Module 4
Lesson 1 - Categorical Explanatory Variables with More Than Two Categories

WESLEYAN UNIVERSITY

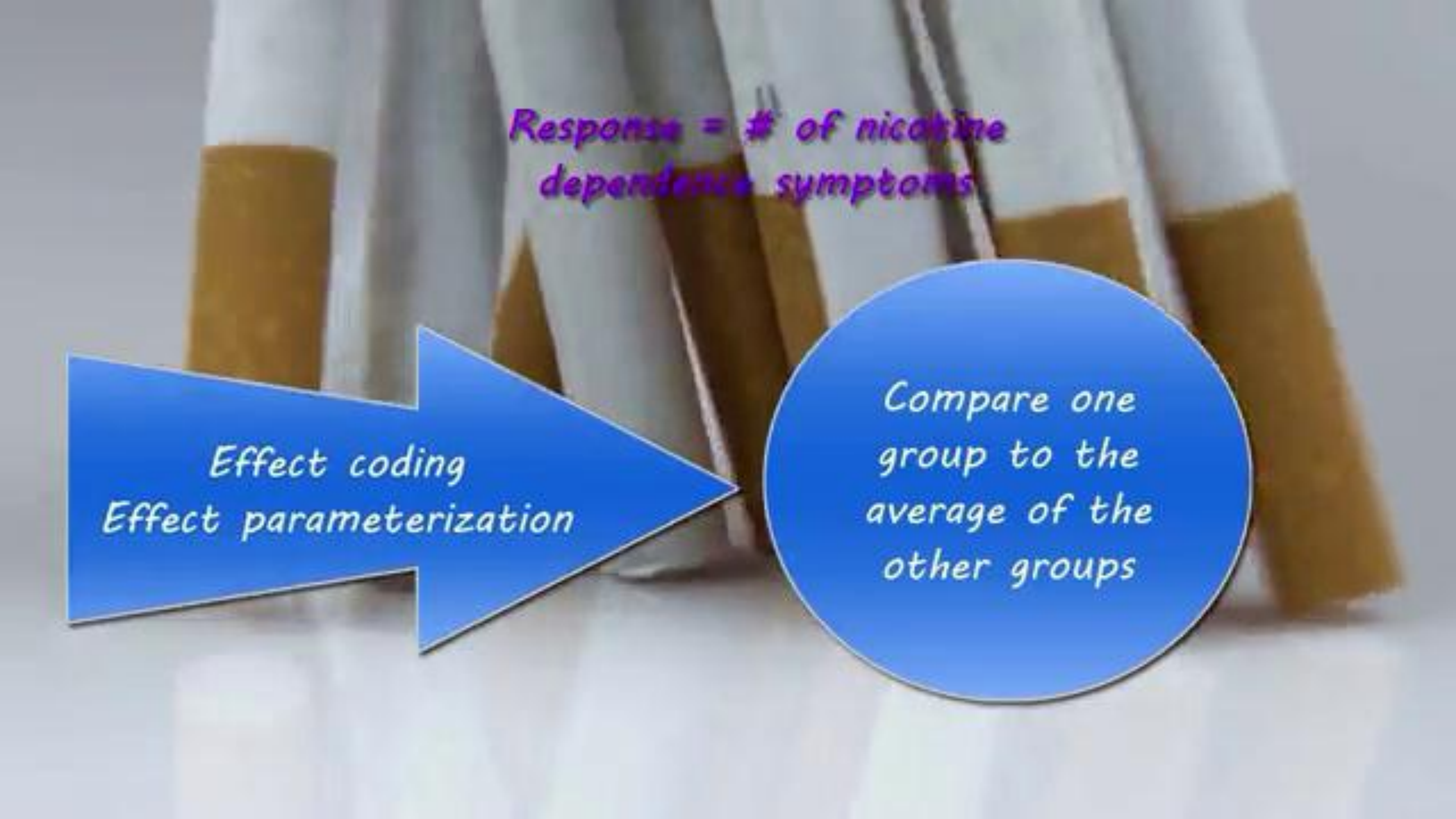
binary (2 categories)

