Int J Child Health Hum Dev 2016;9(2):00-00.

Running title: Global infant mortality

A study of time trend in global infant mortality rates: Regression with autocorrelated data

Ming Yang, MS and Dejian Lai, PhD

Department of Biostatistics, The University of Texas, School of Public Health, Houston, Texas, United States of America

Abstract: Infant mortality rate (IMR) is considered to be one of the most important indices of a country's well being. Countries around the world and international organizations such as the World Health Organization are dedicating their resources, knowledge and energy to reduce the infant mortality rates (IMR). The well-known Millennium Development Goal 4 (MDG 4), which is an example of such commitment, aims to achieve a two-thirds reduction of the under-five mortality rate between 1990 and 2015. Many statistical tools are developed for data analysis by assuming independence of observations. However, IMR data collected over time forms a time series. The repeated observations of IMR time series are usually not statistically independent. In modeling the trend of IMR, it is necessary to account for these autocorrelations. In this article we proposed to use the general linear models to take into account the autocorrelations to model worldwide IMR. We compared results from general linear model with correlation structure to that from ordinary least squares method to investigate how significantly the estimates change. Our analysis showed that results from these two methods were different for global data but not for specific countries except for two special cases and the discrepancy could be significantly different when considering the population size of the countries. We modeled the trends of IMR from the 1950s to 2010s for selected countries. Our results quantified the trends of IMR over time and measured the difference across countries.

Keywords: Infant mortality rate, regression analysis, time series, autocorrelation

Correspondence: Professor Dejian Lai, The University of Texas School of Public Health, RAS 1006, 1200 Herman Pressler, Houston, TX 77030, United States. E-mail: dejian.lai@uth.tmc.edu

Submitted: December 23, 2014. ***Revised:*** January 01, 2015. ***Accepted:*** January 10, 2015.

Introduction

Infant mortality rate (IMR) is defined as the number of deaths within one year of age per 1,000 live births (1). The IMR is considered to be one of the important indicators of a population's health and well being along with life expectancy (LE) and under-five mortality rate (U5MR) (2,3). In the year of 2010, the IMR of the United States was 6.6 per 1000, compared with 66.9 per 1000 in Sudan (4). This tenfold difference in IMR reveals a large gap in health conditions between these two nations. As IMRs from developed countries are generally lower than those in developing countries, the IMR is also a reflection of general socioeconomic conditions across different countries.

With improvement in socioeconomic conditions, improvement in health care for deliveries and newborns, and with recent efforts made by the World Health Organization (WHO), the global IMR has been decreasing in most countries of the world (5). According to the Millennium Development Goal 4 (MDG 4), the global mortality rates for children who are younger than five years old are supposed to be reduced by two-thirds from 1990 to 2015 (6,7).

Although the IMR has declined to a relatively low level in some countries in recent decades, there are still many infants, especially those from the developing countries, who are suffering from many risk factors that lead to high IMR. Our study of time trend of IMR over the past a few decades will provide the health organizations and the general public a clear insight into this global health issue. In our article, we have investigated recent achievements regarding reductions in the IMR, globally with a special attention on specific nations or regions. We modeled the current trend of IMRs and measured gaps from the goal as set in the MDG4 (8), etc. The results from this study will help us to better understand the trend of IMR to formulate strategies for improvement in the IMRs and to decide where to focus our attention and efforts and how to allocate the health resources around the world more effectively (9).

Methods

In this study we used countries as the study units (2). For each of the countries, annual IMR data were collected and the earliest time point in records starts from 1932. In our regression analysis model, the response variable is the IMR and the independent variable is the time (year).

Our data set includes annual infant mortality rate for 195 countries/regions around the world between the times of May 1932 to May 2010. The data set is mainly collected from the Child Mortality Estimates (CME) website: http://www.childmortality.org (4), compiled with those from the World Bank website (7). A closer examination of the data set showed that for most of the

countries/regions, the data were mostly missing before 1950. To deal with this issue, we set the time interval starting from 1950 for our study. Countries with sparse data were not included in the study. Also we took consideration to cover countries from different continents and at different level of socioeconomic status in order to investigate the influence of those factors on the trend of the IMR. To make it representative and comprehensive, we selected the countries and regions that are listed in table 1.

