Identification of factors shaping life expectancy worldwide

1. Presentation of the issue

Life expectancy at birth (also referred to at work as life expectancy) in a given year represents the number of years that a newborn child in a given year expects to live if mortality rates at birth remain the same throughout the child's life. It is an important indicator for assessing the economic and social development of a country or region.

1.1. Description of the phenomenon and previous studies on the development of life expectancy

One of the main goals of any government is to increase the life expectancy of the country's population by reducing the mortality rate to the minimum possible. Over the past 170 years, life expectancy has been steadily increasing1. The lengthening of life expectancy at birth is, on the one hand, a great success for society, but at the same time it is a challenge for us and requires people to adapt to the resulting demographic consequences, so it is an important socio-economic problem. This is because increasing life expectancy, along with declining fertility rates, is contributing to the growth of the elderly population. By 2050, the group of people aged 60 and over is expected to grow to 2 billion worldwide[2]. Population aging poses a serious problem for the economy, especially for the labor market and the pension system, which in most countries is based on a pay-as-you-go system - unsuited to demographic change.

There are still large disparities between developed and developing countries. This disparity in life expectancy is believed to be rooted in socioeconomic differences. The basic rationale is that socioeconomic and environmental factors have an independent as well as interactive effect on the level of life expectancy[3]. Residents of highly developed countries, therefore, on average live longer and have lower mortality rates[4]. Indeed, life expectancy is closely related to a country's level of health, and good population health requires the fulfillment of several socioeconomic conditions such as reducing low levels of education among the population, reducing inequality and improving living conditions[5-6].

Researchers reveal the great importance of a population's level of education on life expectancy at birth - people with higher education lead healthier lives, which in

[1] Münz R, Demographic trend in the world (Innovative Health Initiative, Rovinj, 2012).
[2] United Nations (2009) World Population Ageing
[3] Mohammed Sufian AJ, International Journal of Humanities and Social Science, 3 (2013) 303.
[4] David Cutler, The Determinants of Mortality, https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.20.3.97 (accessed January 9, 2021)
[5] Wilkinson R, Marmot M(2003) Social determinants of health. The solid facts. 2nd edition (WHO Regional Office for Europe, Copenhagen, 2003).
[6] WHO (2009)Milestones in Health Promotion. Statements from Global Conferences.

Consequently, it has a positive effect on their life expectancy at birth[7]. This is because, typically, people with higher education are more aware of healthy lifestyles, work less often physically and often earn more. On the other hand, long education and career-oriented personal values delay marriage and reduce aggregate fertility levels. As mentioned earlier, this can further lead to accelerated aging and a negative impact on the economy[8].

Research indicates that gender inequality also significantly affects life expectancy at birth - women's education, economic status and reproductive autonomy have a significant positive impact on life expectancy[9]. In addition, there is a negative relationship between the number of births per 100 women and life expectancy at birth - teenage motherhood at the age of 15-19 is associated with lower education of young women, lower income and standard of living, and consequently lower life expectancy[10].

Other researchers point to a negative correlation between rural rates and life expectancy - in the U.S. between 2005 and 2009, life expectancy at birth in large metropolitan areas was 79.1 years, compared to 76.9 years in smaller but still urban cities and 76.7 years in rural areas[11]. Reasons for the lower life expectancy at birth in rural areas - identified by the researchers - include higher rates of heart disease, unintentional injuries, chronic obstructive pulmonary disease, lung cancer, strokes, suicide and diabetes.

1.2. Data

For research purposes, data was collected from 156 countries from around the world as of 2018. The data used comes from the World Bank datasets and the Human Development Report. The data refers to 6 demographic and socioeconomic variables. We assume that the indicators measured and calculated to approximate the values of the variables in question are methodologically adequate and correct. In addition, we also assume that both observable variables, as well as variables in the form of expected mean, are characterized only by some random error, i.e. sampling error. The units of statistical observation are individual countries. The following table (Table 1) dwells on the definitions of the variables used and the data sources of each variable:

[7] Blackburn, K. and Cipriani, G.P. (2002) 'A model of longevity, fertility and growth', Journal of Economic Dynamics and Control, Vol. 26, No. 2, pp.187-204.
[8] Čepar, Ž. and Bojnec, Š. (2014) 'Does higher relative participation in higher education also mean a higher absolute number of students?: The case of Slovenia', Eastern European Economics, Vol. 52, No. 2, pp.85-100.
[9] Williamson, J.B. and Boehmer, U. (1997) 'Female life expectancy, gender stratification, health status, and level of economic development: a cross-national study of less developed countries', Social Science & Medicine, Vol. 45, No. 2, pp.305-317.
[10] Singh, S. and Darroch, J.E. (1999) 'Adolescent pregnancy and childbearing: levels and trends in developed countries', Family Planning Perspectives, Vol. 32, No. 1, pp.14-23.
[11] Singh, G.K. and Siahpush, M. (2014) 'Widening rural-urban disparities in life expectancy, US, 1969-2009', American Journal of Preventive Medicine, Vol. 46, No. 2, pp.19-29.

