Stockholms universitet

Demographic impact on childbearing

Regression model fitting and comparison using demographic data

Oliver Murquist

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2016:13
Matematisk statistik
Juni 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2016:13**
http://www.math.su.se

Demographic impact on childbearing

Regression model fitting and comparison using demographic data

Oliver Murquist*

June 2016

Abstract

This thesis presents a study on the effects that demographic non-economic factors such as crime, marriage and population density have on the number of children born each year in Sweden based on data from 2005 to 2014. It also serves as a guide on how to fit and compare regression models of three different types, multiple linear, Poisson and negative binomial, to find the one with best fit. The negative binomial model proved to have the best fit, and after removing insignificant parameters the proportion of refugees, crime, gender distribution and newlyweds all had positive effects on the birthrate. Given a 1% increase from the median while all other variables where fixed, the respective effects of these variables were estimated to be 0.004%, 0.0766%, 0.67% and 0.275%. There was also a significant positive interaction between crime and population density. High education was fitted with a spline function which resulted in a positive but diminishing effect on the number of children born.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: oliver.murquist@gmail.com. Supervisor: Martin Sköld, Mikael Petersson.

Contents

1 Introduction

In Sweden fertility rates have switched from increasing to decreasing and back several times in the last 100 years and the question why has been studied over and over by researchers such as Jan Hoem, a former professor in demometry at Stockholm university who has published numerous articles on the impact of social policies and family economics on fertility [9]. Even though there are many other researchers who have devoted their time and effort to study this question, they have done it, like Hoem, from an almost entirely socioeconomic perspective. In this thesis I will instead study the effect of some demographic non-economic factors on the birthrate in Sweden to try to get a different angle on the matter, and provide a mathematical model to explain the relationship between predictor variables and response. While there are many options on how to construct our model, in this thesis we will utilize regression analysis as it is a very simple tool to learn but hard to master, and is frequently used in a wide variety of fields. It is a powerful tool to someone who uses it correctly, but due the "Simple to learn" nature it is often misused and misinterpreted by people who cut corners and ignore theory without knowing it. For those people this thesis should serve as an example of how to step by step examine data, check assumptions, correctly fit different types of models, transform variables, rank models within and between different types and finally interpret the results correctly. The thesis will also cover some problems that might occur during the fitting process and their solutions or workarounds.

2 Theory

In this section the theory and foundation that the methods used are built upon is described. The reference for each definition is listed before the definition.

2.1 Linear regression

Practical Regression and Anova using R by Julian J. Faraway [10]
Assume that we want to study the relation between a 'response' variable y and a set of 'predictors' x_j. We can do that using a multiple linear regression model if a set of assumptions are met. With assumptions

1. Linearity: Assumed linear relation between predictors and response.
2. No Multicollinearity: predictor variables are assumed to be independent of each other.
3. Normality: Errors are assumed to be normally distributed with mean zero.
4. Homoscedasticity: Errors are assumed to have constant variance.

we define a multiple linear regression model with the formula

$$y_i = \alpha + \sum_{j=1}^{n} \beta_j x_{ji} + \epsilon_i \ , \ \epsilon_i \sim N(0, \sigma^2).$$

Given a 'response' vector y and a set of 'predictors' x_j we can estimate the 'effects' β_j and the 'intercept' α. Simple linear regression is a special case where there is only one predictor, and thus the second assumption disappears.

2.2 Generalized linear models

This definition for the generalized linear model is from Nelder and Wedderburn [8] but updated with more modern notations.
Suppose our observations y come from a distribution with probability mass function

$$f(y; \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

where $a(\phi) > 0$ so that for fixed ϕ we have an exponential family. It can be shown that

$$\begin{aligned} E(Y) &= \mu = b'(\theta), \\ Var(Y) &= \sigma^2 = b''(\theta)a(\phi). \end{aligned} \tag{1}$$

Suppose also that we have a 'design' matirx X where X_{ij} is the value of the j:th variable associated with the i:th observation, a parameter $\eta = X\beta$ and a function g linking μ and η by $\eta = g(\mu)$. Putting these three components together gives us the foundation of the generalized linear model:

1. A response variable Y with distribution from the exponential family.
2. A set of predictor variables X and linear predictor $\eta = X\beta$.
3. A link function g where $E(Y) = \mu = g^{-1}(\eta)$.

Using a algorithm called 'Iterated Re-weighted Least Squares' we can compute the maximum-likelihood estimates of the effects β that the independent variables X have on $g(\mu)$. This algorithm will be defined later on in this section.

2.3 Poisson regression

The Poisson distribution is a discrete distribution and has probability mass function, mean and variance as follows:

$$f(k;\lambda) = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

$$E(X) = Var(X) = \lambda.$$

It can be shown that the Poisson distribution is part of the exponential family, and given the natural logarithm as link function, the theory on generalized linear model can be applied. The resulting regression formula is

$$log(\mu_i) = \alpha + \sum_{j=1}^{n} \beta_j x_{ji}.$$

2.4 Negative binomial regression

The negative binomial distribution is also a discrete distribution, and has probability mass function, mean and variance

$$f(k;r,p) = P(X = k) = \binom{k+r-1}{r} p^k (1-p)^r,$$

$$E(X) = \frac{pr}{1-p} \ , \ Var(X) = \frac{pr}{(1-p)^2}.$$

The 'glm.nb()' function in R that fits a negative binomial regression model uses the parameterization $p = \frac{1}{1+\frac{\mu}{r}}$ to get mean and variance

$$E(X) = \mu \ , \ Var(X) = \mu + \frac{\mu^2}{r}.$$

According to the R-documentation [6], an alternating iteration process is used when fitting the model. For a given r a Poisson model is fitted, then for fixed means r is estimated using score and information iterations. The process is repeated until convergence in both iterations. Using the logarithm as link function, the regression formula is

$$log(\mu_i) = \alpha + \sum_{j=1}^{n} \beta_j x_{ji}.$$

2.5 Iterated Re-weighted Least Squares algorithm

This explanation is from Nelder and Wedderburn [8] and describes the actual algorithm but not the detailed theory behind it. The algorithm is also described in *Practical Regression and Anova using R* by Julian J. Faraway [10].
Given a starting estimate $\hat{\beta}_0$ we can calculate $\hat{\eta}_0 = X\hat{\beta}_0$ and $\hat{\mu}_0 = g^{-1}(\hat{\eta}_0)$. Then we calculate the response and weight vectors

$$\begin{aligned} z_0 &= \hat{\eta}_0 + (y - \hat{\mu}_0)g'(\hat{\mu}_0) \\ w_0^{-1} &= b''(\theta)a(\phi)(g'(\hat{\mu}_0))^2 \end{aligned} \tag{2}$$

Lastly we calculate a new estimate

$$\hat{\beta}_1 = (X'WX)^{-1}X'Wz_0$$

where W is a matrix with diagonal w_0 and rest zeroes. Given this new estimate repeat the process from the start but using β_1 instead of β_0 until convergence is achieved.

