**IS590 BAO Report**

**Danning Du, Jialu Wang, Yinger Tang**

1. **Preparation**

During our first meeting as a group, we determined some of the KPI’s and summary metrics we would like to use for our constituent base.

* Amount raised
* Number of Events and event attendees
* Volunteer
* Demographics (Geographic, Gender, Age, Marital status)
* Family relation

We decided to combine some of the attributes together to represent the donation trend and patterns, but we also kept in mind that our KPI’s should be actionable. In this step, we also made many assumptions on what kind of donors are more likely to donate.

Before we tested our assumptions and dived into SQL, we ran all the six tables in SQL so that we could understand the data first. We found that there are 2,000 distinct constituents in our donor base, but only 1,000 of them made donations. These 1,000 constituents made 6,163 transactions in total. EVENT can only join with EVENT\_ATTENDEE table on EVENT\_ID. We also read the forum posts so that we knew events and volunteers are not related.

1. **Tableau part**

**2.1 Our approach to the problem**

Based on our assumptions, we wrote SQL queries and exported as .csv files for making data visualizations in Tableau. In this part, we had difficulties in writing some of the queries, which is further explained in the “What we found to be both easy and difficult about the assignment?**”** section below.

Some of the visualizations we made in Tableau also required us to combine several tables or use the tables given.

While making visualizations in Tableau, we discovered that some of the KPI’s show obvious insights on giving patterns, while some of the KPI’s we intended to use seemed insignificant when we put them on plots, or obscure other KPI’s while combining together.

We tried to use different kinds of visualizations to tell the stories of our donor base, but we also considered whether the type of visualization we chose could best represent the data, and effectively in conveying the insights.

Regarding the visual representation, we chose red-blue diverging as our colors of plots, and other neutral, grey, and cool tone colors for titles and labels. We also labeled some other data points that are obviously higher, or smaller to show on the plot. We also chose whether to put on the legends depending on the plot.

* 1. **Assumption we had for visualization at the beginning of the project**
* Married individuals are more likely to donate more. Individuals that have a family relationship with another constituent are more likely to donate. We would use the line graph to compare these factors.
* Older generations and females would donate more than younger people and males respectively. We planned to draw a bubble chart to make comparison, using colors to represent the gender and size to the amount.
* Individuals who come from wealth states would donate more than those from states with weaker economy. We can use a map graph to depict these variables with different colors and sizes.
* The more events we hold, the more event attendees we would have. Stacked bar chart was chosen to show the relationship between these two factors.

**2.3．Whether our assumptions were met**

From the revenue by marital status and family relation, we could see that married individuals and those with family relations to another constituent are more likely to donate at event, but those who are not married and have no family relation also are more likely to donate than those who are not married and have one or more relations.

In the practice, the bubble chart for analyzing gender and age group didn’t work as well as we expected. We had planned to use the color of the bubbles to show whether each age group has more male donators or female donators. However, Tableau was only able to reveal the total number of each gender rather than the percentage of that. So, we gave up the bubble graph and turned to the pyramid chart, dividing two genders into two sides to compare the factors. From the dollars raised by age group and gender, we could see those female donors between age 55 and 64 performs better than we expected, not only on the number of donors but also on the amount raised. Which means that our organization could target this group as potential donors.

The third assumption is partly met according to the map we made. We found that most of our event attendees are from Illinois, Wisconsin and Michigan, and the dollars raised are also mainly concentrated in these areas. We built the map layer as Income per capita, so that we could see that except the areas that have many event attendees, areas that have darker gray (higher income level) also have certain amounts of dollars raised. It’s possible that our foundation is located in Midwestern area, which would explain why people from economically under-developed area are more active in donating.

We found that the more events we had, we did have more event attendees on some level. Before the year 1995, we had very few events and almost no attendees. We also found that between 1995 and 1997, there were a significant amount of event attendees.

* 1. **Easy and difficult parts about the assignment**

The SQL queries are both easy and difficult. The part we found difficult was writing JOIN (LEFT JOIN), and UNION statements, because some CONSTITUENT\_ID has more than one AMOUNT, more than one event attended, and more than one family relations, which made us difficult to join them together.

