NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

# Week 8 Assignment – Computational: 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 |

| Texting While Driving | MALE | FEMALE | **Total** |
|---|---|---|---|
| YES | 30 | 34 | **64** |
| NO | 10 | 6 | **16** |
| **Total** | **40** | **40** | **80** |

P (Yes) = 64/80 **0.8**

M: P (Yes) = 30/40 **0.75**

F: P (Yes) = 34/40 **0.25**

The odds ratio for texting while driving for Males vs. Females: (30/10)/(34/6) = **0.529**

Odds of driving while texting for a male is 0.53 times the odds for a female.

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.

```
> table(Relschol = mydataR$RELSCHOL)

            no =   yes =
RelSchol       0       1    Total
   Count     546      80      626
```

Probability of attending religion school is 80/626 = 0.1278 or 12.78%
Overall odds ratio of attending religious school is 0.1278 / (1 – 0.1278) = 0.147; the alternative way to solve is 80/546 = 0.147

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?

```
> table(Race = mydataR$RACE, Relschol = mydataR$RELSCHOL)
```

| | RelSchol | | | | RelSchol | | |
|---|---|---|---|---|---|---|---|
| **Race** | no = 0 | yes = 1 | **Total** | **Race** | no = 0 | yes = 1 | **Total** |
| non-white=0 | 76 | 26 | **102** | non-white=0 | 74.5% | 25.5% | **100.0%** |
| white=1 | 470 | 54 | **524** | white=1 | 89.7% | 10.3% | **100.0%** |
| **Total** | **546** | **80** | **626** | **Total** | **87.2%** | **12.8%** | **100.0%** |

The probabilities that **non-white** students attend religious schools: 26/102 = 0.255 or 25.5%
The probabilities that **white** students attend religious schools: 54/524 = 0.103 or 10.3%

What are the odds that **non-white** students attend religious schools? 0.255/(1 – 0.255) = 0.342
What are the odds that **white** students attend religious schools? 0.103/(1 – 0.103) = 0.115

What is the odds ratio that compares white and non-white students? 0.115/0.342 = 0.336
So, **white** students attending a religious school is 0.3359 times the odds of a Non-White student attending a religious school.

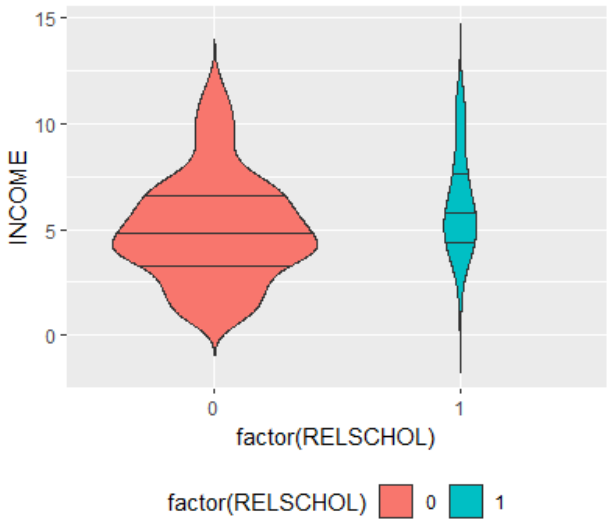
c. Plot RELSCHOL (Y) by INCOME as a scatterplot. The INCOME variable is actually an ordinal variable that is associated with income brackets. This is an old dataset, so for example, INCOME=4 → $20,000-$29,999. Is there a value of INCOME that seems to separate or

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 lower income students and higher income 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?

The scatterplot question is misleading. Per canvas discussions, we will not be using scatterplot. Here is the religion school attendance distribution by income.

```
> table_1 <- table(Relschol = mydataR$RELSCHOL, Income = mydataR$INCOME)
> table_1
        Income
Relschol   1   2   3   4   5   6   7   8   9  10  11  12
       0  38  44  49 101  82  69  58  20  15  18  14   6
       1   1   1   4  13  16  15   6   8   1   5   5   1
> round(prop.table(table_1, 2),2) # column percentages
        Income
Relschol    1    2    3    4    5    6    7    8    9   10   11   12
       0 0.97 0.98 0.92 0.89 0.84 0.82 0.91 0.71 0.94 0.78 0.74 0.86
       1 0.03 0.02 0.08 0.11 0.16 0.18 0.09 0.29 0.06 0.22 0.26 0.14
> round(prop.table(table_1, 1),2) # row percentages
        Income
Relschol    1    2    3    4    5    6    7    8    9   10   11   12
       0 0.07 0.09 0.10 0.20 0.16 0.13 0.11 0.04 0.03 0.04 0.03 0.01
       1 0.01 0.01 0.05 0.17 0.21 0.20 0.08 0.11 0.01 0.07 0.07 0.01
```

