### Hope College

### Hope College Digital Commons

23rd Annual A. Paul and Carol C. Schaap Celebration of Undergraduate Research and Creative Activity (2024)

The A. Paul and Carol C. Schaap Celebration of Undergraduate Research and Creative Activity

4-12-2024

### Examining Regression Assumption Violations in Machine Learning Models Using the Wisconsin Longitudinal Study Dataset

Grace Mooney Anderson *Hope College* 

Melia Brewer Hope College

Follow this and additional works at: https://digitalcommons.hope.edu/curca\_23

Part of the Psychology Commons

### **Recommended Citation**

**Repository citation**: Anderson, Grace Mooney and Brewer, Melia, "Examining Regression Assumption Violations in Machine Learning Models Using the Wisconsin Longitudinal Study Dataset" (2024). 23rd Annual A. Paul and Carol C. Schaap Celebration of Undergraduate Research and Creative Activity (2024). Paper 3.

https://digitalcommons.hope.edu/curca\_23/3

April 12, 2024. Copyright © 2024 Hope College, Holland, Michigan.

This Poster is brought to you for free and open access by the The A. Paul and Carol C. Schaap Celebration of Undergraduate Research and Creative Activity at Hope College Digital Commons. It has been accepted for inclusion in 23rd Annual A. Paul and Carol C. Schaap Celebration of Undergraduate Research and Creative Activity (2024) by an authorized administrator of Hope College Digital Commons. For more information, please contact digitalcommons@hope.edu, barneycj@hope.edu.

### **Hope Examining Regression Assumption Violations in Machine Learning Models** COLLEGE **Using the Wisconsin Longitudinal Study Dataset**

## Introduction

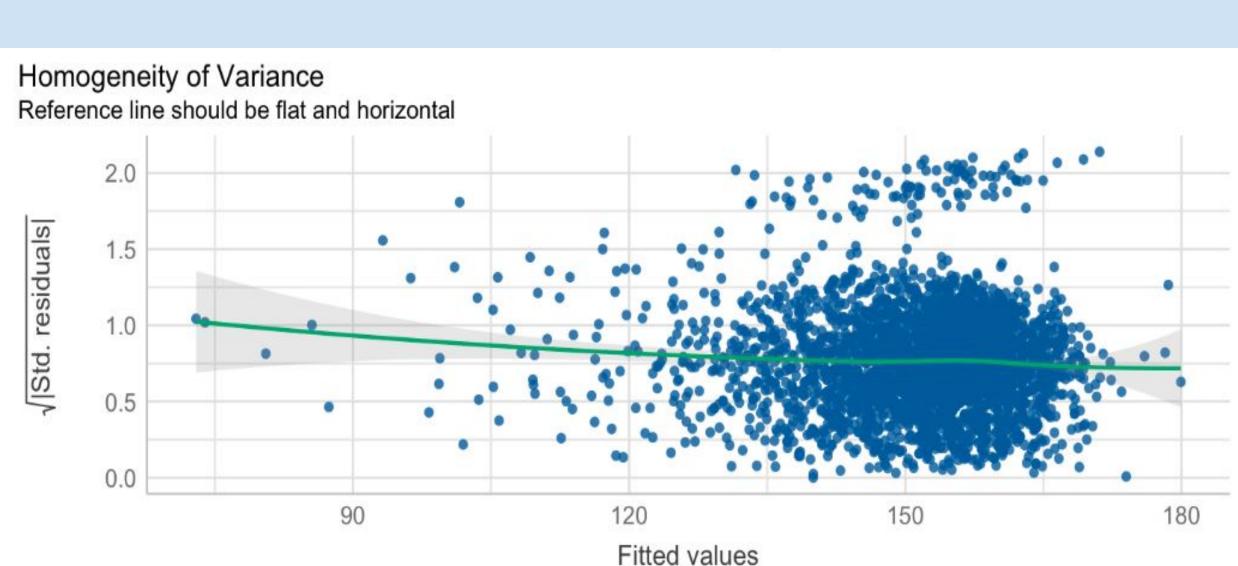
- Machine learning (ML) is becoming increasingly relevant in the social sciences
- Many who use ML models do not verify the assumptions of linear regression (Yarkoni & Westfall, 2017)
- We will use a large dataset (N>2,000) and replicate the findings of an accompanying study and replicate these findings using ML
- We will then simulate a dataset
- We hypothesized that the reliability of ML is dampened with the presence of these violations
  - For instance, the probability of Type I errors increasing when heteroscedasticity exists

## Method

- Utilized data from the Wisconsin Longitudinal Study (WLS)
- Replicated findings from Clark and Lee (2021) which looked into how both early- and later-life variables correlate with later-life subjective well-being using ordinary least squares (OLS) linear regression
- Utilized three supervised learning models:
  - Regularized regression
  - Support vector machine
  - Random forest
- Monte Carlo simulation used with N = 3000 and replications ranging from 20-1000 (depending on the model) to create a dataset with five predictors (X1-X5) and one outcome variable (Y)
- Unstandardized coefficients (see Table) significant at p < .1
- All analyses were performed in R

