

Determinants of rejected mail ballots

Consultant: Xinjie Qian

Client: Dominic Nyhuis

Advisor: Dr. Steve Marron, Dr. Perry Haaland

University of North Carolina at Chapel Hill

4/26/2021

1 Abstract

In Germany, about 2.7 percent of all mail-in ballots are rejected. And most of the rejected ballots are because an affidavit is not enclosed or not signed. This project aims to find out whether the socio-economic status has influence on rejected mail ballots in Germany. The data contains 5 cities and 8 elections. By using partial correlation analysis on 3 pairs of variables with conditioning on different socio-economic variables, we found significant correlation relationships between the percentage of the rejected mail ballots and the socio-economic status. Then we used separate linear regressions for chosen cities and elections under different predictor settings. The results were different under different settings because the high correlation between predictors. Finally, we used Ridge regression to get that a higher taxes per capita in district causes a lower rejection rate, while a higher welfare causes a higher rejection rate.

2 Executive summary

The following shows the overview of major findings:

- The socio-economic status conditioned in each district has a strong influence on rejection rates.

- Rejection rates and voting rates are partially correlated when conditioning on the socio-economic status. And the rate of rejected mail and number of eligible people and voting rates are uncorrelated when conditioning on any socio-economic status.
- Most of the socio-economic variables are correlated with each other.
- Taxes per capita in district have a negative relation with the rejection rate, while welfare and whether young people are eligible to vote have a positive relationship with the rejection rate.

3 Data description

The dataset was provided by Dominic Nyhuis, Department of Political Science and the Center for European Studies, UNC. The dataset was collected in collaboration with several German municipal statistical offices. The dataset consists 5 cities and 8 elections with 418 observations which is for districts within the cities for each election. These included 230 valid observations (data without missing values) from Bremen, a city in Germany, over 3 elections. There are 4 variables about socio-economic status: *foreigners*, *unemployed*, *welfare*, *taxes* (only Bremen has *taxes* data). Also, we have the data about the election type, eligible rules, the number of eligible voters, mail ballots, rejected mail ballots, in-person voters at the district level.

4 Data preprocessing and exploratory data analysis

The raw data can not be used directly because of missing values, and some processing steps should be applied. In addition, the analyses used here need specific assumptions, for example, normality and linearity. Thus exploratory data analysis is essential.

4.1 Data preprocessing

Several preprocessing steps were performed on the raw data. We first examined the presence of missing data and non-standard expression of data, for example, variable *voters_person* has data with comma, just like "1,972". We found 16 missing values and deleted the observations which contain missing values, so our final data has 402 observations and 24 variables. Because proportions are viewed as important as the numerical values we also added four variables:

- $Rate_Rejection_MailVoters = \frac{\text{the number of rejected mail ballots}}{\text{the number of mail voters}}$

- $Rate_Rejection_Eligible = \frac{\text{the number of rejected mail ballots}}{\text{the number of eligible people}}$
- $Rate_Vote_Eligible = \frac{\text{the number of people voted}}{\text{the number of eligible people}}$
- $Rate_MailVoters_Vote = \frac{\text{the number of mail voters}}{\text{the number of people voted}}$

4.2 Exploratory data analysis

First, we used histogram to directly show the distribution of the variables of socio-economic status. After that, we decided whether we should do transformation and which transformation was better.

In Figure 1, we show the histogram of *unemployed*, $\log(\text{unemployed})$ and $\sqrt{\text{unemployed}}$ separately. The distribution of *unemployed* is skewed and the log transformation over corrects the skewness. Hence, we prefer the square root transformation.