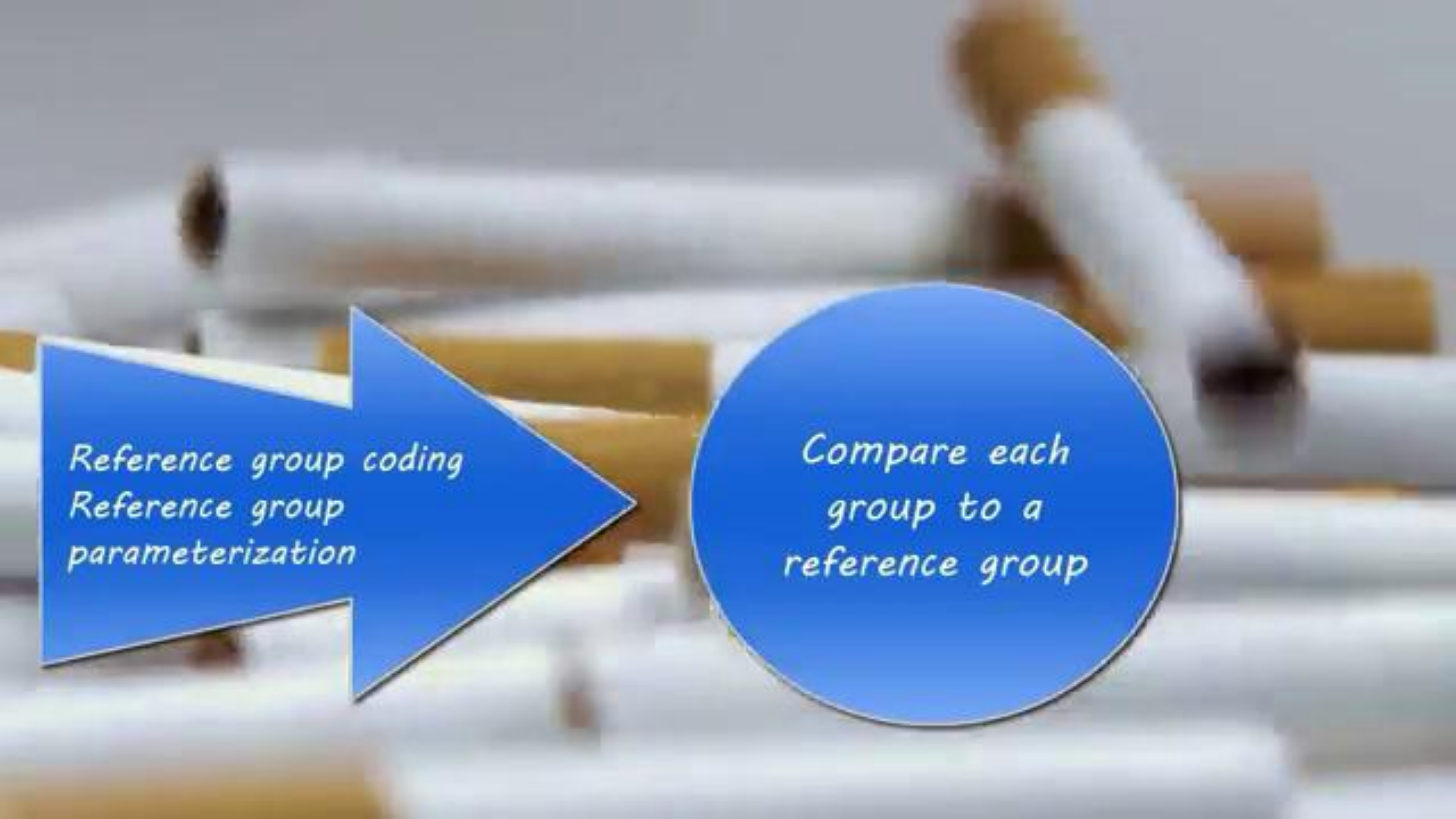
quantitative

categorical with 3+
categories??

dummy coding

parameterization


Response = # of nicotine dependence symptoms
Effect coding
Effect parameterization
Compare one group to the average of the other groups


Reference group coding
Reference group parameterization
Compare each group to a reference group


