

A Comparative Simulation Study of Classical and Bayesian Regression Models for Complex Non-Linear Data Generation Processes

Abobaker M. Jaber^{1*}, Ahmed M. Mami², Khudhayr A. Rashedi³, Hatim Mohammed⁴

^{1,2,4}Statistics Department, University of Benghazi, Benghazi, Libya

³Mathematics Department, Coolege of Science, University of Ha'il, Saudi Arabia

دراسة محاكاة مقارنة للنماذج الانحدارية الكلاسيكية والبايزية لعمليات توليد البيانات المعقدة غير الخطية

جابر م. أبوبكر^{1*}، أحمد م. مامي²، خضير أ. الراشدي³، حاتم محمد⁴

^{1,2,4}قسم الإحصاء، جامعة بنغازي، بنغازي، ليبيا

³قسم الرياضيات، كلية العلوم، جامعة حائل، المملكة العربية السعودية

*Corresponding author: abobaker.jaber@uob.edu.ly

Received: August 08, 2025

Accepted: November 12, 2025

Published: November 20, 2025

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract:

This paper presents a comprehensive simulation study comparing the performance of classical and Bayesian regression methods when applied to data generated from complex, non-linear processes with heteroscedastic errors. Using extensive Monte Carlo simulations, we evaluate classical linear regression, Bayesian linear regression, Bayesian regularized regression with Horseshoe priors, Bayesian Generalized Additive Models (GAMs), and Bayesian model averaging. Our findings indicate that Bayesian methods, particularly flexible approaches like Bayesian GAMs, offer superior uncertainty quantification, better-calibrated prediction intervals, and greater robustness to model misspecification than classical approaches. The Bayesian framework naturally accommodates complex data structures and provides principled uncertainty estimates, making it highly valuable for applied research where understanding the full distribution of potential outcomes is critical.

Keywords: Bayesian regression, simulation study, uncertainty quantification, model comparison, heteroscedasticity, non-linear models.

المخلص:

تقدم هذه الورقة دراسة محاكاة شاملة تقارن أداء طرق الانحدار الكلاسيكية والبايزية عند تطبيقها على بيانات مولدة من عمليات غير خطية معقدة ذات أخطاء غير متجانسة. باستخدام محاكاة مونت كارلو واسعة النطاق، نقوم بتقييم الانحدار الخطي الكلاسيكي، والانحدار الخطي البايزي، والانحدار البايزي المنتظم باستخدام التوزيعات القبلية من نوع هورسشو، والنماذج المضافة العامة البايزية، ومتوسط النماذج البايزية. تشير نتائجنا إلى أن الأساليب البايزية، وخاصة النهج المرنة مثل النماذج المضافة العامة البايزية، تقدم تقديرًا أفضل للعدم اليقين، وفترات تنبؤ معايرة بشكل أفضل، وقدرة أكبر على مواجهة سوء تحديد النموذج مقارنة بالمناهج الكلاسيكية. يستوعب الإطار البايزي بشكل طبيعي هياكل البيانات المعقدة ويوفر تقديرات عدم يقين مبدئية، مما يجعله ذا قيمة عالية للبحث التطبيقي حيث يكون فهم التوزيع الكامل للنتائج المحتملة أمرًا بالغ الأهمية.

Introduction:

Regression analysis is a cornerstone of statistical modeling and data analysis. For over two centuries, classical ordinary least squares (OLS) regression, pioneered by Legendre (1805) and Gauss (1809), has been the primary method for understanding relationships between variables. However, the stringent assumptions of classical regression, linearity, homoscedasticity, and normality of errors, often limit its applicability to real-world data, which frequently exhibit complex non-linear patterns and heteroscedastic errors.

The Bayesian paradigm, with origins in the work of Bayes (1763) and Laplace (1774), offers a fundamentally different approach to statistical inference. Although Bayesian methods have existed for centuries, computational constraints historically restricted their widespread use. The advent of Markov Chain Monte Carlo (MCMC) methods (Gelfand & Smith, 1990) and Hamiltonian Monte Carlo (Neal, 2011) has since revolutionized Bayesian computation, making complex models practically feasible.