Statistical method

The linear regression model is formulated as $\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$, where n indicates the number of observations, p indicates the number of covariate. We assume the error term $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$, which leads to the ordinary-least-squares (OLS) estimator of β (10):

$$\hat{\beta}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{1}$$

with variance-covariance matrix:

$$Var(\hat{\beta}_{OLS}) = \sigma^2(X'X)^{-1} \tag{2}$$

However, in our study this ideal assumption of error terms cannot always be satisfied–there may be some correlations among the error terms, which can be modeled with a more general assumption of $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$, where the error-covariance matrix $\boldsymbol{\Sigma}$ is symmetric and positive-definite. Different diagonal entries in $\boldsymbol{\Sigma}$ represent non-constant error variances and nonzero off-diagonal entries correspond to correlated errors. Based on this more general assumption, our estimator of $\boldsymbol{\beta}$ becomes the generalized-least-squares (GLS) estimator if the variance-covariance matrix is known:

$$\hat{\beta}_{GLS} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \tag{3}$$

with variance-covariance matrix:

$$\mathbf{Var}(\hat{\beta}_{\mathbf{GLS}}) = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1} \tag{4}$$

Time-series data is a typical example in which the error terms from a regression model are correlated: since the data are obtained from multiply measurements on the same subject over time (11). In our study, the observations of infant mortality rates were collected annually from each of the countries, which means that the data are equally spaced with intervals of one-year period. This type of data

belongs to discrete time series case (11). The IMR data measured sequentially over time in a given country are very likely to be correlated due to the similarity of the socioeconomic condition, health care condition, etc. within that country. Ignoring correlations in the model building steps may lead to inaccurate or even misleading results (12).

Time-series regression models

In this study, we assume that the regression errors is stationary (13, 14), that is, the covariance of two error terms depends only upon their separation k in time and is independent of time t, i.e. $cov(\varepsilon_t,\varepsilon_{t+k}) = cov(\varepsilon_t,\varepsilon_{t-k}) = \sigma^2\varrho_k$, where $\varrho_k < 1$ is the error autocorrelation at lag k (14). (With underlying assumptions $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and $var(\varepsilon_t) = \sigma^2$)

Thus, the error-covariance matrix can be written as following:

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \rho_{n-4} & \cdots & 1 \end{bmatrix} = \sigma^2\mathbf{P} \qquad (5)$$

Since the true value of σ^2 and ϱ's are unknown, we cannot directly apply above result to finding the GLS estimator of $\boldsymbol{\beta}$ in this time-series regression setting. In addition, the large number, that is $(n-1)$, of ϱ's makes it impossible to estimate them without specifying any structure for the autocorrelated errors.

The first-order autoregression process, *AR(1)*:

$$\varepsilon_t = \phi\varepsilon_{t-1} + \upsilon_t, \qquad (6)$$

where ε_t and ε_{t-1} are the error terms at time t and $t-1$ respectively; υ_t is the Gaussian white noise, which follows $N(0, \sigma_v^2)$. In this model, it can be shown that $\varrho_1 = \phi$, $\varrho_k = \phi^k$, and $\sigma^2 = \sigma_v^2/(1-\phi^2)$ (14). Because $|\phi|<1$ is assumed for stationary autoregressive processes, the error autocorrelation ϱ_k approaches to 0 exponentially as s increases. The variance-covariance matrix under *AR(1)* model can be expressed as

$$\mathbf{\Sigma} = \sigma^{\mathbf{2}} \begin{bmatrix} 1 & \phi & \phi^2 & \phi^3 & \dots & \phi^{n-1} \\ \phi & 1 & \phi & \phi^2 & \dots & \phi^{n-2} \\ \phi^2 & \phi & 1 & \phi & \dots & \phi^{n-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \phi^{n-3} & \phi^{n-4} & \dots & 1 \end{bmatrix} = \sigma^{\mathbf{2}}\mathbf{P} \qquad (7)$$

In our modeling, we first fitted the data using linear regression model by assuming **P = I** in Model (7); after that and before deciding whether it's appropriate to use GLS method to model the data, we tested the correlation among the observations with two methods: one is the Durbin-Watson test and the other is plotting the correlogram (15). When the existence of autocorrelation was detected, we built the model for the IMR data based on time-series and compared the results with those from OLS model.

Results

Based on data from World Bank (7), the trends of IMR between 1950 and 2010 are depicted in figure 1, in which the three decreasing curves represent the time trends globally, in developed countries/regions, including of Europe, Northern America, Australia/New Zealand and Japan, and developing countries/regions, including all regions of Africa, Asia (excluding Japan), Latin America and the Caribbean plus Melanesia, Micronesia and Polynesia, respectively. Figure 1 shows that the IMR in developed countries is much lower compared with that in developing countries and the overall global IMR, which is calculated by combining all the countries in previous two categories, lies in between. We also plotted the time trend in IMR for each of the selected countries (see figures 3-6) with the IMR values from 1950 to 2010.

Generally, it can be observed from these plots that although the IMR from developing countries was much higher, it decreased more rapidly than the IMR in developed countries. The difference of IMR between them is much smaller in 2010 compared with the difference in 1950 as shown in the figure. The absolute value of the slope of the regression line, modeling the time trend in IMR, is larger in developing countries than that in developed countries.

