BAYESIAN MODEL AVERAGING FOR MATHEMATICS ACHIEVEMENT GROWTH RATE TRENDS

A Thesis

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

---

by

MINSUN KIM

Dr. Athanasios C. Micheas, Thesis Supervisor

MAY 2022

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled

BAYESIAN MODEL AVERAGING FOR MATHEMATICS ACHIEVEMENT

GROWTH RATE TRENDS

presented by Minsun Kim,

a candidate for the degree of master of and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Athanasios C. Micheas

Professor Shih-Kang Chao

Professor Ze Wang

ACKNOWLEDGEMENTS

My most sincere thanks go to Dr. Sakis Micheas, whose perceptive criticism, kind encouragement, and willing assistance helped bring the thesis to a successful conclusion. And a very special thanks to Dr. Shih-Kang Chao and Dr. Ze Wang who has selflessly encouraging my research and offering detailed and invaluable comments.

As always, my family have been there, providing all sorts of tangible and intangible support. I would like to give special thanks to my husband, Sejung Kim, and my lovely son, Joseph.

Finally, I thank God for being with me through all the difficulties and for allowing me to get through them with patience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In this study, we investigated the use of Bayesian model averaging (BMA) for latent growth curve models. We used the Trends in International Mathematics and Science Study (TIMSS) to predict growth rates in 8th-grade students' mathematics achievement. The dataset on male and female students' mathematics achievement contained 6 predictors, meaning that 64 model combinations were generated. Results highlighted science achievement score and teaching years as the most important predictors of both male and female students' growth in mathematics achievement. In this study, the growth rate of mathematical achievement for each country was compared with the predicted density and the density based on actual data. Most countries did not differ significantly in observed and predicted growth rates for male and female groups. For sensitivity analysis, the model prior had the smallest log-predictive score (LPS) value when specified as a binomial model with $m = 4$ for both the male and female data groups, regardless of the parameter prior. When comparing the fixed prior and flexible prior for parameters, the LPS value was relatively small when the fixed prior for parameters was set regardless of the model prior.

Chapter 1

Introduction

The most striking difference between the Bayesian method and the frequentist method in terms of inference is whether uncertainty is considered. In multiple studies (De Beer & Bianchi, 2017; Depaoli & Clifton, 2015; Kim & Wang, 2021; Van De Schoot et al., 2017; van Erp et al., 2018) using the Bayesian method, parameter uncertainty has been explained by setting a prior to each parameter belonging to the model. More precise and richer information is provided as a result. Yet the prior setting for parameters does not fully account for model uncertainty. According to Hoeting et al. (1999), scholars generally specify a model first and then proceed as if the model generated the data; this strategy ignores uncertainty in model selection. In other words, when random predictors are identified based on a research question, then only a few models are compared despite a combination of predictors constituting multiple models. This approach to model selection

allows for no option other than staying within the confines of the model. Furthermore, given over-confidence in the chosen model, risk is inherent in the results (Draper, 1987).

Recognizing these model selection problems presents an opportunity to address model uncertainty. Bayesian model averaging (BMA) has been proposed to mitigate the issue of model uncertainty. This approach integrates models by taking a weighted average among several possible models. Raftery et al. (1997) found the performance of a model derived via BMA to be superior to that of a single model.

Therefore, in this study, an optimal prediction model was developed with BMA. Based on mathematical achievement scores and multiple covariates measured by Trends in International Mathematics and Science Study (TIMSS) over several years, this study confirmed which set of covariates had the most influence on the growth rates of students' math achievement scores. The growth model was adopted because the covariates and outcome variables were scores that were repeated over time. In addition, by setting priors to the model itself as well as to the model parameters, uncertainties regarding the parameters and the model were considered.

This thesis is structured as follows. The next chapter introduces the growth model featured in this study; it also describes the BMA, ways to evaluate it, and priors that can be specified for the parameters as well as for the model itself. Chapter 3 provides an overview of the data and analysis process adopted herein. Chapter 4 presents the BMA results. The results of sensitivity analysis are also summarized to determine whether the findings are robust regardless of the prior used. Finally, in Chapter 5, an overall summary, conclusions, and limitations are discussed.

Chapter 2

Literature Review

This chapter provides a brief description of the latent growth curve model (LGCM) used for the probabilistic predictive model in this study. This introduction is followed by an overview of BMA including scoring rules, Zellner's *g*-priors for parameters, and model priors.

2.1 Latent Growth Curve Model

In essence, the LGCM is a method of describing longitudinal data analysis developed from the tradition of multilevel or hierarchical linear modeling (HLM) within a structural equation framework. The basic latent growth model simply transfers the growth model from the HLM method to structural equation modeling. This basic model can express growth or change over time by repeatedly measuring one variable of interest several times. In so doing, one can monitor the average of the variable of interest at a certain

time point (usually referring to the starting point; intercept) and how the variable has changed on average over time (slope). A latent variable captures a pattern of growth or change, known as a *growth parameter*; the straight line or curve tracking a growth parameter over time is called a *growth trajectory*. The fundamental form of the latent growth model repeatedly measures one variable, but it is also possible to measure several variables repeatedly and integrate them in one model.

The linear growth model in the HLM framework can be defined as follows (Raudenbush & Bryk, 2002). HLM is a multi-level model. The lowest-level model is called a Level-1 model; upper-level models are sequentially expressed as a Level-2 model, a Level-3 model, and so on. The data used in this study contained a structure with repeated measurement data at the lowest level and countries at the upper level. The Level-1 model is called the within-model or intra-model, and it models each country's individual growth trajectory:

$$y_{it} = \pi_{0i} + \pi_{1i} a_i + r_{ti}$$

where y_{it} is the outcome for country i $(i = 1, 2, \cdots, N)$ at time t $(t = 1, 2, \cdots, T)$; π_{0i} is the intercept describing country i's status on the outcome at time t; π_{1i} is the slope for country i; and r_{ti} is the residual term. The term a_i denotes the assessment cycles for country i. Coding for a_i can be expressed in numerous ways. In general, it is coded as $a_i = 0, 1, 2, 3, 4$, but coding may vary with the data properties. For example, considering that the TIMSS data employed in this study were obtained every 4 years, a_i can be coded as $a_i = 0, 4, 8, 12, 16$. Note that careful interpretation is required depending on the coding

value. The above model is a typical simple linear regression model with an independent variable a_i. It is assumed that the residual r_{ti} follows a normal distribution with a mean of 0. In this case, π_0 and π_1 are growth parameters denoting the intercept and growth rate, respectively.

The Level-2 model is called the between- or inter-model. This model attempts to explain the difference in the intercept and the slope between countries (in the case of this study) using covariates. One, multiple, or no covariates may exist. We have:

$$\pi_{si} = \beta_{s0} + \sum_{q=1}^{Q_s} \beta_{sq} x_{qi} + \epsilon_{si},$$

where π_{si} represents the growth parameters, x_{qi} are the values of Q predictors for country i, and β_{sq} are regression coefficients. Assuming a normal distribution with a mean of 0, ϵ_{si} are the residuals or errors of the intercept and slope remaining, which are accounted for by x_{qi}. In the above equation, β_{s0} are key parameters that summarize the growth trajectory. Most programs report them in output as the intercept estimate and the slope estimate. However, this interpretation only applies when there is no covariate; when an exogenous variable exists, attention should be paid to interpretation based on the set value of the exogenous variable. A model in which exogenous variables are added to Level-2 is called a conditional model, and a model in which exogenous variables are not added is called an unconditional model.

In the latent growth model, linear and nonlinear models have their own characteristics. The time points of a_i can be set to fixed values (as in this study) to estimate growth trajectories. The time points of a_i can also be set through the latent basis method,

which uses information from given data. For instance, instead of setting the time points in a_i to $a_i = 0, 1, 2, 3, \cdots$, some time points can be set to $a_i = 0, 1, 2$, while others can be estimated from the data. Such strategies provide a preferable model fit for empirical growth trajectories because the trajectories are estimated with data-based information (Kaplan & Huang, 2021). Two types of covariates are used in latent growth curve modeling: time-invariant and time-varying. Time-invariant covariates (e.g., race, sex, cognitive ability) maintain constant properties over time, whereas time-varying covariates (e.g., attitude, weight, age) fluctuate over time.

2.2 Bayesian Model Averaging

Developing a predictive model typically entails a process of selecting a final model after estimating and comparing several alternatives by setting up a set of models with given predictors. However, selecting a single model ignores the uncertainty inherent in model selection. BMA has been proposed as one way to address the problem of model uncertainty. Model averaging reduces this uncertainty by taking a weighted average (i.e., the posterior model probabilities [PMPs]) of the selected model. BMA has been found to provide better predictive performance than single-model selection (Madigan & Raftery, 1994).

Suppose we have a matrix X containing all explanatory variables in a dataset, where $\tilde{y}$ is the quantity of interest. When X contains K variables, a total of 2^K models are estimated (Madigan & Raftery, 1994). The PMP is calculated as follows according to Bayes' theorem:

$$p(M_\gamma|y, X) = \frac{p(y|M_\gamma, X)p(M_\gamma)}{p(y|X)} = \frac{p(y|M_\gamma, X)p(M_\gamma)}{\sum_{s=1}^{2^K} p(y|M_s, X)p(M_s)}, \qquad \gamma \neq s \qquad (1)$$

The posterior probability of model M_γ, $p(M_\gamma|y,X)$, can vary depending on the model being used. The posterior probability reflects the relative uncertainty of the model (Kaplan & Huang, 2021). The term $p(y|M_\gamma,X)$ in the numerator part of the above equation is the integrated likelihood given model M_γ and can be expressed as

$$p(y|M_\gamma,X) = \int p(y|\theta_\gamma,M_\gamma,X)p(\theta_\gamma|M_\gamma,X)d\theta_\gamma \tag{2}$$

where $p(\theta_\gamma|M_\gamma,X)$ is the prior distribution of the model parameters θ_γ given model M_γ and covariates X (Raftery et al., 1997). By combining the above equations, the model-weighted posterior distribution for $\tilde{y}$ given data y and X can be written as

$$\begin{aligned} p(\tilde{y}|y,X) &= \sum_{\gamma=1}^{2^K} p(\tilde{y}|M_\gamma,y,X)p(M_\gamma|y,X) \\ &= \sum_{\gamma=1}^{2^K} p(\tilde{y}\,|M_\gamma,y,X)\frac{p(y|M_\gamma,X)p(M_\gamma)}{\sum_{s=1}^{2^K} p(y|M_s,X)p(M_s)} \end{aligned} \tag{3}$$

2.3 Implementation Issues for BMA

Although BMA can address model uncertainty, implementing this approach presents some challenges (Hoeting et al., 1999). First, as the number of predictors increases, so does the number of terms in Equation (3). The model space thus expands exponentially (e.g., K predictors make 2^K model spaces). The exhaustive summation of Equation (3) becomes infeasible as a result. For this reason, several methods have been developed to

reduce the overall number of models. *Markov chain Monte Carlo model composition* (MC^3) method was applied in the *Bayesian model sampling* (BMS) package. Details of this method are provided in the following section. Second, a technical issue arises when calculating the implicit integral in Equation (3). Although the advent of MCMC has partly facilitated the integral calculation, difficulties remain when calculating an integral that contains a very large number of terms. The Laplace method (Tierney & Kadane, 1986) offers one solution for this calculation issue that returns a reasonable approximation to $p(y|M_\gamma, X)$. This outcome leads to a very small Bayesian information criterion (BIC) approximation under certain circumstances (Schwarz, 1978). The maximum likelihood estimation (MLE) approximation of Taplin (1993) involves deriving an approximation of $p(\tilde{y}|M_\gamma, y, X)$ by $p(\tilde{y}|M_\gamma, \hat{\theta}, y, X)$, where $\hat{\theta}$ is the maximum likelihood estimate, can also solve the technical issue of computation. Yet other challenges persist when specifying the model's prior distribution (i.e., $p(M_\gamma)$) and determining which model class to average.

2.4 Markov Chain Monte Carlo Model Composition

As mentioned above, the aim of MC^3 is to reduce the number of possible models that can be used in BMA. To clarify the premise of MC^3, assume that $\mathcal{M}$ is the space of the model containing all possible combinations of variables in a linear regression setting. Then the posterior probability, $p(M_\gamma|y, X)$, can be calculated by constructing a Markov chain for the model space $\mathcal{M}$ with a stationary distribution. Indexing the chain as $\{M(t), t = 1,2, \cdots, T\}$, $\hat{\tilde{y}}$ can be estimated as follows:

$$\hat{\tilde{y}} = \frac{1}{T}\sum_{t=1}^{T} g(M(t)), \qquad t = 1, 2, \cdots, T$$

where the function $g(M(t))$ calculates the quantity $\tilde{y}$ for the model $M(t)$, and this quantity converges to the true value of $\tilde{y}$.

Such a chain $M(t)$ can be constructed using the Metropolis–Hastings (M–H) algorithm. At each step of this algorithm, every model is compared to determine which model is retained. Given a current model $M \in \mathcal{M}$, M' can be defined as the neighbor of M containing all models with one additional or fewer covariates than the current model. The M–H algorithm then evaluates whether a new model from the neighborhood can replace the current model with a specified probability of acceptance:

$$P(Accept\ M') = \min(1, \alpha), \qquad where\ \alpha = \frac{p(M'|y)p(M')}{p(M|y)p(M)}.$$

If M' is accepted, M' becomes the current model. Otherwise, the chain is kept as-is in the current model (Fraley & Percival, 2015; Hoeting et al., 1999).

2.4.1 *Scoring Rules*

Multiple predictive models can be implemented based on the use of numerous prior choices for parameters and models. It is important to derive an optimal model by evaluating each model's predictive performance (Dawid & Musio, 2015; Eicher et al., 2011; Zeugner & Feldkircher, 2009). The method of evaluating prediction accuracy is called *scoring rules*, and various such rules have been suggested (Bernardo & Smith, 2009; Carvalho, 2016;

Dawid & Musio, 2015; Gneiting & Raftery, 2007; Merkle & Steyvers, 2013). The log-predictive score (LPS) method built into the BMS package was used in this study. The LPS is defined as follows:

$$LPS = -\sum_{i} \log\left(p\left(\widetilde{y}_i \middle| X, y, \widetilde{X}_i\right)\right),$$

where $p\left(\widetilde{y}_i \middle| X, y, \widetilde{X}_i\right)$ denotes the predictive density for $\widetilde{y}_i$ based on the model information X and y, and $\widetilde{X}_i$ represents the explanatory variables for the prediction model $\widetilde{y}_i$. Note that model comparisons are only meaningful when there are different prediction settings. Generally, the smaller the LPS value, the better the prediction performance.

2.4.2 *Parameter Priors*

One way that the Bayesian framework differs from the frequentist framework is that priors should be specified to the model parameters to derive posterior distributions. A commonly used prior structure is based on Zellner's g priors (Zeugner & Feldkircher, 2015). Taking the simple linear regression equation as an example, a g-prior can be specified as follows:

$$y = \alpha_\gamma + X_\gamma \beta_\gamma + \epsilon, \qquad \epsilon \sim N(0,\ \sigma^2 I).$$

In each Model M_γ presented in the above equation, suppose that the constant term and the error term are specified by as improper prior, such that $p\left(\alpha_\gamma\right) \propto 1$ and $p(\sigma) \propto \sigma^{-1}$. In

other words, these two terms are evenly distributed in their domains. The remaining term, β_γ, is usually the parameter of interest and is typically assumed to follow a normal distribution with certain mean and variance values. When there is no information about the parameter, the mean value is set to 0 and the variance value is set according to Zellner's *g*-prior in a variance–covariance structure that is nearly identical to the data structure. The prior for the parameter β_γ is

$$\beta_\gamma | g \sim N\left(0,\ \ g\sigma^2\left(X_\gamma^T\, X_\gamma\right)^{-1}\right)$$

In this case, the hyperparameter *g* refers to the level of confidence in the specified value (mean = 0 in the example above). If the *g* value is small (i.e., the variance of the parameter is small), the specified value is quite certain. Conversely, the larger the *g* value, the greater the variance in the parameter—indicating weak confidence that the mean value is 0. In addition, the posterior distribution of the parameter reflects the prior uncertainty. Given *g*, the posterior distribution of β_γ follows a *t*-distribution, and the expected posterior and variance posterior distributions of β_γ are then:

$$E\left(\beta_\gamma \middle| y, X, g, M_\gamma\right) = \frac{g}{1+g}\hat{\beta}_\gamma$$

$$Cov\left(\beta_\gamma \middle| y, X, g, M_\gamma\right) = \frac{(y-\bar{y})^T\,(y-\bar{y})}{N-3}\,\frac{g}{1+g}\left(1-\frac{g}{1+g}\,R_\gamma^2\right)\left(X_\gamma^T\, X_\gamma\right)^{-1}$$

where $\hat{\beta}_\gamma$ and R_γ^2 are the standard ordinary least squares (OLS) estimators for M_γ. Both the posterior mean and the posterior variance values are composed of a combination of the *shrinkage factor*[1], $\frac{g}{1+g}$, and each OLS estimator. When the value of g is small, the influence of the prior becomes relatively large and the expected value approaches 0. By contrast, when the value of g increases, the shrinkage factor approaches 1, and the influence of the OLS estimator is relatively large.

The prior g can be specified in two ways under the BMS package: *fixed* and *flexible* priors. Table 1 summarizes how each prior is commanded in the BMS package, and the prior value is specified accordingly.

The fixed prior assigns the positive real scalar value to g, leading to four options (Fernandez et al., 2001): (1) unit information prior (UIP); (2) risk inflation criterion prior (RIC); (3) benchmark risk inflation criterion prior (BRIC); and (4) Hanna and Quinn prior (HQ). A fixed g-prior is commonly used but has unintended consequences. When the value of g is large, the shrinkage factor approaches 1, and the posterior estimation tends to be easily overfitted. This scenario affects not only the estimation coefficient but also the model estimation. Large over-fitting shrinkage factors result in tight PMP concentrations and small model sizes. Therefore, the posterior inclusion probability (PIP) distribution has a skewed distribution with relatively high PIP values for some variables and very low PIP values for the remaining variables. Conversely, when the value of g is small, most

[1] Assume that the posterior mean is given as a weighted combination of the prior mean and the observed data mean, $\hat{\mu} = \frac{\sigma^2}{\sigma^2+n\tau^2}\kappa + \frac{n\tau^2}{\sigma^2+n\tau^2}\bar{y}$, and that these weights are bounded between 0 and 1. These weights constitute the *shrinkage factor* (Kaplan, 2014).

covariates have similar PIP values; they therefore lack discriminatory power in model estimation (Zeugner & Feldkircher, 2015).

To compensate for these problems, the proposed flexible *g*-priors used in the BMS package are as follows: (1) the empirical Bayes local (EBL) prior and (2) the hyper-*g*-prior. In determining the EBL prior, the *g*-prior is estimated via maximum likelihood estimation. Given information contained in the data, this *g*-prior has a different value for each model. Consequently, the value of the shrinkage factor, $\frac{g}{1+g}$, may vary depending on the model. Note that the *g*-prior does not guarantee asymptotic "consistency" (Liang et al., 2008). The hyper-prior allows for prior assumptions about g while relying on the given data. This circumstance retains the benefits and minimizes the pitfalls of using fixed values. The hyper *g*-prior is a Beta prior on the shrinkage factor with $\frac{g}{1+g} \sim B(1, \frac{\alpha}{2} - 1)$, with $E\left(\frac{g}{1+g}\right) = \frac{2}{\alpha}$. Instead of setting the g value directly, the shrinkage factor value is derived by setting the hyperparameter $\alpha \in (2, 4]$. As α approaches 2, the prior expected shrinkage factor is close to 1. When α is 4, the prior distribution on the shrinkage factor is uniformly distributed over $[0, 1]$. In this study, the hyper-*g* could be set to "UIP" and "BRIC." With the hyper-*g* as UIP, the prior expected shrinkage factor is $E\left(\frac{g}{1+g}\right) = \frac{N}{1+N}$, with $\alpha = 2 + \frac{2}{n}$. If the hyper-*g* is BRIC, then the prior expected shrinkage factor $E\left(\frac{g}{1+g}\right)$ is the same as with the BRIC prior. These latter two options guarantee asymptotic consistency (Zeugner & Feldkircher, 2009).

Prior	Argument	Set
Fixed g		
g-prior that specifies a value directly	$g = x$	$g = x$, where x is a positive real scalar.
Unit information prior	g = "UIP"	$g = N$
Risk inflation criterion prior	g = "RIC"	g = Q^2
Hanna and Quinn prior	g = "HQ"	$g = log(N)^3$
Flexible g		
Empirical Bayes (local)	g = "EBL"	$g_\gamma = \max(0, F_\gamma - 1)$
Hyper-g-prior	g = "hyper = x"	$\alpha = x$, where x is a positive real scalar.
	g = "hyper = UIP"	$\alpha = 2 + \frac{2}{n}$
	g = "hyper = BRIC"	$E\left(\frac{g}{1+g}\right)$ to be equivalent to the BRIC prior.

Note. $F_\gamma \equiv \frac{R_\gamma^2(N-1-k_\gamma)}{(1-R_\gamma^2)k_\gamma}$, and R_γ^2 is the OLS R-squared of model M_γ.

Table 1: Overview of g-priors

2.4.3 *Model Priors*

A prior distribution for a model can be assigned to various prior distributions. In this study, the three types of model priors provided by the BMS package for R were investigated, namely (1) the uniform model prior, (2) the binomial model prior, and (3) the binomial-beta model prior.

Uniform Model Prior. Assuming there are K predictors, there are 2^K possible variable combinations and a total of 2^K models. When applying a uniform model prior that gives the distribution of each model as 2^{-K}, a prior expected model size is $\sum_k^K \binom{K}{k} k2^{-K} = K/2$. For example, with six variables, the expected model size is $K/2 = 6/2 = 3$. Given that the probability of the model size being 3 is greater than that for other sizes, the uniform model prior shows more mass around this size ($K/2 = 3$), resulting in a symmetric distribution (Zeugner & Feldkircher, 2015).

Binomial Model Prior. The binomial model prior is mainly taken as a proxy of the uniform model prior (Zeugner & Feldkircher, 2015). Constructing the binomial model prior entails assigning a model size by placing a common and fixed inclusion probability, θ, on each predictor. Then, the prior probability for a model of size k_γ is $p(K_\gamma) = \theta^{k_\gamma}(1-\theta)^{K-k_\gamma}$, and the expected model size is $\bar{m} = K\theta$. When setting $\theta = 0.5$, the prior model size is $\bar{m} = K/2$, which is the same as for the uniform model prior. Thus, θ can be set to less than or greater than 0.5 (note that the theta value lies between 0 and 1 according to the definition of probability).

Binomial-beta Model Prior. Ley and Steel (2009) suggested that specifying a model size neglects prior uncertainty and can lead to unintended results. Therefore, the inclusion probability θ was randomly treated in this study in order to assign a less tight

prior to the prior expected model size. The probability distribution of θ follows the Beta distribution with hyperparameters: $\theta \sim Beta\ (\alpha, \beta)$ for $\alpha, \beta > 0$.

Chapter 3

Data Overview and Methodology

3.1 Sample and variables

This study used data collected from countries that continued to participate in a total of five assessment cycles from TIMSS 2003 to TIMSS 2019. Data consisted of responses from 8th-grade students, their mathematics teachers, and school principals in 16 countries. Other countries that participated in 3 or 4 assessment cycles (i.e., not all cycles) were excluded from analysis due to a lack of information.

Plausible values on mathematics for each cycle were represented as the mathematics achievement scores at five time points in the growth model. TIMSS assesses a limited number of items per student to minimize students' response burden and generates plausible values through a scaling process to accurately estimate the relevant scale score. Therefore, a plausible value cannot be regarded as an individual score for each student.

Plausible values nevertheless have an advantage: they offer unbiased estimates of population characteristics. TIMSS provides five plausible value sets for mathematics in each cycle; the first value set was used in the current study (Foy et al., 2019).