Response = # of nicotine dependence symptoms
Compare each group to a reference group

```
sub4 = sub1[['NDSymptoms','numbercigsmoked','DYSLIFE',
'MAJORDEPLIFE','AGE','SEX']].dropna()

# dysphoria & depression
reg4 = smf.ols('NDSymptoms ~ DYSLIFE + MAJORDEPLIFE', data=sub1).fit()
print (reg4.summary())

# dysphoria & depression + other covariates
sub1['age_c']=(sub1['AGE'] - sub1['AGE'].mean())
print (sub1['age_c'].mean())

reg5 = smf.ols('NDSymptoms ~ DYSLIFE + MAJORDEPLIFE + numbercigsmoked_c + age_c + SEX', data=sub1).fit()
print (reg5.summary())



# adding 4 category ethnicity/race. Default reference group is the first (Hispanic)
reg6 = smf.ols('NDSymptoms ~ DYSLIFE + MAJORDEPLIFE + numbercigsmoked_c + age_c + SEX + C(ETHRACE)',
               data=sub1).fit()
print (reg6.summary())
```

python default parameterization: Reference (Treatment) group coding

Object inspector | Variable explorer | File explorer

IPython console

Console 1/A

```
    ...: print (reg6.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:             NDSymptoms   R-squared:                       0.139
Model:                            OLS   Adj. R-squared:                  0.134
Method:                 Least Squares   F-statistic:                     26.30
Date:                Sat, 14 Nov 2015   Prob (F-statistic):           5.76e-38
Time:                        10:08:36   Log-Likelihood:                -2588.9
No. Observations:                1313   AIC:                             5196.
Df Residuals:                    1304   BIC:                             5242.
Df Model:                           8
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------------
Intercept             2.2970      0.184     12.454      0.000         1.935     2.659
C(ETHRACE)[T.1]      -0.1116      0.134     -0.830      0.407        -0.375     0.152
C(ETHRACE)[T.2]      -0.0851      0.178     -0.478      0.633        -0.435     0.264
C(ETHRACE)[T.3]       0.3260      0.231      1.412      0.158        -0.127     0.779
DYSLIFE               0.2756      0.209      1.322      0.186        -0.133     0.685
MAJORDEPLIFE          1.2881      0.116     11.078      0.000         1.060     1.516
numbercigsmoked_c     0.0371      0.006      6.355      0.000         0.026     0.049
age_c                -0.0406      0.022     -1.837      0.066        -0.084     0.003
SEX                  -0.0279      0.099     -0.281      0.779        -0.223     0.167
==============================================================================
Omnibus:                       69.969   Durbin-Watson:                   2.071
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               47.818
Skew:                           0.353   Prob(JB):                     4.13e-11
Kurtosis:                       2.387   Cond. No.                         50.4
==============================================================================
```

1 = non-Hispanic White
2 = non-Hispanic Black
3= non-Hispanic Other

```
sub4 = sub1[['NDSymptoms','numbercigsmoked','DYSLIFE',
'MAJORDEPLIFE','AGE','SEX']].dropna()

# dysphoria & depression
reg4 = smf.ols('NDSymptoms ~ DYSLIFE + MAJORDEPLIFE', data=sub1).fit()
print (reg4.summary())

# dysphoria & depression + other covariates
sub1['age_c']=(sub1['AGE'] - sub1['AGE'].mean())
print (sub1['age_c'].mean())

reg5 = smf.ols('NDSymptoms ~ DYSLIFE + MAJORDEPLIFE + numbercigsmoked_c + age_c + SEX', data=sub1).fit()
print (reg5.summary())



# adding 4 category ethnicity/race. Reference group coding is called "Treatment" coding in python
# and the default reference catergory is the group with a value = 0 (Hispanic)
reg6 = smf.ols('NDSymptoms ~ DYSLIFE + MAJORDEPLIFE + numbercigsmoked_c + age_c + SEX + C(ETHRACE)',
               data=sub1).fit()
print (reg6.summary())

# can override the default ad specify a different reference group
# non-Hispanic White as reference group
reg7 = smf.ols('NDSymptoms ~ DYSLIFE + MAJORDEPLIFE + numbercigsmoked_c + age_c + SEX + C(ETHRACE, Treatment(reference=1))',
               data=sub1).fit()
print (reg7.summary())
```

```
alled "Treatment" coding in python
 = 0 (Hispanic)
smoked_c + age_c + SEX + C(ETHRACE)',


up

smoked_c + age_c + SEX + C(ETHRACE, Treatment(reference=1))',
```

