American Journal of Theoretical and Applied Statistics
2023; 12(6): 174-179
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20231206.13
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Modelling Stroke Risk Factors Using Classical and Bayesian Quantile Regression Models

Kirui Dennis*, Charity Wamwea, Bonface Malenje, Levi Bor

Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email address:

dkirui54@gmail.com (Kirui Dennis), cwamwea@jkuat.ac.ke (Charity Wamwea), bmalenje@jkuat.ac.ke (Bonface Malenje), leviybor40@gmail.com (Levi Bor)

*Corresponding author

To cite this article:

Kirui Dennis, Charity Wamwea, Bonface Malenje, Levi Bor. Modelling Stroke Risk Factors Using Classical and Bayesian Quantile Regression Models. *American Journal of Theoretical and Applied Statistics*. Vol. 12, No. 6, 2023, pp. 174-179
doi: 10.11648/j.ajtas.20231206.13

Received: October 3, 2023; **Accepted**: October 25, 2023; **Published**: November 11, 2023

Abstract: The assessment of stroke risk and mortality, the second leading global cause of death, is of paramount importance. Stroke prediction is a vital pursuit due to its multifactorial nature, involving variables like age, sex, gender, hypertension, BMI and heart disease, which introduce considerable complexity. These diverse factors often lead to substantial uncertainty in stroke prediction models. Our research delves into the evaluation of two distinct methodologies for quantifying this uncertainty: Bayesian and classical quantiles. Bayesian quantiles are calculated from the posterior distribution of a Bayesian logistic regression model, accounting for prior information and spatial correlations. In contrast, classical quantiles are based on the assumption that stroke probabilities conform to a normal distribution. The results reveal that, across all coefficients, the Bayesian model produces narrower intervals compared to the classical model, indicating higher accuracy and confidence. Hence, we conclude that Bayesian quantiles outperform classical quantiles in the context of stroke prediction in Kenya. We recommend their adoption in future research and applications, acknowledging their superior performance and reliability in enhancing stroke prediction models, ultimately contributing to improved public health outcomes. This research represents a significant step towards a better understanding and management of stroke risks and mortality on a global scale.

Keywords: Bayesian Quantile Regression, Classical Quantile Regression, Potential Scale Reduction Factor, Markov Chain Monte Carlo

1. Introduction

In sub-Saharan Africa, stroke is one of the main causes of mortality and disability, with Kenya being one of the worst-affected nations. Evidence on the epidemiology, treatment, and consequences of stroke in Kenya, as well as the risk factors connected to various quantiles of the stroke distribution, are, however, lacking. It is essential to create statistical models that can account for the complexity and variability of stroke data in order to give insights into preventative and intervention measures.

Utilizing classical and Bayesian binary quantile regression models, which can calculate the effects of variables on various quantiles of the binary answer variable, is one strategy that is plausible. Additionally, accounting for asymmetric and heavy-tailed error distributions, binary quantile regression models are suited for binary results, such as the occurrence or survival of strokes. While Bayesian binary quantile regression models are based on maximizing a likelihood function with an asymmetric Laplace distribution, classical binary quantile regression models are based on minimizing a check function. To cope with high dimensional and collinear variables, these strategies can further use regularization techniques, such as lasso or bridge penalties. The objective here is to model the risk factors of stroke disease in Kenya using Classical and Bayesian binary quantile regression.

BQR has been applied to stroke prediction in different contexts, such as South Africa, China, and Brazil. BQR can identify and quantify the effects of modifiable and non-modifiable risk factors for stroke, such as age, gender, hypertension, diabetes, smoking, obesity, and heart problems.

BQR can also estimate the direct and indirect costs of stroke at different quantiles, which can inform health policy and resource allocation. BQR can provide a richer and more flexible analysis of stroke prediction than CQR, and offer insights into the heterogeneity and variability of stroke outcomes.

A statistics approach dubbed classical quantile regression extends the idea of conditional mean regression to conditional quantile functions. In comparison to the standard least squares approach, it allows for the estimation of the effects of covariates on various percentiles of the response variable and offers a more thorough and reliable analysis. [11], who devised the linear quantile regression model and its asymptotic features, introduced classical quantile regression.