A set of academic-related noncognitive variables, science achievement score, teachers' teaching years, homework frequency, and a school-related variable served as covariates for predictive modeling. Noncognitive variables and science achievement scores were obtained from students' responses; teachers' teaching years and homework frequency were acquired from mathematics teachers' responses. A set of school-related variables, school discipline and safety, came from principals' responses. Variables related to self-concept and value in mathematics were classified as academic noncognitive variables. Self-concept of mathematics, which was composed of one's self-concept of mathematics competence and difficulty, was measured with 4 items scored on a Likert-type scale anchored by 1 (*strongly agree*) and 4 (*strongly disagree*). The degree of value in mathematics was measured with 5 items on the same scale. Science achievement scores also used plausible values estimated based on students' responses. Similar to the mathematics plausible value score, the first plausible value of five science plausible values was taken as the science achievement score. Teachers were asked how many years they had been teaching mathematics, and they answered with their actual number of years spent teaching the subject. The frequency of homework assignment was measured on a scale of 1 (*less than once per week*) to 3 (*more than 3 times per week*). The school-related variable included 11 items scored on a Likert-type scale of 1 (*a lot*) to 4 (*not at all*) to measure school discipline and safety. All scales were recoded so that positive responses had high scores (Fishbein et al., 2021; Gonzalez et al., 2003).

3.2 Methods for Bayesian LGCM

Before implementing Bayesian LGCM, a latent growth curve model that fit the given data well was developed through preliminary analysis. As a result, in this study, fixed time points were used and residual errors were constrained to 1.1 to guarantee convergence. The model adopted in this study is shown in Figure 1.

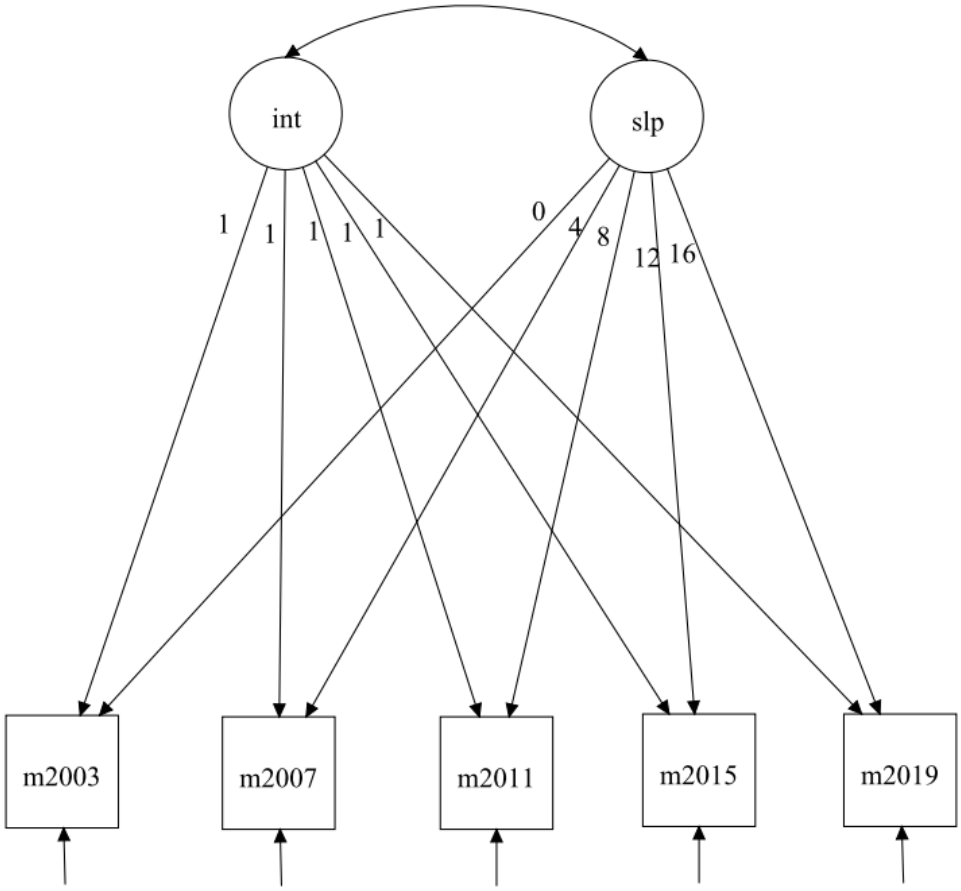


Figure 1: Diagram for Latent Growth Curve Model

Since the latent growth curve model has been extended to the Bayesian framework, prior specification is required for all parameters belonging to the model when applying this approach. The approximate mean of mathematics achievement scores in previous studies was 550; as such, the informative prior in this case was set to $N(550, .1)$ for the intercept. A noninformative prior distribution was specified for all other parameters (see Table 3 in Chapter 4 for details).

The Bayesian growth curve model was estimated using MCMC sampling via the Gibbs sampler. The analysis used 1,000 adaptations, 5,000 burn-in iterations, and 100,000 post-burn-in iterations with 2 chains. In terms of model convergence for the chains, a visual check was also performed through trace plots, density plots, and autocorrelation plots along with Potential Scale Reduction Factor (PSRF) values.

3.3 Analysis of BMA

Thus far, BMA does not seem to address time-varying predictors in longitudinal models. However, because all predictors in this study were time-varying, they needed to be converted properly. Following guidance from Kaplan and Huang (2021), new time-invariant predictors were created by calculating the difference between the variable values of the most recent data (i.e., from 2019) and the oldest data (i.e., from 2003). The new time-invariant predictors were then included in the model instead of time-varying variables. Kaplan and Huang (2021) conceded that this method of handling the time predictor is not ideal but explained that the approach can account for the time predictor's characteristics. The default settings of the BMS package were used when running BMA (Zeugner & Feldkircher, 2015). That is, the optimal model for both male and female datasets was found

using the UIP default prior for the parameters and the uniform model prior for the model. This execution involves evaluating the importance of the predictor included in the model and the optimal model size considering parameter and model uncertainty.

The posterior predictive densities of growth rates were derived from the BMA results. This density plot can visually confirm whether the growth rates of the predictive model match the actual growth rates. If the density of the predictive model and the model based on actual data differs significantly, it is necessary to either revise the predictive model or check the data characteristics.

Finally, sensitivity analysis was conducted by applying different priors for parameters and models. In studies using the Bayesian method, sensitivity analysis to confirm the influence of the prior is essential (Depaoli & van de Schoot, 2017). Therefore, in this study, several steps were taken: comparing the results of applying various priors; verifying whether the result was robust according to the prior; interpreting the prediction model; and developing the model as necessary. The sensitivity of the prior was confirmed using LPS, a scoring rule built into the BMS package described in Chapter 2. The overall analysis process for estimating the Bayesian probabilistic prediction model followed the workflow suggested by Kaplan and Huang (2021).

Chapter 4

Data Analyses and Interpretation

4.1 Bayesian growth curve modeling results

Results of the Bayesian growth curve model for 8th-grade mathematics achievement are presented in Table 2. The findings are reported by gender; the upper panel displays results from males, and results from female students appear in the bottom panel.

Within the data from male students, the Potential Scale Reduction Factor (PSRF) values for all parameters were less than 1.01. The parameter distributions were all acceptable by investigating the parameter trace plots, kernel density plots, and autocorrelation plots. All parameters therefore converged to their respective posterior distributions. The posterior estimate of the growth rate parameter for mathematics achievement was 0.921, and the posterior standard deviation (SD) was 0.270. The 95% highest posterior density (HPD) reflected a 95% probability that the true value of the growth rate for mathematics achievement was between 0.390 and 1.464. It is noted that the

95% HPD is similar to the 95% credible interval (i.e., an equal-tailed interval) when the posterior distribution is unimodal and symmetric, but HPD values are better explained when the distribution is multimodal or skewed (Gelman et al., 2014). The results from female students showed similar trends: all parameters also converged to each posterior distribution with PSRF<1.01 and visual inspection of the parameter distributions. The posterior estimate of the growth rate for female students' mathematics achievement was 0.772, and the posterior SD was 0.271. The 95% HPD ranged from 0.245 to 1.321.

Overall, the growth model that constrained the residual variance to 1.1 and that set a noninformative prior for all parameters except the intercept was well suited to estimating the growth trajectory for both student groups. Upon comparing the posterior estimates, the growth rate estimate for male students was higher than for female students. However, because both growth rates were less than 1, the trend in the growth rate was not pronounced and appeared flat. An overall trend plot (i.e., covering all countries) and trend plots for selected countries are shown in Figures 2–5. Trend plots for each country are available in the Appendix A.1 – A.2.

	Estimate	Post. SD	HPD.025	HPD.975	PSRF	Prior
Male students' growth parameters						
Intercept	548.248	3.115	542.367	554.465	1.000	dnorm(550, .1)
Slope	0.922	0.270	0.388	1.461	1.000	dnorm(0, 1e-2)
Pre(Intercept)	2573.880	981.311	1048.870	4471.620	1.000	dgamma(1, .5)
Pre(Slope)	1.048	0.424	0.432	1.878	1.000	dgamma(1, .5)
Female students' growth parameters						
Intercept	548.297	3.103	542.201	554.336	1.000	dnorm(550, .1)
Slope	0.772	0.271	0.245	1.321	1.000	dnorm(0, 1e-2)
Pre(Intercept)	2682.837	1042.916	1175.480	4788.936	1.000	dgamma(1, .5)
Pre(Slope)	1.044	0.426	0.420	1.866	1.000	dgamma(1, .5)

Note. Pre() refers to the precision of the parameter, where precision = 1 / variance. *dnorm* is the normal distribution, $N(\mu, \sigma^2)$, where μ is the location and σ is the scale. *dgamma* is the gamma distribution, $G(\alpha, \beta)$, where α is the shape parameter and β is the inverse scale.

Table 2: Bayesian growth curve modeling results

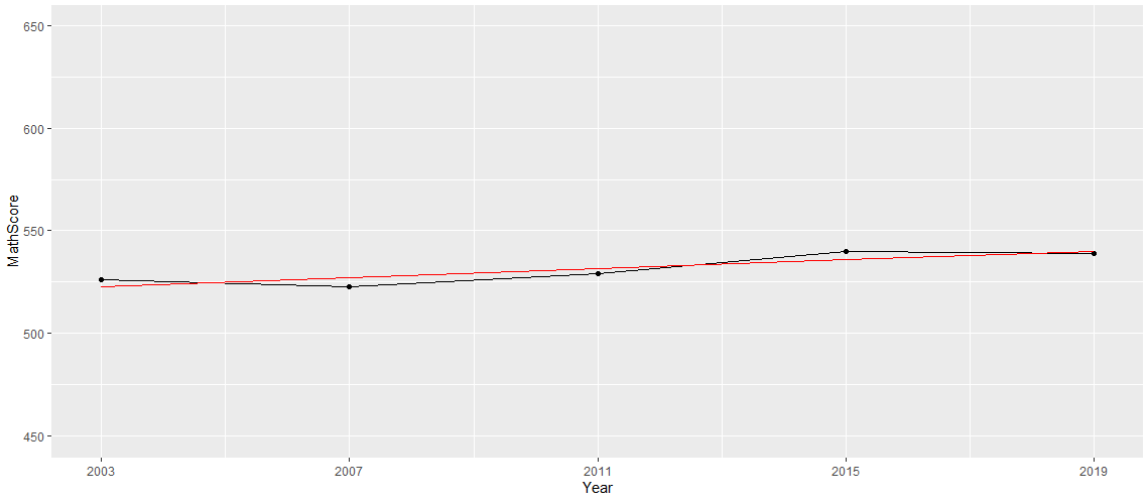


Figure 2: Fitted Trend Plot for Mathematics Achievement for Male Students in All Countries

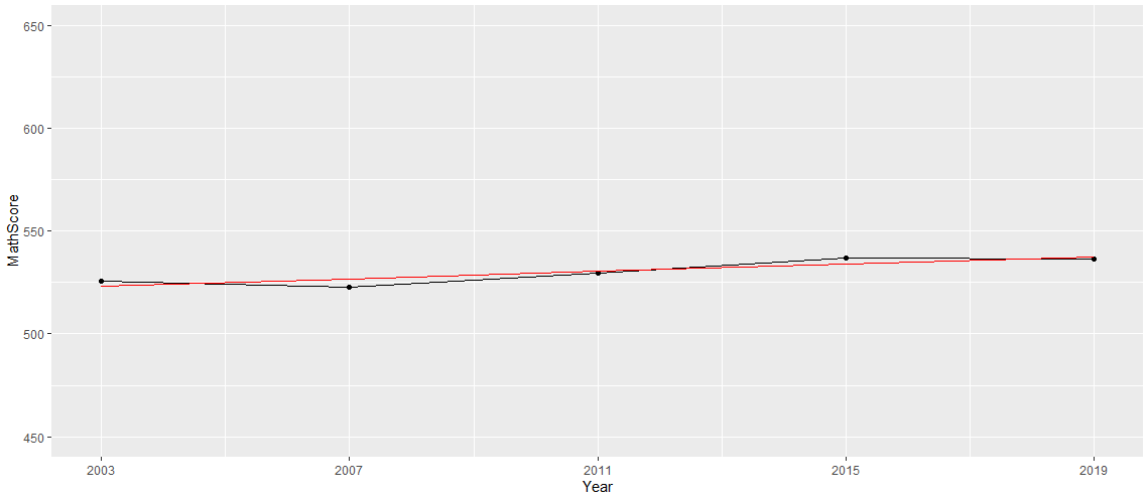


Figure 3: Fitted Trend Plot for Mathematics Achievement for Female Students in All Countries

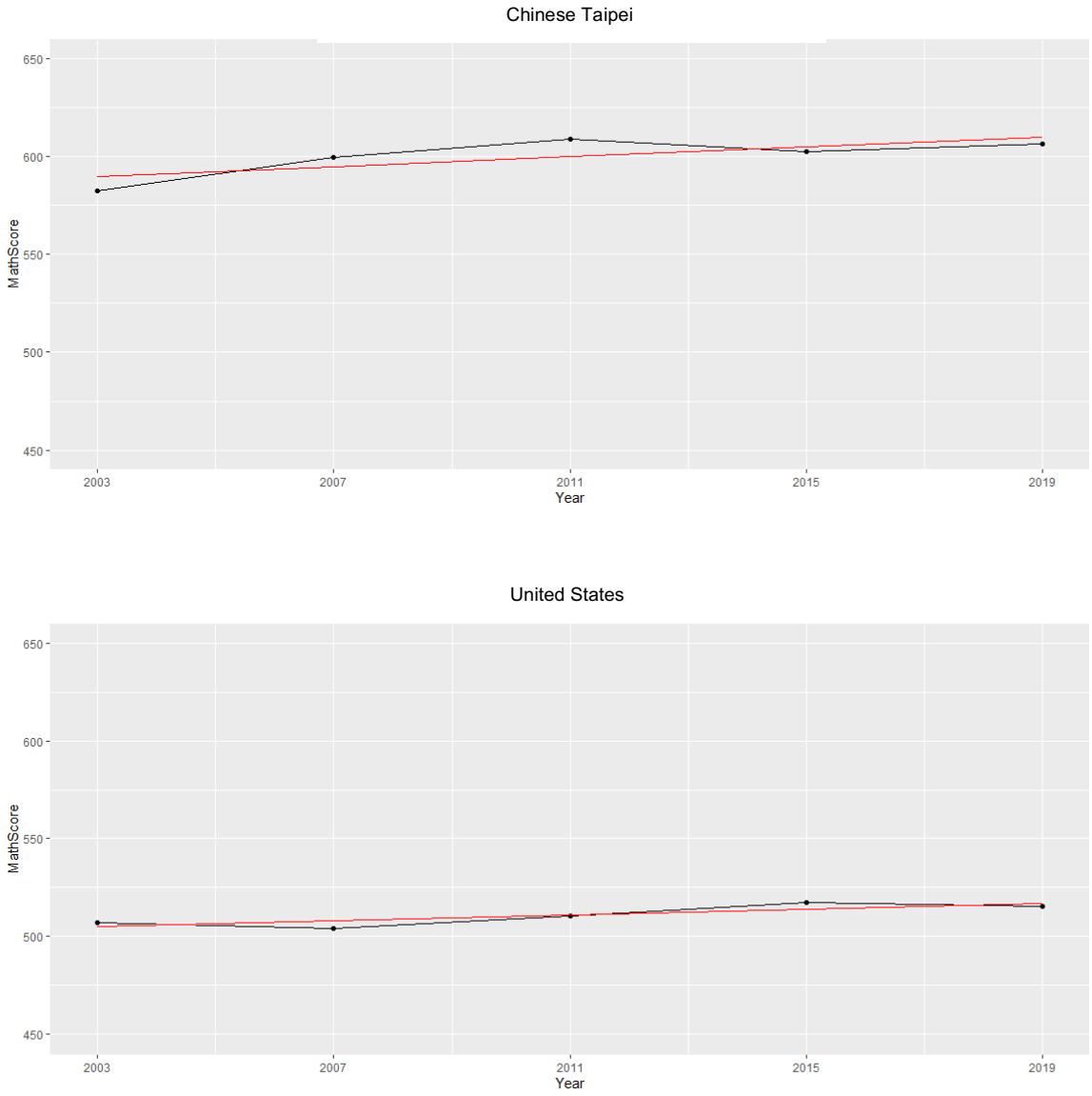


Figure 4: Fitted Trend Plots for Mathematics Achievement for Male Students in Selected Countries

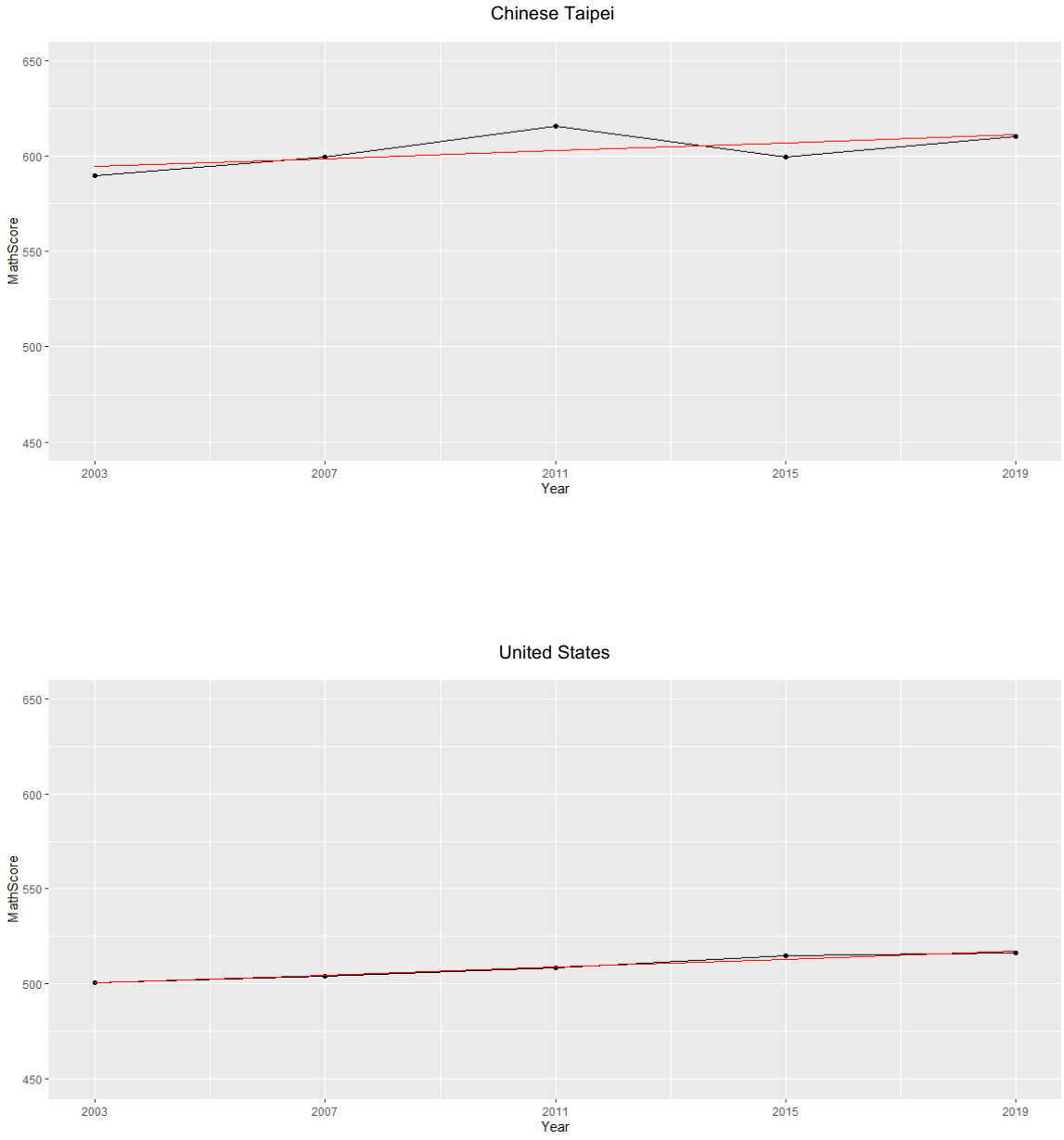


Figure 5: Fitted Trend Plots for Mathematics Achievement for Female Students in Selected Countries

4.2 Bayesian model averaging results

The dataset on male and female students' mathematics achievement contained 6 predictors, meaning that $2^6 = 64$ model combinations were generated. The growth rate (i.e., slope) obtained from Bayesian growth curve modeling was taken as an outcome variable. Bayesian model averaging was performed using the default prior setting for both parameters (i.e., the unit information prior on Zellner's g) and the model (i.e., a uniform model prior). The results for Bayesian model averaging are displayed in Table 3.

Each column of Table 3 is interpreted as follows. The posterior inclusion probability (PIP) column describes the importance of that variable, and its value is equal to the sum of the posterior model probabilities (PMPs) for all models that include that variable. Another statistic that can explain a variable's importance, the posterior mean (i.e., the "Post Mean" column), is the averaged coefficient of that variable over all models (i.e., a total of 64 models). In models that do not include this variable, the coefficient value of the variable is assumed to be 0. The "Post SD" column lists coefficients' posterior standard deviations. The "Cond. Pos. Sign" column indicates the posterior probability that the coefficient of this variable is positive in the model that includes this variable.

	PIP	Post Mean	Post SD	Cond. Pos. Sign
Males				
Science score	0.87	0.43	0.24	1.00
Teaching years	0.79	0.41	0.28	1.00
School discipline	0.67	0.28	0.26	1.00
Math Self-concept	0.34	0.07	0.16	1.00
Freq. homework	0.28	0.05	0.15	0.95
Math value	0.28	0.05	0.15	0.95
Females				
Science score	0.92	0.42	0.19	1.00
Teaching years	0.91	0.52	0.25	1.00
School discipline	0.77	0.30	0.22	1.00
Math Self-concept	0.71	0.22	0.18	1.00
Math value	0.49	0.14	0.19	1.00
Freq. homework	0.26	0.03	0.11	0.94

Note. PIP = posterior inclusion probability; Post Mean = expected a posterior estimate; Post SD = posterior standard deviation; Cond. Pos. Sign = probability that the sign of the estimate is positive conditional on inclusion in the model. Values for Post Mean and Post SD are standardized. The order of covariates is sorted by PIP.

Table 3: Bayesian Model Averaging Results

Among the data from males, the PIP of the science score was 87%; that is, 87% of the posterior model mass included science score predictors. This variable seems to be important in predicting the growth rate of male students' math achievement. Relatedly, the number of years a teacher had taught math was key to the math-score growth rate based on the PIP score of 79%. By contrast, the homework frequency and math value variables each only had a PIP value of 28%. Since the Post Mean values of these variables were quite small, the coefficient values of the variables were 0 in most models. Also, because the Cond.Pos.Sign of these variables was not 1, the coefficient of these variables was occasionally negative.