To investigate the relationship between IMR and time quantitatively, we implemented the method of regression analysis and built models of IMR in terms of time by estimating the regression coefficient in the model. We first tried simple OLS method without taking into account the autocorrelation to model the IMR time trend, and then we built new models that take into account autocorrelation. The results from these two methods were compared. We took the United Stated as an example to conduct the above analyses. The results of the other selected countries were obtained similarly.

Fitted models with OLS method

Without considering the autocorrelation among the observations, first we simply tried to fit the regression model for the IMR data by OLS method. Table 2 (upper) shows the estimations of coefficients using the OLS regression of United States (1950–2010) and Figure 2 (left) shows the IMR trend in time of United States between the years of 1950 to 2010, along with the regression line fitted to the data. The function below shows the relationship between IMR and time:

$$\text{IMR} = -0.456 * \text{year} + 917.8335,$$

which means that the IMR of United States decreased 0.456 per 1000 per year during this time interval.

Fitted models with GLS method

The R function arima was used to build the time-series regression model. For the data from United States we chose the order (1,0,0), which is equivalent to the first order autoregressive model, that is *AR(1)*. Table 2 (lower) shows the output, and the resulting model is:

$$\text{IMR} = -0.427 * \text{year} + 864.2053,$$

which means that the annual decreasing rate of IMR in United States was 0.427 per 1000 between 1950 and 2010 with $\varrho = 0.9926$, $\sigma^2 = 0.06588$. The regression line is shown in Figure 2 (right) along with the IMR trend curve similar as in Figure 2.

Comparison of time trends of the IMR from the selected countries

Figures 3-6 present country-wise time trends of the IMR by continent. To compare the equality of the decreasing rate (the slope) of the IMR between developed and developing countries, we applied the Wilcoxon rank sum test. The Wilcoxon rank sum test gave us p-values of 0.0008 in comparing the equality of slopes for developed countries and developing countries from OLS method and 0.0003 from the GLS method respectively.

To compare the slopes for two specific countries, we implemented two-sample t-test (based on the relatively large sample size, which is around 60 for the selected countries). And the results are listed in Table 3. Note that all slopes are negative. Positive difference indicates that first country has smaller slope, in absolute value,

than the second, indicating smaller decreasing rate of IMR, vice versa.

Comparison of the results from two model-building methods

In Table 4 we listed the slopes and corresponding standard errors for all the countries from both models. As mentioned previously, regardless of the close value of the coefficients, they still can be statistically different. To check if the regression coefficients from two regression methods for each specific country are statistically different, we assumed that the two estimates were independent and conducted a series of *z*-tests. The test results are listed in Table 5. To check the overall similarity of the two model-building methods, we treated the 19 slope estimates from OLS and GLS methods as two groups respectively and conducted the two-sample paired *t*-test.

Discussion

Time trend of the IMR were studied for selected countries in this study, in which the graphical representation showed how the IMR has declined during the past six decades. After investigating the autocorrelations among the observations using Durbin-Watson test as well as the autocorrelation function (acf) and partial-autocorrelation function (p-acf), both produced significant results, i.e. the IMR data are statistically correlated (not shown in this paper). We proceeded to build up the time-series regression models for the IMR in term of time based on the results from autocorrelation tests and then compared with those results from OLS methods.

Time trends of the IMR for different regions

Among all the selected countries, the situation of IMR was the worst in Africa and the best in Europe and Oceania. Although they all have achieved a huge decline in IMR for the past six decades, the general IMR in Africa is still around the level that the European countries achieved in 1950s (around 30 per 1000). One special case is in Zambia, instead of a decrease, the IMR increased during the 1980s to 1990s. In Asia and America the case was similar. There was still a gap of IMR between developing and developed countries, however the developing countries are catching up.

The baseline level of IMR was much higher in developing countries compared with developed countries; however the difference became smaller during the past six decades since the IMR decreased faster in those developing countries. Wilcoxon rank sum tests showed that the decreasing rates between developing countries were significantly higher than those in developing countries either based on OLS or GLS method (with p-values 0.00081 and 0.0003 respectively).

From table 3, for example, we can see that the slopes for China and India are significantly different (with difference equals to 0.6751 and p-value <0.0001),

which indicates China has smaller IMR decreasing rate than India (in absolute value of the slopes). The same conclusion can be drawn form the comparison between US and China (US has smaller IMR decreasing rate than China). Comparisons can also be done for other countries using similar method. Although great achievement had been made in reducing the IMR in developing countries for the past few decades, attentions, efforts, and resources are still desirable and necessary in the future. For the countries that already have lower IMR, more effective methods need to be considered and applied to lower the IMR further.

