**Challenge Problem I**

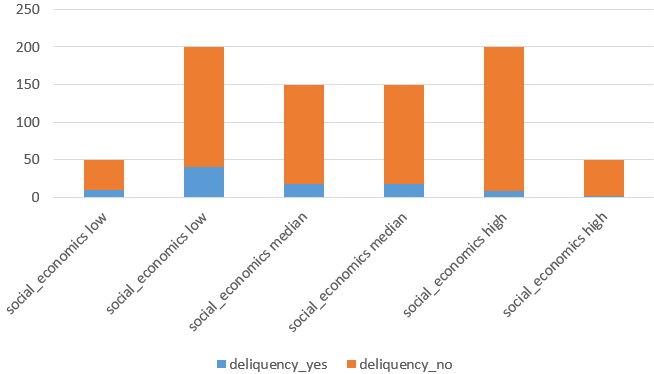
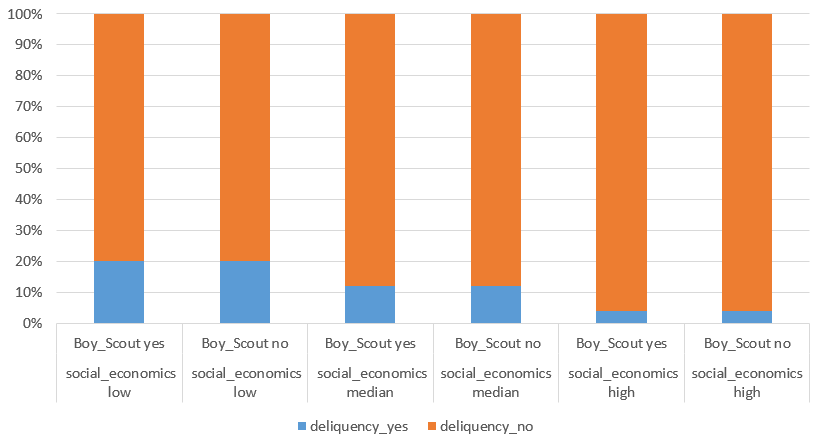
Yc3356 Yi Chen

**Abstract:** In this report, I will use the method of logistic regression to make an analysis of the problem about whether a boy will be delinquent or not based on two potential factors (Social Economic Status and In Boy Scout or not). The analysis focus on four parts: exploratory data analysis, logistic regression estimation, likelihood ratio tests and prediction.

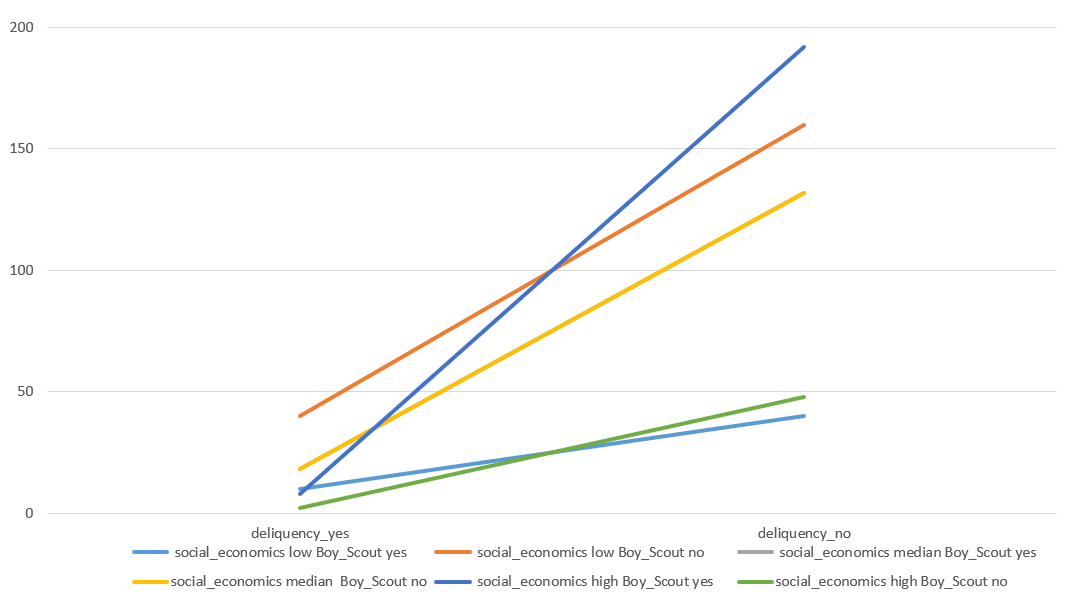
**Key words:** Logistic Regression; Likelihood Ratio Test; Prediction

**1 Exploratory Data Analysis**

In this challenge problem, we are given the data with three variables (social economics status, whether in boy scout and delinquency). And for each situation, we are provided with the corresponding frequency.

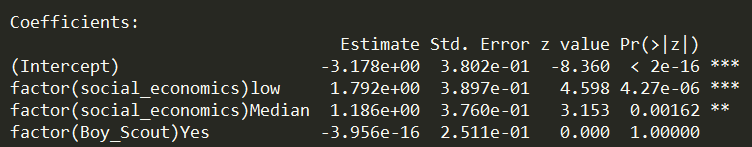
The first question I faced is to decide is among these three variables, what is the respondence? Clearly, in reality, social economics status and whether in boy scout is the unchangeable background for every individual. These two factors may will inference the probability of delinquency. And we may also want to know how much these two factors may correlated with delinquency. (Regression model cannot tell the real causations since the data we get usually not based on the Completely Randomized Design, but we may can calculate the associations). Thus, for this question, I choose delinquency (Yes, No) to be the response variable.

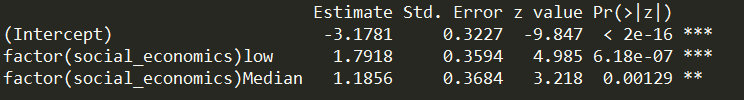
As we can see from the plot, we may find some possible trends. First, relatively, boy from higher social economics status may have lower percentage of being delinquent. Second, the relative difference in delinquency for boy in scout and not in every social status are is very small. Third, for the boy in scout or not the difference of the number of being delinquent have relative small difference in median social economics status but high in other two status. (new question: whether there may exist some interaction effect?)

The plot on the left suggests that there may have some interaction effect between whether in boy scout and social economics status. Based on these finding, I first build the following model:

**2 Logistic Regression and Estimation**

Using R, we can quickly get the results from the model and data. (To see the code in appendix).

As we can see from the result. The p-value of the whether in boy scout is almost equal to 1, which indicates that this factor would not inference the response. Thus, we redo the regression without this factor

We get the new model:

Here, let’s assume this model is reliable (I will do the hypothesis tests later). We can draw the conclusions that:

1. for (intercept): a boy who has high level social economics status, on average, we have 95% confidence estimate that the odd of this boy is delinquent is . For confidence interval: on average, we have 95% confidence estimate that this multiplicative factor would between and .

2. for (social economics low): on average, we have 95% confidence estimate the odds odd of a boy has low level social economics status is delinquent is to be times the odds that boy who has high level social economics status (a 500% increase roughly). For confidence interval: on average, we have 95% confidence estimate that this multiplicative factor would between and .

3. for social economics median): on average, we have 95% confidence estimate the odds odd of a boy has median level social economics status is delinquent is to betimes the odds that boy who has high level social economics status (a 227% increase roughly). For confidence interval: on average, we have 95% confidence estimate that this multiplicative factor would between and .

**3 Likelihood Ratio Tests**

4 Prediction