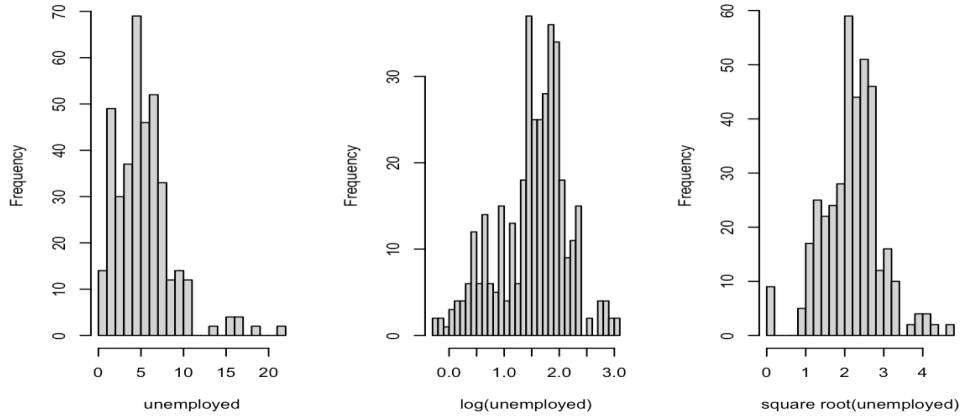


Figure 1: Histogram of unemployed(left panel) and its log(center) and square root(right panel) transformation. It shows square root transformation is most appropriate.

Similarly, we also chose the square root transformation for *welfare*, according to Figure 2.

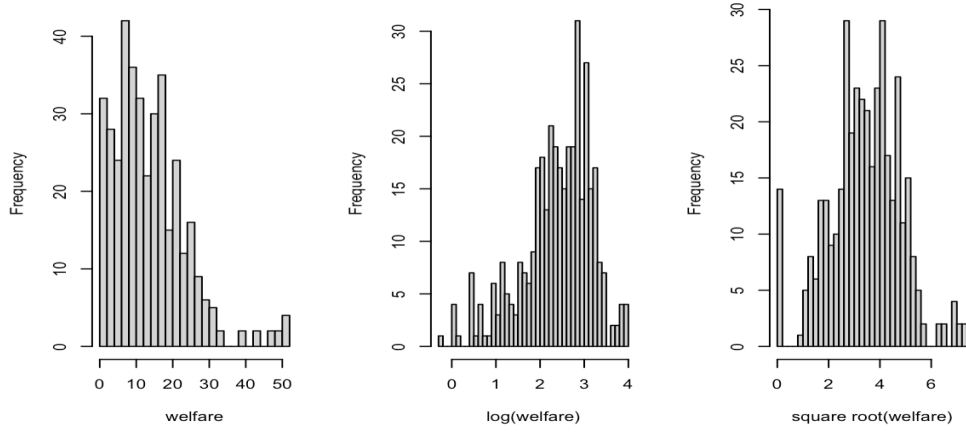


Figure 2: Histogram of welfare(left panel) and its log(center) and square root(right panel) transformation. It shows square root transformation is most appropriate.

After examining the distribution of *foreigners*, we decided not to do any transformation, while for *taxes*, a log transformation was necessary due to the skewness of the raw distribution (Figure 3).

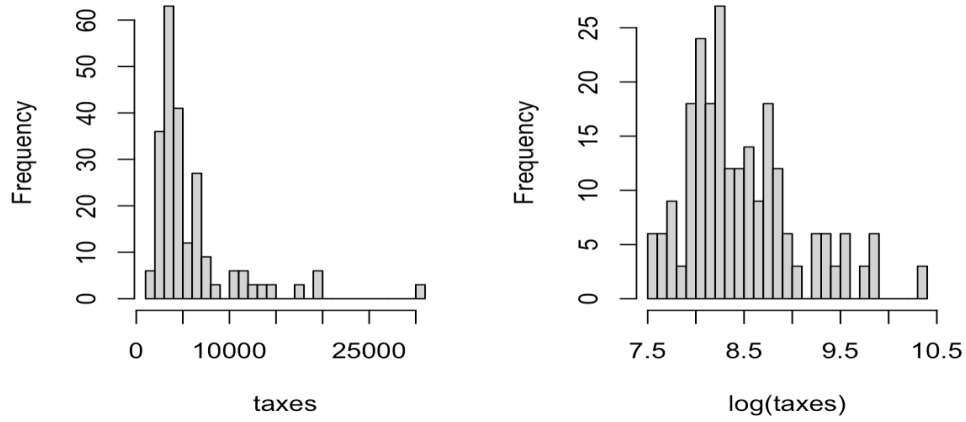


Figure 3: Histogram of taxes(left panel) and its log(right panel) transformation. It shows log transformation is most appropriate.

Before using partial correlation analysis, we needed to examine the normality. Take *Rate_Rejection_MailVoters* as example. Figure 4 is a Q-Q plot for the distribution of *Rate_Rejection_MailVoters* and the red line is a reference for the data if it is normally distributed. As shown in Figure 4, it tells us that our data are approximately normally distributed.