The difference in two model-building methods

Two regression lines were presented in Figure 3 to Figure 6, graphically showing the comparison between these two model-building methods (solid line stands for OLS regression and dashed line stands for GLS regression). Neither from the numeric values in Table 4 nor from the graphs the results from OLS method are greatly different from those from the GLS method at the first glance – they have similar absolute values or the two regression lines lie close to each other. Country-wise statistical tests for comparison of the difference in the estimates of slopes also turned out to be statistically insignificant expect for two countries as we can see from Table 5: Zambia and Morocco had significant results, which means the regression slopes from OLS and GLS methods are significantly different with each other. For all other countries we failed to reject the null hypothesis and both OLS and GLS methods produced similar estimates of slopes.

However the overall two-sample paired *t*-test showed that the regression coefficients from the OLS and GLS methods are statistically significantly different from each other with *p-value* turned out to be 0.0435, which was slightly less than 0.05, indicating the slops from OLS and GLS methods were statistically different with a significance level of 5%. In addition, when we take account of the whole population size, this "small" difference could result in huge discrepancy: for example, the population of US was around 308,745,538 in 2010. If we apply above two methods separately the difference in IMR for the year of 2010 can be calculated as

$$|(-0.427 * 2010 + 864.2053) - (-0.456 * 2010 + 917.8335)| = 4.6618/1000$$

If we multiply this value with the population of US, the difference is 1,439,310 deaths of infant, which is not minor. Same for other nations, when we take into account the population size, the OLS and GLS methods actually lead to different results that cannot be ignored. As we have shown that the IMR data are highly correlated, it is highly necessary to choose GLS method instead of simply apply OLS regression method.

References

1. National Center for Health Statistics. Health, United States, 2007 with

chartbook on trends in the health of Americans. Hyattsville, MD: NCHS, 2007.

2. Carl OS, Marie R, Hans R, Stefan P, Anna ME. Socioeconomic determinants of infant mortality: A worldwide study of 152 low-, middle-, and high-income countries. Scand J Public Health 2007;35:288–97.

3. Singh GK, Yu SM. Infant mortality in the United States: Trends, differentials, and projections, 1950 through 2010. Am J Public Health 1995;85(7):957–64.

4. Child Mortality Estimates (CME) Info. Child mortality report 2011. Accessed 2014 May 20. URL: http://www.childmortality.org

5. UNICEF. Statistics by area/child survival and health. Under-five mortality. Accessed 2014 May 20. URL: http://www.childinfo.org

6. WHO. Millennium development goals (MDGs). Geneva: World Health Organization, 2000.

7. The World Bank. World development indicators. Level and trends in child mortality. Report 2011. Accessed 2014 May 20. URL: http://data.worldbank.org/indicator

8. Julie KR, Jake RM, Abraham DF, Wang H, Alison LR, Laura D, et al. Neonatal, postneonatal, childhood, and under-5 mortality for 187 countries, 1970–2010: a systematic analysis of progress towards Millennium Development Goal 4. Lancet 2010;375(9730):1988–2008.

9. Singh GK, van Dyck PC. Infant mortality in the United States, 1935–2007: Over seven decades of progress and disparities. A 75th anniversary publication. Rockville, MD: Health Resources Services Administration, Maternal Child Health Bureau, US Department Health Human Services, 2010.

10. Fox J. Time-series regression and generalized least squares. Appendix to: An R and S-plus companion to regression. Thousand Oaks, CA: Sage, 2002.

11. Chatfield C. Time-series forecasting. London: Chapman Hall/CPC, 2000.

12. Diggle PJ. Time series: A biostatistical introduction. New York: Oxford University Press, 1990.

13. Peter JBl, Richard AD. Introduction to time series and forecasting, second edition). New York: Springer, 2002.

14. Brockwell PJ, Davis RA. Time series: Theory and methods, second edition. New York: Springer, 1991.

15. Bracewell R. The autocorrelation function. The Fourier transform and its applications. New York: McGraw-Hill, 1965:40-5.

Tables and Figures

Tables

Table 1: Selected countries for the study

Continent	Countries Selected	Development Status
Asia	Japan, Singapore	developed (countries)
	China, India, Thailand	developing (countries)
Europe	UK, Italy, Sweden	developed
	Russia, Poland	developing
Africa	Algeria, Sudan, Zambia,Morocco, Libya	developing
America	US	developed
	Brazil, Mexico	developing
Oceania	Australia	developed

Table 2: Regression coefficients of IMR v.s. time from OLS (upper) & GLS (lower) models, United States (1950∼2010)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	917.8335	29.8878	30.71	0.0000
imr$Year	-0.4556	0.0151	-30.19	0.0000

