NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Computational Assignment #5: Logistic Regression Computations

MSDS 410

In this assignment we will be calculating the various summative statistics that are associated with logistic regression, as well as fitting logistic regression models and interpreting the results. Students are expected to show all work in their computations. A good practice is to write down the generic formula for any computation and then fill in the values need for the computation from the problem statement. Throughout this assignment keep all decimals to three places, i.e. X.xxx. Students are expected to use correct notation and terminology, and to be clear, complete and concise with all interpretations of results. This computational assignment is worth 50 points. The points associated with each problem are given with the specific question.

Any computations that involve "the log function", denoted by log(x), ***are always meant to mean the natural log function (which will show as ln() on a calculator).*** The only time that you should ever use a log function other than the natural logarithm is if you are given a specific base.

1. For the 2x2 table, determine the odds and the probabilities of texting while driving among males and females. Then compute the odds ratio of texting while driving that compares males to females. (5 points)

Texting While Driving	MALE	FEMALE
YES	30	34
NO	10	6

Probability of men texting and driving: 30/ (30+10) =0.75

Probability of women texting and driving: 34/ (34+6) =0.85

Odds of men texting and driving 30/10=3.0 or 3:1

Odds of women texting and driving 34/6=5.667

2. Download the data file RELIGION.CSV and import it into R. Use R and your EDA skills to gain a basic understanding of this dataset. Please note, there is a variable labeled RELSCHOL. This variable indicates

if a survey respondent attends a religiously affiliated private secondary school (1) or not (0). Use this dataset to address the following questions: (10 points)

a. Compute the overall odds and probability of attending a religious school, assuming this data is from a random sample.

Probability of going to a religious school:

```{r}
sum(mydata$RELSCHOL)/nrow(mydata)
```

```
[1] 0.1277955
```

Odds of going to religious school:

```{r}
sum(mydata$RELSCHOL)/(nrow(mydata)-sum(mydata$RELSCHOL))
```

```
[1] 0.1465201
```

Probability of not going to a religious school:

```{r}
1-(sum(mydata$RELSCHOL)/nrow(mydata))
```

```
[1] 0.8722045
```

Odds of not going to a religious school:

```{r}
(nrow(mydata)-sum(mydata$RELSCHOL))/sum(mydata$RELSCHOL)
```

```
[1] 6.825
```

b. Cross-tabulate RELSCHOL with RACE (coded: 0=non-white, 1=white). What are the probabilities that non-white students and white students attend religious schools? What are the odds that white students and non-white students attend religious schools? What is the odds ratio that compares white and non-white students?

```
          Race
Religious   0   1
        0  76 470
        1  26  54
```

;

```{r}
prop.table(relrace, 2)
```

```
          Race
Religious         0         1
        0 0.7450980 0.8969466
        1 0.2549020 0.1030534
```

Probability non-white students attend religious school:

0.255

Probability white students attend religious school:

0.103

Probability Non-white students do not attend religious school:

0.745

Probability white students do not attend religious school:

0.897

Odds non-white students attend religious school:

0.342

Odds white students attend religious school:

0.317

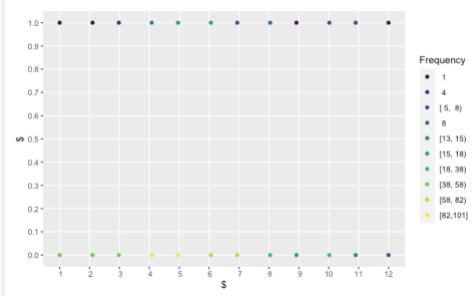
Odds Non-white students do not attend religious school:

2.923

Odds white students do not attend religious school:

