COMPUTATIONAL ASSIGMENT #5
BRANDON MORETZ

COMPUTATION

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	
1	30	34	*64*
0	10	6	*16*
	40	*40*	*80*
Probability Ratio:	0.750	0.850	
Odds Ratio:	3.000	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.*

	Religious School
1	80
0	546
Total	*626*

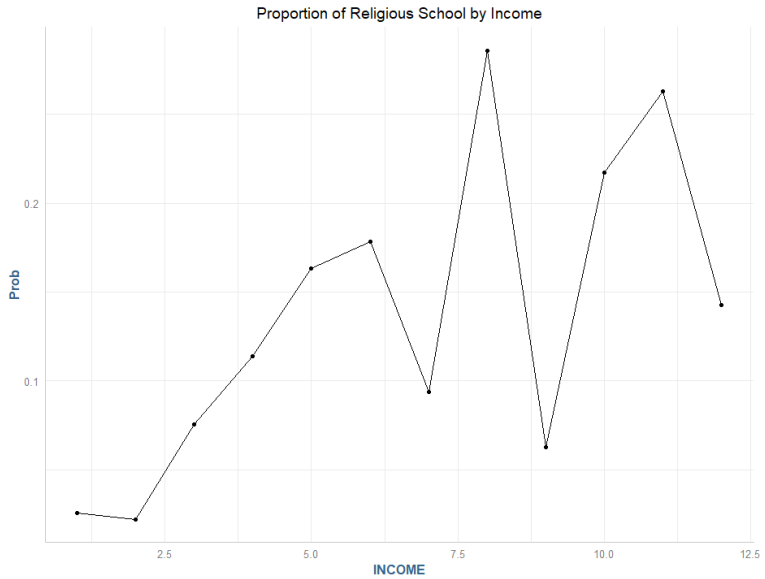
	Yes	No
Probability Ratio:	0.128	0.872
Odds Ratio:	0.147	6.825

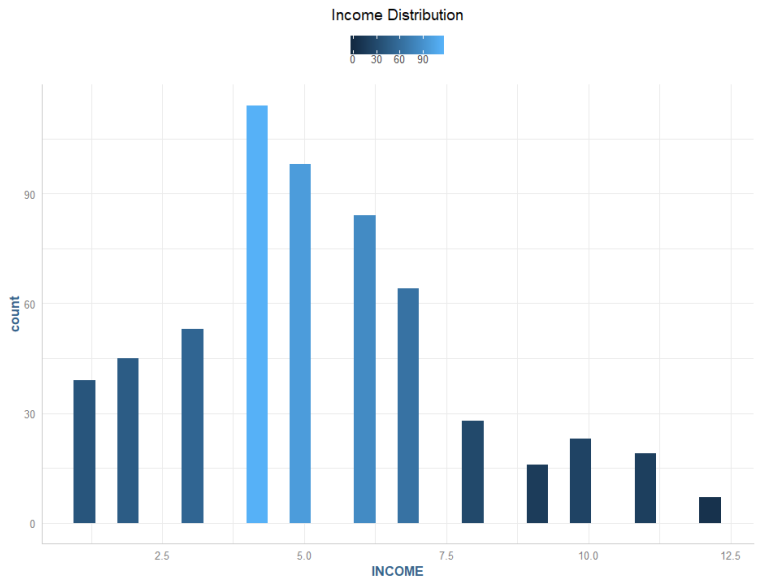
b. *Cross-tabulate RELSCHOL with RACE (coded: 0=non-white, 1=white). What are the probabilities that non-white students and white students attend religious schools? What are the odds that white students and non-white students attend religious schools? What is the odds ratio that compares white and non-white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

Religious School	Race 0	Race 1
0	76	470
1	26	54
	102	524

	Religious School	White? No	White? Yes
Probability Ratio:	No	0.745	0.897
	Yes	0.255	0.103
Odds Ratio:	No	2.923	8.704
	Yes	0.342	0.115

c. Plot *RELSCHOL* (Y) by *INCOME* as a scatterplot.





