Predicting Political Donations through Exploratory Model Selection

Elisa Cui and Samuel O'Neill
Lab Section: Wednesday 4pm

Political donations can be influenced by a multitude of factors. Our goal was to test and identify predictor variables that had statistically significant effects on the amount a person donates to their political party, ultimately formulating an optimal multivariate model. Several multiple linear regression techniques were used to perform our analysis and develop our final model. We analyzed the following predictor variables from a dataset with 100 subjects: annual income, years of education, and political party. Annual income (coded as *income*) is a quantitative variable with units of thousands of dollars. Years of education (coded as *educyrs*) is a quantitative variable that takes on a value 12, 16, or 20 to represent a high school diploma, college diploma, or doctorate degree. Political party (coded as *party*) is a qualitative variable that takes on a value 0 or 1, where 0 corresponds to a Democrat and 1 corresponds to a Republican. Donation amount (our response variable coded as *donate*) is a quantitative variable with units of dollars.

Our regression analysis mainly consisted of the following methods: first order fitting, assumptions evaluation, ANOVA model comparison, and step-wise model selection. Our first step was to create a scatter-plot matrix (Figure 1) of the full data set to visually examine relationships between variables. This visual analysis quickly made it clear that income and donation amounts have a strong linear relationship.

Scatter-Plot Matrix w/Outlier

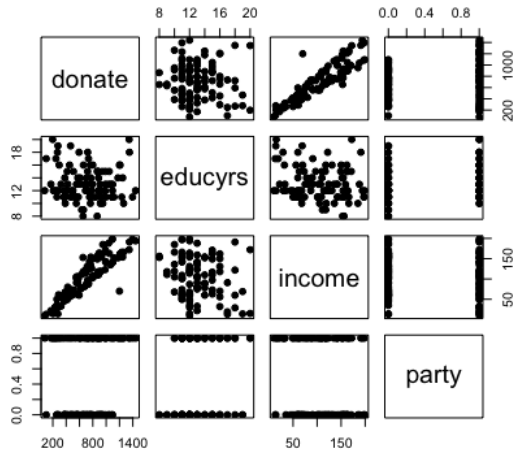

FIG. 1: This scatter-plot matrix shows the individual relationships between amount donated and all predictor variables. An outlier is clearly visible in the plot with *donate* and *income*.

It was also clear that there was a distinct outlier in the data set as seen in Figures 1 and 2. Before proceeding, we created a new dataset without the outlier. However, we did not ignore the outlier and after selecting a model based on the new dataset, we repeated the same steps to confirm our final model.

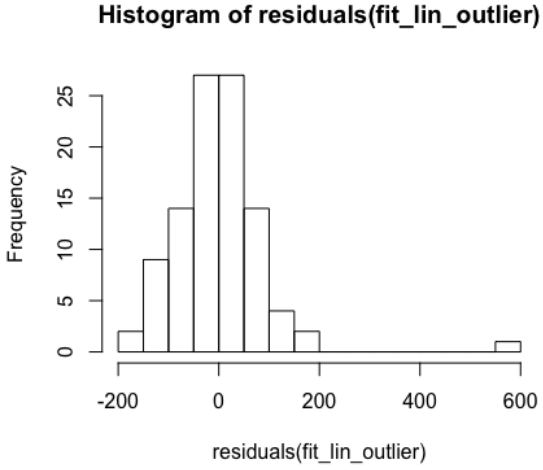

FIG. 2: The histogram reaffirms the existence of an outlier (present on the far right of the graph) in the original dataset. It appears not to follow the same normal distribution as the other residuals. For these reasons, we removed the outlier from the dataset.

With the altered dataset, we fit a first order linear model titled *fit_lin* with all predictor variables (written below). The coefficients were derived using the least squares estimation technique. The intercept will later be referred to as β_0, and similarly variable coefficients will be referred to as $\beta_1, \beta_2, \beta_3, ...$

$$\begin{aligned} donate = & -131.41 + 7.61 * educyrs \\ & + 6.10 * income + 194.85 * party \end{aligned} \quad (1)$$

We then performed the following hypothesis test to check the significance of this initial model.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$
$$H_A : \text{At least one } \beta_{i=1,2,3} \neq 0$$

With an F-statistic of 624.9 with $df1 = 3, df2 = 95$, our p-value was less than $2.2{*}10^{-16}$ and, consequently, we rejected the null hypothesis. Thus, *fit_lin* is a statistically significant model and at least one predictor variable has a statistically significant effect on donation amounts. It is also worth noting that the adjusted R^2 for *fit_lin* is about 95%.