The Tableau part was also more difficult than we thought. We needed to create calculation fields for some of the variables we had, and also built the group for marital status and family relations. In addition, when put all the components on the dashboard, there was a lot of formatting than we imagined.

The most difficult part was to write SQL queries and choose the type of visualizations based on what we would like to best represent each set of data.

1. **Model part**
   1. Hypothesis

Independent variables: age, gender, marital status, state, family, event, volunteer

Dependent variable: amount

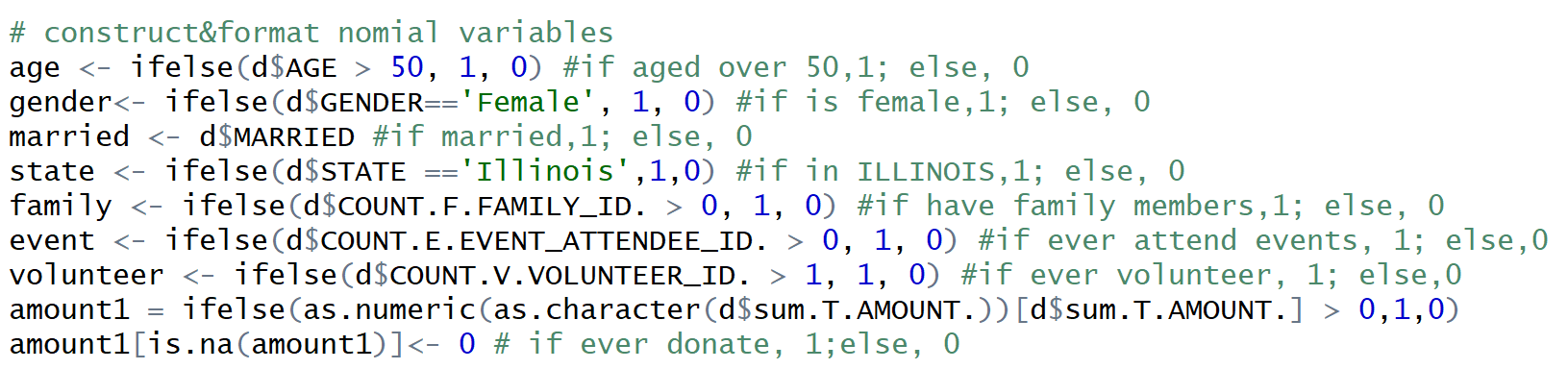
* 1. **In MySQL**

Use queries to get the dataset exported to R.