### Grace Mooney Anderson, Melia Brewer, & Robert D. Henry (faculty mentor)

### Results

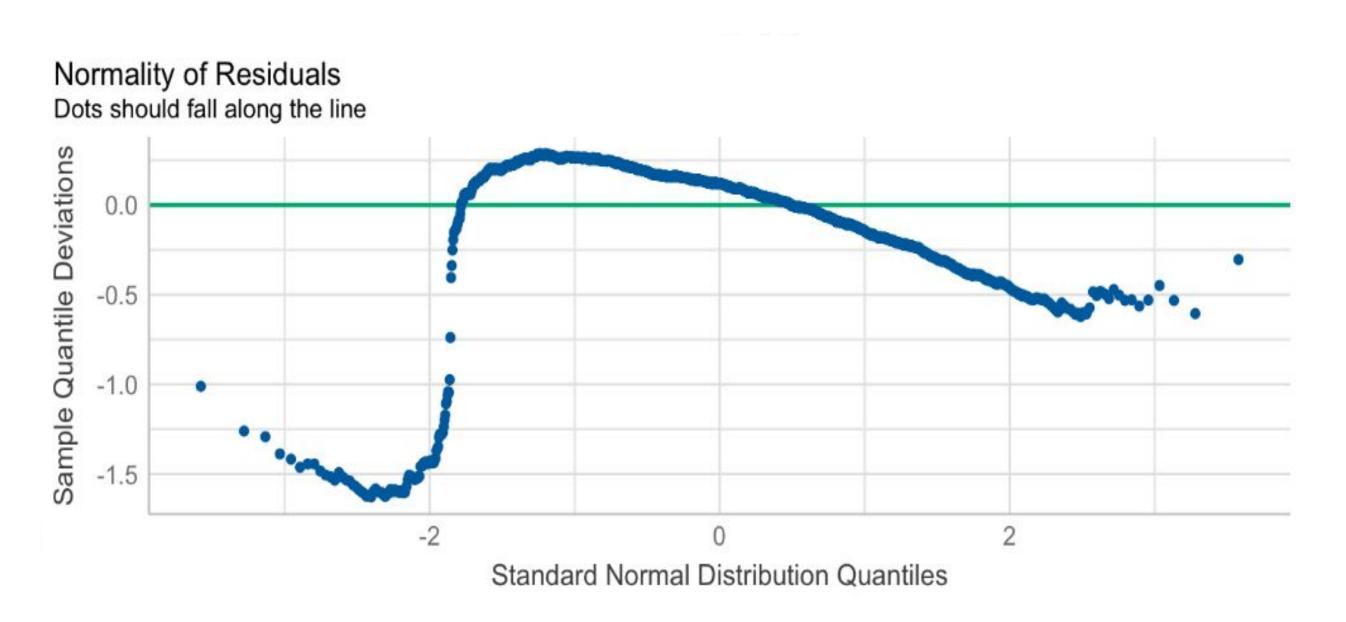


Within the WLS dataset, we found violations of the assumptions of homoscedasticity (above) and normality of residuals (right). All other assumptions were met.

Clark and Lee (2021)	Linear Regression <sup>*</sup>	Ŭ	Support Vector Machine <sup>^</sup>	Random Forest^
Mental Health	-0.7	-9.9	100	100
Social Participation	0.6	2.4	17.1	32.8
Education	1.0	1.5	10.1	15.8
Physical Health	25.6	0.2	1.9	5.6
Never Married	-3.3	-0.5	1.4	0.7
IQ	-0.1	-0.1	2.1	36.3
Female	4.7	1.3	0.5	5.1
Number of Siblings		-0.3	0.6	19.2
Single Parent Household		0.03	0.2	1.5
Mom Age at Birth		-0.1		23.8
Retired		-0.1		4.7
Separated		0.1		3.1
Simulated Dataset	Linear Regression <sup>*</sup>	Ŭ	Support Vector Machine <sup>^</sup>	Random Forest^
X1 (0.5)	0.8	0.4	46.6	38.5
X2 (0.6)	0.7	0.5	60.7	25.9
X3 (0.3)		0.2	19.1	20.6
X4 (0.1)		0.05	5.0	3.5
X5 (0.8)	1.1	0.6	82.5	99.8

\* unstandardized coefficients | ^ standardized variable importance (0-100)





## Discussion

- assumption violations
- of models

# References

Clark, A.E. & Lee, T. (2021). Early-life correlates of later-life well-being: Evidence from the Wisconsin Longitudinal Study. Journal of Economic Behavior & Organization, 181, 360-368. https://doi.org/10.1016/j.jebo.2017.11.013

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science, 12(6), 1100-1122.

• Overall, when regression assumptions are violated in ML, the risk for false positive/negative results may be *less* than in the original regression model • ML may be a useful tool for linear regression

• Understanding these implications of assumption

violations in ML can significantly improve replicability

• More work to be done to understand the implications of these violations for model fit

• Future directions: attempt to "fix" these violations and re-run regression and machine learning models