Analysis and modeling of child mortality

Project report

Joel Berhane
Emil Godonou
5/21/2014
# Table of Contents

ABSTRACT ................................................................................................................................................ 3

INTRODUCTION ...................................................................................................................................... 4

THEORY .................................................................................................................................................... 5
  
  *Multiple linear regression* .......................................................................................................................... 5
  
  *Ordinary least squares* .................................................................................................................................. 5
  *Gauss-Markov assumptions* .......................................................................................................................... 6
  
  *Goodness-of-fit: R²* .......................................................................................................................................... 7
  
  *OLS: Problems & Remedies* ......................................................................................................................... 7
    *Multicollinearity* ........................................................................................................................................ 7
    *VIF* ............................................................................................................................................................ 8
    *Heteroskedasticity* ...................................................................................................................................... 8
    *Variable transformation* ............................................................................................................................... 9
    *White’s consistent estimator* ....................................................................................................................... 10
    *Endogeneity* .......................................................................................................................................... 10
    *Residual plots* ........................................................................................................................................... 11
  
  *Hypothesis testing & Inference* .................................................................................................................... 12

METHOD ................................................................................................................................................. 12

  
  *Modeling* ................................................................................................................................................... 13
    *Dependent variable* ..................................................................................................................................... 13
    *Covariates* ............................................................................................................................................... 13
  
  *Data acquisition* ....................................................................................................................................... 18

  
  *Regression Software* ................................................................................................................................ 18

ANALYSIS ................................................................................................................................................ 19

ENDOGENEITY ............................................................................................................................................ 19

  
  *Heteroskedasticity* .................................................................................................................................. 22
  
  *Multicollinearity* ..................................................................................................................................... 25
  
  *Model selection* ..................................................................................................................................... 25
Abstract

Child mortality has always been a strong indicator on the general wealth of a country therefore was decreasing it in the world one of the millennium goals. Since child mortality is such an important measure we have decided to make a structural model of it. This is an investigation regarding the causes of child mortality.

The definition of child mortality is the number of children that died before the age of five. Not to be mixed with infant mortality which is a measure of those who died before the age of one. In contrast to infant mortality we think that child mortality strongly depends on other factors than just sickness, that’s why we chose to investigate child mortality among the two of them.

We emanated from the assumption that it was possible to describe child mortality with the model:

\[ Y = X\beta + e \]

With that in mind we collected data from the UN and the World Bank for child mortality and all possible factors that could affect child mortality.

The analysis is made with the method “multiple linear regression”. At the end it all came down to a structural model consisting of six explanatory variables which are:

- U-nation
- GNI in PPP-terms
- Improved sanitation
- Average precipitation
- Help organization
- Colony

All of these variables got a positive coefficient in their effect on child mortality except for “GNI in PPP-terms” and “improved sanitation” which rationally got a negative sign. We also made a model which could be used for predicting child mortality. The difference between the prediction model and the structural model is the variable birthrate which is included in the prediction model but not in the structural model. In the prediction model birthrate got a positive sign.
Introduction

The human race has always made everything in their command to save their children. In one of the largest gatherings of world leaders in human history, known as the Millennium Summit, the United Nations and a number of major independent international organizations established the Millennium Development Goals: eight goals to improve the quality of life for the poor and generate conditions for global sustainable development in the world. Of the eight cases involving two children, one of which is to reduce child mortality worldwide. Child mortality is of enormous interest for many purposes and is often used as an indicator of health status within countries and regions.

This project is about how to come up with a structural model of the causes of child mortality. The definition of child mortality is the number of children that die before the age of 5. Not to be mixed with infant mortality which is the amount of children that do not survive past their first birthday. We chose to investigate child mortality instead of infant mortality since we think that there are more causes to rule in if the child lives up to five years instead of just one.

The method to do this investigation is called multiple linear regression. The basic model is:

\[ y = x\beta + e \]

where \( y \) is the dependent variable, \( x \) is a vector with possible covariates on which \( y \) is dependent, \( \beta \) is a vector with coefficients and \( e \) is the error residual. The regression estimates the coefficients \( \beta \) with the ordinary least squared method.

The dependent variable in this project is child mortality. The task is to find suitable covariates which explain child mortality.
Theory

Multiple linear regression

Multiple linear regression is a method for describing a variable $y$ with several explanatory variables $x_i$ using a linear model:

$$y_i = \sum_{j=0}^{k} x_{ij} \beta_j + e_i, \quad i = 1, \ldots, n,$$

where $y_i$ is the dependent variable, $x_{ij}$ are the covariates on which $y_i$ is dependent on and $e_i$ is the error term. $j = 0$ corresponds to the intercept and hence $x_{i0} = 1$ for all $i$. $i = 1, \ldots, n$ represents the observations on which the model is based on. The model can also be written in the more compact matrix notation:

$$Y = X\beta + e$$

(1)

The advantage of multiple regressions analysis, compared to simple regression analysis, is the ceteris paribus interpretation of the coefficients. Including multiple covariates in the model allows for explicit control of the factors that affect the dependent variable. This enables the study of each covariates effect on the dependent variable while keeping the others fixed.

