Computing quantities of interest and their uncertainty using Bayesian simulation [*]

ANDREAS MURR, RICHARD TRAUNMÜLLER AND JEFF GILL

When analyzing data, researchers are often less interested in the parameters of statistical models than in functions of these parameters such as predicted values. Here we show that Bayesian simulation with Markov-chain Monte Carlo tools makes it easy to compute these quantities of interest with their uncertainty. We illustrate how to produce customary and relatively new quantities of interest such as variable importance ranking, posterior predictive data, difficult marginal effects, and model comparison statistics to allow researchers to report more informative results.

[*] Andreas Murr, corresponding author, Department of Politics and International Studies, University of Warwick, Coventry, UK (a.murr@warwick.ac.uk). Richard Traunmüller, School of Social Sciences, University of Mannheim, Mannheim, DE. Jeff Gill, Departments of Government, Mathematics & Statistics, American University, Washington, DC, USA. The research was partly supported by an International Visiting Fellowship from the Institute of Advanced Study, University of Warwick. A previous version of this paper was presented at the Society for Political Methodology XXXVI Meeting. We thank Paul Kellstedt, Jonathan Katz, Jacob Montgomery, Justin Esarey, Walter Mebane, and the anonymous reviewers for constructive comments.

Answering empirical research questions requires estimating quantities of interest and their uncertainty. For instance, researchers typically would like to know which explanatory variable is the most important one or whether its effect depends on the value of another variable. While estimating these quantities of interest is analytically straightforward most of the time, it is less often so for their measures of uncertainty. To address this issue, King, Tomz, and Wittenberg (2000) were the first in political science to explicitly advocate the use of simulation in such situations. In the same work they also recognized that "[f]ully Bayesian methods, using Markov-Chain Monte Carlo techniques, are more powerful than our algorithms" (352). Essential criteria for obtaining reliable inferences with Markov-Chain Monte Carlo (MCMC) tools include convergence and mixing (Gill 2014). Assuming that these criteria have been met, below we illustrate how to use the output of MCMC estimation to easily estimate quantities of interest and their uncertainty, often by simply sorting saved simulation values for a parameter of interest. Table 1 lists the illustrated quantities as well as example research questions for which these quantities are of interest. Table 2 shows which data and models we use to illustrate each quantity.

TABLE 1 *Quantities of interested illustrated below and example research questions for which these quantities are of interest.*

ID	Quantity of interest	Example question
Estimate level		
1	Coefficient estimates	What is the effect?
2	Relative effect size	Which effect is largest?
3	Marginal effects	How does the marginal effect vary?
Observation level		
4	Residuals	Does the data meet the normality assumption?
5	Posterior predictive checks	Does the model adequately capture key features of the data?
Model level		
6	Explained variance	How well does the model fit the data?
7	Predictive error	How well does the model predict new data?

TABLE 2 *Data and models used to illustrate quantities of interest.*

Data	Type	Model	Quantity (ID)
Candidate ratings	Experimental	Hierarchical linear model	1, 2
Voter turnout	Observational	Logit model	3, 5, 7
Union density	Observational	Linear model	4, 6

Estimate level

Coefficient estimates

Consider the conjoint experiment on candidate ratings from Hainmueller, Hopkins, and Yamamoto (2014). This experiment asked respondents to rate their support of hypothetical presidential candidates who differed in eight attributes: gender, age, race, education, profession, income, religion, and military service. Since each respondents rated twelve candidate profiles, the experimental data has a hierarchical structure with profile ratings nested in respondents. We use a hierarchical linear model with the candidate attributes as profile-level treatments and allow the intercept to vary by respondents. (This hierarchical linear model gives tighter credible intervals of the treatment effects than the linear model with clustered standard errors used by Hainmueller, Hopkins, and Yamamoto (2014). The reason is that it captures respondents' heterogeneity more appropriately in the variation of the intercepts instead of in the standard errors of the coefficients.)

