R-squared for Bayesian regression models*

Andrew Gelman† Ben Goodrich‡ Jonah Gabry‡ Aki Vehtari§

4 Nov 2018

Abstract

The usual definition of R^2 (variance of the predicted values divided by the variance of the data) has a problem for Bayesian fits, as the numerator can be larger than the denominator. We propose an alternative definition similar to one that has appeared in the survival analysis literature: the variance of the predicted values divided by the variance of predicted values plus the expected variance of the errors.

1. The problem

Consider a regression model of outcomes y and predictors X with predicted values $\mathrm{E}(y|X,\theta)$, fit to data $(X,y)_n$, $n = 1,\dots,N$. Ordinary least squares yields an estimated parameter vector $\hat{\theta}$ with predicted values $\hat{y}_n = \mathrm{E}(y|X_n,\hat{\theta})$ and residual variance $V_{n=1}^N \hat{y}_n$, where we are using the notation,

$$V_{n=1}^N z_n = \frac{1}{N-1}\sum_{n=1}^N (z_n - \bar{z})^2, \text{ for any vector } z.$$

The proportion of variance explained,

$$\text{classical } R^2 = \frac{V_{n=1}^N \hat{y}_n}{V_{n=1}^N y_n}, \tag{1}$$

is a commonly used measure of model fit, and there is a long literature on interpreting it, adjusting it for degrees of freedom used in fitting the model, and generalizing it to other settings such as hierarchical models; see, for example, Xu (2003) and Gelman and Pardoe (2006).

Two challenges arise in defining R^2 in a Bayesian context. The first is the desire to reflect posterior uncertainty in the coefficients, which should remove or at least reduce the overfitting problem of least squares. Second, in the presence of strong prior information and weak data, it is possible for the fitted variance, $V_{n=1}^N \hat{y}_n$ to be higher than total variance, $V_{n=1}^N y_n$, so that the classical formula (1) can yield an R^2 greater than 1 (Tjur, 2009). In the present paper we propose a generalization that has a Bayesian interpretation as a variance decomposition.

*To appear in *The American Statistician*. We thank Frank Harrell and Daniel Jeske for helpful comments and the National Science Foundation, Office of Naval Research, Institute for Education Sciences, Defense Advanced Research Projects Agency, and Sloan Foundation for partial support of this work.

†Department of Statistics and Department of Political Science, Columbia University.

‡Institute for Social and Economic Research and Policy, Columbia University.

§Department of Computer Science, Aalto University.

