Multiple Regression and Geographically Weighted Regression (GWR)

Your results should be delivered as an HTML webpage generated using R Markdown. Make sure to include all code and results along with the answers to the questions. You are only required to deliver text to answer the questions; you do not need to explain the methods.

Grading Criteria

- Correctness and completeness of code (16 Points)
- Answer to questions (16 Points)
- Webpage formatting (4 Points)
- Map outputs (4 Points)

The goal of this modeling exercise is to predict the percent of voters that voted for the democratic presidential candidate at the county level during the 2012 presidential election. You will explore both multiple regression and geographically weighted regression (GWR). You will investigate the following model:

per_dem~log(pop_den)+per_gt55+per_notw+per_vac+per_pov

“per_dem”: percent of voters voting democrat (dependent variable)

“pop_den”: population density

“per_gt55”: percent of the population older than 55

“per_notw”: percent of the population that is not Caucasian

“per_vac”: percent of homes that are vacant

“per_pov”: percent of the population that is below the poverty line

All predictor variables were derived from US Census data. Note that you will apply a log transformation on the population density variable so that the relationship between this variable and the dependent variable is more linear. Also, percent of the population that voted democrat is not easy to predict, so don’t expect to achieve low RMSE or high R^2 metrics.

Multiple Regression

T1: Create a map of the “per_dem” data using **tmap** to visualize the dependent variable.

T2: Generate the multiple regression model. Remember to use a log transform for the population density variable. Note that you may need to create a copy of the data without the geometry column to perform the multiple regression analysis.

T3: Is a statistically significant F-statistic obtained for the multiple regression model? What does a statistically significant F-score indicate?

T4: Which independent variables are suggested to be statistically significantly correlated with “per_dem”?

T5: What adjusted R-squared value is obtained for the multiple regression model?

T6: What RMSE value is obtained for the multiple regression model?

T7: Are the residuals normally distributed? Is there an issue with this regression assumption?

T8: Is the homoscedasticity assumption met?

T9: Are there issues with multicollinearity between the predictor variables?

T10: Are there any outlier data points?

T11: Are there any high leverage points?

T12: Based on the results of your diagnostics, can multiple linear regression be applied to generate this model without further manipulation or preparation of the data? Why or why not?

T13: Make a map with **tmap** of the residuals for the multiple regression model.

Geographically Weighted Regression (GWR)

T14: Use the gwr.sel() function to determine an optimal bandwidth. What bandwidth is selected as optimal?

T15: Produce the GWR model using the optimal bandwidth.

T16: What RMSE value is obtained for the GWR model? Does this suggest improvement in comparison to the multiple regression model?

T17: Make a map with **tmap** of the predicted values for "per_dem" obtained using GWR.