In the Bayesian quantile regression (BQR) method, the parameters of the quantile regression model are estimated using Bayesian inference. Prior knowledge may be included, complicated models and hierarchical structures can be handled, and the estimates' uncertainty can be quantified using BQR, [7].

Variational inference (VI) minimizes the Kullback-Leibler divergence and approximates the posterior distribution by a more straightforward distribution. Compared to Markov Chain Monte Carlo (MCMC), VI can offer quicker and more scalable inference, albeit at the cost of some accuracy and dependability. The VI for Bayesian Quantile Regression (BQR) was created by [14] using a combination of normal distributions and by [12] using a normal approximation.

Choosing an appropriate likelihood function for the quantile regression model is one of the primary issues of BQR. The most popular option is the asymmetric Laplace (AL) distribution, which has a density function corresponding to the frequentist QR's negative check function, [4]. The location parameter, which defines the conditional quantile, and the scale parameter, which regulates the errors' dispersion, are the two parameters that make up the AL distribution. The AL distribution may be implemented using Markov chain Monte Carlo (MCMC) techniques for BQR because it can be represented as a scale mixture of normal distributions with exponential mixing weights.

Distributed computing, divides the data into smaller subgroups and does computations in parallel or sequentially on each portion. Distributed computing may make use of the parallelism of current technology while lowering memory and transmission expenses. [13] suggested distributed computing for Bayesian Quantile Regression (BQR) utilizing consensus Monte Carlo algorithms and divide-and-conquer tactics, respectively.

The regularization or shrinking of the parameters in sparse modeling is done to accomplish variable selection and dimension reduction. By avoiding overfitting and multicollinearity issues, sparse modeling can improve the interpretability and predictive accuracy of Bayesian Quantile Regression (BQR). Studies on sparse modeling for BQR include [9], which utilized an adaptable Lasso prior, [3], which employed a horseshoe prior.

BQR has been used in several disciplines, including engineering, finance, ecology, and biostatistics. Several instances are:

In an economic growth study, the causes of economic growth at various income quantiles were examined using Bayesian Quantile Regression (BQR). According to BQR, Ordinary Least Squares (OLS) regression failed to account for the diverse impacts of explanatory factors on various income distribution segments. This application was presented by [1], 2012, who used cross- country data from 1960 to 2000.

Bayesian Quantile Regression (BQR) has been utilized in ecological niche modeling to simulate the link between species occurrence and environmental factors across various quantiles of occurrence probability. BQR allowed for the evaluation of the uncertainty and variability of species responses and offered a more flexible and comprehensive description of species-environment interactions than logistic regression. This use was demonstrated by [6], who employed data from 226 bird species in North America.

2. Methods

2.1. Data Source

The performance of BQR and Classical quantiles in determining stroke risk in Kenya is compared in the study. The research employed data from the 2015 Kenya Stepwise Survey for Non-Communicable Diseases Risk Factors, a nationally representative survey of Kenyan households. The study was then repeated using a random simulation of the data set to compare the outcomes and efficiency. The study was then repeated using a random simulation of the data set to compare the outcomes and efficiency. Age, sex, hypertension, diabetes, smoking, alcohol use, and physical activity are some examples of possible risk factors for stroke that may be identified using the information.

2.2. Classical Binary Quantiles

The link between a response variable and one or more explanatory factors is described using quantiles in the classical quantiles model [15]. Quantiles, like the median (the 50th percentile) or the quartiles (the 25th and 75th percentiles), are points that split a distribution into equal portions. The conditional quantile of the response variable given the explanatory factors is assumed by the classical quantiles model to be a linear function of the explanatory variables. For the τ-th quantile, the classical quantile regression model is:

$$Q_\tau(B|A) = \beta_0(\tau) + \beta_1(\tau)A_1 + \cdots + \beta_p(\tau)A_p$$

The classical binary quantile regression estimator is established by minimizing the sum of the check function over the data set.

$$\beta(\tau) = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau \left(b_i - x_i^T \beta\right)$$

Under some regularity limitations such as the rank and identification conditions, this estimator is consistent and asymptotically regular, [10]. A prior distribution is the