Table 1: Description of variables and data sources

Variable	Definition of	Source
Birthsi	Index fertility rate among teenagers is the number of births per 1,000 women aged 15-19.	The World Bank (World Bank): World Development Indicators database
Educationi	The number of years of schooling a child of school starting age can expect to receive if the dominant age patterns.	Human Development Report 2019
Life_expectancyi	The number of years a newborn child expects to live if mortality rates at birth remain the same level throughout the child's life.	Human Development Report 2019
Ratioi	Ratio of women's labor force participation rate to men's labor force participation rate (in percentage %). The ratio of the female labor force participation rate to the male labor force participation rate is calculated by dividing the female labor force participation rate by the male labor force participation rate and Multiplying it by 100.	World Bank (World Bank): World Development Indicators database
Urbani	Urban population (in percentage %). The number of people living in the area defined as "urban" per 100 total residents. The variable refers to people living in urban areas as defined by the national statistical offices.	World Bank (World Bank): World Development Indicators database
Glli	An indicator that reflects the potential losses from female-male gender inequality in three dimensions: reproductive health, empowerment and the labor market.	Human Development Report 2019

The transformation of the original data consisted of removing observations (countries) with missing values in any of the variables. This is because the missing data can be considered as occurring randomly due to the large disparity in economic levels in the removed countries, so removing observations with missing values should not disturb further results of the analysis. Moreover, in the remaining 156 observations, there is no situation where a particular group of countries is unrepresentable in this sample. In addition, the sample can be considered large, so the reduction in test power (statistical power), that is, the increase in the risk of type II error, resulting from the removal of observations from the

of the original data is not problematic. Given this - removing entire observations with missing data in any of the variables from the original datasets is the optimal technique for working with missing data and produces unencumbered [results12]. The removed observations are 70 countries and represent about 31% of the original data.

1.3. Methodology

In further econometric analysis, life expectancy at birth was used as an explanatory variable, and 5 demographic and socioeconomic variables were used as explanatory variables. The general notation of the econometric model is of the form:

$$Life\_expectancy_i = \alpha_1 + \alpha_2 Education_i + \alpha_3 GIIi + \alpha_4 Ratio_i + \alpha_5 Urbani + \alpha_6 Births_i + \epsilon_i$$

2. Elicitation of a priori parameters

The a priori parameters were determined based on the work of researchers Žig Čepar and Aleš Trunk, "The role of expected years of schooling among life expectancy determinants," from January 2016. In the study, a linear regression model was set up using data from 187 countries in 2010, where life expectancy is the explanatory variable. By introducing such a priori information on parameter estimates of variables based on 2010 data from another study, we are able to derive estimates that include information older than just that of 2018. Such a priori information is useful in terms of parameter estimates that also use past data from other studies.

The a priori beta parameters and their variances were set in the study at the same level as in the study by Žig Čepar and Aleš Trunk. Since the study does not have exact information about the correlation between the parameters, we set zero correlations between the a priori parameters for simplicity. The authors point out that variables that are not highly correlated with each other were used as explanatory variables, so this simplification should not cause a large burden on the results. Another simplification due to the lack of precise information is that we set the inverse of the variance of the random component a priori at the same level as this statistic for our estimated linear regression model using 2018 data. Both models (using 2010 and 2018 data) use the same explanatory variables, so the difference in the inverse of the variance of the random component should be small, and simplification should therefore not lead to a bias in the results. We set the value of 187 as the v a priori parameter, since Žig Čepar and Aleš Trunk's study was conducted using data from 187 countries.

[12] Acock, A. C (2005) Working With Missing Values, JOURNAL OF MARRIAGE AND FAMILY, VOL 67; NUMBER 4, pages 1012-1028

3. Expected value of a posteriori parameters and their a posteriori marginal distributions The expected values of a posteriori parameters are as follows:

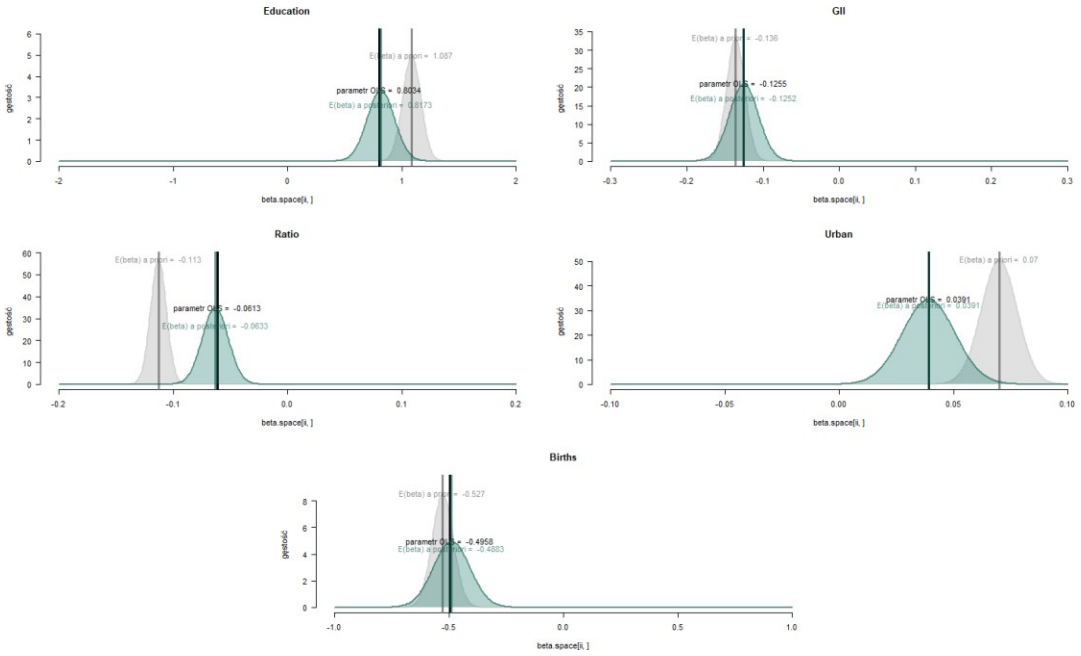
Table 2: Expected value of a posteriori parameters