2.6 Splines

The following explanation of Splines is on purpose very brief and will only cover the regression aspect of spline fitting. The reason for this is that the theory behind the process can be complicated and requires a good amount of time and effort to understand. The reference used is *Regression: Models, Methods and Applications* by Fahrmeir, Kneib, Lang and Marx [12]
There are three things that characterize a B-spline function, a vector of values called the 'knot' vector $t = (t_0 < t_1 < t_2 ... < t_m)$, a vector $\beta = (\beta_0, \beta_1, \beta_2 ..., \beta_m)$ with unknown values that will be estimated, and a degree n of which the basis functions will be. The B-spline function is continuous and has continuous derivatives up to degree $n-1$. A B-spline function of a variable X is then given

by the sum of the weighted basis functions

$$S_n(x) = \sum_{i=0}^{m} \beta_i B_{i,n}(x)$$

where $B_{i,n}(x)$ is calculated with Cox-de Boor recursion formula

$$B_{i,n}(x) = \begin{cases} 1 & \text{if } t_i \leq x < t_{i+1} \\ 0 & \text{otherwise} \end{cases},$$

$$B_{i,k}(x) = \frac{x - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(x) + \frac{t_{i+k} - x}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(x).$$

In this thesis we will use 'natural cubic' B-splines, which are B-splines of degree 3 with the added requirement that the second derivative is equal to 0 at the first and last knot. In regression the vector β is estimated along the coefficients of the other variables in the regression.
We will not go into further details but if you wish to learn more about the topic then read the referenced material.

2.7 Model selection algorithm

Due to the fact that there is no universal method to finding the best fitting model, the method described below is my personal method, which is but one of countless methods of finding a suitable model for a given data set. This method is a mix of backward elimination on the normal variables and forward inclusion on the interactions. The thought behind this process is that we want to have as few variables as possible, but still retain a good fit. While it is possible that there exists interactions between excluded variables and included ones, we want as few interactions as possible to be able to interpret the model better.

1. Fit a model containing all variables/covariates but excluding all interactions.
2. Look for possible transformations of the included variables.
3. With each new model, try to reduce it as much as possible while still retaining a similar fit.
4. When no more reductions or transformations improve the model fit, start introducing potential interaction terms between the variables.
5. When no more interactions improve the model fit, reduce the model as much as possible.
6. We are now left with the final model.

2.8 AIC and BIC

Practical Guide to Logistic Regression by Joseph M. Hilbe [11]
Akaikes Information Criterion(AIC) and the Bayesian information criterion(BIC) are two alternative goodness of fit measurements based of log likelihood, and are defined as

$$\begin{aligned} AIC &= -2\log(L(M)) + 2k(M) \\ BIC &= -2\log(L(M)) + 2k(M)\log(obs) \end{aligned} \tag{3}$$

where $L(M)$ denotes the likelihood of model M, $k(m)$ the number of parameters in the model and *obs* the number of observations. For both these measurements a lower value implies a better fit, but to compare two or more models they all have to use the same data set or have the same response scale. From the definitions it is trivial to see that BIC penalizes larger models more than AIC.

Given two comparable models we define $\Delta_i = AIC_i - AIC_{min}$ and with the rule of thumb described in 'Model Selection and Multimodel Inference' [7] by Kenneth P. Burnham and David R. Anderson, we can get an indicator of when to choose one model over the other.

> "As a rough rule of thumb, models having Δ_i within 1–2 of the best model have substantial support and should receive consideration in making inferences. Models having Δ_i within about 4–7 of the best model have considerably less support, while models with Δ_i>10 have either essentially no support and might be omitted from further consideration or at least fail to explain some substantial structural variation in the data."

This rule can be used for BIC in the same way and will be the basis of our AIC/BIC related arguments.

2.9 Cook's distance

Practical Regression and Anova using R by Julian J. Faraway [10]
Cook's distance is a measurement of the influence of a given observation.
Faraway describes influence as follows: "An influential point is one whose removal from the dataset would cause a large change in the fit. An influential point may or may not be an outlier and may or may not have large leverage but it will tend to have at least one of those two properties.". Therefore a high Cook's distance value does not imply that we should directly remove and label a observation as a outlier, but rather investigate it further before continuing with

the analysis. The Cook's distance is given by

$$D_i = \frac{r_i^2}{k} \frac{h_i}{1 - h_i}$$

where r_i is the i-th residual, k is the number of parameters in the model and h_i is the i-th diagonal element in the 'hat'-matrix $H = X(X^T X)^{-1} X^T$. A $D_i > 1$ suggests that the observation should be investigated.

2.10 VIF

Practical Regression and Anova using R by Julian J. Faraway [10]
The Variance Inflation Factor(VIF) is a measurement of multicollinearity, or rather how much a variable can be explained by the other variables in the regression model. A value above 4 indicates that further investigation is suggested, and a value above 10 indicates that corrections are required. The VIF value of variable x_j is calculated by regressing x_j on all other predictors, taking the resulting coefficient of determination R_j^2 and using it on the formula

$$VIF_j = \frac{1}{1 - R_j^2}.$$

2.11 Pearson dispersion statistic

Practical Guide to Logistic Regression by Joseph M. Hilbe [11]
The Pearson dispersion statistic is a measurement of the dispersion parameter in a generalized linear model. It is given by dividing the sum of squared Pearson residuals by the residual degrees of freedom. The Formula for this statistic is therefore

$$\phi = \frac{\sum_{i=1}^{n} \left(\frac{(y_i - \mu_i)^2}{V(\mu_i)} \right)}{df(residal)}$$

where $V(\mu_i) = b''(\theta)$ for distributions from the exponential family, and specifically $V(\mu_i) = e^\theta = \mu_i$ for the Poisson distribution. The dispersion parameter itself is the relation between the mean and variance of the model distribution, so a value larger than one indicates that the observed variance is larger than the observed mean.