For *fit_lin*, we checked the multiple linear regression model assumptions, namely constant residual variance, normal distribution and linearity. We had already addressed issues with identical distribution by initially removing the outlier. Figure 3 demonstrates non-constant variance violations, however, it confirms the linearity assumption with its even spread around the residual zero line. Figure 4 exhibits a violation of the normality assumption.

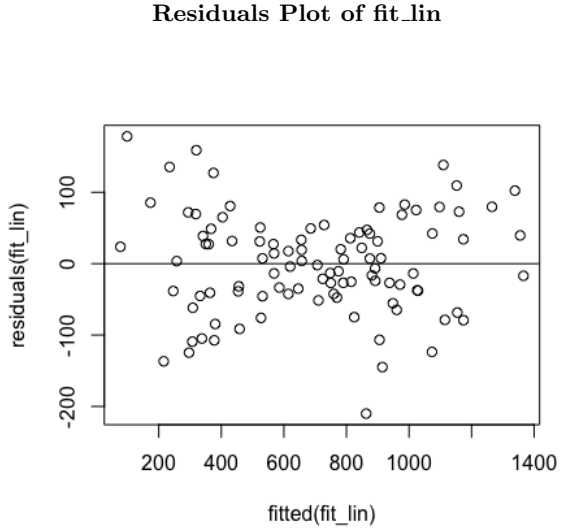

FIG. 3: The residuals plot exhibits a non-constant variance along its domain. The variance of residuals is wide at the end and narrow in the middle. Although the variance is not constant, the model appears linear since the residuals are evenly distributed about the zero line.

Considering there are multiple assumption violations, we attempted to remediate them by adding an interaction term. We checked for an interaction between *income* and *donate* and concluded there was one due to the visual evidence presented in Figure 5. We then created a new model titled *fit_int* that included the interaction term (written below).

$$\begin{aligned} donate = & -5.46 + 7.06 * educyrs \\ & + 5.04 * income - 20.90 * party \\ & + 2.05 * income * party \end{aligned} \quad (2)$$

To confirm the validity of *fit_int*, we performed an ANOVA model comparison test between *fit_int* and

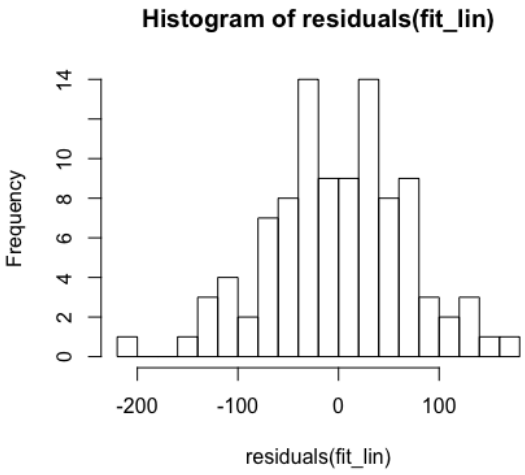

FIG. 4: The histogram almost appears to satisfy the normality assumption. However, the middle of the residual distribution is not characteristic of a normal distribution.

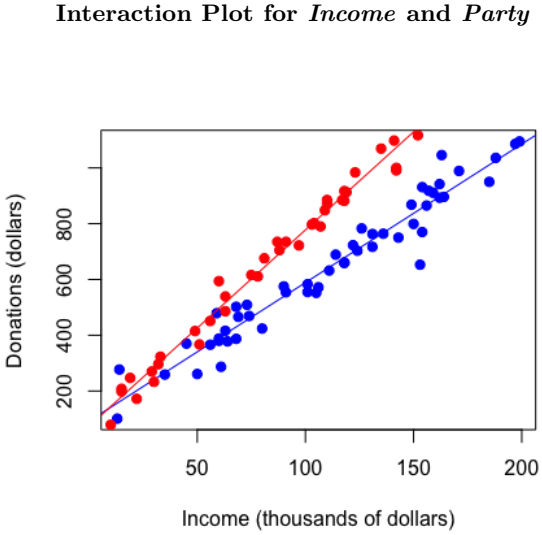

FIG. 5: The plot of annual income vs. donation amount was coded to check an interaction with a third variable, political affiliation. The data points were coded to distinguish between Republicans (colored red) and Democrats (colored blue). There are distinct trends for the two different parties, strongly suggesting there is an interaction between the *income* and *party* predictor variables.

fit_lin. The null and alternative hypotheses are as follows (note β_4 is the coefficient for $income * party$ term).

$$H_0 : \beta_4 = 0$$
$$H_A : \beta_4 \neq 0$$

With an F-statistic of 111.51 with $df1 = 95, df2 = 94$, our p-value was less than $2.2 * 10^{-16}$ and, consequently,

we rejected the null hypothesis. Thus, adding the interaction term has a statistically significant effect. Moreover, the adjusted R^2 for *fit_int* is 98% compared to 95% for *fit_lin*. Therefore, *fit_int* is likely a better fitting model than *fit_lin*.