Ordinary least squares

The OLS method solves the optimization problem of finding the vector $\beta$ that minimizes the sum of the squares of the residuals (SSR):

$$\hat{e} = Y - X\hat{\beta}$$

$$\min_{\beta \in \mathbb{R}^k} \hat{e}^T \hat{e} = \min_{\beta \in \mathbb{R}^k} |\hat{e}|^2$$

where $\hat{\beta}$ is the OLS estimate of $\beta$. This is accomplished by solving the normal equation:
\[ X^t \hat{e} = 0 \]

From the normal equation one can derive the following expression for \( \hat{\beta} \):

\[
\hat{\beta} = (X^t X)^{-1} X^t Y
\]  

(2)

From this can it be proved that \( \hat{\beta} \) is an unbiased estimate of \( \beta \):

\[
\hat{\beta} = (X^t X)^{-1} X^t Y = (X^t X)^{-1} X^t (X \beta + e) = (X^t X)^{-1} (X^t X) \beta + (X^t X)^{-1} X^t e
\]

\( X \) is the data matrix with the observed values for all covariates. The expected value of the error \( e \) is zero. Therefore the expected value of \( \hat{\beta} \) is, conditional on \( X \):

\[
E(\hat{\beta} | X) = E(\beta | X) + E(e | X) = \beta
\]

Thus \( \hat{\beta} \) is an unbiased estimator of \( \beta \).

**Gauss-Markov assumptions**

The Gauss-Markov theorem states that OLS is the best linear unbiased estimator (BLUES) of \( \beta \) if the Gauss-Markov assumptions are valid. The Gauss-Markov assumptions are:

- The model is of the following form: \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e \) where \( \beta_0, \beta_1, \ldots, \beta_k \) are the parameters to estimate and \( e \) is an unobserved error term.

- Data is a random sample of \( n \) observations \((x_{i1}, x_{i2}, \ldots, x_{ik}; y_i)\): \( i = 1, 2, \ldots, n \)

- None of the independent variables in the sample is constant and there is no exact linear relationship among the independent variables.

- The expected value of the error \( e \) is zero given any value of the independent variables.

- The model has to be homoskedastic which means that the variance of the error term is the same for all explanatory variables.

The covariance matrix of the estimated coefficients is described as:

\[
\text{Cov}(\hat{\beta}) = (X^t X)^{-1} \sigma^2
\]  

(3)

where \( \sigma^2 \) is the variance of \( \hat{\beta} \). An unbiased estimate of \( \sigma^2 \) is given by:
\[ s^2 = \frac{|\mathbf{e}|^2}{(n-k-1)} \]

Where \( n \) is the number of observations and \( k \) is the number of covariates.

**Goodness-of-fit: \( R^2 \)**

There are different tools to perform a regression (OLS) with. Besides the estimations of the coefficients, the estimation of the standard error of the coefficients and the estimation of the standard error of the residual most tools do present following result:

\( \text{SST} \) is the sum of squares total and is defined as:

\[ \text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2. \]

\( \text{SSR} \) is the sum of squares residual and is defined as:

\[ \text{SSR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{e}_i^2 = |\mathbf{e}|^2, \]

for observations \( i = 1 \ldots n \). The \( R^2 \) value is defined as:

\[ R^2 = 1 - \frac{\text{SSR}}{\text{SST}}. \]

\( R^2 \) is a measure of the proportion of the sample variation in \( y \) that is described by the model and is hence a measure of the goodness-of-fit.

**OLS: Problems & Remedies**

*Multicollinearity*

The term multicollinearity describes that there is a correlation between covariates in the model. The effect of multicollinearity is that the standard errors of some of the estimated coefficients are very large. This can be seen in the formula below:
\[
\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\text{SST}_j (1-R_j^2)}.
\]

(4)

The standard error is the square root of the variance. \(\sigma^2\) is the variance of the residual. \(\text{SST}_j\) is given by the equation:

\[
\text{SST}_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2,
\]

where \(\bar{x}_j\) is the mean of \(x_{ij}\) for all \(i\). \(R_j^2\) is the R-squared from regressing \(x_j\) on the other covariates. If \(R_j^2 = 1\) there is a perfect correlation between \(x_j\) and the other covariates and accordingly the standard error of \(\hat{\beta}_j\) will be infinitely large.

**VIF**

VIF is an abbreviation for *Variance Inflation Factor* and it is a measure of how the variance of the estimated coefficients is related to multicollinearity. Equation (4) shows the three components affecting the variance

\[
\text{var}(\hat{\beta}_j) = \frac{1}{\text{SST}_j} * \sigma^2 * \frac{1}{1-R_j^2}.
\]

The last factor in the product is the definition of VIF and hence,

\[
\text{VIF}(R_j) = \frac{1}{1-R_j^2}.
\]

Another remedy for the problem of multicollinearity is to exclude the covariate that is correlated with the others. If a covariate is well correlated with some other covariates included in the model then it is already accounted for and therefore not necessary in the model.

A remedy for the multicollinearity problem is to exclude the covariate that is correlated with the others. If a covariate is well correlated with some other covariates included in the model then it is already accounted for and therefore not necessary in the model.

**Heteroskedasticity**

A heteroskedastic model is a model where it cannot be assumed that the error terms for each observation has the same variance:
Heteroskedasticity: \( E(e_i^2) = \sigma_i^2 \)

Homoskedasticity: \( E(e_i^2) = \sigma^2 \),

The Gauss Markov assumption of homoskedasticity is necessary because of two main reasons:

- The formulas are simplified
- Gives the OLS an efficient property

OLS will return inconsistent estimates of the standard deviation of the coefficients if used on a heteroskedastic model. Inconsistent estimates of the standard deviation make the F-test invalid.