The posterior densities of the coefficients (or treatment effects in this case) are shown in Figure 1 with a vertical line at zero. Notice the location and spread of the coefficients. The coefficients are reliably estimated, though, as expected, the reliability increases with fewer treatment levels per attribute (more respondents per treatment level). For instance, we see a small bias against Catholic candidates, whose estimated level of support is 0.005 lower when compared to a baseline candidate with no stated religion. Despite the relatively diffuse nature of this posterior distribution, most researchers would still conclude that the effect is negative and statistically reliable. The corresponding coefficient has 95% of its density to the left of zero.

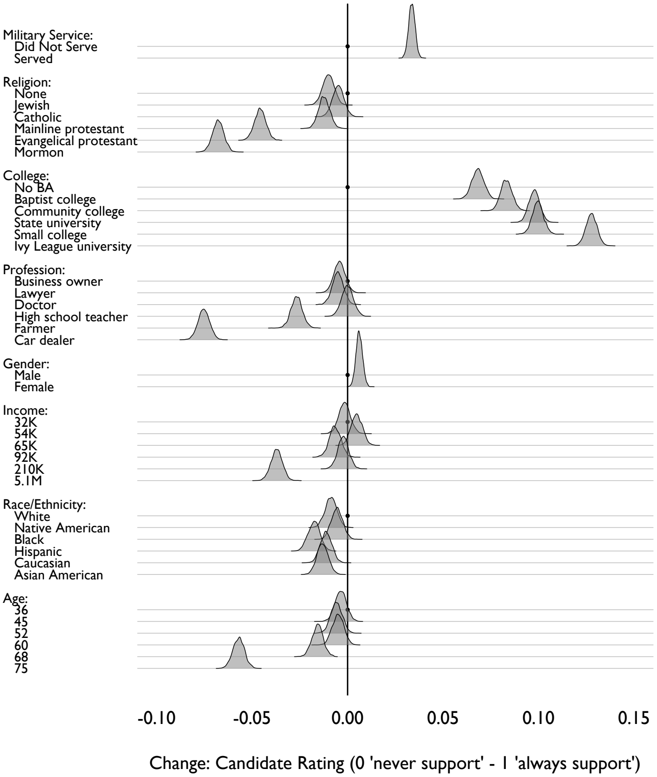


Figure 1. *The posterior density for coefficients in the hierarchical linear model of candidate ratings vary in location and spread. While candidates with a degree from a small college are rated higher than candidates with a degree from a state university, it is unclear how certain we can be of this statement (but see Figure 2).*

Relative explanatory variable importance

In regression a key question that should be asked is actually which explanatory variables are most influential in affecting variability in the outcome variable (effect size), not which have the smallest p-values. Fixation with p-values and star-counting is a well documented disease but remains prevalent nonetheless. Figure 1 shows that a candidate who attended an Ivy League university has higher ratings compared to a baseline candidate without a BA, and indeed any college degree. However, for candidates who attended a small college and candidates who attended a state university the posterior distributions overlap to some degree, showing uncertainty which of them has higher ratings. To quantify this uncertainty we go beyond Hainmueller, Hopkins, and Yamamoto (2014) and compute the probability that a variable has a certain rank of importance. For instance, to find the most important variable, we compute the probability that its absolute coefficient is largest. In other words, we compute the proportion of times the variable had the largest absolute coefficient of all variables across simulations. To find the second most important variable, we compute the proportion of times a variable had the second largest absolute coefficient, and so on. As a result, the uncertainty of this quantity flows from coefficient to relative variable importance.

Figure 2 shows the posterior probability of importance for each variable. There is a virtually 100% probability that the effect of a degree from an Ivy League university compared to no BA is largest. The effect of a degree from a small college compared to no BA has the highest probability of being the second most important. The probability is about 74%. Further, the probability that the effect of a degree from a state university compared to no BA is the third most important is also 74%. In other words, the ranks of variables are much more certain than one would expect from looking at Figure 1 alone.