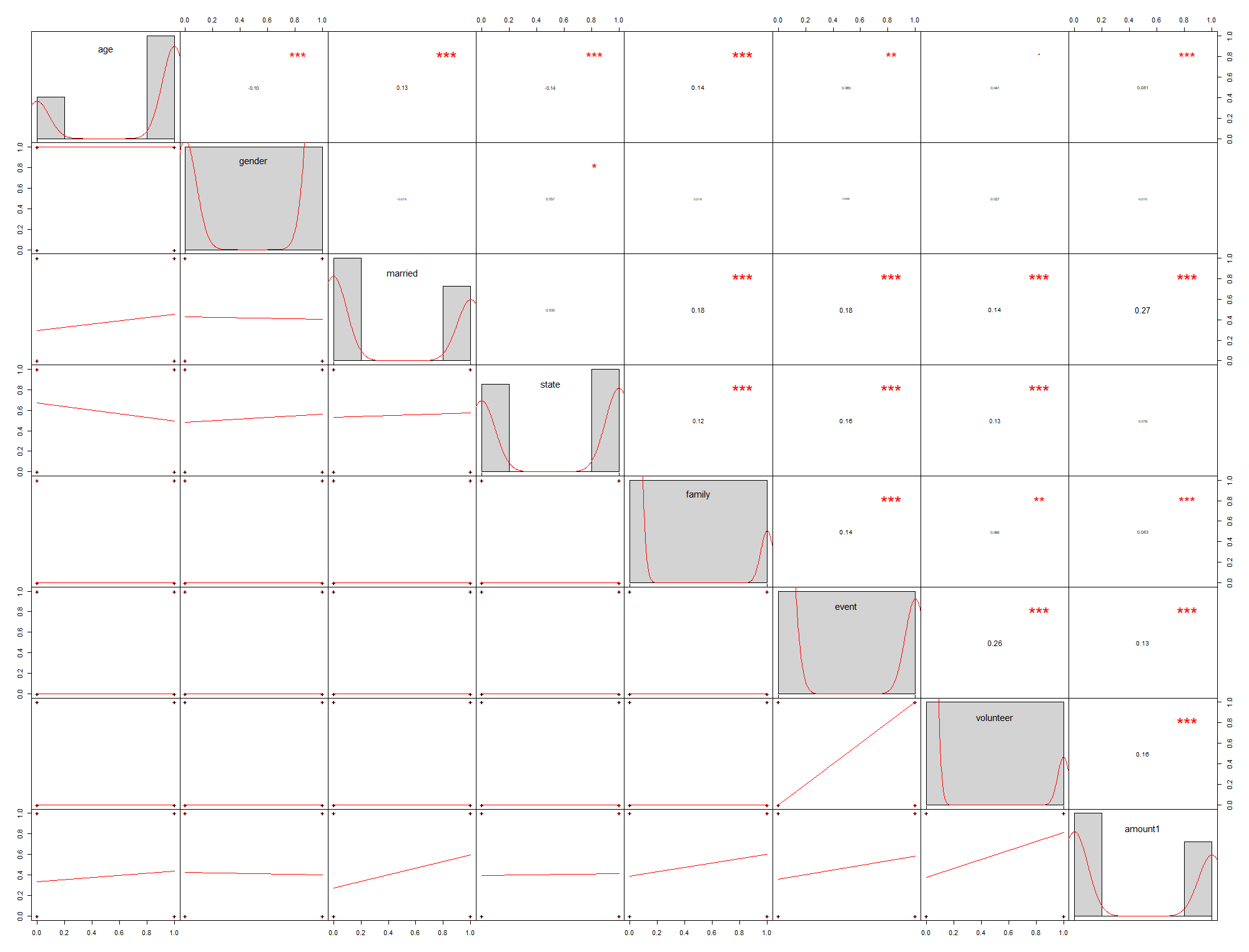
* 1. In R
* OUTPUT:
* Hypothesis

With significance level alpha==0.05, a constituent with attributes: age 50+, male, married, have family members, attend events, do volunteer job, is more likely to donate

