MSPA PREDICT 411

Bonus Problem: Chapter 3

In [1]:
```
#!pip install sas7bdat

import numpy as np
import pandas as pd
import statsmodels.api as sm
import scipy

from patsy import dmatrices
from sas7bdat import SAS7BDAT

import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.pylab as pylab

sns.set_style('darkgrid')
%matplotlib inline
```

Introduction

This document presents the results of second set of bonus problems for the Masters of Science in Predictive Analytics course: PREDICT 411. This assessment required the student to w ork through the problem set of Chapter 3 of Hoffmann (2004), Generalized Linear Models, An Applied Approach.

Question 1

For the follow ing 2x2 table, determine the odds and the probabilities of marijuana use among males and females. Then compute the odds ratio of marijuana use that compare males to females.

Marijuana Use	Male	Female
Yes	10	6
No	30	34

The *probability* of an event occuring can be defined as:

$$\text{probability} = \frac{\text{events}}{\text{number of outcomes}}$$

In [2]:
```
count_yes_male = 10
count_no_male = 30
count_yes_female = 6
count_no_female = 34

p_yes_male = count_yes_male / sum([count_yes_male, count_no_male])
print('p_yes_male:', p_yes_male)

p_yes_female = count_yes_female / sum([count_yes_female, count_no_female])
print('p_yes_female:', p_yes_female)
```

```
p_yes_male: 0.25
p_yes_female: 0.15
```

While *odds* can be defined as:

$$\text{odds} = \frac{p}{1-p}$$

In [3]:
```
odd_yes_male = p_yes_male / (1- p_yes_male)
print('odd_yes_male:', odd_yes_male)

odd_yes_female = p_yes_female / (1- p_yes_female)
print('odd_yes_female:', odd_yes_female)
```

```
odd_yes_male: 0.3333333333333333
odd_yes_female: 0.17647058823529413
```

The *odds ratio* forumla:

$$OR_{\text{a vs b}} = \frac{\text{odds}_{\text{a}}}{\text{odds}_{\text{b}}}$$

In [4]:
```
odd_yes_malevsfem = odd_yes_male / odd_yes_female
print('odd_yes_malevsfem:', odd_yes_malevsfem)
```

```
odd_yes_malevsfem: 1.888888888888886
```

Question 2

Loading the Data

In [5]:
```
with SAS7BDAT('data/religion.sas7bdat') as f:
    df_rel = f.to_data_frame()
```

In [6]:
```
df_rel.head(5)
```

Out[6]:

	ID	SEX	AGE	EDUC	INCOME	RELSCHOL	MARRIED	ATTEND	AGESQ	RACE
0	2	1	30	6	11	0	1	6	900	1
1	3	1	32	6	6	0	1	5	1024	1
2	4	1	51	2	11	0	1	2	2601	1
3	5	1	18	2	3	0	0	6	324	1
4	6	1	37	5	6	0	1	6	1369	1

Part A

Compute the overall odds and probability of attending a religious school.

In [7]:
```
count_no_relschool = df_rel['RELSCHOL'].value_counts()[0]
count_yes_relschool = df_rel['RELSCHOL'].value_counts()[1]

p_yes_relschool = count_yes_relschool / sum([count_yes_relschool, count_no_relschool])
print('p_yes_relschool:', p_yes_relschool)

odd_yes_relschool = p_yes_relschool / (1- p_yes_relschool)
print('odd_yes_relschool:', odd_yes_relschool)
```

```
p_yes_relschool: 0.127795527157
odd_yes_relschool: 0.14652014652
```

Part B

Cross-tabulate *relschol* w ith *race* (coded 0 as non-w rite, 1 as w hite). What are the probabilities that non-w hite students and w hite students attend religious schools? What are the odds that w hite students and non-w hite students attend religious schools? What is the odds ratio that compares w hite and non-w hite students?

