Incorporating Substantive Knowledge Into Regression Via A Bayesian Approach To Modeling

Roy Levy **Aaron V. Crawford**
Arizona State University

A multiple regression example is used to illustrate advantages of a Bayesian approach that incorporates situation-specific substantive information over frequentist and Bayesian approaches that ignore such information. Frequentist and Bayesian analyses of a traditional regression model produced nearly identical results. A Bayesian analysis of a modified model yielded preferred estimates of parameters and quality of prediction.

Regression models are useful for characterizing patterns and quantifying the relationships that exist among observable variables. An important and widespread application of regression is to facilitate predictions for an outcome. In applied analyses, regression model parameters are estimated based on sample data; frequently the estimated parameters are then used to make predictions for future cases whose outcomes are not known. The utility of a regression model for making predictions for future cases is therefore limited by the information that is available when it is constructed. The current work illustrates how using a Bayesian approach allows the researcher to incorporate substantive information about the problem to augment the information available from sample data to obtain preferred estimates of the parameters and the quality of prediction. Specifically, it will be shown in the context of a regression model for educational achievement tests that incorporating boundary constraints via a Bayesian approach to regression modeling yields preferred estimates of parameters and measures of prediction accuracy compared to traditional approaches.

Comparisons of frequentist and Bayesian approaches typically highlight the presence of prior distributions in the Bayesian framework. A common criticism of the Bayesian approach is that it is "only as good as the priors", meaning that if the prior distributions poorly match the structure of the data in the population, the Bayesian approach will suffer relative to a frequentist approach. On the other hand, as demonstrated in the current work, prior distributions can be a mechanism for incorporating substantive information into the model. While this is certainly one of the main ways that the two approaches differ, we will demonstrate that prior distributions are not the only way to incorporate characteristics of the substantive problem into the analysis. In the current example, it is argued that placing substantively motivated boundaries on the prior distribution and the likelihood—which are easily incorporated in a Bayesian approach with flexible estimation routines—yields preferred estimates of parameters and prediction quality. This is illustrated in the context of regression with small samples, where the substantive information that is brought to bear augments the information in the data.

Context and Data

The data used in the analyses come from the first three end-of-chapter exams associated with the course Networking Basics, the first of a four-course curriculum in the Cisco Networking Academy Program. Students in this program come from a wide variety of educational backgrounds, and are typically progressing toward certification that will allow them to work as computer networking professionals servicing home or business settings. For researchers of the Cisco Networking Academy Program, operational work in this context frequently involves characterizing relationships between performance on early exams and performance on later exams using regression. Moreover, the complexities of the online administration of exams yields situations in which sample sizes for such analyses vary considerably. As such, the regression analyses in operational work may employ small samples. This work illustrates the usage of Bayesian approaches to modeling that allow for the incorporation of substantive knowledge to improve data analysis in such contexts. The primary data used in the analyses consist of total scores from 50 students on the three exams. For each exam, scores in the population had the potential to range from zero to the number of items on the exam. There were 16 items on the first exam; in this sample, total scores ranged from 4 to 16, (M = 14.10, SD = 2.02). There were 18 items on the second exam; in the sample, total scores ranged from 3 to 18, (M = 14.34, SD = 3.29). The third exam had 15 items; in the sample, total scores ranged from 1 to 15 (M = 12.22, SD = 2.96). The zero-order correlations between the chapter exams are as follows: Chapters 1 and 2, 0.58; Chapters 1 and

3, 0.69; Chapters 2 and 3, 0.68. A second data set, consisting of test scores from 1950 students, was used in a follow-up analysis as described below.

Classical and Bayesian Analyses of a Traditional Regression Model

In each analysis, the scores on the first and second exams in the curriculum were used to predict scores on the third exam.

Classical Analysis. A traditional model regressing the third exam on the first and second exams is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \tag{1}$$

where X_{i1}, X_{i2}, and Y_i denote the total scores on the first, second, and third exams, respectively, for subject i, and $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. A classical approach to model estimation treats the parameters as fixed unknowns, commonly employing maximum likelihood (ML) or equivalently least squares estimation. Following the model in (1) and assumptions regarding errors, the likelihood function may be written as

$$L(\beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2 \mid \mathbf{X}, \mathbf{Y}) = \prod_{i=1}^{N} N(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}, \sigma_\varepsilon^2) \tag{2}$$

where **X** and **Y** are the full collections of predictor and dependent variables, respectively. Straightforward differentiation and analysis yields well-known closed form solutions for ML estimators of the parameters (e.g., Rencher, 2000).