Object inspector | Variable explorer | File explorer

IPython console

Console 1/A

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             NDSymptoms   R-squared:                       0.139
Model:                            OLS   Adj. R-squared:                  0.134
Method:                 Least Squares   F-statistic:                     26.30
Date:                Sat, 14 Nov 2015   Prob (F-statistic):           5.76e-38
Time:                        10:36:41   Log-Likelihood:                -2588.9
No. Observations:                1313   AIC:                             5196.
Df Residuals:                    1304   BIC:                             5242.
Df Model:                           8
Covariance Type:            nonrobust
=============================================================================================================
                                             coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------------------------------------
Intercept                                  2.1855      0.163     13.385      0.000         1.865     2.506
C(ETHRACE, Treatment(reference=1))[T.0]    0.1116      0.134      0.830      0.407        -0.152     0.375
C(ETHRACE, Treatment(reference=1))[T.2]    0.0265      0.150      0.177      0.860        -0.267     0.320
C(ETHRACE, Treatment(reference=1))[T.3]    0.4376      0.209      2.091      0.037         0.027     0.848
DYSLIFE                                    0.2756      0.209      1.322      0.186        -0.133     0.685
MAJORDEPLIFE                               1.2881      0.116     11.078      0.000         1.060     1.516
numbercigsmoked_c                          0.0371      0.006      6.355      0.000         0.026     0.049
age_c                                     -0.0406      0.022     -1.837      0.066        -0.084     0.003
SEX                                       -0.0279      0.099     -0.281      0.779        -0.223     0.167
==============================================================================
Omnibus:                       69.969   Durbin-Watson:                   2.071
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               47.818
Skew:                           0.353   Prob(JB):                     4.13e-11
Kurtosis:                       2.387   Cond. No.                         40.0
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Permissions: RW   End-of-lines: CRLF   Encoding: UTF-8   Line: 322   Column: 1   Memory: 48 %

Module 3

Lesson 6 - A Few Things to Keep in Mind

WESLEYAN UNIVERSITY

Bad data
will produce
meaningless
results

Module 4

Lesson 3 - Logistic Regression for a Binary Response Variable

WESLEYAN UNIVERSITY


MULTIPLE REGRESSION
for Quantitative
Variables

```
"""

import numpy
import pandas
import statsmodels.api as sm
import statsmodels.formula.api as smf

data = pandas.read_csv('nesarc_pds.csv',  low_memory=False)

#setting variables you will be working with to numeric

data['IDNUM'] = pandas.to_numeric(data['IDNUM'], errors='coerce')
data['TAB12MDX'] = pandas.to_numeric(data['TAB12MDX'], errors='coerce')
data['SOCPDLIFE'] = pandas.to_numeric(data['SOCPDLIFE'], errors='coerce')
data['MAJORDEPLIFE'] = pandas.to_numeric(data['MAJORDEPLIFE'], errors='coerce')

# subset data
sub1=data[(data['AGE']<=25) & (data['CHECK321']==1) & (data['S3AQ3B1']==1)]


# create binary nictoine dependence variable
def NICOTINEDEP (x):
    if x['TAB12MDX']==1:
        return 1
    else:
        return 0
sub1['NICOTINEDEP'] = sub1.apply (lambda x: NICOTINEDEP (x), axis=1)
print (pandas.crosstab(sub1['TAB12MDX'], sub1['NICOTINEDEP']))
```

```
    if x['TAB12MDX']==1:
        return 1
    else:
        return 0
sub1['NICOTINEDEP'] = sub1.apply (lambda x: NICOTINEDEP (x), axis=1)
print (pandas.crosstab(sub1['TAB12MDX'], sub1['NICOTINEDEP']))


# logistic regression social phobia
lreg1 = smf.logit(formula = 'NICOTINEDEP ~ SOCPDLIFE', data = sub1).fit()
print (lreg1.summary())
```

IPython console

Console 1/A

```
In [28]: lreg1 = smf.logit(formula = 'NICOTINEDEP ~ SOCPDLIFE', data = sub1).fit()
    ...: print (lreg1.summary())
Optimization terminated successfully.
         Current function value: 0.664381
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:            NICOTINEDEP   No. Observations:                 1320
Model:                          Logit   Df Residuals:                     1318
Method:                           MLE   Df Model:                            1
Date:                Sat, 07 Nov 2015   Pseudo R-squ.:                0.009574
Time:                        12:35:30   Log-Likelihood:                -876.98
converged:                       True   LL-Null:                       -885.46
                                        LLR p-value:                 3.829e-05
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      0.3776      0.057      6.569      0.000         0.265     0.490
SOCPDLIFE      1.2318      0.335      3.674      0.000         0.575     1.889
==============================================================================

In [29]:
```