Variable	*Expected value a posteriori parameters*
Intercept	70.4187
Education	0.8173
GII	-0.1252
Ratio	-0.0633
Urban	0.0391
Births	-0.4883

For the Education and Urban variables, the expected value of the a posteriori parameters is positive. An increase in the expected years of education in a country will therefore cause an increase in life expectancy in that country ceteris paribus. Similarly, an increase in the number of people living in an area defined as "urban" will cause an increase in life expectancy in that country ceteris paribus. The variables GII, Ratio and Births, on the other hand, show a negative relationship with the explained variable. Intuitively, one would expect that an increase in the ratio of female to male labor force participation ($Ratio_i$) would result in an increase in life expectancy ($Life\_expectancy_i$). However, it turns out that the opposite is true, and an increase in the ratio of female to male labor market participation is associated with a decrease in life expectancy[13]. This can be explained by the fact that a higher ratio may on the one hand be associated with higher female employment, but on the other hand it may also be due to lower male employment. A lower male employment rate, and thus a higher female-to-male labor force participation rate ($Ratio_i$), could mean a worsening of the labor market, and thus a lower overall economic situation and a lower standard of living in a country, as well as a lower life expectancy in that country. Moreover, higher female employment - and thus a higher $Ratio_i$ - is often associated with women's employment in industry, especially in developing countries with lower living standards. The estimation results therefore agree with previous studies on the determinants of life expectancy and can be considered reasonable.

[13] Novak, A., Čepar, Ž. and Trunk, A. (2016) 'The role of expected years of schooling among life expectancy determinants', Int. J. Innovation and Learning, Vol. 20, No. 1, pp.85-99.

Figure 1: Boundary distributions of a posteriori parameters



According to the graphs showing the a posteriori marginal distributions (Figure 1), it can be seen that for each of the variables, the parameter value based on the sample model, the a priori and a posteriori expected parameter values are similar. Based on this, it can be concluded that over time the relationships between the selected explanatory variables and the life expectancy variable have changed slightly - in fact, the estimates for the variables based on the 2018 sample model and the expected values of the parameters based on the 2010 model are similar.

In addition, it is worth noting that for each of the variables, the expected value of the a posteriori parameters are more similar to the sample parameter estimates than to the expected value of the a priori parameters. This is due to the fact that for each of the variables, the variance of the a priori parameter estimates is higher than the variance of the sample parameter estimates.

4. Importance of individual variables using HPDI and Bayes factors

The study yields the following 95% HPDI ranges and Bayes factors:

Table 3: 95% HPDI interval

Variable	Lower limit	Upper limit
Education	0.59649123	1.03759398
GII	-0.16165414	-0.09097744
Ratio	-0.08471178	-0.04260652
Urban	0.01679198	0.06090226
Births	-0.64411028	-0.33834586

Figure 2: Edge distributions of a posteriori parameters with designated 95% HPDI intervals.