And here is the visualization. We can see that there is not a value of INCOME that seems to separate or discriminate between those attending religious schools and those that don't. However, we can see that about 60% of religious attendance comes from income buckets 4, 5, and 6.

We can split the income into two groups: incomes less than 5 grouped as lower-income and incomes larger than or equal to 5 grouped as a higher income. Please note that 36/626 = 5.75% of income is missing or NA. Since the step by step instructions did not specify cleaning the data, I left it as it is.

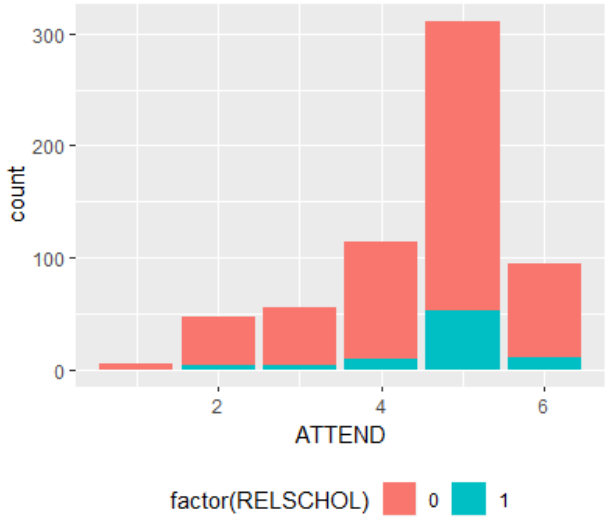
After grouping, about 43% of data falls to lower-income and 57% falls to higher-income buckets.

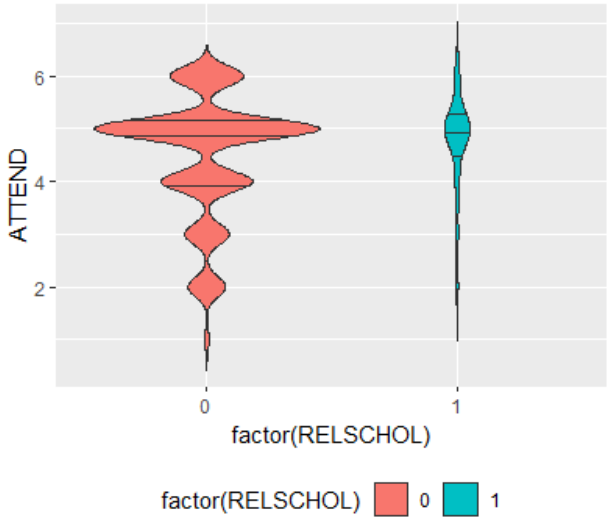
| | RelSchol | | | | | RelSchol | | |
|---|---|---|---|---|---|---|---|---|
| **D_Income** | no = 0 | yes = 1 | **Total** | | **D_Income** | no = 0 | yes = 1 | **Total** |
| low = 0 | 232 | 19 | **251** | | low = 0 | 92.4% | 7.6% | **100.0%** |
| high = 1 | 282 | 57 | **339** | | high = 1 | 83.2% | 16.8% | **100.0%** |
| **Total** | **514** | **76** | **590** | | **Total** | **87.1%** | **12.9%** | **100.0%** |

- What are the probabilities that lower-income students attend religious schools? 19/251 = 0.076 or 7.6%
- What are the probabilities that higher-income students attend religious schools? 57/339 =0.168 or 16.8%
- What are the probabilities that both lower-income and higher-income students attend religious schools? 76/590 = 0.129 or 12.9%
- What are the odds that lower-income students attend religious schools? 0.076 / (1 - 0.076) = 0.082
- What are the odds that higher-income students attend religious schools? 0.168 / (1 - 168) = 0.202
- What is the odds ratio that compares lower and higher-income students? 0.082/0.202 = 0.405