**Variable transformation**

Taking the logarithm of the dependent variable can sometimes work as a remedy for heteroskedasticity. Consider the model

\[
y = x\beta + e.
\]

Taking the logarithm as a remedy requires that there are reasons to believe that the standard error of \( e \) is about proportional to \( E(y) \). Larger values of \( E(y) \) would then produce larger standard deviations for \( e \). Since the expected value of \( e \) is zero, \( e \) is (or rather observations of the stochastic variable \( e \)) proportional its standard deviation and hence approximately proportional its standard error. If \( y \) changes so will \( SE(e) \) and accordingly the model is heteroskedastic.

\[
SE(e) = kE(y) = kE(x\beta + e) = kE(x\beta) + kE(e) = kE(x\beta) = kx\beta.
\]

where \( k \) is a constant. And since \( SE(e) \) is also proportional to \( e \) we have:

\[
blkx\beta = e; \ bk = v,
\]

where \( b \) is a constant.

Now we can rewrite the model:

\[
y = x\beta(1 + v),
\]

where the variance of \( v \) is independent of \( x \) and therefore also \( y \). Taking the logarithm:

\[
\ln(y) = \ln(x\beta) + u; \ u = \ln(1 + v),
\]
where the variance of $u$ is independent of $x$ and accordingly independent of $y$ therefore the model is no longer necessary heteroskedastic.

**White’s consistent estimator**

White’s consistent estimator is another way to obtain consistent standard errors for the estimated coefficients of a heteroskedastic model. White’s consistent estimator will always give consistent estimates of the standard errors even when used on a homoscedastic model. The estimator is defined as:

$$\text{Cov}(\hat{\beta}) = (X^tX)^{-1}\left(\sum_{i=1}^{n} e_i^2 x_i^t x_i\right) (X^tX)^{-1}.$$  

**Endogeneity**

The phenomenon endogeneity appears when the error term is correlated with at least one of the covariates. This is a problem that causes inconsistent OLS estimates of the coefficients. For example consider the model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

$$\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}),$$

where $\hat{e}_i$, $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are the OLS estimates obtained by choosing the $\hat{\beta}_j$ that minimizes:

$$S = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}))^2$$

Differentiating $S$ with respect to $\beta_j$ and setting the derivative equal to zero will give us the following OLS first order conditions:

$$\hat{\beta}_2: \quad \frac{1}{n} \sum_{i=1}^{n} x_{2i} \hat{e}_i = 0$$

$$\hat{\beta}_1: \quad \frac{1}{n} \sum_{i=1}^{n} x_{1i} \hat{e}_i = 0$$
\[ \hat{\beta}_0: \frac{1}{n} \sum_{i=1}^{n} \hat{e}_i = 0. \]

To observe if the estimates seem to be close to the true parameter values one should rather evaluate the conditions above at the true parameter values:

\[ \beta_2: \frac{1}{n} \sum_{i=1}^{n} x_{2i}e_i = a \]
\[ \beta_1: \frac{1}{n} \sum_{i=1}^{n} x_{1i}e_i = b \]
\[ \beta_0: \frac{1}{n} \sum_{i=1}^{n} e_i = c. \]

If \( a, b \) and \( c \) are all equal to zero then the true parameter values and the estimates have the same first order conditions and they should be equal to each other, assumed there is a unique solution. But if one or several of \( a, b \) and \( c \) are not equal to zero the estimates will be inconsistent since the true parameter and the estimates no longer has the same conditions. This occurs for endogenous variables. For example consider \( x_{1i} \) to be correlated with the error term and thus endogenous:

\[ \beta_2: \frac{1}{n} \sum_{i=1}^{n} x_{2i}e_i \rightarrow E(x_2e) \text{ as } n \rightarrow \infty; \ E(x_2e) = \text{cov}(x_2, e) = 0 \]
\[ \beta_1: \frac{1}{n} \sum_{i=1}^{n} x_{1i}e_i \rightarrow E(x_1e) \text{ as } n \rightarrow \infty; \ E(x_1e) = \text{cov}(x_1, e) \neq 0 \]
\[ \beta_0: \frac{1}{n} \sum_{i=1}^{n} e_i \rightarrow E(e) \text{ as } n \rightarrow \infty; \ E(e) = 0, \]

where \( \text{cov}(x_1, e) \neq 0 \) is the definition of \( x_2 \) being correlated with \( e \).

**Residual plots**

A way to sometimes spot endogeneity is to simply plot the residual against the covariates to visualize if there is any correlation between them. A strong correlation would mean that the covariate is endogenous. There are no simple remedies for endogeneity but one possible solution is to find some new covariates to replace the endogenous one. These new covariates, called instrumental variables, should be well correlated with the endogenous covariate and uncorrelated with the residual.
Hypothesis testing & Inference

A t-statistic for the hypothesis that $\beta_j = \beta_j^0$ is:

$$t = \frac{\bar{\beta}_j - \beta_j^0}{SE(\beta_j)},$$ \hspace{1cm} (5)

where the p-value for this hypothesis is $Pr(X > t)$ and $X$ has the t-distribution in equation (5). $X$ has this distribution if the covariate has a normal distribution, which they have if the residual has a normal distribution. A low p-value should be interpreted as a small chance that the hypotheses could be true, under the null distribution from the hypothesis. Therefore the hypothesis is rejected if the p-value is low enough. How small the p-value should be for the hypothesis to be rejected is subjective.

