Financial Independence Survey - Logistic Regression Model

Barbara Flores

Table of contents

1 Overview

The objective of this project is to address the research question:
What factors contribute to an individual's perception of financial independence?
To accomplish this objective, we will perform logistic regression analysis using a dataset that is a subset of the official results from the 2020 Financial Independence Survey on Reddit. This subset includes responses from individuals who represent themselves (excluding contributions from other household members) and excludes responses from retired individuals. The dataset encompasses **1,998 rows** and **65 variables**, covering information such as income contributors, the financial impact of the pandemic, political affiliation, demographics, details about financial independence, employment status, and various financial aspects. The data has been sourced from Reddit. For additional details regarding the data dictionary and source information, please refer to the www.openintro.org website.
In our original dataset, 147 individuals, which accounts for 8% of our sample, reported considering themselves financially independent. Therefore, we have unbalanced data, which we should take into consideration in our future analyses.

2 Data cleaning

In the data cleaning process, the following transformations were implemented. For more details on the specific variables that experienced these transformations, please refer to the appendix 6.1.

- The dependent variable **fin_indy** (Are you financially independent?) was transformed from a string variable to a binary numeric variable (0 and 1).
- The variable **political** had 24 categories, which were reduced to 4: 'Democrat,' 'Libertarian Party,' 'Republican,' and 'Other,' because the data had many categories with very few observations. Finally, it was transformed from string to factor.
- The variable **age** was transformed from a string variable to a numeric variable to simplify analysis, with each category assigned a corresponding numeric value (1-8) to represent age brackets.
- The variable **edu** had 11 categories, which were reduced to 4: "High School or Less", "Some College or Trade School", "Bachelor's or Associate's Degree", "Doctorate or Master's Degree", because the data had many categories with very few observations. Then, it was transformed from string to factor.
- The variable **housing** (What is your current housing situation?) had 19 categories, which were reduced to 2: "Own", "Not Own", because the data had many categories with very few observations. Finally, it was transformed from string to factor.
- Out of a total of 29 variables, which constituted numeric-type variables such as Children Expenses, Luxury Expenses, Transportation Expenses, Taxes, Medical Debt, etc., it was observed that they did not contain the number 0 but did have many NA values. The assumption was made that individuals who responded with NA in fields such as Children Expenses did so because they had no associated expenses in that category. For this reason, in these 29 variables, NA values were replaced with 0. The complete detail of the variables can be reviewed on the appendix page.
- The variable **total_assets** was created, which is formed by the sum of 8 variables such as cash, investment accounts, crypto, etc. For more details, refer to the appendix.
- The variable **total_debts** was created, which is formed by the sum of 7 variables such as student loans, mortgage, medical debt, etc. For more details, refer to the appendix.
- The variable **total_expenses** was created, which is formed by the sum of 12 variables such as necessities expenses, children expenses, transportation expenses, etc.
- Finally, observations with NA values for any of the above variables were removed, resulting in 11 observations out of the original total of 1,998, leaving us with 1,987 final observations.

3 Modeling

3.1 Logistic Regression Model

To address our research question, we will employ a Logistic Regression Model. This model is suitable for this situation as it allows us to examine how various predictor variables influence the probability of an individual perceiving financial independence. Given that the variable of interest is binary (yes/no), logistic regression will provide us with estimations of log probabilities and coefficients that will aid in understanding the direction and strength of the relationship between the considered factors and the perception of financial independence.

3.2 Variable Selection

After analyzing the variables provided in the dataset, those that could be considered to influence the financial independence variable were identified. The following variables were selected a priori as predictors: **Age**, **Political** (With which political party do you most closely identify?), **Education** (What is the highest level of education you have completed?), **Housing** (What is your current housing situation? **Rent**, Own…) , **Total Debts**, **Total Assets**, **Total Expenses**, **2020 Gross Income**, **2020 Investments, 2020 savings**.

