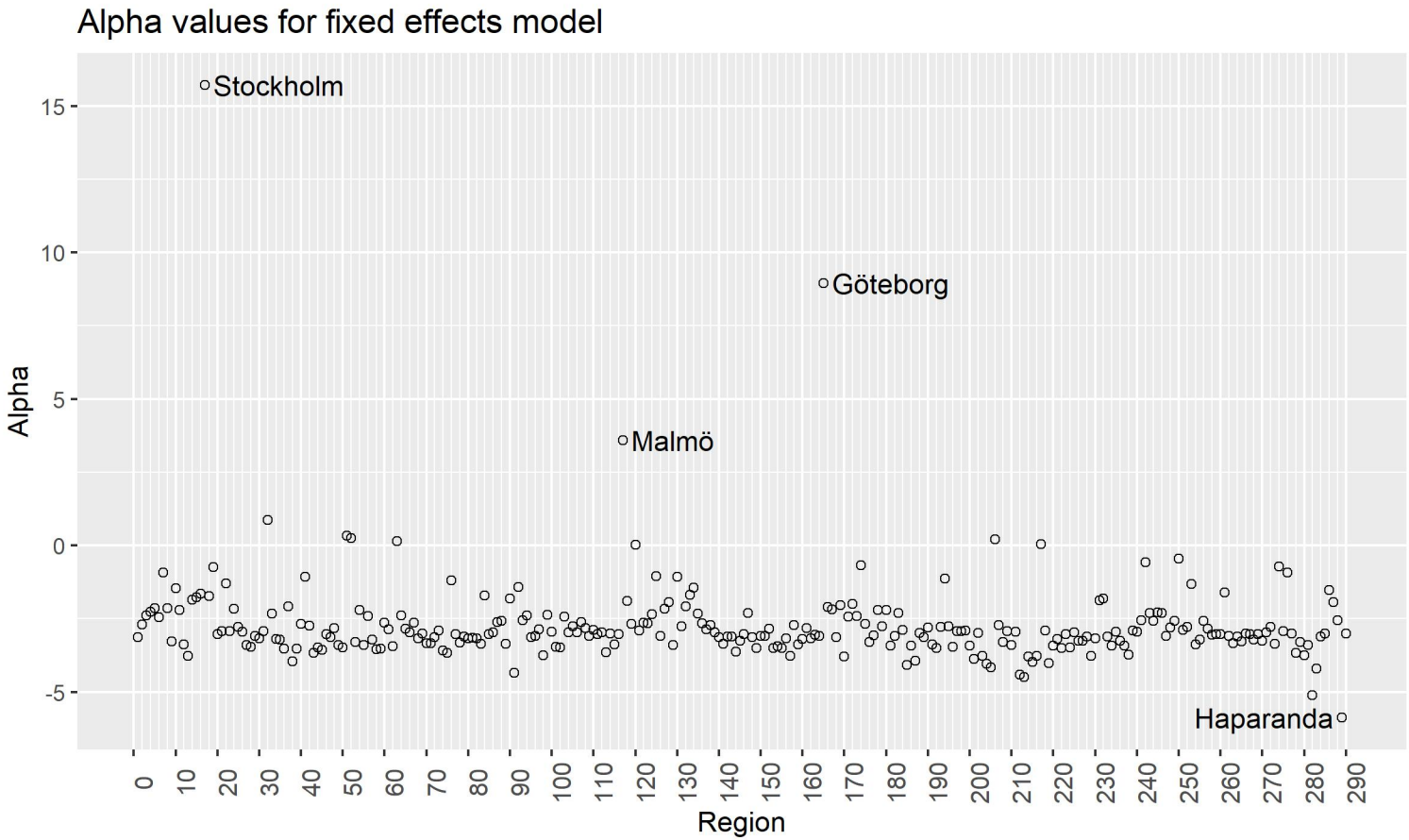

Figure 6: The alpha-values for our fixed effects predictive model

A.4 Random effects model

	Estimate	Std. Error	z-value	Pr(>|z|)
Intercept	-3.4306	1.8789e-01	-18.2585	< 2.2e-16
Municipal avg age	6.8818e-02	4.5653e-03	15.0741	< 2.2e-16
Municipal population	3.0767e-06	7.0156e-07	4.3855	1.157e-05
Municipal proportion non-natives	4.0766	1.9913e-01	20.4724	< 2.2e-16
Number of asylumseekers in Sweden	-1.1590e-07	1.0189e-07	-1.1375	0.2553
Municipal avg income (thousand kronor)	1.6285e-03	4.0367e-05	40.3431	< 2.2e-16
Municipal number of people with tertiary education	-9.1137e-06	1.8542e-06	-4.9153	8.865e-07
R-Squared: 0.69937 Adj. R-Squared: 0.69896				

Table 7: Summary results for random effects model.