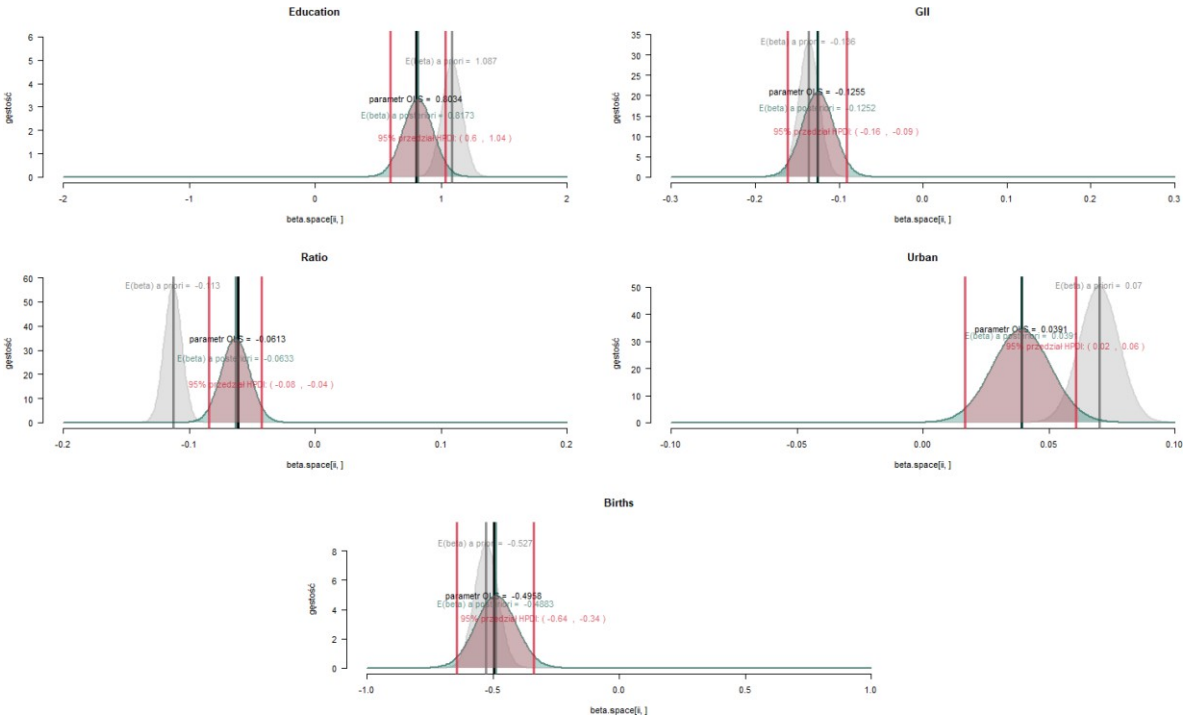


Table 4: Bayes Factors (BF).

Variable	Bayes factor (BF)
Education	44211883.71
GII	2103278.70
Ratio	1599.51
Urban	0.76
Births	563526.98

Based on the obtained HPDI intervals (Table 3 and Figure 2), it can be seen that for the Education and Urban variables, both the left and upper limits of the 95% HPDI interval are positive. This indicates a positive effect of these variables on life expectancy. In contrast, for the remaining variables - GII, Ratio and Births - the left and upper limits of the 95% HPDI interval are negative. This means that an increase in the female-male gender gap, an increase in the female-to-male labor force participation rate, and an increase in the number of births to women aged 15-19 in a country will respectively cause a decrease in life expectancy in that country. The estimation results agree with previous studies on the determinants of life expectancy and can be considered reasonable.

Bayes factors were determined "for individual variables" (Table 4) - that is, more precisely, for models with and without these variables. In the calculation of Bayes factors, as the first model $_{M1}$ we define a model with all explanatory variables, and as the second model $_{M2}$ we define several times a model without a given explanatory variable. Based on the obtained Bayes factors, it can be concluded that for each of the explanatory variables, except for the variable

Urban, the strength of evidence is classified as very strong according to the Kass and Rafter scale (because BF>150 values) and as conclusive according to the Jeffreys scale (because BF>100 values). The results therefore support the model with these variables far more than the model without them, which can be interpreted as a strong significant effect of each of the Education, GII, Ratio and Births variables on life expectancy. For the Urban variable, on the other hand, the strength of evidence is classified as negative (because the value of BF<1) - the results therefore support the model without the Urban variable more than the model containing this variable, which can be interpreted as an insignificant effect of the Urban variable on life expectancy.

1. Presentation of the issue

1.1. Description of the phenomenon

The issue addressed is the same topic that I addressed in homework 1 (Appendix 1), that is, the identification of factors shaping life expectancy in the world. The following paper uses the same data as the first homework, and the work of researchers Žig Čepar and Aleš Trunk "The role of expected years of schooling among life expectancy determinants" (2016) was similarly used as the available a priori knowledge prior to obtaining the sample.

1.2. Methodology

In further analysis, life expectancy at birth was used as the explanatory variable, and 5 demographic and socioeconomic variables were used as explanatory variables (see Appendix 1 for variable descriptions). The general notation of the econometric model is of the form:

$$Life\_expectancy_i = \beta_1 + \beta_2 Education_i + \beta_3 GII_i + \beta_4 Ratio_i + \beta_5 Urban_i + \beta_6 Births_i + \epsilon_i$$

We will use the Hamiltonian Monte Carlo (HMC) method with the No-U-Turn Sampler (NUTS) extension to estimate a posteriori parameter distributions. HMC is an MCMC-type algorithm that improves on the Metropolis-Hastings algorithm in terms of efficiency, avoiding random behavior by performing subsequent algorithm steps based on first-order gradient information. These features allow the results to converge faster than the random walk of the Metropolis algorithm. However, the performance of HMC is very sensitive to two parameters: the step size and the desired number of steps. The HMC No-U- Turn Sampler (NUTS) extension eliminates the need to determine the number of steps. NUTS uses a recursive algorithm to build a set of likely candidate points, continuing until the trajectory turns back on itself - then it stops (hence the name "No U-Turn"). Details can be found in Hoffman & Gelman (2011).