For the variable age, it is expected that as a person gets older, they are more likely to have achieved greater economic stability and, therefore, are more inclined to being financially independent. Regarding the political variable, it is anticipated that individuals belonging to parties that oppose, for example, tax increases, see economic independence as feasible through their own assets. It is expected that with higher education, there is a greater likelihood of being financially independent. Concerning the housing variable, it is expected that owning a home will significantly impact the dependent variable. In the case of the variables debt, assets, expenses, 2020 income, and 2020 investments, they are considered closely related to the dependent variable, as the available wealth of an individual, i.e., income minus expenses, assets minus liabilities, will determine whether they can achieve economic independence or not.

4 Results

4.1 Model results

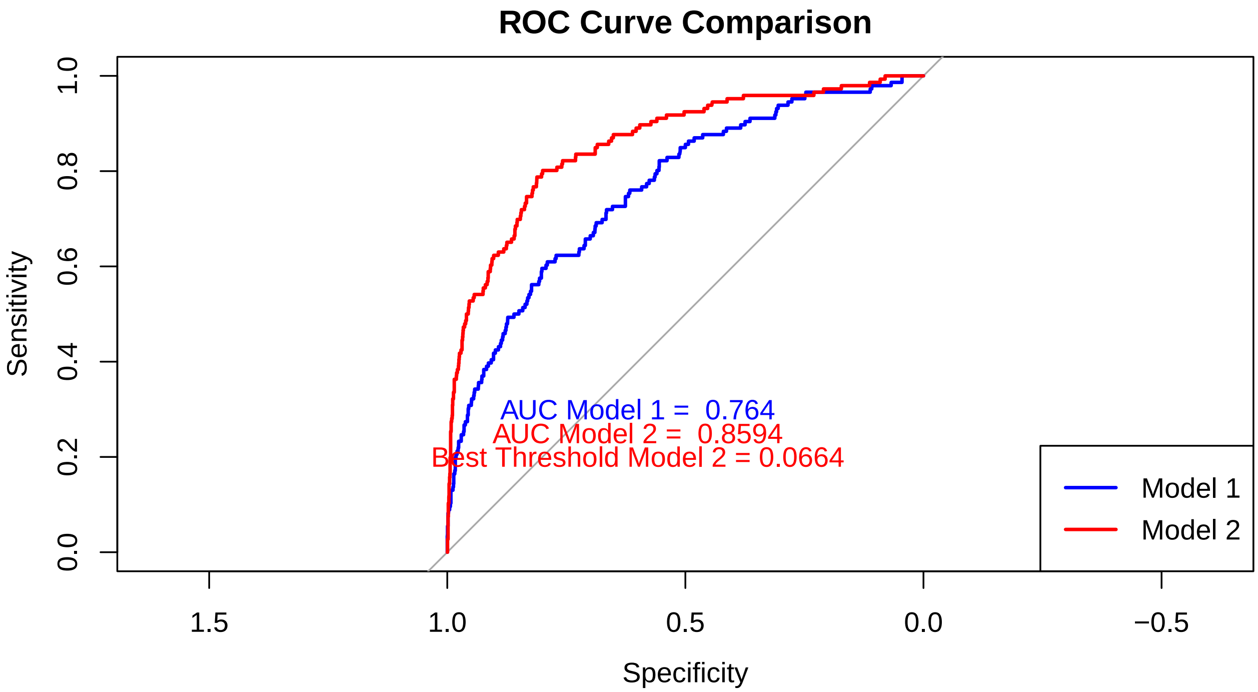
An initial Model 1 was constructed using the specified variables, but it was observed that some of these variables lacked significance. Model 1 had an AIC of 925 (you can refer to the details of Model 1 in Appendix 6.2). Following this, influential data points were identified through Cook's distance analysis and subsequently removed from the dataset. This process led to the development of Model 2, the results of which are presented below. The Cook's distance plots for both models can be found in Appendix 6.3.

```
Call:
glm(formula = factor(fin_indy) ~ age + political_grouped + edu_grouped +
    housing_grouped + total_debts + total_assets + total_expenses +
    `2020_gross_inc` + `2020_invst_save`, family = "binomial",
    data = reddit_finance_sub2)

Coefficients:
                                           Estimate Std. Error z value
(Intercept)                              -4.585e+00  7.490e-01  -6.121
age                                       4.441e-01  7.067e-02   6.284