A.5 The pooled model's residuals

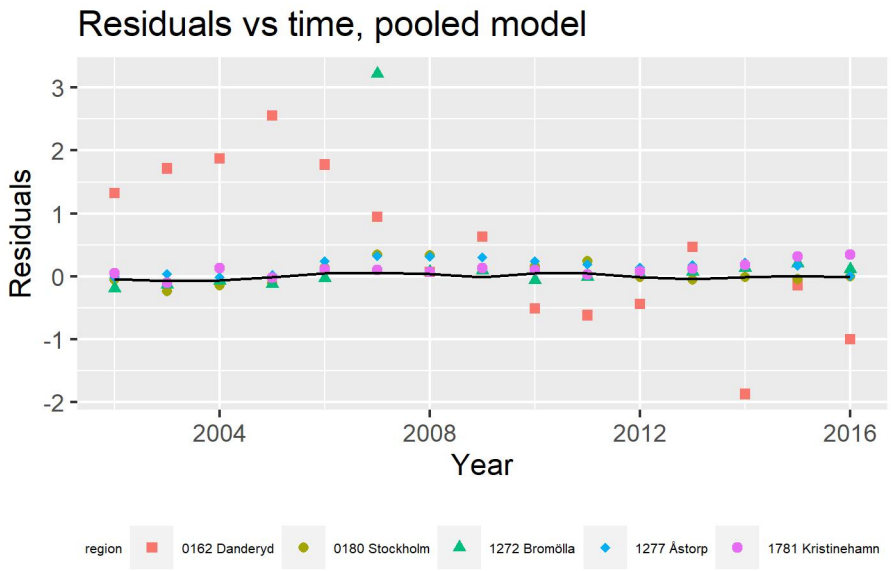

Figure 7: Residuals vs time for the pooled model

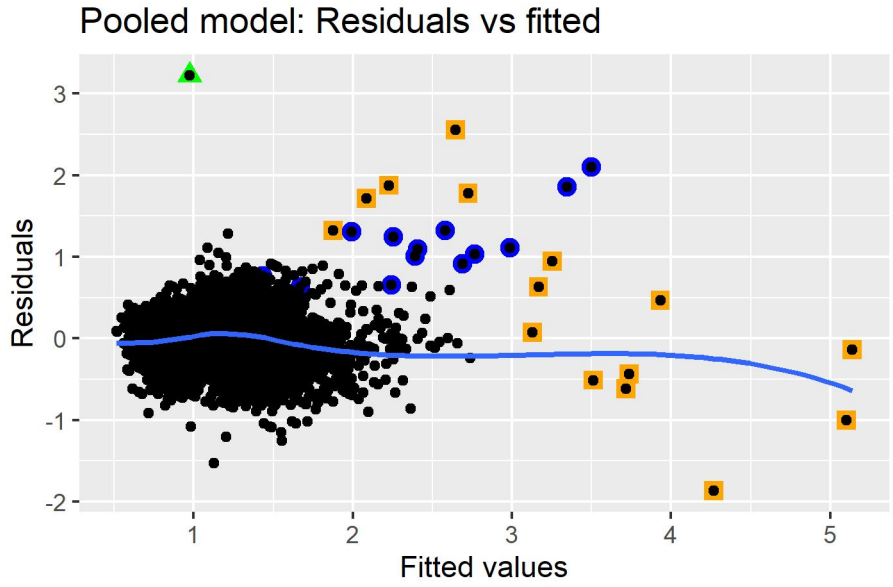

Figure 8: Residuals vs fitted plot for the pooled model
Orange squares: Danderyd, Blue circles: Lidingö, Green triangles: Bromölla.