The INCOME variable is 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?

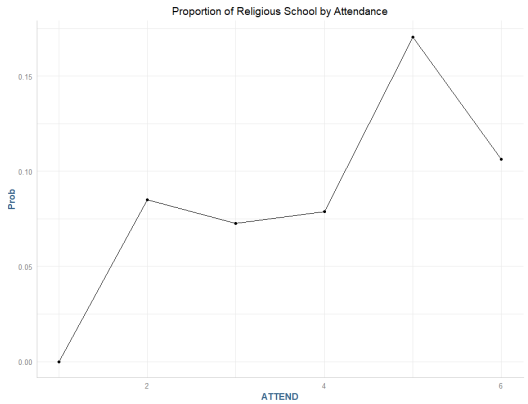
Those students in the income bracket 8 and over have an average change of 19.43% for attending a religious school, compared to a 9.61% for those students with less than an income of the 8th bracket.

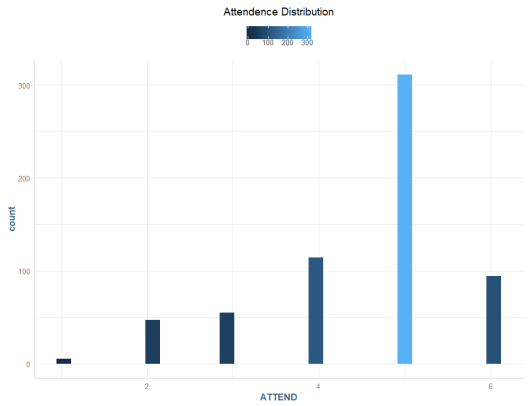
Create a variable that dichotomizes INCOME based on this value you observed. Call this new variable D_INCOME. Cross-tabulate RELSCHOL with D_INCOME. What are the probabilities that low-income students and higher students attend religious schools? What are the odds that lower income students and higher income students attend religious schools? What is the odds ratio that compares lower and higher income students?

	D_INCOME	
Religious School	0	1
0	441	73
1	56	20
	497	93

		High Income?	
	Religious School	No	Yes
Probability Ratio:	No	0.887	0.785
	Yes	0.113	0.215
Odds Ratio:	No	7.875	3.650
	Yes	0.127	0.274

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 profile for attendance varies across the number of days attended, however, we see the largest spike on day 5.

	ATTEND					
Religious School	1	2	3	4	5	6
0	5	43	51	105	258	84
1	0	4	4	9	53	10
	5	*47*	*55*	*114*	*311*	*94*

		Attendance (Days)					
Religious School		1	2	3	4	5	6
Probability Ratio:	No	1.000	0.915	0.927	0.921	0.830	0.894
	Yes	0.000	0.085	0.073	0.079	0.170	0.106
Odds Ratio:	No	1.000	10.750	12.750	11.667	4.868	8.400
	Yes	0.000	0.093	0.078	0.086	0.205	0.119

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.*

(**3 points**)

Model 1

$$\hat{Y} = -1.073 - 1.091\beta_1$$

Where β_1 is a binary variable that represents a white vs non-white student (white = 1), and $\hat{Y}$ is the log odds of a person attending a religious school. The intercept here can be interpreted as:

$$exp(-1.073)/(1 + \exp(-1.073)) = \mathbf{.255}$$

which denotes roughly a **25.5**% chance that a non-white student (**race = 0**) attends a religious school. The X coefficient here is interpreted as for a given student, if they are white (**race = 1**), then the probably of that student attending a religious school further decrease by

$$\exp(-1.09)/(1 + \exp(-1.09)) = \mathbf{.251}$$

an additional **25.1**%. The model information loss statistics are summarized in the table below.