political_groupedLibertarian Party        2.788e-01  4.194e-01   0.665
```

```
political_groupedOther                       4.535e-01  2.212e-01   2.051
political_groupedRepublican                  7.911e-01  3.065e-01   2.581
edu_groupedSome College or Trade School     -1.292e+00  8.583e-01  -1.506
edu_groupedBachelor's or Associate's Degree -6.132e-01  6.904e-01  -0.888
edu_groupedDoctorate or Master's Degree     -5.694e-01  6.968e-01  -0.817
housing_groupedOther                        -1.164e+01  4.252e+02  -0.027
housing_groupedOwn                          -7.762e-02  2.431e-01  -0.319
total_debts                                 -1.551e-06  4.303e-07  -3.604
total_assets                                 5.117e-07  7.025e-08   7.284
total_expenses                              -2.515e-06  9.701e-07  -2.592
`2020_gross_inc`                             1.034e-06  5.952e-07   1.737
`2020_invst_save`                           -2.230e-06  1.266e-06  -1.762
                                            Pr(>|z|)
(Intercept)                                 9.29e-10 ***
age                                         3.29e-10 ***
political_groupedLibertarian Party          0.506296
political_groupedOther                      0.040301 *
political_groupedRepublican                 0.009852 **
edu_groupedSome College or Trade School     0.132158
edu_groupedBachelor's or Associate's Degree 0.374383
edu_groupedDoctorate or Master's Degree     0.413827
housing_groupedOther                        0.978169
housing_groupedOwn                          0.749489
total_debts                                 0.000314 ***
total_assets                                3.25e-13 ***
total_expenses                              0.009533 **
`2020_gross_inc`                            0.082349 .
`2020_invst_save`                           0.078138 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1042.89  on 1983  degrees of freedom
Residual deviance:  794.55  on 1969  degrees of freedom
AIC: 824.55

Number of Fisher Scoring iterations: 13
```