Laplace distribution, which has a density function given by;

$$f(\beta) = \frac{1}{2b}\exp\left(-\frac{|\beta|}{b}\right)$$

Given b > 0 is a scale parameter. This prior distribution induces sparsity in the estimation as it penalizes large values. The posterior distribution (y, x) of the given data is proportional to the product of the likelihood function and the prior distribution

$$f(\beta|y,x) \propto L(y|x,\beta) * f(\beta)$$

L(y|x, β) is the likelihood function of y given x. For the binary data, the likelihood function can be written as;

$$L(y|x,\beta) = \prod_{i=1}^{n} F\,(x_i^T\beta)^{y_i} * \left(1 - F(x_i^T\beta)\right)^{1-y_i}$$

The explanatory variables are βX = (X1, X2….. Xp), the response variable is Y, and the QR coefficients are β0(τ),..., βp(τ). According to the model, many sets of coefficients reflect how the explanatory variables influence the τ-th quantile of the response variable for each value of τ.

By reducing the total weighted absolute errors between the response variable's observed and projected values, the classical quantile regression model may be computed. The weights are determined by the value of tau and by whether the mistake is positive or negative. The desired outcome to be minimized is

$$\sum_{i=1}^{n} \rho_\tau\left(y_i - x_i^T\beta(\tau)\right)$$

Where yi is the observed value of Y for observation i, xi is the vector of explanatory variables for observation i, β(τ) is the vector of quantile regression coefficients for quantile τ, and ρτ (u) is the check function.

2.3. Bayesian Binary Quantiles

According to [2], when estimating the quantile regression coefficients using Bayesian inference, Bayesian quantile regression is a methodology that takes into account previous knowledge and uncertainty. Assume that Y is the response variable, and that X is a matrix of predictor variables, that β is a vector of regression coefficients, and that ϵ is the error term. The following is the representation of the conditional quantile regression function at the τ-th quantile of Z:

$$Q(Z) = X\beta(\tau)$$

Where βτ is the vector of regression coefficients at the τ-th quantile. In Bayesian Quantile Regression, we place a prior distribution on βτ and ϵ and estimate the posterior distribution ofβτ and ϵ given the data. The general form of a Bayesian quantile regression model for the τ-th quantile is:

$$Q_\tau(Z) = \beta_0(\tau) + \beta_1(\tau)X_1 + \cdots + \beta_p(\tau)X_p$$

Where Y is the response variab*le, X = (X1,*...., Xp) are the explanatory variables, and β0(τ),..., βp(τ) are the quantile regression coefficients that depend on τ. The model implies that for each value of τ, there is a different set of coefficients that describe how the explanatory variables affect the τ -th quantile of the response variable**.** Given is the basic Bayesian binary response model:

$$Z^* = X\beta(\tau) + \epsilon$$

$$\epsilon \sim ALD(\mu,\sigma,\tau)$$

$$Y = \begin{cases} 1 & : Z^* > 0 \\ 0 & : Z^* \le 0 \end{cases}$$

Priors: $\beta(\tau) \sim N(\beta(\tau)_0, \Sigma_0)$

where Z is the binary response variable determined by the latent variable y∗, X is a n× p matrix of predictor variables, β(τ) is n× 1 vector of unknown regression parameters for specific quantile and ϵ is a vector of random error terms, β(τ)0 is the vector of prior means and Σ0 is the prior covariance matrix. The posterior distribution that results from identifying the quantile of interest tau and integrating the priors into the model components is given as;

$$(\beta(\tau)|Z^*,x,\tau) \propto \pi\big(\beta(\tau)\big)\prod_{i=1}^{n} \text{ALD}\,(Z_i^*|x_i^T\beta(\tau),\tau)$$

The asymmetric Laplace distribution (ALD), whose density function is defined by, is a popular basis for a likelihood function used in Bayesian quantile regression, [17]:

$$f(y|\mu,\sigma,\tau) = \tau(1-\tau)\frac{1}{\sigma}\exp\left(-\rho_\tau\frac{y-\mu}{\sigma}\right)$$

Where μ is the location parameter, σ is the scale parameter, τ is the skewness parameter, and ρτ (u) is the check function.