* Construct variables



* Examine data

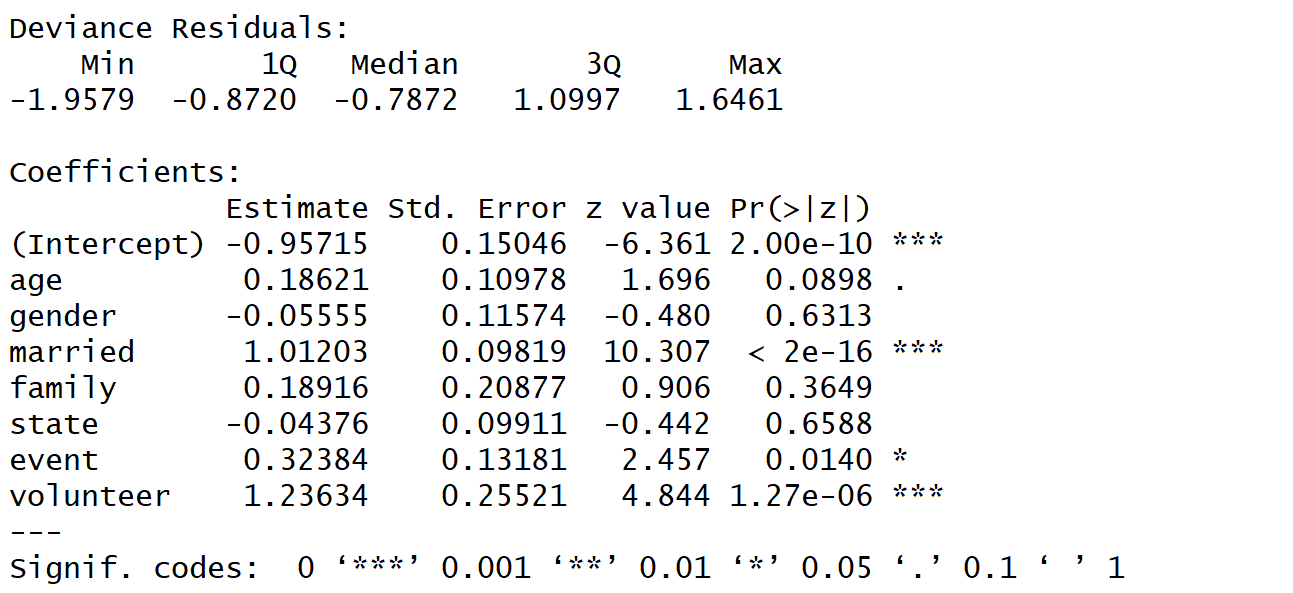


The variables are somehow correlated. The more "\*" it has, the stronger the two variables have the association. Positive correlation coefficient shows positive association and negative correlation coefficient show the negative association.

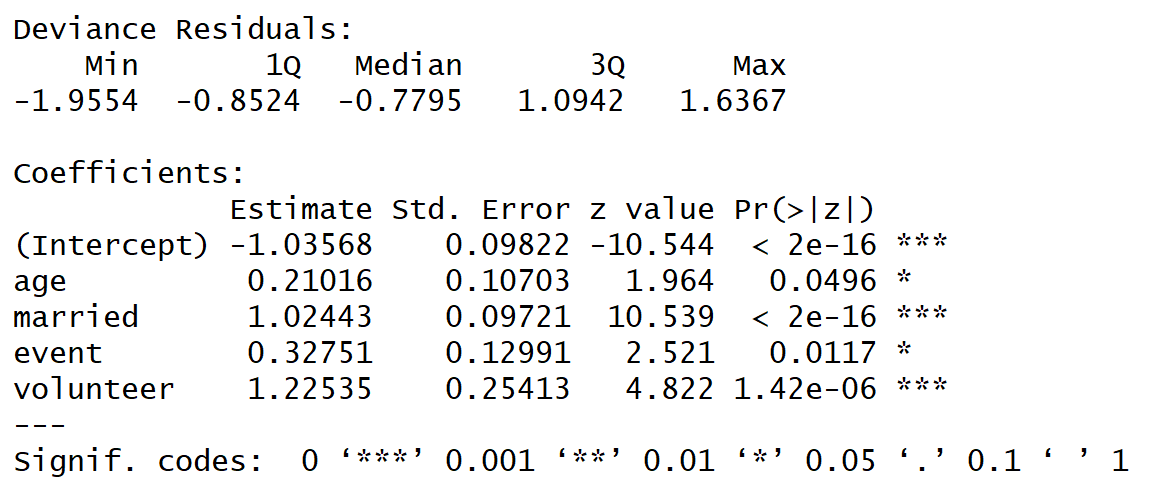
Some variables have significantly more records in one value. (gender: female；family: those have no constituent as family members; event: those who do not attend event; volunteer: those who do not volunteer)