Bayesian Analysis. A Bayesian approach to modeling differs from the classical approach by treating each entity as a random variable that can be characterized via probability distributions (Gelman, Carlin, Stern, & Rubin, 1995). A prior distribution is specified for unknown model parameters and the posterior distribution is given by Bayes' theorem:

$$P(\beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2 \mid \mathbf{X}, \mathbf{Y}) = \frac{P(\beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2) P(\mathbf{Y} \mid \beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2, \mathbf{X})}{\int_{\beta_0}\int_{\beta_1}\int_{\beta_2}\int_{\sigma_\varepsilon^2} P(\beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2) P(\mathbf{Y} \mid \beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2, \mathbf{X})}$$

$$\propto P(\beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2) P(\mathbf{Y} \mid \beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2, \mathbf{X}) \tag{3}$$

where $P(\mathbf{Y} \mid \beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2, \mathbf{X})$ is the conditional distribution of the data or likelihood function given in (2) and $P(\beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2)$ is the prior distribution for the model parameters.

The prior distribution is constructed via specifying independent components. Frequently, diffuse prior distributions are employed in situations where prior knowledge is limited. The current analysis adopts this approach to highlight the comparability between the classical approach and the Bayesian approach to the traditional model under such specifications. We specify diffuse generalized prior distributions (Press, 1989) in the form of normal distributions for the intercept and coefficients and an inverse-gamma distribution for the residual variance (for alternative specifications of prior distributions in regression and related contexts see Gelman et al., 1995; Gill, 2007; Lee, 2007)

$$\begin{aligned} P(\beta_0) &\sim N(0, 10{,}000); \\ P(\beta_1) &\sim N(0, 10{,}000); \\ P(\beta_2) &\sim N(0, 10{,}000); \\ P(\sigma_\varepsilon^2) &\sim Inv - G(.01, .01). \end{aligned} \tag{4}$$

Though analytical solutions to the model are available under certain choices of distributional forms (e.g., Gelman et al., 1995), they are frequently intractable for complex problems. The current work employs Markov chain Monte Carlo (MCMC; e.g. Gilks, Richardson, & Spiegelhalter, 1996) estimation to conduct the analyses, as MCMC algorithms capitalize on the proportionality relationship in (3) to

provide a flexible framework that allows for the estimation of complex models. MCMC consists of taking a series of draws to form a chain such that, in the limit, the chain converges to a stationary distribution such that subsequent draws may be viewed as draws from the stationary distribution (see Gilks et al., 1996 for details and an overview of popular MCMC algorithms). In a Bayesian analysis, we construct the chain so that the stationary distribution is the posterior distribution of interest.

MCMC estimation was conducted in WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2007) via the package R2WINBUGS (Sturtz, Ligges, & Gelman, 2005) in the R statistical environment (R Core Development Team, 2008). Annotated WinBUGS code for running this model and later models are contained in the Appendix. Steps in an MCMC analysis include monitoring the convergence of the chain(s), determining the number of iterations to discard as burn-in, and summarizing the remaining draws for the parameters.

Bayesian Analysis of a Modified Model

The modified model incorporates existing knowledge about the range of actual outcome possibilities in a way that the traditional regression model neglects. Specifically, the values of the criterion variable necessarily fall between zero and 15, the lowest and highest possible scores, respectively, on the Chapter 3 exam. A more thorough Bayesian analysis includes such substantive knowledge in the probability model (Gelman et al., 1995). The model modifications used here to incorporate that knowledge include changes to the prior distribution and changes to the likelihood.

The prior distribution for the intercept (β_0) is changed from a normal distribution to a uniform distribution bounded by the potential response range on the outcome variable. The prior distributions for the remaining parameters are made less diffuse to facilitate convergence of the more complex modified model, though these priors are still quite diffuse:

$$\begin{aligned} P(\beta_0) &\sim U(0, 15); \\ P(\beta_1) &\sim N(0, 1000); \\ P(\beta_2) &\sim N(0, 1000); \\ P(\sigma_\varepsilon^2) &\sim Inv-G(1,\ 1). \end{aligned} \tag{5}$$

The likelihood is altered by modifying the regression model, where the predicted values are adjusted to take into account the maximum possible score on the criterion. For students who would otherwise be predicted to score above 15 by the prediction equation (1), the out-of-bounds predicted score is changed to equal 15, which we designate to be the "adjusted predicted score". Given the positive bivariate relationships, the prior distribution for β_0 effectively serves to bound the predicted values below by 0. Estimation and convergence assessment were conducted using the same tools reported above for the original model.

We note that this modified model is similar in spirit to censored regression models (Tobin, 1958) for which ML and Bayesian approaches to estimation have been developed (Chib, 1992). However, censored regression models are limited in that they do not directly constrain the regression parameters and the proper interpretation of the parameters concerns the relationships between the predictors and the latent dependent variable. In the model adopted here, the use of the prior distribution in (5) directly constrains the intercept in accordance with substantive theory and yields parameters that concern the relationships between the predictors and the observed dependent variable.

