**Challenge Problem I**

Yc3356 Yi Chen

**Report**

* **Summary**

In this report, I will make an analysis about the delinquency problem based on two potential factors (Social Economic Status and Boy Scout). The whole report involved four parts: Summary, Exploratory data analysis, conclusion and appendix (include all the mathematical analysis and code). The main mathematical method I used in this report is logistic regression.

* **Exploratory data analysis**

The table below gives frequencies for whether or not in boy scout, delinquency status and socioeconomic status.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Social Economics Status | Boy Scout | frequency | | Odd of Delinquent |
| **Delinquent** | **Not Delinquent** |
| low | Yes | 10 | 40 | 0.25 |
| low | No | 40 | 160 | 0.25 |
| Median | Yes | 18 | 132 | 0.136 |
| Median | No | 18 | 132 | 0.136 |
| high | Yes | 8 | 192 | 0.042 |
| high | No | 2 | 48 | 0.042 |

Here are some basic finding we can get from this table (more plot to see the appendix):

First, boy from higher social economics status have lower odd of being delinquent. And the difference is obvious.

Second, for each social economics status, there is no difference between whether in boy scout or not.

* **Conclusion**

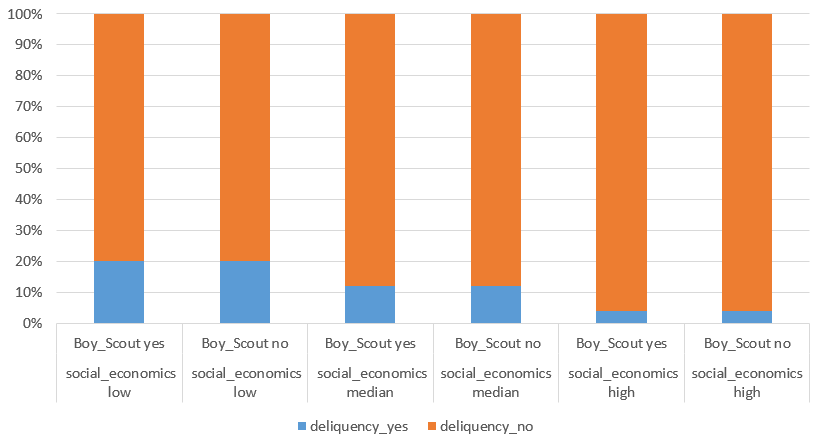
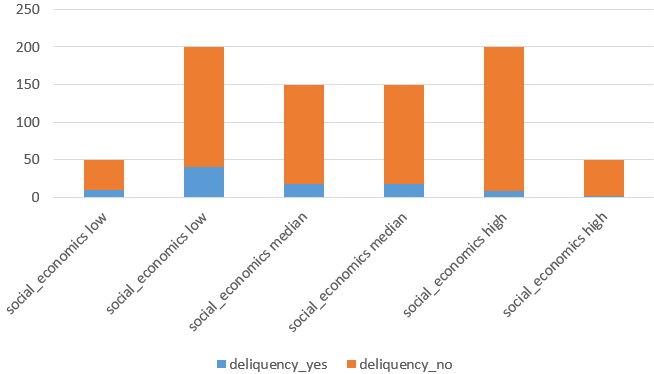
Based on the logistic regression, here are the main conclusions (more detail to see appendix):

1. Social Economics Status has significant inference on the probability of delinquency while the boy scout has not.
2. The odd of delinquency for a boy from median social economics status is about 3.3 times than the odd of delinquency for a boy from high social economics status.
3. The odd of delinquency for a boy from low social economics status is about 6 times than the odd of delinquency for a boy from high social economics status.
4. For a boy who is from high level social economics background, the on average, the probability for him to be delinquent is 4%.
5. For a boy who is from median level social economics background, the on average, the probability for him to be delinquent is 12%.
6. For a boy who is from low level social economics background, the on average, the probability for him to be delinquent is 20%.

**Appendix**

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.