3 Data

To get a true comparison between the model fits, the data used is the actual year by year data produced by the three public authorities Statistiska Centralbyrån(Statistics Sweden)[1], Brottsförebyggande rådet/BRÅ(The Swedish

National Council for Crime Prevention)[2] and Migrationsverket(The Migration Agency)[3]. The data was produced according to The Official Statistics Act[4] and consists of one observation of each explanatory variable per municipality and year during the period 2005-2014. Data was manually collected from each source and then imported and processed in R[5].

3.1 Data set

The starting data set contains one observation of the following variables per municipality.

Year : The year of data collection, 2005-2014.

Born : Number of successful childbirths.(SCB)

Marriage : Number of newlywed women between ages 20-39.(SCB)

Refugees : Number of received refugees.(MIG)

Density : Population per square kilometer.(SCB)

Crime : Number of committed crimes.(BRÅ)

HEducation : Number of citizens with a post tertiary education >3 years.(SCB)

GenderDist : The number of men divided by the number of women.(SCB)

Population : Population of the municipality.(SCB)

3.2 Data analysis

Before beginning with the analysis it is always a good idea to try and get an overview of the data set. We do this mainly to detect faulty values, high correlation and other problems that would disrupt model fitting.

When inspecting the mean, median, minimum and maximum for each variable we do not find any anomalies in any of the variables, and apart from 'Refugees' no variable has any missing values.

In the data file containing the number of received refugees per municipality we find that not all municipalities are present in the table, in other words we find out that the municipalities with missing values in our data set do not have the value 'NA' in the data file but are missing altogether. It is also discovered that there are no observations equal to zero in the entire data file. These two

discoveries combined strongly suggests that the data file is in fact a list of number of received refugees for each municipality that received at least one refugee, meaning those that received no refugees have been left out. By using this interpretation we simply replace all the missing values for the variable 'Refugees' with zero.

The next step is to do a Matrix-Plot over our data set. A Matrix-plot is essentially a matrix of all variables(including the response variable) plotted against each other in pairs. The plots are found above the diagonal in Figure 1 below, and as an addition the correlation coefficients of each pair have been listed below the diagonal.

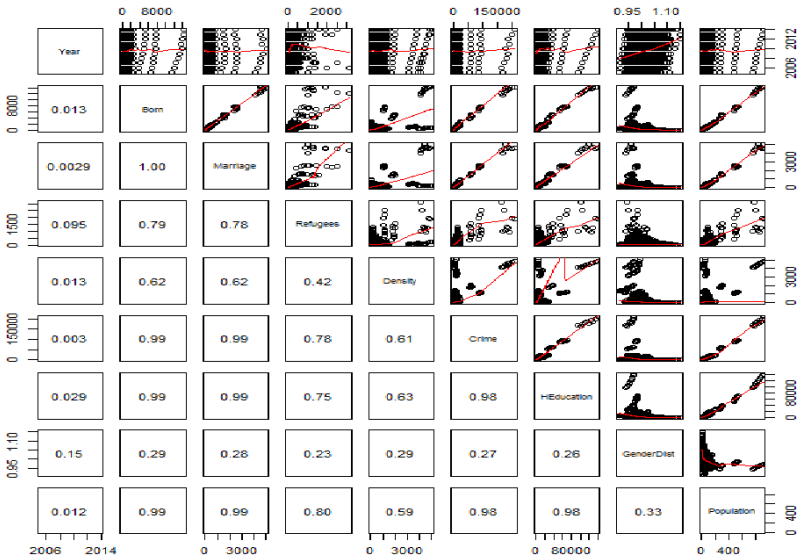

Figure 1: Plot of variables vs each other in pairs.

By examining the plots and the correlation coefficients we observe that there seems to be a high to very high correlation between all the variables except for 'Year' and 'GenderDist'. These correlations are to be expected due to all variables being based on the same concept of number of citizens that got married, are refugees, have a high education etc. With this thought in mind when looking at the correlation coefficients we come to the conclusion that the common divisor of all the variables is 'Population'. We come to this conclusion due to the fact that it has a correlation of around 0.99 with most other variables, and that it is logical that things like number of highly educated citizens and number of crimes etc. increase with the population size.

To combat this correlation and get explanatory variables not directly dependent on population size we divide all variables except for 'Year', 'Born' and 'GenderDist' by the corresponding observation of 'Population'. We then create another data set where we have divided 'Born' by 'Population' as well. This is done due to Poisson and negative binomial models using count data and the logarithm of 'Population' as an added offset, whereas simple linear and multiple linear models use continuous data and therefore can handle the divided response variable.

To get an overview of the data set now that it has been modified we make two matrix-plots, one for each data set. Figure 2 contains the plot for the simple and multiple linear regression data set.

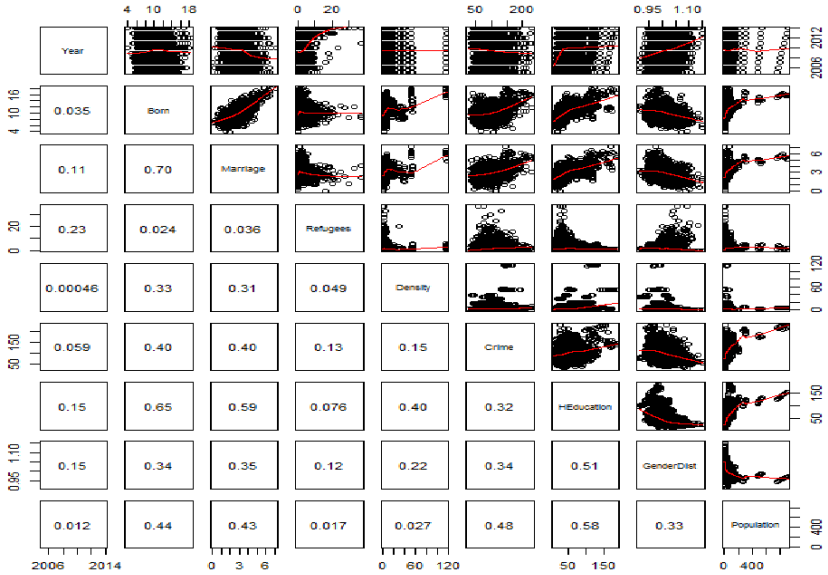

Figure 2: Plot of variables vs each other in pairs(divided data set).

4 Analysis

To avoid repetition we try to find the best fitting model in linear regression and then fit models using the same variables for Poisson and negative binomial regression. This gives a good comparison between the three model types but as a consequence we might not end up with the overall best model for any given type.

To make formulas more compact we will use the first letter in each variable as notation for that variable, β for the effects, α for the intercept, ϵ for the error

term and ':' to denote a interaction term.