IPython console

Console 1/A

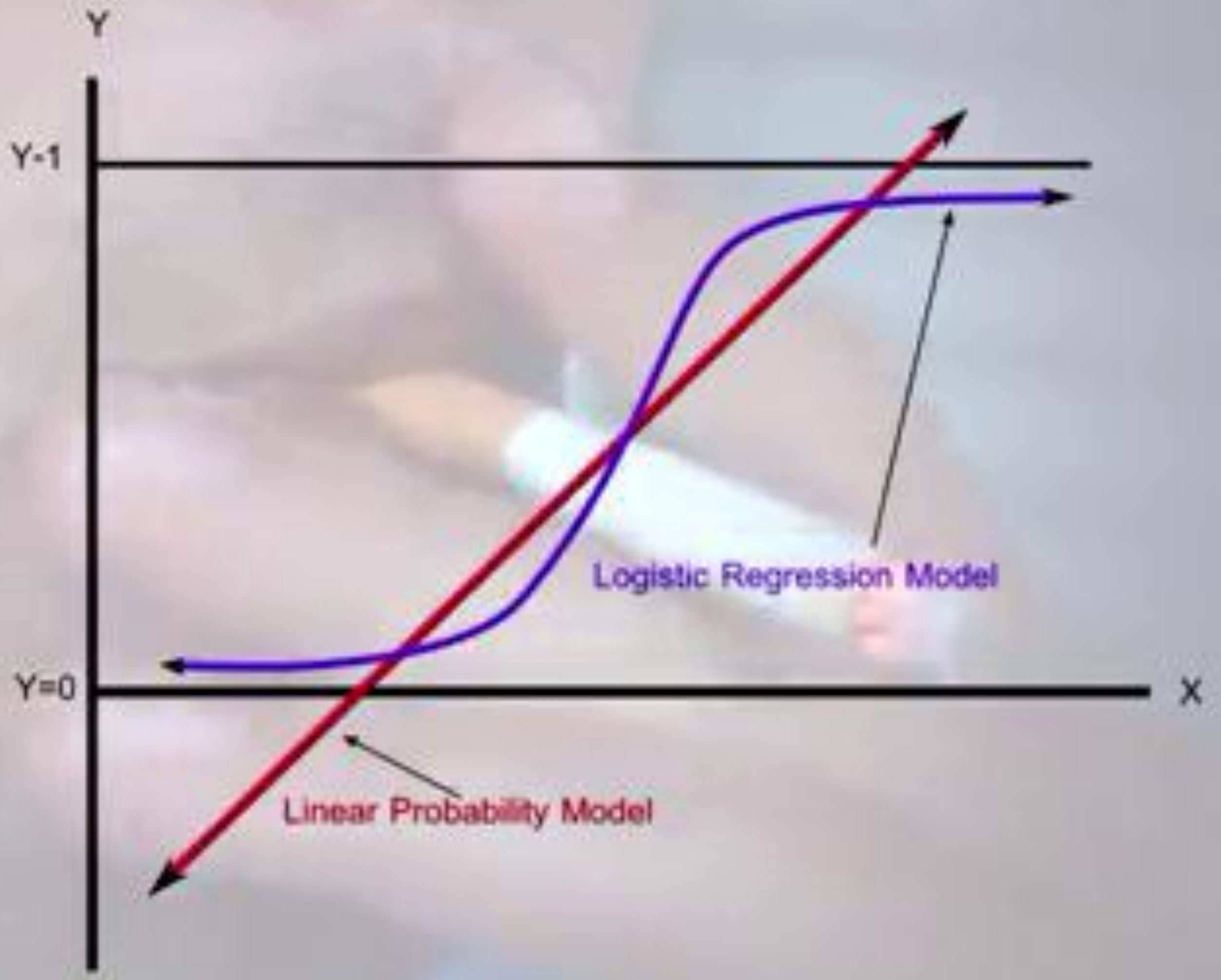
```
In [28]: lreg1 = smf.logit(formula = 'NICOTINEDEP ~ SOCPDLIFE', data = sub1).fit()
    ...: print (lreg1.summary())
Optimization terminated successfully.
         Current function value: 0.664381
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:            NICOTINEDEP   No. Observations:                 1320
Model:                          Logit   Df Residuals:                     1318
Method:                           MLE   Df Model:                            1
Date:                Sat, 07 Nov 2015   Pseudo R-squ.:                0.009574
Time:                        12:35:30   Log-Likelihood:                -876.98
converged:                       True   LL-Null:                       -885.46
                                        LLR p-value:                 3.829e-05
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      0.3776      0.057      6.569      0.000         0.265     0.490
SOCPDLIFE      1.2318      0.335      3.674      0.000         0.575     1.889
==============================================================================

In [29]:
```