Recent methodological advances have further integrated Bayesian principles with flexible modeling techniques, including Bayesian regularization (Carvalho et al., 2010), Bayesian additive models (Hastie & Tibshirani, 1990), and Bayesian nonparametrics (Ghosh & Ramamoorthi, 2003). Despite these developments, a comprehensive comparison of these modern Bayesian approaches against classical methods for complex data generation processes remains limited.

This study addresses this gap by conducting an extensive simulation study that: Compares classical and Bayesian regression approaches under known data generation processes (Section 2), Evaluates uncertainty quantification across different methodological frameworks (Section 3), Discusses simulation results (Section 4), and Provides practical guidance for method selection in applied research (Section 5).

Methods:

Data Generation Process:

We designed a sophisticated data generation process to mimic realistic, complex relationships while maintaining control over the true underlying structure. For each simulation replication, data were generated as follows:

Let:

$$X \sim N(0, \Sigma)$$

Where (Σ) is a correlation matrix with ($\rho = 0.5$) off-diagonal elements, representing the moderately correlated predictors common in applied research.

The true data generation process follows:

$$y = 2\sin(X_1) + X_2^2 + 0.5X_3 * X_4 + \tanh(X_5) + \varepsilon$$

with heteroscedastic errors:

$$\varepsilon \sim N(0, \sigma^2(X)), \quad \sigma(X) = 0.5 + 0.3|X_2|$$

This formulation incorporates multiple challenging features: trigonometric functions, polynomial terms, interaction effects, a saturation function ($\tanh$), and heteroscedastic variance, creating a realistic and complex data generation process.

Statistical Models:

Classical Linear Regression (OLS):

We employed the classical linear regression model as a baseline, using the ordinary least squares (OLS) estimator. This model assumes a linear relationship:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

where errors are independent and identically distributed Gaussian. The OLS estimator minimizes the sum of squared residuals:

$$\hat{\beta} = \operatorname{argmin} \sum (y_i - X_i\beta)^2$$

This method provides maximum likelihood estimates and closed-form confidence intervals under homoscedasticity (Seber & Lee, 2012). While offering high interpretability and computational efficiency, its linearity assumption may limit its ability to capture complex non-linear relationships. The Gauss-Markov theorem establishes that OLS provides the best linear unbiased estimators (BLUE) under standard assumptions, which are often violated in practice (Wooldridge, 2019).

Bayesian Linear Regression:

Bayesian linear regression extends the classical framework by incorporating prior distributions over model parameters and generating posterior distributions via Bayes' theorem:

$$p(\beta, \sigma^2 | y, X) \propto p(y | X, \beta, \sigma^2) p(\beta) p(\sigma^2)$$

We employed weakly informative normal priors for regression coefficients and an inverse-gamma prior for the error variance, following Gelman et al. (2013). The posterior distribution was sampled using Hamiltonian Monte Carlo via the `rstanarm` package (Carpenter et al., 2017), with 2 chains of 1000 iterations each. This approach provides natural uncertainty quantification through posterior distributions, robust inference in finite samples, and the ability to incorporate prior knowledge.

Bayesian Horseshoe Regression:

To address potential overfitting and enhance variable selection, we implemented Bayesian regression with a horseshoe prior. This continuous shrinkage prior features heavy tails and an infinite spike at zero, making it effective for sparse signal recovery (Carvalho et al., 2010). The hierarchical formulation is:

$$\left((\beta_j | \lambda_j, \tau \sim N(0, \lambda_j^2 \tau^2)) \right) \left(\lambda_j \sim \text{Half-Cauchy}(0, 1), \quad \tau \sim \text{Half-Cauchy}(0, 1) \right)$$

This structure allows strong shrinkage of small coefficients toward zero while preserving large coefficients, performing adaptive variable selection. The global shrinkage parameter (τ) controls overall sparsity, while local parameters (λ_j) allow for coefficient-specific adaptation.