To confirm that *fit_int* is the best descriptive model of the data, we performed forward and backward stepwise model selection between a null and a full equation (written below).

Null Model

$$donate = \beta_0 \tag{3}$$

Full Model

$$\begin{aligned} donate = \beta_0 &+ \beta_1 * educyrs + \beta_2 * income + \beta_3 * party \\ &+ \beta_4 * income * educyrs \\ &+ \beta_5 * party * educyrs \\ &+ \beta_6 * income * party \end{aligned} \tag{4}$$

Our model selection code added and subtracted variables based on the greatest negative change in AIC (Akaike's Information Criteria). Under this criteria, both the forward and backward stepwise model selection arrived at *fit_int* (see Appendix). This offered strong support for *fit_int* so we decided to not add or remove any variables from the model.

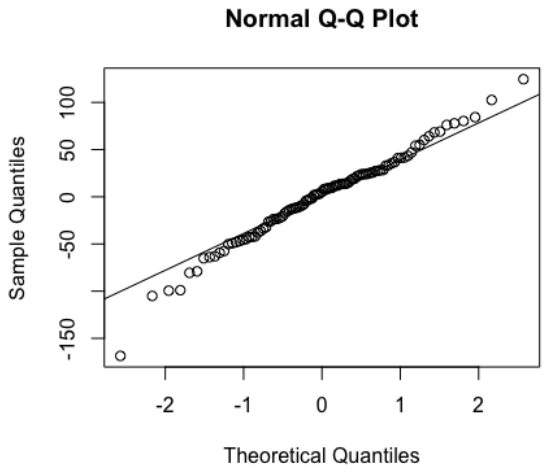

FIG. 6: The QQ-Plot compares the observed distribution to an expected normal distribution (represented by the line). The observed data generally matches the normal distribution making it unlikely for normality to be violated.

We proceeded to reevaluate the model assumptions for *fit_int* to confirm if it would be a valid final model. In Figures 6, 7, and 8, we checked normality, constant variance, and linearity which were not satisfied in the previous model, *fit_lin*. The figures offered visual evidence

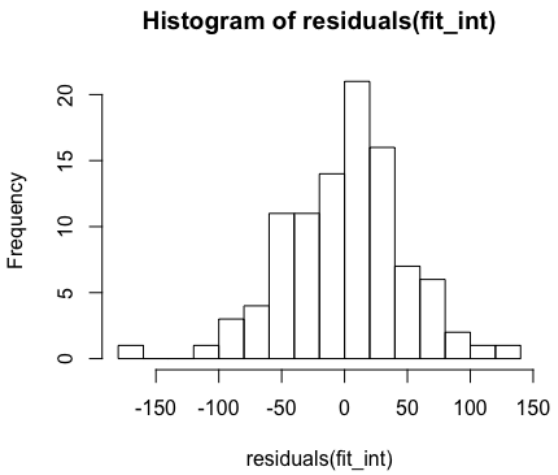

FIG. 7: The histogram generally has the shape of a normal distribution. We can conclude the normality assumption is satisfied.

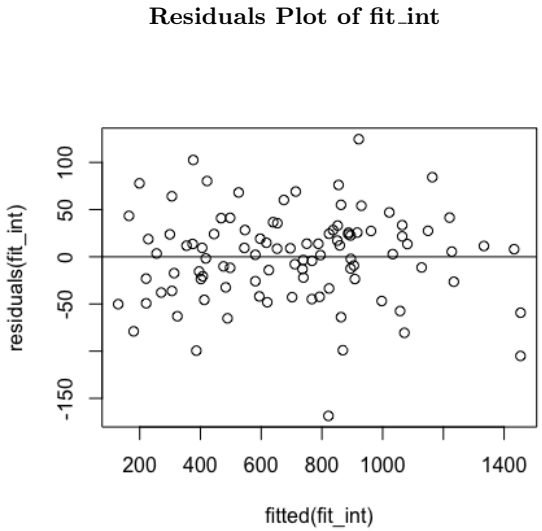

FIG. 8: This residuals plot has constant variance compared to the residuals plot for *fit_lin*. Linearity is satisfied as well.

that *fit_int* did, in fact, satisfy the assumptions. This further reinforced our selection of *fit_int* as our potential final model.