In [8]:
```
pd.crosstab(df_rel['RELSCHOL'], df_rel['RACE'])
```

Out[8]:

RACE	0.0	1.0
RELSCHOL		
0	76	470
1	26	54

The *probability* of an event occuring can be defined as:

$$\text{probability} = \frac{\text{events}}{\text{number of outcomes}}$$

In [9]:
```
count_white_relschool = 54
count_black_relschool = 26
count_white_norelschool = 470
count_black_norelschool = 76

p_white_relschool = count_white_relschool / sum([count_white_relschool, count_white_norelschool]
)
print('p_white_relschool:', p_white_relschool)

p_black_relschool = count_black_relschool / sum([count_black_relschool, count_black_norelschool]
)
print('p_black_relschool:', p_black_relschool)
```

```
p_white_relschool: 0.10305343511450382
p_black_relschool: 0.2549019607843137
```

While *odds* can be defined as:

$$\text{odds} = \frac{p}{1-p}$$

In [10]:
```
odd_white_relschool = p_white_relschool / (1- p_white_relschool)
print('odd_white_relschool:', odd_white_relschool)

odd_black_relschool = p_black_relschool / (1- p_black_relschool)
print('odd_black_relschool:', odd_black_relschool)
```

```
odd_white_relschool: 0.11489361702127661
odd_black_relschool: 0.3421052631578947
```

The *odds ratio* forumla:

$$OR_{\text{a vs b}} = \frac{\text{odds}_{\text{a}}}{\text{odds}_{\text{b}}}$$

In [11]:
```
odd_relschool_whitevsblack = odd_white_relschool / odd_black_relschool
print('odd_relschool_whitevsblack:', odd_relschool_whitevsblack)
```

```
odd_relschool_whitevsblack: 0.3358428805237317
```

Question 3

Estimate tw o logistic regression models that are designed to preict *relschol*. In the first model include only the variable *race*. In the second model, include *Race*, *attend* (religious service attendence), and *income* (family income), treating the latter tw o as continuous variables.

Part A

Based on the first model, w hat is the odds ratio that compares w hite and non-w hite students? Compare this to the odds ratio computed in Excercise 2.B.

In [12]:
```
y, X = dmatrices('RELSCHOL ~ C(RACE)', data=df_rel, return_type='dataframe')
model = sm.Logit(y, X)
results = model.fit()
```

```
Optimization terminated successfully.
         Current function value: 0.370181
         Iterations 6
```

In [13]:
```
results.summary()
```

Out[13]:

Logit Regression Results

Dep. Variable:	RELSCHOL	**No. Observations:**	626
Model:	Logit	**Df Residuals:**	624
Method:	MLE	**Df Model:**	1
Date:	Tue, 12 Jul 2016	**Pseudo R-squ.:**	0.03138
Time:	17:11:22	**Log-Likelihood:**	-231.73
converged:	True	**LL-Null:**	-239.24
		LLR p-value:	0.0001066

	coef	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**
Intercept	-1.0726	0.227	-4.721	0.000	-1.518 -0.627
C(RACE)[T.1.0]	-1.0911	0.269	-4.059	0.000	-1.618 -0.564

In [14]:
```
results.summary2()
```

Out[14]:

Model:	Logit	Pseudo R-squared:	0.031
Dependent Variable:	RELSCHOL	AIC:	467.4662
Date:	2016-07-12 17:11	BIC:	476.3449
No. Observations:	626	Log-Likelihood:	-231.73
Df Model:	1	LL-Null:	-239.24
Df Residuals:	624	LLR p-value:	0.00010659
Converged:	1.0000	Scale:	1.0000
No. Iterations:	6.0000		

	Coef.	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-1.0726	0.2272	-4.7211	0.0000	-1.5179	-0.6273
C(RACE)[T.1.0]	-1.0911	0.2688	-4.0589	0.0000	-1.6180	-0.5642

In [15]:
```
y, X = dmatrices('RELSCHOL ~ C(RACE) + ATTEND + INCOME', data=df_rel, return_type='dataframe')
model = sm.Logit(y, X)
results = model.fit()
```

```
Optimization terminated successfully.
         Current function value: 0.353214
         Iterations 7
```

In [16]:
```
results.summary()
```

Out[16]:

Logit Regression Results

Dep. Variable:	RELSCHOL	**No. Observations:**	590
Model:	Logit	**Df Residuals:**	586
Method:	MLE	**Df Model:**	3
Date:	Tue, 12 Jul 2016	**Pseudo R-squ.:**	0.08047
Time:	17:11:22	**Log-Likelihood:**	-208.40
converged:	True	**LL-Null:**	-226.63
		LLR p-value:	5.943e-08

	coef	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**
Intercept	-3.5831	0.719	-4.987	0.000	-4.991 -2.175
C(RACE)[T.1.0]	-1.2893	0.290	-4.449	0.000	-1.857 -0.721
ATTEND	0.3316	0.130	2.558	0.011	0.078 0.586
INCOME	0.2007	0.049	4.118	0.000	0.105 0.296

In [17]:
```
results.summary2()
```

Out[17]:

Model:	Logit	Pseudo R-squared:	0.080
Dependent Variable:	RELSCHOL	AIC:	424.7930
Date:	2016-07-12 17:11	BIC:	442.3135
No. Observations:	590	Log-Likelihood:	-208.40
Df Model:	3	LL-Null:	-226.63
Df Residuals:	586	LLR p-value:	5.9434e-08
Converged:	1.0000	Scale:	1.0000
No. Iterations:	7.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-3.5831	0.7185	-4.9868	0.0000	-4.9914	-2.1749
C(RACE)[T.1.0]	-1.2893	0.2898	-4.4488	0.0000	-1.8573	-0.7213
ATTEND	0.3316	0.1297	2.5577	0.0105	0.0775	0.5858
INCOME	0.2007	0.0487	4.1183	0.0000	0.1052	0.2962

Part B

What are the AIC and BIC for the tw o models? Based on these measures of fit, w hich model do you prefer?

	AIC	BIC
Model_1	467.4662	476.3449
Model_2	424.7930	442.3135

The usual rule of thumb is to compare the AIC and BIC values and choose the model in w hich the values are low er. In this case, w e w ould prefer Model_2.

Part C

For tose w ho attend religious services five days per month (*attend* = 5) and have a family income of $20,000 - 29,999$ predicted odds of attending a religious school for w hite and non-w hite students?

In [18]:
```
odds_white  = np.exp(-3.5831 + 1.0*-1.2893 + 5.0*0.3316 + 4.0*0.2007)
odds_non_white = np.exp(-3.5831 + 0.0*-1.2893 + 5.0*0.3316 + 4.0*0.2007)
print('odds_white:', odds_white, 'odds_non_white:', odds_non_white)
```

```
odds_white: 0.0896717050032 odds_non_white: 0.325530213444
```

Part D

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

We w ill calculate the adjusted odds ratio for gender by plugging the mean values in and predicting the odds for w hite and non-w hite.

In [19]:
```
df_rel.mean()
```

Out[19]:
```
ID           503.487220
SEX            0.584665
AGE           47.077047
EDUC           3.323718
INCOME         5.215254
RELSCHOL       0.127796
MARRIED        0.635783
ATTEND         4.535144
AGESQ       2518.495987
RACE           0.837061
dtype: float64
```

In [20]:
```
m_odds_white = np.exp(-3.5831 + 1.0*-1.2893 + 4.535144*0.3316 + 5.215254*0.2007)
m_odds_non_white = np.exp(-3.5831 + 0.0*-1.2893 + 4.535144*0.3316 + 5.215254*0.2007)
print('m_odds_white:', m_odds_white / m_odds_non_white , np.exp(-1.2893))
```

```
m_odds_white: 0.275463540095 0.275463540095
```

Question 4

Re-estimate the tw o models outlined in Excercise 3, but use a probit model.

In [21]:
```
y, X = dmatrices('RELSCHOL ~ C(RACE)', data=df_rel, return_type='dataframe')
model = sm.Probit(y, X)
results = model.fit()
results.summary2()
```

```
Optimization terminated successfully.
         Current function value: 0.370181
         Iterations 6
```

Out[21]:

Model:	Probit	Pseudo R-squared:	0.031
Dependent Variable:	RELSCHOL	AIC:	467.4662
Date:	2016-07-12 17:11	BIC:	476.3449
No. Observations:	626	Log-Likelihood:	-231.73
Df Model:	1	LL-Null:	-239.24
Df Residuals:	624	LLR p-value:	0.00010659
Converged:	1.0000	Scale:	1.0000
No. Iterations:	6.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-0.6591	0.1344	-4.9040	0.0000	-0.9226	-0.3957
C(RACE)[T.1.0]	-0.6052	0.1535	-3.9439	0.0001	-0.9060	-0.3044

Part A

Based on the first model, w hat is the predicted probability that w hite and non-w hite students attend a religious school? Compare these results to those found in Exercise 2.B.

