Prediction Question

How do demographic and socioeconomic factors, such as poverty levels, income, median rent, and the proportion of white population, influence the Democratic margin of victory in Virginia counties during a presidential election year?

Comparison of Models

The results (Figure I) show that LASSO did not exclude any features because the coefficients remained non-zero in the LASSO model.

Figure I

	variable	linear_coefficient	lasso_coefficient
0	income_log	6724.682626	6690.825736
1	poverty_ratio	1660.241808	1696.903113
2	median_rent_log	4195.132712	4206.051610
3	pctwhite	-3849.206533	-3856.592050

Key Observations

1. Consistency Across Models: The LASSO coefficients are very close to the Linear Regression coefficients. Some of those adjustments contributed to variables such as, poverty_ratio, median_rent_log, and whitepop increased slightly, indicating a slightly greater positive effect under the LASSO model. Therefore, this could indicate that the regularization penalty applied by LASSO was relatively small and didn't significantly shrink the coefficients. Moreover, the predictors are relevant to the target variable (margin), so LASSO didn't need to exclude any.

2. LASSO Retained all Features: Since all the features were retained, it indicates that each predictor contributes to the target variable and that there isn't significant multicollinearity or noise in the data that would result in irrelevant features or overfitting in the data.
3. Model Interpretation: Both models suggest that all four variables (income_log, poverty_ratio, median_rent_log, and whitepop) are significant in explaining the variation in the Democratic margin of victory. The LASSO model has slightly different coefficients which indicate that it applied a mild regularization effect – reduces some regression coefficients to zero to prevent overfitting without much bias. Thus, improving the predictive performance model.
4. Significance of Log-Scaling: The inclusion of log scaled variables, such as income_log and median_log highlights the importance of accounting for skewness or non-linear relationships within the data. This allows both models to become more interpretable and better illustrate the relationship between predictive variables and the Democratic margin of victory.

Why Did LASSO Not Exclude Any Features?

LASSO did not exclude any features due to several factors including strong predictors, low regularization strength, and minimal multicollinearity. Strong predictors reveal that all features may have significant correlations with the target variable; thus, improving predictive accuracy. As for the low regularization strength, it indicates that the penalty term's influence was small and the model retains its flexibility, leading to low bias. Finally, minimal multicollinearity reveals that LASSO will retain all variables, by doing so it improves predictive performance. This is due to low redundancy in which the coefficients directly represent each feature's

predictability (better interpretability) and the model is less sensitive to small changes in the data (stronger stability).

Key Takeaways

There are three important takeaways that can be drawn. First, linear regression and LASSO produced similar R^2 values and coefficients which indicates that all predictors are relevant to the model. As a result, it contributes to the explanation of the variance in the target variable. Furthermore, it indicates that our model is robust without needing to eliminate a lot of features. Secondly, LASSO retained all features, but slightly shrunken coefficients – reinforcing the robustness of the features – without sacrificing predictive accuracy and performance. Additionally, it suggests that there is minimal multicollinearity in the predictive model because LASSO did not make coefficients to zero. Finally, the incorporation of LASSO's built-in regularization also adds another layer of robustness by preventing overfitting, which is important because there could be a lot of noise in our model. The slight shrinkage of coefficients further indicates that it is generalized efficiently while ensuring that our model is easily interpreted and reliable for future predictions.

By understanding that LASSO achieves comparable performances, we can now transition into the interpretation of the coefficients produced by the LASSO model. This will allow us to better understand the relationship between our predictors and the margin of victory, as shown below.

Feature Selection and Coefficients

We used an 80 / 20 train-test split for the LASSO model, and our data included four features (Appendix A: Codebook). Because the LASSO model did not exclude any features and applied only a small penalty to the linear coefficients, we will use the LASSO model coefficients to interpret the results of our project.

To reduce the impact of widely distributed variables, income and median rent were log-transformed to reduce skewness, and all predictors were subsequently standard-scaled (mean = 0, standard deviation = 1) to facilitate direct comparisons between coefficients.