1.3. A priori information and assumptions made

The aforementioned work by researchers Žiga Čepar and Aleš Trunk (2016) was used as a priori information. In the study, a linear regression model was estab- lished using data from 187 countries from 2010, where the explanatory variable is life expectancy. The researchers point out that "for the model, the distributions of all variables included in the analysis are sufficiently close to a normal distribution," so the paper assumes that the a priori beta distribution is normal for each beta. The researchers further write in the paper that "regression error distributions, with a mean close to zero, also confirm a normal distribution," so in the paper we use the information that for each observation the random component has a normal distribution. For the precision of the random component, we assume that it has a gamma distribution with hyperparameters 1, 1 - this represents a relatively uninformative variant.

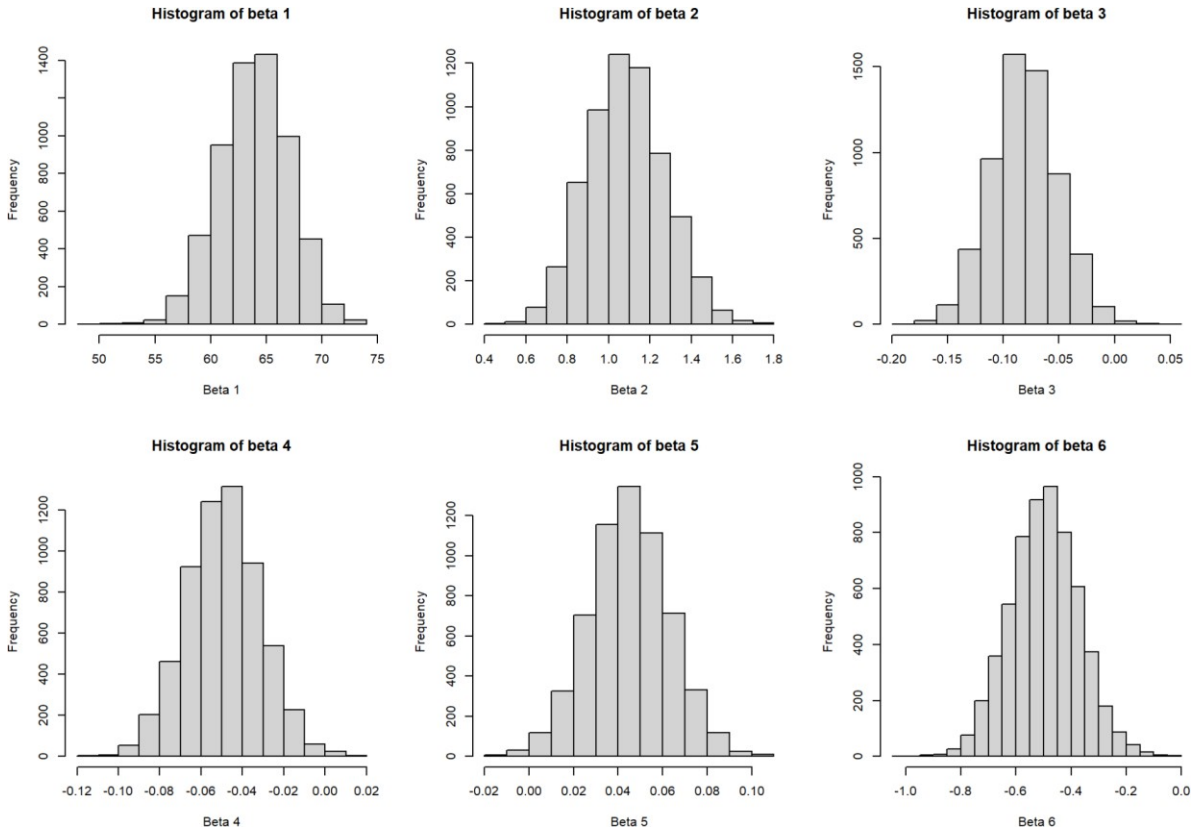
Using the Hamiltonian Monte Carlo method with the NUTS variant, we assume the use of 4 Markov chains, 2000 iterations for each chain (including warmup) and 500 warmup iterations per chain.

2. Analysis results

The results for a posteriori parameter distributions based on 4 chains of 1,500 post-warmup observations, or 6,000 post-warmup observations in total, are as follows:

Parameter	Average	Deviation std.	2.5%	25%	50%	75%	97.5%
beta 1	63.9373	3.1374	57.7070	61.8164	64.0141	66.1263	69.7883
beta 2	1.0847	0.1887	0.7373	0.9532	1.0816	1.2102	1.4576
beta 3	-0.0809	0.0299	-0.1386	-0.1006	-0.0812	-0.0612	-0.0221
beta 4	-0.0492	0.0180	-0.0845	-0.0613	-0.0493	-0.0373	-0.0136
beta 5	0.0450	0.0179	0.0095	0.0332	0.0451	0.0572	0.0800
beta 6	-0.4970	0.1253	-0.7390	-0.5823	-0.4959	-0.4135	-0.2505