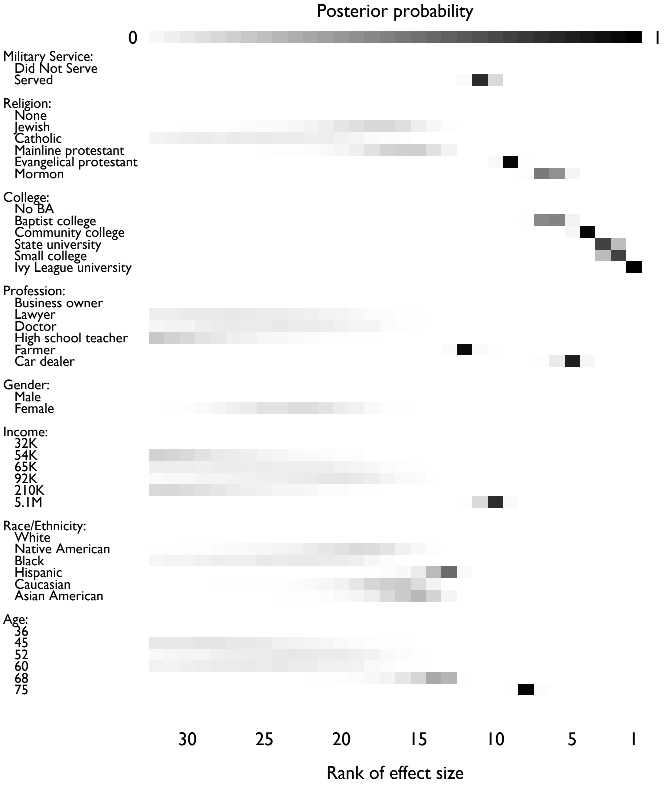


Figure 2. *Posterior probability of having the k-th largest effect size ('rank') for each treatment in the hierarchical linear model of candidate ratings. Most of the first twelve ranks are highly certain. The remaining ranks are much more uncertain.*

Note that one major advantage of Bayesian inference is that model results can be analyzed in explicitly *probabilistic* terms as done here.

Marginal effects

Marginal effects in regression modeling are well-studied in political science, including possible misinterpretations. Here we provide some ideas about extracting additional information about marginal effects. Consider modeling the probability of turning out to vote, $\pi_i = P(Y_i = 1|X_i)$, with a logistic specification. For illustrative purposes, we estimate the same model as King, Tomz, and Wittenberg (2000):

$$M_1 : \ \operatorname{logit}(\pi_i) = \beta_1 + \beta_2 \cdot \texttt{income}_i + \beta_3 \cdot \texttt{white}_i + \beta_4 \cdot \texttt{age}_i + \beta_5 \cdot \texttt{age}_i^2 + \beta_6 \cdot \texttt{educate}_i. \quad (1)$$

King, Tomz, and Wittenberg (2000, 355) included both age and age-squared "to test the hypothesis that turnout rises with age until the respondent nears retirement, when the tendency reverses itself." They tested this hypothesis by estimating the predicted probability of turnout, and its uncertainty, as a function of age for two different levels of education, while holding the other variables at their means.

However, estimating probabilities at mean values can be less meaningful when variables are binary (`white`) or spread out (`income`). In addition, the hypothesis can be tested more directly by looking at the marginal effect of age on turnout, which is defined as how much the predicted probability of voting changes when age changes. Hence, below we estimate the marginal effect of age on turnout, and its uncertainty, for each respondent in the data set. We easily compute the uncertainty in the marginal effects from the MCMC samples of the posterior density of $\boldsymbol{\beta}$.

The top-panel of Figure 3 displays the probability of turnout at mean values, the bottom-panel displays the marginal effects at observed values. Comparing both panels, it is clear that marginal effects at observed values provide more information. First, marginal effects at observed values unmask substantial heterogeneity in respondents with the same age and level of education. Consider respondents with a high school degree at the age of 34. At mean values, such a hypothetical respondent has the most precisely estimated probability of turnout: the credible interval is shortest among all age groups. However,

