From Black Box to Shining Spotlight:

Using Random Forest Prediction Intervals to Illuminate the Impact of Modeling Assumptions in Linear Regression

Andrew J. Sage, Yang Liu, and Joe Sato

Lawrence University
Appleton, Wisconsin

Background

Two new approaches for constructing random forest prediction intervals have been introduced in recent statistics literature.

- Approach of Zhang et al. (2019) relies on symmetry and constant variance assumptions.
- More flexible approach of Lu & Hardin (2021) does not depend on these assumptions.
- Random forests do not depend on linearity assumption, as linear regression models do.

Random forests are often taught in statistical learning classes, separate from traditional modeling techniques.

Real Data App

Figure 1: Real data app applied to Ames Housing dataset

Our Study

We created two Shiny Apps to visualize random forest prediction intervals alongside those produced by regression models.

Users simulate data under specified conditions and compare/contrast prediction intervals produced by regression models and the two random forest approaches. Apps can also be used on real data.

Apps facilitate integration of random forests into undergraduate statistical modeling courses, so that students can compare and contrast regression models with more flexible alternatives.

Simulation App

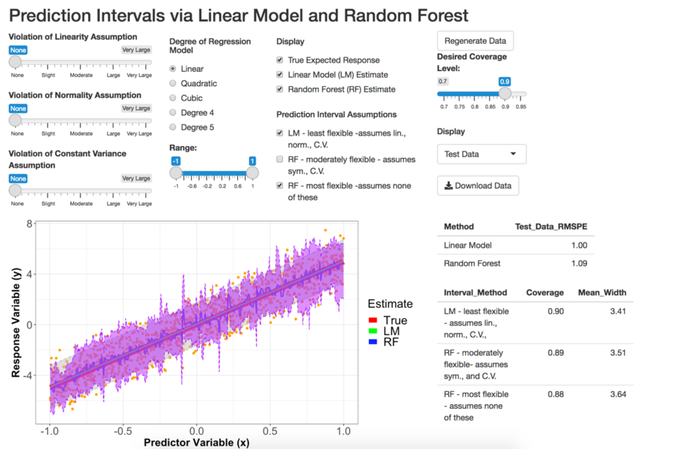
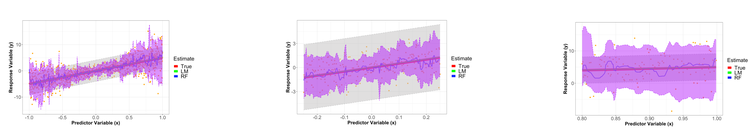


Figure 2: Simulation app and interface

Non-constant Variance Scenario



-1≤ x ≤ 1 -0.2≤ x ≤ 0.2 0.8≤ x ≤ 1

Learning Outcome and Hypothesis

The apps were implemented and assessed in 2 sections of 200-level statistical modeling course at a liberal arts college.

Learning Outcome: Explain the effect of violations of linear regression assumptions on predictions and prediction intervals.

Hypothesis: Comparing prediction intervals produced by regression models to those produced by more flexible techniques helps students understand the impact of assumptions in regression.

Results and Conclusions

A survey was administered before and after students completed a lab activity using the labs.

Questions:

1. If you are primarily interested in using a linear regression model to make accurate predictions, how concerned would you be about a violation of the (insert) assumption?
2. If you are primarily interested in using a linear regression model to obtain reliable prediction intervals, how concerned would you be about a violation of the (insert) assumption?

Figure 3: Pre and post survey results

- **Pro:** Students showed more careful discernment in post-survey.
- **Con:** Some students became less concerned about violations when they should be.

Our results suggest that integrating new techniques in undergraduate statistical modeling classes, for the purpose of comparison, can help students think critically about assumptions behind models traditionally taught in those classes.

References

[1] Andrew J Sage, Yang Liu, and Joe Sato. From black box to shining spotlight: Using random forest prediction intervals to illuminate the impact of assumptions in linear regression. *The American Statistician*, 76(4):414–429, 2022.

[2] Benjamin Lu and Johanna Hardin. A unified framework for random forest prediction error estimation. *The Journal of Machine Learning Research*, 22(1):386–426, 2021.

[3] Haozhe Zhang, Joshua Zimmerman, Dan Nettleton, and Daniel J Nordman. Random forest prediction intervals. *The American Statistician*, 2019.