For the female student data, the PIP of the science score and the number of years teaching was 92% and 91%, respectively, showing significantly high values. Those variables thus appeared important for predicting the growth rate of female students' math achievement. Conversely, homework frequency had 26% PIP and 0.03 Post Mean values. The value of this variable was therefore 0 in most models, and its role in predicting mathematical achievement was quite low. In the same vein, among data on female students, the Cond.Pos.Sign of all variables other than homework frequency was 1. Taking the school discipline and safety variable as an example, the coefficient of this variable was positive in virtually all models that included it.

When comparing male and female results, science score and teaching years were critical in predicting growth in math achievement in both groups of data. However, a difference was observed in the degree of PIP values. Those for the female data were relatively large (92% and 91%, respectively), indicating that these variables played more significant roles in predicting the growth rate of female students' mathematical

achievement. Also, the PIP for math value was 0.49 for the female student group, which was more influential than for male students. The results for both data groups showed that Cond.Pos.Sign was less than 1 for the homework frequency variable and that the coefficient sign was positive and sometimes negative in models including this variable.

The marginal densities of the posterior coefficient distribution for several variables are displayed below (see Figure 6 and Figure 7). These plots are a visual depiction of the data in Table 3. The posterior density was normal given this model, and the mean and median values of coefficients were well aligned; 95% posterior probability intervals are denoted by dashed lines.

Table 4 presents the five models with the highest PMP per dataset. For male students, the best model (with 21% posterior model probability) included the science score, teaching years, and school discipline and safety variables. For female students, the best model (also with 21% posterior model probability) included math self-concept, school discipline and safety, teaching years, and science score.

	Model1	Model2	Model3	Model4	Model5
PMP(Male)	.21	.09	.07	.07	.06
PMP(Female)	.21	.21	.08	.07	.05

Table 4: Posterior Model Probabilities (PMPs) for Top Five Models

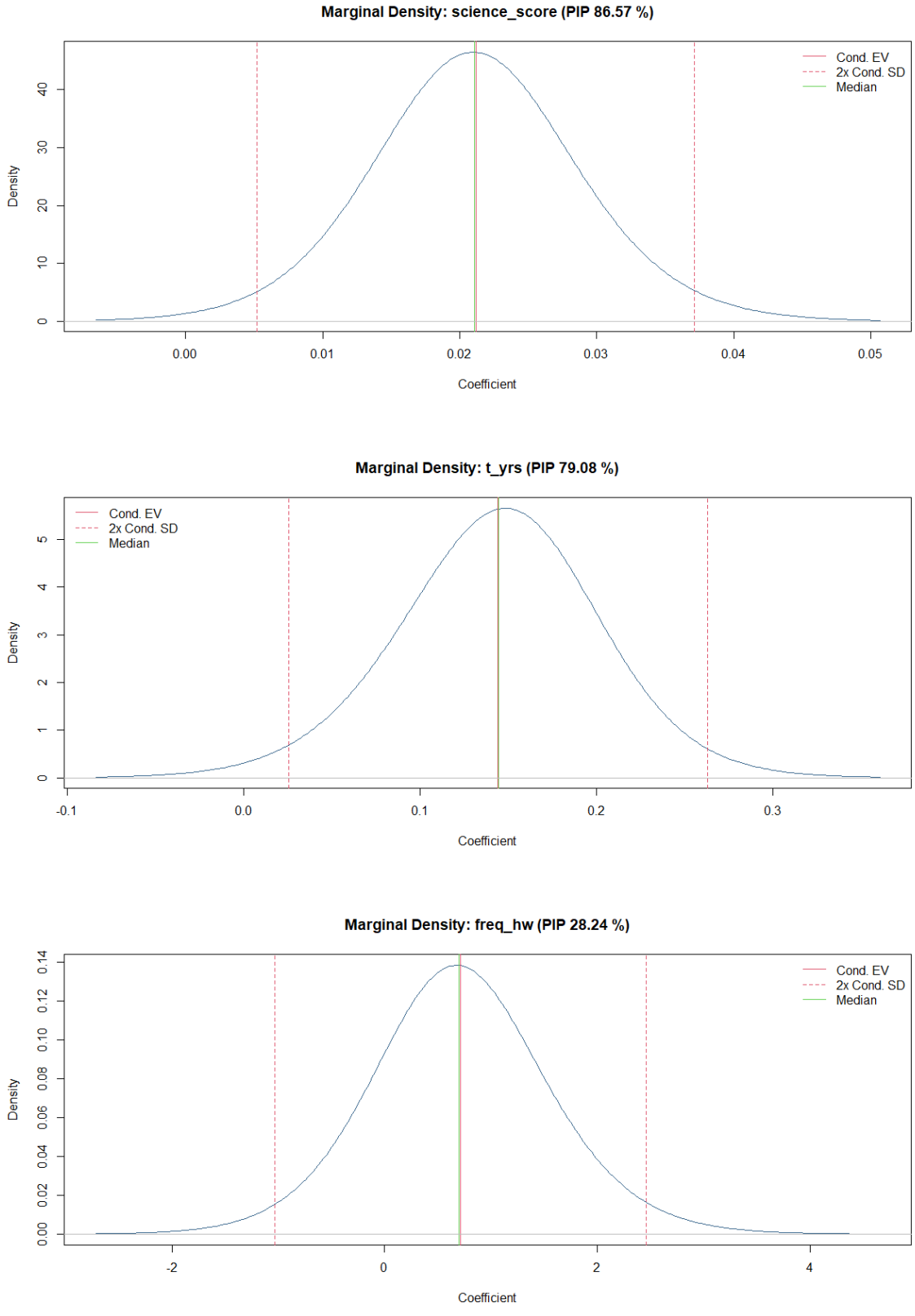


Figure 6: Posterior Coefficient Density Plots for Selected Variables for Male Students

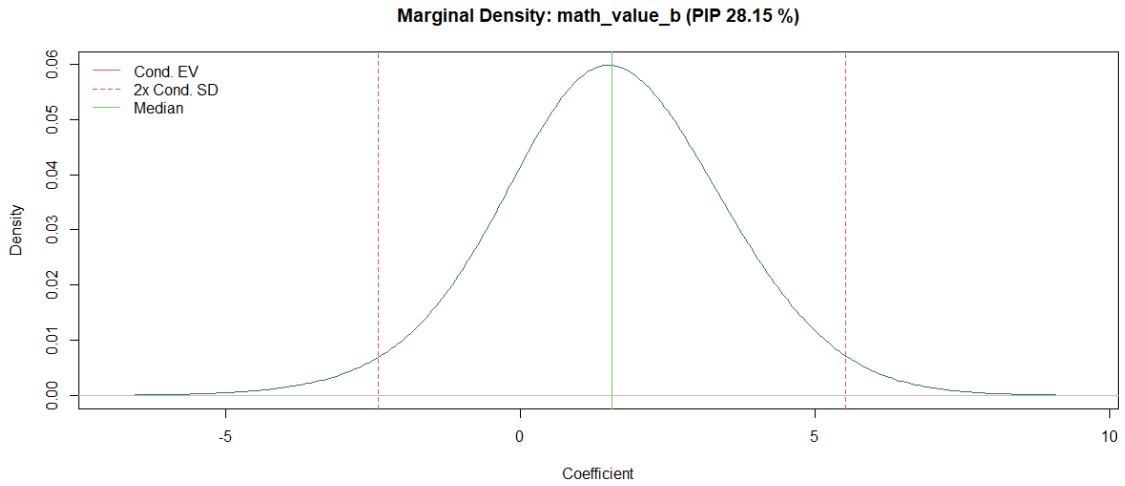


Figure 6 (Continued): Posterior Coefficient Density Plots for Selected Variables for Male Students

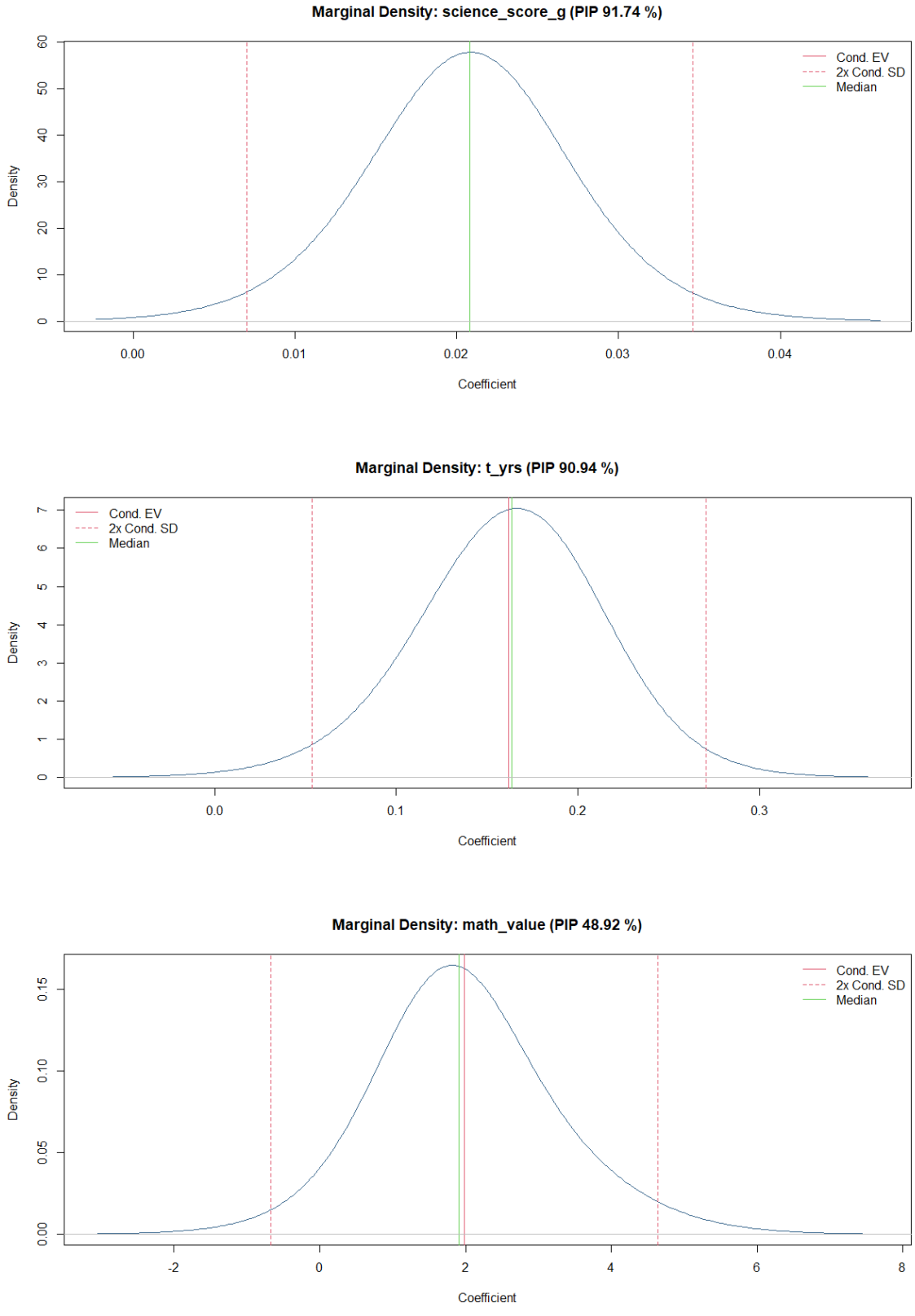


Figure 7: Posterior Coefficient Density Plots for Selected Variables for Female Students

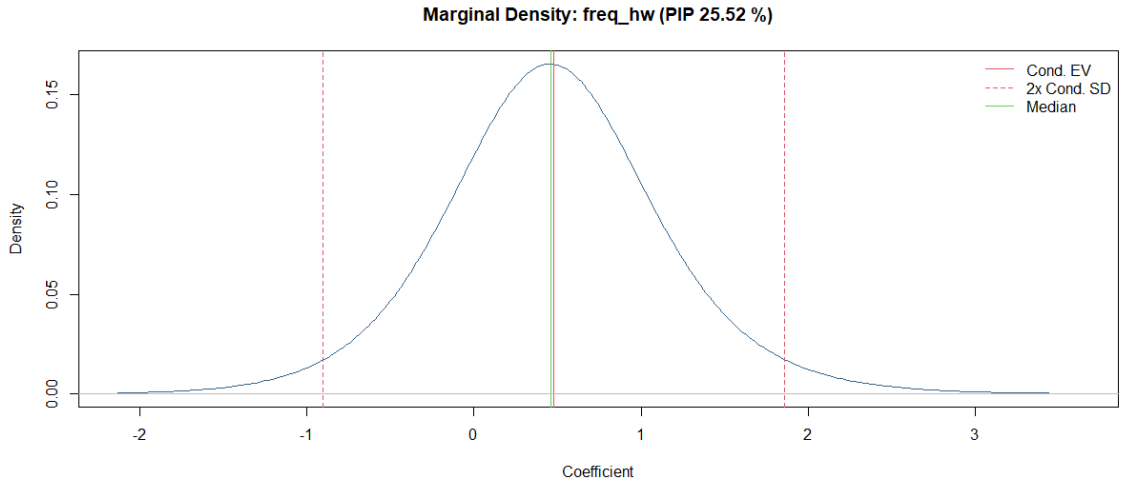


Figure 7 (Continued): Posterior Coefficient Density Plots for Selected Variables for Female Students

Figures 8 and 9 provide a more detailed illustration of the models for each dataset. The horizontal axis indicates models in the order of PMP values and reflects cumulative probability. Each variable is expressed with three colors representing the coefficient sign or inclusion of that variable in the models: blue signifies that the coefficient of a variable is positive, red signifies that the coefficient is negative, and white signifies that the variable is not included in the model. Taking male student data as an example, 0.21 on the horizontal axis represents the mass of the model with the largest PMP value (i.e., the best model). The value of 0.3, on the horizontal axis, is the cumulative PMP value of the model with the largest PMP value and the model with the next largest PMP value. The best model for male students included the science score, years of teaching, and school discipline and safety variables.

Bayesian growth curve models generate predictive densities, and a mixture of them yields the BMA predictive density. Comparing this BMA predictive density with the actual

density from data can determine whether the predictive density is adequate. In this study, the growth rate of mathematical achievement for each country was compared with the predicted density and the density based on actual data. Most countries did not differ significantly in observed and predicted growth rates for male and female groups. The predictive model therefore demonstrated a good fit. Figures 10 and 11 are a posterior density plot of the growth rate of math achievement for male and female students in the United States.

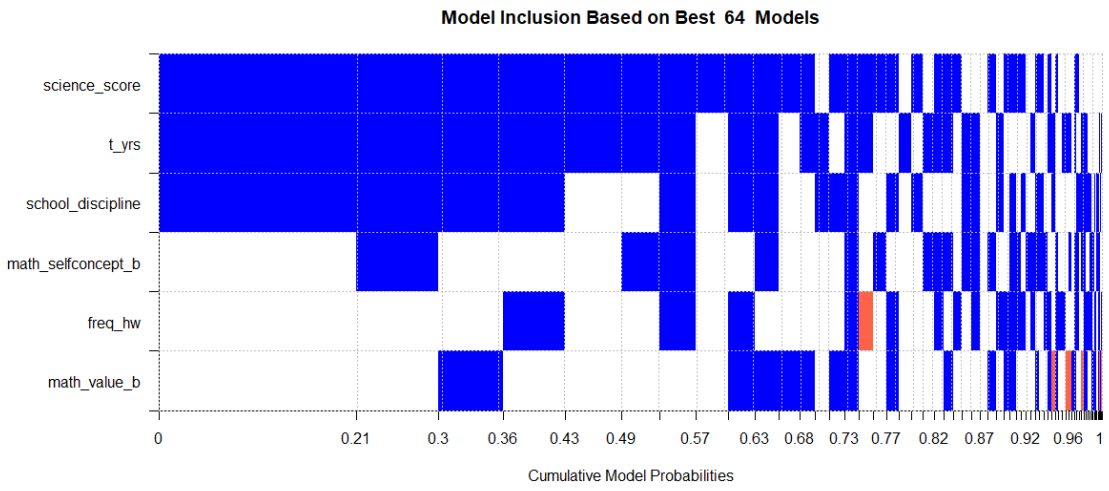


Figure 8: Posterior Model Probabilities for Each Model for Male Students

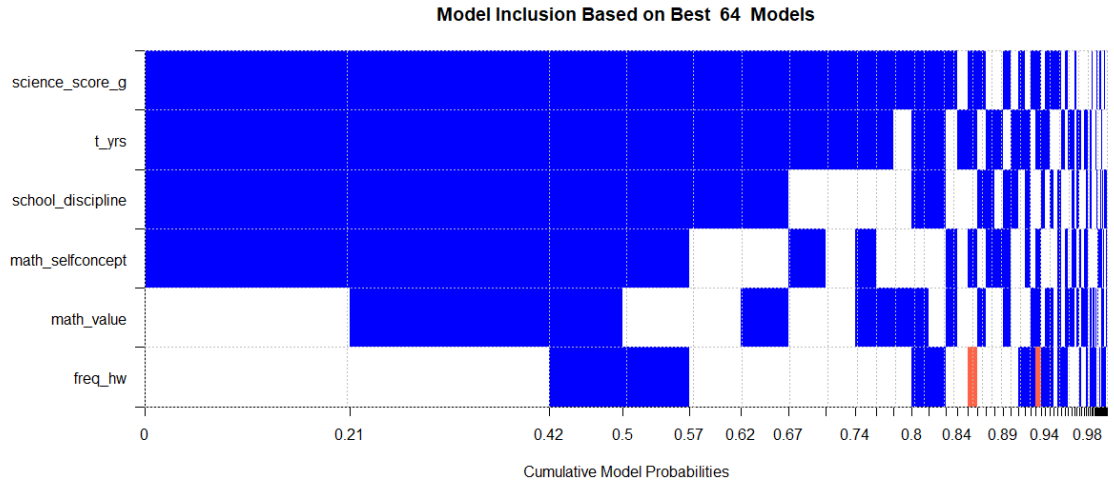
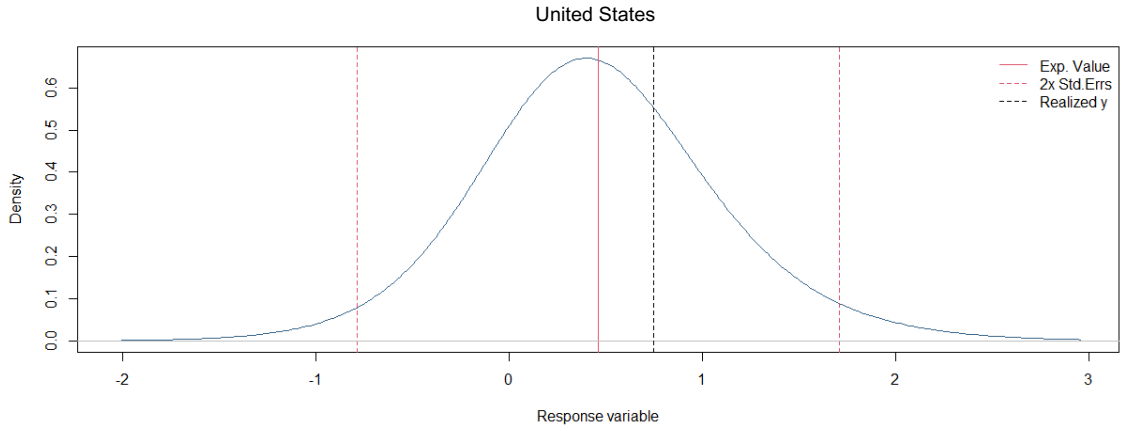
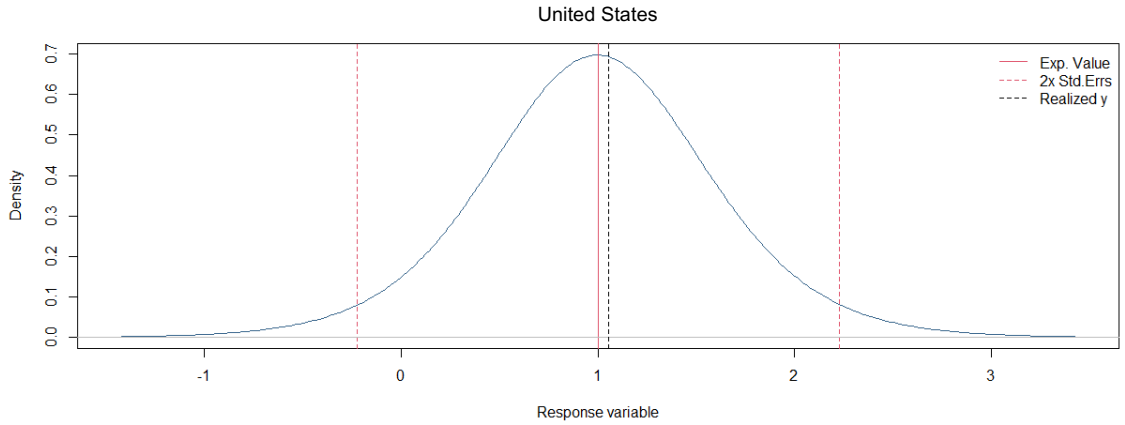


Figure 9: Posterior Model Probabilities for Each Model for Female Students



Note. The dashed line is the actual growth rate; the solid line is the model-predicted growth rate based on BMA.

Figure 10: Posterior Density Plots for Male Students in United States



Note. The dashed line is the actual growth rate; the solid line is the model-predicted growth rate based on BMA.

Figure 11: Posterior Density Plots for Female Students in United States

4.3 Sensitivity Analysis

The results of the sensitivity analysis are summarized in Table 5. The upper panel contains male student data, and the lower panel contains female student data. Note that UIP, RIC, BRIC, and HQ are fixed priors, and EBL and Hyper-*g* are flexible priors for parameters.

The male data showed the smallest LPS value when the model prior was specified as a binomial model with $m = 4$ when the parameter was set as a fixed prior (regardless of type). Female students' data showed the same results: the smallest LPS value appeared when the binomial model with $m = 4$ was used as the model prior in all fixed prior types. However, these LPS values were much smaller in comparison to male students. Parallel findings were also noted when the flexible prior was set to parameters. For both the male

and female data groups, regardless of the parameter prior, the model prior had the smallest LPS value when specified as a binomial model with $m = 4$.

When comparing the fixed prior and flexible prior for parameters, the LPS value was relatively small when the fixed prior for parameters was set regardless of the model prior. Additionally, in terms of gender, the LPS values for female students were much smaller.

This study included 6 predictors and 16 sample sizes, leading the g-value of BRIC (i.e., $g = max(N, Q^2)$) and the g-value of RIC (i.e., $g = Q^2$) to be the same. Thus, the LPS results of RIC and BRIC, which are fixed priors, were the same irrespective of the model prior. In the case of a flexible prior, hyper = RIC was therefore excluded.

	UIP	RIC	BRIC	HQ	EBL	HG2.1	HG4	HG-UIP	HG-BRIC
Male									
Uniform	0.907	0.923	0.923	0.909	0.960	1.006	1.076	1.005	1.011
Bin.(m = 2)	1.002	1.033	1.033	1.009	1.058	1.115	1.147	1.108	1.140
Bin.(m = 4)	0.846	0.849	0.849	0.844	0.911	0.945	1.023	0.946	0.944
Bin.-Beta	0.905	0.980	0.980	0.924	0.951	1.044	1.053	1.033	1.083
Female									
Uniform	0.679	0.673	0.673	0.670	0.742	0.795	0.901	0.796	0.797
Bin.(m = 2)	0.826	0.849	0.849	0.826	0.888	0.962	1.030	0.957	0.983
Bin.(m = 4)	0.596	0.574	0.574	0.583	0.658	0.699	0.804	0.700	0.697
Bin.-Beta	0.587	0.587	0.587	0.579	0.653	0.723	0.802	0.718	0.744

Note. UIP = unit information prior; RIC = risk inflation criterion; BRIC = benchmark risk inflation criterion; HQ = Hanna-Quinn criterion; EBL = empirical Bayes local; HG2.1 = hyper-g prior with α = 2.1; HG4 = hyper-g prior with α = 4; HG-UIP = hyper-g prior with UIP setting; HG-BRIC = hyper-g prior with BRIC setting; Uniform = uniform model prior; Bin. (m = 2) = binomial model prior with model size = 2; Bin. (m = 4) = binomial model prior with model size = 4; Bin.-Beta = beta-binomial model prior.