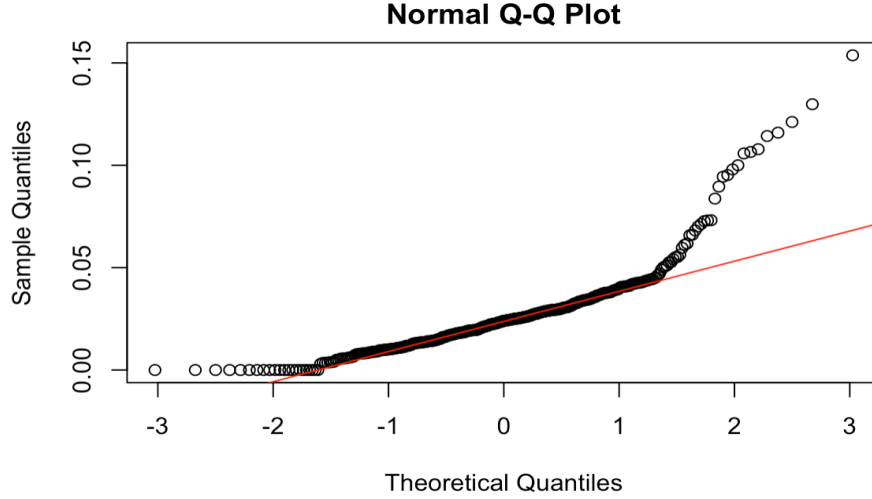


Figure 4: The Q-Q plot for *Rate_Rejection_MailVoters*. It shows that our data are approximately normally distributed.

5 Analysis

5.1 Partial correlation analysis

Partial correlation measures the degree of association between two random variables, with the effect of a set of controlling random variables removed. If we are interested in finding to what extent there is a numerical relationship between two variables of interest, using their conventional correlation coefficient will give misleading results if there is another confounding variable that is numerically related to both variables of interest (a good example is Table 3). This misleading information can be avoided by conditioning on the confounding variable.

For the reason that only Bremen has *taxes* data and the client would focus on the elections in Bremen, we put emphasis on Bremen first.

We measured the association between *Rate_Rejection_MailVoters* and *Rate_Rejection_Eligible* with conditioning on 1) all socio-economic variables; 2) each socio-economic variable separately; 3) no condition. The following table (Table 1) shows the partial correlation coefficient (denoted by r) and the p-value which measures the strength of the evidence in favor of the correlation coefficient being different from zero under the control of those variables.

controlling variables	r value	p value
foreigners, sqrt(unemployed), sqrt(welfare), log(taxes)	0.853	$3.30 * 10^{-65}$
foreigners	0.717	$1.99 * 10^{-37}$
sqrt(unemployed)	0.787	$1.61 * 10^{-49}$
sqrt(welfare)	0.800	$1.94 * 10^{-52}$
log(taxes)	0.820	$5.04 * 10^{-57}$
none	0.633	$< 2.2 * 10^{-16}$

Table 1: The partial correlation coefficient between *Rate_Rejection_MailVoters* and *Rate_Rejection_Eligible* and the p-value under the control of different variables. All partial correlations are very strongly statistically significant.

As we can see in Table 1, the rate of the rejected_mail and mail_voters and the rate of the reject_mail and eligible people are significantly correlated because p-value is very low. When we conditioned on the socio-economic variables, the correlation is even higher. This suggests that the socio-economic status conditioned in each district has a strong influence in rejection rates. Each one of these socio-economic variable has a strong correlation between these two rates. But none of them completely determine the correlation.

Next, we measured the association between *Rate_Rejection_MailVoters* and *Rate_Vote_Eligible* with the same controlled variables:

controlling variables	r value	p value
foreigners, sqrt(unemployed), sqrt(welfare), log(taxes)	-0.276	$2.63 * 10^{-5}$
foreigners	-0.383	$2.10 * 10^{-9}$
sqrt(unemployed)	-0.413	$7.41 * 10^{-11}$
sqrt(welfare)	-0.368	$9.31 * 10^{-9}$
log(taxes)	-0.264	$5.35 * 10^{-5}$
none	-0.429	$1.06 * 10^{-11}$

Table 2: The partial correlation coefficient between *Rate_Rejection_MailVoters* and *Rate_Vote_Eligible* and the p-value under the control of different variables. All partial correlations are statistically significant.