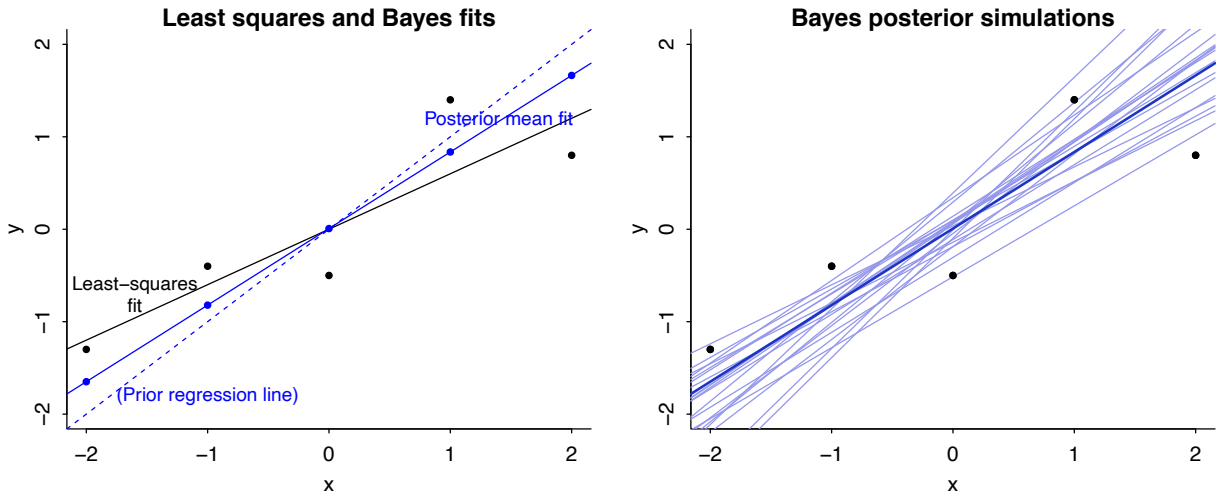

Figure 1: *Simple example showing the challenge of defining R^2 for a fitted Bayesian model.* Left plot: *data, least-squares regression line, and fitted Bayes line, which is a compromise between the prior and the least-squares fit. The standard deviation of the fitted values from the Bayes model (the blue dots on the line) is greater than the standard deviation of the data, so the usual definition of R^2 will not work.* Right plot: *posterior mean fitted regression line along with 20 draws of the line from the posterior distribution. To define the posterior distribution of Bayesian R^2 we compute equation* (3) *for each posterior simulation draw.*

2. Defining R^2 based on the variance of estimated prediction errors

Our first thought for Bayesian R^2 is simply to use the posterior mean estimate of θ to create Bayesian predictions $\hat{y}_n$ and then plug these into the classical formula (1). This has two problems: first, it dismisses uncertainty to use a point estimate in Bayesian computation; and, second, the ratio as thus defined can be greater than 1. When $\hat{\theta}$ is estimated using ordinary least squares, and assuming the regression model includes a constant term, the numerator of (1) is less than or equal to the denominator by definition; for general estimates, though, there is no requirement that this be the case, and it would be awkward to say that a fitted model explains more than 100% of the variance.

To see an example where the simple R^2 would be inappropriate, consider the model $y = \alpha + \beta x +$ error with a strong prior on (α, β) and only a few data points. Figure 1a shows data and the least-squares regression line, with R^2 of 0.77. We then do a Bayes fit with informative priors $\alpha \sim \text{N}(0, 0.2^2)$ and $\beta \sim \text{N}(1, 0.2^2)$. The standard deviation of the fitted values from the Bayes model is 1.3, while the standard deviation of the data is only 1.08, so the square of this ratio—R^2 as defined in (1)—is greater than 1. Figure 1b shows the posterior mean fitted regression line along with 20 draws of the line $y = \alpha + \beta x$ from the fitted posterior distribution of (α, β).

Here is our proposal. First, instead of using point predictions $\hat{y}_n$, we use expected values

conditional on the unknown parameters,

$$y_n^{\text{pred}} = \text{E}(\tilde{y}_n | X_n, \theta),$$

where $\tilde{y}_n$ represents a future observation from the model with predictors X_n. For a linear model, y_n^{pred} is simply the linear predictor, $X_n\beta$; for a generalized linear model it is the linear predictor transformed to the data scale. The posterior distribution of θ induces a posterior predictive distribution for y^{pred}.

Second, instead of working with (1) directly, we define R^2 explicitly based on the distribution of future data $\tilde{y}$, using the following variance decomposition for the denominator:

$$\text{alternative } R^2 = \frac{\text{Explained variance}}{\text{Explained variance} + \text{Residual variance}} = \frac{\text{var}_{\text{fit}}}{\text{var}_{\text{fit}} + \text{var}_{\text{res}}}, \tag{2}$$

where

$$\begin{aligned} \text{var}_{\text{fit}} &= V_{n=1}^N \text{E}(\tilde{y}_n|\theta) = V_{n=1}^N y_n^{\text{pred}} \text{ is the variance of the modeled predictive means, and} \\ \text{var}_{\text{res}} &= \text{E}(V_{n=1}^N (\tilde{y}_n - y_n^{\text{pred}})|\theta) \text{ is the modeled residual variance.} \end{aligned}$$

This first of these quantites is the variance among the expectations of the new data; the second term is the expected variance for new residuals, in both cases assuming the same predictors X as in the observed data. We are following the usual practice in regression to model the outcomes y but not the predictors X. As defined, var_{fit} and var_{res} are defined conditional on the model parameters θ, and so our Bayesian R^2, the ratio (2), depends on θ as well.