Table 5: Summary of LPS for Parameter and Model Prior Settings

Chapter 5

Conclusions

In this study, the Bayesian probabilistic prediction distribution with BMA was investigated to address model uncertainty. BMA was applied to the TIMSS, an international large-scale assessment, to predict growth rates in 8th-grade students' mathematics achievement. Results highlighted science achievement score and teaching years as the most important predictors of both male and female students' growth in mathematics achievement. Several limitations and implications emerged from these findings.

First, during preliminary analysis, different models were found to be best suited to the male and female datasets. The model in this study which used fixed time points and constrained residual variances was optimal for male student data, whereas female student data were best managed with the latent basis method and no constraints. Nevertheless, because these datasets were compared based on the same model, more accurate conclusions

remain to be drawn for female students. An optimal model for female student data will capture these students' mathematics achievement growth more precisely.

Second, the latent basis method generally outperforms approaches where the time points have fixed values (Kaplan & Huang, 2021). However, the fixed method was adopted in this study for a specific reason: when the latent basis method was applied to determine the optimal growth curve model in preliminary analysis, the time points displayed negative values at the inappropriate time point. For example, out of 5 time points, when the former time points were given as fixed values 0, 1, and 2, negative values such as -4 and -7 were chosen for the latter time point based on the data. Unlike the positive value given in the former, if a negative value is set in the latter, then the time-point interpretation does not capture the meaning of the data. The growth curve model was thus adopted in this study with the fixed method to ensure the simplicity and accuracy of analysis.

Third, the correlations among model predictors were not checked during preliminary analysis. BMA includes numerous variables to find the optimal model, and this approach is highly sensitive to collinearity issues. Draper (1999) described the effects of such issues with the following example. Suppose there are predictors x_1, x_2, and x_3, where x_2 and x_3 are collinear. When x_1 and x_2 are included in model space M_1, 4 models can be created; that is, $M_1 = \{$ no predictors, $x_1, x_2, (x_1, x_2)\}$. Using the indifference prior on the model space, each model can be given $1/4$ weight. Assume that the model space M_2 contains x_1, x_2, and x_3. Eight models are created in this case—$M_2 = \{$no predictors, $x_1, x_2, x_3, (x_1, x_2), (x_1, x_3), (x_2, x_3), (x_1, x_2, x_3)\}$—and each is given a weight of $1/8$. In this instance, the estimation of the last two models fails due to collinearity. Apart from these two models, the M_2 space can also be updated with six models, $M_2{}' =$

{no predictors, $x_1, x_2, x_3, (x_1, x_2), (x_1, x_3)$}, and weighted by $1/6$ using an indifferent prior across them. Because the $M_2{}'$ space is essentially identical to the M_1 space in terms of predictions, the weight of the M_1 space is given as $\{1/6, 1/6, 2/6, 2/6\}$. The weights vary even for different versions of the same model when a collinear predictor is included. Therefore, to strengthen BMA results, collinearity diagnostics (Belsley et al., 2005) should be assessed first.

Fourth, the BMS package provides PMP values for each model. In this study, the top 5 models were reported in the order of the largest PMP value (see Table 4). This result requires additional explanation. Despite being the PMP of the top model, its value for Model 1 was too small at 0.21. The model with the largest PMP value is often chosen. Yet if the largest PMP value is too small (i.e., if the model uncertainty is still large), then it is risky to draw inferences based on that model. Draper (1987) stated that inferences based on models with small PMP values tend to be overconfident. Making decisions based on these inferences hence carries greater risk than expected. In addition, the sum of the PMP values for the top five models was far less than 1. This outcome may have arisen because the number of predictors in this study was limited. This finding also implies uncertainty about the model space. Therefore, if the PMP value is not large enough when choosing a model via BMA, then the selected model will contain substantial uncertainty and will need to be updated.

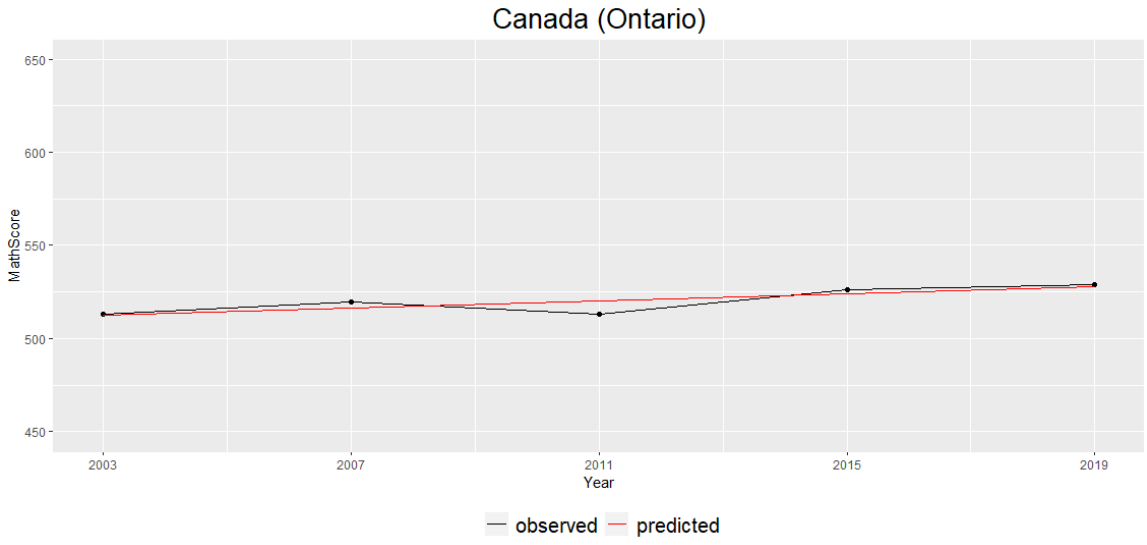
Fifth, Bayesian predictive densities for growth were checked for each country to verify how well the growth rate predictions and the actual data aligned. As mentioned in Ch. 4 (Results), in most countries, the difference between the model-predicted growth rate and the actual growth rate did not exceed $SD = \pm 1$. However, the difference between them

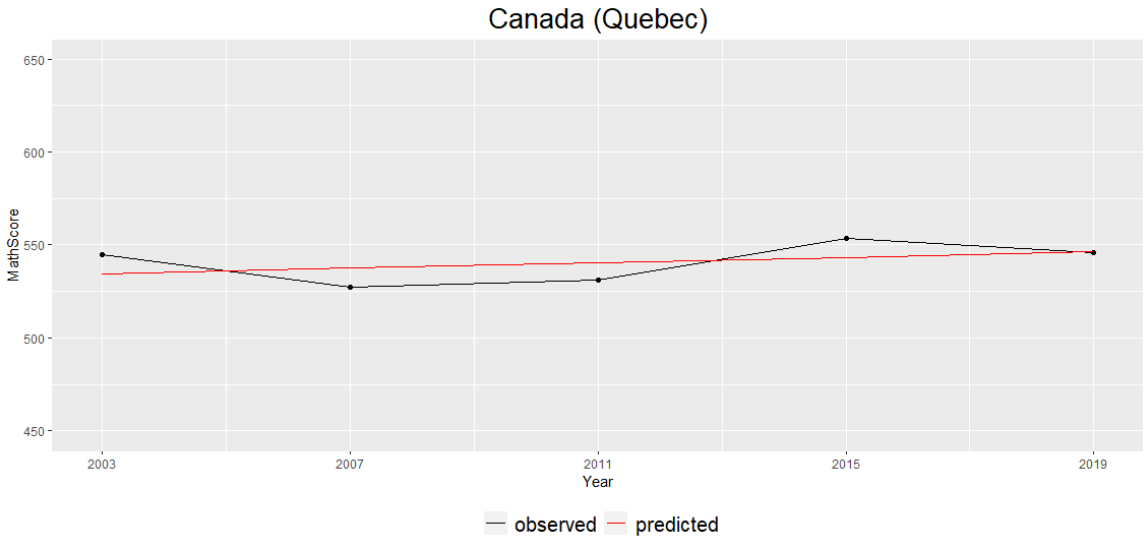
was quite large in the data on Israeli male students. The prediction model was thus incorrect for this dataset. The data on Israeli male students must be scrutinized accordingly. Compared with other countries, some characteristics may only be explained by Israeli male student data. The outlier's meaning must be determined as well. All Bayesian predictive densities for male and female students in all countries are presented in the Appendix A.3 – A.4. Moreover, the predictive densities for growth where the model-predicted growth rate and actual growth rate were consistent only found in a few countries. A different prior setting should be specified in place of the default prior (used in this study) to further reduce the discrepancy between these two growth rates.

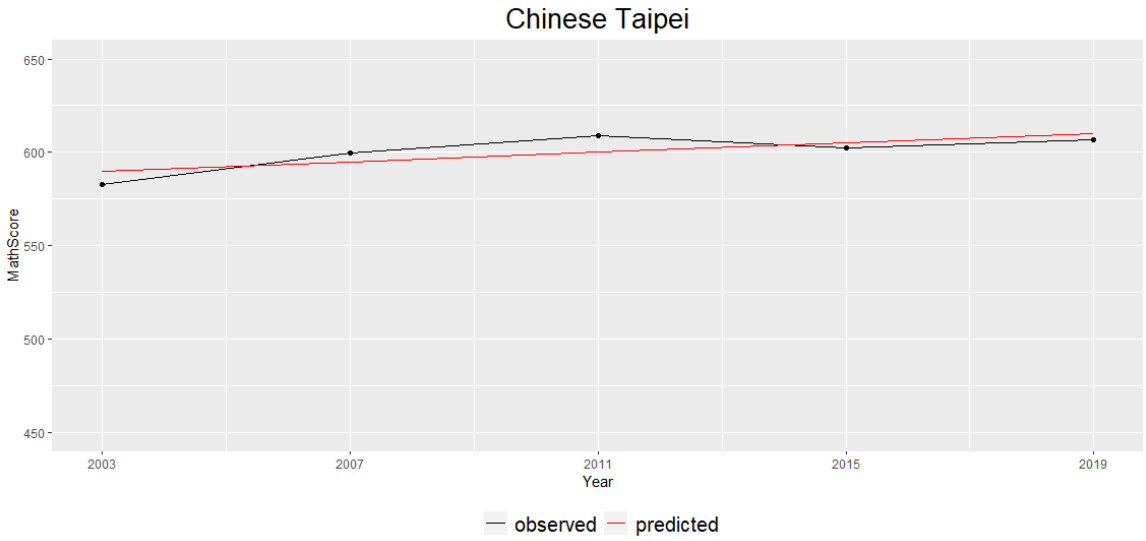
Appendix A

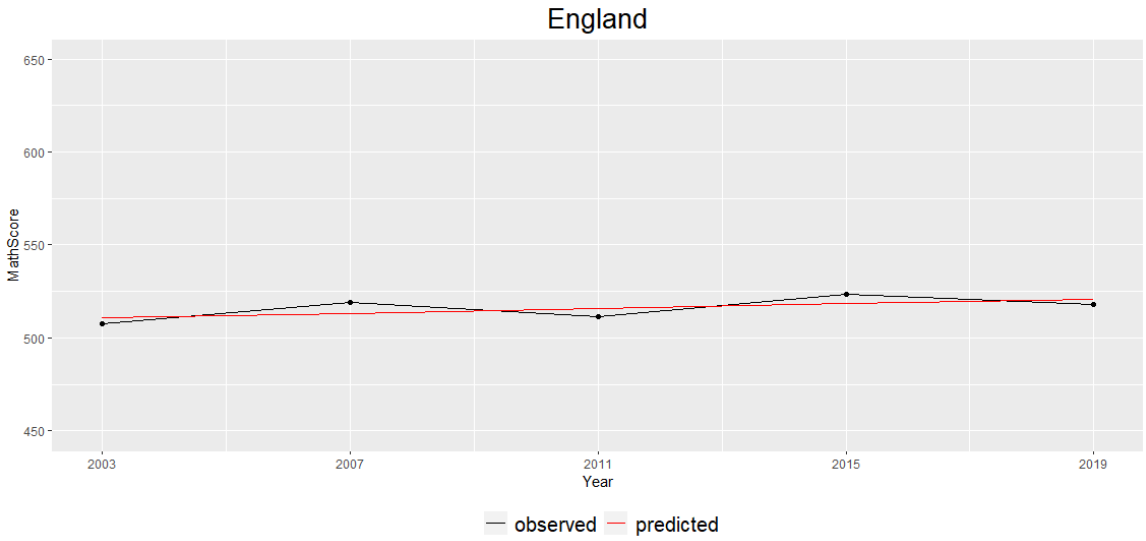
Appendix A.1

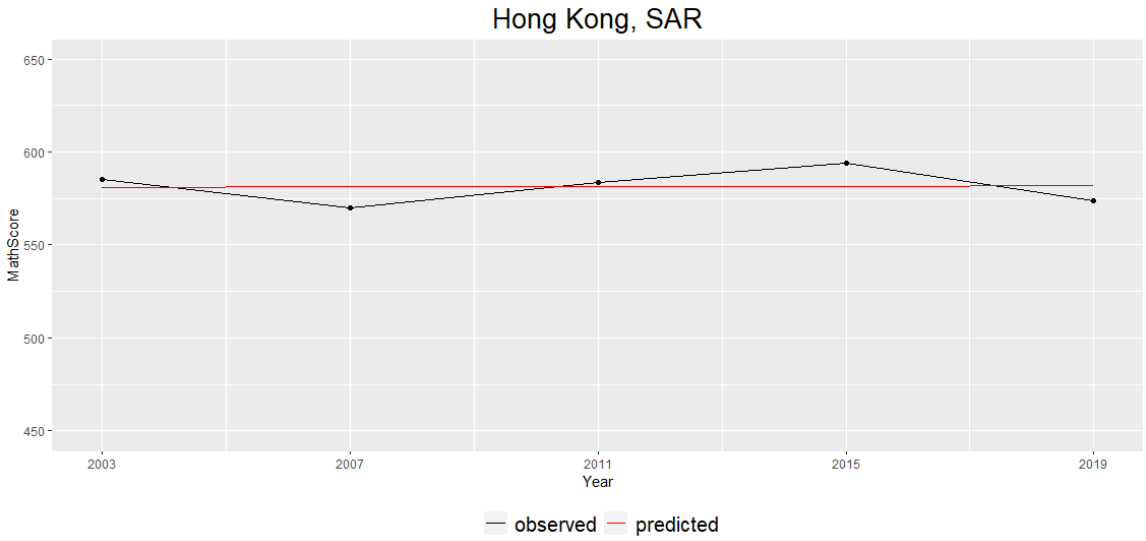
The following provides fitted and actual trend plots of the male students' data for each country.