To test several hypotheses simultaneously one could use the F-test. The F-statistic for the hypothesis that several coefficients are all equal to zero is

$$F = \frac{(n-k-1)}{r} \cdot \left( \frac{|\hat{e}^2|^2}{|\hat{e}|^2} - 1 \right),$$ \hspace{1cm} (6)

where $\hat{e}$, is the residual from the restricted model and $\hat{e}$ is the residual from the full model. $K$ is the number of covariates, $n$ is the number of observations and $r$ is the number of restrictions. The F-statistic has a $F(r,n-k-1)$ distribution. P-value for the hypotheses that the model could be restricted is $Pr(X > F)$ where $X$ has the distribution in equation (6). Requirements for the F-test is homoscedasticity and normally distributed residuals. If the distribution of the error terms is unknown but they are iid the F-test is asymptotically valid for large values of $n$. But if the model is heteroskedastic the F-test is invalid. To test several hypotheses at a heteroskedastic model one can use Wald’s test.

Wald’s test uses a Wald statistic which is approximate $X^2 (r)$ distributed where $r$ is the number of restrictions. The Wald statistic is $W = \bar{\beta}_2 V^{-1}_2 \bar{\beta}_2$ where $V_2$ is the estimated covariance matrix for $\beta_2$ according to white’s consistent variance estimator. This test is used for hypothesis involving linear constraints of the coefficients, when the model is heteroskedastic.

Method

In the beginning of any form of statistical analysis the problem of modeling consist of two different parts: formulation of a model, with clear and precise definitions of all constituents, and the acquisition of data. In the field of econometrics, and particularly in quantitative social sciences,
these two parts are often hard to match. The inevitable truth is that a statistician is always limited to the available data and this must be kept in mind at all times during the process of modeling.

This section describes how the modeling process was conducted, how data was acquired and how it was handled to produce results.

**Modeling**

*Dependent variable*

The first step of the modeling was to determine the dependent covariate. The main purpose of the project was to model child mortality (Code: CMR). But child mortality can be defined in a few different ways, which all have different causes. Therefore the effects that will be taken into account and studied must be determined beforehand in order to choose the appropriate definition.

Early in the process it was decided that the covariates to be studied would not be the “direct” effects that cause child mortality, such as disease (malaria, pneumonia, diarrhea etc.), malnutrition in the area, droughts etc. This is reasonable to assume since the life expectancies of infants is highly sensitive to these factors (Danzhen, et al., 2013). Instead the focus would be on studying covariates with a more subtle effect, such as the general wealth of the country, crime rates, mean years of schooling and climate. In order for these covariates’ influence on child mortality to actually be detectable, they must be studied during a longer period of time. Therefore the appropriate definition for this project’s purpose is child mortality as death of children of ages between 0 and 4 (up to the 5th birthday).

*Covariates*

A lot of effort was spent on constructing the preliminary model in order to ensure that all covariates suspected to have an effect on child mortality were included. The list of the covariates which were deemed to be relevant and for which data was accessible is presented in the table below.

<table>
<thead>
<tr>
<th>Name of covariate</th>
<th>Definition &amp; Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aid organizations (Code: Helporg)</td>
<td>Dummy variable where 1 means that one or several of the four largest help organisations (UNICEF, SC, RC, MSF) are offering support in the country.</td>
</tr>
<tr>
<td>Average precipitation (Code: AvgPre)</td>
<td>The average precipitation in depth (mm) year 2011.</td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Drought, Epidemic famine, Natural disasters (Code: <strong>Chaos</strong>)</td>
<td>Dummy variable where 1 means that the country has suffered from any catastrophe, in last years, were a significant part of the population’s life standard was affected.</td>
</tr>
<tr>
<td>War (Code: <strong>War</strong>)</td>
<td>Dummy variable where 1 means that the country has taken part in a domestic conflict or war in the last years, with at least 1000 fatalities per year.</td>
</tr>
<tr>
<td>Human Development Index (Code: <strong>HDI</strong>)</td>
<td>A measure of the population health including life expectancy, mean year of schooling and income.</td>
</tr>
<tr>
<td>Colony 1950 (Code: <strong>Colony</strong>)</td>
<td>Dummy variable where 1 means that the country was a colony year 1950.</td>
</tr>
<tr>
<td>Number of hospitals/doctors per capita</td>
<td>More hospitals/doctors probably results in a larger part of the population having access to health care.</td>
</tr>
<tr>
<td>Difference in social classes</td>
<td>A covariate intended to reveal larger diversities in life standard in the population.</td>
</tr>
<tr>
<td>GNI per capita, PPP (Code: <strong>GNI</strong>)</td>
<td>A measure of national wealth and economic power per capita (see appendix) and PPP (see appendix) in the country.</td>
</tr>
<tr>
<td>Continents or areas (Code: <strong>World</strong>)</td>
<td>Several dummy variables to estimate differences in child mortality due to difference in location.</td>
</tr>
<tr>
<td>Maternal mortality</td>
<td>A measure of complications at birth, availability of medically trained personnel at time of birth and disease carried by the mother</td>
</tr>
<tr>
<td>Crude birth rate (Code: <strong>Birth</strong>)</td>
<td>The number of live births occurring during the year, per 1,000 population estimated at midyear</td>
</tr>
<tr>
<td>Unemployment</td>
<td>Levels of long-term unemployment in the</td>
</tr>
<tr>
<td>Covariate</td>
<td>Country</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Population (Code: Population)</td>
<td>Number of inhabitants in the country.</td>
</tr>
<tr>
<td>Homicide rate (per 100 000) (Code: Homicide)</td>
<td>The number of homicides in the country as a measure of level of security and law enforcement in the country.</td>
</tr>
<tr>
<td>Education (Code: Education)</td>
<td>Expected mean years of schooling of parents as a measure of general education levels in the country.</td>
</tr>
<tr>
<td>Illiteracy</td>
<td>Rate of illiteracy in the population</td>
</tr>
<tr>
<td>Improved sanitation (Code: Sani)</td>
<td>Percentage of the inhabitants in the country that have access to improved sanitation (see appendix).</td>
</tr>
</tbody>
</table>

Table 1. A list of all the covariates originally proposed, their explanation and their codes.