8.704

c. Plot RELSCHOL (Y) by INCOME as a scatterplot. The INCOME variable is actually an ordinal variable that is associated with income brackets. This is an old dataset, so for example, INCOME=4 → $20,000-$29,999. Is there a value of INCOME that seems to separate or discriminate between those attending religious schools and those that don't? Create a variable that dichotomizes INCOME based on this value you observed. Call this new variable D_INCOME. Cross-tabulate RELSCHOL with D_INCOME. What are the probabilities that low income students and higher students attend religious schools? What are the odds that lower income students and higher income students attend religious schools? What is the odds ratio that compares lower and higher income students?



```{r}
mydata$D_Income=ifelse(mydata$INCOME>=6,1,0)
```

```{r}
D_IncRace=table(mydata$D_Income, mydata$RELSCHOL)
names(dimnames(D_IncRace)) <- c("D_Income", "RELSCHOL")
D_IncRace
```

```
        RELSCHOL
D_Income   0   1
       0 314  35
       1 200  41
```

```{r}
prop.table(D_IncRace, 2)
```

```
        RELSCHOL
D_Income         0         1
       0 0.6108949 0.4605263
       1 0.3891051 0.5394737
```

***Includes proportion table (D_Income 1=High income, RELSCHOL 1= Yes to religious school)

Odds high-income students attend religious school:

0.205

Odds high-income students do not attend religious school:

4.878

Odds low-income students attend religious school:

0.111

Odds low-income students do not attend religious school:

8.971

d. Plot RELSCHOL (Y) by ATTEND as a scatterplot. The ATTEND variable is the number of times the survey respondent attends a service during a month. Cross-tabulate RELSCHOL with ATTEND. Are the proportion profiles the same for those attending religious school versus not, across the values of the ATTEND variable? Is there a value of ATTEND that seems to separate or discriminate between those attending religious schools and those that don't? Save this value for later.

The proportion of students who go to religious schools and attends service is lower than those that do not attend religious schools. The biggest change occurs in the ATTEND value of 5.

3. First, fit a logistic model to predict RELSCHOL (Y) using only the RACE (X) variable. Call this Model 1. Report the logistic regression model and interpret the parameter estimates for Model 1. Report the AIC and BIC values for Model 1. (3 points)

AIC: 391.12

```{r}
model1=glm(mydata$RELSCHOL~mydata$RACE)
model1
```

```
Call:  glm(formula = mydata$RELSCHOL ~ mydata$RACE)

Coefficients:
(Intercept)  mydata$RACE
     0.2549      -0.1518

Degrees of Freedom: 625 Total (i.e. Null);  624 Residual
Null Deviance:      69.78
Residual Deviance: 67.81        AIC: 391.1
```

```{r}
summary(model1)
```

```
Call:
glm(formula = mydata$RELSCHOL ~ mydata$RACE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2549  -0.1031  -0.1031  -0.1031   0.8970

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.25490    0.03264   7.810 2.44e-14 ***
mydata$RACE -0.15185    0.03568  -4.256 2.40e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.1086661)

    Null deviance: 69.776  on 625  degrees of freedom
Residual deviance: 67.808  on 624  degrees of freedom
AIC: 391.12

Number of Fisher Scoring iterations: 2
```

```{r}
BIC(model1)
```

```
[1] 404.4345
```

4. Next, fit a logistic model to predict RELSCHOL (Y) using only the INCOME(X) variable. Call this Model 2. For Model 2, do the following: (6 points)

    a. Report the logistic regression model and interpret the parameter estimates for Model 2. Report the AIC and BIC values for Model 2. How do these compare to Model 1?

```r
mydata=na.omit(mydata)
```
```r
model2=glm(mydata$RELSCHOL~mydata$INCOME)
model2
print(paste("BIC",BIC(model2)))
```

```
Call:  glm(formula = mydata$RELSCHOL ~ mydata$INCOME)

Coefficients:
  (Intercept)  mydata$INCOME
      0.02907        0.01924

Degrees of Freedom: 586 Total (i.e. Null);  585 Residual
Null Deviance:      66.16
Residual Deviance: 64.77        AIC: 378
[1] "BIC 391.099130663688"
```

Both the AIC and BIC values for model 2 are lower.

b) Use the logit predictive equation for Model 2 to compute PI for each record. Plot PI (Y) by INCOME(X). At what value of X, does the value of PI exceed 0.50? How does this value compare to your visual estimate from problem 2c)?