3. Results

3.1. Feature Selection

When it comes to data reduction or lowering the size of the coefficients, the regression model known as the Least Absolute Shrinkage and Selection Operator (Lasso) is a valuable tool. By using it, some of the coefficients are set to zero, essentially conducting feature selection and bringing down the model's complexity. Even in situations when there are exponentially more irrelevant features than in the training samples, the L_1 regularization is still successful. The penalty indicates a trait that predicts sparse estimations of the parameter vector. To comprehend shrinkage, we describe it as an instance in which data values are drawn inward, often toward the mean.

The absolute value of the coefficients, multiplied by a user-defined parameter (λ), is the repercussions term in Lasso regression. The parameter's value influences how harsh the penalty is; higher values result in a larger shrinking of the coefficients, [16].

To prevent over-fitting in the model, Lasso as a regularized ML model imposes a penalty on the magnitude of the regression coefficients. It functions by including a penalty term in the function and attempts to reduce coefficients to zero, aiding in the removal of irrelevant predictors from a model, [5]. The best-tune λ in this instance was discovered to be

0.01277063, i.e., the model with this value of λ has the optimum complexity-accuracy proportion. With this value of λ, certain characteristics were chosen as important to the model and others were eliminated.

Displayed below are the features which were of importance and those which were discarded.

Table 1. *Variable importance.*

Variable	Coefficients
(Intercept)	0.53535
sex	-0.00197
age	.
hypertension	0.06273
heart_disease	0.08426
ever_married	0.07518
bmi	0.02391
Residence_type	.
avg_glucose_level	0.07749

3.2. Classical Quantiles

The R output shows the results of a quantile regression model with a classical quantile level of 0.75, which means that the model estimates the conditional 75th percentile of the response variable (stroke) given the predictor variables (sex, hypertension, heart, married, glucose, BMI). The output below shows the coefficients of the model and their lower and upper confidence bounds.

The intercept term is the constant term in the model, representing the 75th percentile of stroke when all the predictors are zero. The coefficient for the intercept is -1.85262, which means that the 75th percentile of stroke is -1.85262 units when all the predictors are zero. The lower and upper confidence bounds for this coefficient are -5.3925 and 0.5042, respectively, which means that we are 95% confident that the true value of the coefficient lies in this interval.

Table 2. *Classical Quantiles at 0.5.*

Variable	Bayes Estimate	Lower C.I	Upper C.I
(Intercept)	-0.24325	-5.35497	-1.13640
Sex	-0.01955	-2.08729	-0.43319
Hypertension	0.50241	0.12596	4.58932
Heart-att	0.44869	0.07435	3.54558
Marital-st	0.05680	0.2922	5.11172
Glucose-lvl	0.00349	0.00294	0.04575
BMI	-0.00034	-0.07358	0.00371

Table 3. *Classical Quantiles at 0.75.*

Variable	Lower C.I	Upper C.I
(Intercept)	-1.85262	-5.3925
Sex	-1.35197	-2.7753
Hypertension	4.27658	2.20058
Heart-att	2.08125	0.00594
Marital-st	3.72934	1.7207
Glucose-lvl	0.02315	0.00856
BMI	-0.02144	-0.0889

3.3. Bayesian Quantiles

The results show the posterior estimates of the coefficients and their 95% credible intervals for a Bayesian quantile regression model with a quantile level of 0.75, which means that the model estimates the conditional 75th percentile of the response variable (stroke prediction) given the predictor variables (sex, hypertension, heart, marital status, glucose, BMI). The model is based on the asymmetric Laplace distribution, which is a flexible and robust distribution for modeling quantiles.

The intercept term is the constant term in the model, representing the 75th percentile of stroke prediction when all the predictors are zero. However, this term may not have a meaningful interpretation in this context, since some of the predictors (sex and marital status) are categorical variables that cannot take zero values. The posterior mean for the intercept is -2.09586, which means that the 75th percentile of stroke prediction is -2.09586 units when all the predictors are zero. The lower and upper bounds for this coefficient are -4.39948 and 0.1042, respectively, which means that there is a 95% probability that the true value of the coefficient lies in this interval.