In [22]:
```
p_white = scipy.stats.norm.cdf(-0.6591-0.6051)
print('p_white:', p_white)

p_non_white = scipy.stats.norm.cdf(-0.6591)
print('p_non_white:', p_non_white)
```

```
p_white: 0.103079125424
p_non_white: 0.25491577784
```

Part B

What are the AIC and BIC for the tw o models? Compare these to the AIC and BIC computed in Exercise 3.B.

In [23]:
```
results.summary2()
```

Out[23]:

Model:	Probit	Pseudo R-squared:	0.031
Dependent Variable:	RELSCHOL	AIC:	467.4662
Date:	2016-07-12 17:11	BIC:	476.3449
No. Observations:	626	Log-Likelihood:	-231.73
Df Model:	1	LL-Null:	-239.24
Df Residuals:	624	LLR p-value:	0.00010659
Converged:	1.0000	Scale:	1.0000
No. Iterations:	6.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-0.6591	0.1344	-4.9040	0.0000	-0.9226	-0.3957
C(RACE)[T.1.0]	-0.6052	0.1535	-3.9439	0.0001	-0.9060	-0.3044

In [24]:
```
y, X = dmatrices('RELSCHOL ~ C(RACE) + ATTEND + INCOME', data=df_rel, return_type='dataframe')
model = sm.Probit(y, X)
results = model.fit()
results.summary2()
```

```
Optimization terminated successfully.
         Current function value: 0.351750
         Iterations 6
```

Out[24]:

Model:	Probit	Pseudo R-squared:	0.084
Dependent Variable:	RELSCHOL	AIC:	423.0652
Date:	2016-07-12 17:11	BIC:	440.5857
No. Observations:	590	Log-Likelihood:	-207.53
Df Model:	3	LL-Null:	-226.63
Df Residuals:	586	LLR p-value:	2.5609e-08
Converged:	1.0000	Scale:	1.0000
No. Iterations:	6.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-2.0718	0.3849	-5.3824	0.0000	-2.8263	-1.3174
C(RACE)[T.1.0]	-0.7261	0.1631	-4.4527	0.0000	-1.0457	-0.4065
ATTEND	0.1858	0.0689	2.6965	0.0070	0.0508	0.3209
INCOME	0.1156	0.0271	4.2612	0.0000	0.0624	0.1687

	AIC	BIC
Model_1	467.4662	476.3449
Model_2	424.7930	442.3135
Model_3	467.4662	476.3449
Model_4	423.0652	440.5857

The AIC and BIC between Logit and Probit models is very similar. However, Model_4 has slightly lower AIC and BIC.

Part C

For those who attend religious services five days per month (*attend* = 5) and have a family income of $20,000 - 29,999$ predicted probabilities of attending a religious school for white and non-white students?

In [25]:
```
p_white  = scipy.stats.norm.cdf(-2.0718 + 1.0*-0.7261 + 5.0*0.1858 + 4.0*0.1156)
print('p_white:', p_white)

p_non_white = scipy.stats.norm.cdf(-2.0718 + 0.0*-0.7261 + 5.0*0.1858 + 4.0*0.1156)
print('p_non_white:', p_non_white)
```

```
p_white: 0.0797878523018
p_non_white: 0.248125610515
```

Part D

Compute the discrete change in probability under the following scenario: A non-white student whose *attend* value equals 4 with a shift in family income (*income*) rom a value of 4$(20,000 - 29,999$

In [26]:
```
p_4 = scipy.stats.norm.cdf(-2.0718 + 0.0*-0.7261 + 5.0*0.1858 + 4.0*0.1156)
p_10 = scipy.stats.norm.cdf(-2.0718 + 0.0*-0.7261 + 5.0*0.1858 + 10.0*0.1156)

print(p_4 - p_10)
```

```
-0.257140274665
```

Question 5

In plain English, what do you conclude about the relationship between a student's race/ehnicity, religious service attendance, family income, and attending a religious school?

Each have an impact on the probability that a student w ill attend a religious school.