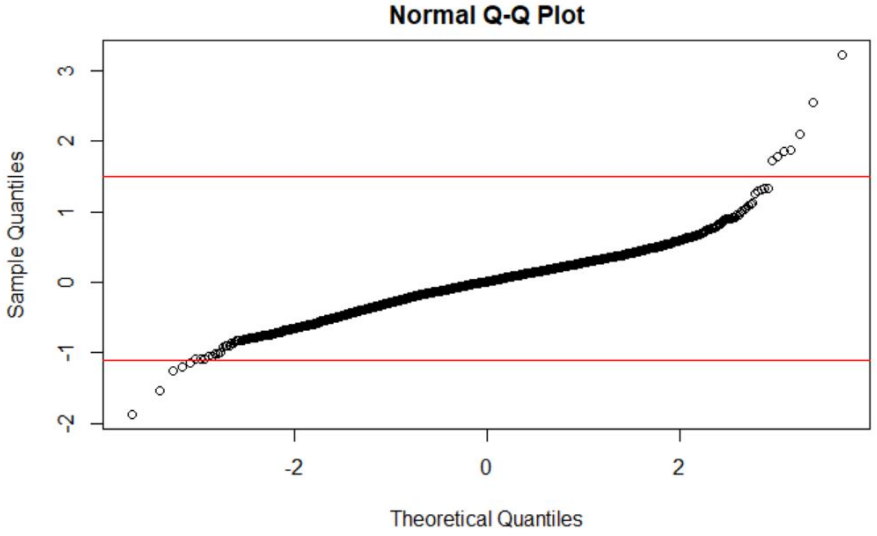

Figure 9: Q-Q plot for the pooled model

A.6 List of municipalities

Table 8: List of municipalities.

Beginning of Table	
Number	Municipality
1	0114 Upplands Väsby
2	0115 Vallentuna
3	0117 Österåker
4	0120 Värmdö
5	0123 Järfälla
6	0125 Ekerö
7	0126 Huddinge
8	0127 Botkyrka
9	0128 Salem
10	0136 Haninge
11	0138 Tyresö
12	0139 Upplands-Bro
13	0140 Nykvarn
14	0160 Täby
15	0162 Danderyd
16	0163 Sollentuna
17	0180 Stockholm
18	0181 Södertälje
19	0182 Nacka
20	0183 Sundbyberg
21	0184 Solna
22	0186 Lidingö
23	0187 Vaxholm
24	0188 Norrtälje
25	0191 Sigtuna
26	0192 Nynäshamn
27	0305 Håbo
28	0319 Älvkarleby
29	0330 Knivsta
30	0331 Heby
31	0360 Tierp
32	0380 Uppsala
33	0381 Enköping
34	0382 Östhammar
35	0428 Vingåker
36	0461 Gnesta
37	0480 Nyköping
38	0481 Oxelösund
39	0482 Flen

Continuation of Table 8	
Number	Municipality
40	0483 Katrineholm
41	0484 Eskilstuna
42	0486 Strängnäs
43	0488 Trosa
44	0509 Ödeshög
45	0512 Ydre
46	0513 Kinda
47	0560 Boxholm
48	0561 Åtvidaberg
49	0562 Finspång
50	0563 Valdemarsvik
51	0580 Linköping
52	0581 Norrköping
53	0582 Söderköping
54	0583 Motala
55	0584 Vadstena
56	0586 Mjölby
57	0604 Aneby
58	0617 Gnosjö
59	0642 Mullsjö
60	0643 Habo
61	0662 Gislaved
62	0665 Vaggeryd
63	0680 Jönköping
64	0682 Nässjö
65	0683 Värnamo
66	0684 Sävsjö
67	0685 Vetlanda
68	0686 Eksjö
69	0687 Tranås
70	0760 Uppvidinge
71	0761 Lessebo
72	0763 Tingsryd
73	0764 Alvesta
74	0765 Älmhult
75	0767 Markaryd
76	0780 Växjö
77	0781 Ljungby
78	0821 Högsby
79	0834 Torsås
80	0840 Mörbylånga
81	0860 Hultsfred
82	0861 Mönsterås

Continuation of Table 8	
Number	Municipality
83	0862 Emmaboda
84	0880 Kalmar
85	0881 Nybro
86	0882 Oskarshamn
87	0883 Västervik
88	0884 Vimmerby
89	0885 Borgholm
90	0980 Gotland
91	1060 Olofström
92	1080 Karlskrona
93	1081 Ronneby
94	1082 Karlshamn
95	1083 Sölvesborg
96	1214 Svalöv
97	1230 Staffanstorp
98	1231 Burlöv
99	1233 Vellinge
100	1256 Östra Göinge
101	1257 Örkelljunga
102	1260 Bjuv
103	1261 Kävlinge
104	1262 Lomma
105	1263 Svedala
106	1264 Skurup
107	1265 Sjöbo
108	1266 Hörby
109	1267 Höör
110	1270 Tomelilla
111	1272 Bromölla
112	1273 Osby
113	1275 Perstorp
114	1276 Klippan
115	1277 Åstorp
116	1278 Båstad
117	1280 Malmö
118	1281 Lund
119	1282 Landskrona
120	1283 Helsingborg
121	1284 Höganäs
122	1285 Eslöv
123	1286 Ystad
124	1287 Trelleborg
125	1290 Kristianstad