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 scatterplot question is misleading. Per canvas discussions, we will not be using a scatterplot. There are no missing values in attendance service predictor

| Attendance Service | RelSchol | | |
|---|---|---|---|
| | no = 0 | yes = 1 | **Total** |
| 1 | 5 | 0 | **5** |
| 2 | 43 | 4 | **47** |
| 3 | 51 | 4 | **55** |
| 4 | 105 | 9 | **114** |
| 5 | 258 | 53 | **311** |
| 6 | 84 | 10 | **94** |
| **Total** | **546** | **80** | **626** |

| Attendance Service | RelSchol | | |
|---|---|---|---|
| | no = 0 | yes = 1 | **Total** |
| 1 | 100.0% | 0.0% | **100.0%** |
| 2 | 91.5% | 8.5% | **100.0%** |
| 3 | 92.7% | 7.3% | **100.0%** |
| 4 | 92.1% | 7.9% | **100.0%** |
| 5 | 83.0% | 17.0% | **100.0%** |
| 6 | 89.4% | 10.6% | **100.0%** |
| **Total** | **87.2%** | **12.8%** | **100.0%** |

| Attendance Service | RelSchol | | |
|---|---|---|---|
| | no = 0 | yes = 1 | **Total** |
| 1 | 0.9% | 0.0% | **0.8%** |
| 2 | 7.9% | 5.0% | **7.5%** |
| 3 | 9.3% | 5.0% | **8.8%** |
| 4 | 19.2% | 11.3% | **18.2%** |
| 5 | 47.3% | 66.3% | **49.7%** |
| 6 | 15.4% | 12.5% | **15.0%** |
| **Total** | **100.0%** | **100.0%** | **100.0%** |

- Are the proportion profiles the same for those attending religious school versus not, across the values of the ATTEND variable? We can see that the distribution of attending religion schools increases as the attendance service increases. According to the bar chart above, we can also see that most of the students attend religious schools when they have 5 service attendance in a month.

- Is there a value of ATTEND that seems to separate or discriminate between those attending religious schools and those that don't? Attendance service more than 3 times in a month can be grouped as a high attendance service, and anything less than or equal to 3 can be grouped as a low attendance service.

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)

According to the summary logistic regression function below, the coefficient associated with the dummy variable is negative, and the associated p-value is statistically significant. Here is our model:

```
Logit_Y = -1.073 - 1.091* RACE
AIC: 467.4662
BIC: 476.3449
exp (-1.091) - 1 = -0.664
```

The estimated coefficients of the logistic regression model that predicts the probability of religious school attendance using race predictor. This indicates that whites tend to have less likely attended religious schools than non-whites; the odds of attending religious school decreases by 66.42%.

```
Call:
glm(formula = RELSCHOL ~ RACE, family = binomial, data = mydataR)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7671  -0.4664  -0.4664  -0.4664   2.1319

Coefficients:
            Estimate Std. Error z value   Pr(>|z|)
(Intercept)  -1.0726     0.2272  -4.721 0.00000235
RACE         -1.0911     0.2688  -4.059 0.00004930

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 478.48  on 625  degrees of freedom
Residual deviance: 463.47  on 624  degrees of freedom
AIC: 467.47

Number of Fisher Scoring iterations: 4
```

**In task 2b, we calculated the odds ratio of non-white in 2b task, and it was 26/76 = 0.342. The intercept of -1.0726 is the log odds for non-white which is log(26/76) = log(0.342) = -1.0726. We also calculated the odds ratio that compared white and non-white students. The odds ratio was 0.336. If we take the log of 0.336, we will get the race coefficient number -1.0911. When we subtract 1 from 0.336, we will get the -0.664.**

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?

   According to the summary logistic regression function below, the coefficient associated with the income variable is positive, and the associated p-value is statistically significant.

```
Logit_Y = -2.821 + 0.162 * Income
AIC: 445.324
BIC: 454.084
exp (0.162) - 1 = 0.176
```

   Estimated coefficients of the logistic regression model that predicts the probability of religious school attendance using household income. A one-unit increase in household income is associated with an increase in the log-odds of attending religious school by 0.162 units. Or, a one-unit increase in household income is associated with an increase in the odds of attending religious school by 17.6%

   The model 2 has lower AIC and BIC values than model 1. Both models are significant. Income has a positive impact and race has a negative impact.