* Build binomial logistic regression model with 2 factor levels

Result: some variables show no significant association based on α=0.05

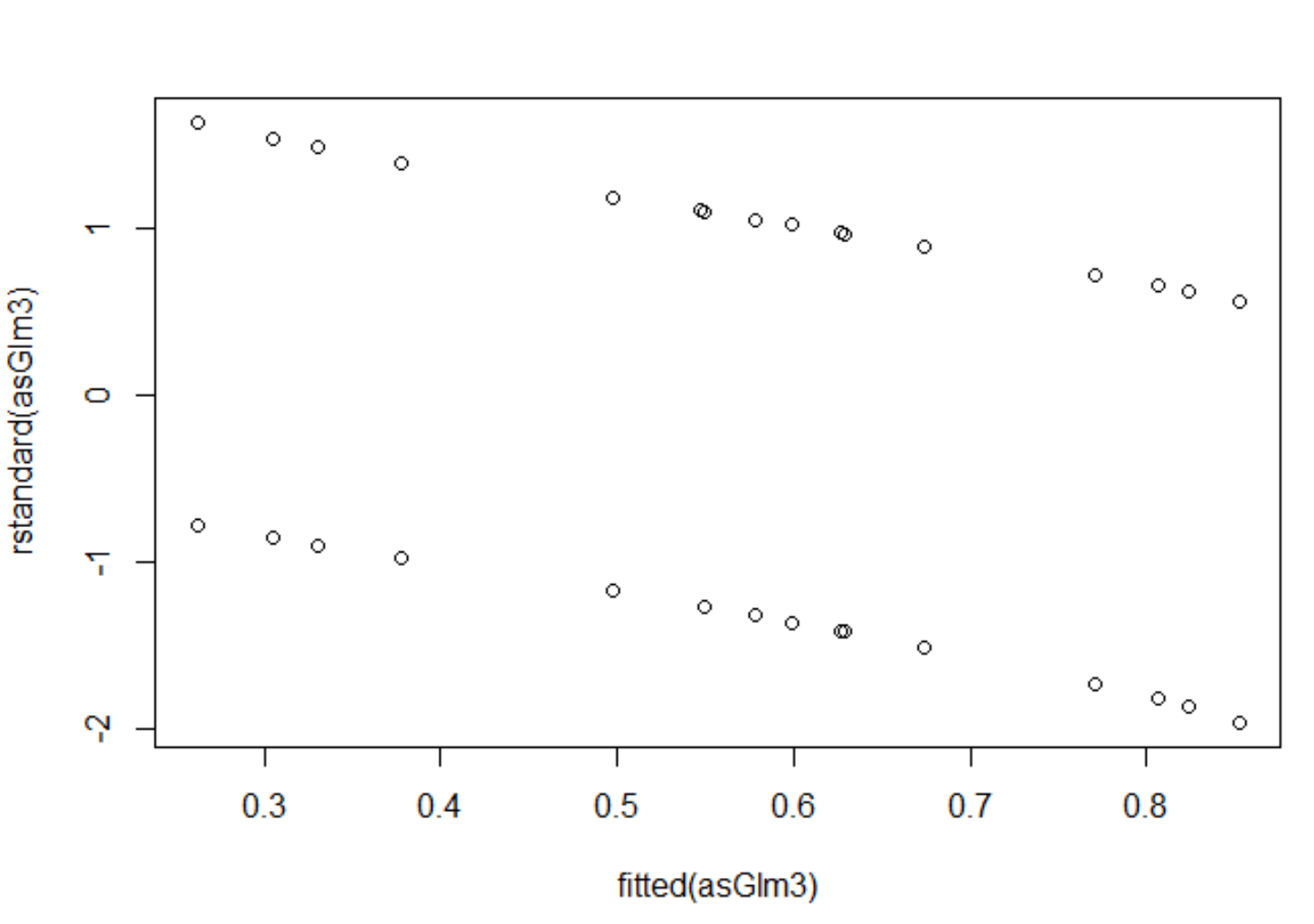


Eliminate the variable with p-value which is higher than 0.05 from highest to smallest each time and generate new GLMs. After eliminating state, gender, family, we’ve got:



* Standardized Residuals

Within ±2, no significant x/y bias

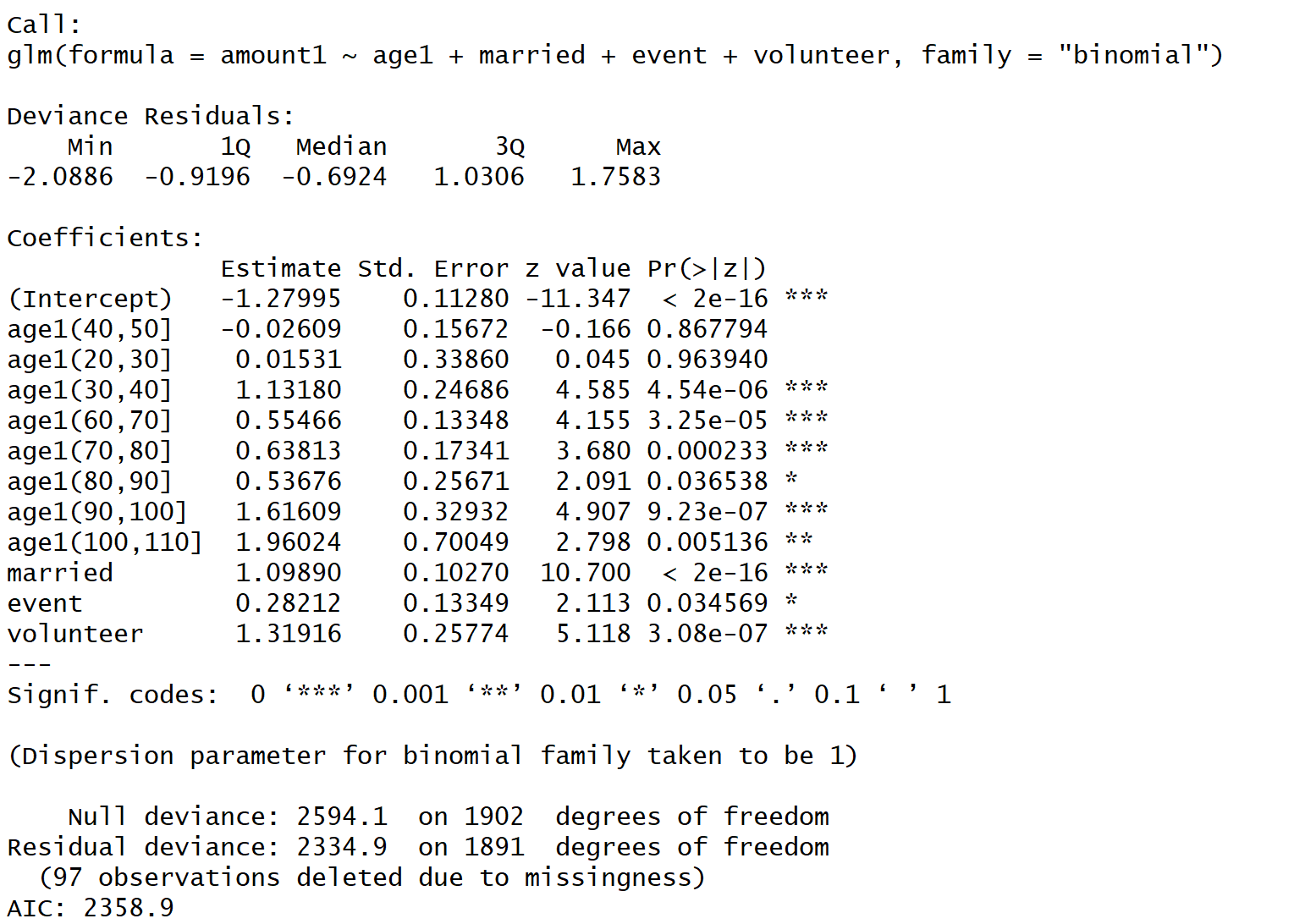


* Result for binomial GLM with 2 factor levels

So our model is with age, marital status, event attendance and volunteer participants. **The result is**: **constituents age 50+, married, actively attending events and actively participate volunteer activities are more likely to donate on a significance level of 0.05.**

* Re-build model with multiple factor levels

We also tried to show divide age and state into smaller-sized levels to improve the model. Age group (50-60] and state ‘Illinois' are selected as reference respectively. We follow the previous operations by eliminating variables with p-value>0.05 each time. After building and testing the model, we get:



Within age, ages with class 30-40,60-70, 70-80, 80-90, 90-100, 100-110 are more likely to donate than age class 50-60 at a 0.05 significance level. Level (100-110] are most likely to donate more.

As to state, unfortunately, we can only say that this factor is still not significantly associated with donation at α=0.05.