From Table 2, it's easy to get that *Rate_Rejection_MailVoters* and *Rate_Vote_Eligible* are correlated under the socio-economic status is controlled, but there is a lot less correlation than that in Table 1 and the correlation is negative.

Finally, we measured the association between *Rate_Rejection_Eligible* and *Rate_Vote_Eligible* with the same controlled variables:

controlling variables	r value	p value
foreigners, sqrt(unemployed), sqrt(welfare), log(taxes)	-0.111	0.10
foreigners	0.062	0.35
sqrt(unemployed)	-0.108	0.10
sqrt(welfare)	-0.102	0.13
log(taxes)	-0.005	0.94
none	0.18	0.006

Table 3: The partial correlation coefficient between *Rate_Rejection_Eligible* and *Rate_Vote_Eligible* and the p-value under the control of different variables. All partial correlations are not statistically significant, while the conventional correlation is statistically significant.

From Table 3, we would see something interesting. When the p value is bigger than 0.05, we say it's not statistically significant. So *Rate_Rejection_Eligible* and *Rate_Vote_Eligible* are uncorrelated under any socio-economic status is controlled, but the conventional correlation is statistically significant.

5.2 Linear regression

We got the result that the socio-economic conditioned in each district have a strong influence in rejected rates in section 5.1. In order to get the direction of this influence, we try to use linear regression to deal with it.

For a chosen city-election(we first chose Bremen and the state election of 2019), we set *Rate_Rejection_MailVoters* as response variable and socio-economic variables as predictor to do linear regression. The following is the result:

	coefficient	p value
sqrt(unemployed)	-0.0157	0.0027
sqrt(welfare)	0.0056	0.0306
log(taxes)	-0.0095	0.0003
foreigners	0.0002	0.4866

Table 4: The result of the first linear regression model. *unemployed*, *welfare* and *taxes* are significant.

In the Table 4, the p-value of *unemployed*, *welfare* and *taxes* are all less than 0.05, which indicates that all of these three variables are significant. Also the higher the tax or the unemployment rate is, the lower rejection rate is. The higher the welfare rate is, the higher rejection rate is. And the number of foreigners doesn't seem to have a significant impact for this model. However, when we changed the predictor setting, which added *Rate_Vote_Eligible* and *Rate_MailVoters_Vote* into account, the result became totally different:

	coefficient	p value
Rate_Vote_Eligible	-0.0208	0.4731
Rate_MailVoters_Vote	-0.0326	0.4534
sqrt(unemployed)	-0.0153	0.0132
sqrt(welfare)	0.0036	0.2881
log(taxes)	-0.0055	0.1285
foreigners	0.0003	0.2881

Table 5: The result of the second linear regression model. Only *unemployed* is significant.

As we can see in Table 5, only *unemployed* is significant. The reason of this situation was that we ignored the correlation between predictors when doing linear regression. This is called multicollinearity which causes the following two basic types of problems:

- 1) The coefficient estimates can swing wildly based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model.
- 2) Multicollinearity reduces the precision of the estimate coefficients, which weakens the statistical power of our regression model. We might not be able to trust the p-values to identify independent variables that are statistically significant.

This motivates detecting and checking the correlations between those variables.

