Predicting Systolic Blood Pressure of Pregnant Women

FURMAN UNIVERSITY

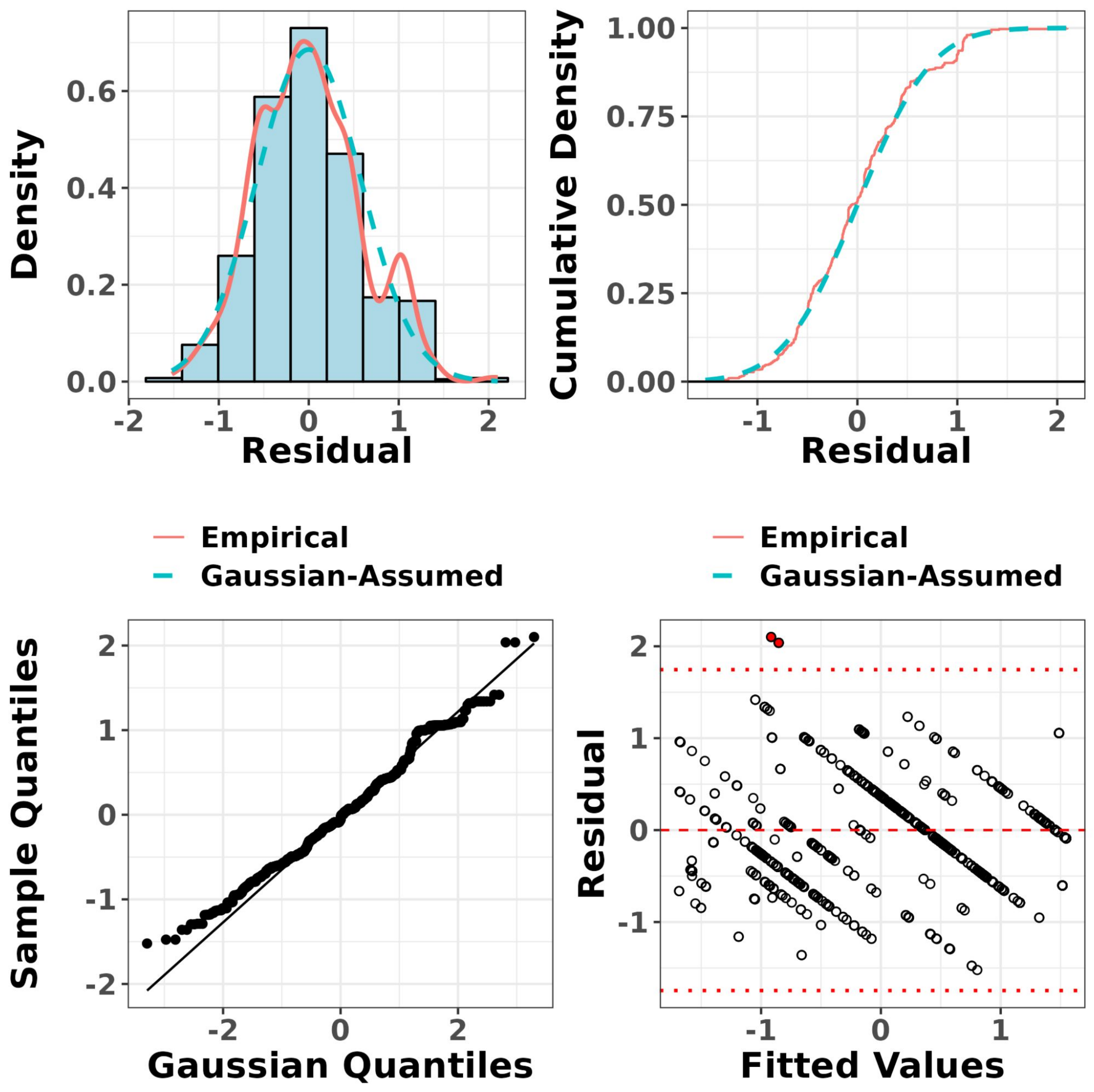
Victoria Fenwick, Helen Jacobson, Jack Roberts