4.1 Simple linear regression

To start off the regression analysis we apply the theory of Simple Linear Regression to see how well the covariates can explain the response one by one. When fitting these regression lines we make three assumptions

1. Independent observations of the response(municipalities are independent).
2. Independent errors with constant variance, $\epsilon_i \sim N(0, \sigma^2)$.
3. $E(Y_i)$ is a linear function of the predictor X_i for all values x_i.

When fitting simple linear model our goal is mainly to find out what variables explain the response well on their own as well as finding possible variable transformations that could be useful later in the analysis. Due to this and the fact that these models will not be used for any predictions och explanations we will not check the assumption of constant variance(homoscedasticity).
The models we fit are all of the form

$$\frac{B}{P} = \alpha + \beta Covariate + \epsilon$$

where the transformation of the response is described at the end of the 'Data' section. The results of these fitted regression models are presented in Table 1.

Table 1: Results of simple linear regression models

Variable	Estimate	P-value	R^2	AIC
Year	0.0230	0.06	0.0012	11993.77
Marriage	1.5680	<2e-16	0.4883	10054.30
Refugees	0.0148	0.19	0.0006	11995.54
Density	0.0690	<2e-16	0.1107	11656.97
Crime	0.0252	<2e-16	0.1581	11498.22
HEducation	0.0507	<2e-16	0.4161	10436.86
GenderDist	-20.705	<2e-16	0.1153	11641.95
Population	0.0129	<2e-16	0.1894	11388.17

Reading the table we notice that 'Marriage' and 'HEducation' have the best fits with R squared values of 48.8% and 41.6%. It is also noticed that the variables 'Year' and 'Refugees' have coefficients not significantly different from zero. To check if this logical we examine the respective plots in Figure 2 and come to the conclusion that the statement is indeed logical due to the lack of a close to linear relationship between the variables and the response.
While inspecting the other plots in Figure 2 we notice that both 'HEducation'

and 'Population' seem to have non-linear relationships with the response. To analyze this further we enlarge the plots and add the simple regression line, which can be seen in Figure 3 below.

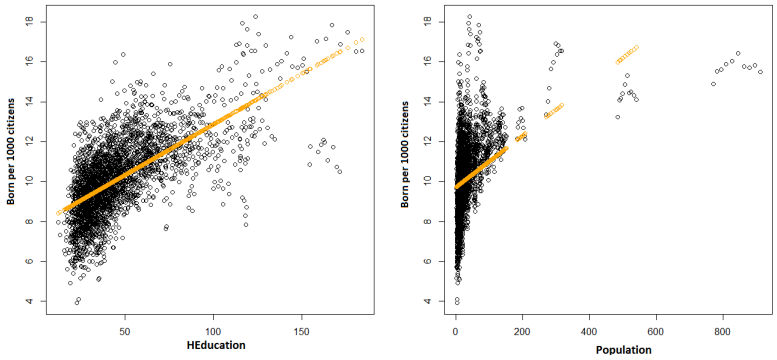

Figure 3: Heducation(left) and Population(right) against response with added simple regression points(orange).

Both relationships look logarithmic, so therefore we fit simple regression models of the form

$$Response = \alpha + \beta \log(Covariate) + \epsilon$$

Table 2 consists of a comparison between the new log-models and the previous ones.

Table 2: Simple regression with and without log transformed covariates.

Variable	P-value	R^2	AIC
HEducation	<2e-16	0.4161	10436.86
log(HEducation)	<2e-16	0.4543	10240.78
Population	<2e-16	0.1894	11388.17
log(Population)	<2e-16	0.3446	10772.13

When examining the values in Table 2 we come to the conclusion that the log transformed variables produce models with much better fits than their counterparts, but Figure 4 shows that the slope of the log-function still flattens quicker than preferred.

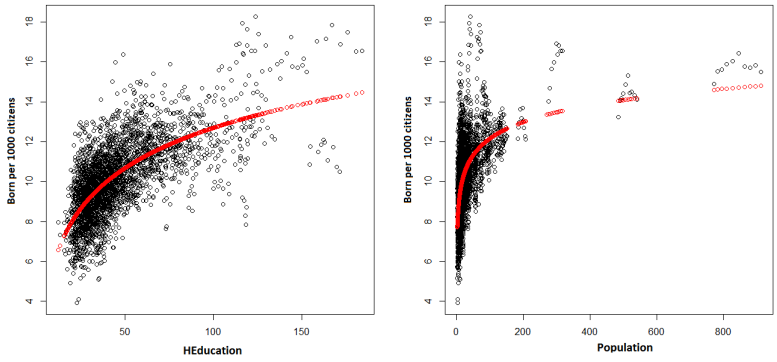

Figure 4: Heducation(left) and Population(right) against response with added log(covariate) regression points(red).

In the theory section we introduces the concept of B-Splines, mainly the R-function $'ns\{Splines\}'$ which produces a B-Spline basis for fitting natural cubic splines between knots. This theory is mainly used for fitting curves to data with complex relationships between response and explanatory variables not easily explained by relatively simple transformations like $log(x)$, x^a or e^x, but is still usable in rather simple cases like ours.
We fit two models using the function 'ns()' on the explanatory variable and increase the number of knots until the models have lower or the same AIC values as the log-models. For both variables we stopped at four knots with AIC values for 'HEducation':10221.76 vs previous 10240.78, and 'Population':10774.46 vs previous 10772.13. For the 'Population' models AIC of the log-model is still lower than the other, but only by 2.33. Using the rule of thumb presented in Section 2 both models of 'Population' are assumed to be equal in terms of AIC. The natural cubic spline fitting introduces five new polynomials to the regression, which are extremely difficult to interpret. Below in Figure 5 we have added the fitted values of the Spline regressions(blue) to the plots in Figure 4. Now that we have some potential transformations for non-linear-fitting variables we continue our analysis by moving on to multiple regression.

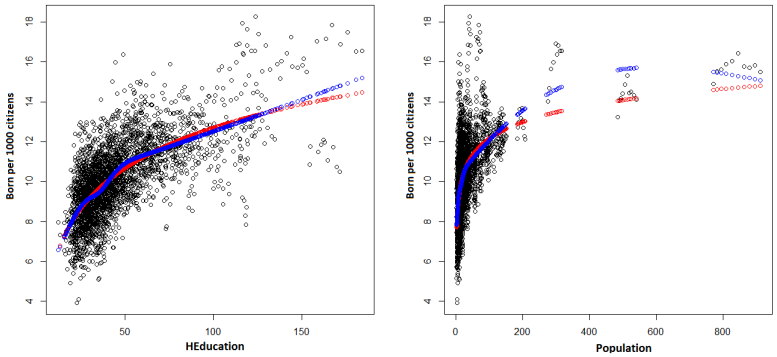

Figure 5: Same as Figure 4 but with fitted values from spline regression(blue).