Table 4. *Bayesian Quantiles at 0.5 and 0.75.*

Variable	Bayes Estimate	Lower C.I	Upper C.I	Variable	Lower C.I	Upper C.I
(Intercept)	-2.6457	-4.39114	-1.0036	(Intercept)	-2.09586	-4.39948
Sex	-0.9640	-1.55514	-0.3823	Sex	-1.12447	-1.95043
Hypertension	2.1893	1.23147	3.2691	Hypertension	3.51141	1.90967
Heart-att	0.9378	-0.16264	2.1309	Heart-att	1.66338	-0.07663
Marital-st	2.2443	1.15481	3.5106	Marital-st	2.69188	1.45101
Glucose-lvl	0.0117	0.00538	0.0181	Glucose-lvl	0.01780	0.00851
BMI	-0.0129	-0.06130	0.0340	BMI	-0.00806	-0.07738

3.4. Comparative Analysis

Based on the table, we can see that the Bayesian model has smaller widths than the classical model for all the coefficients, which means that the Bayesian model is more precise and certain than the classical model. This may be because the Bayesian model incorporates prior information and produces posterior distributions that are more informative and robust than the classical model, which relies on asymptotic approximations and produces point estimates and confidence intervals that are more sensitive to outliers and violations of assumptions. Therefore, based on the widths of the intervals, we can conclude that the Bayesian model is a better model than the classical model for stroke prediction at all the quantile levels.

Table 5. *Width difference at 0.5 and 0.75 quantiles levels.*

Variable	Classical	Bayesian	Variable	Classical	Bayesian
(Intercept)	5.9040	4.5037	(Intercept)	4.3187	3.3875
Sex	2.6681	1.6432	Sex	1.6541	1.1723
Hypertension	4.6283	3.5192	Hypertension	4.4634	2.0376
Heart-att	4.7254	3.8083	Heart-att	3.4712	2.2935
Marital-st	4.5186	2.5883	Marital-st	4.8195	2.3557
Glucose-lvl	0.0400	0.0201	Glucose-lvl	0.0428	0.0127
BMI	0.1803	0.1388	BMI	0.0773	1.1946

3.5. Tests of Convergence

After a certain number of iterations, a test of convergence may be utilized to determine if the Bayesian quantile regression model has arrived at a stable posterior distribution. Using the Gelman-Rubin diagnostic, which evaluates the variance within and across the model's multiple chains, is a typical technique for determining if convergence has occurred. For each quantile parameter, the G-R diagnostic outputs a potential scale reduction factor (PSRF), which expresses how much the chains may contract if they were run infinitely, [8]. If the PSRF is near 1, the chains have converged to the same posterior distribution, while a PSRF larger than 1 indicates that the chains have not converged and need more iterations. Upon executing iterations, it was discovered that the shrinkage levels were near to one across different quantile levels, demonstrating that the traces of the variables were converging as shown below in the marital status sample.