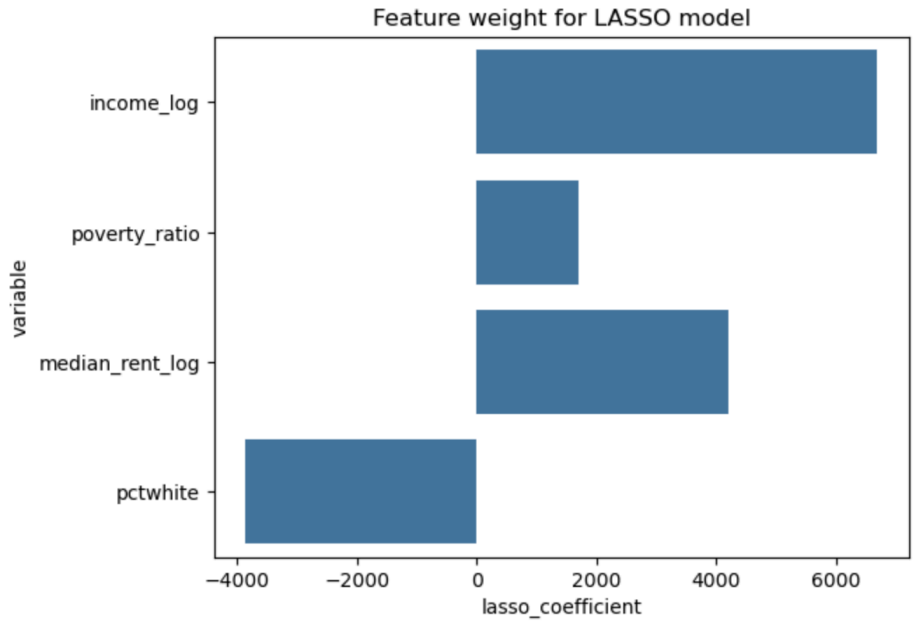
The coefficients of each predictor in the LASSO model are shown in Figure II. These results can be interpreted as follows: "A one standard-deviation increase in variable *X* is associated with a β_1 change in voting margin, holding all else constant." Variables with large positive coefficients, like log-scaled income, strongly increase the model's predicted margin for Democrats. The only variable that increases the model's predicted margin for Republicans is the percentage of the population identifying as white, with an effect of ~4,000 votes.

The model's performance is moderate, with an R^2 of 0.329. This indicates that approximately 32.9% of the variance in the Democratic margin of victory is explained by the features included in the model (Appendix A). The root mean squared error (RMSE) of 9,9704 votes suggests that the average error for the predicted margin across counties is relatively large. This reflects the limited number of features selected and the limited number of data available, with only 532 observations in the Commonwealth of Virginia across four election cycles.

Among the selected features, percent white was the only statistically significant predictor at the 0.05 level, with a coefficient of -3849.21. This can be interpreted that a one-standard-deviation increase in the proportion of the white population is associated with a