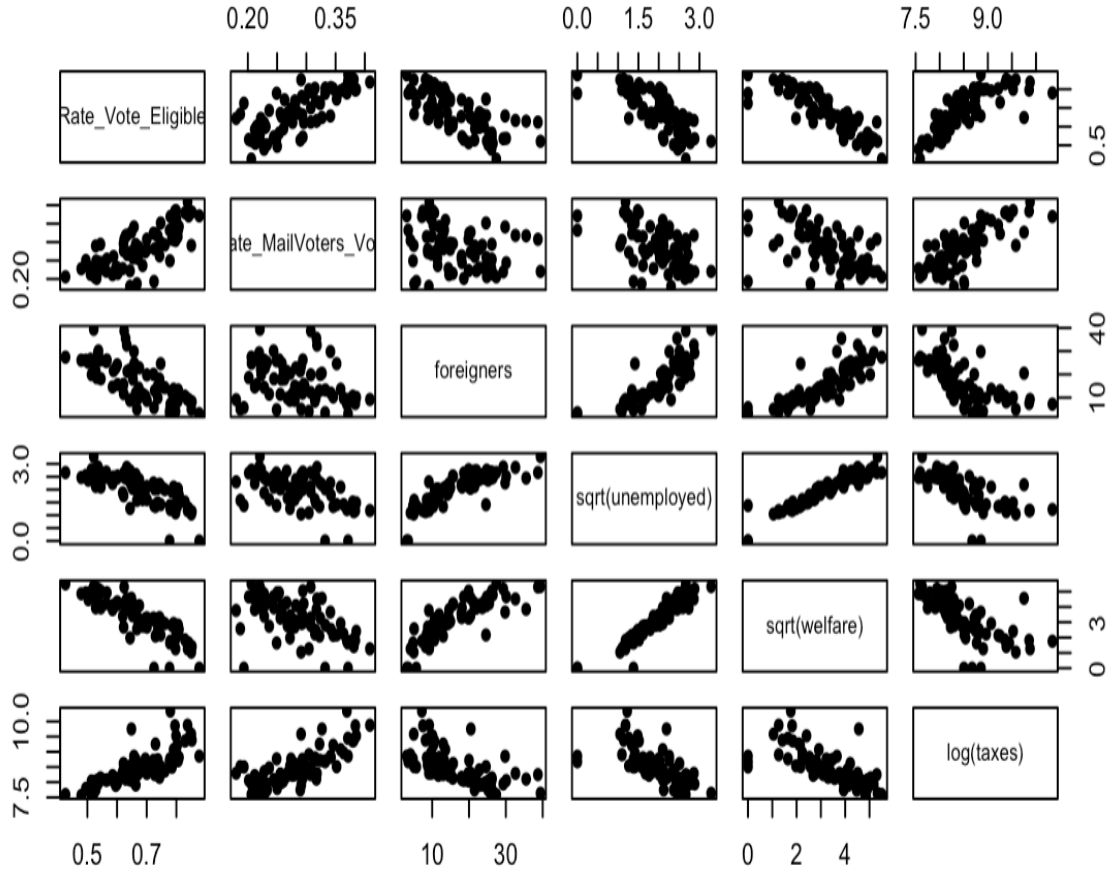


Figure 5: The all pairwise scatter plots of the predictors. There is very strong correlation between lots of those variables.

Figure 5 is the all pairwise scatter plots(a type of plot which displays values for typically two variables for a set of data) of the predictors which mentioned in the second linear regression model. We can easily see the linearity between different variables and almost all of the predictors are correlated to each other, thus the multicollinearity is very severe. Linear regression is not a wise method in this situation and that's why we got different result by using linear regression directly. We can see a more clear version by using the heat map:

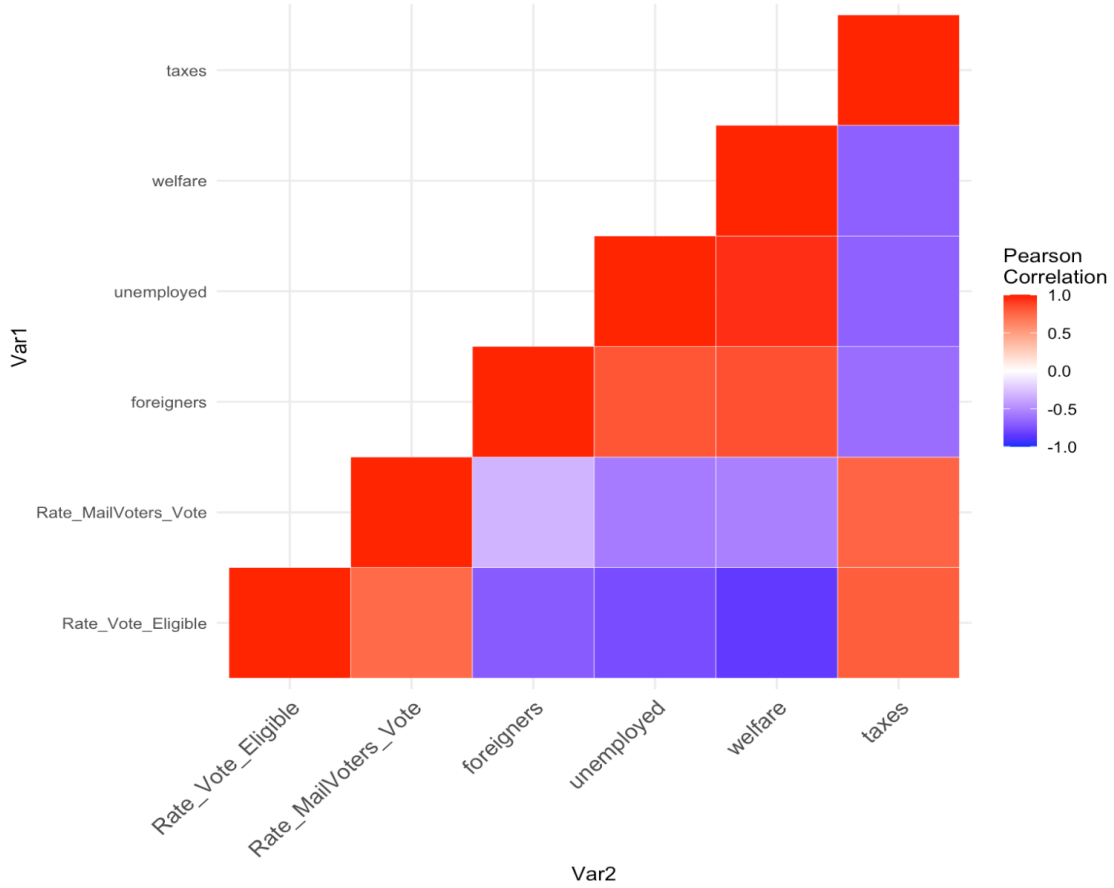


Figure 6: The heat map of the correlation matrix. It shows strong correlations.

This is a heat map of the correlation matrix. In the Figure 6, negative correlations are in blue color and positive correlations in red. We can approximately see the correlation between variables through the depth of color and the Pearson correlation coefficient. Lessons here are very similar to what we learned from the scatter plot matrix.

5.3 Ridge regression

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where independent variables are highly correlated. So it was useful in our situation. We combined all 8 elections and transfer election type and city to numerical ID. The data was separated by training data and testing data equally and randomly. Ridge regression is driven by a tuning parameter called Lambda, which gives various levels of control on the multicollinearity. We set 100 different lambda from 10^{-2} to 10^{10} evenly and used cross-validation to get the best lambda. The goal of cross-validation is to test the model's ability to predict new data that was not used in

estimating it. After that, we used this best lambda to get our optimal model and the result.

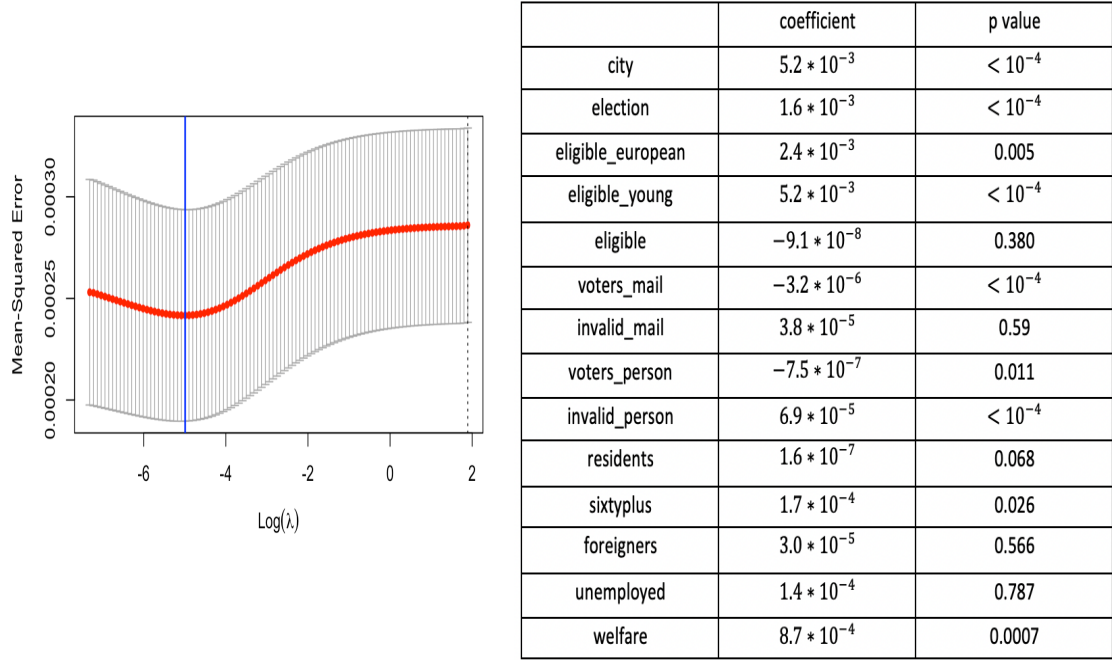
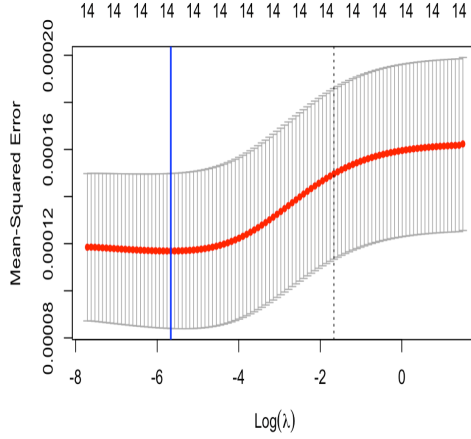