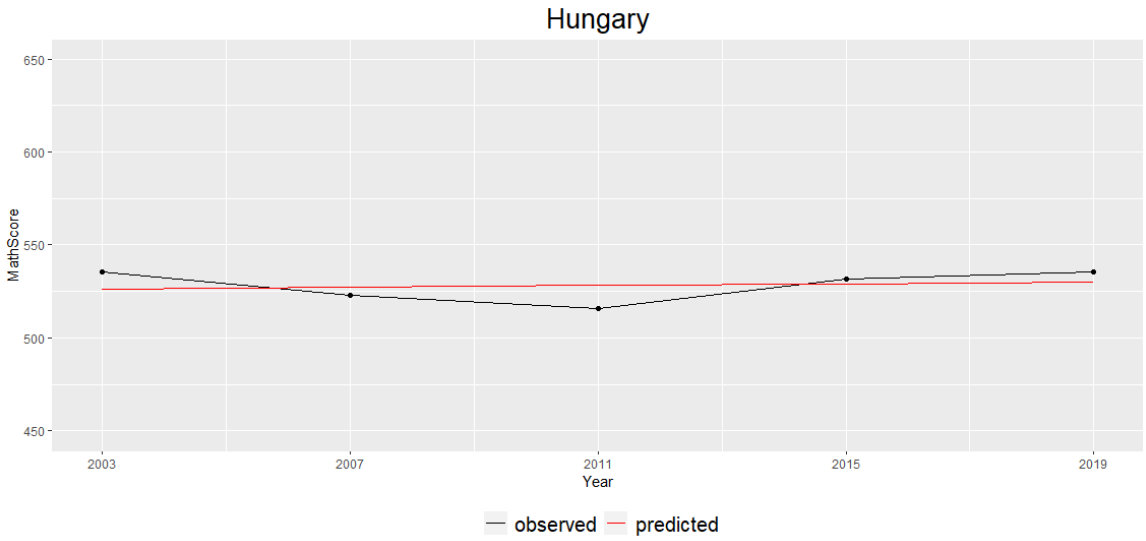


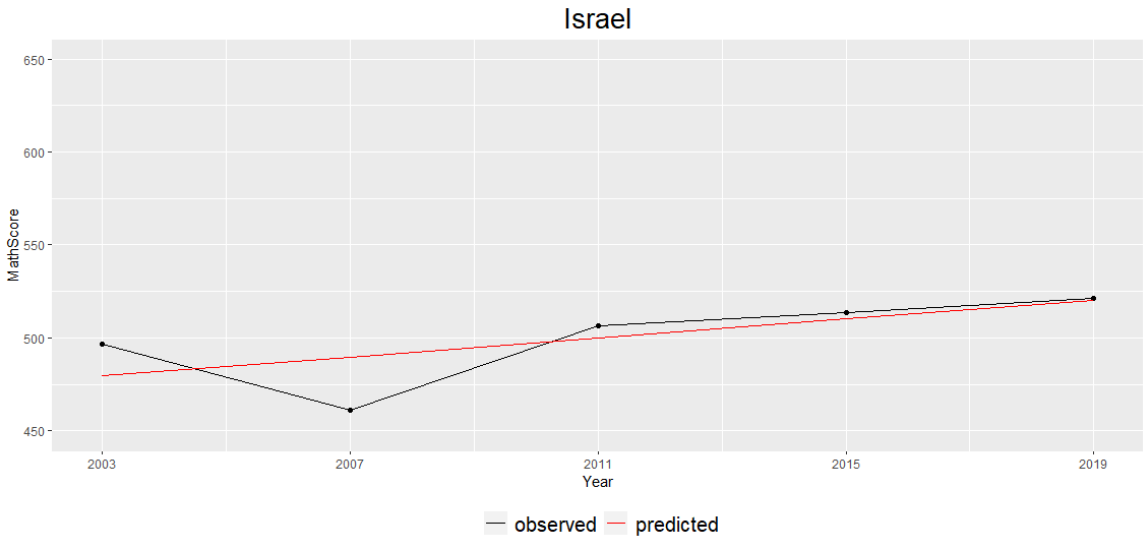

Canada (Quebec)
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

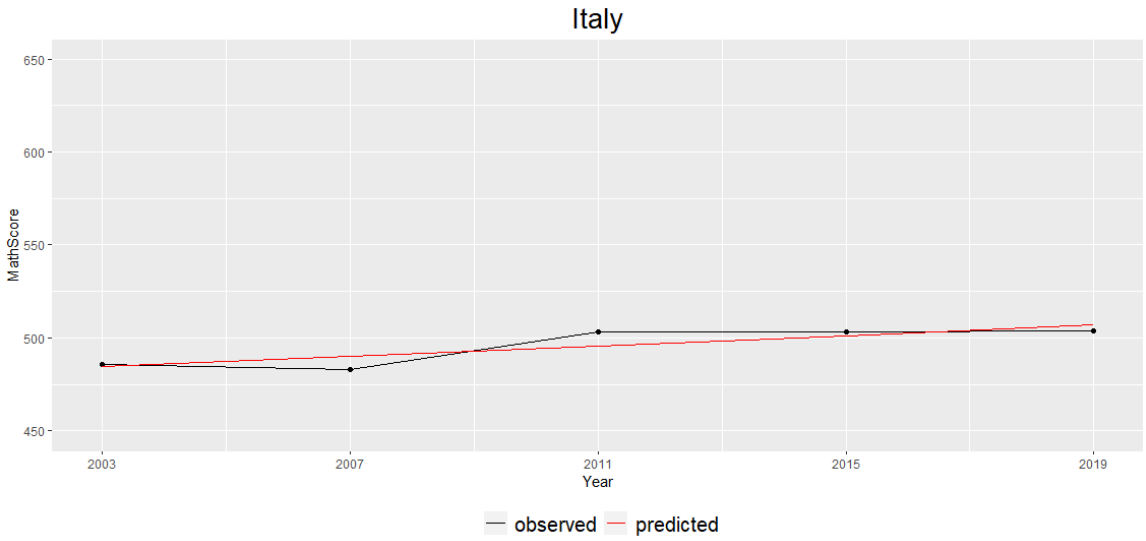

Chinese Taipei
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

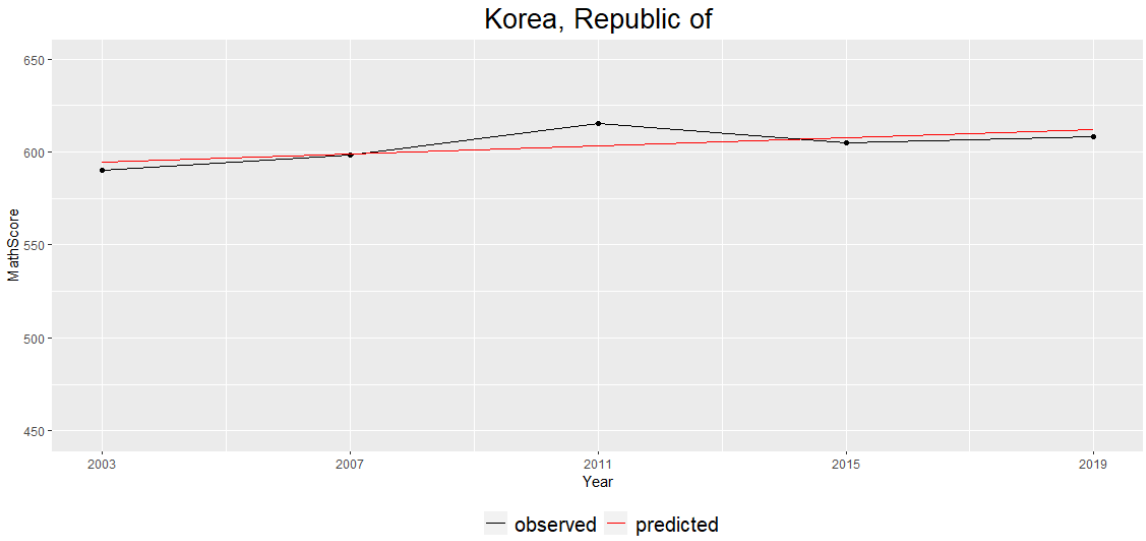

England
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

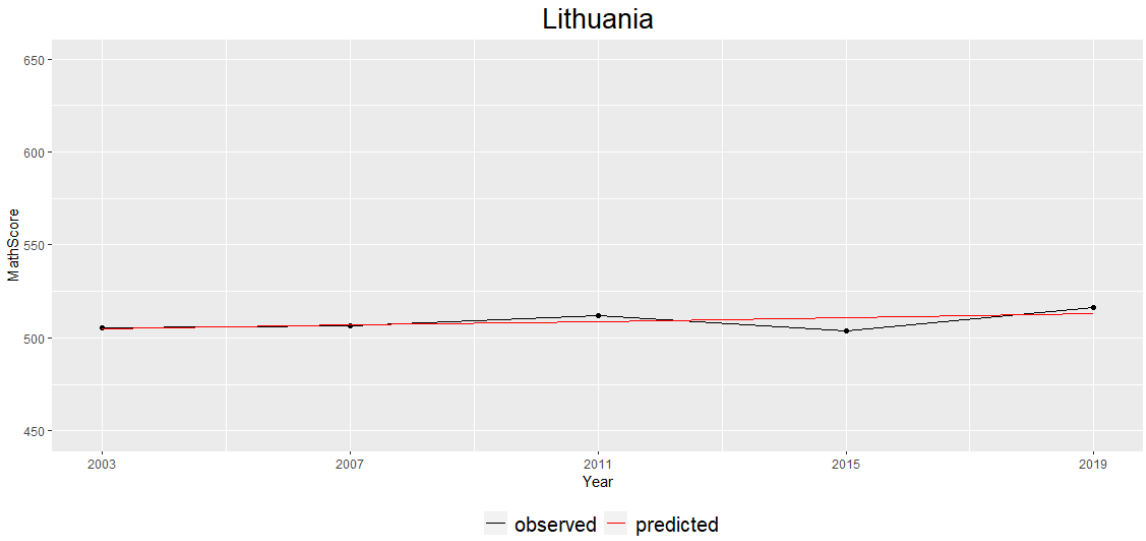

Hong Kong, SAR
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

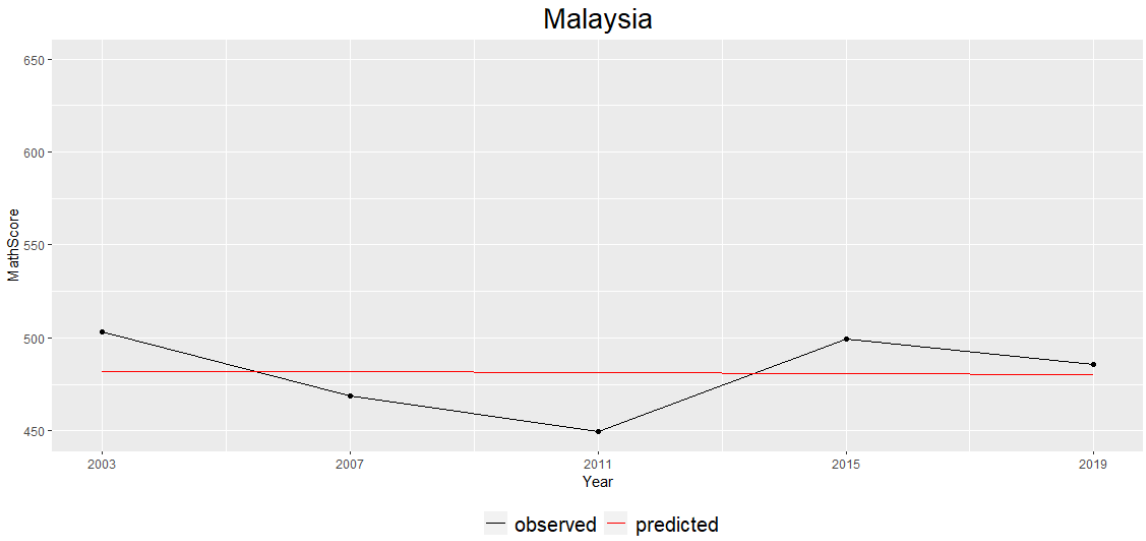

Hungary
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

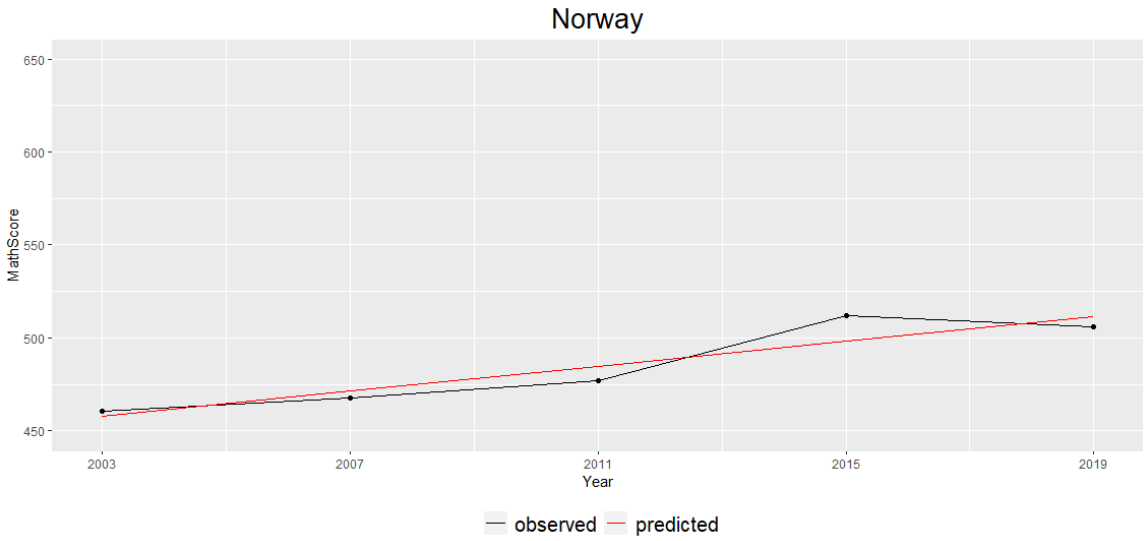

Israel
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

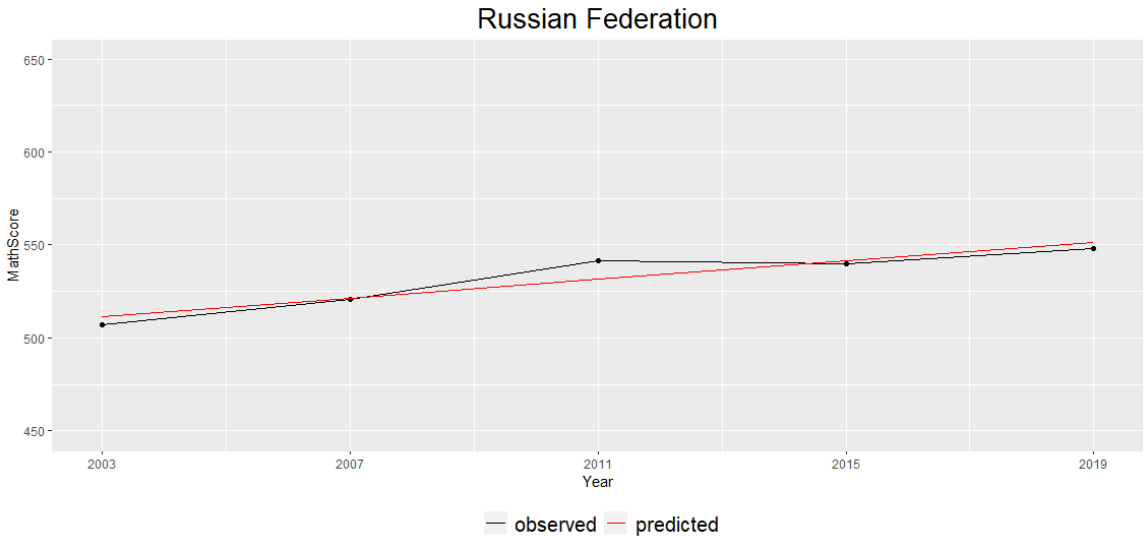

Italy
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

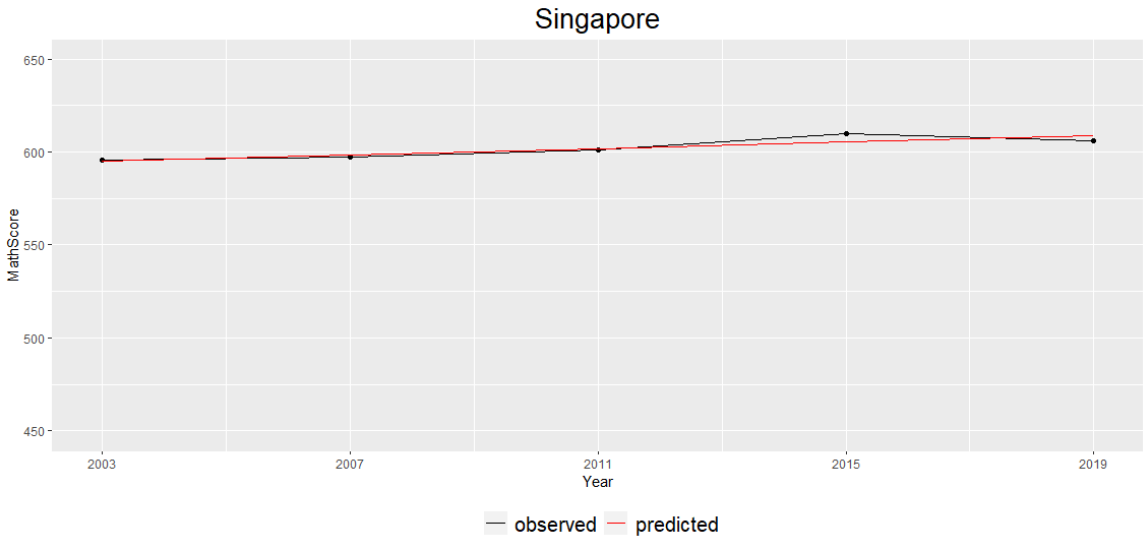

Korea, Republic of
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

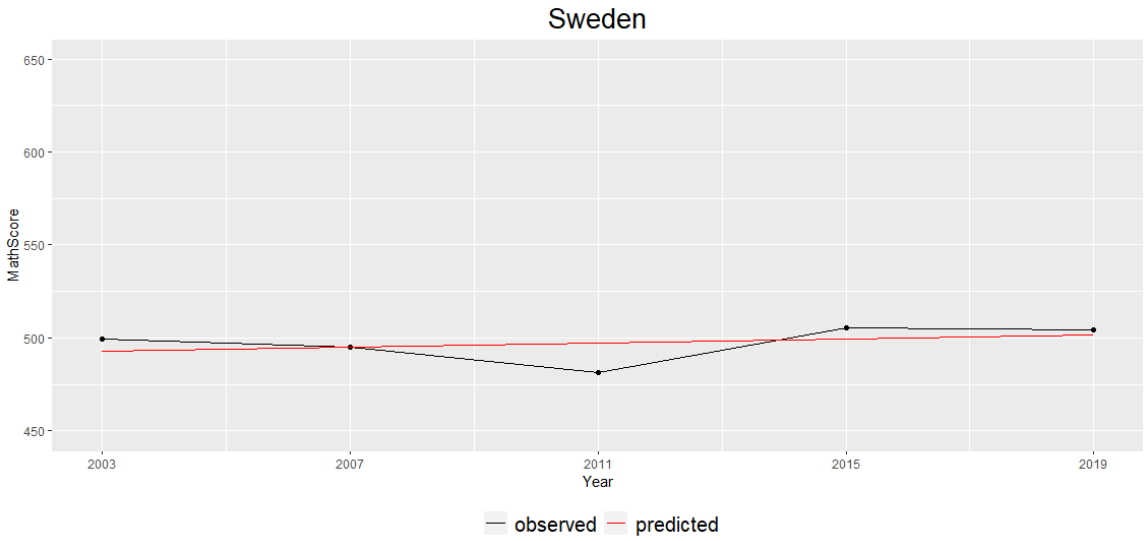

Lithuania
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

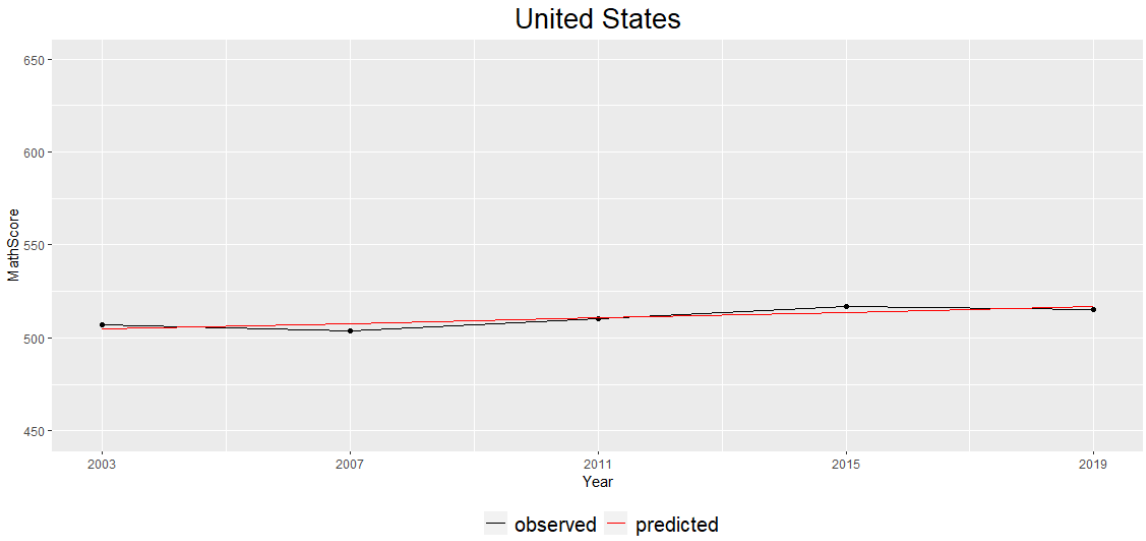

Malaysia
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted


Norway
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted


Russian Federation
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

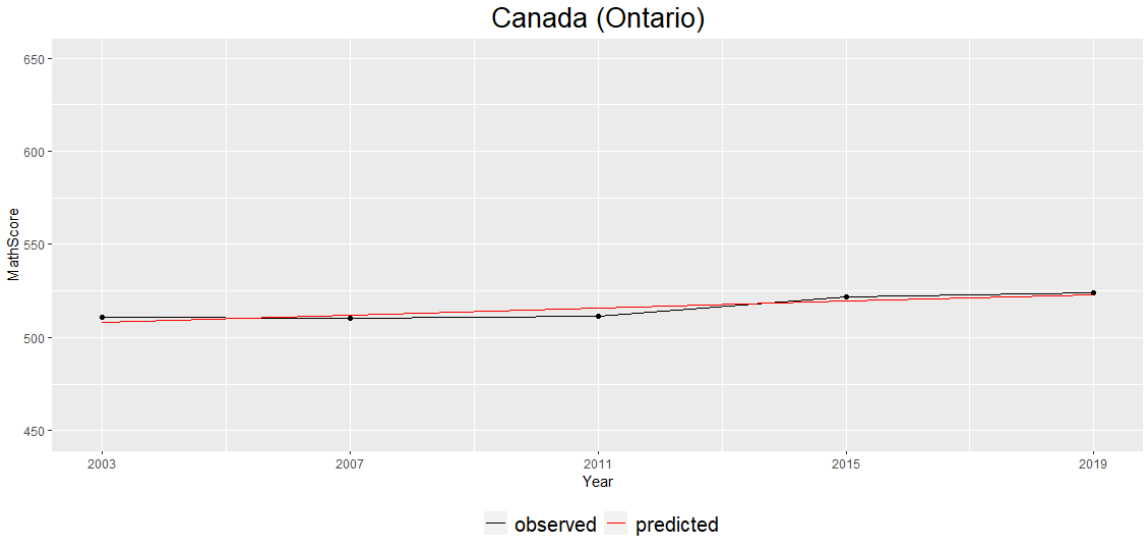

Singapore
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

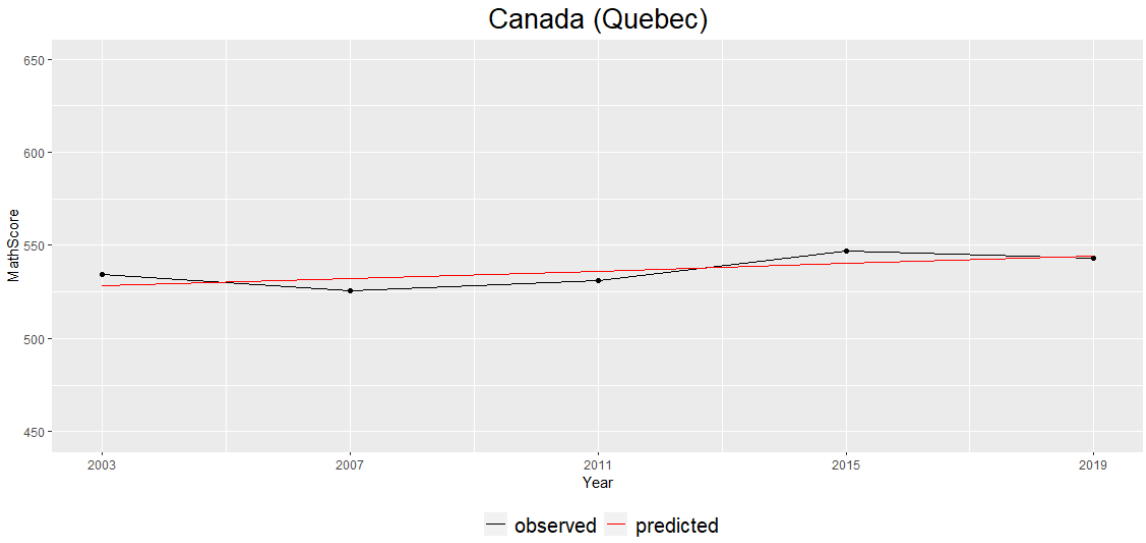

Sweden
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

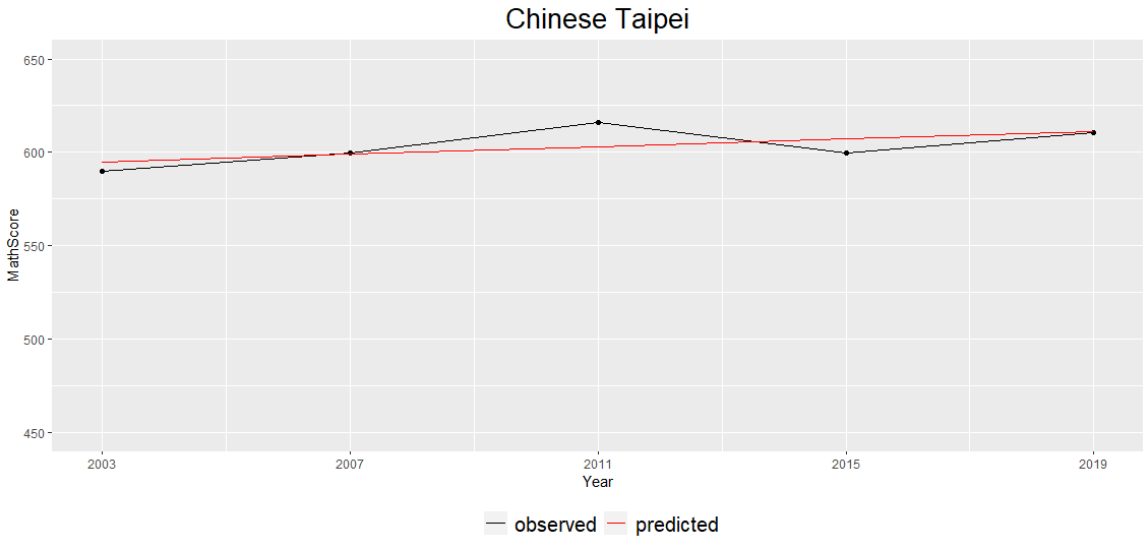

United States
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

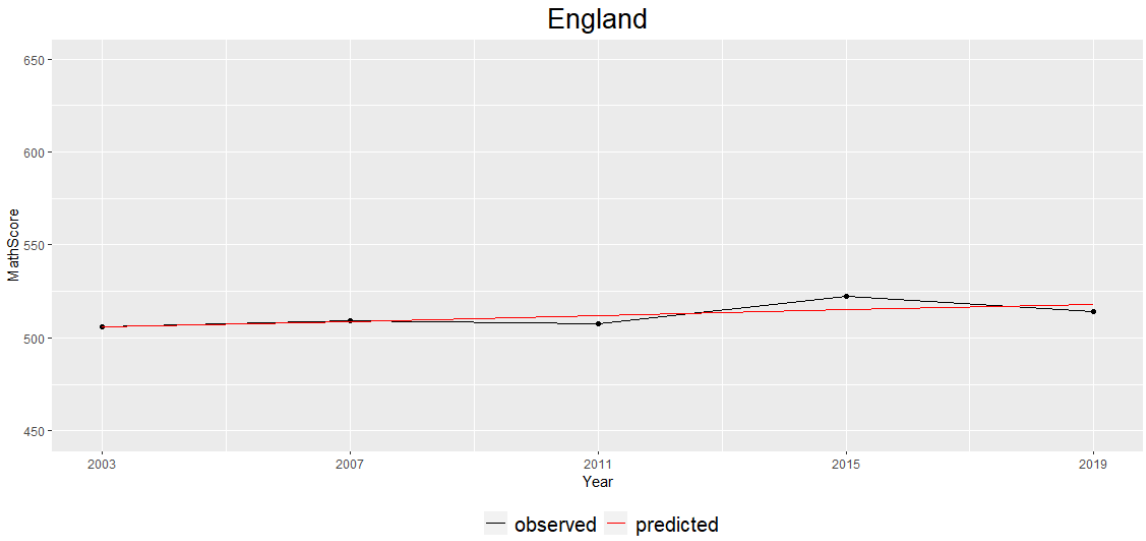
Appendix A.2

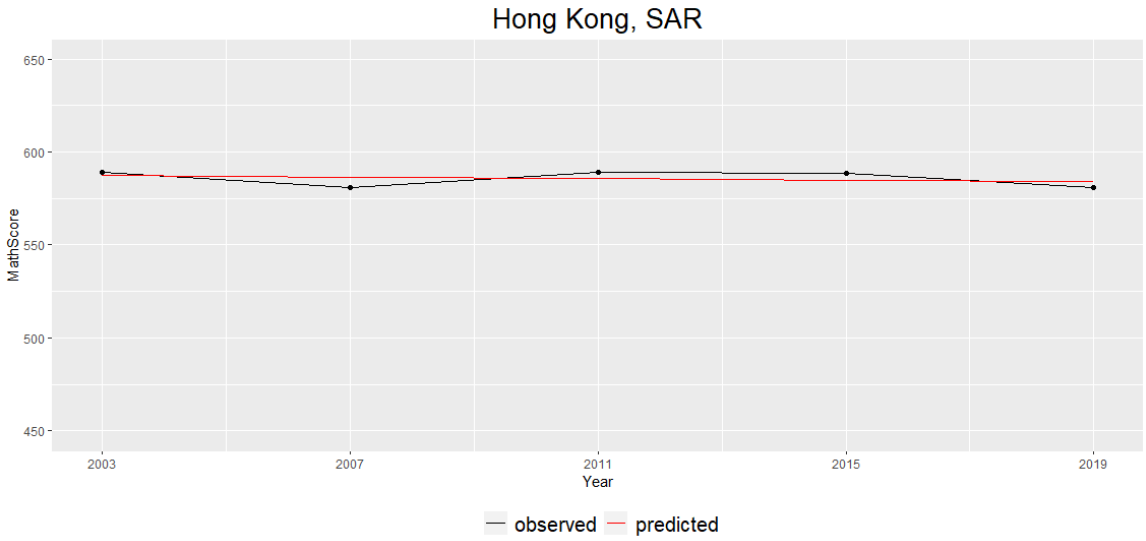
The following provides fitted and actual trend plots of the female students' data for each country.