4.2 Multiple linear regression

When comparing models in multiple regression we have to take into account the number of parameters in the models since we want a model as simple as possible while still having a good fit. Because of that we will not only compare AIC values but also BIC values since BIC penalizes large models more than AIC, and the rule of thumb we will use in our decisions is defined in Section 2.

We start off by investigating models without interactions, and then include them later on. The first step we take is to verify that the transformations of 'HEducation' and 'Population' made in the previous section result in a better fit than regression with the normal non transformed variables and regression with fitted natural cubic B-splines. In accordance with the theory section our multiple linear regression models will be of the form

$$\frac{B}{P} = \alpha + \sum \beta Covariate + \epsilon$$

where the transformation of the response is described at the end of the 'Data' section. Table 3 consists of R^2, AIC and BIC for seven different models, and indicate that both log-transformations at the same time gives a better fit than not transforming, but that the model with splines fitted to 'HEducation' but not to 'Population' had the best fit of them all.

Table 3: Results of fitted multiple linear regression models.

Transformed variable	R^2	AIC	BIC
None	0.5855	9457.317	9517.042
log(HEducation)	0.6093	9285.876	9345.6
log(Population)	0.5914	9415.732	9475.456
log(Both)	0.6099	9281.202	9340.927
Spline(HEducation)	0.614	9254.438	9326.108
Spline(Population)	0.5864	9453.16	9518.857
Spline(Both)	0.6141	9256.178	9333.82

To check the assumptions of normality and constant variance of the residuals we utilize a normal QQ plot and a residual plot. The QQ-plot on the right in Figure 6 gives a perception of the normality of our standardized residuals by plotting them and the corresponding residuals from a perfect normal distribution(diagonal dotted line) against the theoretical quantiles. It shows that the points are aligned well will the diagonal and therefore that the assumption of normality of the residuals seem to be fulfilled. The residual plot on the left in the same figure indicates that apart from the interval 14-18, the assumption of homoscedasticity is correct since the vertical spread of the points is even along the x-axis. The specified interval does not interfere with the assumption because of the fact that the decrease in variance is almost certainly due to a lack of observations in that interval rather than a sign of heteroscedasticity. Lastly we check the Variance Inflation Factors(VIF) to make sure that we do not have high levels of multicollinearity, and get no values above the threshold 4.

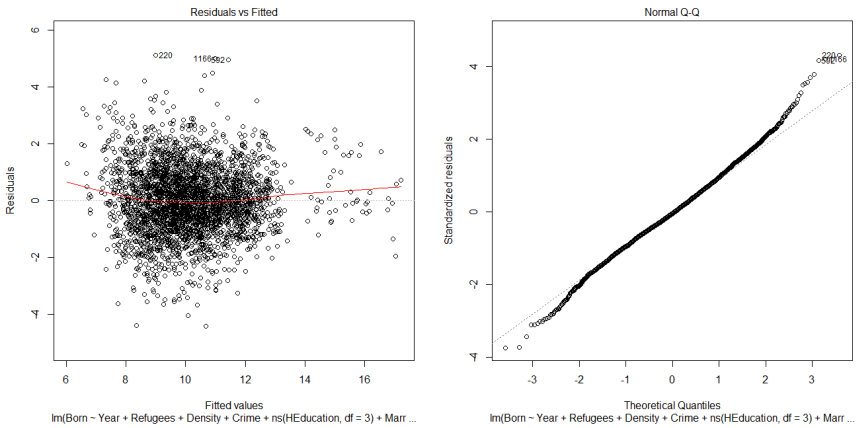

Figure 6: Residual-plots of the current model

The next step in our analysis is to discard insignificant or otherwise unnecessary variables from the model. Looking at the regression output from R, our

model has no insignificant variables on level 5%, 'Year' is the least significant with a P-value of 0.0219. The result of reducing the model anyway is an increase in AIC by 3.28 but a decrease in BIC by 2.69. Using our rule of thumb both models are almost equal, but since we want to have as few variables as possible we stick with the reduction and move on with the analysis.

Now we have reduced the model as much as possible while still retaining a good fit. The next step is to introduce interactions between parameters. Due to the lack of previous studies on this particular subject, we have no starting ground or previous results to base our testing on. Apart from testing all combinations of interactions, our only way of selecting which parameters might have potential interactions is to construct personal hypotheses.

One hypothesis is that the effect of crime is increased in tightly populated areas due to a greater awareness of the crime rates. An example is that given two areas of different size but with the same crime per citizen ratio, the risk of a crime being committed in a close vicinity of you is higher in the more densely populated area than the other. Introduction of this interaction parameter results in a decrease in AIC and BIC by 32.02 and 26.05 respectively, and that the parameter's coefficient is positive, which is in line with the hypothesis.

We cannot come up with any arguments for investigation of other interactions, but decide to test some interactions that, if proven to be significant, we would find interesting. While testing we discover a strange behavior were AIC is lowered after almost all new introductions, but as consequence the previously included interactions and almost all main parameter effects shift between being significant and not seemingly at random. We then look at the BIC values instead and notice that they are all roughly the same or higher than that of our current best model. This strange occurrence could be due to the larger models over-fitting data and letting the interactions falsely try to explain random noise. We cannot be sure that that is the case but after experimenting with different interactions and not finding any pattern, it is my opinion.

Since no expansion, reduction or applied transformation results in a model with better fit judging by AIC, BIC, residual normality and simplicity, we have found the final multiple linear regression model to be

$$\frac{B}{P} = \alpha + \beta_M M + \beta_R R + \beta_D D + \beta_C C + \beta_H S(H) + \beta_G G + \beta_P P + \beta_{DC}(D:C) + \epsilon$$

In Figure 7 we have three plots, normal QQ, residual and Cook's distance plot.

With the same motivations as we had when we discussed Figure 6 our residuals seem to fit the normality assumption well apart from the right tail(QQ plot), and there seems to be no interference with the assumption of homoscedasticity(residual plot). The Cook's distance plot gives us an indication of how influential the observations are, and from it we conclude that there seem to be no major outliers worth removing(Cook's distance >1). Table 4 consists of the regression coefficient estimate, standard error and P-value of each parameter in the final model(except splined 'HEducation'), and will be compared to the other models in Section 4.5.