NICOTINEDEP = 0.38 + 1.23 * SOCPDLIFE

Are those with social phobia more or less likely to be nicotine dependent?

Comparing the LP and Logit Models



Odds Ratio

$0 \rightarrow \infty$

OR = 1 model statistically non-significant

OR > 1 as explanatory variable increases, response variable more likely.

OR < 1 as explanatory variable increases, response variable is less likely.

Console 1/A

```
Method:                           MLE   Df Model:                            1
Date:                Sat, 07 Nov 2015   Pseudo R-squ.:                0.009574
Time:                        12:35:30   Log-Likelihood:                -876.98
converged:                       True   LL-Null:                       -885.46
                                        LLR p-value:                 3.829e-05
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      0.3776      0.057      6.569      0.000         0.265     0.490
SOCPDLIFE      1.2318      0.335      3.674      0.000         0.575     1.889
==============================================================================

In [29]: print ("Odds Ratios")
    ...: print (numpy.exp(lreg1.params))
    ...:
    ...:
Odds Ratios
Intercept    1.46
SOCPDLIFE    3.43
dtype: float64

In [30]:
```

Permissions: RW   End-of-lines: CRLF   Encoding: UTF-8   Line: 56   Column: 1   Memory: 32 %

```
Model:                          Logit   Df Residuals:                     1318
Method:                           MLE   Df Model:                            1
Date:                Sat, 07 Nov 2015   Pseudo R-squ.:                0.009574
Time:                        12:30:19   Log-Likelihood:                -876.98
converged:                       True   LL-Null:                       -885.46
                                        LLR p-value:                 3.829e-05
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      0.3776      0.057      6.569      0.000         0.265     0.490
SOCPDLIFE      1.2318      0.335      3.674      0.000         0.575     1.889
==============================================================================
Odds Ratios
Intercept   1.46
SOCPDLIFE   3.43
dtype: float64

In [26]: params = lreg1.params
    ...: conf = lreg1.conf_int()
    ...: conf['OR'] = params
    ...: conf.columns = ['Lower CI', 'Upper CI', 'OR']
    ...: print (numpy.exp(conf))
    ...:
           Lower CI  Upper CI   OR
Intercept      1.30      1.63 1.46
SOCPDLIFE      1.78      6.61 3.43

In [27]:
```

Permissions: RW   End-of-lines: CRLF   Encoding: UTF-8   Line: 65   Colu

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
SOCPDLIFE	3.427	1.777	6.611

```
params = lreg1.params
conf = lreg1.conf_int()
conf['OR'] = params
conf.columns = ['Lower CI', 'Upper CI', 'OR']
print (numpy.exp(conf))

#social phobia and depression
lreg2 = smf.logit(formula = 'NICOTINEDEP ~ SOCPDLIFE + MAJORDEPLIFE', data = sub1).fit()
print (lreg2.summary())

# odd ratios with 95% confidence intervals
print ("Odds Ratios")
params = lreg2.params
conf = lreg2.conf_int()
conf['OR'] = params
conf.columns = ['Lower CI', 'Upper CI', 'OR']
print (numpy.exp(conf))
```

```
                    Logit Regression Results
==============================================================================
Dep. Variable:            NICOTINEDEP   No. Observations:                 1320
Model:                          Logit   Df Residuals:                     1317
Method:                           MLE   Df Model:                            2
Date:                Sun, 08 Nov 2015   Pseudo R-squ.:                 0.05758
Time:                        14:43:10   Log-Likelihood:                -834.47
converged:                       True   LL-Null:                       -885.46
                                        LLR p-value:                 7.177e-23
================================================================================
                   coef    std err          z      P>|z|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------
Intercept        0.0939      0.065      1.444      0.149        -0.034     0.221
SOCPDLIFE        0.8393      0.347      2.416      0.016         0.158     1.520
MAJORDEPLIFE     1.3072      0.152      8.588      0.000         1.009     1.606
================================================================================
Odds Ratios
              Lower CI  Upper CI        OR
Intercept     0.967033  1.247795  1.098480
SOCPDLIFE     1.171534  4.573507  2.314740
MAJORDEPLIFE  2.742580  4.980617  3.695909

In [17]:
```