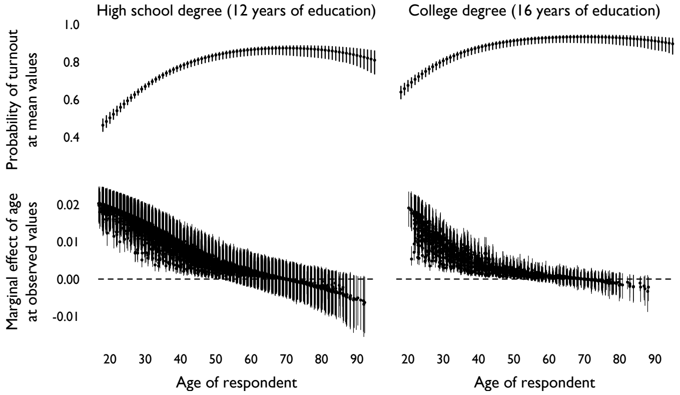


Figure 3. Probability of turnout at mean values (top panel replicated from King, Tomz, and Wittenberg (2000, Figure 1 on p. 355)) and marginal effects at observed values (bottom panel). The marginal effects unmask heterogeneity in respondents with the same age and level of education, show the location of the turnout plateau more clearly, and convey the observed values. Note: *Dots represent posterior medians. Vertical bars indicate 99% credible intervals. In the bottom panels the age values are jittered slightly to increase readability. The horizontal dashed line indicates a marginal effect of 0.*

at observed values, the actual respondents have the highest variation in marginal effects: their posterior medians have the highest variance among all age groups.

Second, the marginal effects at observed values pinpoint the location of the turnout plateau more clearly. When looking at the probability of turnout at mean values, King, Tomz, and Wittenberg (2000) see a “plateau between the ages 45 and 65”. However, when looking at the marginal effects of age at observed values, we see that the plateau begins later, around the age of 52 (credible intervals begin to include 0), or much later, around the age of 70 (posterior medians begin to equal 0).

Finally, the marginal effects at observed values convey what values of the variables were actually observed. For instance, they show that there are no 18 or 95 year olds with 16 years of education. Overall, the marginal effects at observed values, and their uncertainty, add substantial information relative to predicted probabilities at mean values.

Observation level

Residuals

The most common way to evaluate the properties of a linear model is to inspect its residuals, $r_i = y_i - \hat{y}_i$. This includes checking the full distribution of the residuals to see whether they are distributed as assumed by the model, particularly with regard to skewness and fixed patterns, as well as fit to specific observations. Our Bayesian simulation approach focuses on the vector of residuals for each observation with length equal to the number of simulations post-convergence.

To illustrate this approach we rely on a classic example of applying a linear model of union density in 20 OECD countries (Western and Jackman 1994) where union density is modeled with three explanatory variables (government control by leftist parties, size of the labor force, and economic concentration) whose coefficients are given informative normal priors. In this case the required choice of prior distribution can actually matter due to the sample size. Our purpose here is not to focus on the influence that priors can have with moderately sized data, but instead we assume that a reasonable choice has been made and the researcher is interested in producing additional revealing posterior quantities such as the distribution of residuals.

Figure 4 assesses whether the distribution of the residuals follows an approximately normal distribution using a density plot as well as a QQ-plot. Importantly, we not only look at *one* distribution of the residuals but at a *thousand* residual distributions–one for each iteration of the MCMC sampler. The density plot shows that the residual distributions are all roughly symmetric, having no major deviations. The QQ-plot shows that the distributions of all residual quantiles squarely include the reference line. Again, we see no

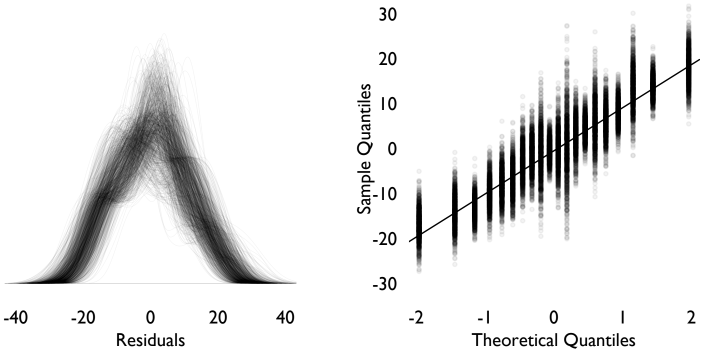


Figure 4. *Posterior distributions of residuals in linear model of union density to check for normality. The distributions look symmetric and conform with theoretical quantiles.* Note: *Grey lines/dots are 1000 simulation draws from the posterior, the black line in the right plot indicates the posterior means.*

clear sign of misfit with the normality assumption.