Results

Table 1 summarizes the results of the models. For the classical analysis, ML estimates, standard errors, and 95% confidence intervals are reported, as is R^2. For the Bayesian analyses, history plots of the draws for each parameter and the Brooks-Gelman-Rubin diagnostic (Brooks & Gelman, 1998; Gelman & Rubin, 1992) were examined to determine that 1000 iterations were sufficient to burn-in the chains for both the traditional and modified regression model. For each model, the results in Table 1 were thus computed using iterations 1001 to 4000 for each of the three chains, for a total of 9000 iterations. Posterior means, standard deviations, and 95% credibility intervals are reported for the parameters and R^2.

Table 1. Summary of Results of Classical and Bayesian Analyses for the primary dataset.

	Classical Analysis of Traditional Model			Bayesian Analysis of Traditional Model			Bayesian Analysis of Modified Model		
	Estimate	SE	95% Confidence Interval	Posterior Mean	Posterior SD	95% Credibility Interval	Posterior Mean	Posterior SD	95% Credibility Interval
β_0	-2.54	1.93	(-6.41, 1.34)	-2.54	1.96	(-6.32, 1.31)	1.07	0.93	(0.03, 3.47)
β_1	0.66	0.17	(0.33, 0.99)	0.66	0.17	(0.33, 0.99)	0.40	0.12	(0.15, 0.63)
β_2	0.38	0.10	(0.18, 0.59)	0.38	0.10	(0.17, 0.58)	0.39	0.10	(0.18, 0.59)
σ_ε	1.95	0.28	(1.60, 2.37)	1.94	0.21	(1.59, 2.40)	1.98	0.21	(1.63, 2.43)
R^2	0.60			0.59	0.06	(0.45, 0.68)	0.48	0.05	(0.35, 0.56)

Discussion

The results of the Bayesian analysis of the traditional model—in terms of point estimates and intervals—closely mirrored those of the classical analysis, as expected given the use of diffuse priors. By contrast, the results of the Bayesian analysis of the modified model differed from those of the other analyses. These differences are highlighted by the results for β_1 and β_0. In terms of the latter, whereas the classical and Bayesian analysis of the traditional model allows β_0 to take on any real value, the use of the prior distribution in the modified model restricts the posterior distribution to be between zero and 15. This difference is summarized by the point estimates. The ML estimate and the posterior mean for β_0 for the traditional model is –2.54 and the posterior mean for β_0 for the modified model is 1.07. It is problematic to interpret the negative value for β_0 in the traditional model, as it is impossible for a student to score less than zero on the third exam. By construction, this is precluded in the modified model via the prior distribution for β_0.

Interestingly, the R^2 values for each model make it appear at first glance that the modified model (posterior mean of $R^2 = 0.44$) does not perform as well as the traditional models using ML or Bayesian analysis ($R^2 = 0.60$ and 0.59, respectively). This is a necessary result, as the ML solution to the traditional model maximizes R^2 in the sample on which the estimates are derived. However, to explore the difference in the quality of prediction, a second sample of 1950 students' tests scores was employed. For each student, the point estimates (ML estimates or posterior means in Table 1) from each of the models were used to generate a prediction. The squared correlations between these predictions and the true values were then calculated as R^2 statistics for this second dataset. When the regression model based on the original 50 sample scores are used to predict the 1950 scores in this dataset, the R^2 for each of the models is as follows: ML analysis of the traditional model, $R^2 = 0.43$, Bayesian analysis of the traditional model, $R^2 = .43$, Bayesian analysis of the modified model, $R^2 = 0.44$. These three R^2 values are not meaningfully different; the models performed equally well in predicting the outcome on the third exam in the larger sample. For all the models, using the prediction equation from the original sample to form predictions for new data naturally lowers each of these R^2 values relative to the values in the original sample. However, the differences in the amount of the reduction in R^2 when cross-validated with the second sample are revealing. The modified model displayed much less of this reduction than did the traditional models. This is interpreted as indicating that—in the original sample—the modified model provided the most realistic view of the predictive utility of the predictors in the population and future samples. Put another way, the traditional model capitalizes on variation in the sample data with which it is estimated and suffers when cross-validated on another dataset, whereas the modified model performs almost as well in estimating the cross-validating dataset as it does in the original sample. Note that using adjusted R^2 for the analyses of the original model yielded 0.58 and 0.57 for the classical and Bayesian analyses, respectively. Though these values are smaller than the values of R^2 reported in Table 1 (0.60 and 0.59), they still indicate considerably inflated predictive quality relative to the cross-validation. By incorporating existing substantive knowledge of the population, the modified model (necessarily) sacrifices predictive power in the original sample yet provides a more accurate estimate of the predictive accuracy for future samples.