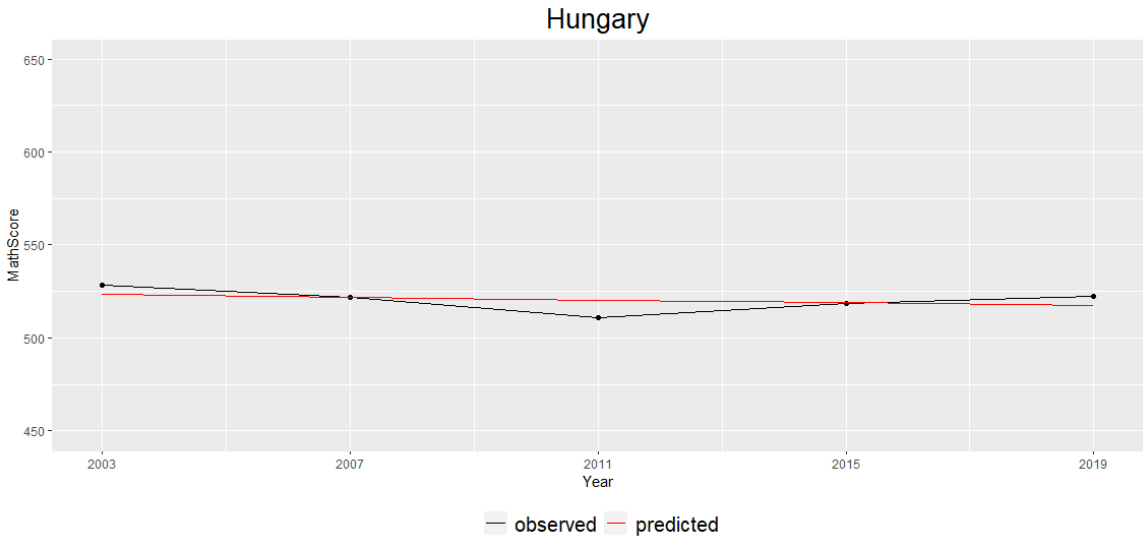


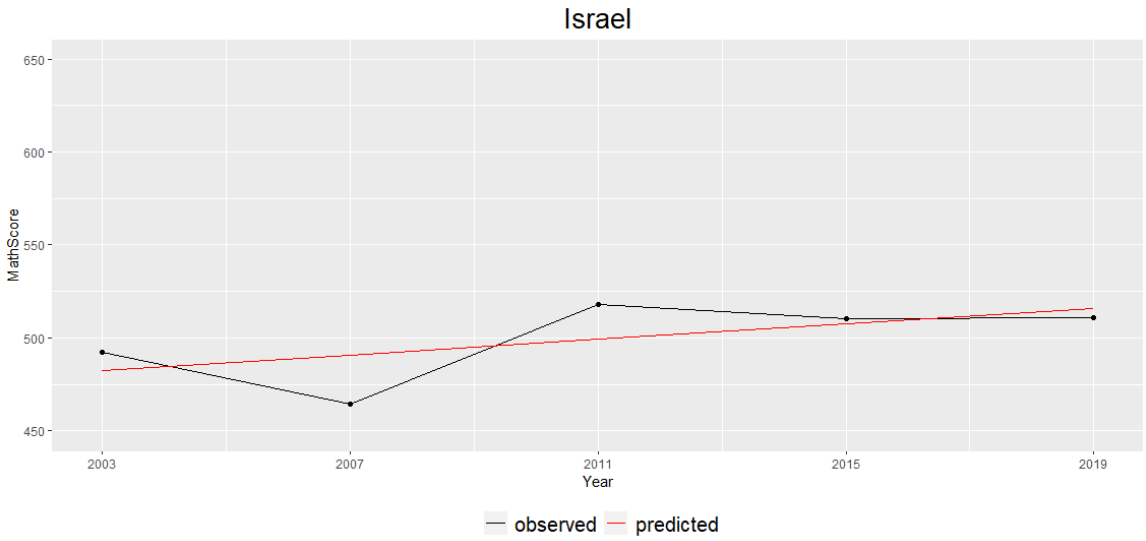


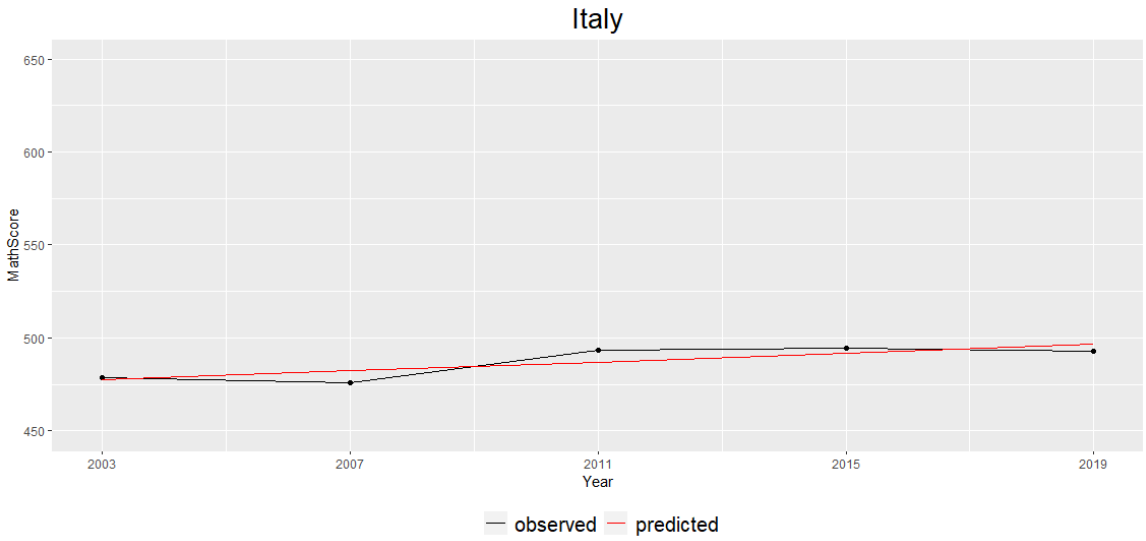

Chinese Taipei
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

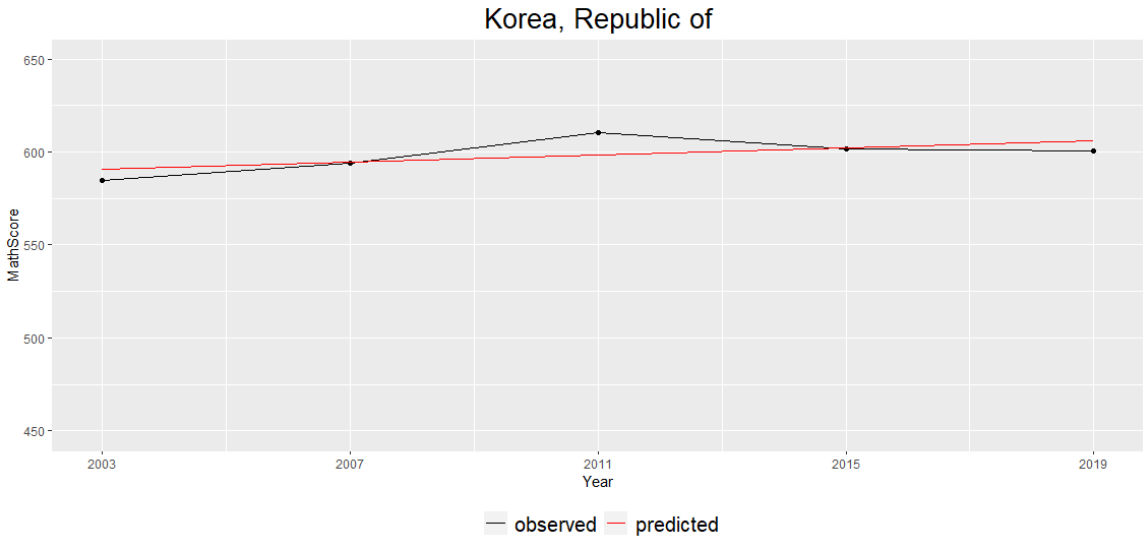

England
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

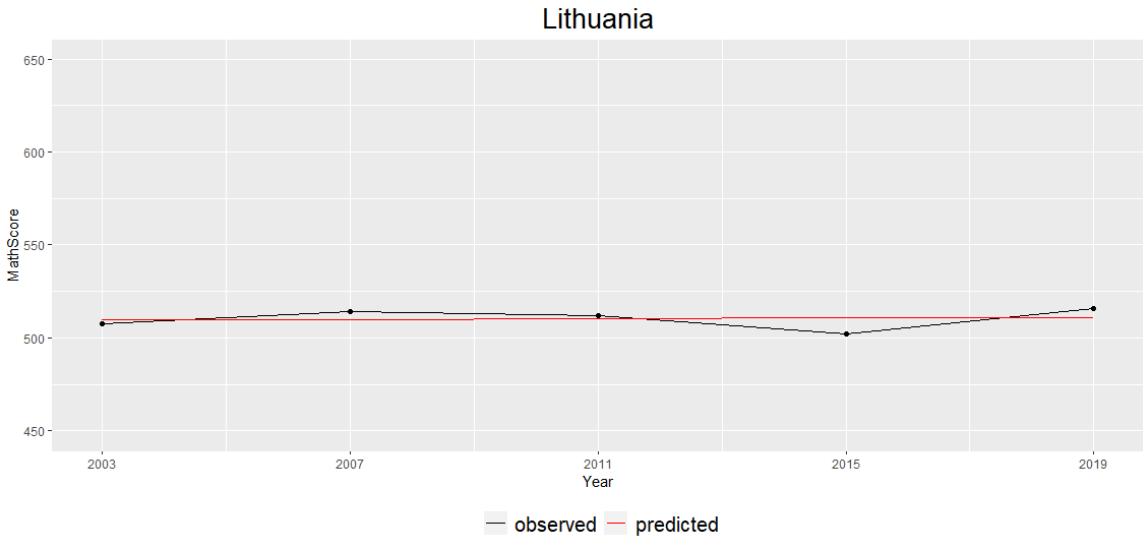

Hong Kong, SAR
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

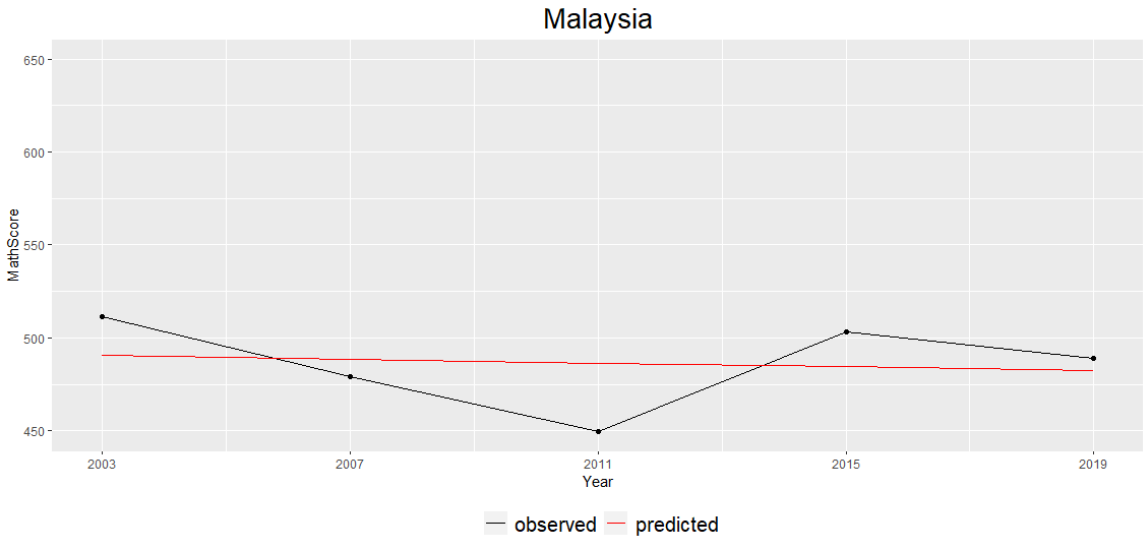

Hungary
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

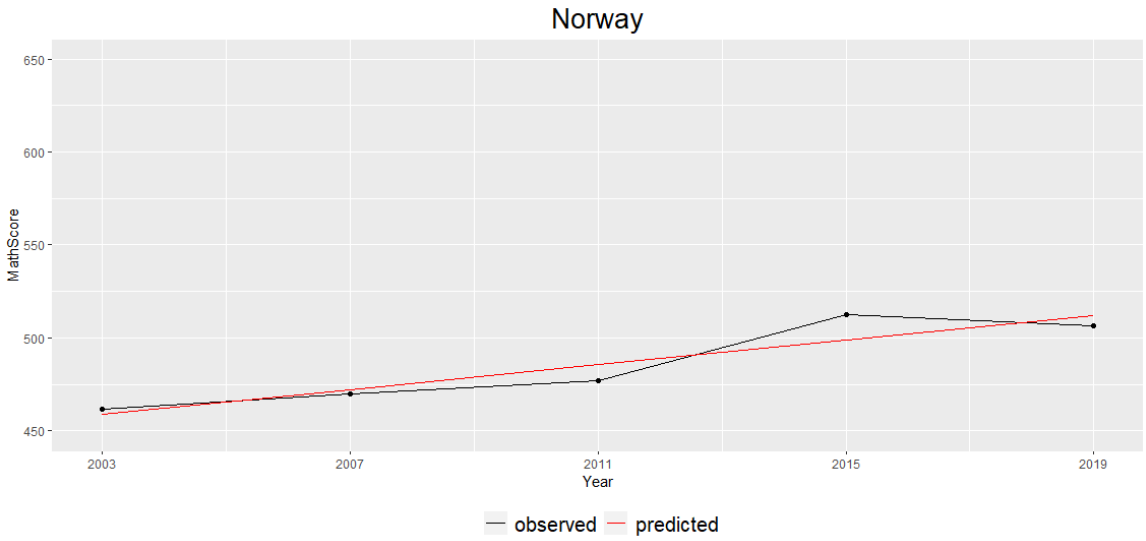

Israel
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

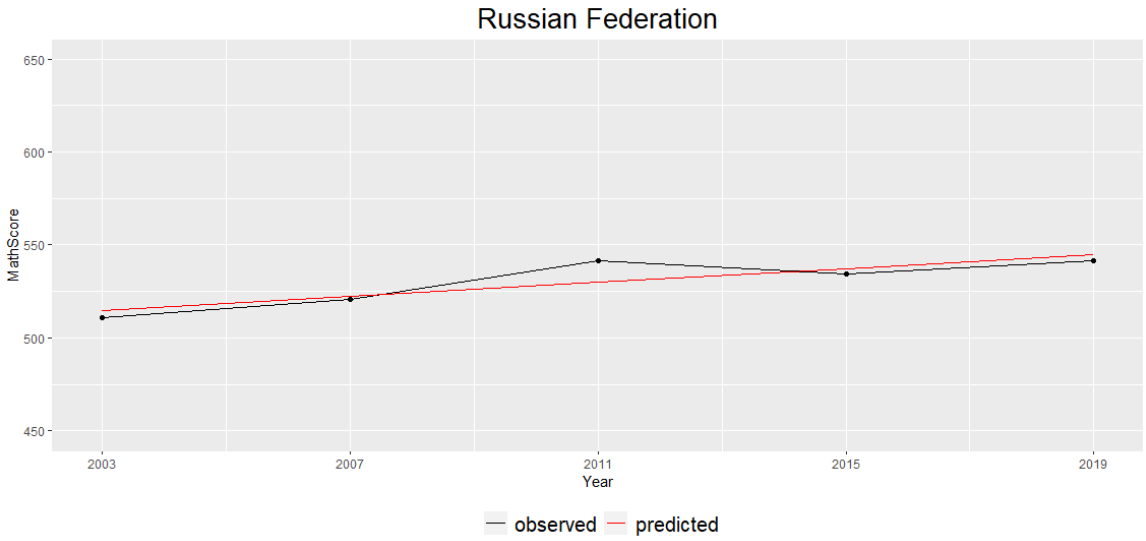

Italy
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

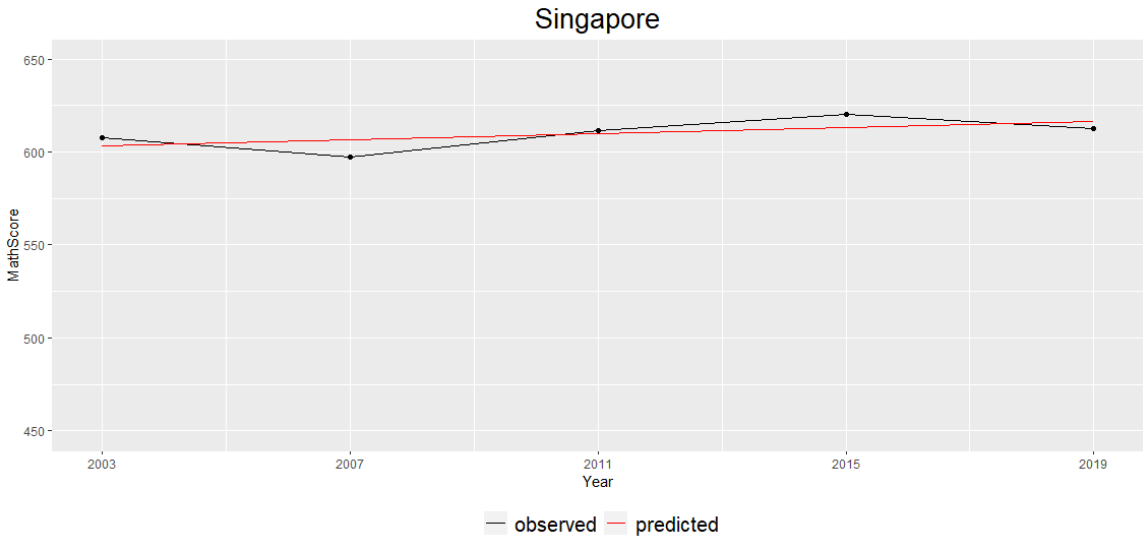

Korea, Republic of
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

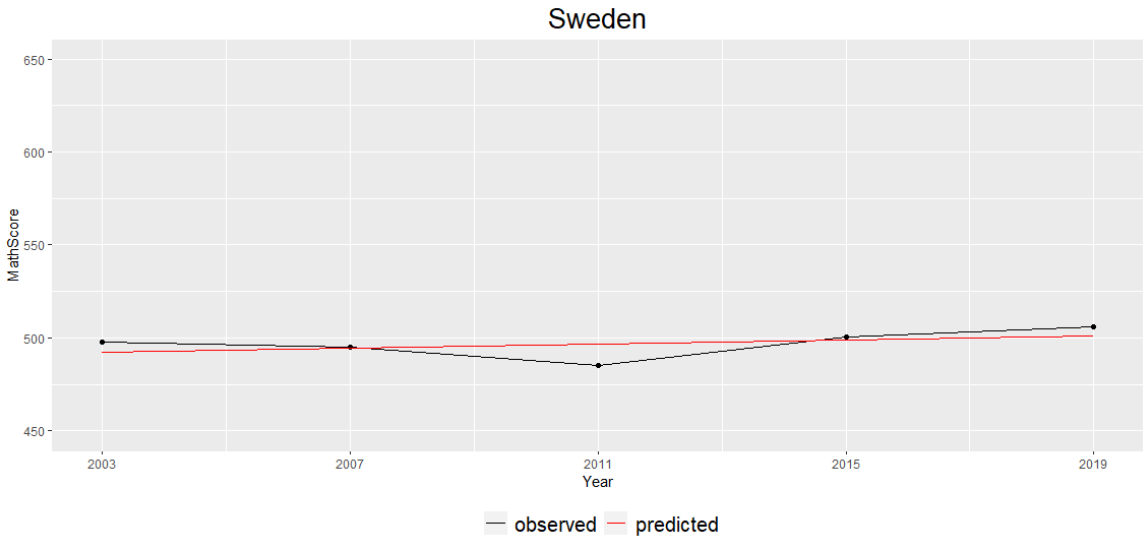

Lithuania
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

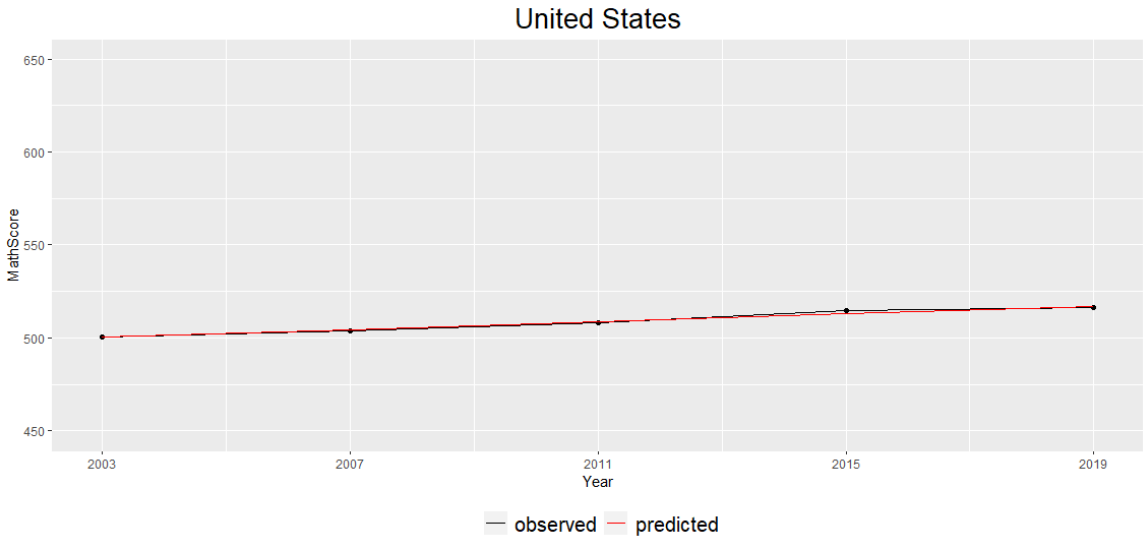

Malaysia
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted


Norway
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted


Russian Federation
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted


Singapore
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

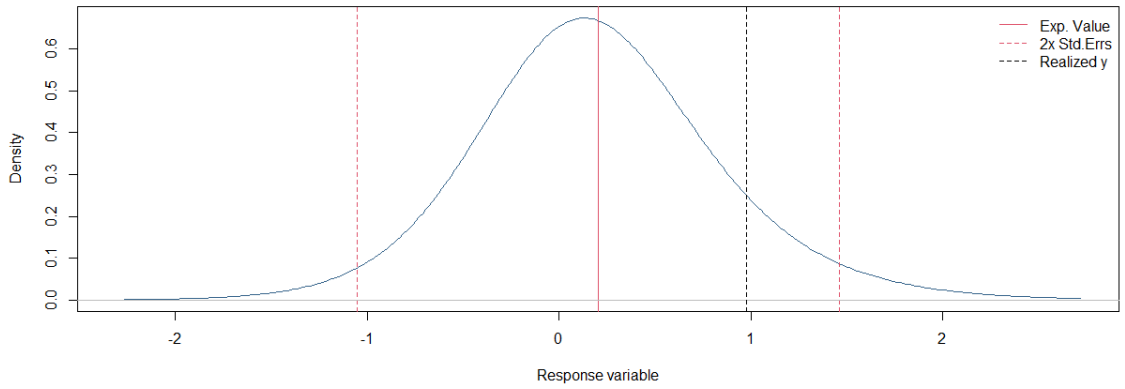

Sweden
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted


United States
650
600
550
500
450
MathScore
2003
2007
2011
2015
2019
Year
observed
predicted

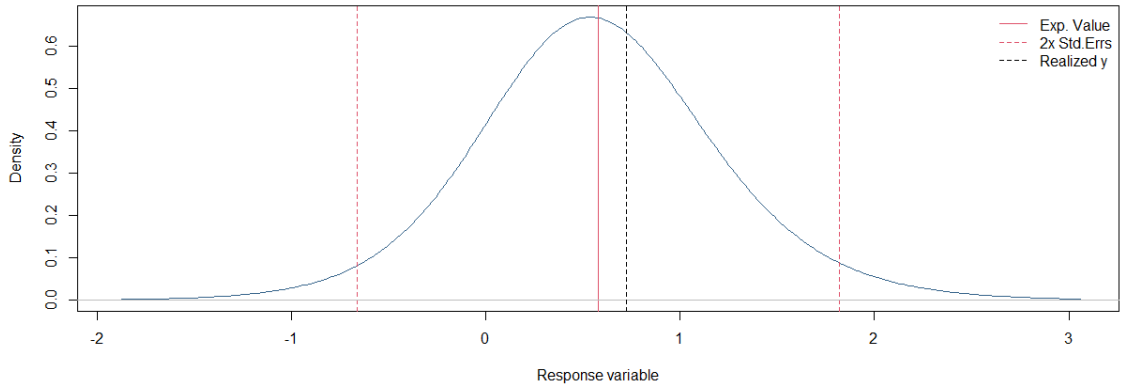
Appendix A.3

The following provides posterior density plots for male students in each country.