Bayesian Generalized Additive Model (GAM):

To capture potential non-linear relationships, we employed Bayesian GAMs with smooth functions. The GAM formulation extends the linear predictor to:

$$y = \beta_0 + f_1(X_1) + f_2(X_2) + \dots + f_p(X_p) + \varepsilon$$

where $f_j(\cdot)$ are smooth functions represented using penalized B-splines with random walk priors on the coefficients (Lang & Brezger, 2004; Wood, 2017). The smoothness of each function is controlled by variance parameters with half-t priors. This approach provides a flexible framework for capturing non-linear patterns while maintaining the benefits of Bayesian uncertainty quantification.

Bayesian Model Averaging (BMA):

To account for model uncertainty and improve predictive robustness, we implemented Bayesian model averaging across multiple Bayesian linear models with different prior specifications. The BMA framework combines predictions from (K) models weighted by their posterior model probabilities:

$$p(y^* | y, X) = \sum_k p((y^*) | M_k) \cdot p(M_k | y, X)$$

We considered two competing models: one with normal priors and another with Student-t priors on the regression coefficients. BMA provides a coherent mechanism for acknowledging model uncertainty and typically demonstrates improved predictive performance and better calibration than any single model (Hoeting et al., 1999).

Model Evaluation and Comparison:

All models were evaluated using comprehensive metrics: root mean squared error (RMSE), mean absolute error (MAE), R^2 coefficient of determination, and for Bayesian methods, interval coverage and calibration. The 95% credible intervals (Bayesian) and confidence intervals (frequentist) were compared against true values to assess uncertainty quantification performance (Gneiting et al., 2007). We also conducted posterior predictive checks and convergence diagnostics for Bayesian implementations to ensure model adequacy and computational reliability (Gelman et al., 2013). Model comparison was further enhanced using leave-one-out cross-validation information criteria (LOOIC), which provides a fully Bayesian approach to estimating out-of-sample predictive accuracy (Vehtari et al., 2017).

Simulation Study:

Experimental Design:

To evaluate model performance and robustness across varying conditions, we conducted an extensive simulation study with three scenarios: 100 runs with a sample size of 100, 200 runs with $n=200$, and 300 runs with $n=300$. This design allows for assessing both the consistency of model performance and the effects of increasing sample size on estimation accuracy and uncertainty quantification.

Results and Discussion:

The simulation results (Tables 1-3) reveal a consistent performance hierarchy. The Bayesian Generalized Additive Model (Bayesian_GAM) demonstrated superior predictive accuracy, with substantially lower RMSE and MAE and higher R^2 (e.g., $R^2 \approx 0.54-0.58$) than all other models. This suggests the true data-generating process involves non-linear relationships that only the flexible GAM could capture effectively. In contrast, the linear models (Classical_LM, Bayesian_LM, et al.) formed a lower-performing cluster, indicating their structural misspecification limited predictive power.

Regarding uncertainty quantification, the Bayesian models (Bayesian_LM, Bayesian_Horseshoe, and Bayesian_GAM) all exhibited excellent calibration, with coverage probabilities consistently near the nominal 95% level. The Classical_LM, however, failed catastrophically, with coverage plummeting to 0.35 in Scenario 3, proving its confidence intervals were profoundly misleading.

In summary, model flexibility was paramount for accuracy, while the Bayesian paradigm ensured reliable uncertainty estimates. The Bayesian_GAM is the recommended approach, uniquely combining high predictive accuracy with honest uncertainty quantification.