The problem of finding data only applied to the level (not dummy) covariates. After a thorough search of different statistical data banks, some covariates had to be eliminated due to insufficient data. This resulted in a drastically reduced choice of covariates to include in the model.

The first proposed model was the following, were more or less complete data was found for the covariates included:

\[
(\text{Child mortality rate}) = (\text{Intercept}) + (\text{War}) + (\text{Helporg}) + (\text{World}) + (\text{GNI}) + (\text{Sani}) + (\text{Birth}) + (\text{Pop}) + (\text{Education}) + (\text{HDI}) + (\text{AvgPre}) + (\text{colony}) + (\text{homicide}) + \text{error}
\]

After consideration a few problems were detected:

- How should the *World* dummy be defined?
- How should *War* be defined?
- What becomes of the ceteris paribus interpretation of the coefficients when including *Birth* and *Pop*? What would it mean to keep one fixed while letting the other vary? This also applies to *Edu* and *HDI* since the latter contains, not only an income index based on GNI, but an education index based on years of schooling as well.
- Will endogeneity occur if the covariate *Birth* is included?
• Should the model be linear or is it more likely to be nonlinear?

It was believed that grouping similar areas in dummies would detect differences between groups that are not already measured by the model. This was to ensure that even though some relevant covariates could be missing from the model, the addition to child mortality due to these effects could be measured. Also, these groups could help divide larger countries with big differences within the country into smaller and more homogenous areas. This proved to be very challenging since, depending on the criteria used, several of different groups could be used and making the division more detailed means creating more dummies which in turn decreases the accuracy of the estimates.

One of the preferred divisions was a grouping based on climate. Regions with similar climates usually produce similar crops and suffer from the same types of weather induced problems on life standards. This was also coincided with some of the reason for using the covariates Average Precipitation and Chaos in the model. Many different climate classification systems exist and one of the most famous and widely used is the Köppen climate classification system (Wiki: Köppen). It classifies climate system in five large groups and several of subgroups within the larger. But the main problem is that many countries, and specially countries with large areas, belong to several of the different categories, which renders categorizing practically impossible. This led to the decision to include Average Precipitation in the model and transforming the World dummy into the new dummy U-nation (Code: Uland), describing whether the country is a developing or developed (industrialized). Many different definitions exist to determine whether a country is developing or not based on for example, HDI, GDP per capita or Gross national happiness. In choosing the definition relevant to the problem three factors were considered: HDI is dependent on both GNI and Education, data for the covariate Education was not complete and finally, using GDP to describe whether a country is developed or not can be very misleading since it does not reveal skewness in income distribution. This would mean that a country may have a relatively high GDP per capita while the majority of the inhabitants may belong to the low-level income group.

The suggested solution to the problems mentioned in the paragraph above is to remove Education, due to the data insufficiencies, and HDI from the model and instead use a country's HDI value to determine if it is a developing country. This in turn, with a pre-determined threshold for the maximum HDI value that constitutes a developed country, would be used to form the U-nation dummy with 1 meaning that the country is developing.

In determining the covariate War, similar problems as those mentioned above were discovered. It proved challenging not only to find trustworthy data, but influence of war on child mortality rate is also debatable. Obviously there would be a direct effect if civilian casualties, at least some of them children, occurred frequently but the general suffering of a country due to an ongoing war might be harder to estimate. Whether the war is domestic or not, as in the case on a nation sending troops to aid an ally should also change in which ways war affects child mortality. The end of a war does not necessarily mean that the living standards return to pre-war levels, but on the contrary it is
reasonable to suspect that war has various long term effects on a country’s well being. In summary, a definition accounting for not only the duration of the war but also the location and number of fatalities would need to be created and based on the available data. It was not deemed to be feasible to create such a covariate. Therefore the covariate War was excluded from the model.

A more subtle problem in the modeling was discovered in the interpretation of the coefficients for the covariates Birth and Population. One of the advantages of multiple linear regression is the ceteris paribus interpretation of the coefficients. Ceteris paribus enables detailed study of the effect at the dependent variable from one covariate in particular when keeping all other covariates fixed. The interpretation of the coefficient for Birth would be the difference keeping all other factors fixed which would include keeping the covariate Population fixed. But varying the country’s birth rate while keeping the population fixed would inevitably result in strange interpretations. For example, increasing the birth rate and keeping the population fixed requires that a part of the population, corresponding to the increase in birth rate, cease to exist which can be attributable to two different causes: death or emigration. There are no guarantees that the increase in birth rate is matched by the two. An increase in the number of deaths or emigrants would also suggest that the conditions in the country have been altered, which could be attributable to something not included in the model and therefore residing in the error term, rendering the covariate endogenous. Hence one of the covariates is redundant and causes more problems than it solves. Out of the two Birth was deemed to be more relevant, which led to the exclusion of Population from the model.