I choose delinquency (Yes, No) to be the response variable. And the social economics status and whether in boy scout as the predictors.

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.

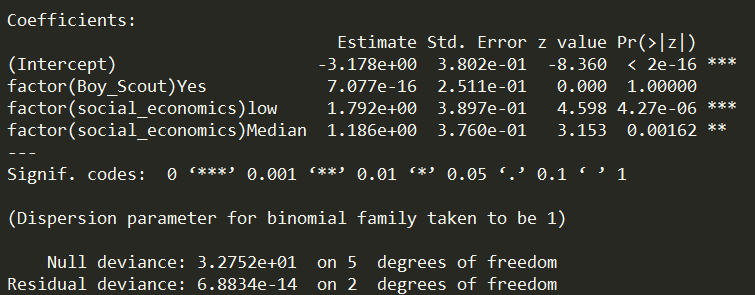
Based on these finding, I first build the following model, where I take Social Economics High and not in boy scout as the base category:

1. **Logistic Regression**

Using R, we can get the estimated:

Just looking at the result, we can find that the parameter for whether in boy scout is almost 0. Thus, we need to do the hypothesis test to find whether we need to keep this.

1. **Hypothesis Tests**

****The result of the regression can be seen from the plot and I will do several hypothesis tests based on this:

* **Test One:**

The first test, I need to do is to ensure whether the form of current model is reliable or not. As we can see from the result, the residual deviance. Thus, we fail to rejectand conclude that: this is the suitable form of model.

* **Test Two:**

Based on likelihood ratio test and the result of the regression: the null deviance is 32.752 with 5 degrees of freedom while the residual deviance is . Thus, the test statistics is 32.752, since the . We can get that: . Thus, we can rejectand conclude that: at least one predictor would have significant inference.

* **Test Three:**

We can see from the result of the table that the p-value of this test is 1. This means we can conclude that and whether in boy scout would have no significant difference.Another way to do this test is use the idea of partial F test. The result would be the same. In the same way, we know that and **.** Social economics status would have a significant inference.

Based on the result of these tests, now I update our model in a new form. Actually, the estimated value of parameters do not change while the degree of freedom changed.

1. **Point Estimation and Confidence Interval**

1. for (intercept): for 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 .

1. **Prediction**

As a prediction based on the new model we can get that:

1. for a boy who is from high level social economics background, the on average the probability for him to be delinquent is 4%.

2. for a boy who is from median level social economics background, the on average the probability for him to be delinquent is 12%.

3. for a boy who is from low level social economics background, the on average the probability for him to be delinquent is 20%.

In order to know how much the result of this finding is reliable, I do the risk analysis. This is shown that all of the predictions have the deviance residual equal to 0. This means that out prediction fit the real data.

1. **Extended thinking**

How can we improve this research?

1. Involve more data (i.e. increase sample size);

2. Involve more predictor (there must have more factor would influence delinquency like education);

3. Multicategory model (distinguish different types of delinquency)

1. **Code**

#input the data (original data)

social\_economics <- c(rep('low',4),rep('Median',4),rep('high',4))

Boy\_Scout <- c(rep(c(rep('Yes',2),rep('No',2)),3))

deliquency <- c(rep(c('Yes','No'),6))

frequency <- c(10,40,40,160,18,132,18,132,8,192,2,48)

data <- as.data.frame(cbind(social\_economics,Boy\_Scout,deliquency,frequency))

# data without boy scout

new\_data <- social\_economics <- factor(c(rep('low',1),rep('Median',1),rep('high',1)))

deliquency\_yes <- c(50,36,10)

deliquency\_no <- c(200,264,240)

data\_1 <- as.data.frame(cbind(social\_economics,deliquency\_yes,deliquency\_no))

# logistic regression

logit1 <- glm(formula = cbind(deliquency\_yes,deliquency\_no)~factor(social\_economics),family = binomial)

summary(logit1)

# prediction

predict (logit1, type = 'response')

# risk analysis

residuals (logit1, type="deviance")