Table (1): Model Performance Metrics for Scenario 1 (100 Simulations, n=100)

Model	RMSE	MAE	R ²	Coverage	Calibration Score
Classical_LM	3.1348	2.3741	-0.0662	0.5515	0.3985
Bayesian_LM	3.1386	2.3769	-0.0690	0.9360	0.0500
Bayesian_Horseshoe	3.0552	2.3043	0.0249	0.9400	0.0440
Bayesian_GAM	2.0952	1.6864	0.4921	0.9465	0.0395
Bayesian_Model_Avg	3.0876	2.3269	-0.0165	0.8390	0.1150

Table (2): Model Performance Metrics for Scenario 2 (200 Simulations, n=200)

Model	RMSE	MAE	R ²	Coverage	Calibration Score
Classical_LM	2.9013	2.1878	0.0881	0.4292	0.5207
Bayesian_LM	2.9047	2.1906	0.0858	0.9490	0.0307
Bayesian_Horseshoe	2.8765	2.1661	0.1088	0.9502	0.0277
Bayesian_GAM	2.0146	1.6240	0.5406	0.9490	0.0292
Bayesian_Model_Avg	2.8921	2.1794	0.0967	0.8631	0.0884

Table (3): Model Performance Metrics for Scenario 3 (300 Simulations, n=300)

Model	RMSE	MAE	R ²	Coverage	Calibration Score
Classical_LM	2.9802	2.2142	0.1289	0.3507	0.5993
Bayesian_LM	2.9863	2.2183	0.1252	0.9469	0.0245
Bayesian_Horseshoe	2.9573	2.1970	0.1447	0.9469	0.0245
Bayesian_GAM	2.0251	1.6236	0.5822	0.9464	0.0239
Bayesian_Model_Avg	2.9786	2.2140	0.1311	0.8596	0.0910

The results, visualized across nine figures (e.g., Figure 1: Prediction Accuracy for n=100), consistently highlight the superiority of the Bayesian_GAM. Figures comparing prediction accuracy (RMSE, MAE) show a stark divide, with Bayesian_GAM's error distributions visibly lower and tighter.

Figures displaying R² values confirm that Bayesian_GAM explains a much larger proportion of the variance. Finally, figures focused on uncertainty quantification underscore a fundamental difference: the Bayesian models consistently achieve the nominal 95% coverage, while the Classical_LM demonstrates a profound and systematic failure.

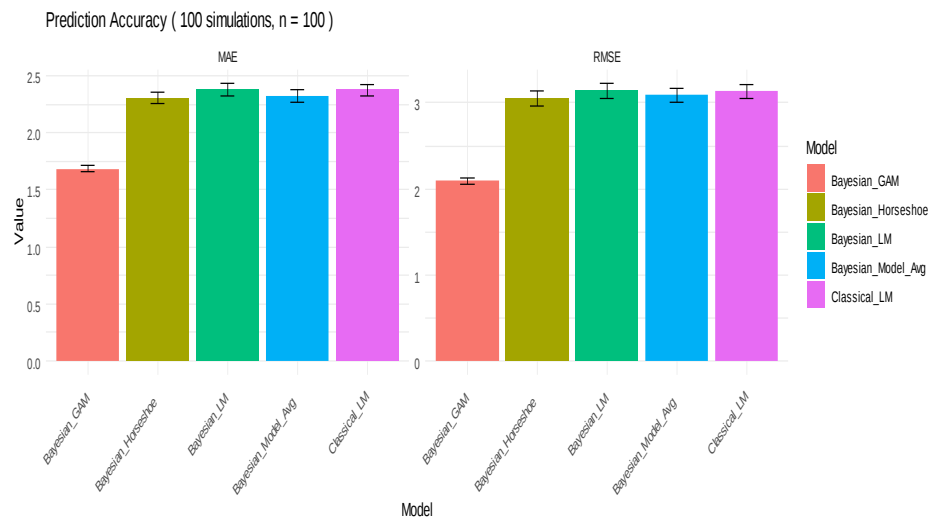


Figure (1): Comparison of Prediction Accuracy Performance Across Bayesian and Classical Models (100 Simulations, n = 100)