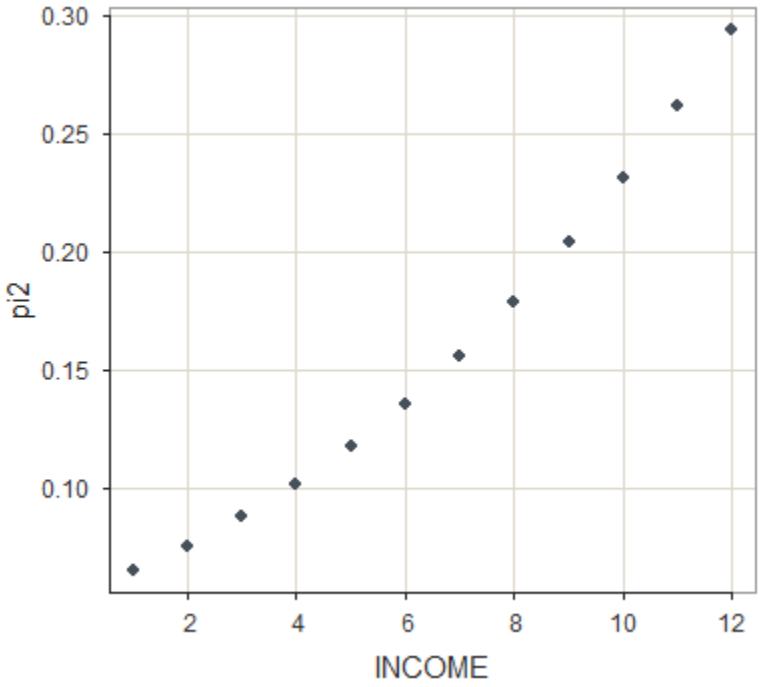
```
Call:
glm(formula = RELSCHOL ~ INCOME, family = binomial, data = mydataR)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8352  -0.5411  -0.4646  -0.3979   2.3352

Coefficients:
            Estimate Std. Error z value             Pr(>|z|)
(Intercept) -2.82110    0.30603  -9.218 < 0.0000000000000002
INCOME       0.16228    0.04669   3.476             0.000509

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 453.27  on 589  degrees of freedom
Residual deviance: 441.32  on 588  degrees of freedom
  (36 observations deleted due to missingness)
AIC: 445.32

Number of Fisher Scoring iterations: 4
```

   **Intercept:** In this case, the estimated coefficient for the intercept is the log odds of a student with a household income of zero being in a religious school. In other words, the odds of being in religious school when the household income is zero is exp(-2.821) = 0.06. These odds are very low, but if we look at the distribution of the variable income, we will see that most of the

students' household income centered around 5. So the intercept in this model corresponds to the log odds of being in a religious school when income is at the hypothetical value of zero.

**Coefficient:** We can say now that the coefficient for income is the difference in the log odds. In other words, for a one-unit increase in income, the expected change in log odds is 0.1622. So we can say for a one-unit increase in income, we expect to see about 17.6% increase in the odds of being in religious school. This 17.6% increase does not depend on the value that income is held at.

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)?




According to the chart, our highest pi value is around 0.3. As our income increases there will be a higher chance of going to religious schools. However, our threshold is 0.5 which is greater than 0.3. Our pi value never exceeds the 0.5 threshold and all twelve income buckets predict 0 values only, attendance to the non-religious schools. Since it is predicting all zeros, it is not a good predictor.

At the problem 2c, we grouped income into two groups: lower and higher incomes. We noticed that people who have higher incomes send their children to religious schools. However, according to pi graph, this problem 2c assumption does not meet 0.5 thresholds.

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?

According to the summary logistic regression function below, the coefficient associated with attend variable is positive, and the associated p-value is not statistically significant at 0.95 confidence interval.