The value of PI does not exceed 0.5, as the max is at around 0.25.

5. Next, fit a logistic model to predict RELSCHOL (Y) using only the ATTEND(X) variable. Call this Model
3. For Model 3, do the following: (6 points)

a. Report the logistic regression model and interpret the parameter estimates for Model 3. Report the AIC and BIC values for Model 3. How do these compare to Models 1 and 2?

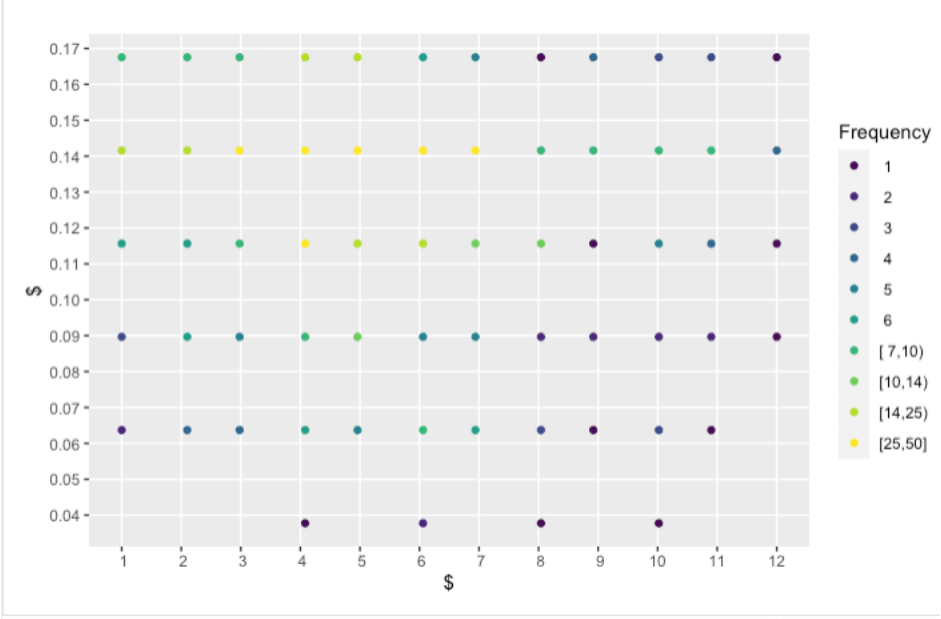
```r
model3=glm(mydata$RELSCHOL~mydata$ATTEND)
model3
print(paste("BIC",BIC(model3)))
```

```
Call:  glm(formula = mydata$RELSCHOL ~ mydata$ATTEND)

Coefficients:
  (Intercept)  mydata$ATTEND
      0.01177        0.02596

Degrees of Freedom: 586 Total (i.e. Null);  585 Residual
Null Deviance:      66.16
Residual Deviance: 65.65        AIC: 385.9
[1] "BIC 399.065650168081"
```

THE AIC and BIC are still below that of model 1 but above that of model 2.

b) Use the logit predictive equation for Model 3 to compute PI for each record. Plot PI (Y) by INCOME(X). At what value of X, does the value of PI exceed 0.50? How does this value compare to your visual estimate from problem 2d)?

```{r}
mydata$PI2=predict(model3)

ggfreqScatter(mydata$INCOME,mydata$PI2)
```