Both variance terms can be computed using posterior quantities from the fitted model: var_{fit} is determined based on y^{pred} which is a function of model parameters (for example, $y_n^{\text{pred}} = X_n\beta$ for linear regression and $y_n^{\text{pred}} = \text{logit}^{-1}(X_n\beta)$ for logistic regression), and var_{res} depends on the modeled probability distribution; for example, $\text{var}_{\text{res}} = \sigma^2$ for simple linear regression and $\text{var}_{\text{res}} = \frac{1}{N}\sum_{n=1}^N (\pi_n(1-\pi_n))$ for logistic regression.

By construction, the ratio (2) is always between 0 and 1, no matter what procedure is used to construct the estimate y^{pred}. Versions of (2) have appeared in the survival analysis literature (Kent and O'Quigley, 1988; Choodari-Oskoo et al., 2010), where it makes sense to use expected rather than observed data variance in the denominator, as this allows one to compute a measure of explained variance that is completely independent of the censoring distribution in time-to-event models. Our motivation is slightly different but the same mathematical principles apply, and our measure could also be extended to nonlinear models.

In Bayesian inference, instead of a point estimate $\hat{\theta}$, we have a set of posterior simulation draws, θ^s, $s = 1, \ldots, S$. For each θ^s, we can compute the vector of predicted values $y_n^{\text{pred}\,s} = \text{E}(\tilde{y}|X_n, \theta^s)$ and the expected residual variance $\text{var}_{\text{res}}^s$, and thus the proportion of variance explained is,

$$\text{Bayesian } R_s^2 = \frac{V_{n=1}^N y_n^{\text{pred}\,s}}{V_{n=1}^N y_n^{\text{pred}\,s} + \text{var}_{\text{res}}^s}, \tag{3}$$

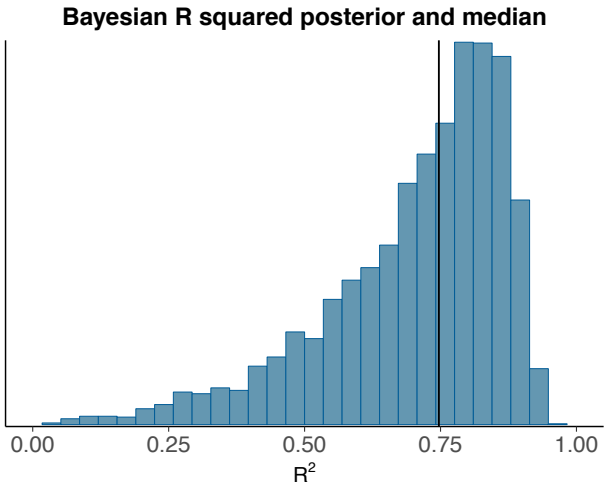

Figure 2: *The posterior distribution of Bayesian R^2 for the simple example shown in Figure 1 computed using equation* (3) *for each posterior simulation draw.*

where $\text{var}_{\text{res}}^s = (\sigma^2)^s$ for a linear regression model with equal variances.

For linear regression and generalized linear models, expression (3) can be computed using the `posterior_linpred` function in the `rstanarm` package and a few additional lines of code, as we demonstrate in the appendix, or see Gelman et al. (2018) for further development. For the example in Figure 1, we display the posterior distribution of R^2 in Figure 2; this distribution has median 0.75, mean 0.70, and standard deviation 0.17.

3. Discussion

R^2 has well-known problems as a measure of model fit, but it can be a handy quick summary for linear regressions and generalized linear models (see, for example, Hu et al., 2006), and we would like to produce it by default when fitting Bayesian regressions. Our preferred solution is to use (3): predicted variance divided by predicted variance plus error variance. This measure is model based: all variance terms come from the model, and not directly from the data.