Model	AIC	BIC
Model 1	467.4662	476.3449

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?*

Model 2

$$\hat{Y} = -2.821 + 0.162\beta_1$$

Where β_1 is an ordinal categorical variable that represents the income bracket for the family the student comes from, and $\hat{Y}$ is the log odds of a person attending a religious school. The intercept here can be interpreted as:

$$\exp(-2.821)/(1 + \exp(-2.821)) = \mathbf{0.056}$$

which denotes a roughly **5.6%** chance of a given student attending a religious school ignoring the income bracket of the family. The coefficient in this model denotes a

$$\exp(0.162) = \mathbf{1.255}$$

or an approximately **25.5%** increased chance of a student attending religious school per increase in income bracket. If we were to predict the probabilities of a student attending religious school by income bracket, we would predict a steady increase in religious school enrollment through each income bracket.

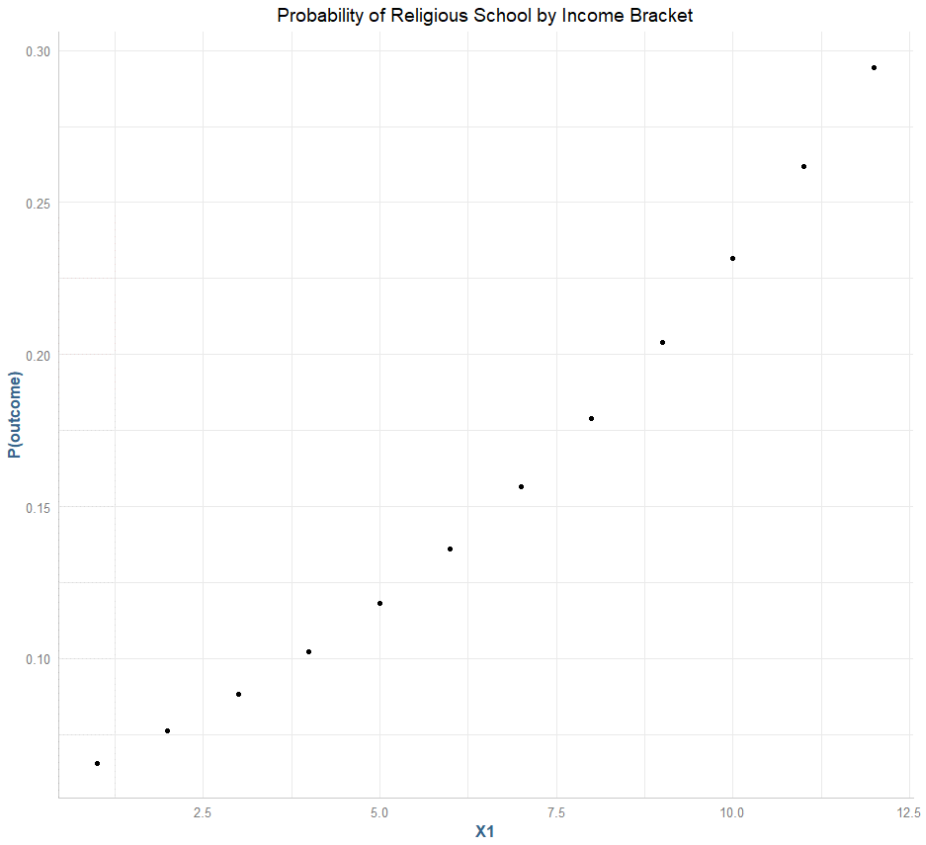
The model fit statistics can be found summarized in the following table:

Model	AIC	BIC
Model 2	445.3235	454.0837

This model is different than the first model in that we see progressively different values for predicted religious school enrollment for various levels of income brackets, where the first model was by a dichotomous variable, race.

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 plot of PI vs Income can be seen in the following section. The value of PI does not exceed 50% at any point along the income brackets, as the maximum value of PI for all income groups is **29.4%** at income bracket **12**.