Permissions: RW   End-of-lines: CRLF   Encoding: UTF-8

```
                           Logit Regression Results
==============================================================================
Dep. Variable:            NICOTINEDEP   No. Observations:                 1320
Model:                          Logit   Df Residuals:                     1317
Method:                           MLE   Df Model:                            2
Date:                Sun, 08 Nov 2015   Pseudo R-squ.:                 0.05758
Time:                        14:43:10   Log-Likelihood:                -834.47
converged:                       True   LL-Null:                       -885.46
                                        LLR p-value:                 7.177e-23
================================================================================
                   coef    std err          z      P>|z|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------
Intercept        0.0939      0.065      1.444      0.149        -0.034     0.221
SOCPDLIFE        0.8393      0.347      2.416      0.016         0.158     1.520
MAJORDEPLIFE     1.3072      0.152      8.588      0.000         1.009     1.606
================================================================================
Odds Ratios
              Lower CI  Upper CI        OR
Intercept     0.967033  1.247795  1.098480
SOCPDLIFE     1.171534  4.573507  2.314740
MAJORDEPLIFE  2.742580  4.980617  3.695909

In [17]:
```

Permissions: RW   End-of-lines: CRLF   Encoding: UTF-8

```
params = lreg2.params
conf = lreg2.conf_int()
conf['OR'] = params
conf.columns = ['Lower CI', 'Upper CI', 'OR']
print (numpy.exp(conf))

# logistic regression panic
lreg3 = smf.logit(formula = 'NICOTINEDEP ~ PANIC', data = sub1).fit()
print (lreg3.summary())

# odds ratios
print ("Odds Ratios")
print (numpy.exp(lreg3.params))

# odd ratios with 95% confidence intervals
params = lreg3.params
conf = lreg3.conf_int()
conf['OR'] = params
conf.columns = ['Lower CI', 'Upper CI', 'OR']
print (numpy.exp(conf))
```

IPython console

Console 1/A

```
   ...:
Optimization terminated successfully.
         Current function value: 0.662762
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:            NICOTINEDEP   No. Observations:                 1320
Model:                          Logit   Df Residuals:                     1318
Method:                           MLE   Df Model:                            1
Date:                Sun, 08 Nov 2015   Pseudo R-squ.:                 0.01199
Time:                        14:26:47   Log-Likelihood:                -874.85
converged:                       True   LL-Null:                       -885.46
                                        LLR p-value:                 4.079e-06
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      0.3202      0.061      5.278      0.000         0.201     0.439
PANIC          0.7590      0.172      4.423      0.000         0.423     1.095
==============================================================================
Odds Ratios
Intercept    1.377399
PANIC        2.136134
dtype: float64
           Lower CI  Upper CI        OR
Intercept  1.222987  1.551306  1.377399
PANIC      1.526024  2.990167  2.136134

In [13]:
```

Console 1/A

```
Current function value: 0.633241
```

lreg4 = smf.logit(formula = 'NICOTINEDEP ~ PANIC + MAJORDEPLIFE', data = sub1).fit()