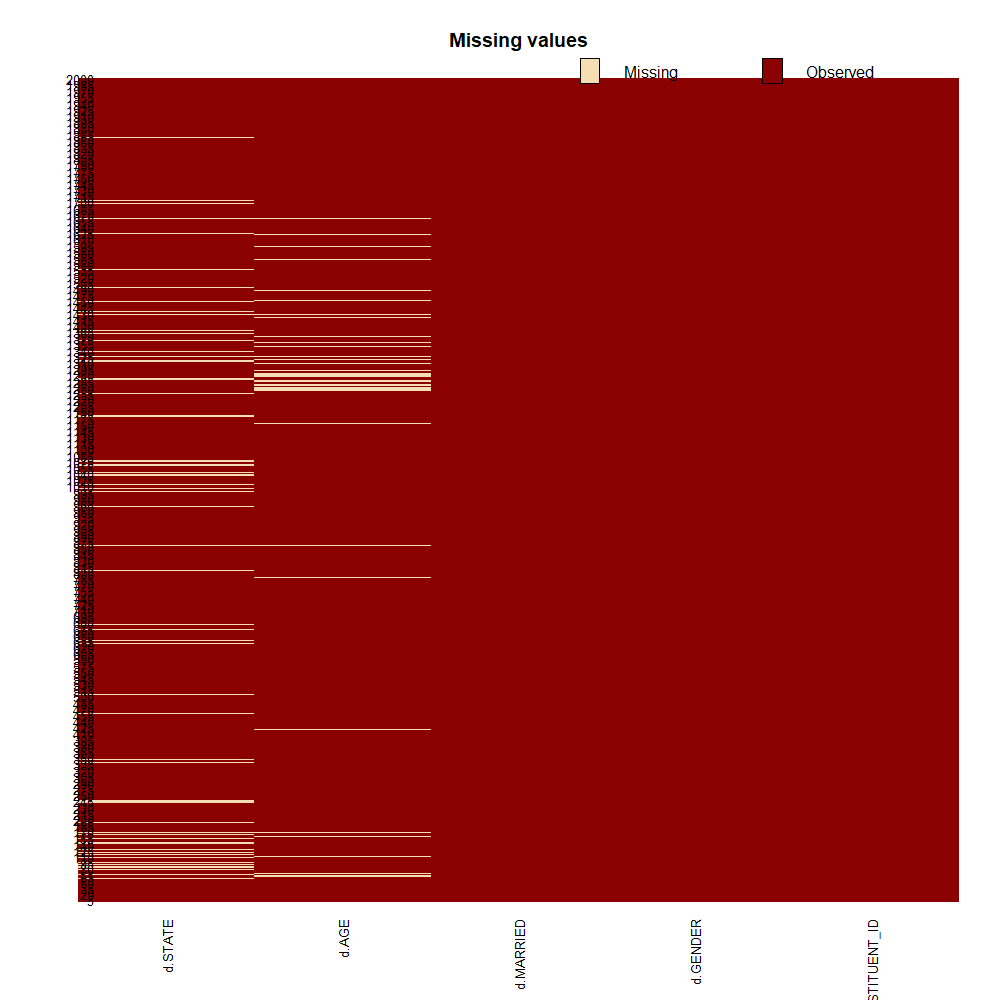
Other variables do not change much as gender and family are still not significant in the new model.

* Attempt on building multinomial logistic regression

We tried function multinom() and build training and test data, but only get a misclassification error of NaN, which means something I have no idea of goes wrong…

* Missing Value

We also find that there are missing values in the table CONSTITUENT, so we generate a miss map of the table, which shows that STATE and AGE contain NULL values.



* 1. **Limitations**
* A multinomial logistic model may have better implications. We tried the function multinom in library nnet but it is way into machine learning and beyond our scope.
  1. **Difficulties**
* how to classify each attribute
* how to determine/deal with the correlation

**URL of Finished Tableau Dashboard**

<https://tableau.admin.uillinois.edu/views/DanningJialuYinger-DonorBaseKPIsDashboard/DonorDashboard?:iid=10>