Continuation of Table 8	
Number	Municipality
126	1291 Simrishamn
127	1292 Ängelholm
128	1293 Hässleholm
129	1315 Hylte
130	1380 Halmstad
131	1381 Laholm
132	1382 Falkenberg
133	1383 Varberg
134	1384 Kungsbacka
135	1401 Härryda
136	1402 Partille
137	1407 Öckerö
138	1415 Stenungsund
139	1419 Tjörn
140	1421 Orust
141	1427 Sotenäs
142	1430 Munkedal
143	1435 Tanum
144	1438 Dals-Ed
145	1439 Färgelanda
146	1440 Ale
147	1441 Lerum
148	1442 Vårgårda
149	1443 Bollebygd
150	1444 Grästorp
151	1445 Essunga
152	1446 Karlsborg
153	1447 Gullspång
154	1452 Tranemo
155	1460 Bengtsfors
156	1461 Mellerud
157	1462 Lilla Edet
158	1463 Mark
159	1465 Svenljunga
160	1466 Herrljunga
161	1470 Vara
162	1471 Götene
163	1472 Tibro
164	1473 Töreboda
165	1480 Göteborg
166	1481 Mölndal
167	1482 Kungälv
168	1484 Lysekil

Continuation of Table 8	
Number	Municipality
169	1485 Uddevalla
170	1486 Strömstad
171	1487 Vänersborg
172	1488 Trollhättan
173	1489 Alingsås
174	1490 Borås
175	1491 Ulricehamn
176	1492 Åmål
177	1493 Mariestad
178	1494 Lidköping
179	1495 Skara
180	1496 Skövde
181	1497 Hjo
182	1498 Tidaholm
183	1499 Falköping
184	1715 Kil
185	1730 Eda
186	1737 Torsby
187	1760 Storfors
188	1761 Hammarö
189	1762 Munkfors
190	1763 Forshaga
191	1764 Grums
192	1765 Årjäng
193	1766 Sunne
194	1780 Karlstad
195	1781 Kristinehamn
196	1782 Filipstad
197	1783 Hagfors
198	1784 Arvika
199	1785 Säffle
200	1814 Lekeberg
201	1860 Laxå
202	1861 Hallsberg
203	1862 Degerfors
204	1863 Hällefors
205	1864 Ljusnarsberg
206	1880 Örebro
207	1881 Kumla
208	1882 Askersund
209	1883 Karlskoga
210	1884 Nora
211	1885 Lindesberg

Continuation of Table 8	
Number	Municipality
212	1904 Skinnskatteberg
213	1907 Surahammar
214	1960 Kungsör
215	1961 Hallstahammar
216	1962 Norberg
217	1980 Västerås
218	1981 Sala
219	1982 Fagersta
220	1983 Köping
221	1984 Arboga
222	2021 Vansbro
223	2023 Malung-Sälen
224	2026 Gagnef
225	2029 Leksand
226	2031 Rättvik
227	2034 Orsa
228	2039 Älvdalen
229	2061 Smedjebacken
230	2062 Mora
231	2080 Falun
232	2081 Borlänge
233	2082 Säter
234	2083 Hedemora
235	2084 Avesta
236	2085 Ludvika
237	2101 Ockelbo
238	2104 Hofors
239	2121 Ovanåker
240	2132 Nordanstig
241	2161 Ljusdal
242	2180 Gävle
243	2181 Sandviken
244	2182 Söderhamn
245	2183 Bollnäs
246	2184 Hudiksvall
247	2260 Ånge
248	2262 Timrå
249	2280 Härnösand
250	2281 Sundsvall
251	2282 Kramfors
252	2283 Sollefteå
253	2284 Örnsköldsvik
254	2303 Ragunda

Continuation of Table 8	
Number	Municipality
255	2305 Bräcke
256	2309 Krokom
257	2313 Strömsund
258	2321 Åre
259	2326 Berg
260	2361 Härjedalen
261	2380 Östersund
262	2401 Nordmaling
263	2403 Bjurholm
264	2404 Vindeln
265	2409 Robertsfors
266	2417 Norsjö
267	2418 Malå
268	2421 Storuman
269	2422 Sorsele
270	2425 Dorotea
271	2460 Vännäs
272	2462 Vilhelmina
273	2463 Åsele
274	2480 Umeå
275	2481 Lycksele
276	2482 Skellefteå
277	2505 Arvidsjaur
278	2506 Arjeplog
279	2510 Jokkmokk
280	2513 Överkalix
281	2514 Kalix
282	2518 Övertorneå
283	2521 Pajala
284	2523 Gällivare
285	2560 Älvsbyn
286	2580 Luleå
287	2581 Piteå
288	2582 Boden
289	2583 Haparanda
290	2584 Kiruna
End of Table	