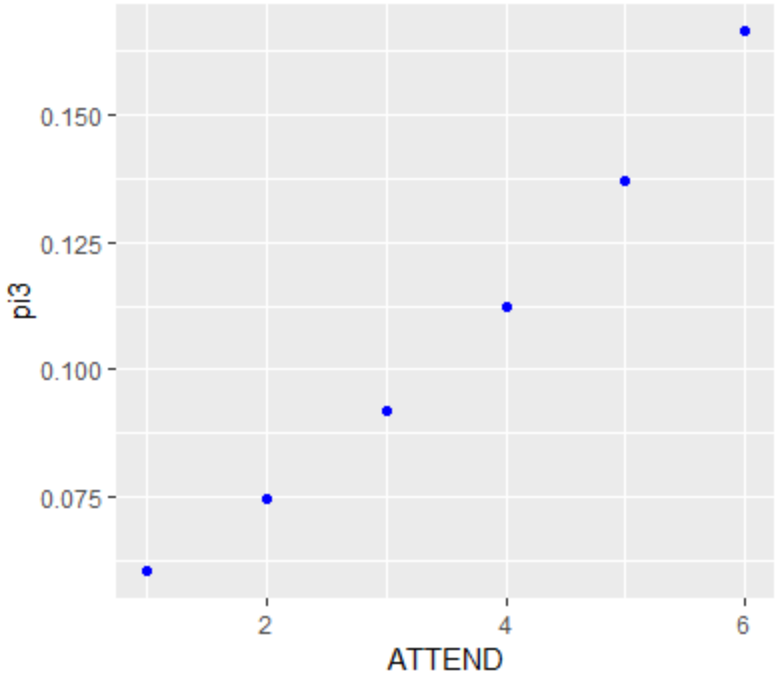
```
Call:
glm(formula = RELSCHOL ~ ATTEND, family = binomial, data = mydataR)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6033  -0.5433  -0.5433  -0.4388   2.2788

Coefficients:
            Estimate Std. Error z value    Pr(>|z|)
(Intercept)  -2.9727     0.5723  -5.194 0.000000206
ATTEND        0.2269     0.1182   1.920      0.0549

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 478.48  on 625  degrees of freedom
Residual deviance: 474.50  on 624  degrees of freedom
AIC: 478.5

Number of Fisher Scoring iterations: 5
```

```
Logit_Y = -2.973 + 0.227 * Attend
AIC: 478.504
BIC: 487.382
exp (0.0.227) - 1 = 0.255
```

Estimated coefficients of the logistic regression model that predicts the probability of religious school attendance using attend predictor. A one-unit increase in attend predictor is associated with an increase in the log-odds of attending religious school by 0.227 units. Or, a one-unit increase in attendance service is associated with an increase in the odds of attending religious school by 25.5%

This model 3 has the highest AIC and BIC values than models 1 and 2, and the attend variable is not a strong predictor of religious school attendance.

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)?




According to the chart, our highest pi value is around 0.17. As our attendance service increases, there will be a higher chance of going to religious schools. However, our threshold is 0.5 which is greater than 0.16. Our pi value never exceeds the 0.5 threshold and all six attendance services predict 0 values only, attendance to the non-religious schools. Since it is predicting all zeros, it is not a good predictor.

At the problem 2d, we grouped attend into two groups: attendance service more than 3 times in a month grouped as a high attendance service and anything less than or equal to 3 grouped as a low attendance service. We noticed that students who had a higher attendance service had a higher chance of going to religious schools. However, according to pi graph, this problem 2d assumption does not meet 0.5 thresholds.

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 Modesl 1, 2 and 3?

According to the summary logistic regression function below, all p-values are statistically significant.

```
glm(formula = RELSCHOL ~ RACE + INCOME + ATTEND, family = binomial,
    data = mydataR)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5371  -0.5507  -0.4177  -0.3204   2.5477

Coefficients:
            Estimate Std. Error z value    Pr(>|z|)
(Intercept) -3.58314    0.71852  -4.987 0.000000614
RACE        -1.28931    0.28981  -4.449 0.000008634
INCOME       0.20068    0.04873   4.118 0.000038171
ATTEND       0.33164    0.12966   2.558      0.0105

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 453.27  on 589  degrees of freedom
Residual deviance: 416.79  on 586  degrees of freedom
  (36 observations deleted due to missingness)
AIC: 424.79

Number of Fisher Scoring iterations: 5
```

```
Logit_Y = -3.583 - 1.289*Race + 0.201*Income + 0.332*Attend
AIC: 424.793
BIC: 442.314

Race: exp(-1.289)-1 = -0.725
Income: exp(0.201)-1 = 0.223
Attend: exp(0.332)-1 = 0.394
```