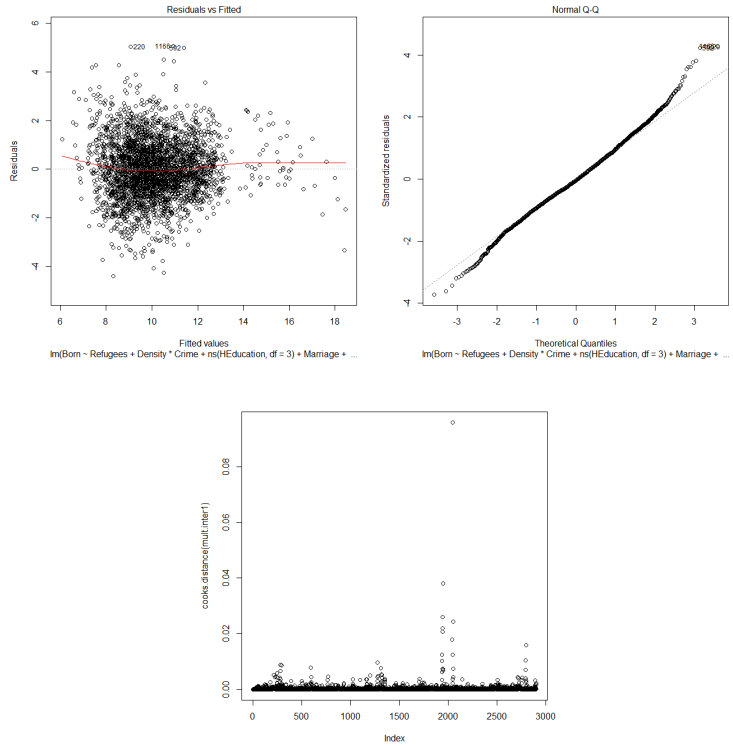

Figure 7: Residual, normal QQ and cooks distance plots for the final model.

Table 4: Parameter summary for the final multiple linear model.

Parameter	Estimate	std.error	P-value
Marriage	8.778e-01	3.476e-02	<2e-16
Refugees	2.487e-02	7.247e-03	0.0006
Density	-3.284e-02	9.889e-03	0.0009
Crime	5.147e-03	9.689e-04	1.17e-07
GenderDist	5.462e+00	8.935e-01	1.11e-09
Population	2.029e-03	4.958e-04	4.39e-05
Density:Crime	3.977e-04	7.069e-05	2.02e-08

4.3 Poisson regression

As previously stated we won't repeat the process of finding a suitable model due to the repetition not adding any valuable information and being a tedious process to read through. Instead we will fit a model with the same parameters as the final multiple linear model to analyze the fit and suitability of a Poisson regression model on our data. Comparison between the models will be covered in Section 4.5.

The reason behind us exploring the possibility of a Poisson regression model is that we have count data, meaning that our response variable is discrete non negative. While Poisson regression results in predictions that are possible in the real world, i.e no 3.5 or 1.7 children born, we must however deal with the strong assumption of equidispersion, which means that the mean should equal the variance. While it is a strong assumption, in practice Poisson models are often fitted even if the assumption does not hold completely. To check the assumption we calculate the Pearson dispersion, which we explained in Section 2, and check if it is 1. To get the desired count data we will, as stated in the last paragraph of Section 3, counter the high correlation between population and all other variables including the response by adding a offset parameter instead of dividing all observation by the corresponding population value. From the theory section we recall that a Poisson model with a added offset variable has the formula

$$log(\mu) = \alpha + \sum \beta Covariate + \text{offset}$$

and therefore the model with the same parameters as the final multiple linear model is given by the formula

$$log(\mu_B) = \alpha + \beta_M M + \beta_R R + \beta_D D + \beta_C C + \beta_H S(H) + \beta_G G + \beta_P P + \beta_{DC}(D:C) + log(P)$$

Now we can fit our model and while doing so we calculate the Pearson dispersion to be 2.5. Since it is considerably larger than 1 we have *Over*-dispersion, which is described in detail in Section 2. Now that we know that the variance is not equal to the mean we might find a negative binomial model to be more suitable, and in the next section we will explore that option further. Moving on we evaluate the fit and leave discussion on the topic to Section 5.

When looking at the parameter summary in Table 5 we note that all parameters except the interaction 'Crime:Density' and the main effect of 'Density' are significant on levels way below 5%. When we remove the interaction we get an increase in AIC from 27715.46 to 27716 and decrease in BIC from 27781.16

to 27775.36. Using the rule of thumb for delta AIC/BIC we do not favor the smaller model at all for AIC($\Delta AIC = 0.54$), but for BIC we might prefer the smaller model($\Delta BIC = 5.8$). Since $\Delta BIC < 7$ and having the same set of parameters makes comparison between model types easier, we simply note that the interaction is insignificant, but stick to the non reduced model.

Table 5: Parameter summary for the Poisson model

Variable	Estimate	Std.error	P-value
Refugees	3.666e-03	4.440e-04	<2e-16
Density	4.583e-04	5.072e-04	0.366195
Crime	8.768e-04	4.670e-05	<2e-16
Marriage	9.638e-02	1.898e-03	<2e-16
GenderDist	6.557e-01	4.913e-02	<2e-16
Population	2.800e-05	8.287e-06	0.000727
Density:Crime	5.135e-06	3.486e-06	0.140777

Looking at the parameter estimates in the same table it is necessary to keep in mind that due to the different response structure in Poisson regression, we cannot directly compare estimates with the ones acquired from linear regression. What we can do and will do in Section 4.5 is to compare if the effects have switched from positive to negative or the opposite.
The Cook's distance plot in Figure 8 does not indicate that any potential outliers should be removed since the all are well below the threshold $D_i > 1$. Lastly we check the VIF values and discover that we do in fact have medium amount of multicollinearity(above 4), however none of the values are above 10, which is the threshold for serious collinearity that requires immediate correction.

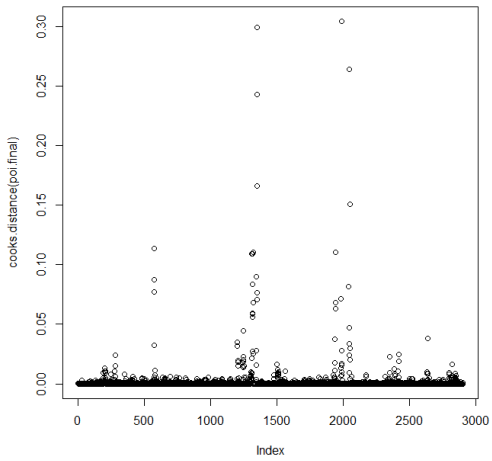

Figure 8: Cooks distance for the Poisson model.