There is not value of X at which PI exceeds 0.5

6. Finally, fit a logistic model to predict RELSCHOL (Y) using RACE, INCOME and ATTEND as explanatory (X) variables. Please consider INCOME and ATTEND to be continuous variables. Call this Model 4. For Model 4, do the following: (9 points)

a. Report the logistic regression model and interpret the parameter estimates for Model 4. Report the AIC and BIC values for Model 4. How does this model compare to Models 1, 2 and 3?

```{r}
model4=glm(mydata$RELSCHOL~mydata$RACE+mydata$INCOME+mydata$ATTEND)
model4
print(paste("BIC",BIC(model4)))
```

```
Call:  glm(formula = mydata$RELSCHOL ~ mydata$RACE + mydata$INCOME +
    mydata$ATTEND)

Coefficients:
  (Intercept)    mydata$RACE  mydata$INCOME  mydata$ATTEND
     0.006259      -0.165732       0.022734       0.031397

Degrees of Freedom: 586 Total (i.e. Null);  583 Residual
Null Deviance:      66.16
Residual Deviance: 61.89        AIC: 355.2
[1] "BIC 377.117749986208"
```

Both the AIC and BIC are the lowest compare to the other models.

b. For those who attend religious service 5 days per month (attend=5) and have a family income of $20-$29,000 (INCOME=4), what are the predicted odds of attending a religious school for white and non-white students?

Non-white:

0.022734(4)+0.031397(5)-0.165732(0)=0.248 or 24.8%

White:

0.022734(4)+0.031397(5)-0.165732(1)=0.082 or 8.2%

c. What is the adjusted odds ratio for race? Interpret this odds ratio.

Odds of a white student going to religious school:

0.331

Odds of a non-white student going to religious school:

3.024

A non-white student is more likely to go to a religious school than a white student.

7. For Models 1, 2 and 3, use the logit models to make predictions for RELSCHOL. Note, you will have to calculate the estimated logit and then convert it into PI_estimates for each module. The classification rule is: If PI < 0.50, predict 0; otherwise predict 1 for RELSCHOL. Obtain a cross-tabulation of RELSCHOL with the predicted values for each model. Compare the correct classification rates for each of the three models. (6 points)

```{r}
mydata$P=predict(model1)
predictTbl=table(mydata$P,mydata$RELSCHOL)
names(dimnames(predictTbl)) <- c("Model1", "RELSCHOL")
predictTbl
```

Model1	RELSCHOL 0	1
0.104508196721312	437	51
0.252525252525253	74	25

```{r}

predictTbl=table(mydata$PI,mydata$RELSCHOL)
names(dimnames(predictTbl)) <- c("Model2", "RELSCHOL")
predictTbl
```

Model1	RELSCHOL 0	1
0.048304241255153	37	1
0.0675393685798301	44	1
0.0867744959045071	49	4
0.106009623229184	101	13
0.125244750553861	82	16
0.144479877878538	67	15
0.163715005203215	58	6
0.182950132527892	20	8
0.202185259852569	15	1
0.221420387177246	18	5
0.240655514501923	14	5
0.2598906418266	6	1

```{r}

predictTbl=table(mydata$PI2,mydata$RELSCHOL)
names(dimnames(predictTbl)) <- c("Model3", "RELSCHOL")
predictTbl
```

Model3	RELSCHOL 0	1
0.0377370488973681	5	0
0.0637005745951718	41	3
0.0896641002929755	47	4
0.115627625990779	100	9
0.141591151688583	239	50
0.167554677386387	79	10

The predictions do not go above 0.5 for any of the models ever, as there is not many who go to religious school in general. Model 2 seemed to have the most diverse probabilities of going to religious school, however, when using INCOME as a distinguishing factor. This also provided the highest possibility of going to religious school with a percentage of about 0.26.

8. In plain English, what do you conclude about the relationship between a student's race/ethnicity, religious service attendance, family income and attending a religious school? (5 points)

When it comes to race/ethnicity, non-white students are more likely to go to religious school. When looking at service attendance, those who go to service more are less likely to go to religious school. When looking at income, those with higher income appear to go to religious schools.

9. Conclusion.

This assignment introduced me to linear regression. This allowed me to familiarize myself with more models other than just linear regression. Sometimes that is not one that fits the best, and when working with categorical or binary numbers, this may work better than a linear regression model. It is always better to have more techniques to analyze trends in data, as some may fit specific data better than others.