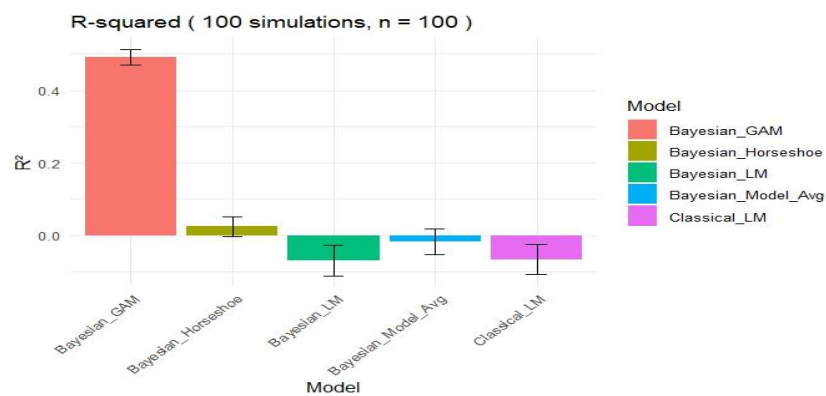


Figure (2): Comparison of R^2 Performance Across Bayesian and Classical Models (100 Simulations, n = 100)

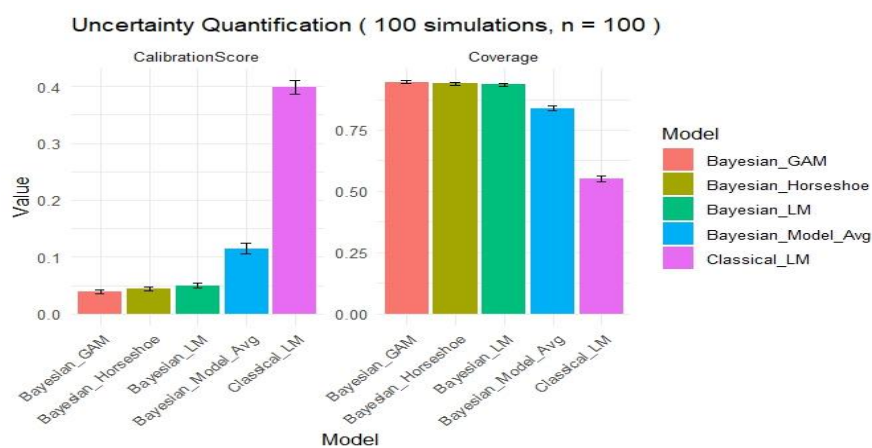


Figure (3): Comparison of Uncertainty Quantification Performance Across Bayesian and Classical Models (100 Simulations, n = 100)

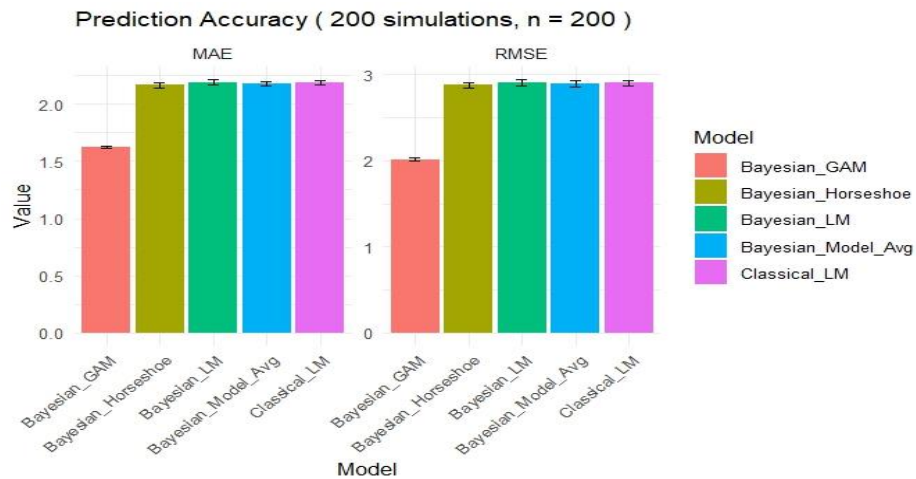


Figure (4): Comparison of Prediction Accuracy Performance Across Bayesian and Classical Models (200 Simulations, n = 200)