Model 2 exhibits a significantly improved AIC of 824.55 compared to the previous Model 1 (925), indicating a better model fit and greater explanatory power. The results suggest that age, political affiliation (specifically, affiliation with the Republican party), total debts, total assets, total expenses, 2020 gross income, and 2020 savings and investments are factors that influence the likelihood of perceiving financial independence. However, education and housing situation did not show significant associations in this model. The coefficients in the logistic regression model represent the log odds. For example, in our model, an increase in the age bracket by 1 unit results in a log odds increase of 0.4441. This means that for each step up in age bracket, the odds of perceiving financial independence increase by a factor of exp(0.4441), which can be interpreted as a percentage change in the odds.

4.2 ROC Curve



To assess the logistic regression of the model, ROC curves were plotted for both Model 1 and Model 2. It is observed that the AUC of Model 1 is 0.764, while the AUC of Model 2 is 0.8594. This indicates that Model 2 exhibits superior performance in terms of its ability to distinguish between the categories of the variable of interest (in this case, the perception of financial independence).
In this plot, we can observe that our "Best Threshold Model 2" is 0.0664. This value represents the optimal threshold for classification in Model 2.

4.3 Confusion Matrix

We used this threshold to create our confusion matrix (Appendix 6.4). The main performance indicators are:

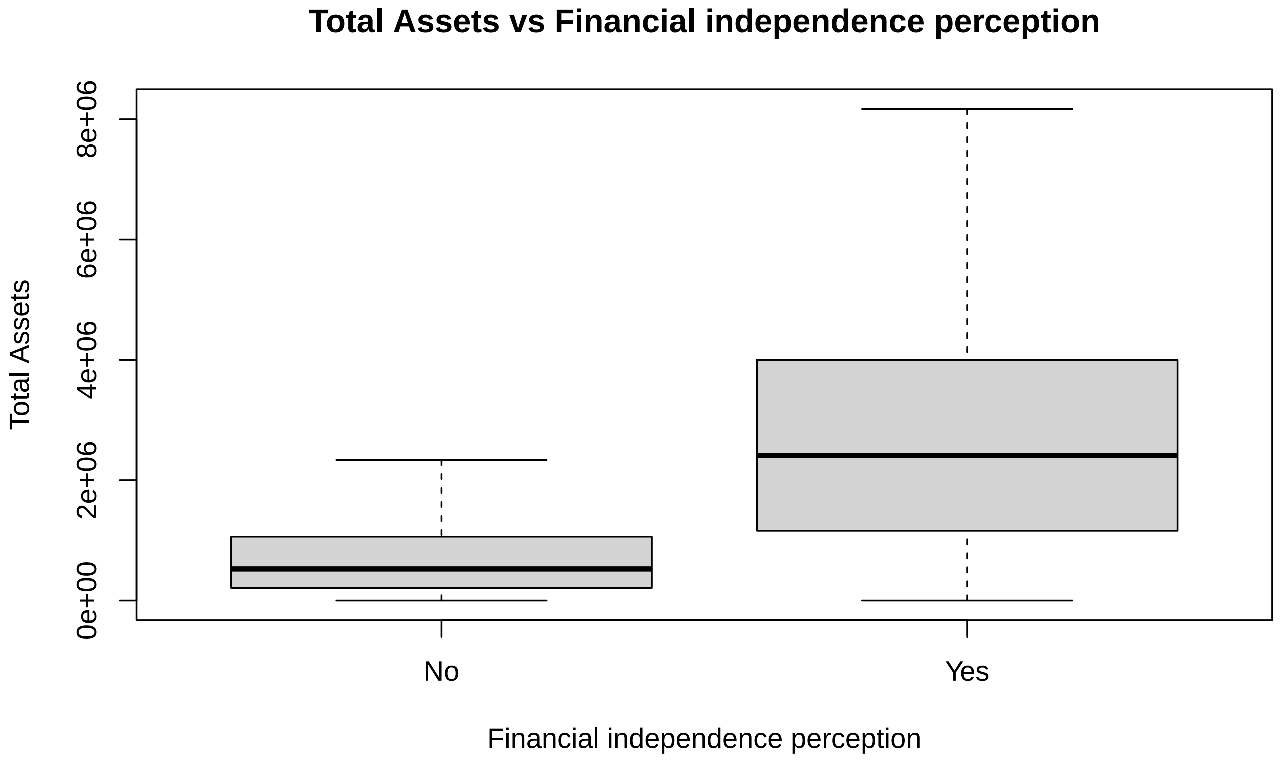
Prediction	No	Yes
No	1318	53
Yes	524	93

Accuracy : 0.7098 Kappa : 0.1418 Sensitivity : 0.63699 Specificity : 0.71553

First, we look at **Accuracy** which measures the overall correctness of our model's predictions. In this case, our model achieves an accuracy of 70.98%, indicating that it correctly predicts whether an event will occur or not in nearly 71% of cases. Additionally, we consider the **Kappa statistic**, which assesses the agreement between our model's predictions and random chance. A Kappa value of 0.1418 suggests a slight agreement beyond what would be expected by chance. Practical applicability is revealed by a **Sensitivity** of 63.7%, meaning it correctly identifies 63.7% of positive cases, and a **Specificity** of 71.553%, correctly identifying 71.6% of negatives. These indicators collectively provide insights into the model's performance and its ability to make accurate predictions.

4.4 Results figure

After generating our model, we can, observe the behavior of one of the variables that turned out to be most influential in predicting our dependent variable, Total Assets. We can see how the group of individuals who are financially independent indeed has significantly higher assets.



5 Conclusion and Future work

5.1 Conclusion

The analysis looks to investigate the factors contributing to an individual's perception of financial independence. The findings indicate that various variables play a crucial role in shaping this perception. Age was a significant factor, with older individuals more likely to perceive themselves as financially independent. Political affiliation, particularly with the Republican party, emerged as an influential factor, suggesting that certain political ideologies may influence one's financial perspective. Furthermore, total debts, total assets, total expenses, 2020 gross income, and 2020 savings and investments were identified as significant contributors to financial independence perception. These variables reflect the complex interplay between economic resources, expenses, and income.