In the above table and subsequent analysis, the beta designation is consistent with the description of the equation in section 1.2 Methodology. The mean value of the a posteriori parameter for free expression is therefore approximately 63.94, for education 1.08, for the GII index -0.08, for the female-to-male ratio in the labor market -0.05, for the urbanization index 0.05, and for the fertility rate per 1,000 women aged 15-19 -0.5. The standard deviations indicate not much deviation from the mean, as also confirmed by the histograms of the a posteriori distributions for each variable:



The study also determined 95% HPDI confidence intervals:

Parameter	Lower limit	Upper limit
beta 1	57,8520	69,8782
beta 2	0,7350	1,4533
beta 3	-0,1368	-0,0212
beta 4	-0,0858	-0,0156
beta 5	0,0083	0,0782
beta 6	-0,7427	-0,2570

Based on the obtained HPDI intervals, it can be seen that for the Education and Urban variables, both the left and upper limits of the 95% HPDI interval are positive. This indicates a positive effect of these variables on life expectancy. In contrast, for the remaining variables - GII, Ratio and Births - the left and upper limits of the 95% HPDI interval are negative. This means that an increase in the female-male gender gap, an increase in the female-to-male labor force participation rate, and an increase in the number of births to women aged 15-19 in a country will respectively cause a decrease in life expectancy in that country. The estimation results agree with intuition and can be considered reasonable.

3. Evaluation of results

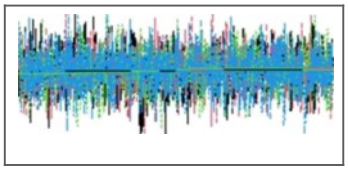
In order to evaluate the above results, the numerical standard error was calculated, that is, the error that will account for the phenomenon of autocorrelation. The following results were obtained:

Parameter	Average	Numerical error standard
beta 1	63.93732	0.0604146
beta 2	1.08473	0.0034195
beta 3	-0.08090	0.0005273
beta 4	-0.04918	0.0002765
beta 5	0.04501	0.0002645
beta 6	-0.49701	0.0019929

Based on the numerical standard error, we can say that the obtained results were obtained with high accuracy, since the numerical standard error is low.

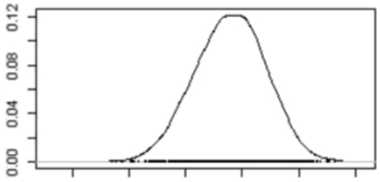
The following analysis shows the waveforms of each of the 4 chains and
The densities of the distributions for the a posteriori parameters of each variable:

Trace of beta[1]



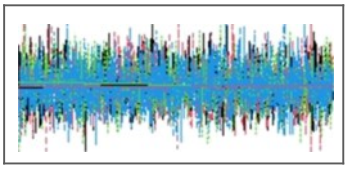
Iterations

Density of beta[1]

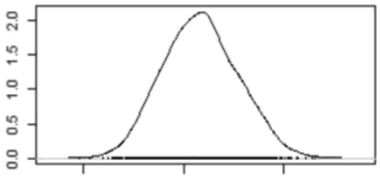


Bandwidth = 0.5838

I race of beta[2].

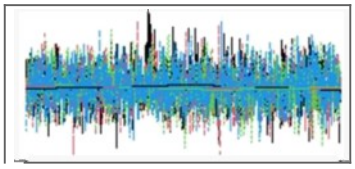


Density of beta[2]



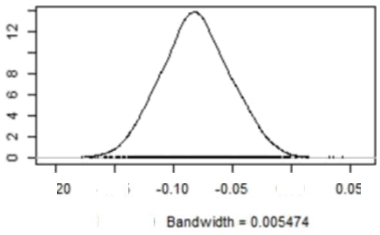
Bandwidth = 0.03511

I race of beta[3].



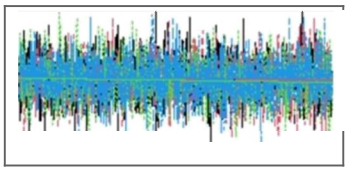
fie rat ie r s

Density of beta[3]

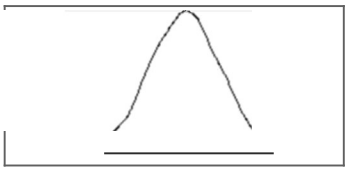


Bandwidth = 0.005474

Trace of beta[4].



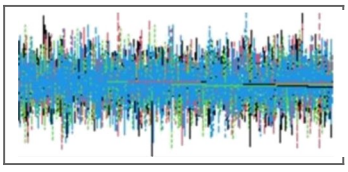
Density of beta[4]



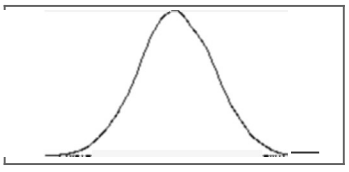
-C6E c r L

B a- .'. -c'n = 0 CC ? 2 * "

Trace of beta[5].