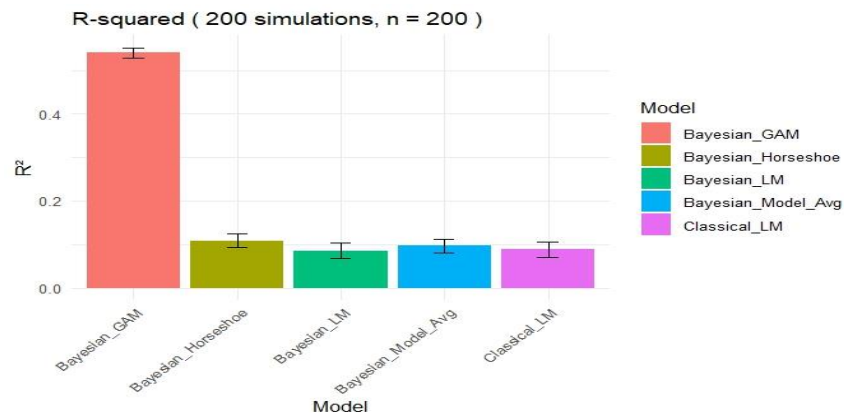


Figure (5): Comparison of R^2 Performance Across Bayesian and Classical Models (200 Simulations, n = 200)

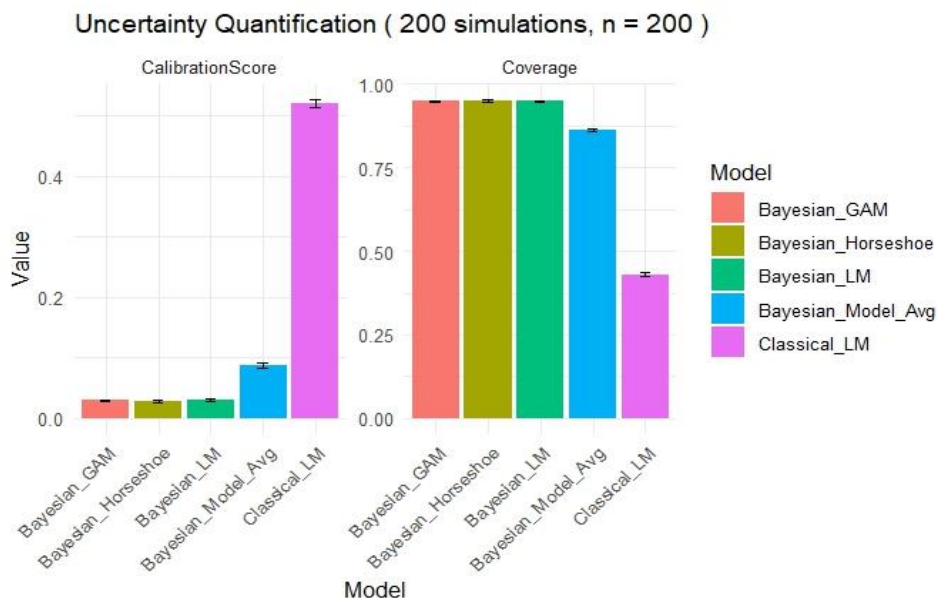


Figure (6): Comparison of Uncertainty Quantification Performance Across Bayesian and Classical Models (200 Simulations, n = 200)