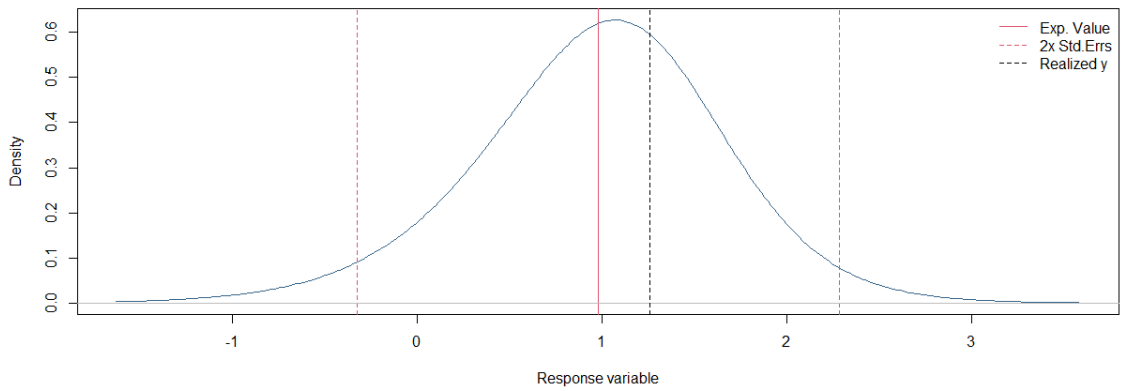
Canada (Ontario)



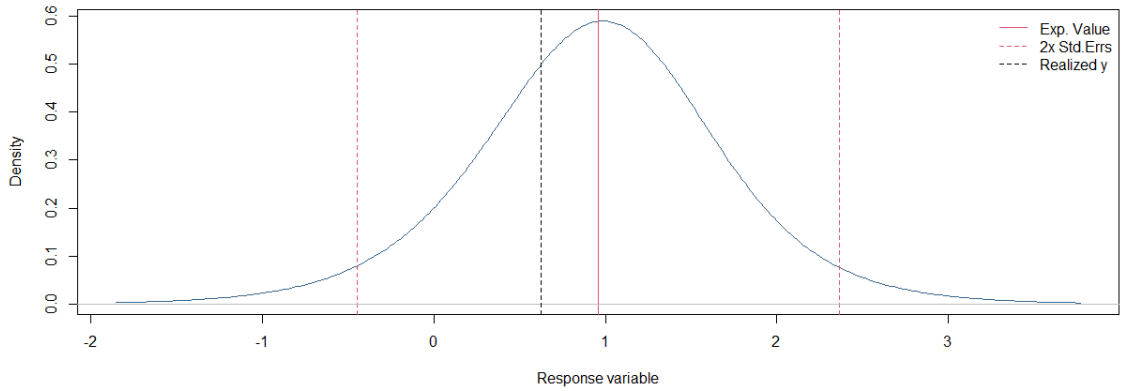
Canada (Quebec)



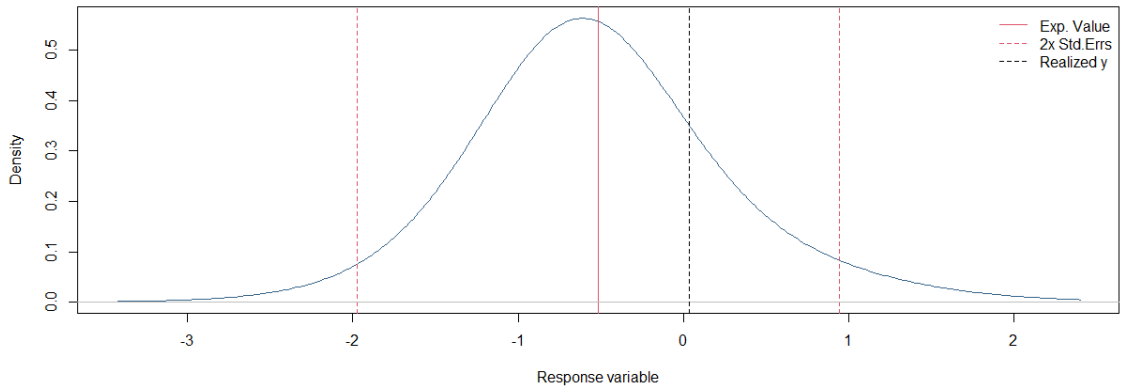
Chinese Taipei



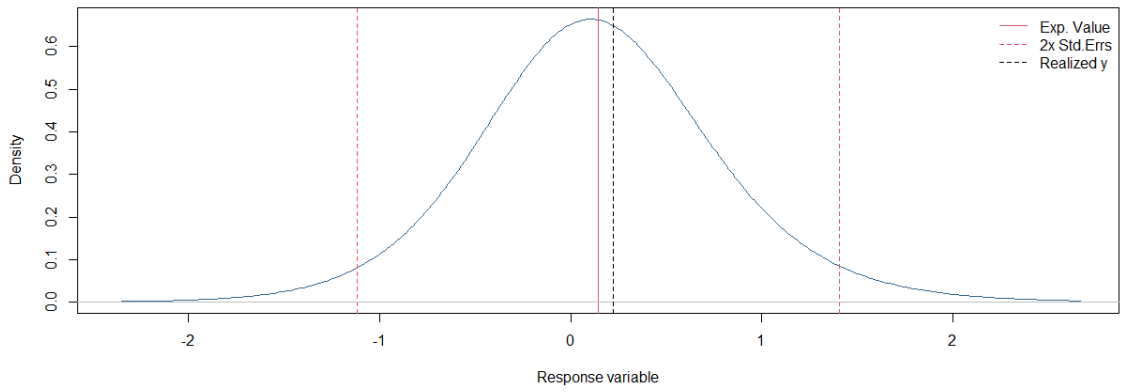
England



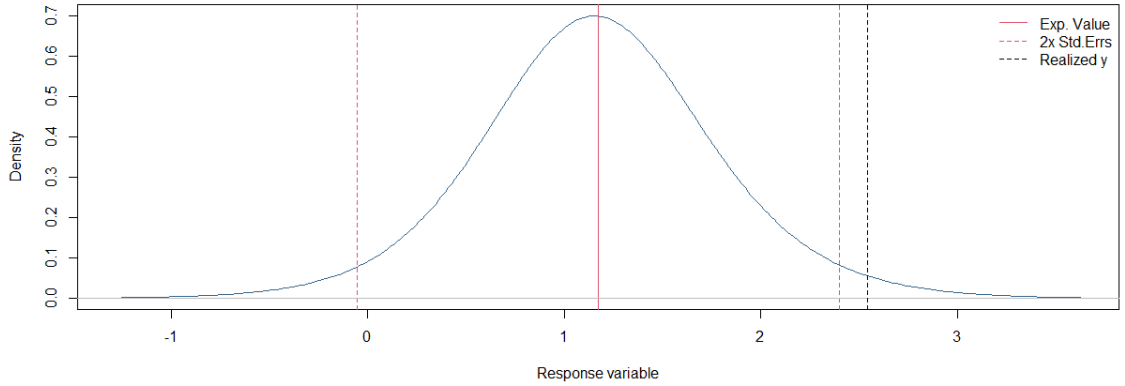
Hong Kong, SAR



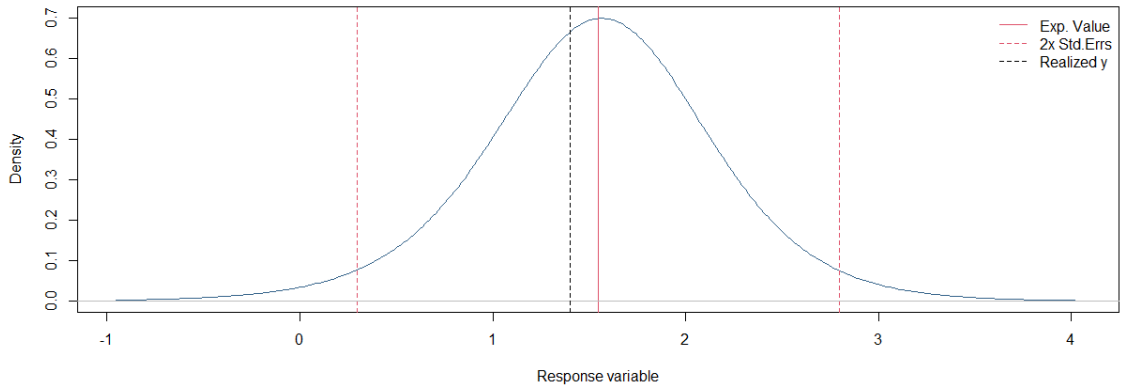
Hungary



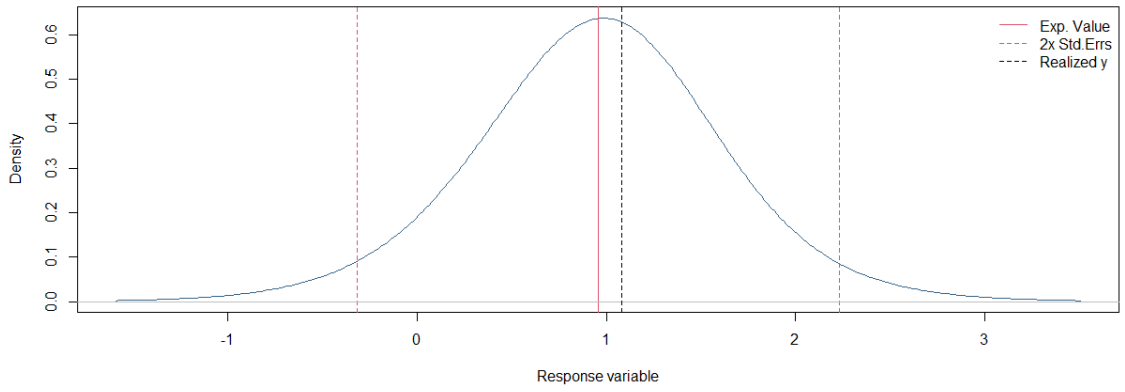
Israel



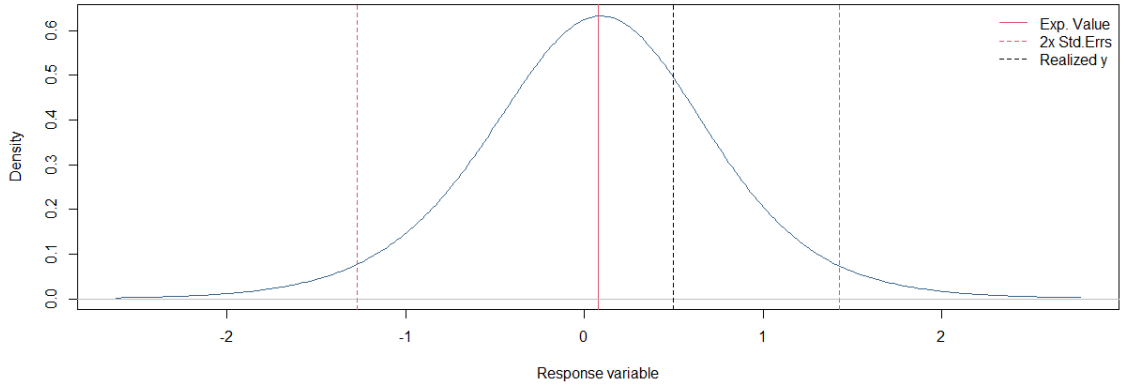
Italy



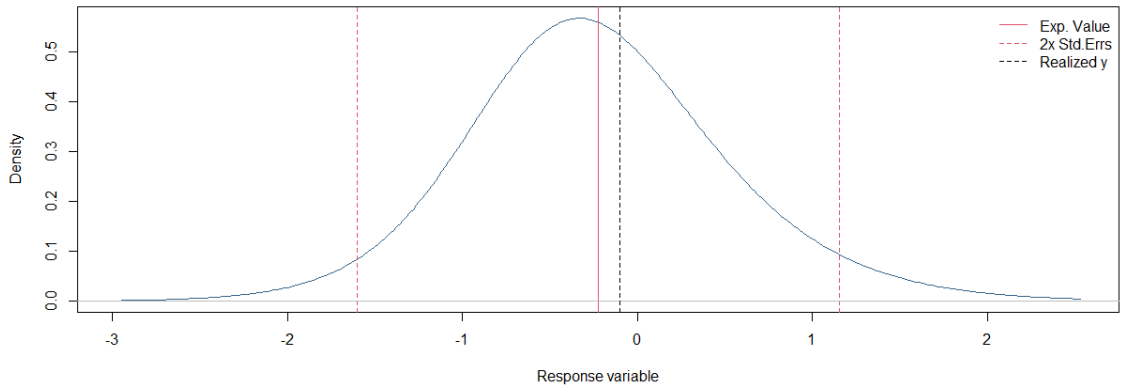
Korea, Republic of



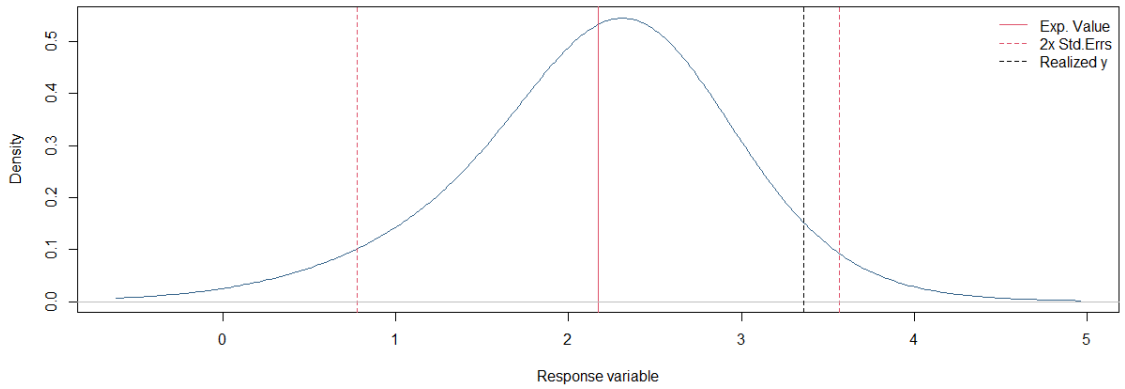
Lithuania



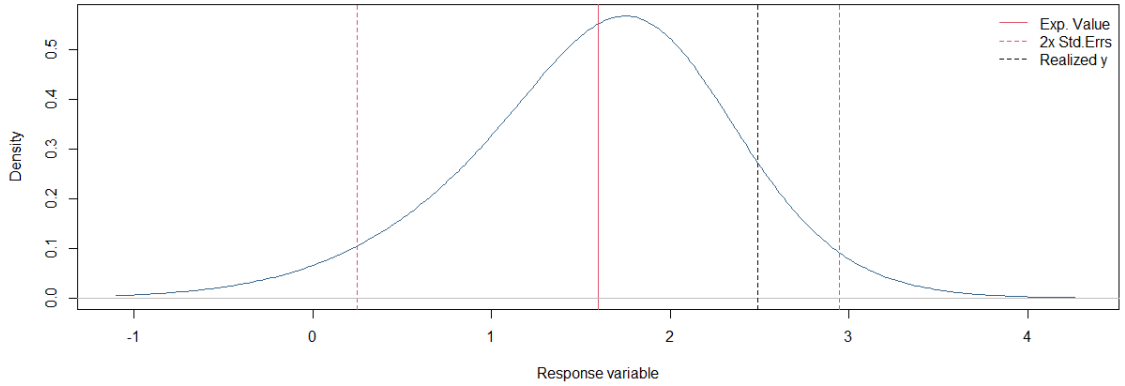
Malaysia



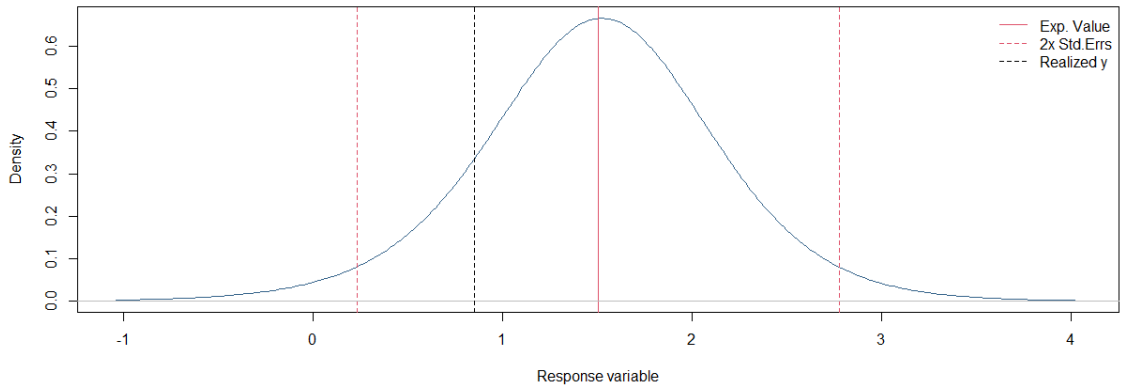
Norway



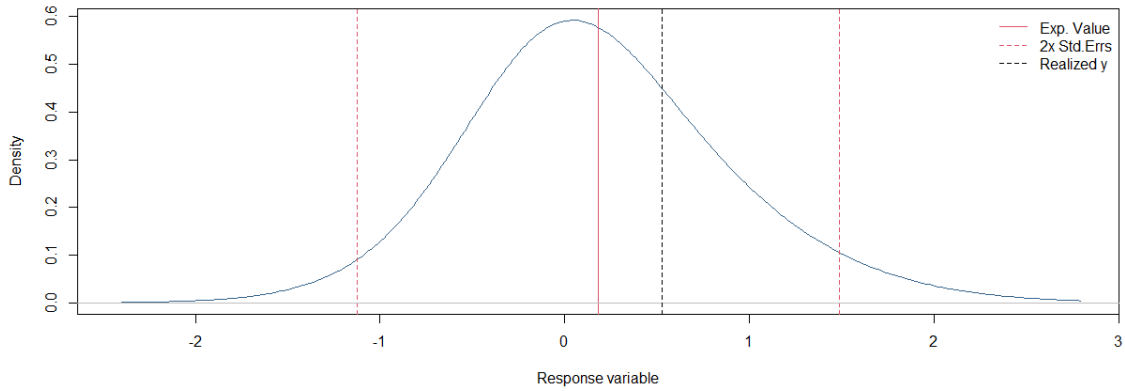
Russian Federation



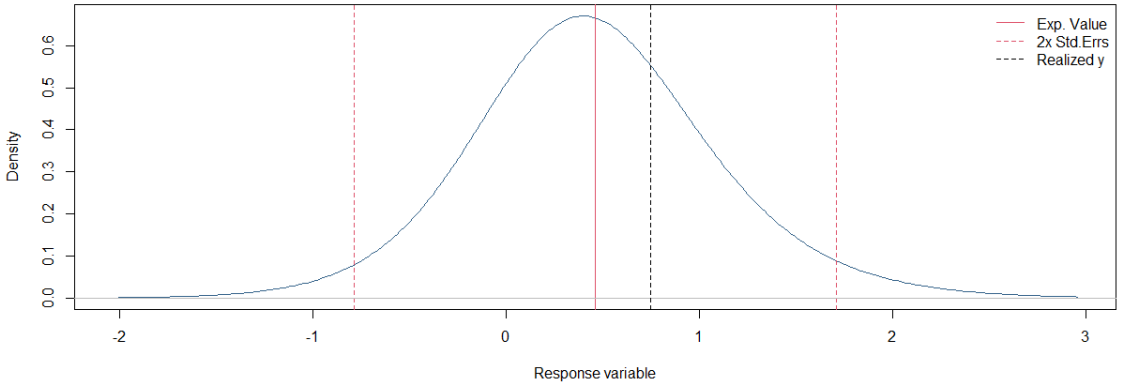
Singapore



Sweden



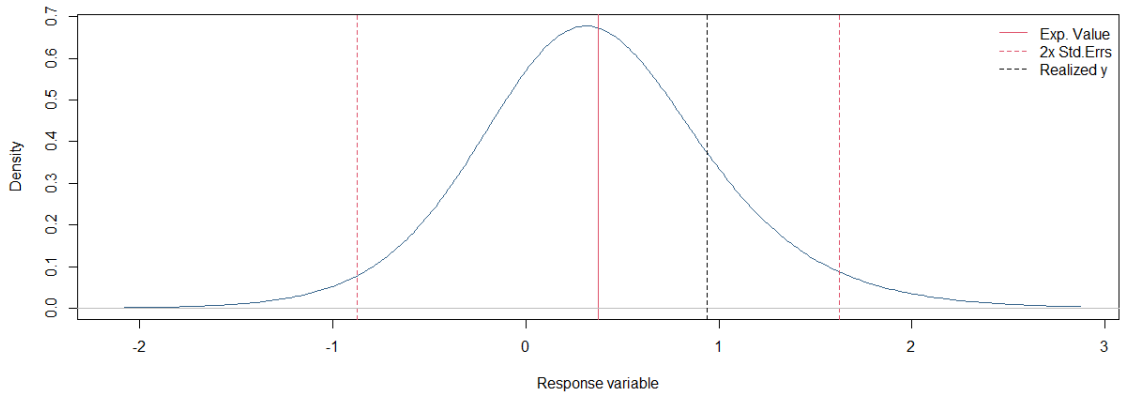
United States



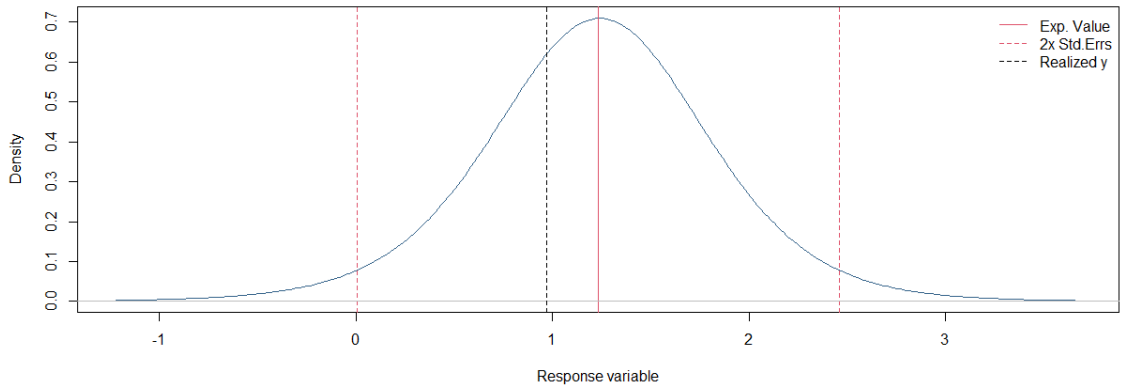
Appendix A.4

The following provides posterior density plots for female students in each country.