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?*

Model 3

$$\hat{Y} = -2.972 + 0.227\beta_1$$

Where β_1 is an ordinal variable that represents how many days / weeks a student attends religious service. The intercept here can be interpreted as:

$$\exp(-2.972)/(1+\exp(-2.972)) = \mathbf{0.049}$$

which denotes roughly a **4.9**% chance of attending religious school without attending religious services. The coefficient in this model denotes a

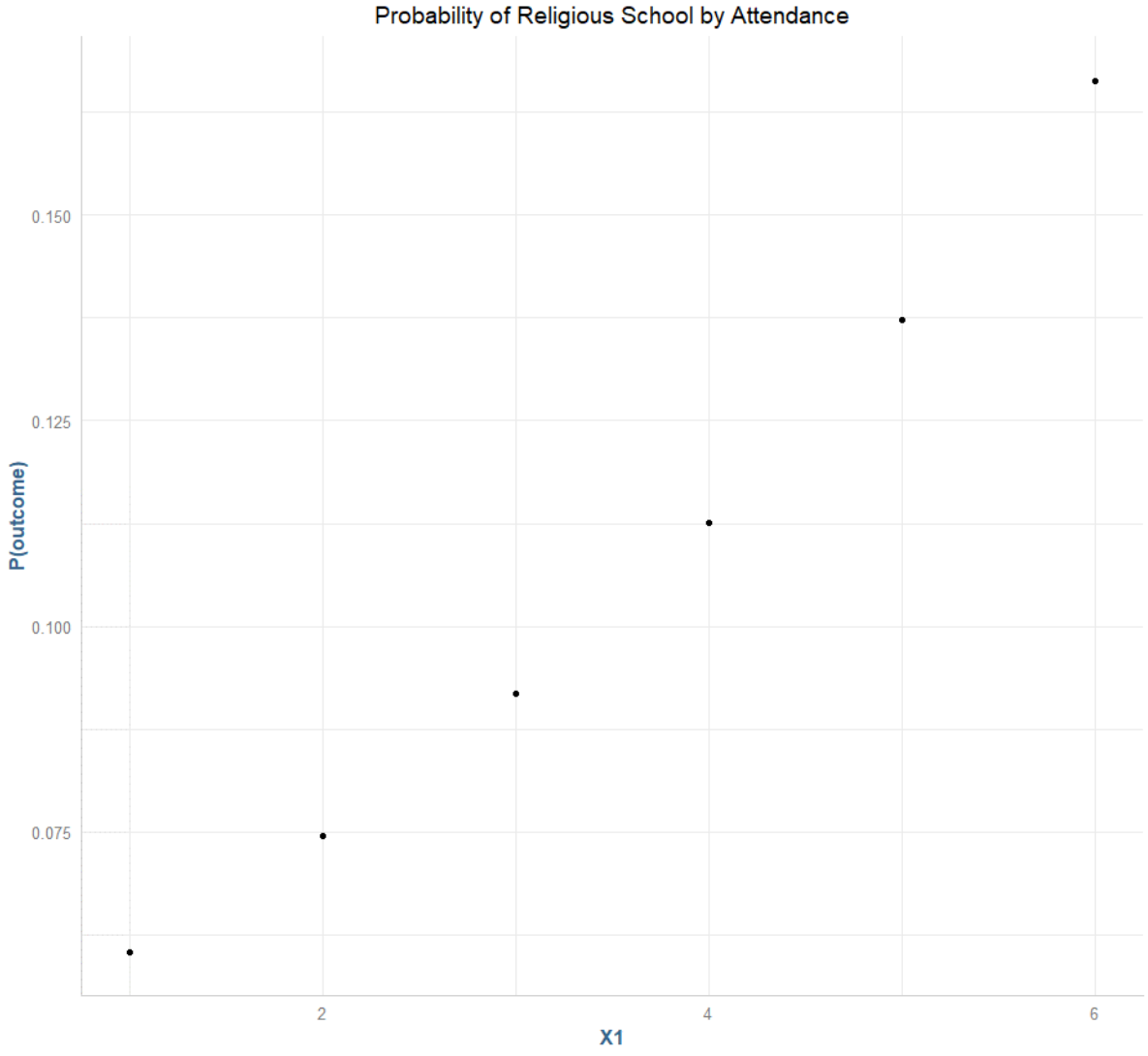
$$\exp(.227) = \mathbf{1.255}$$

or an approximate **25.5%** increased chance of attending religious school per days attending religious services every week. The model fit statistics are:

Model	AIC	BIC
Model 3	478.5036	487.3823

This model is similar to the other two in that it has a negative intercept, meaning that there is a low probability of attending religious services without any coefficients, and as in the case with model 2, the coefficient increases the probability of religious school as the values increase. The AIC and BIC scores are lower in model 2 than in model 3, indicating income is a more informative modeling variable than 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****)?*



The PI value never crosses the **.5** threshold like in the other two models, and there is an overall low probability of attending religious school as we saw in the previous exercises.