4.4 Negative binomial regression

Due to overdispersion in the case of Poisson regression we explore the possibility that a negative Binomial regression model will fit data better. The reason for this is that Negative Binomial models, unlike Poisson models, have a second parameter that can be used to adjust the variance independently of the mean. In theory this should be the best of two worlds since it is a count data model and handles situations where it is not appropriate to assume equidispersion.

Just like in Poisson regression we use an added offset parameter and so our model has the formula

$$log(\mu_B) = \alpha + \beta_M M + \beta_R R + \beta_D D + \beta_C C + \beta_H S(H) + \beta_G G + \beta_P P + \beta_{DC}(D:C) + log(P)$$

In Table 6 we list the coefficient estimate, standard error and P-value for each parameter, and due to Poisson and negative binomial regression having the same response scale, we can directly compare these estimates with those acquired from Poisson regression, but not those from linear regression.

Table 6: Parameter summary for the negative binomial model

Variable	Estimate	Std.error	P-value
Refugees	3.095e-03	7.036e-04	1.08e-05
Density	-8.604e-04	8.295e-04	0.2996
Crime	6.930e-04	8.400e-05	<2e-16
Marriage	9.434e-02	3.194e-03	<2e-16
GenderDist	6.695e-01	8.308e-02	7.76e-16
Population	6.042e-05	3.163e-05	0.0561
Density:Crime	1.442e-05	5.797e-06	0.0129

We observe that the main effects of 'Density' and 'Population' are insignificant on level 5%. We research if it is possible and logical to remove a main effect while keeping the interaction term, but come up with no serious mathematical evidence that supports that decision, and therefore refrain from it. Removing 'Population' increased AIC by 1.66 and reduced BIC by 4.31. This suggests that we favor the reduced model only by a very small amount, but on grounds of easy comparison we keep the non reduced model.

The Cooks distance plot in Figure 9 does not indicate that any major outliers exist and should be removed($D_i > 1$), and all VIF values are well below 4. With that we end the sections about model fitting and move on to comparison between model types.

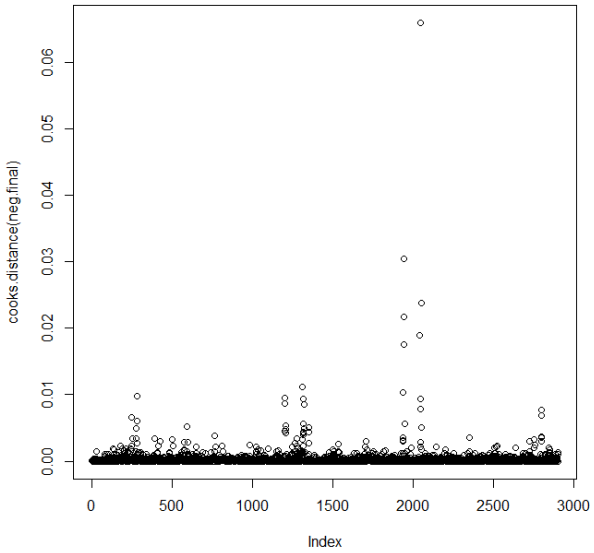

Figure 9: Cooks distance plot for the non-reduced negative binomial model.

4.5 Comparison

4.5.1 Effects

By examining Table 7 we get a clear view of any variables that switch effect from positive to negative or the opposite between model types. We observe that 'Population', 'Density' and its interaction term with 'Crime' all switch between significance and not, but no variable changes sign between model types.

Table 7: Sign of estimated effects for each model.

Parameter	Linear	Poisson	Neg.Bin
Refugees	+	+	+
Density	-	0	0
Crime	+	+	+
S(HEducation)1	+	+	+
S(HEducation)2	+	+	+
S(HEducation)3	+	+	+
Marriage	+	+	+
GenderDist	+	+	+
Population	+	+	0
Density:Crime	+	0	+

4.5.2 Best model

Before we can decide on which model is the best we must take a closer look at the definitions of AIC and BIC. By definition AIC is given by

$$AIC = -2\log(L(M)) + 2k(M)$$

where $L(M)$ is the notation for the likelihood of model M, and $k(m)$ is the number of parameters in the model. Similarly BIC is given by

$$BIC = -2\log(L(M)) + 2k(M)\log(obs)$$

where *obs* denotes the number of observations in the data set. Both these criteria are based on the likelihood function, and therefore comparison of AIC or BIC values require the response to be on the same scale. You can illustrate this by yourself if you compare a multiple linear regression model with response Y_i and the multiple linear regression model with response $log(Y_i)$.
At the end of the 'Data' section of this thesis we stated that the response in our Simple and Multiple regression models was Y_i/n_i and that the response in our Poisson and Negative Binomial regressions was Y_i. Due to this we can only directly compare our Poisson model to our negative binomial model, but with some theory we can actually calculate the log likelihood that our multiple linear model would have if it was on the same scale. In our multiple linear model the response is distributed

$$\frac{Y_i}{n_i} \sim N(\mu, \sigma^2)$$

which by properties of the normal distribution is

$$Y_i \sim N(\mu n_i, \sigma^2 n_i^2)$$

which now is on the same scale as the other regression models. We now use the general formula for calculating the log likelihood of a normal distribution to scale AIC and BIC of our final multiple regression model and construct Table 8 with the values off all three final models.

Table 8: AIC and BIC comparison between the three models.

Model	AIC	BIC
Multiple Linear	26140.09	26307.43
Poisson	27715.46	27781.16
Negative Binomial	25788.69	25860.36

5 Discussion

We will begin the discussion by deciding on which model is the best overall, and then interpret the estimated effects in that model. After that we will bring the thesis to a close by discussing some flaws and disadvantages with how our analysis was carried out as well as some potential improvements that could be made.