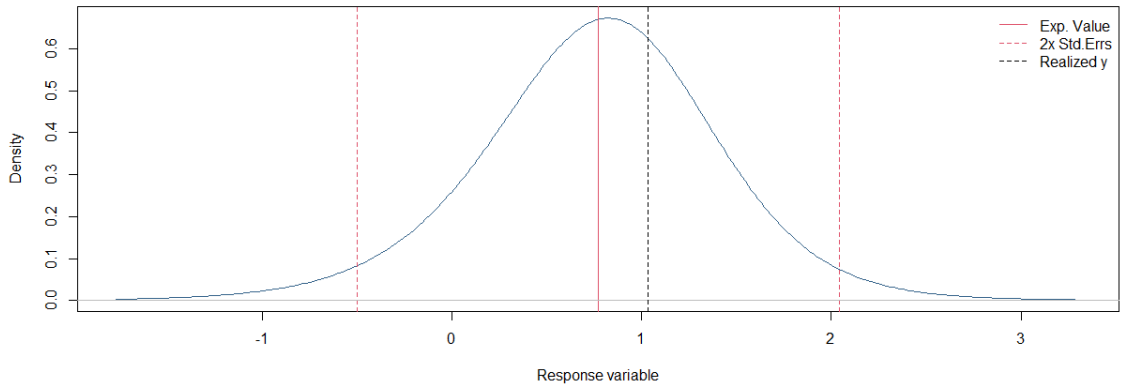
Canada (Ontario)



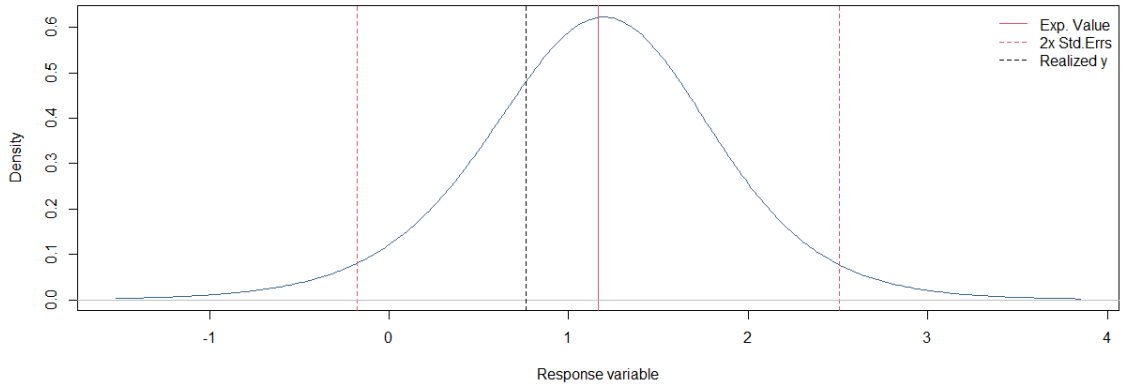
Canada (Quebec)



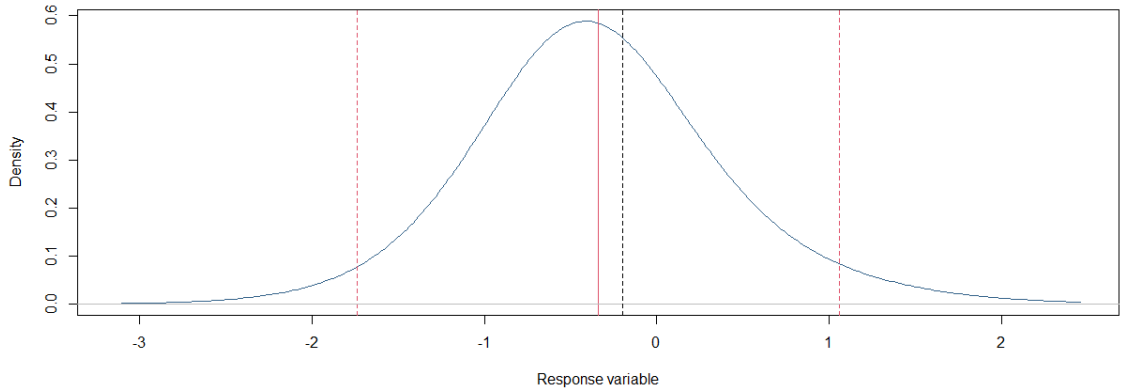
Chinese Taipei



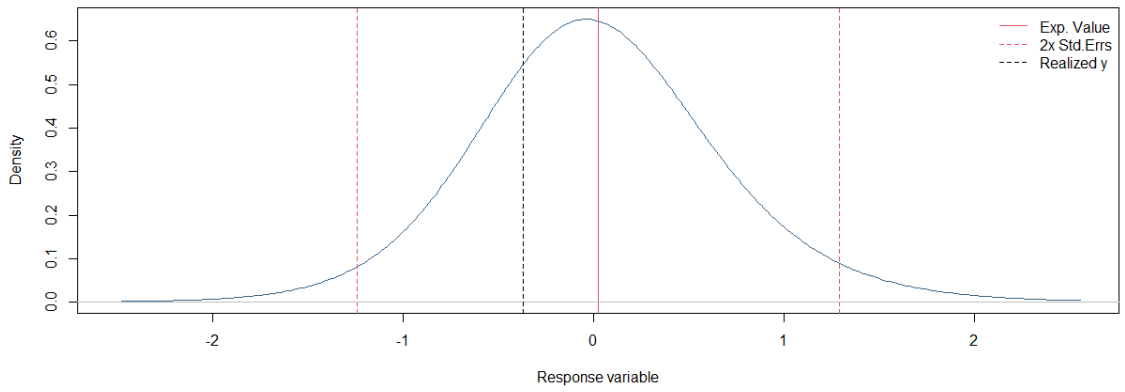
England



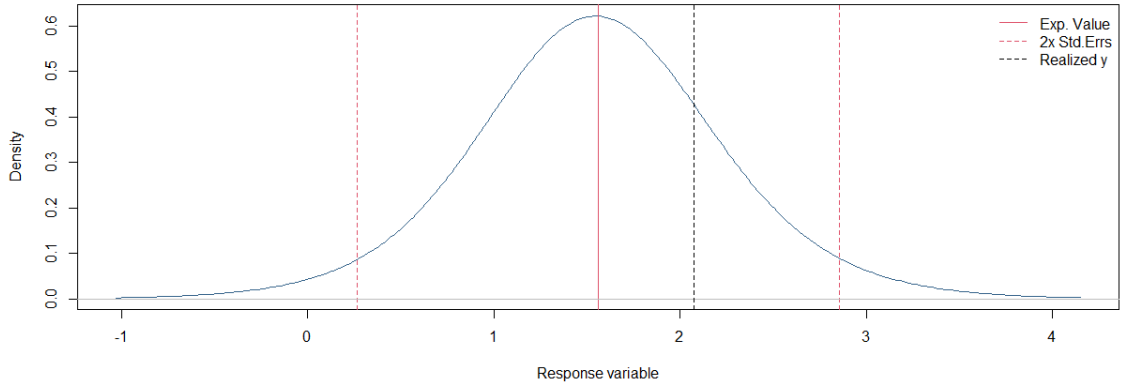
Hong Kong, SAR



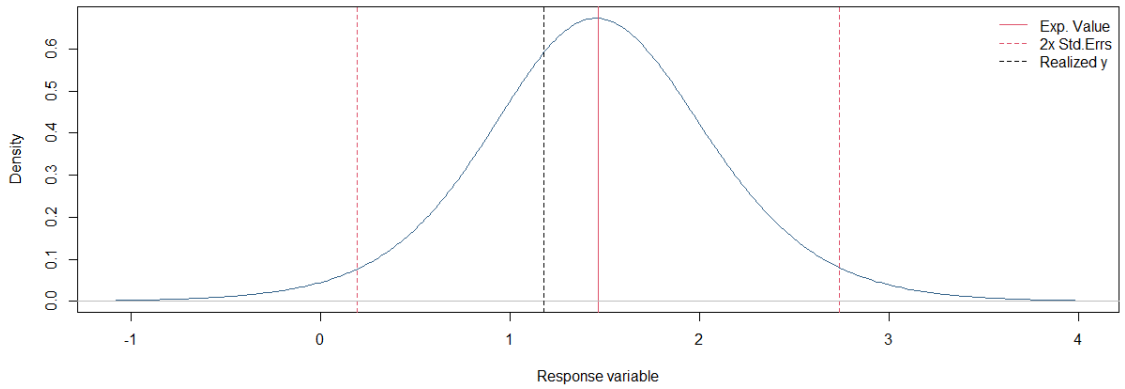
Hungary



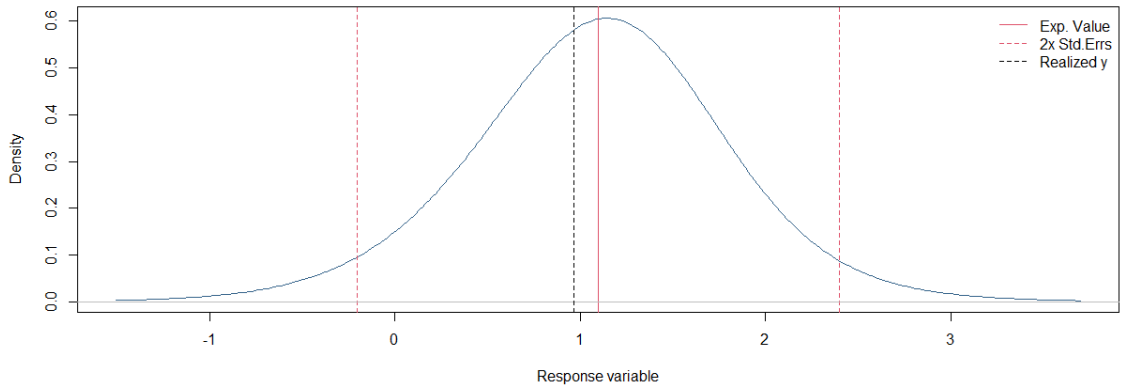
Israel



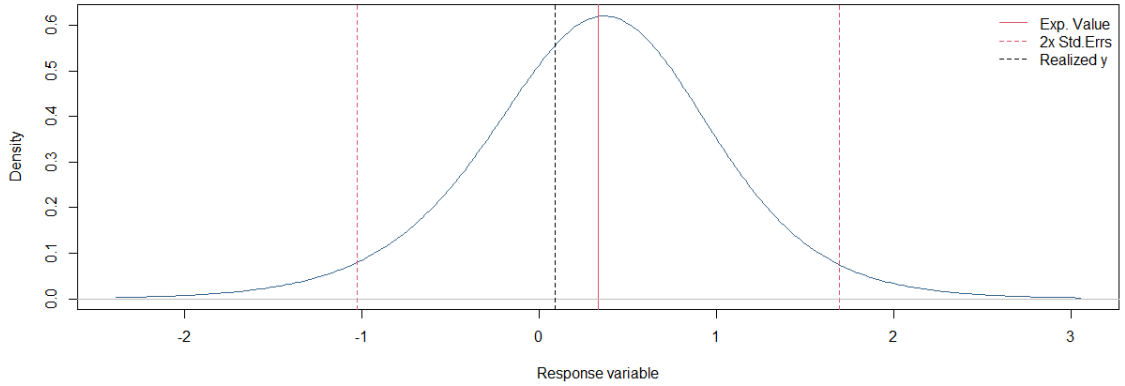
Italy



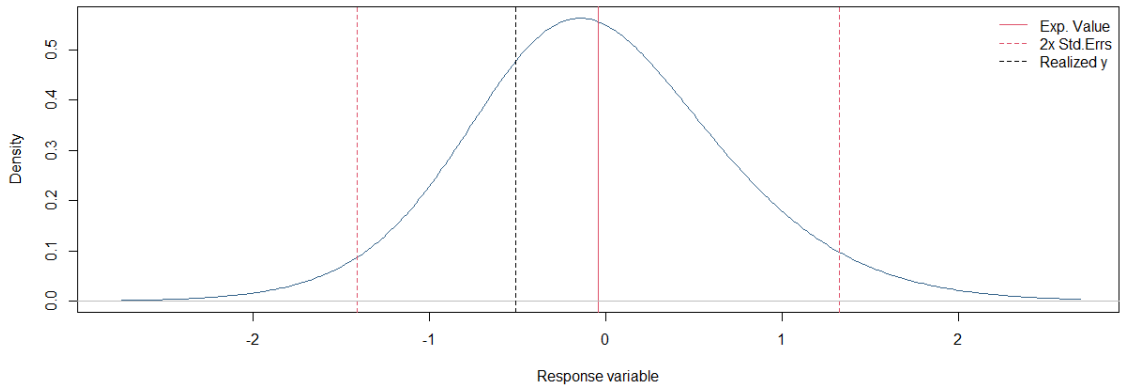
Korea, Republic of



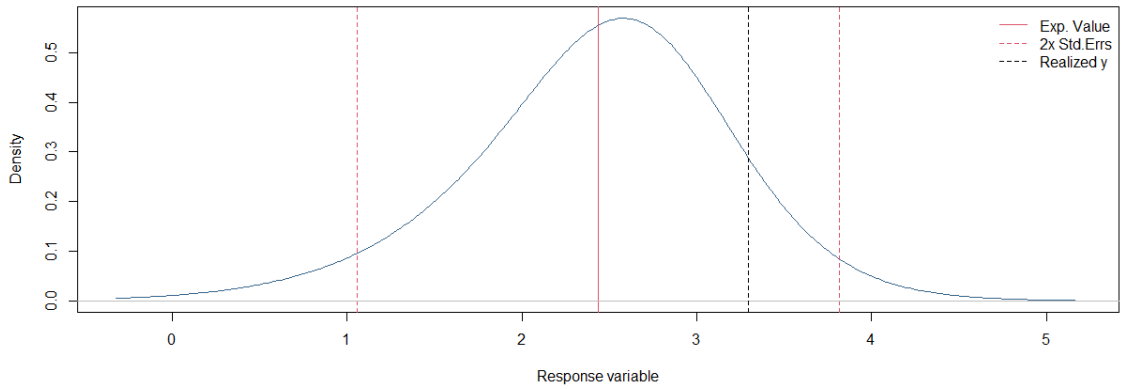
Lithuania



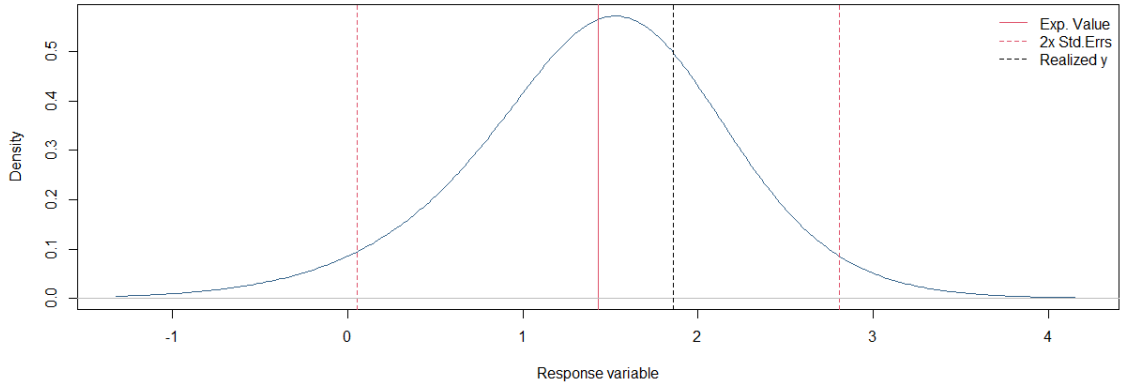
Malaysia



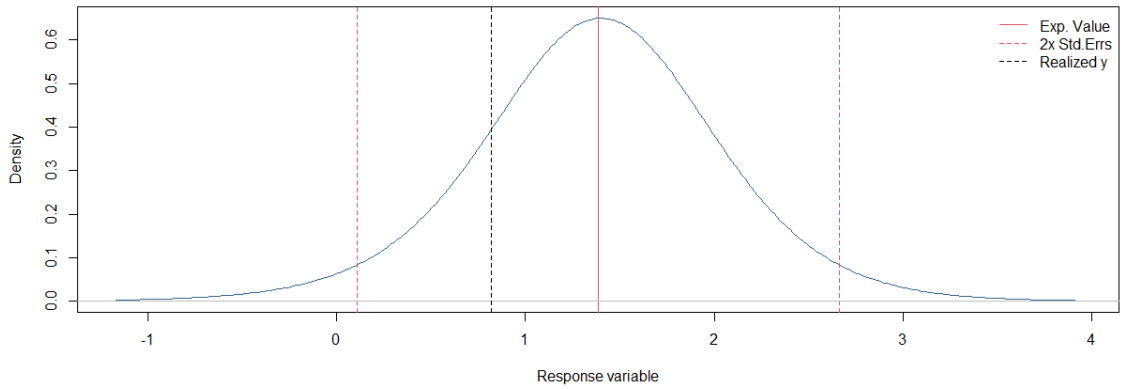
Norway



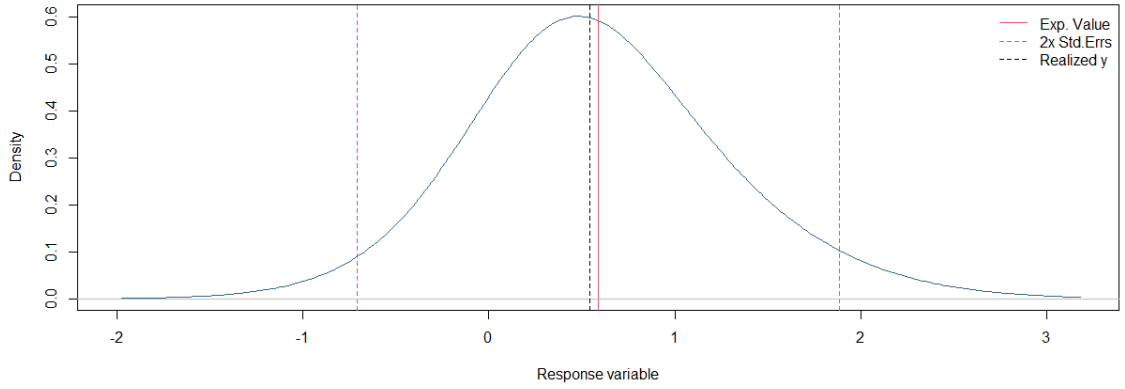
Russian Federation

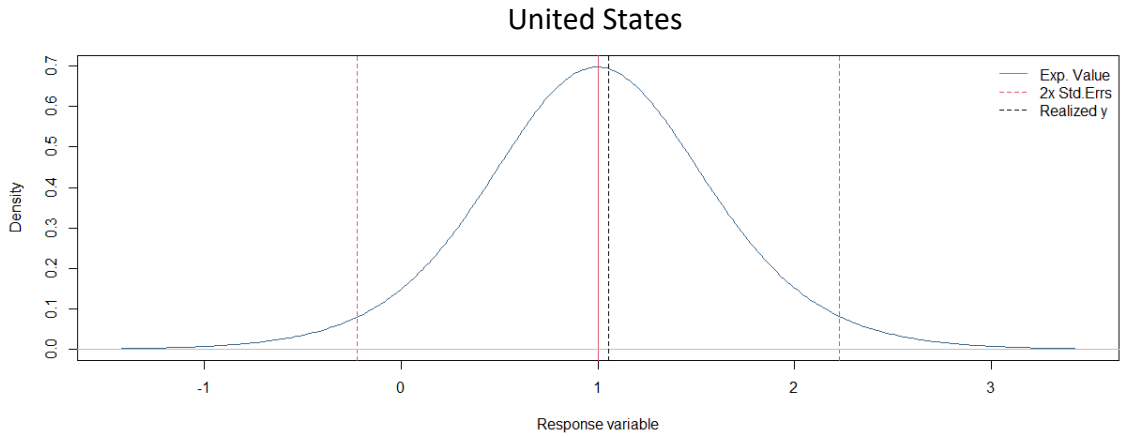


Singapore



Sweden




United States
Exp. Value
2x Std.Errs
Realized y
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
-1
0
1
2
3
Response variable

Bibliography

Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.

Bernardo, J. M., & Smith, A. F. (2009). *Bayesian theory* (Vol. 405). John Wiley & Sons.

Carvalho, A. (2016). An overview of applications of proper scoring rules. *Decision Analysis, 13*(4), 223-242.

Dawid, A. P., & Musio, M. (2015). Bayesian model selection based on proper scoring rules. *Bayesian Analysis, 10*(2), 479-499.

De Beer, L. T., & Bianchi, R. (2017). Confirmatory factor analysis of the maslach burnout Inventory. *European Journal of Psychological Assessment*.

Depaoli, S., & Clifton, J. (2015). A Bayesian Approach to Multilevel Structural Equation Modeling With Continuous and Dichotomous Outcomes. *Structural Equation Modeling: A Multidisciplinary Journal, 22*, 1-25. https://doi.org/10.1080/10705511.2014.937849

Depaoli, S., & van de Schoot, R. (2017, Jun). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychol Methods, 22*(2), 240-261. https://doi.org/10.1037/met0000065

Draper, D. (1987). *A research agenda for assessment and propagation of model uncertainty* (Vol. 2683). Rand.

Draper, D. (1999). Model uncertainty yes, discrete model averaging maybe. *Statistical Science, 14*, 405-409.

Eicher, T. S., Papageorgiou, C., & Raftery, A. E. (2011). Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics, 26*(1), 30-55.

Fernandez, C., Ley, E., & Steel, M. F. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics, 16*(5), 563-576.

Fishbein, B., Foy, P., & Yin, L. (2021). TIMSS 2019 user guide for the international database. *Hentet fra https://timssandpirls. bc. edu/timss2019/international-database*.

Foy, P., Fishbein, B., von Davier, M., & Yin, L. (2019). Implementing the TIMSS 2019 scaling methodology. *Methods and Procedures: TIMSS*, 12.11-12.146.

Fraley, C., & Percival, D. (2015). Model-averaged ℓ1 regularization using Markov chain Monte Carlo model composition. *Journal of statistical computation and simulation, 85*(6), 1090-1101.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Third edition. ed.). CRC Press.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359-378.

Gonzalez, E. J., Galia, J., & Li, I. (2003). Scaling methods and procedures for the TIMSS 2003 mathematics and science scales. *TIMSS*, 253-273.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science, 14*(4), 382-401.

Kaplan, D., & Huang, M. (2021). Bayesian probabilistic forecasting with large-scale educational trend data: a case study using NAEP. *Large-scale Assessments in Education, 9*(1), 1-31. https://doi.org/10.1186/s40536-021-00108-2

Kim, M., & Wang, Z. (2021). Factor Structure of the PANAS With Bayesian Structural Equation Modeling in a Chinese Sample. *Evaluation & the Health Professions*, 0163278721996794.

Ley, E., & Steel, M. F. J. (2009). On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression. *Journal of Applied Econometrics, 24*(4), 651-674. https://doi.org/10.1002/jae.1057

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association, 103*(481), 410-423.

Madigan, D., & Raftery, A. E. (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association, 89*(428), 1535-1546. https://doi.org/10.2307/2291017

Merkle, E. C., & Steyvers, M. (2013). Choosing a strictly proper scoring rule. *Decision Analysis, 10*(4), 292-304.

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association, 92*(437), 179-191. https://doi.org/10.2307/2291462

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). sage.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6*(2), 461-464.

Taplin, R. H. (1993). Robust Likelihood Calculation for Time Series. *Journal of the Royal Statistical Society. Series B (Methodological), 55*(4), 829-836.

Tierney, L., & Kadane, J. B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association, 81*(393), 82-86. https://doi.org/10.2307/2287970

Van De Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods, 22*(2), 217.

van Erp, S., Mulder, J., & Oberski, D. L. (2018, Jun). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychol Methods, 23*(2), 363-388. https://doi.org/10.1037/met0000162

Zeugner, S., & Feldkircher, M. (2009). *Benchmark priors revisited: on adaptive shrinkage and the supermodel effect in Bayesian model averaging*. International Monetary Fund.

Zeugner, S., & Feldkircher, M. (2015). Bayesian Model Averaging Employing Fixed and Flexible Priors: The BMS Package for R. *Journal of Statistical Software, 68*(1), 1-37. https://doi.org/10.18637/jss.v068.i04