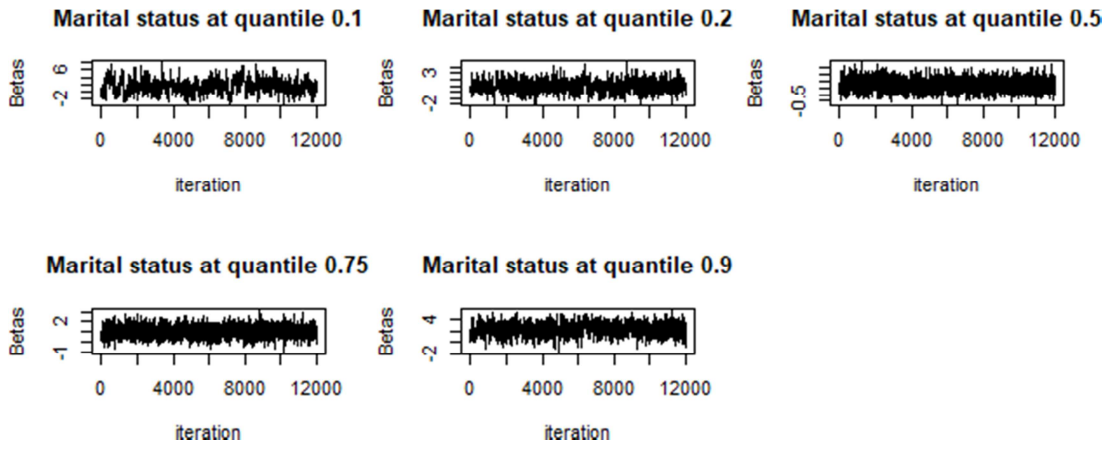

Figure 1. *Marital Status Convergence at different quantiles.*

4. Conclusion

The findings indicate that for stroke prediction in Kenya at various quantiles, the Bayesian quantile regression model performs better than the classical quantile regression model. Compared to the classical model, the Bayesian model offers more substantial fit, performance, precision, and interpretation. It may also offer more flexible and useful inferences concerning the effects of various predictors on stroke risk at various quantile levels. The findings additionally suggest that stroke risk factors at all quantiles include hypertension, heart disease, diabetes, and BMI, whereas the quantile level effects of sex and marital status differ. According to the findings, the Bayesian quantile regression model can be a valuable tool in assessing stroke risk and controlling it.

Some recommendations based on the above results are:

To use the Bayesian quantile regression model rather than the classical one.

Advise high-risk individuals to seek medical attention and adopt healthy lifestyle choices to reduce their risk of stroke and enhance their quality of life.

Investigate and contrast several approaches and priors for Bayesian quantile regression, and provide findings clearly and thoroughly.

References

[1] Alhamzawi, R., Yu, K., & Benoit, D. F. (2012). Bayesian adaptive lasso quantile regression. Statistical Modelling, 12 (3), 279–297.

[2] Benoit, D. F., & Van den Poel, D. (2009). Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services. Expert Systems with Applications, 36 (7), 10475–10484.

[3] Datta, J., & Ghosh, J. K. (2013). Asymptotic properties of bayes risk for the horseshoe prior.

[4] Gómez, G., Calle, M. L., & Oller, R. (2004). Frequentist and bayesian approaches for intervalcensored data. Statistical Papers, 45, 139–173.

[5] Hainmueller, J., & Hazlett, C. (2014). Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. Political Analysis, 22 (2), 143–168.

[6] He, F., Zhou, J., Feng, Z.-k., Liu, G., & Yang, Y. (2019). A hybrid short-term load forecasting model based on variational mode decomposition and long short-term memory networks considering relevant factors with bayesian optimization algorithm. Applied energy, 237, 103–116.

[7] Hothorn, T., & Everitt, B. S. (2014). A handbook of statistical analyses using r. CRC press.

[8] Kim, J. S., Shah, A. A., Hummers, L. K., & Zeger, S. L. (2021). Predicting clinical events using bayesian multivariate linear mixed models with application to scleroderma. BMC medical research methodology, 21, 1–12.

[9] Klau, S., Jurinovic, V., Hornung, R., Herold, T., & Boulesteix, A.-L. (2018). Priority-lasso: A simple hierarchical approach to the prediction of clinical outcome using multi-omics data. BMC bioinformatics, 19 (1), 1–14.

[10] Koenker, (2004). Quantile Regression.

[11] Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. Econometrica: journal of the Econometric Society, 33–50.

[12] Koenker, R., & Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. Journal of the American Statistical Association, 109 (506), 674–685.

[13] Kozumi, H., & Kobayashi, G. (2011). Gibbs sampling methods for bayesian quantile regression.

[14] Leung, A. A., Daskalopoulou, S. S., Dasgupta, K., McBrien, K., Butalia, S., Zarnke, K. B., Nerenberg, K., Harris, K. C., Nakhla, M., Cloutier, L., et al. (2017). Hypertension canada's 2017 guidelines for diagnosis, risk assessment, prevention, and treatment of hypertension in adults. Canadian Journal of Cardiology, 33 (5), 557–576.

[15] Liu, X., Saat, M. R., Qin, X., & Barkan, C. P. (2013). Analysis of us freight-train derailment severity using zero-truncated negative binomial regression and quantile regression. Accident Analysis & Prevention, 59, 87–93.1565–1578.

[16] Meinshausen, N., & Bühlmann, P. (2010). Stability selection. Journal of the Royal Statistical Society Series B: Statistical Methodology, 72 (4), 417–473.

[17] Sala-i-Martin, X., Doppelhofer, G., & Miller, R. I. (2004). Determinants of long-term growth: A bayesian averaging of classical estimates (bace) approach. American economic review, 94 (4), 813–835.