Posterior predictive checks

Posterior predictive checks can be used to explore the model fit (e.g., Gelman and Hill 2006). We first generate replicated data sets from the posterior predictive distribution of the logit model of turnout using the MCMC simulations. We then compare these replicated data sets with the observed data with regards to an interesting aspect of the data. The functions of the *data* used to compare with the model are called *test variables* $T(y)$. We will compare the $T(y)$ with the test variable in the *replications* $T(y^{\text{rep}})$. Specifically, we consider the proportion of voters in subpopulations: the turnout rate at each value of a predictor.

The first and third column of Figure 5 displays the turnout rate among subpopulations defined by unique values of the predictors. The solid black line shows the data, $T(y)$, and the grey lines represent 20 simulated replications from the model $T(y^{\text{rep}})$. The second and fourth column of Figure 5 shows $T(y) - T(y^{\text{rep}})$. Systematic differences from the horizontal line represent aspects of the data that are not captured by the model. For most subpopulations the model captures their turnout rate, with one exception: education. A close inspection of Figure 5 reveals three important features of the data that are not captured by the model. First, the actual turnout rate is larger than predicted among respondents with fewer years of education. Second, the actual turnout rate rises more quickly after 12 years of education than before. Finally, the actual turnout rate rises non-linearly—this is particularly evident for citizens with more than 12 years of education.

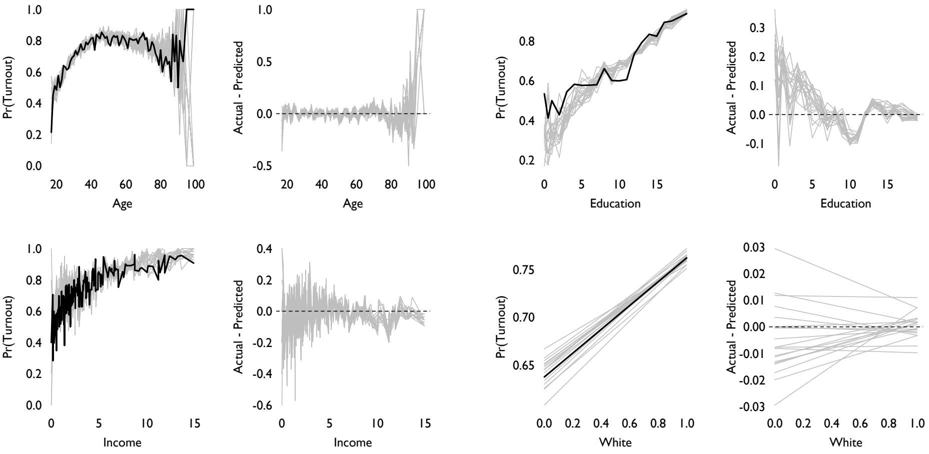
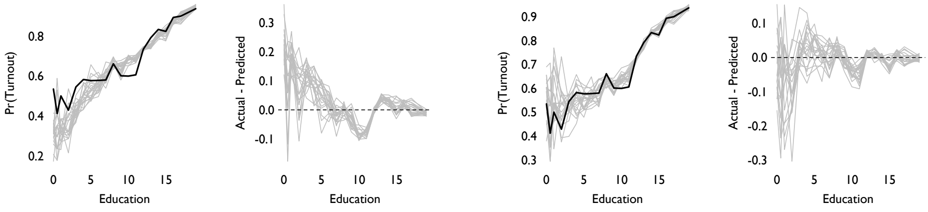


Figure 5. Posterior predictive checks of the logit model of turnout. The model fails to adequately captures the turnout rate among education subgroups. Note: *The solid black line (———) shows the data and the grey lines (———) represent 20 simulated replications from the model.*

To better model the turnout rate, M_1 is augmented by adding four predictors: education-squared, a dummy variable that indicates whether a citizen at least started a college degree



Model M_1 Model M_2

Figure 6. Posterior predictive checks of the original and revised logit models of turnout. The revised model better captures the turnout rate among education subgroups. Notes as in Figure 5.