5.2 Future work

However, the analysis had limitations, including the removal of influential data points to enhance model performance. This process might have excluded valuable information. Moreover, variables such as education and housing situation did not show significant associations in this model, suggesting that other factors may be at play. To further strengthen the analysis, future work should involve refining data collection, expanding the dataset and implementing **cross-validation**. These steps will contribute to addressing limitations and enhancing the analysis's robustness and accuracy.

6 Appendix

6.1 Variable Transformation

Variable	**Transformation**
political	• From 24 original categories, it was grouped into 4: “Democrat” “Libertarian Party” “Republican” “Other” –> (“American Solidarity Party”,“Citizens Party of the United States”,“Communist Party USA”,“Constitution Party”,“Freedom Socialist Party”,“Green Party”,“Humane Party”,“Independent American Party”,“Legal Marijuana Now Party”,“N/A”,“NA”,“National Socialist Movement”,“None”,“Objectivist Party”,“Peace and Freedom Party”,“Socialist Action”,“Socialist Alternative”,“Socialist Equality Party”,“Socialist Party USA”,“United States Marijuana Party”,“United States Pirate Party”,“Workers World Party”) • Transformed into a factor.
age	• It was transformed from string to numeric (1-10)
edu	• From 11 original categories, it was grouped into 4: “High School or Less” –> (“High School diploma / GED”, “Less than high school”, “Some high school”) “Some College or Trade School” –> (“Some college, no degree”, “Trade School Degree”) “Bachelor’s or Associate’s Degree” –> (“Associate’s Degree”, “Bachelor’s Degree”) “Doctorate or Master’s Degree” –> (“Doctorate / Post Graduate”, “Doctorate / Post-Graduate”, “Graduate degree”) • Transformed into a factor.
fin_indy	• It was transformed from string to binary

housing	• From the original 19 categories, it was grouped into 2: “Own” –> (Own”, “Own a house but currently living in a RV”, “Own a house that I rent out but I rent an apartment in another state”, “Rent and own”, “Rent my apartment, but own a home I rent out”, “Rent temporarily while building new home”) “Not Own” –> (“Government housing”, “Homeless”, “House owned by parents”, “Houseless/nomadic - living in a car voluntarily”, “Live with family or friends”, “Live with family while building a house”, “Live with partner and split costs of owning”, “Live with partner in their house”, “Liveaboard boat”, “Living with parents”, “NA”, “Own and live in a van”, “Provided by employer or school”, “Rent”) • Transformed into a factor.
home_value	• NA replaced with 0 • sum in total_assets variable
broke rage_accts_tax	• NA replaced with 0 • sum in total_assets variable
retire ment_accts_tax	• NA replaced with 0 • sum in total_assets variable
cash	• NA replaced with 0 • sum in total_assets variable
invst_accts	• NA replaced with 0 • sum in total_assets variable
spec_crypto	• NA replaced with 0 • sum in total_assets variable
invs t_prop_bus_own	• NA replaced with 0 • sum in total_assets variable
other_val	• NA replaced with 0 • sum in total_assets variable
student_loans	• NA replaced with 0 • sum in total_debts variable

mortgage	• NA replaced with 0 • sum in total_debts variable
auto_loan	• NA replaced with 0 • sum in total_debts variable
credi t	• NA replaced with 0 • sum in total_debts variable
_personal_loan	
medical_debt	• NA replaced with 0 • sum in total_debts variable
invst_pr o p_bus_own_debt	• NA replaced with 0 • sum in total_debts variable
other_debt	• NA replaced with 0 • sum in total_debts variable
2020_gross_inc	• NA replaced with 0
2 0 20_housing_exp	• NA replaced with 0 • sum in total_expenses variable
202 0	• NA replaced with 0 • sum in total_expenses variable
_utilities_exp	
2 020_transp_exp	• NA replaced with 0 • sum in total_expenses variable
2020_n ecessities_exp	• NA replaced with 0 • sum in total_expenses variable
2020_lux_exp	• NA replaced with 0 • sum in total_expenses variable
2020_child_exp	• NA replaced with 0 • sum in total_expenses variable