The cause and effect relationship between child mortality rates and birth rates is not absolute. It might be reasonable to assume that a larger child mortality rate is related to some unobserved factor not included in the model that is correlated with birth rate. For example, in countries with larger child mortality rates it could be common that women give birth to more children. This would mean that birthrate is dependent on child mortality as well as the other way around. A two way correlation like this could be an indication of endogeneity. To spot endogeneity, the residuals from the regressions can be regressed on each of the covariates to study if there are any trends. If there seems to be any form of relationship between the residual and any of the covariate, the model is endogenous. This will be examined in the analysis section.

It might also be reasonable to assume that the effects of multiple of the covariates are not of linear nature. The influence of more levels of annual precipitation might not change the mortality rates by a fixed amount, but by a percentage. This is possible if the logarithm of dependent variable is used instead. Another reason to take the logarithm of child mortality might be to reduce any heteroskedasticity in the model (see Heteroskedasticity p.7). It can also be legitimated to use the logarithm when the dependent variable is a positive quantity (Lang, 2013, p.41).

For all the reasons discussed above the model was reformulated into the following form:

\[
\log(\text{Child mortality rate}) = (\text{Intercept}) + (\text{Helporg}) + \log(\text{GNI}) + (\text{Sani}) + (\text{Birth}) \\
+ (\text{Colony}) + (\text{Uland}) + (\text{Homicide}) + (\text{AvgPre}) + \text{error} \\
\]

(7)
Data acquisition

The next phase in the process was to collect data for the chosen covariates. A full set of data for every country proved very difficult to find. By searching on several large statistical databases, the ones presenting the most coherent data with citations and references were the United Nations Development Programme databank and the World DataBank. UNDP is a branch of the United Nations organization offering comprehensive data of various nature and type, aid and crisis programs for developing nations and commissioning investigations and reports on world health. WDB is a branch of the World Bank Group, which is a collaboration offering technical and financial assistance to developing countries around the world in an effort to reduce worldwide poverty.

A majority of the used data was downloaded from UNDP but much of the same data could also be found on WDB as well. Some difference in the completeness of the data was also discovered, showing lacking data points for different years. It was the year with most coherent and complete data was found to be 2010 and hence this year was chosen for the creation of the model.

Regression Software

After collecting and structuring all data, regressions were calculated. Essentially all regressions and calculations were made using Microsoft Excel. Q-Q plots and plots of the residuals regressed on each covariate were made using the software R.

All the work regarding discussions, remedies and changes that were made in between the first and the final model are presented in the next two chapters Analysis and Discussion.
Analysis

Endogeneity

A regression was made in excel for the first model (equation 6). The residual was saved for this regression. There was performed a test to see if there were any more arguments for birth being an endogenous covariate. To study if endogeneity is present in the model, the residuals from the regression were plotted against each covariate using R (see figure 1).
Figure 1. Plots of the residuals regressed on each covariates. The dotted red line is the OLS line. The green curve is a nonparametric-regression curve calculated by the LOWESS, or LOESS, method (“Locally weighted scatterplot smoothing”).

The plot of the residuals against *Birth* reveals an approximately linear relationship between the two, providing more reason to the belief that including birth rate in the model introduces
endogeneity. It is therefore removed. The new model and the new residual plots are presented below:

$$\log(CMR) = (\text{Intercept}) + (\text{Colony}) + (\text{Homicide}) + (\text{Helporg}) + (\log\text{GNI}) +$$

$$+ (\text{Sani}) + (\text{Uland}) + (\text{AvgPre}) + \text{error} \quad (8)$$

Figure 2. Plots of the residuals regressed on each covariates for the model without birth rate.
From figure 2, it is reasonable to assume that there is no clear linear correlation between the residual and any of the covariates and accordingly the conclusion is made that there should no longer be any endogenous covariates in the model. The covariate Sani is displaying some correlation with the residual but it is not as evident as for birth rate. Since there are no other reasons to believe that Sani could be endogenous it was decided to keep it in the model.

**Heteroskedasticity**

The remedy for heteroskedasticity is White’s consistent estimator which was used to calculate the consistent estimates of the variances of the coefficients. Another remedy, which was also mentioned in the section addressing heteroskedasticity, is the one taking the logarithm of the dependent variable. Taking the logarithm might work if the standard error of the residuals $e_i$ are suspected to be proportional to the expected value of the dependent variable $E(y)$. It is easy to believe that higher values of child mortality rates are accompanied by a larger uncertainty in their values. Most of the countries showing higher values of child mortality rates are developing countries. Due to underdeveloped infrastructure in these countries it is plausible that reported CMR carry larger inaccuracies. Therefore it is believed that the standard error of $e$ is about proportional to the expected values of the dependent variable.