(more than 12 years), and interactions between the college and education variables:

$$\begin{aligned} M_2: \quad \text{logit}(\pi_i) = {} & \beta_1 + \beta_2 \cdot \texttt{income}_i + \beta_3 \cdot \texttt{white}_i + \beta_4 \cdot \texttt{age}_i + \beta_5 \cdot \texttt{age}_i^2 \\ & + \beta_6 \cdot \texttt{educate}_i + \beta_7 \cdot \texttt{educate}_i^2 + \beta_8 \cdot \texttt{college} \\ & + \beta_9 \cdot \texttt{educate}_i \cdot \texttt{college} + \beta_{10} \cdot \texttt{educate}_i^2 \cdot \texttt{college}. \end{aligned} \tag{2}$$

After fitting the model we computed the same posterior predictive checks.

The first two panels of Figure 6 display the turnout rate by years of education in the data and in 20 simulation replications from model M_1. For comparison the last two panels of Figure 6 displays the same for the new model M_2. Figure 6 demonstrates that the new model captures the turnout rate much better than the previous one. As we will see in the next section, despite having more parameters to estimate, model M_2 also has a lower expected out-of-sample predictive error than model M_1. Note that the tools in this section could also be used with an out-of-sample procedure (Vehtari, Gelman, and Gabry 2017).

Model level

A requirement when statistical models are fit is that the researcher provides a suite of measures demonstrating that the data simplification is sympathetic with the original data. These vary by modeling approach but often involve building summaries from residuals, theoretical quantities, and outcome values. In this section we apply our Bayesian simulation approach to some common measures of fit to show how to easily obtain additional information.

Explained variance

Linear regression results include the classical R-squared, which measures the proportion of variance explained. While this measure has some problems (e.g., adding explanatory variables never decreases the measure), it can be a handy summary of the model fit. From a Bayesian perspective, there are, however, two issues with this measure: first, we would like to include the posterior uncertainty of the coefficients in the computation of the model fit; and, second, with strong priors and weak data the classical R-squared can be larger than 1. To address both issues, Gelman et al. (2019) propose an alternative Bayesian R-squared. Its computation is based on the set of posterior simulations draws $\boldsymbol{\theta}^{(s)}$, $s = 1, \ldots, S$ and so it accounts for the posterior uncertainty in the coefficients. They define the Bayesian R-squared as the variance of the predicted values over the variance of the predicted values plus the expected residual variance. Hence it always ranges from 0 to 1. For a linear regression model the proportion of variance explained for new data is:

$$\text{Bayesian } R_s^2 = \frac{V_{i=1}^{n} \hat{y}_i^{(s)}}{V_{i=1}^{n} \hat{y}_i^{(s)} + \hat{\sigma}_s^2}, \tag{3}$$

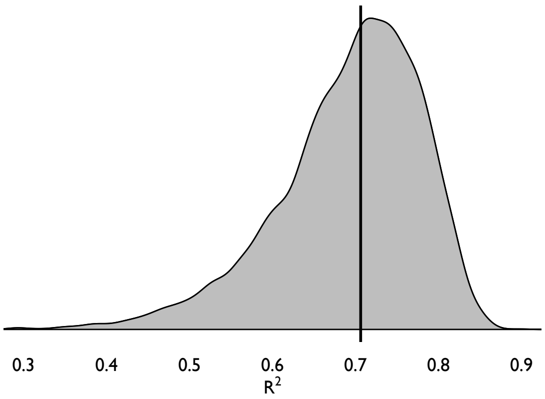


Figure 7. *Posterior distribution and median of the Bayesian R^2 in the linear model of union density.*

where $V_{i=1}^{n}$ represents the sample variance, $V_{i=1}^{n} z_i = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})^2$.

We illustrate the Bayesian R^2 using the above mentioned linear regression model fit to the union density data set (Western and Jackman 1994). Figure 7 shows the posterior distribution of the R^2 from 10,000 MCMC draws post-convergence. We find the posterior mean and median of $R^2 = .69$ and .71, respectively, and a 95% credible interval between .49 and .82.