	ρ	Intercept	Slope
	0.9926	864.2053	-0.4271
s.e.	0.0071	59.1964	0.0299

Table 3: Comparison of the slopes (from GLS) for different countries

Comparison	*Diff. in slopes*	*Std of diff.*	$z-value$	$p-value$	*lower*	*upper*
US-UK	0.0113	0.042	0.266	0.790	-0.072	0.095
Australia-Sweden	-0.036	0.031	-1.173	0.241	-0.096	0.024
Japan-Singapore	0.0822	0.164	0.502	0.616	-0.239	0.403
China-India	0.6751	0.163	4.145	**<0.00001**	0.356	0.994
Sudan-Zambia	-0.1078	0.184	-0.586	0.558	-0.468	0.252
US-China	1.2523	0.135	9.266	**<0.00001**	0.987	1.517

Table 4: Regression coefficients from two model-building methods

country	*development_status*	*ols_slope*	*s.e.*	*gls_slope*	*s.e.*	ρ
Japan	developed	-0.7292	0.05377	-0.8995	0.1145	0.9126
Singapore	developed	-0.8231	0.04683	-0.9817	0.1172	0.8773
Thailand	developing	-2.06	0.06825	-2.066	0.1504	0.9035
India	developing	-2.275	0.04358	-2.3545	0.0957	0.8918
China	developing	-1.454	0.05659	-1.6794	0.1318	0.8563
Algeria	developing	-2.635	0.06008	-2.5626	0.1308	0.9185
Sudan	developing	-1.15	0.05225	-1.3219	0.1156	0.9068
Zambia	developing	-0.8197	0.05924	-1.2141	0.1429	0.9383
Libya	developing	-3.046	0.1895	-3.7296	0.4409	0.8859
Morocco	developing	-2.267	0.02545	-2.1171	0.0602	0.9256
UK	developed	-0.4382	0.0122	-0.4384	0.0302	0.8991
Sweden	developed	-0.3094	0.009713	-0.3149	0.0214	0.9253
Italy	developed	-0.9982	0.04317	-1.0559	0.093	0.9125
Poland	developing	-1.008	0.05758	-1.263	0.15	0.8592
Russia	developing	-0.5489	0.01509	-0.5686	0.0289	0.9181
US	developed	-0.4556	0.01509	-0.4271	0.0299	0.9367
Brazil	developing	-2.427	0.04272	-2.4057	0.093	0.9132
Mexico	developing	-1.741	0.02986	-1.6635	0.0587	0.9235
Australia	developed	-0.3784	0.008724	-0.3509	0.022	0.9310

Table 5: Statistical test of the equivalence of the slope estimates from OLS and GLS methods

Country	*Diff.*	*Std of Diff.*	$z-value$	$p-value$	*lower*	*upper*
Japan	0.1703	0.126	1.346	0.177	-0.078	0.418
Singapore	0.1586	0.126	1.257	0.208	-0.089	0.406
Thailand	0.006	0.165	0.036	0.971	-0.318	0.330
India	0.0795	0.105	0.756	0.450	-0.127	0.286
China	0.2254	0.143	1.571	0.116	-0.056	0.507
Algeria	-0.0724	0.144	-0.503	0.615	-0.355	0.210
Sudan	0.1719	0.127	1.355	0.175	-0.077	0.421
Zambia	0.3944	0.155	2.550	**0.011**	0.091	0.698
Libya	0.6836	0.480	1.424	0.154	-0.257	1.624
Morocco	-0.1499	0.065	-2.294	**0.022**	-0.278	-0.022
UK	0.0002	0.033	0.006	0.995	-0.064	0.064
Sweden	0.0055	0.024	0.234	0.815	-0.041	0.052
Italy	0.0577	0.103	0.563	0.573	-0.143	0.259
Poland	0.255	0.161	1.587	0.113	-0.060	0.570
Russia	0.0197	0.033	0.604	0.546	-0.044	0.084
US	-0.0285	0.033	-0.851	0.395	-0.094	0.037
Brazil	-0.0213	0.102	-0.208	0.835	-0.222	0.179
Mexico	-0.0775	0.066	-1.177	0.239	-0.207	0.052
Australia	-0.0275	0.024	-1.162	0.245	-0.074	0.019