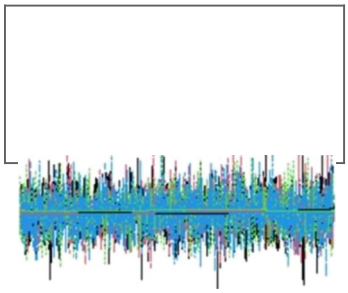


Density of beta[5]

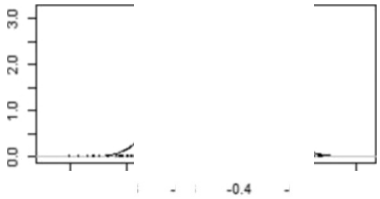


Bandwidth = 0.003327

Trace of beta[6].



Density of beta[6]



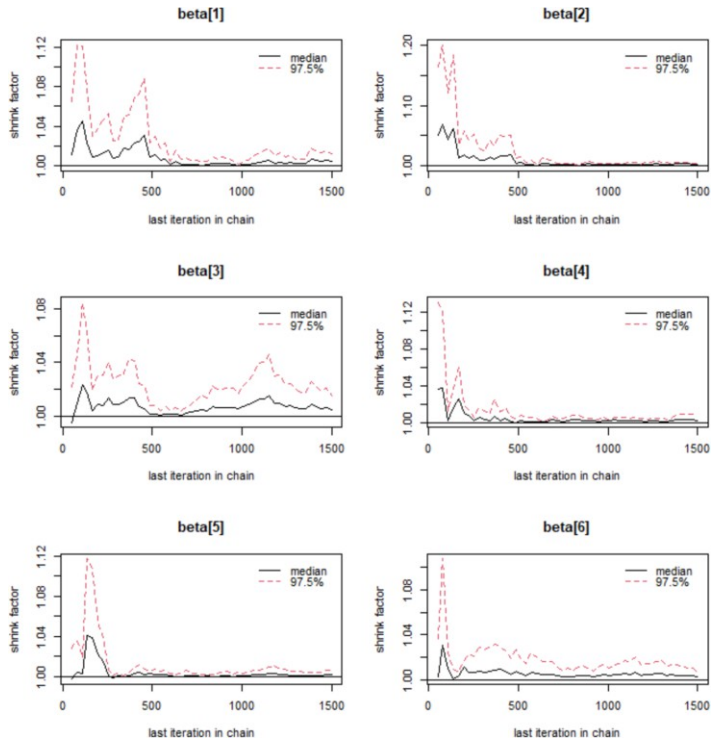
Bandwidth = 0.02331

According to the graph, the course of all chains for each parameter looks stationary - it oscillates with equal strength around the same level and shows no trends. These are satisfactory results for us, showing that the number of warmup iterations was sufficiently chosen and that the total number of iterations was satisfactory. Looking at the densities for each parameter a posteriori, the distributions are unimodal, indicating the true form of the distribution a posterori. This is because a multimodal distribution would indicate that the densities gather by chains in different ways. In contrast, in the present study, the distributions for each a posteriori parameter are qualitatively close to a normal distribution. Considering the graphical analysis, one can speak of a convergence of chains.

In addition to visual interpretation, tests for convergence and stationarity of the results were also performed. For the Gelman-Rubin criterion, both for individual variables for the upper confidence interval and synthetically for the entire vector of parameters, the values of the statistic are below 1.2 - this indicates the presence of convergence:

Parameter	Point estimate	Upper compartment
beta 1	1	1.01
beta 2	1	1.00
beta 3	1	1.02
beta 4	1	1.01
beta 5	1	1.01
beta 6	1	1.01
total	1.01	