Expected out-of-sample predictive error

Whereas effect sizes indicate variable importance in causal settings, changes in the expected out-of-sample error when dropping a variable from a model do so in predictive settings. To illustrate how to add information from simulation when estimating expected out-of-sample predictive error, we revisit the two turnout models specified above. We seek to show that model outcomes that are normally given only as point estimates can reveal more information when described probabilistically from simulations. Given the four additional predictors in M_2, one may worry about over-fitting the data. Hence, we compare the expected out-of-sample predictive errors of the two models to gauge the value of these additional predictors. To show how simulation adds information, we assess both models using two popular information criteria: the Akaike Information Criterion (AIC), and the Watanabe–Akaike information criterion (WAIC). Both information criteria estimate a model's expected out-of-sample-prediction error using a bias-corrected adjustments of within-sample error. But the computation of the WAIC relies on simulation, whereas the computation of the AIC does not.

The expression of AIC contains the log likelihood evaluated at the maximum likelihood estimate and the number of parameters to correct for bias due to overfitting:

$$\text{AIC} = -2 \sum_{i}^{n} \log p(y_i | \hat{\theta}_{\text{MLE}}) + 2k, \tag{4}$$

where k is the number of parameters. Two issues with the AIC are that (1) the number of parameters is a questionable penalty term for models with informative priors and hierarchical structures, and (2) inference for θ is summarized by a point estimate not by a full posterior distribution.

The WAIC addresses both issues: its expressions contains the log likelihood evaluated at the posterior simulations of the parameter values and the effective number of parameters. The WAIC can be estimated using posterior simulations (Vehtari, Gelman, and Gabry 2017):

$$\text{WAIC} = -2 \sum_{i}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i | \theta^s) \right) + 2 \sum_{i}^{n} V_{s=1}^{S} \left(\log p(y_i | \theta^s) \right). \tag{5}$$

The benefit of using simulations to compute the WAIC is that we get approximate standard errors for both the estimated predictive errors and the estimated difference in predictive errors between two models.

TABLE 3 *Akaike Information Criterion (AIC) and Watanabe–Akaike Information Criterion (WAIC) of logit models M_1 and M_2 of turnout and their difference. Despite having more predictors, the revised model M_2 predicts better that M_1. The Bayesian simulation of the WAIC enables computing standard errors (in parentheses) around the expected out-of-sample predictive errors and their difference.*

	M_1	M_2	$M_2 - M_1$
AIC	15836	15758	79
WAIC	15842 (132)	15763 (133)	79 (19)

Table 3 shows estimates of the expected out-of-sample predictive errors and their difference between models. Note that in contrast to the AIC, for the WAIC we also get standard errors as a measure of uncertainty. Looking at the WAIC, model M_2 improves upon M_1 with great certainty. The improvement in WAIC is 79 with a standard error of 19. Because its expected out-of-sample predictive error is lower, we favor this model M_2 over model M_1. The Bayesian simulation approach could also be applied to cross-validation measures of model quality.

Conclusion

A common task for empirical researchers is to obtain ancillary quantities of interest from a regression model, which should be accompanied by measures of uncertainty. Often these measures of uncertainty are difficult to calculate analytically, but easy to compute from MCMC output. This comes down to simple arithmetic operations and then merely sorting the results. While this work illustrates this point on a small set of models, it is applicable to a wide range of statistical settings (see Online Appendix). Our main message is that researchers are now free to tap into the full potential of Bayesian stochastic simulation to creatively summarize model results. Annotated code and data for use with this approach is provided in our `Dataverse` archive corresponding to this research note.

References

Gelman, Andrew, Ben Goodrich, Jonah Gabry, and Aki Vehtari. 2019. "R-squared for Bayesian Regression Models." *The American Statistician* 73 (3): 307–309.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press.

Gill, Jeff. 2014. *Bayesian Methods for the Social and Behavioral Sciences.* Third Edition. Chapman & Hall/CRC.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22 (1): 1–30.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44 (2): 341–355.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical Bayesian Model Evaluation Using Leave-one-out Cross-validation and WAIC." *Statistics and Computing* 27:1413–1432.

Western, Bruce, and Simon Jackman. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88 (2): 412–423.