Figures

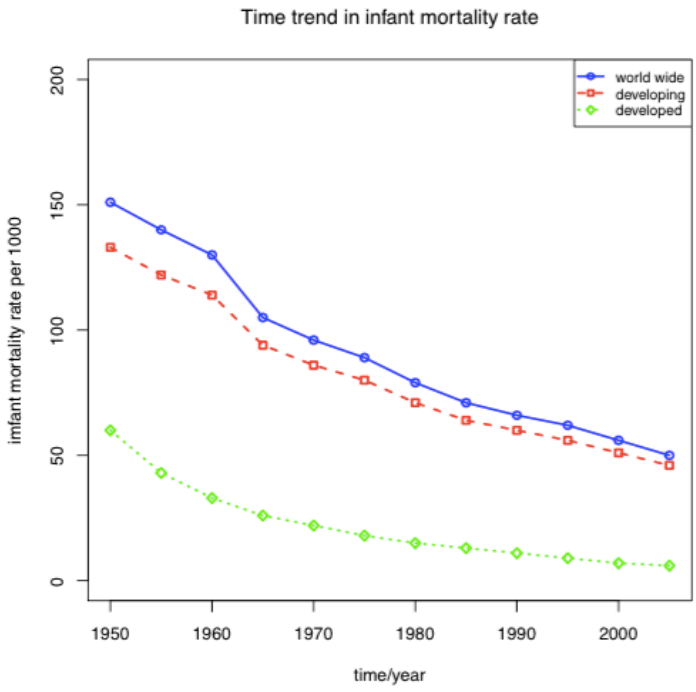

Figure 1: Time trends of IMR for different development status

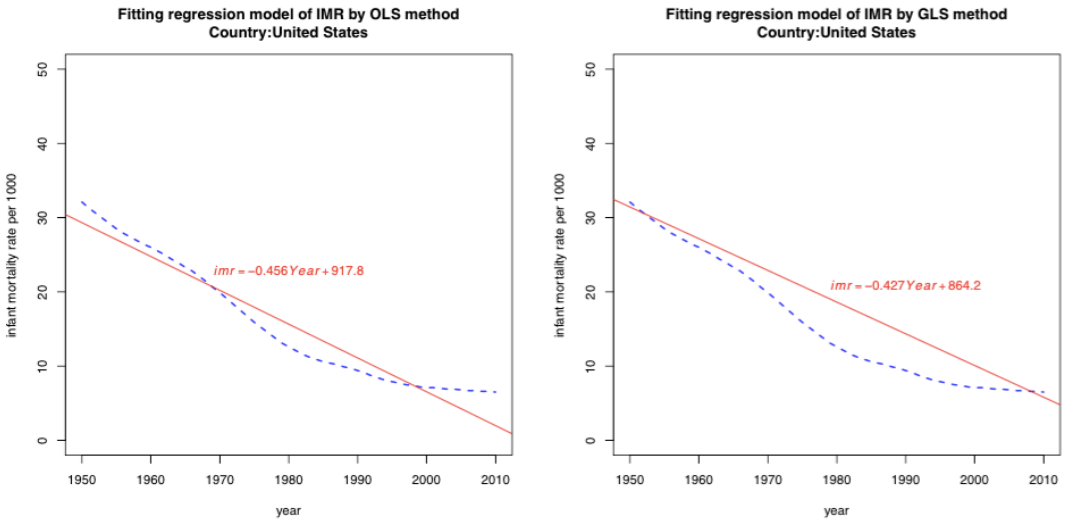

Figure 2: Regression line of IMR v.s. time from OLS (left) & GLS (right) models, United States (1950–2010)