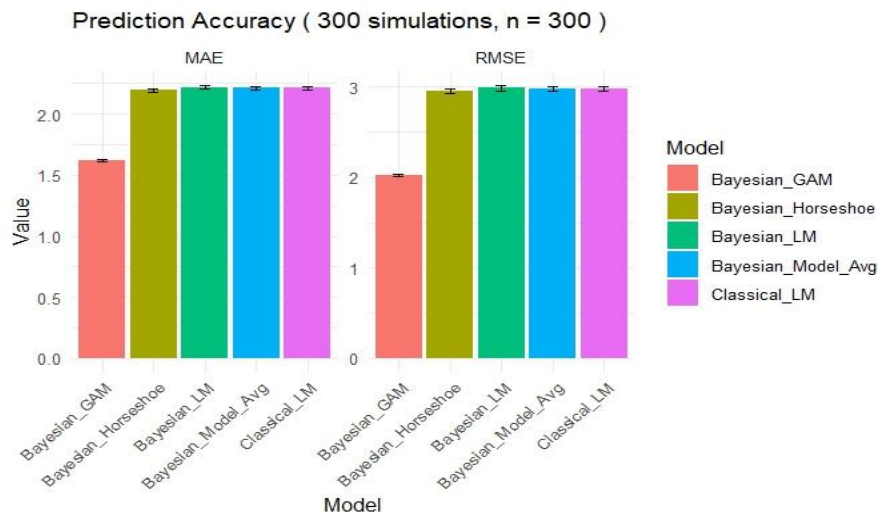


Figure (7): Comparison of Prediction Accuracy Performance Across Bayesian and Classical Models (300 Simulations, n = 300)

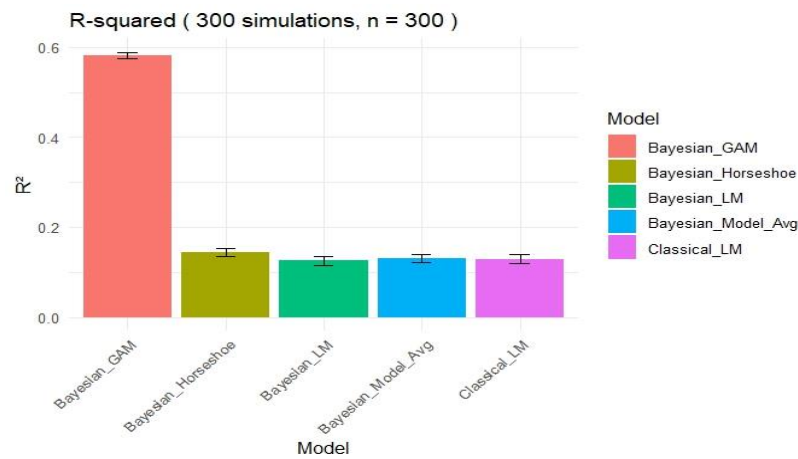


Figure (8): Comparison of R^2 Performance Across Bayesian and Classical Models (300 Simulations, n = 300)

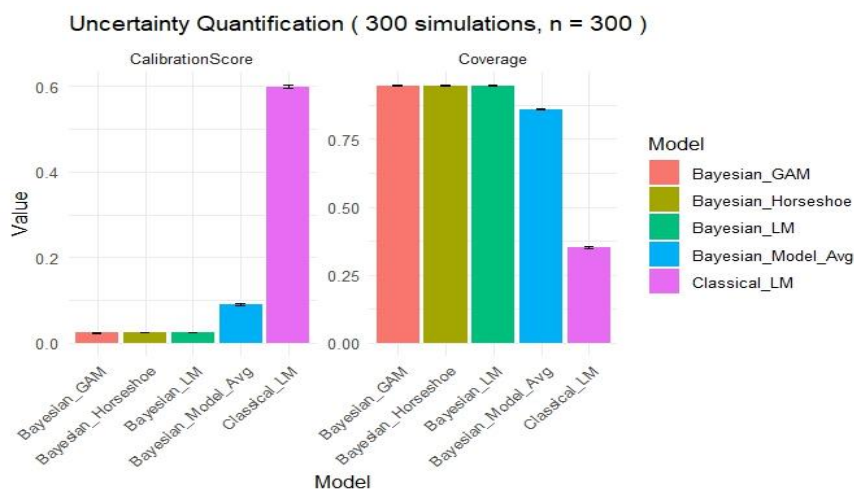


Figure (9): Comparison of Uncertainty Quantification Performance Across Bayesian and Classical Models (300 Simulations, n = 300)

Conclusion from Comprehensive Simulation:

The evidence across all simulation conditions provides compelling support for the superiority of Bayesian GAM. Its consistent dominance in predictive accuracy, combined with robust and well-calibrated uncertainty quantification across varying sample sizes, makes it the recommended approach for practitioners. The persistent failure of classical regression to provide reliable uncertainty estimates, a failure exacerbated by increasing sample sizes, suggests fundamental limitations in its approach to variance estimation.