Department of Mathematics, Furman University, Greenville, SC

Introduction

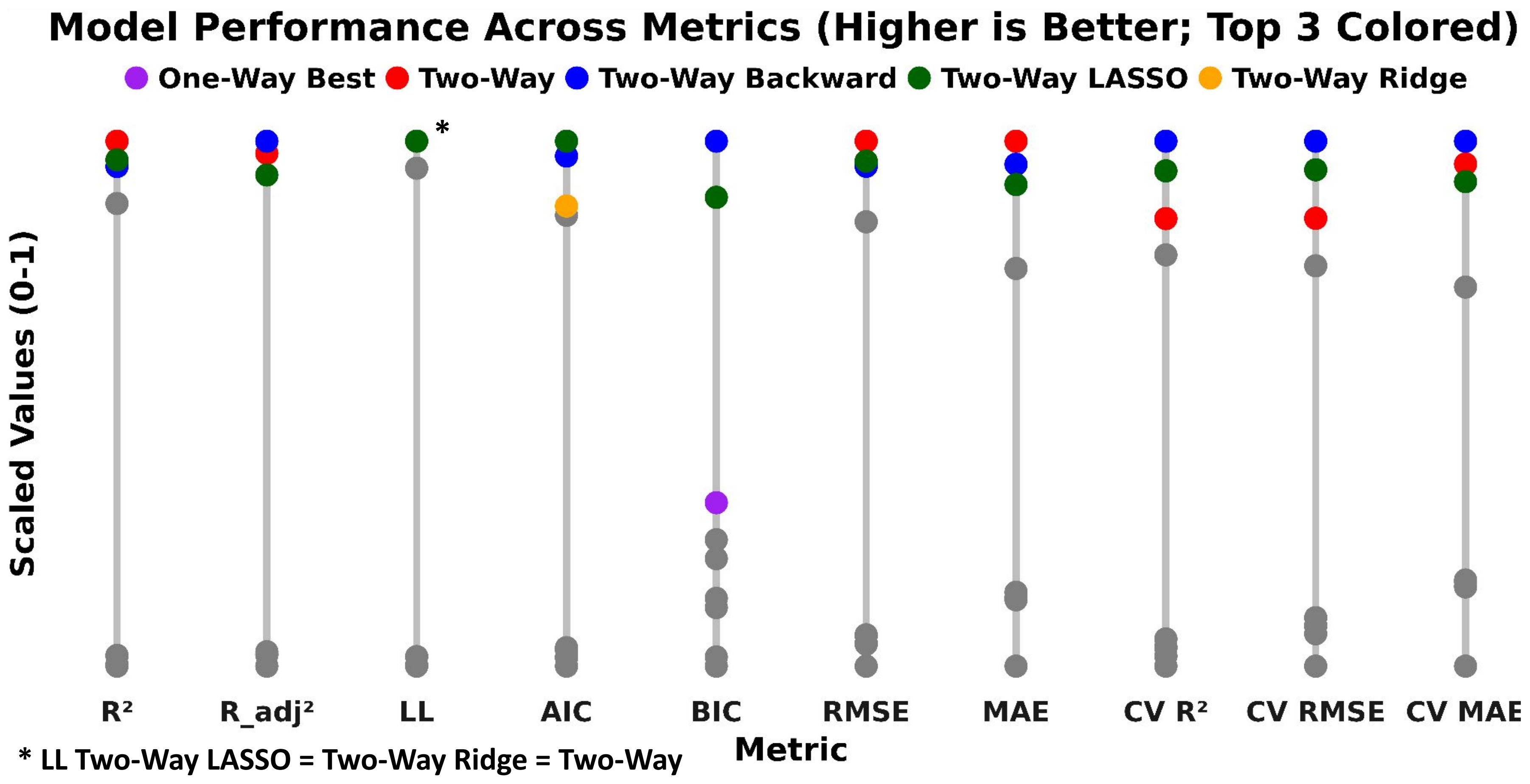
According to the American Heart Association, hypertension during pregnancy is the second leading cause of maternal death worldwide. In the United States alone, almost 15% of maternal deaths are due to hypertension. High systolic blood pressure is a key component of hypertension, and accurate prediction and management of systolic blood pressure levels can significantly reduce the risk of hypertension-related complications during pregnancy.

The purpose of this study is to determine the individual variables and their interactions which most accurately predict systolic blood pressure. We used a linear regression model to identify these variables.