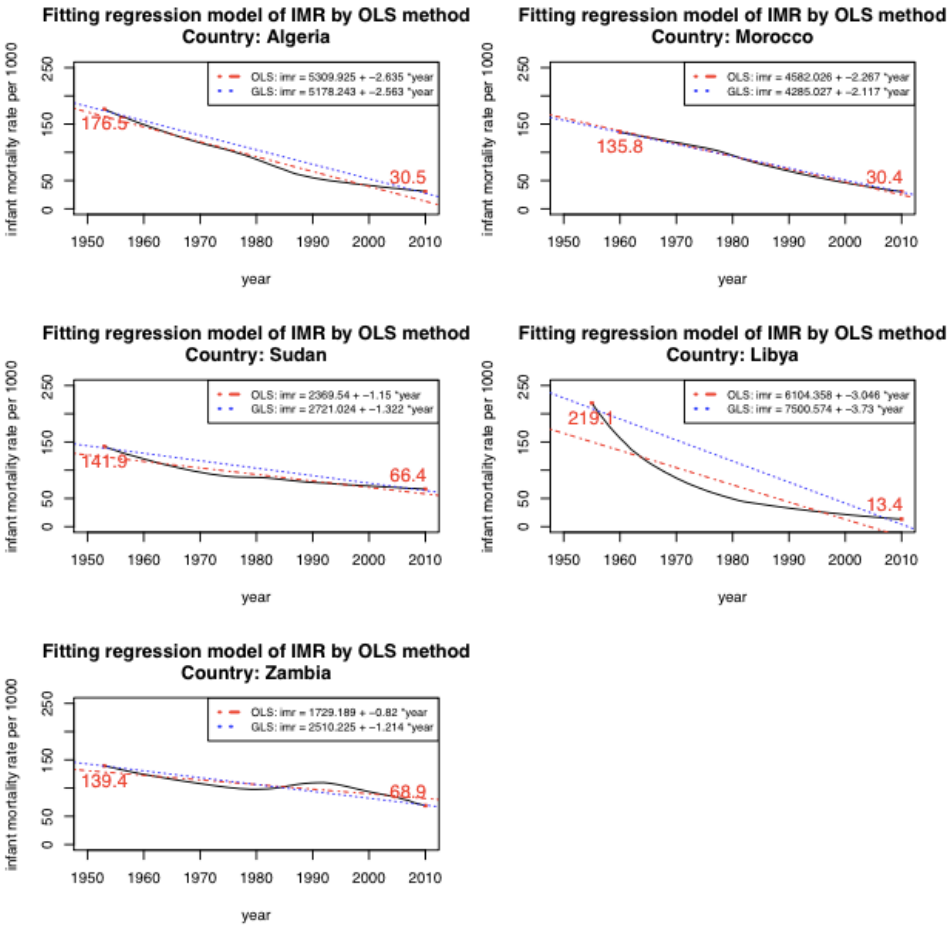

Figure 2: IMR time trends for African countries: Algeria, Sudan, Zambia, Morocco, and Libya

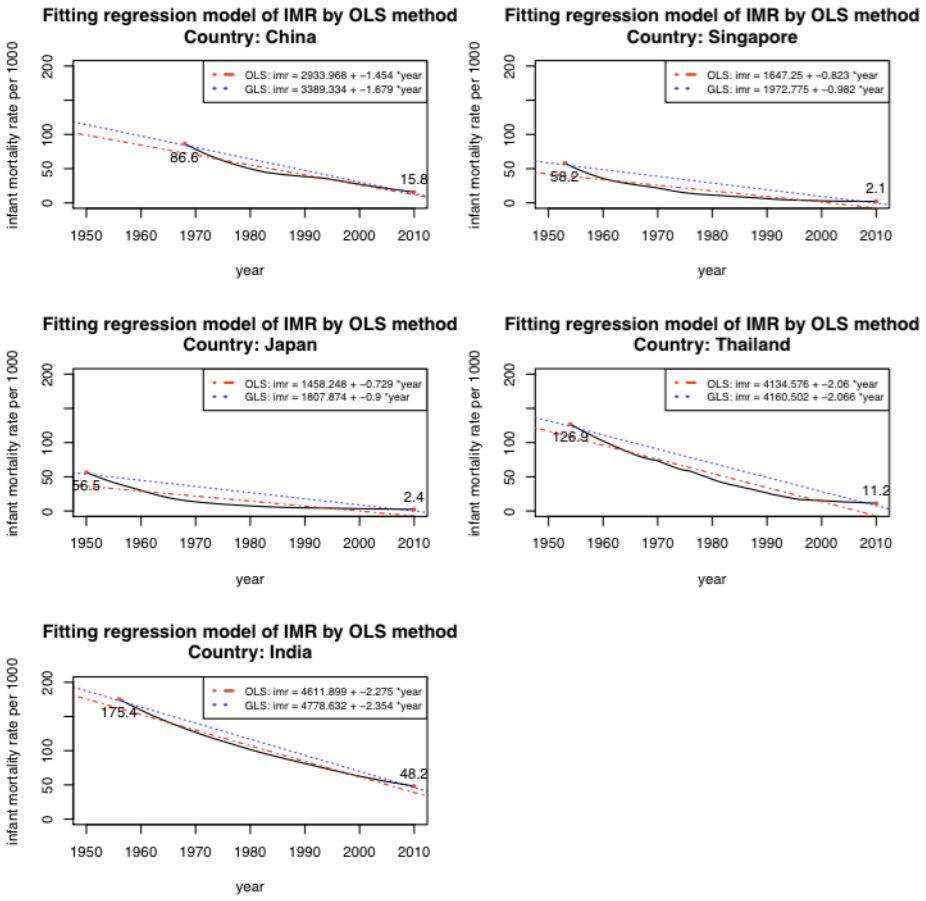

Figure 4: IMR time trends for Asian countries: China, Japan, India, Singapore, and Thailand