Since all models have the same number of parameters and observations, the model rankings will be the same for AIC as for BIC. From Table 8 we get the result that the negative binomial regression model has the best fit judging by AIC/BIC, and we stated earlier that it fit well theoretically due to it being a count data model. The conclusion we arrive at is that the negative binomial model is the best, but since we could reduce that model and still get the same, if not slightly better, fit we choose to do that. Therefore the best model is the reduced negative binomial model with formula

$$log(\mu_B) = \alpha + \beta_M M + \beta_R R + \beta_D D + \beta_C C + \beta_H S(H) + \beta_G G + \beta_{DC}(D:C) + log(P).$$

In that model all significant estimated parameter coefficients are positive. 'Refugees', 'Marriage' and 'GenderDist' are the only variables that are not involved in an interaction or transformation, and are therefore the only ones that can be interpreted in the normal way. The effect C of a 1% increase in a variable v_i with coefficient β_i can be calculated as

$$\begin{aligned} C * Born_{before} &= Born_{after} \\ &= e^{(\sum^{i-1} \beta_j v_j)} * e^{\beta_i (1.01 v_i)} * e^{(\sum_{i+1} \beta_j v_j)} * Population \\ &= e^{\beta_i 0.01 v_i} * e^{(\sum \beta_j v_j)} * Population \\ &= e^{\beta_i 0.01 v_i} * Born_{before} \\ C &= e^{\beta_i 0.01 v_i} \end{aligned} \tag{4}$$

Where $Born_{after}$ and $Born_{before}$ are the estimated number of children born before and after the 1% increase. Applying the results from the equation and choosing a baseline value for each variable, we calculate the effect of a 1% increase in a variable while the others are fixed. As baseline value we choose the median since the variables have some max or min values that skews the mean. The results of the calculations are presented in Table 9. The interpretation of the variable 'GenderDist' is however a bit different from the others. A 1% increase in that variable is interpreted as a 1% increase in the number of men, given that the number of women stay the same, or that the fraction of the factor increase in men divided by the factor increase in women is equal to 1.01.

Table 9: Estimated effects of a 1% increase in a variable given a baseline value.

Variable	Effect	Baseline value
Refugees	0.004%	1.294
Marriage	0.275%	2.885
GenderDist	0.67%	1.0099

Regarding 'Crime', 'Density' and their interaction, we can rewrite the sum

$$\beta_D D + \beta_C C + \beta_{DC}(DC) = (\beta_C + \beta_{DC} D)C + \beta_D D$$

and observe that by thinking of $\beta_D D$ as an intercept, the interpretation of the interaction coefficient is the change in slope of the effect of 'Crime'. By fixing 'Density' at its baseline value we can calculate the effect of a 1% increase in 'Crime' just as we did previously. The result we get is that a 1% increase in crime translates to a 0.0766% increase in number of children born. If we do the same for 'Density' but using $\beta_D = 0$ due to parameter insignificance, we get an increase by 0.2036%.

The effect of 'HEducation' is the only effect that we cannot calculate using the same method. A 1 unit increase has different effects depending on the value of the variable due to the non linear relationship with $\log(\mu)$, and therefore a 1% increase given a baseline value would depend even more on the chosen value. We can however inspect the spline-function and interpret its shape. In Figure 10 we have plotted the function given that all other variables are 0 except the offset 'Population' which we set to 1. This means that the numbers on the Y-axis are the number of children born in an area with a population of 1000. We can clearly see that the effect diminishes as 'HEducation' increases.

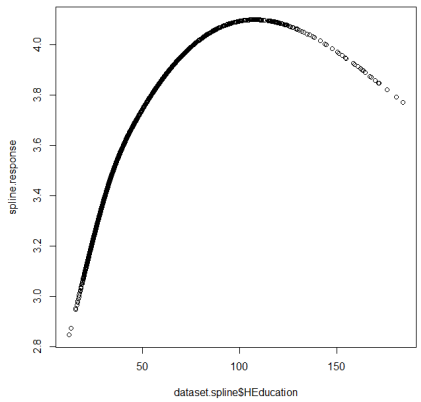

Figure 10: Plotted $\exp(\beta_H S(HEducation))$ given a population of 1000.

Moving on to disadvantages and flaws with the study, one is that there is no included variable that takes into account geographic location or separates urban and rural areas. The 'Density' variable was intended to have a similar effect, but due to it being continuous the clear lines between groups could not be drawn. Another example of a missed effect is some kind of measurement of religion, but it was excluded due to lack of easily obtainable data.

Due to the fact that it takes time between when a couple chooses to have a child and when it actually is born, a time series analysis might have been a better choice, but I personally wanted a practical test on regression analysis and that is why I used it. With that said, perhaps fertility would be a more fitting response variable, but based on a personal hypothesis that it would require data on a individual level, I chose to study the number of children born instead.

Lastly, when fitting the Poisson model we encountered the issue of overdispersion and chose to fit a negative binomial model, but made no attempt to adjust the Poisson model. While there are 'robust' methods to work around this problem using 'Quasi-Likelihood', we deemed those methods to be beyond the intended scope of this thesis, and instead chose to narrow our analysis down to the three model types most people are familiar with.

Acknowledgments

I would like to thank my supervisors Matrin Sköld and Mikael Petersson for their opinions, guidance and patience during the whole process of this project. I would also like to thank Sebastian Bergström for the productive discussions on the topic of B-splines.

References

[1] Statistiska Centralbyrån: Statistikdatabasen
http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/?rxid=579ec7f6-7a98-47f0-9ab7-c27bc76afc3a

[2] Brottsförebyggande rådet: Brottsdatabasen
http://statistik.bra.se/solwebb/action/index

[3] Migrationsverket: Kommunmottagna enligt ersättningsförordningen
http://www.migrationsverket.se/Om-Migrationsverket/Statistik/Oversikter-och-statistik-fran-tidigare-ar/Kommunmottagna--tidigare-ar.html

[4] The Official Statistics Act
http://www.scb.se/en_/About-us/Official-Statistics-of-Sweden/

[5] R: Software environment for statistical computing
https://www.r-project.org/

[6] glm.nb(): Fit a Negative Binomial Generalized Linear Model. R documentation. https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/glm.nb.html and http://stats.stackexchange.com/a/103579

[7] Kenneth P. Burnham, David R. Anderson: Model Selection and Multimodel Inference. 2002. Springer

[8] J.A.NELDER, R.W.M.WEDDERBUR: Generalized Linear Models. 1972. Journal of the Royal Statistical Society. Wiley

[9] Hoem Jan M: Social Policy and Recent Fertility Change in Sweden. 1990 Population and Development Review 16.4 (1990). Population Council.

[10] Julian J. Faraway: Practical Regression and Anova using R. 2002. R Documentation.

[11] Joseph M. Hilbe: Practical Guide to Logistic Regression. 2016. CRC Press.

[12] Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, Brian Marx: Regression: Models, Methods and Applications. 2013. Springer