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?*

Model 4

$$\hat{Y} = -3.583 - 1.289\beta_1 + 0.2\beta_2 + 0.332\beta_3$$

Where β_1 denotes white/non-white (1 = white), β_2 is the income bracket of the family and β_3 is the number of religious services attended on a weekly basis. The intercept term here can be interpreted as:

$$\exp(-3.583)/(1+\exp(-3.583)) = \mathbf{0.027}$$

Which denotes a **2.7**% chance of attending religious school ignoring all other terms in the model. We see that race has a negative association with religious school attendance, meaning that white students (race =1) have:

$$\exp(-1.289)/(1+\exp(-1.289)) = \mathbf{0.216}$$

or a **21.6**% lower chance of attending religious school compared to non-whites (**race** = 0). For income, we see a positive association meaning that overall students with higher income brackets tend to have a higher probability of attending religious school, converting the log odds to probability:

$$\exp(.2) = \mathbf{1.222}$$

meaning that for every 1 unit increase in income we see a 22% increase in the chance of a student attending religious school. Factoring religious service attendance into the equation, we see that the β_3 coefficient is:

$$\exp(.332) = \mathbf{1.393}$$

which we can interpret as that for every additional day a person attends religious service during the week, there is an associated **39.3**% increased chance that they attended religious school.

Model	AIC	BIC
Model 4	424.793	442.3135

Overall, this model compares intuitively to what we would expect the outcomes and associations to be. The model denotes that overall, there is a relatively low chance of someone attending a religious school, with non-whites (race=0) having an overall higher chance of attending religious school. Income and religious attendance are both positively corelated to religious school attendance, and people with higher income and religious service attendance showing a higher probability of attending religious school.

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?*

For non-white students we can plug the following values into the model:

Non-white student:

Log odds = -3.586 – 1.289*0 + .2*4 + .332*5 = **-1.122**

Odds = exp(**-1.122**) = **.326**

Probability = .326 / (1 + .326) = **.245**

White student:

Log odds = -3.586 – 1.289*1 + .2*4 + .332*5 = **-2.412**

Odds = exp(**-1.122**) = **.09**

Probability = .326 / (1 + .326) = **.082**

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

For race, we see that there is a has a negative association with religious school attendance, meaning that white students (**race** =1) have an overall lower probability of attending religious school. The change in odds for white students can be found:

exp(-1.289) = **.275**

Which can also be interpreted as the following probability:

exp(-1.289)/(1+exp(-1.289)) = **0.216**

or a **21.6**% lower chance of attending religious school compared to non-whites (**race** = 0).

RESEARCH

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**)

Model	Yes	No
model1_pred	0	590
model2_pred	0	590
model3_pred	0	590

These prediction rates make sense given the low overall probability of attending a religious school and we have explored above. None of the three models generate a predicted probability of greater than .5, with most of them capping out at around 30%. There are a few outliers in the fourth model that have a > .5 chance, however, they are on extreme values in the data set.

CONCLUSION

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**)

In this lab we have explored the relationship between religious school attendance and sex, income bracket and religious service attendance. From our modeling and initial data exploration, we found that there is a low likelihood of attending religious school regardless of the independent factors under consideration, at least in this data set. We found that whites (race = 1) have an overall lower percentage chance of attending religious school compared to non-whites (race=0). Additionally, we found positive associative relationships between income and religious service attendance and religious school. Simply, the more income the family has and the higher the number of days they attended religious services throughout the week, the higher their overall chance of attending religious school is.