One potential issue with *fit_int* was the additional complexity of the model compared to *fit_lin*. We took this into account by comparing the BIC's (Bayesian Info Criteria) of the models. The BIC penalizes models with increased complexity but rewards better fitting models. Although *fit_int* has more complexity, its BIC is lower than *fit_lin* (1071.27 compared to 1144.12). This is a testament to *fit_int*'s significantly improved fit.

At this point, we performed the same exploratory process for the dataset with the outlier including an ini-

tial linear fit, ANOVA comparison, assumptions evaluation, step-wise model selection and BIC comparison. We wanted to make sure that removing the outlier did not alter our potential final model. Fortunately, our analysis led to a model titled *fit_aic* (written below) with the same predictor variables as *fit_int*.

$$\begin{aligned} donate = & -36.17 + 9.28 * educyrs \\ & + 5.06 * income - 4.06 * party \\ & + 1.91 * income * party \end{aligned} \quad (5)$$

Compared to Equation 2 (*fit_int*), Equation 5 (*fit_aic*) has quite different coefficient values. This is because the outlier strongly skews the data and resulting linear model. Due to this effect, we chose our final model to be *fit_int*.

A more detailed description of our final model *fit_int* is written below.

$$\begin{aligned} donate = & -5.46 + 7.06 * educyrs \\ & + 5.04 * income - 20.90 * party \\ & + 2.05 * income * party \end{aligned}$$

95% Confidence Intervals and P-Values for Coefficients:
$\beta_0 = -5.46$ has CI $[-68.79, 57.87]$ and $p = .864$
$\beta_1 = 7.059$ has CI $[3.17, 10.94]$ and $p = .000498$
$\beta_2 = 5.03$ has CI $[4.75, 5.31]$ and $p < 2 * 10^{-16}$
$\beta_3 = -20.89$ has CI $[-65.98, 24.18]$ and $p = .360$
$\beta_4 = 2.05$ has CI $[1.66, 2.43]$ and $p < 2 * 10^{-16}$
-Note that the p-values are based off the partial sum of squares
-The large p-values and confidence intervals for β_0 and β_3 imply that the intercept of our model is around zero and has large variability. This is a limiting feature of our model because predicted individual responses can considerably vary.

fit_int effectively splits into two models: one for Republicans and another for Democrats.

For Democratic subjects ($party = 0$), the regression model becomes

$$donate = -5.46 + 7.06 * educyrs + 5.04 * income$$

-Based on the intercept coefficient, if $educyrs = income = 0$ for a Democratic subject, then the expected amount donated will be $-\$5.46$. This is not a realistic expectation, but it can be inferred that the subject's donation amount will be close to zero.
-For every extra year of education, a Democratic subject is expected to donate \$7.06 more, holding annual income constant.
-For every extra thousand dollars of annual income, a Democratic subject is expected to donate \$5.04 more, holding years of education constant.

For Republican subjects ($party = 1$), the regression model becomes

$$donate = -26.36 + 7.06 * educyrs + 7.09 * income$$

-Based on the intercept coefficient, if $educyrs = income = 0$ for a Republican subject, then the expected amount donated will be $-\$26.36$. This is not a realistic expectation, but it can be inferred that the subject's donation amount will be close to zero.
-For every extra year of education, a Republican subject is expected to donate \$7.06 more, holding annual income constant.
-For every extra thousand dollars of annual income, a Republican subject is expected to donate \$7.09 more, holding years of education constant.

Now that we formulated a final model, we explored its predictions with actual predictor values. We examined predicted donation amounts for both a Democratic and Republican subject group with an annual income of $\$100,000$ and 16 years of eduction (AKA college diploma).
-The predicted individual response for Democrats was \$611.14 with a 95% CI of $[513.40, 708.87]$
-The predicted mean response for Democrats was \$611.14 with a 95% CI of $[592.66, 629.61]$
-The predicted individual response for Republicans was \$795.15 with a 95% CI of $[697.68, 892.63]$
-The predicted mean response for Republicans was \$795.15 with a 95% CI of $[778.12, 812.19]$

Even if the statistics support the model, it is important that the model has a reasonable interpretation. The presence of an interaction term makes sense, because we would not expect identical donation behavior from people with differing political ideologies. Exploring the nature and cause of this interaction term would be an intriguing focus for further research. It is interesting that high-earning Republicans donate more than high-earning Democrats. Furthermore, the positive relationship between donation amounts and the predictor variables *income* and *educyrs* is quite realistic. We would expect wealthier people to donate more due to a larger disposable income. We would also expect well educated people to donate more due to their knowledge of the political process. Combined with the regression techniques that optimize our model's fit and complexity, these factors make our model an effective predictive tool for forecasting political donations of Republicans and Democrats.