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):

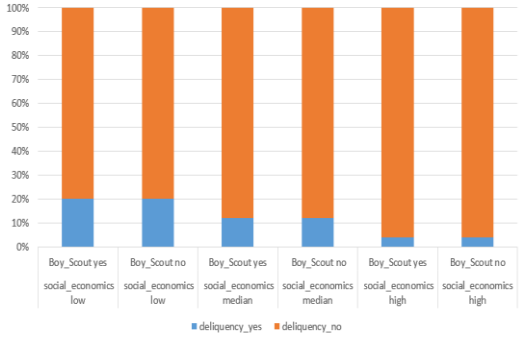
1. Social Economics Status has significant inference on the probability of delinquency while the boy scout has not.
2. The odd of delinquency for a boy from median social economics status is about 3.3 times than the odd of delinquency for a boy from high social economics status.
3. The odd of delinquency 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%.

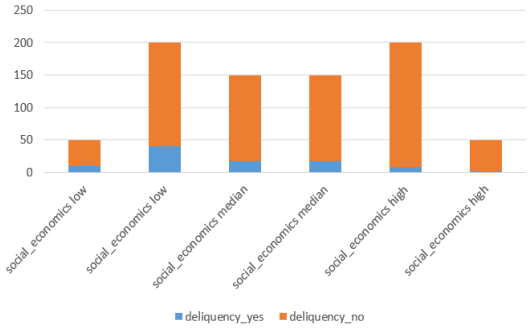
Appendix

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:

$$y = \begin{cases} 1, & being\ deliquent \\ 0, & not\ being\ deliquent \end{cases}, \quad x_{11} = Social - Economic_{low},\ x_{12} = Social - Economics_{median}, \quad x_2 = in\ boy\ socut$$

$$\mathbf{p(y = 1 \mid x_1, x_2) = \frac{e^{\beta_0+\beta_{11}x_{11}+\beta_{12}x_{12}+\beta_2x_2}}{1 + e^{\beta_0+\beta_{11}x_{11}+\beta_{12}x_{12}+\beta_2x_2}}} \qquad \boldsymbol{or} \qquad \boldsymbol{logit(\pi(y = 1)) = \beta_0 + \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_2x_2}$$

2. Logistic Regression

Using R, we can get the estimated:

$$\boldsymbol{\beta_0 = -3.178054\ ;\ \beta_{11} = 1.791759\ ;\ \beta_{12} = 1.185624\ ;\ \beta_2 = 7.076634 \times 10^{-16} \approx 0}$$

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.

3. Hypothesis Tests

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

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -3.178e+00  3.802e-01  -8.360  < 2e-16 ***
factor(Boy_Scout)Yes        7.077e-16  2.511e-01   0.000  1.00000
factor(social_economics)low     1.792e+00  3.897e-01   4.598 4.27e-06 ***
factor(social_economics)Median  1.186e+00  3.760e-01   3.153  0.00162 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.2752e+01  on 5  degrees of freedom
Residual deviance: 6.8834e-14  on 2  degrees of freedom
```

- **Test One:** $H_0: logit(\pi(y=1)) = \beta_0 + \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_2 x_2$

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 $6.8834 \times 10^{-14} \approx 0$. Thus, we fail to reject $\mathbf{H_0}$and conclude that: this is the suitable form of model.

- **Test Two:** $H_0: \beta_{11} = \beta_{12} = \beta_2 = 0$

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 $6.8834 \times 10^{-14} \approx 0$. Thus, the test statistics is 32.752, since the $\mathrm{p} = 3$. We can get that: $32.752 \geq \chi_2^2(0.95)$. Thus, we can reject $\mathbf{H_0}$and conclude that: at least one predictor would have significant inference.

- **Test Three:** $H_0: \beta_2 = 0$

We can see from the result of the table that the p-value of this test is 1. This means we can conclude that $\boldsymbol{\beta_2 = 0}$ 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 $\boldsymbol{\beta_{11} \neq 0}$ and $\boldsymbol{\beta_{12} \neq 0}$. 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.

$$\mathbf{p(y=1 \mid social\_economics)} = \frac{e^{-3.178+1.792x_{11}+1.186x_{11}}}{1+e^{-3.178+1.792x_{11}+1.186x_{11}}} \quad or \quad logit(\pi(y=1))$$

$$= -3.178 + 1.792x_{11} + 1.186x_{11}$$

4. Point Estimation and Confidence Interval

1. for $b_0 = -3.178$(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 $e^{-3.178} = 0.04166891$. For confidence interval: on average, we have 95% confidence estimate that this multiplicative factor would between $e^{-3.9776309} = 0.01872996$ and $e^{-2.4766143} = 0.08402724$.

2. for $b_{11} = 6.001443$ (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 $e^{1.792} = 6.001443$ 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 $e^{0.4832948} = 2.895996$ and $e^{2.6037768} = 13.51468$.

3. for $b_{12} = -3.178$ 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 be $e^{1.186} = 3.273959$ times 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 $e^{-0.4971858} = 0.60824$ and $e^{0.4897023} = 1.63183$.

5. Prediction

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

$$\frac{e^{-3.178}}{1+e^{-3.178}} = 4\% \ ; \ \frac{e^{-3.178+1.792\times1}}{1+e^{-3.178+1.792\times1}} = 12\% \ ; \ \frac{e^{-3.178+1.186\times1}}{1+e^{-3.178+1.186\times1}} = 20\%$$

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.

6. 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)

7. 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")
```