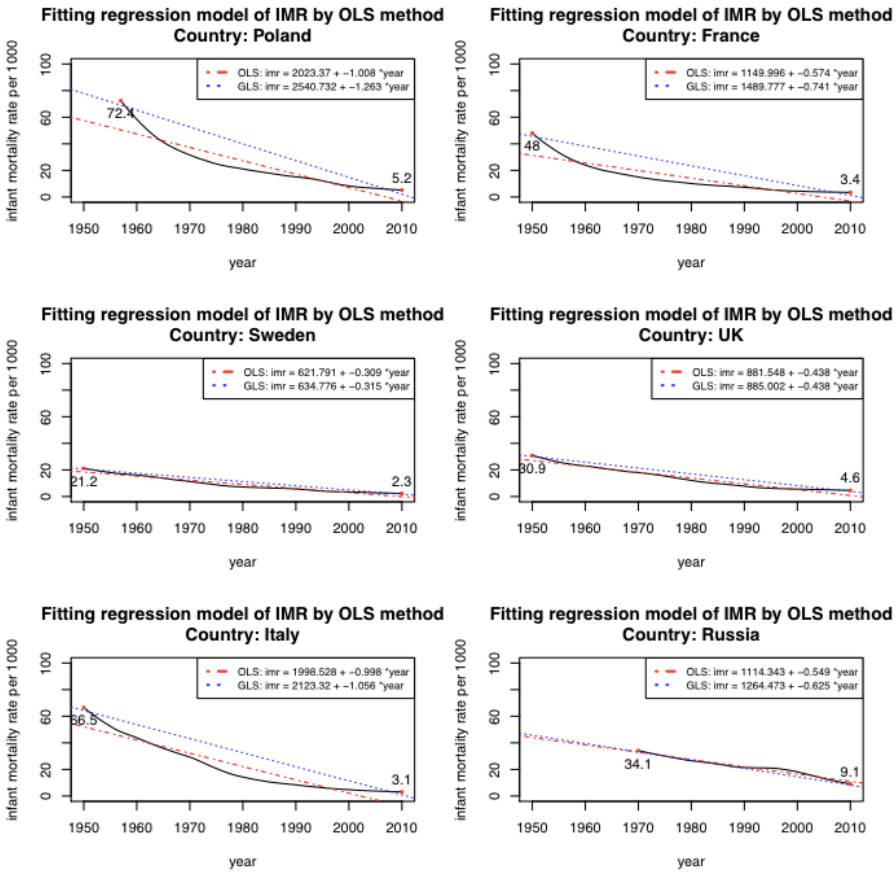

Figure 5: IMR time trends for European countries: Poland, Sweden, Italy, France, UK, and Russia

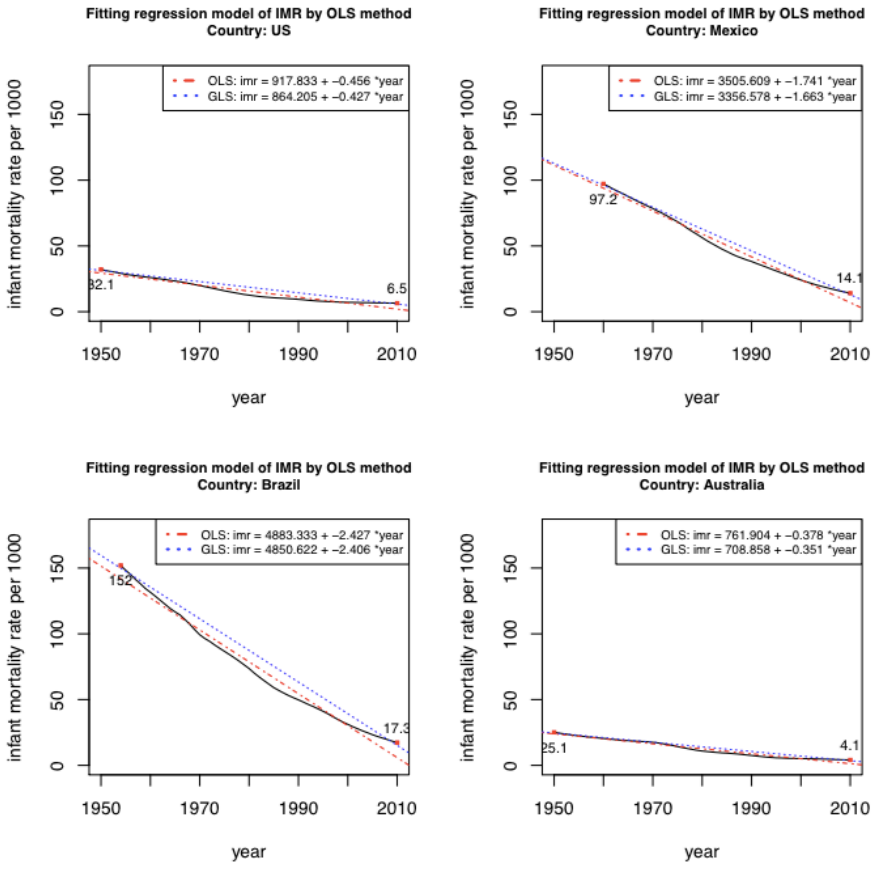

Figure 6: IMR time trends for American and Oceanian countries: US, Brazil, Mexico, and Australia