2 020_debt_repay	• NA replaced with 0 • sum in total_expenses variable
2 020_invst_save	• NA replaced with 0
2020_charity	• NA replaced with 0 • sum in total_expenses variable
2020_ healthcare_exp	• NA replaced with 0 • sum in total_expenses variable
2020_taxes	• NA replaced with 0 • sum in total_expenses variable
2020_edu_exp	• NA replaced with 0 • sum in total_expenses variable
2020_other_exp	• NA replaced with 0 • sum in total_expenses variable

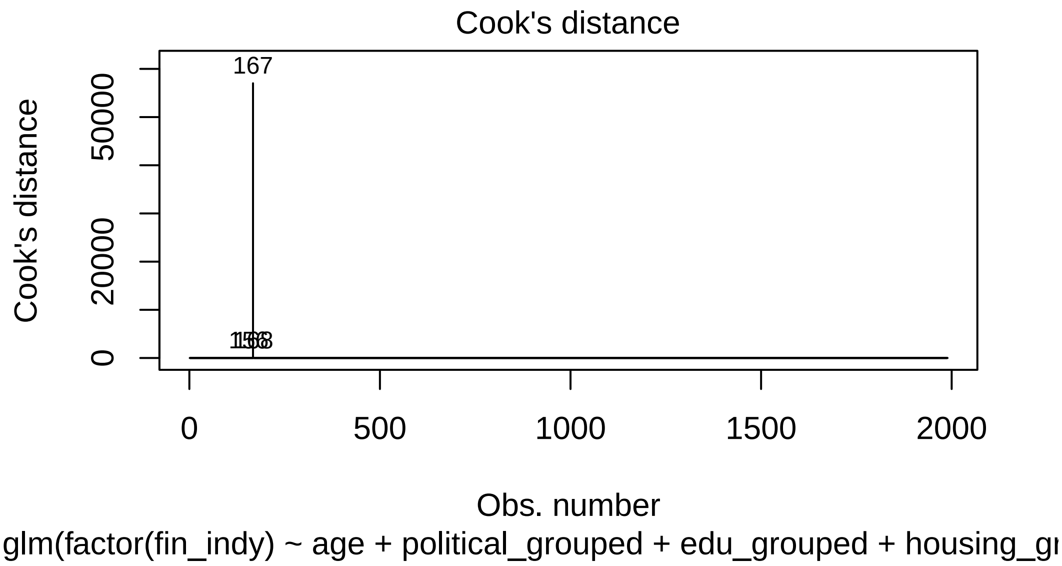
6.2 Results model 1

```
Call:
glm(formula = factor(fin_indy) ~ age + political_grouped + edu_grouped + 
    housing_grouped + total_debts + total_assets + total_expenses + 
    `2020_gross_inc` + `2020_invst_save`, family = "binomial", 
    data = reddit_finance_sub)

Coefficients:
                                              Estimate Std. Error z value
(Intercept)                                 -4.647e+00  6.588e-01  -7.053
age                                          6.334e-01  6.420e-02   9.867
political_groupedLibertarian Party           1.749e-01  3.925e-01   0.446
political_groupedOther                       4.381e-01  2.066e-01   2.120
political_groupedRepublican                  8.530e-01  2.847e-01   2.996
edu_groupedSome College or Trade School     -1.894e+00  7.783e-01  -2.433
edu_groupedBachelor's or Associate's Degree -9.887e-01  5.993e-01  -1.650
edu_groupedDoctorate or Master's Degree     -1.004e+00  6.052e-01  -1.659
housing_groupedOther                        -1.189e+01  4.261e+02  -0.028
housing_groupedOwn                          -1.586e-02  2.310e-01  -0.069
total_debts                                 -5.509e-07  3.867e-07  -1.425
total_assets                                -2.907e-11  1.759e-10  -0.165
total_expenses                               3.282e-07  2.404e-07   1.365
`2020_gross_inc`                            -7.097e-10  5.591e-08  -0.013
`2020_invst_save`                           -1.888e-07  6.126e-07  -0.308
                                            Pr(>|z|)    
(Intercept)                                 1.75e-12 ***
age                                          < 2e-16 ***
political_groupedLibertarian Party           0.65582    
political_groupedOther                       0.03397 *  
political_groupedRepublican                  0.00273 ** 
edu_groupedSome College or Trade School      0.01497 *  
edu_groupedBachelor's or Associate's Degree  0.09901 .  
edu_groupedDoctorate or Master's Degree      0.09718 .  
housing_groupedOther                         0.97773    
housing_groupedOwn                           0.94526    
total_debts                                  0.15427    
total_assets                                 0.86873    
total_expenses                               0.17218    
`2020_gross_inc`                             0.98987    
`2020_invst_save`                            0.75798    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1043.50  on 1987  degrees of freedom
Residual deviance:  895.23  on 1973  degrees of freedom
```