Quantitatively the heteroskedasticity in the model was analyzed by studying the difference between the standard errors of the coefficients reported from the regressions and the standard errors calculated using White’s method. This was done for the model in equation (6) and the same model but with Level CMR as regressand.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0,2133</td>
<td>0,2775</td>
<td>0,0641</td>
<td>0,3009</td>
</tr>
<tr>
<td>Colony</td>
<td>0,0322</td>
<td>0,0356</td>
<td>0,0034</td>
<td>0,1049</td>
</tr>
<tr>
<td>Homicide rate</td>
<td>0,0013</td>
<td>0,0011</td>
<td>-0,0001</td>
<td>-0,108</td>
</tr>
<tr>
<td>Help-org</td>
<td>0,0382</td>
<td>0,0440</td>
<td>0,0059</td>
<td>0,1513</td>
</tr>
<tr>
<td>log(GNIpc in PPP terms)</td>
<td>0,0596</td>
<td>0,0723</td>
<td>0,0127</td>
<td>0,2136</td>
</tr>
</tbody>
</table>
### Table 2
The reported errors from Excel can be found in the column named standard error. The second column contains the errors calculated using White’s consistent estimators. The third column is the difference between the first and second and the fourth is the relative difference between White’s errors and the standard errors.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Standard Errors, Level: CMR</th>
<th>White’s estimates, Level: CMR</th>
<th>Difference (White’s - Standard)</th>
<th>Difference / Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>25.0919</td>
<td>32.2833</td>
<td>7.1914</td>
<td>0.2866</td>
</tr>
<tr>
<td>Colony</td>
<td>3.7935</td>
<td>3.8472</td>
<td>0.0537</td>
<td>0.0148</td>
</tr>
<tr>
<td>Homicide rate</td>
<td>0.1487</td>
<td>0.1162</td>
<td>-0.0325</td>
<td>-0.2186</td>
</tr>
<tr>
<td>Help-org</td>
<td>4.4956</td>
<td>4.0180</td>
<td>-0.4776</td>
<td>-0.1062</td>
</tr>
<tr>
<td>log(GNIpc in PPP terms)</td>
<td>7.0133</td>
<td>8.0354</td>
<td>1.0221</td>
<td>0.1457</td>
</tr>
<tr>
<td>Improved sanitation (%)</td>
<td>0.1246</td>
<td>0.1381</td>
<td>0.0135</td>
<td>0.1086</td>
</tr>
<tr>
<td>U-Nation</td>
<td>0.0021</td>
<td>0.0019</td>
<td>-0.0002</td>
<td>-0.1102</td>
</tr>
<tr>
<td>Average precipitation</td>
<td>5.8788</td>
<td>4.2775</td>
<td>-1.6014</td>
<td>-0.2724</td>
</tr>
</tbody>
</table>

### Table 3
This table is identical to Table 2, but the values are for the model with Level CMR.

White’s consistent estimator will always give the consistent estimates of the standard deviation of the coefficients, regardless of the model being heteroskedastic or homoscedastic. If the model would have become homoscedastic after taking the logarithm of CMR there would only have been a slight difference between the standard errors with and without White’s estimator. Since white’s changed the standard error of the coefficients with as much as 30% the conclusion was made that the model is heteroskedastic. Therefore the t-tests have to be calculated using the standard errors obtained with White’s and the F-test has to be replaced by Wald’s test. The t-test and the F-test also assume normally distributed residuals:
Quantile-Quantile plots for the errors are presented below. A curve close to the marked line indicates that the two distributions are linearly related.

Figure 3. Normal QQ-plot for the model in equation 8.

The QQ-plot suggests that the residuals are normally distributed.
**Multicollinearity**

To study linear relationships between the covariates, regressions were run for each of the covariates as the dependent variable regressed on the others. The results are presented in the table below.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>VIF</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colony</td>
<td>1.17436</td>
<td>0.148472</td>
</tr>
<tr>
<td>Homicide</td>
<td>1.307395</td>
<td>0.23512</td>
</tr>
<tr>
<td>Helporg</td>
<td>1.611075</td>
<td>0.379296</td>
</tr>
<tr>
<td>log(GNI)</td>
<td>4.801183</td>
<td>0.791718</td>
</tr>
<tr>
<td>Sani</td>
<td>3.020575</td>
<td>0.668937</td>
</tr>
<tr>
<td>Uland</td>
<td>2.106071</td>
<td>0.525182</td>
</tr>
<tr>
<td>AvgPre</td>
<td>1.09484</td>
<td>0.086625</td>
</tr>
</tbody>
</table>

Some of the covariates display a rather large $R^2$ (marked red in **TABLENUMBER**) suggesting that the multicollinearity is of notable size. One of the red marked covariates might accordingly be removed from the model. T-tests were made to further investigate whether or not to exclude these covariates.

**Model selection**

T-tests from the regression, with White’s consistent estimates of the standard errors of the coefficients, are presented below:

<table>
<thead>
<tr>
<th>Covariates</th>
<th>t-statistic (Using White’s errors)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>10.245</td>
<td>1.18082E-19</td>
</tr>
<tr>
<td>Colony</td>
<td>3.657</td>
<td>0.000335462</td>
</tr>
<tr>
<td>Homicide</td>
<td>0.566</td>
<td>0.572173012</td>
</tr>
</tbody>
</table>
Table 4. t-statistics and p-values for each covariate corresponding to the hypothesis that the covariate should be excluded.

Because of the low p-values for $\text{log(GNI)}$, $Sani$ and $Uland$ none of these can be excluded from the model. The hypothesis that $AvgPre$, $Helporg$ or $Homicide$ should be excluded separately had the largest p-values. These values suggest that further investigation is needed for these three covariates. Using Wald tests, three different restricted models were compared. For each model two of the insecure covariates mentioned above had been removed simultaneously.

Table 5. Wald-statistics and p-values for hypothesis with different variables excluded.

The p-value from the Wald-test shows that all three hypotheses could be rejected at a 5% level of significance. This means that two covariates cannot be excluded from the model simultaneously at this level of significance. The large p-value for $Homicide$ compared to $AvgPre$ and $Uland$ indicates that $Homicide$ might be removed. Since BIC is not applicable on a heteroskedastic model values for $R^2$ and $SE(e)$ were compared between the full model and the model without $Homicide$ to determine whether to exclude $Homicide$ or not:

Table 6. Table with R square values and standard errors for the model with $Homicide$ and the model without.