Figure 7: Get the best lambda by cross validation(left). The best lambda has the minimal mean squared error. The summary of final model using the best lambda for all elections(right). *eligible_european*, *eligible_young* and *welfare* have a positive relationship with the rejected rate.

In Figure 7, the blue line in the left plot means the best lambda, which has the minimal mean squared error. The table at the right hand shows the coefficient and p value of each variable in the final model. As we can see, *city*, *election*, *eligible_european*, *eligible_young*, *voters_mail*, *voters_person*, *invalid_person*, *sixtyplus* and *welfare* are statistically significant. And *eligible_european*, *eligible_young* and *welfare* have a positive relationship with the rejected rate.

Now let's focus on Bremen, so that we could take *taxes* into consideration. The procedure was similar as before, and we just plotted the result and analysed it.



	coefficient	p value
election	$2.0 * 10^{-3}$	$< 10^{-4}$
eligible_european	$1.1 * 10^{-3}$	0.076
eligible_young	$5.4 * 10^{-3}$	$< 10^{-4}$
eligible	$-1.5 * 10^{-7}$	0.099
voters_mail	$-3.8 * 10^{-6}$	$< 10^{-4}$
invalid_mail	$6.9 * 10^{-5}$	0.008
voters_person	$-1.6 * 10^{-6}$	$< 10^{-4}$
invalid_person	$4.3 * 10^{-5}$	$< 10^{-4}$
residents	$6.6 * 10^{-7}$	0.455
sixtyplus	$1.5 * 10^{-4}$	0.175
foreigners	$5.5 * 10^{-5}$	0.015
unemployed	$-2.9 * 10^{-3}$	0.235
welfare	$6.5 * 10^{-4}$	0.001
taxes	$-4.4 * 10^{-3}$	$< 10^{-4}$

Figure 8: Get the best lambda by cross validation(left). The best lambda has the minimal mean squared error. The summary of final model for Bremen using the best lambda for all elections(right). *eligible_young*, *foreigners* and *welfare* have a positive relationship with the rejected rate, while *taxes* has a negative relationship with the rejected rate.

In Figure 8, *election*, *eligible_young*, *voters_mail*, *invalid_mail*, *voters_person*, *invalid_person*, *foreigners*, *welfare* and *taxes* are statistically significant. And *eligible_young*, *foreigners* and *welfare* have a positive relationship with the rejected rate, while *taxes* has a negative relationship with the rejected rate.

In both models, *eligible_young* and *welfare* are statistically significant and have a positive relationship with the rejected rate. *foreigners* is only statistically significant in the model of Bremen. And *taxes* has a negative relationship with the rejected rate.

6 Discussion

We conjectured that lower socio-economic status leads to more rejected ballots before doing this project. And the results seem to prove most of it. A higher taxes causes a lower rejection rate. A higher welfare, which means a lower socio-economic status causes a higher rejection rate in both models. Although *foreigners* is only statistically significant in model of Bremen, it has a positive

relationship with the rejected rate. *unemployed* isn't statistically significant in both models. In the future, we should collect more data on other 4 cities, because if we have complete data on them, we may get a more stable result.