For comparative purposes the traditional model using ML was fit on this cross-validation dataset; the results are given in Table 2. Viewing the results from this larger dataset as more representative of the population, note that the estimates from this model are quite close to those from the results from the modified model of the original data set. Additionally, for β_0, β_1, and R^2, the results are much closer to the modified model than the traditional model. From the perspective of the results from the second dataset, the estimates of the parameters (particularly β_0 and β_1) and the estimate of the quality of prediction (in terms of R^2) of the modified model of the original dataset yield more accurate results than those from the traditional model. This is because the modified model augments the data by incorporating known properties of the substantive problem into the model. On a criterion that ranges from zero to 15 in the population, it is intellectually unsatisfying if not contradictory to allow a predicted value outside this range for the range of possible values of the predictors. In the current context, the intercept represents such a prediction. Substantively, as researchers knowledgeable about the context, we know that it is impossible for a student to have a negative total score on the third exam, regardless of performance on the first two exams. Yet the traditional models do not allow us to incorporate this substantive knowledge. The fact that the model-implied intercepts for the traditional models were negative in the original data emphasizes the point that those models capitalized on chance when fitting the best line for the observed data. A Bayesian approach—supported by the flexibility of MCMC estimation—allows this prior knowledge to be brought to bear in modeling.

Table 2. Results of Classical Analysis for the cross-validation dataset.

	Estimate	SE	95% Confidence Interval
β_0	1.39	0.32	(0.76, 2.02)
β_1	0.39	0.03	(0.33, 0.45)
β_2	0.38	0.02	(0.34, 0.42)
σ_ε	2.17	0.00	(2.17, 2.17)
R^2	0.44		

In summary, this paper is intended to highlight an understudied advantage of a Bayesian approach to regression modeling, namely, the ease and flexibility with which substantive information may be incorporated to augment the information in the sample data when fitting models. We illustrate how that information can be modeled in the prior distribution (in the example, via the choice of the support of the prior distribution) and via the likelihood (in the example, by adjusting predicted values). The advantages manifest themselves in supporting inferences consistent with the population, which is particularly beneficial in the case of small sample analyses, in which sampling variability is more profound.

References

Brooks, S., & Gelman, A. (1998). Some issues in monitoring convergence of iterative simulations. *Computing Science and Statistics.*

Chib, S. (1992). Bayes inference in the Tobit censored regression model. *Journal of Econometrics, 51,* 79-99.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457-472.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall: London.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Gill, J. (2007). *Bayesian methods: A social and behavioral sciences approach* (2nd ed.). New York: Chapman and Hall/CRC.

Lee, S. Y. (2007). *Structural equation modeling: A Bayesian approach.* West Sussex: Wiley & Sons.

Press, S. J. (1989). *Bayesian statistics: Principles, models, and applications*. New York: Wiley & Sons.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Rencher, A. C. (2000). *Linear models in statistics*. New York: John Wiley & Sons Ltd.

Spiegelhalter, D. J., Thomas, A., Best, N. G., Lunn, D. (2007). *WinBUGS user manual: version 1.4.3*. Cambridge: MRC Biostatistics Unit. Online at: http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software, 12*, 1-16.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrika, 26,* 24-36

Acknowledgment

The authors wish to thank three anonymous reviewers for their helpful comments regarding earlier versions of the manuscript.

Send correspondence to: Roy Levy
Arizona State University
Email: Roy.Levy@asu.edu

Appendix

Annotated WinBUGS code for running the traditional model

```
model{

beta.0 ~ dnorm(0, .0001);              # prior for the intercept
beta.1 ~ dnorm(0, .0001);              # prior for coefficient 1
beta.2 ~ dnorm(0, .0001);              # prior for coefficient 2
tau.e ~ dgamma(.01, .01);              # prior for the error precision
sigma.e <- 1/sqrt(tau.e);          # standard deviation of the errors

for(i in 1:N){
    y.prime[i] <- beta.0 + beta.1*x1[i] + beta.2*x2[i];     # predicted value
    y[i] ~ dnorm(y.prime[i], tau.e);              # conditional distribution of y
}

}
```

Annotated WinBUGS code for running the modified model

```
model{

beta.0 ~ dunif(0, 15);              # prior for the intercept
beta.1 ~ dnorm(0, .001);      # prior for coefficient 1
beta.2 ~ dnorm(0, .001);      # prior for coefficient 2
tau.e ~ dgamma(1, 1);              # prior for the error precision
sigma.e <- 1/sqrt(tau.e);      # standard deviation of the errors

for(i in 1:N){
    y.prime[i] <- beta.0 + beta.1*x1[i] + beta.2*x2[i];     # predicted value, adjusted next
    y.prime.adj[i] <- step(y.prime[i] - 15)*15 + (1-step(y.prime[i] - 15))*y.prime[i]
    y[i] ~ dnorm(y.prime.adj[i], tau.e);              # conditional distribution of y
}

}
```