Since the differences in $R^2$ and $SE(e)$ between the models are very small there is no reason to keep $homicide$ in the model. The final model is:
\[ \log(CMR) = 2.834 + 0.312(Uland) - 0.330(\log(GNI)) - 0.00682(Sani) - 0.0000446(AvgPre) + 0.113(Helporg) + 0.132(Colony) + error \quad \text{(9)} \]

**Discussion**

Due to insufficient data, several of the originally proposed covariates had to be dropped from the model. Neither literacy, unemployment nor any covariate measuring general health were included in the model and these are believed to greatly influence child mortality. *HDI* was dropped from the model. The R² value for the final model also shows some room for improvement. It is possible that for some of the covariates, proxy variables or instrumental variables could have been found to further help explain child mortality. This is certainly the case for the covariate *Birth*, which was excluded from the model only because it displayed symptoms of endogeneity. Finding suitable instrumental variable to replace it was out of the scope for this project. But as one can see below there is an improvement in R² with birth rate included:

<table>
<thead>
<tr>
<th>Model with:</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth rate included</td>
<td>0.8542</td>
</tr>
<tr>
<td>Birth rate excluded</td>
<td>0.8867</td>
</tr>
</tbody>
</table>

Table 7. R square values for the model with and for the model without the endogenous covariate birth rate included.

The conclusion is that one should include birth rate if the model is used for prediction but as for a structural model, birth rate should be excluded.

Different reasonable combinations of variable transformations could also have been used to study if the model could have been improved regarding increased R² value, reduced standard error for the residual or as a remedy for multicollinearity between some covariates. Some transformations were investigated (see appendix) but none of these showed any significant improvements.

All data used was obtained from two large international and independent organizations. Information on how the data was originally collected was found for all covariates included in the model. When looking at the model it has to be kept in mind that the model is based on observational data. Despite the fact that the data is from reliable sources, the method used to obtain data will inevitably affect the error and the uncertainty in data. Parts of data, specifically some data points for CMR and Sani, are not measured values. They are estimated values calculated using some more or less advanced model (Alkema, New, 2013). This could mean that any form of modeling using the estimated values only "re-discover" the original equation data was estimated from, hence showing great compatibility between model and data. But the model used to estimated child
mortality is much more complex than the regular linear regression (Alkema, New, 2013). And it was used primarily for underdeveloped countries where data was available but inconsistent among different sources.

Further investigation of the cause and effect relationship between birth rates and mortality rates could also have led to new insights about missing covariates and it is plausible that new covariates could have been found. The covariates used may not be the optimal but data exists, obviously, for all of them. And the covariates used are fairly general values and not specific nor overrepresented for any country in general. This means that estimates for child mortality rates could be calculated using easily accessible values. And due to the rather large $R^2$ value, these estimated would offer quite valuable information on how high the child mortality rates could be, making it particularly useful for undeveloped countries or agencies with limited resources. Although these values should not be taken for absolute truths since they originate from a crude and simple model, they offer guidelines for mortality rates. The worth of any model is based on its complexity relative to its accuracy and the model suggested in equation (9) is a compromise between the two.

**Conclusion**

This is the final model:

$$
\log(\text{Child mortality rate}) = 2.834 + 0.312(Uland) - 0.330(\log(GNI)) - 0.00682(Sani) - 0.0000446(AvePre) + 0.113(\text{Helporg}) + 0.132(\text{Colony}) + error
$$

Where the signs for the covariates are as expected. The positive sign for Help should not be interpreted as “help organizations are increasing child mortality”. Instead it should be interpreted as “help organizations are present in countries with higher child mortality rate”.

Improvements could be made on all stages of the modeling to increase the accuracy of the model. But results show that it can be useful for not only a structural interpretation, but as a predictive model as well. Nine data points, approximately 5% of total data, were removed from the regression to study the predicted values and this showed good agreement with the actual values (see Appendix).
References

Books:

Lang, H. *Topics on Applied Mathematical Statistic*, version 0.97. (Nov.2013). KTH Teknikvetenskap


Reports:


World Wide Web pages:


Figure 4. A plot of the residuals versus the fitted values. This visualizes the spread of the residuals.
<table>
<thead>
<tr>
<th>Country</th>
<th>y pred</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>zimbabwe</td>
<td>2.019463</td>
<td>1.90309</td>
</tr>
<tr>
<td>zambia</td>
<td>2.077772</td>
<td>2.045323</td>
</tr>
<tr>
<td>Yemen</td>
<td>1.699203</td>
<td>1.886491</td>
</tr>
<tr>
<td>vietnam</td>
<td>1.304275</td>
<td>1.361728</td>
</tr>
<tr>
<td>venezuela</td>
<td>1.281383</td>
<td>1.255273</td>
</tr>
<tr>
<td>Vanuatu</td>
<td>1.567161</td>
<td>1.146128</td>
</tr>
<tr>
<td>Uzbekistan</td>
<td>1.490129</td>
<td>1.716003</td>
</tr>
<tr>
<td>Uruguay</td>
<td>1.08302</td>
<td>1.041393</td>
</tr>
<tr>
<td>United states</td>
<td>0.713857</td>
<td>0.90309</td>
</tr>
</tbody>
</table>

Table 8. Predicted values compared to the actual values. The countries data points excluded from the regression.