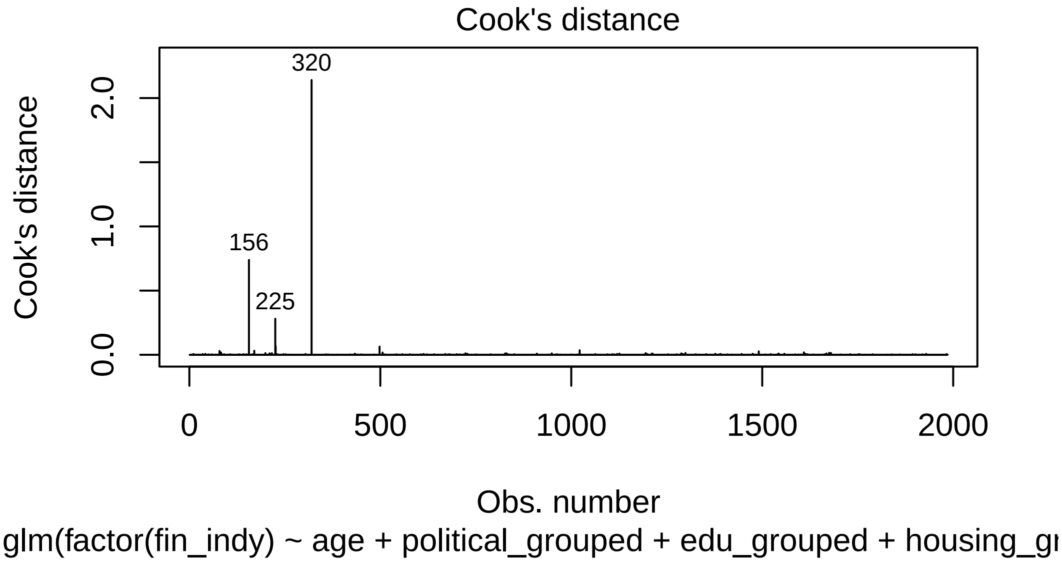
```
AIC: 925.23

Number of Fisher Scoring iterations: 13
```

6.3 Cook's distance

6.3.1 Model 1



6.3.2 Model 2



6.4 Confusion Matrix

6.4.1 Model 2

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1318   53
       Yes  524   93

               Accuracy : 0.7098
                 95% CI : (0.6893, 0.7296)
    No Information Rate : 0.9266
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1418

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.63699
            Specificity : 0.71553
         Pos Pred Value : 0.15073
         Neg Pred Value : 0.96134
              Precision : 0.15073
                 Recall : 0.63699
                     F1 : 0.24377
             Prevalence : 0.07344
         Detection Rate : 0.04678
   Detection Prevalence : 0.31036
      Balanced Accuracy : 0.67626

       'Positive' Class : Yes
```