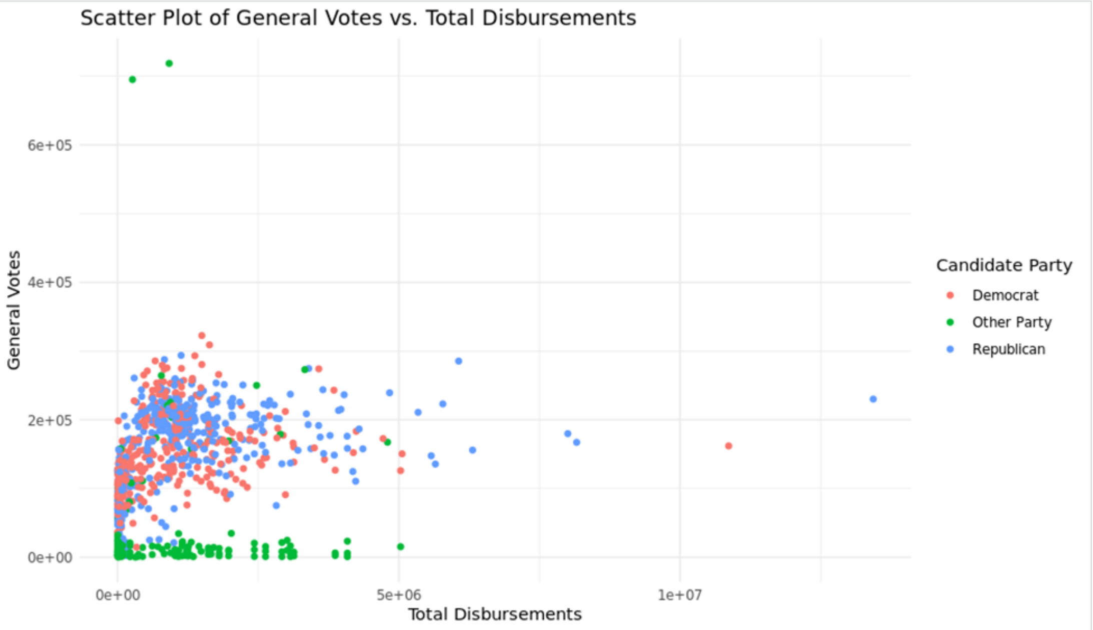
Regression – Evaluating Large Sample Assumptions

We produced a linear regression model with the outcome general_votes on ttl_disb and candidate_party and now, we evaluate the large-sample linear model assumptions:



Assumption 1: Identically and Independently Distributed (I.I.D.) Data Points

The data consists of campaign finance and vote counts from the 2016 election cycle, where each data point represents a different candidate's campaign. These observations are collected independently, meaning that the outcome of one candidate's campaign does not influence another. The scatter plot of general votes versus total disbursements, colored by candidate party, shows no obvious clustering or patterns that would suggest dependency between observations. Additionally, the residuals summary statistics do not show any strong deviations that would indicate violations of independence. Therefore, based on the independent nature of the data collection and the lack of patterns in the scatter plot, we can reasonably assume that the data points are I.I.D.

Assumption 2: Best Linear Predictor (BLP) Exists and is Unique

The linear relationship between spending (total disbursements) and votes (general votes) is a common assumption in political science, supported by theories that higher spending generally increases votes. The scatter plot shows a general trend that higher disbursements tend to be associated with higher vote counts, especially within each candidate party category, supporting the presence of a linear relationship. The regression summary indicates significant coefficients for total disbursements and candidate party "Other Party," with an R-squared value of 0.3589, suggesting that the model explains a substantial portion of the variance in general votes. These findings confirm the existence and uniqueness of the BLP, as the linear relationship is the best approximation of the true underlying relationship between the predictors and the response variable. Therefore, the evidence from the scatter plot and the statistical significance of the model coefficients supports the validity of this assumption.