A new issue then arises, though, when fitting a set of a models to a single dataset. Now that the denominator of R^2 is no longer fixed, we can no longer interpret an increase in R^2 as a improved fit to a fixed target. We think this particular loss of interpretation is necessary: from a Bayesian perspective, a concept such as "explained variance" can ultimately only be interpreted in the context of a model. The denominator of (3) can be interpreted as an estimate of the expected variance of predicted future data from the model under the assumption that the predictors X are held fixed; alternatively the predictors can be taken as random, as suggested by Helland (1987) and Tjur (2009). In either case, we can consider our Bayesian R^2 as a data-based estimate of the proportion of variance explained for new data. If the goal is to see continual progress of the fit to existing data, one can simply track the decline in the expected error variance, σ^2.

Another issue that arises when using R^2 to evaluate and compare models is overfitting. As with other measures of predictive model fit, overfitting should be less of an issue with Bayesian inference because averaging over the posterior distribution is more conservative than taking a least-squares or maximum likelihood fit, but predictive accuracy for new data will still on average be lower, in expectation, than for the data used to fit the model (Gelman et al., 2014). One could construct an overfitting-corrected R^2 in the same way that is done for log-score measures via cross-validation (Vehtari et al., 2017). In the present paper we are trying to stay close to the sprit of the original R^2 in quantifying the model's fit to the data at hand.

References

Choodari-Oskoo, B., P. Royston, and M. K. B. Parmar (2010). A simulation study of predictive ability measures in a survival model I: Explained variation measures. *Statistics in Medicine 31*, 2627–2643.

Gelman, A., B. Goodrich, J. Gabry, and A. Vehtari (2018). Bayesian R^2. `https://avehtari.github.io/bayes_R2/bayes_R2.html`.

Gelman, A., J. Hwang, and A. Vehtari (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing 24*, 997–1016.

Gelman, A. and I. Pardoe (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics 48*, 241–251.

Helland, I. S. (1987). On the interpretation and use of R^2 in regression analysis. *Biometrics 43*, 61–69.

Hu, B., M. Palta, and J. Shao (2006). Properties of R^2 statistics for logistic regression. *Statistics in Medicine 25*, 1383–1395.

Kent, J. T. and J. O'Quigley (1988). Measures of dependence for censored survival data. *Biometrika 75*, 525–534.

Tjur, T. (2009). Coefficient of determination in logistic regression models—A new proposal: The coefficient of discrimination. *American Statistician 63*, 366–372.

Vehtari, A., A. Gelman, and J. Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing 27*, 1413–1432.

Xu, R. (2003). Measuring explained variation in linear mixed-effects models. *Statistics in Medicine 22*, 3527–3541.

Appendix

This simple version of the `bayes_R2` function works with Bayesian linear regressions fit using the `stan_glm` function in the `rstanarm` package.

```
# Compute Bayesian R-squared for linear models.
#
# @param fit A fitted linear or logistic regression object in rstanarm
# @return A vector of R-squared values with length equal to
#      the number of posterior draws.
#
bayes_R2 <- function(fit) {
  y_pred <- rstanarm::posterior_linpred(fit)
  var_fit <- apply(y_pred, 1, var)
  var_res <- as.matrix(fit, pars = c("sigma"))^2
  var_fit / (var_fit + var_res)
}

## Example from Figure 1 of the paper
x <- 1:5 - 3
y <- c(1.7, 2.6, 2.5, 4.4, 3.8) - 3
xy <- data.frame(x,y)

## Bayes fit with strong priors
library("rstanarm")
fit_bayes <- stan_glm(y ~ x, data = xy,
  prior_intercept = normal(0, 0.2, autoscale = FALSE),
  prior = normal(1, 0.2, autoscale = FALSE),
  prior_aux = NULL)

## Compute Bayesian R2
rsq_bayes <- bayes_R2(fit_bayes)
hist(rsq_bayes)
print(c(median(rsq_bayes), mean(rsq_bayes), sd(rsq_bayes)))
```

Expanding the code to work for other generalized linear models requires some additional steps, including setting `transform=TRUE` in the call to `posterior_linpred` (to apply the inverse-link function to the linear predictor), the specification of the formula for var_{res} for each distribution class, and code to accomodate multilevel models fit using `stan_glmer`.