Problem Set 2: Linear Regression

Student Name

Due to COVID-19 pandemic, many states made alternatives in voting, such as voting by mail, more widely available for the 2020 U.S. election. The general consensus was that voters who were more Democratic leaning would be more likely to vote by mail, while more Republican leaning voters would largely vote in-person. This was supported by multiple surveys, including this survey conducted by Pew Research.

The goal of this analysis is to use regression analysis to explore the relationship between a county's political leanings and the proportion of votes cast in-person in 2020. The ultimate question we want to answer is **Did counties with more Republican leanings have a larger proportion of votes cast in-person in the 2020 election?**

We will use the proportion of votes cast for Donald Trump in 2016 (pctTrump_2016) as a measure of a county's political leaning. Counties with a higher proportion of votes for Trump in 2016 are considered to have more Republican leanings.

We will focus on the following variables for this analysis:

- `inperson_pct`: The proportion of votes cast in-person in the 2020 election in a given county.
- `pctTrump_2016`: The proportion of a given county's votes cast for Donald Trump in the 2016 election.

Importing the Data

Up to this point in this course we have been using data frames that we either created in R or loaded from some package. In this problem set, you will import a dataset in csv format into your RStudio cloud and into your quarto work space before you proceed. The data set is available on this link

Problem 1 (data visualization & summary stats)

a) Create a visualization for the variable `inperson_pct` and calculate at least two summary statistics to measure the central tendency, as well as one summary statistic to measure the spread. Then, interpret both the visualization and the summary statistics in context of the data.

b) Create a visualization of the relationship between `inperson_pct` and `pctTrump_2016`. Based on the visualization, how would you describe the relationship?

c) Compute the correlation coefficient between `inperson_pct` and `pctTrump_2016`. Interpret the value in the context of the data.

Problem 2 (linear regression)

a) Fit the linear model to understand variability in the percent of in-person votes based on the percent of votes for Trump in the 2016 election. Write down the equation of the model.

b) Interpret the slope of the model. The interpretation should be written in a way that is meaningful in the context of the data.

c) Does it make sense to interpret the intercept of the model? If so, write the interpretation in the context of the data. Otherwise, briefly explain why not.

d) Calculate the R-squared value for the model. Interpret the value of the R-squared in the context of the data.

e) Use the linear model to predict the proportion of in-person votes in a county where 60% of the votes were cast for Trump in 2016.

Problem 3 (model diagnostics)

a) Calculate the residuals for the linear model. Create a histogram of the residuals and describe the distribution of the residuals.

b) Create a residual plot - scatterplot of the residuals against the fitted values. Describe the relationship between the residuals and the fitted values.

c) Assuming that the cases (counties) in the dataset are independent, explain whether the assumptions of the linear regression model are met.

Problem 4 (categorical predictor)

a) Create a binary variable called`pctTrump_2016_high` with levels `pro_trump` and `anti_trump` that indicates whether the proportion of votes for Trump in 2016 was above 50% in a county.

b) Fit a linear regression model to predict the proportion of in-person votes based on the binary variable `pctTrump_2016_high`. Write down the equation of the model.

c) Interpret the slope of the model. The interpretation should be written in a way that is meaningful in the context of the data.

d) Which model (the one in Problem 2 or the one in Problem 4) is more ''appropriate'' for predicting the proportion of in-person votes in the 2020 election? Justify your answer.