Data

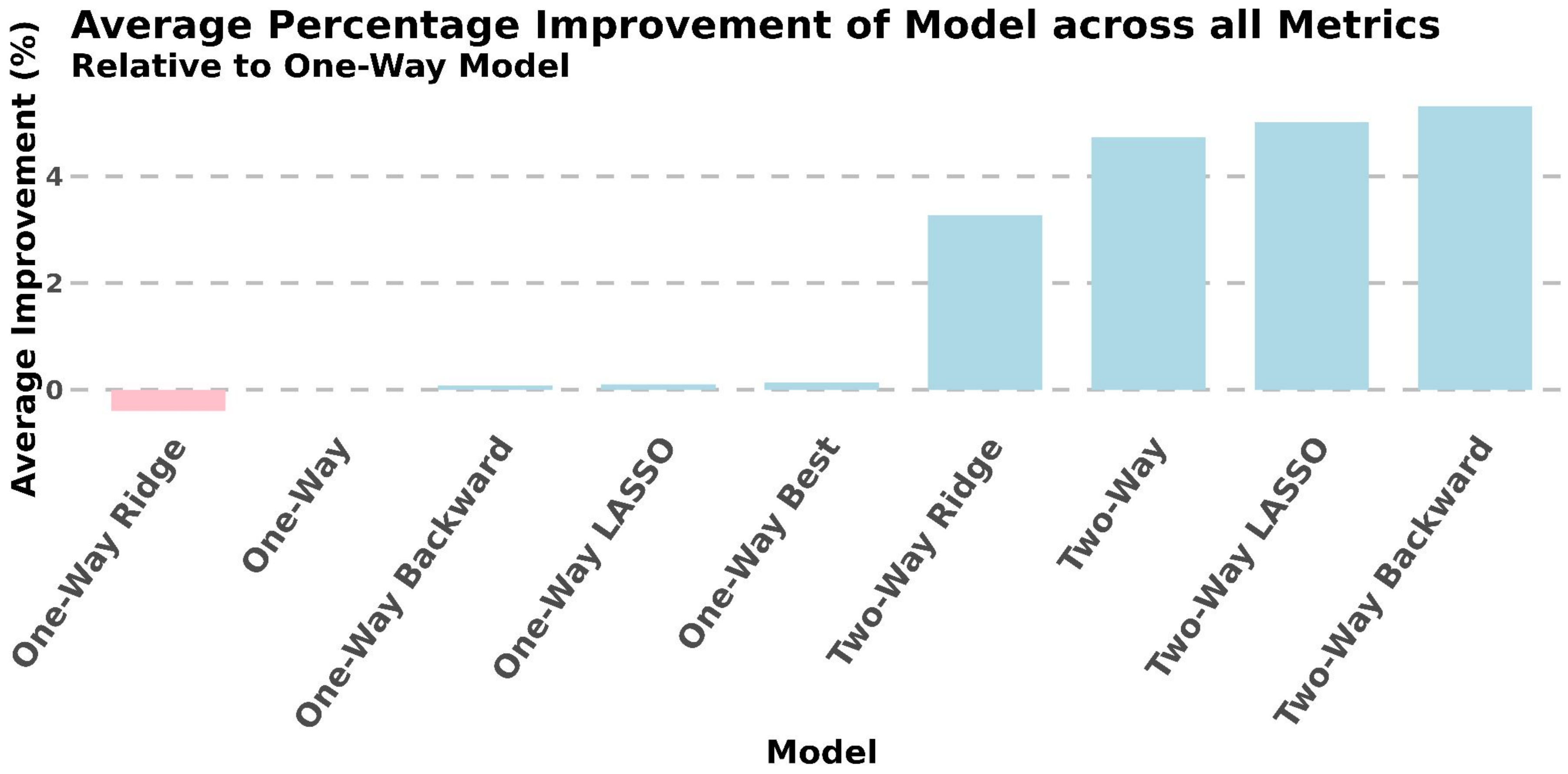
Our data, sourced from Kaggle, was collected in 2023 via a random sample of 1,206 pregnant women at rural hospitals and clinics in Bangladesh. The predictor variables include age (years), diastolic blood pressure (mmHg), blood sugar (mmol/L), heart rate (bpm), body temperature (°F), and risk level. Our response variable is systolic blood pressure (mmHg). Risk level is the only categorical variable and spans low, medium, and high. Because assumptions of normality and constant variance were met (see chart above), no transformations were applied. Lastly, we performed standardization and minor data cleaning.



Model Selection

To ensure predictive accuracy and robustness, we conducted a thorough model selection process. In addition to individual variables ("one-way"), we considered the interactions between variables ("two-way") in case the importance of a variable depended upon the presence of another. To improve accuracy and reduce overfitting, we applied feature selection techniques to these variables, including forward selection, backward elimination, stepwise regression, LASSO regression, Ridge regression, and best-subset regression. This process resulted in a total of thirteen models, with two sets of three duplicates. We discarded two of each of the duplicates and omitted two-way best-subset regression due to computational complexity. The nine remaining models were then evaluated across the ten measurements of performance ("metrics") in the chart above.

These metrics were chosen to provide a holistic view of model performance, encapsulating variance explanation, observation likelihood, and predictive accuracy. Leave-One-Out Cross-Validation ("CV") metrics were valued the most as they best reflect real-world performance. The two-way backward elimination model excelled in the three CV metrics and also had the highest average improvement across all metrics at +5.31% (see chart below). Finally, we ranked each model from one to nine in each metric, and averaged the ranks. Again, the two-way backward elimination model was the best with an average rank of 1.9. These three factors led us to concluding the two-way backward elimination model to be the best for our data.



Results

Our research found that the two-way backwards regression model is the most predictive and accurate model. It identifies diastolic blood pressure (t-value = 13.2154), low risk level (t-value = -5.4471), and the interaction between diastolic blood pressure and body temperature (t-value = 5.6983) as the most significant predictors. Conversely, the least significant predictors were blood sugar (p-value = 0.4504), age (p-value = 0.3320), and the interaction between age and low risk level (p-value = 0.6486).

These findings suggest that traditional indicators such as age and blood sugar may have limited predictive value for systolic blood pressure. Conversely, diastolic blood pressure, risk level, and body temperature appear to be crucial factors. These insights could be pivotal in developing more targeted and effective approaches to managing hypertension, leading to safer pregnancies. Note that due to our data's demographics, results may not hold for all pregnancies.

Future Research

For further research, an ideal dataset would include additional parameters such as patient weight, number of births, and race. We would also like to consider patients of other countries and backgrounds to generalize our findings. Lastly, using a best subset regression on interaction parameters and selecting parameters based on average improvement rather than AIC would refine our model.

Acknowledgements

We would like to thank Dr. Roy Bower for his facilitation of this project and mentorship. In addition, we would like to thank the Furman University Mathematics Department.