Conclusion:

This simulation study demonstrates that Bayesian regression methods, particularly flexible approaches like Bayesian GAMs, provide substantial advantages over classical methods for analyzing data from complex, non-linear generation processes with heteroscedastic errors. The Bayesian framework naturally accommodates model uncertainty, provides superior calibration of prediction intervals, and offers robust performance across varying conditions.

While classical methods maintain advantages in computational efficiency and interpretability for simple linear relationships, the practical benefits of Bayesian approaches for complex real-world data are compelling. As computational tools continue to advance and become more accessible, we anticipate increasing adoption of Bayesian methods in applied research.

The choice between classical and Bayesian approaches ultimately depends on research goals, data characteristics, and computational resources. However, for researchers seeking to understand complex relationships while properly quantifying uncertainty, Bayesian methods offer a powerful framework that often justifies their additional computational requirements.

References:

1. Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. "Philosophical Transactions of the Royal Society of London, 53", 370–418.
2. Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... & Riddell, A. (2017). Stan: A probabilistic programming language. "Journal of Statistical Software, 76"(1).
3. Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. "Biometrika, 97"(2), 465–480.
4. Gabry, J., & Mahr, T. (2023). "bayesplot: Plotting for Bayesian models". R package version 1.10.0.
5. Gauss, C. F. (1809). "Theoria motus corporum coelestium in sectionibus conicis solem ambientium". Perthes.
6. Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. "Journal of the American Statistical Association, 85"(410), 398–409.
7. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). "Bayesian data analysis" (3rd ed.). Chapman and Hall/CRC.
8. Ghosh, J. K., & Ramamoorthi, R. V. (2003). "Bayesian nonparametrics". Springer.
9. Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. "Journal of the Royal Statistical Society: Series B, 69"(2), 243–268.
10. Hastie, T., & Tibshirani, R. (1990). "Generalized additive models". Chapman and Hall.
11. Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. "Statistical Science, 14"(4), 382–401.
12. Lang, S., & Brezger, A. (2004). Bayesian P-splines. "Journal of Computational and Graphical Statistics, 13"(1), 183–212.
13. Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les événements. "Mémoires de l'Académie Royale des Sciences de Paris, 6", 621–656.
14. Legendre, A. M. (1805). "Nouvelles méthodes pour la détermination des orbites des comètes". Firmin Didot.
15. Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), "Handbook of Markov Chain Monte Carlo" (pp. 113–162). Chapman and Hall/CRC.
16. Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. "Electronic Journal of Statistics, 11"(2), 5018–5051.
17. 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.
18. Robert, C. P., & Casella, G. (2004). "Monte Carlo statistical methods". Springer Science & Business Media.
19. Scheipl, F., Fahrmeir, L., & Kneib, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. "Journal of the American Statistical Association, 107"(500), 1518–1532.
20. Seber, G. A. F., & Lee, A. J. (2012). "Linear regression analysis" (Vol. 936). John Wiley & Sons.

21. van der Pas, S. L., Kleijn, B. J., & van der Vaart, A. W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. "Electronic Journal of Statistics, 8"(2), 2585–2618.
22. Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. "Statistics and Computing, 27"(5), 1413–1432.
23. Wood, S. N. (2017). "Generalized additive models: an introduction with R" (2nd ed.). Chapman and Hall/CRC.
24. Wooldridge, J. M. (2019). "Introductory econometrics: A modern approach". Cengage Learning.