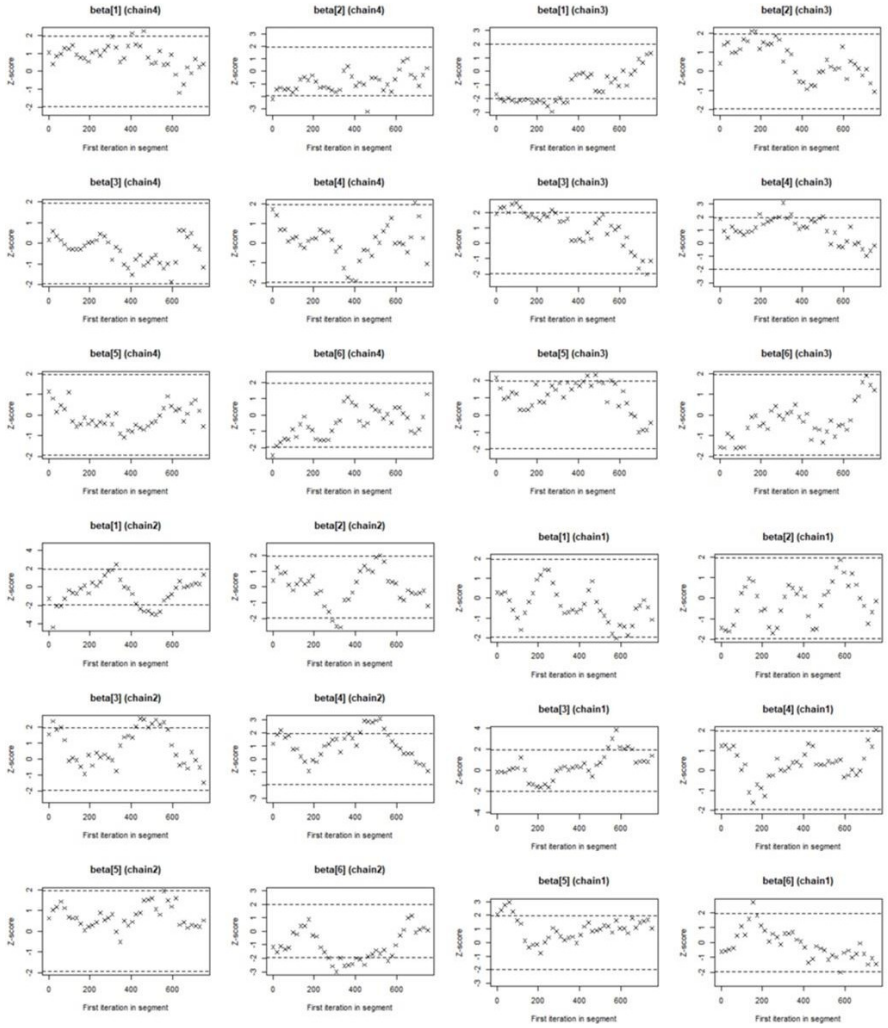
The Gelman-Rubin criterion decreases with the number of iterations, and for the number of iterations selected in the paper
has dropped to a level very close to 1:



Another statistic used is the Geweke statistic, in which it was assumed that the chain was divided into 3 fragments given by fractions of its length: 10%, 50%, 40%. According to the results, it is possible to speak of convergence within each of the chains, since for each parameter in the chain the statistic does not exceed the value of 1.96 or is minimally above this value. Therefore, it cannot be said that the target escapes us inside any of the chains:

Chain	Beta 1	Beta 2	Beta 3	Beta 4	Beta 5	Beta 6
1	0.2814	-1.4379	-0.1580	1.2163	2.0562	-0.6153
2	-1.2581	0.4039	1.5588	1.1818	0.6300	-1.1634
3	-1.6664	0.4194	1.9235	1.8273	2.1782	-1.5603
4	1.0348	-2.1851	0.1529	1.7116	1.1403	-2.4864

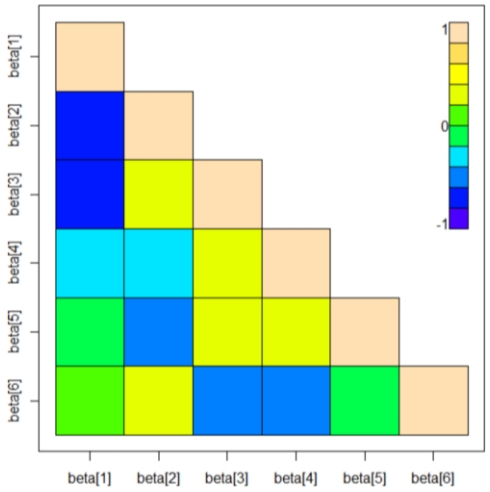
Since we have obtained single cases for which the value is above 1.96, we can also iteratively search for the moment from which the chain is considered to converge. To do this, we discard 50% of the initial observations and determine the value of the Geweke statistic for this case. It can be seen that the values of the Geweke statistic fluctuate around the value of the
-1.96 a 1.96 for each case and there is no clear trend of the value approaching 0, so it can be concluded that lengthening the chains by increasing the number of iterations would yield similar results:



To verify stationarity, the Heidelberg-Welch criterion was used. The results show that in each chain for a given parameter the p-value is above 0.05, so it can be concluded that each chain is stationary:

Parameter	p-value for chain 1	p-value for chain 2	p-value for chain 3	p-value for chain 4
beta 1	0.507	0.182	0.151	0.122
beta 2	0.589	0.908	0.318	0.103
beta 3	0.151	0.086	0.079	0.468
beta 4	0.973	0.208	0.434	0.334
beta 5	0.197	0.480	0.091	0.472
beta 6	0.470	0.142	0.567	0.259

The autocorrelation of the parameters was also examined during the evaluation. The results show high autocorrelation of parameter beta 1 with parameters beta 2 and beta 3. The presence of autocorrelation between parameters slows down the occurrence of convergence and requires more iterations to show a posteriori distribution. Increasing the number of iterations, however, does not get rid of the occurrence of autocorrelation, and since we have no doubt that convergence was achieved, as shown by the Geweke criterion and the Gelman-Rubin criterion, we conclude that a sufficiently long chain was used. In further studies, chain thinning could be used as a method to eliminate the occurrence of autocorrelation.



In summary, we are satisfied with the test results, as the chains are stationary and have reached convergence. This shows that the number of warmup iterations was sufficiently chosen and the number of total iterations was satisfactory, and that the assumptions were well made. Thus, the test results prove the true form of a posterior distribution, so the results of the analyses lead to the right conclusions.