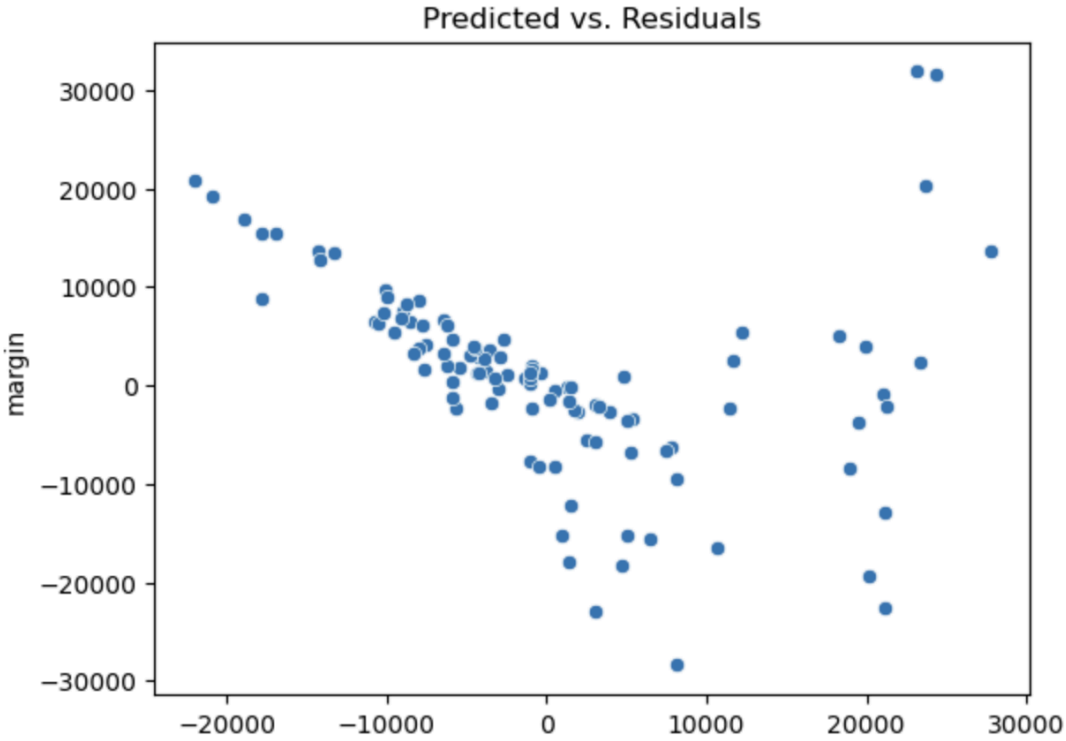
3,849 increase for the Republican margin, holding all else constant. This variable remains largely unchanged between the two models, suggesting its strength as a key critical predictor. Other predictors, such as percent living under the poverty line (1,660.24) and scaled median rent (4,195.13), demonstrated positive associations with the Democratic margin, albeit with smaller magnitudes.

Figure II



The residual plot (Figure III) reveals that the model's residuals are not evenly distributed around zero, suggesting heteroscedasticity. The variance of the residuals increases greatly at higher predicted margins, indicating that the model performs quite inconsistently for counties with extreme Democratic margins. This reflects the limited set of predictors and the unique variance of specific counties (e.g., Fairfax County) that are not fully captured by the model.

Figure III



Summary of Findings

These results emphasize the relationships between demographic and socioeconomic factors and voting outcomes in Virginia. The influence of the percent White variable on the Republican margin of victory suggests a pivotal role in the composition of certain demographics and how it shapes electoral outcomes. On the other hand, the income_log variable has a negative relationship with the Democratic margin of victory, possibly indicating that higher-income areas are more likely to align with Democratic preferences and the overall wealth gap. When looking at the last two variables of poverty_ratio and median_rent_log, indications of economic well-being may be presumed, and while influential, are secondary in demographic factors within the models conducted. Lastly, a moderate R^2 value suggests that these features don't capture the

entire story, but remain important predictors, providing valuable information into trends and tendencies among voter characteristics.

Among the selected features, only percent identifying as white was statistically significant at the 0.05 level, with a strong positive relationship to the Republican margin. The other predictors showed no statistically significant relationships but still may offer some explanatory value in specific contexts.

Appendix A: Codebook

Features

Variable Name	**Description**
income_log	Log-transformed median income of the county
poverty_ratio	Proportion of the county's population living below the poverty line (bounded between 0 and 1)
median_rent_log	Log-transformed median rent price of the county
pct_white	Proportion of the county's population identifying as white (bounded between 0 and 1).

Target Variable

Variable Name	**Description**
margin	Margin of victory (Democratic votes - Republican votes) in the county for a given Presidential election year