The table shows the coefficient estimates for a logistic regression model that uses race, income, and attend to predict the probability of attending religious schools. The p-values associated with income, attend, and the dummy variable for the race is very small, indicating that each of these variables is associated with the probability of attending religious school. The negative coefficient for the race in the multiple logistic regression indicates that for a fixed value of income and attend, white is less likely to join religious schools than non-whites.

Whites tend to have less likely attended to religious schools than non-whites; the odds of attending religious school decreases by 72.5%, (at single variable logistic regression it used to be 66.42%).

A one-unit increase in household income is associated with an increase in the log-odds of attending religious school by 0.201 units. Or, a one-unit increase in household income is associated with an increase in the odds of attending religious school by 25.1%, (at single variable logistic regression it used to be 17.6%).

A one-unit increase in attend predictor is associated with an increase in the log-odds of attending religious school by 0.332 units. Or, a one-unit increase in attendance service is

associated with an increase in the odds of attending religious school by 39.4%, (at single variable logistic regression it used to be 25.5%).

This fitted model says that holding income and attendance service at a fixed value, the odds of attending to a religious school for whites (white = 1) over the odds of attending to a religious school for non-whites (non-whites = 0) is exp(-1.28931) = 0.28. In terms of percent change, we can say that the odds for white students are 72.45% (0.28-1) lower than the odds for non-white students. The coefficient for income says that holding white students and attend service at a fixed value, we will see 22.22% (exp(0.20068)-1) increase in the odds of attending to religious schools for a one-unit increase in income.

We have the lowest AIC and BIC values, which means that this model has strong predictors when they used together.

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?

White = 1: -3.583 - 1.289*1 + 0.201*4 + 0.332*5 = -2.408;
exp(-2.408)=0.09

Non-white = 0: -3.583 - 1.289*0 + 0.201*4 + 0.332*5 = -1.119;
exp(-1.119)=0.327

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

Unadjusted Odds Ratio (OR) is a simple ratio of probabilities of outcome in two groups: OR = p1/(1-p1) ÷ p2/(1-p2). Let's assume this question is the continuation of question 6b task since we derived odds for white and non-white; we can derive the adjusted odds as following: 0.09/0.327 = 0.275. Therefore, white students attending a religious school is 0.275 times the odds of a non-white student attending a religious school.

Please note that in this case, there are no predictor variables but only the outcome variable. In logistic regression, we can include other confounding variables to control their influence on our outcome variable, and if we do so, what we can get is, OR that is adjusted for the influence of confounders. In the assignment, if we use the predicted probabilities from the Logistic Regression to calculate the OR, then it is called adjusted OR. We can compute unadjusted OR just by considering the outcome variable. Then we would be able to compare the 2 OR's and that would give us more insights about the problem.

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)

| | RelSchol | | |
|---|---|---|---|
| **Model** | no = 0 | yes = 1 | Total |
| Actual | 546 | 80 | 626 |
| Model 1 (race) | 626 | 0 | 626 |
| Model 2 (income) | 590 | 0 | 590 |
| Model 3 (attend) | 526 | 0 | 526 |
| Model 4 (all) | 526 | 0 | 526 |

The probability value of all models did not exceed 0.5 thresholds.

We have a highly skewed response because 13% (80 out of 626) of students are attending religious school. There is data imbalance: more data collected about non-religious schools than religious schools. Almost all variables had 'small' probability and it can be treated/regressed differently due to the rare event.

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)

There is a higher chance for non-white people with higher income and more religious service attendance to attend religious schools. To diversify our mix, we should start attracting white and/or low-income families. We should also promote more religious services for students because it proved that students who had more attendance service had a higher chance of attending religious schools.

This is my hypothesis. I believe that the first and second generations of non-white immigrants encourage their children to attend religious schools. This involvement will help their children not to forget the religion, culture, and language of their parents or grandparents.