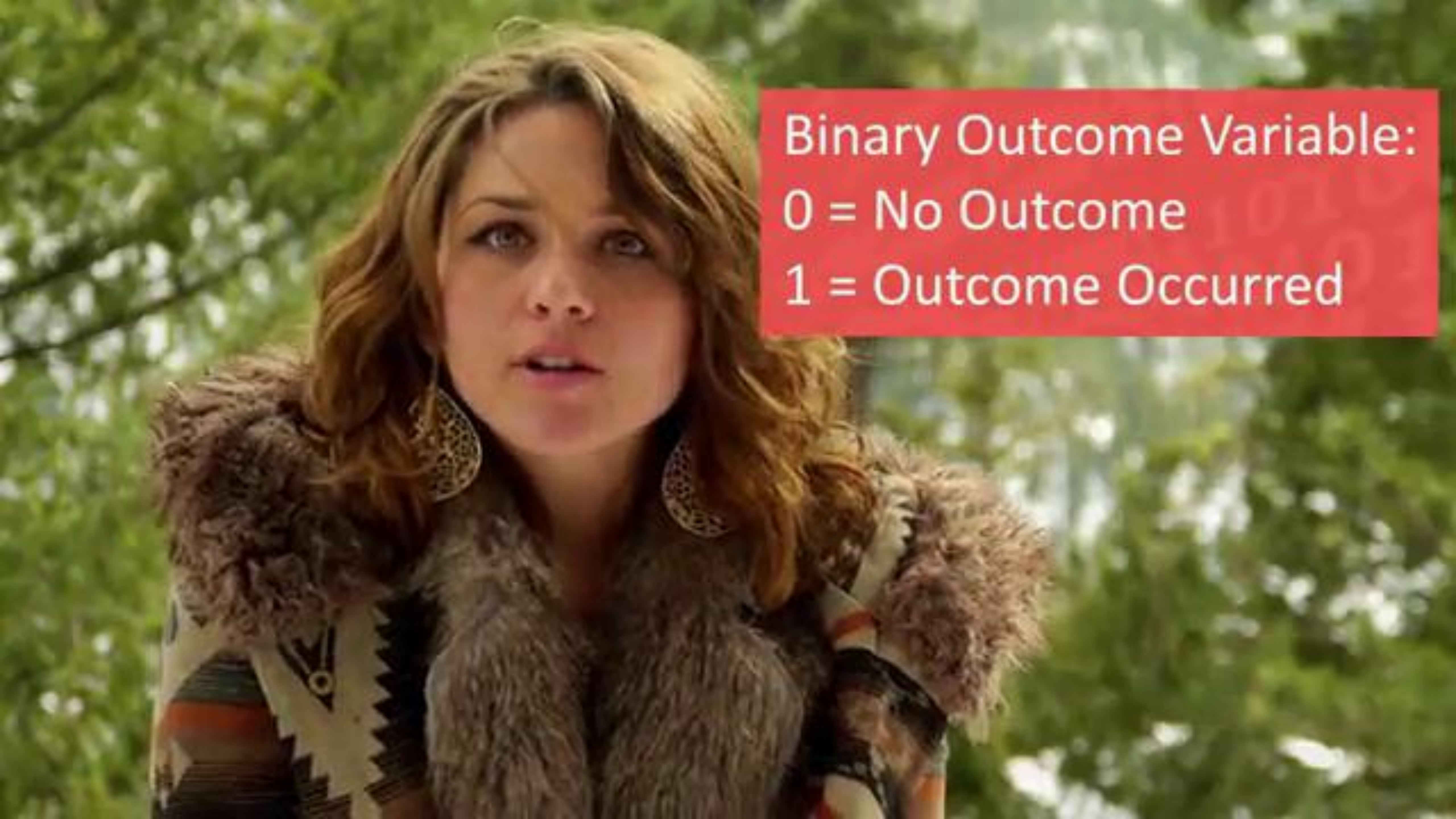
```
Model:                          Logit   Df Residuals:                     1317
Method:                           MLE   Df Model:                            2
Date:                Sun, 08 Nov 2015   Pseudo R-squ.:                 0.05600
Time:                        14:30:41   Log-Likelihood:                -835.88
converged:                       True   LL-Null:                       -885.46
                                        LLR p-value:                 2.930e-22
================================================================================
                   coef    std err          z      P>|z|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------
Intercept        0.0826      0.066      1.243      0.214        -0.048     0.213
PANIC            0.3554      0.183      1.941      0.052        -0.003     0.714
MAJORDEPLIFE     1.2848      0.155      8.266      0.000         0.980     1.589
================================================================================
Odds Ratios
Intercept       1.086115
PANIC           1.426815
MAJORDEPLIFE    3.613795
dtype: float64
              Lower CI  Upper CI        OR
Intercept     0.953426  1.237270  1.086115
PANIC         0.996509  2.042933  1.426815
MAJORDEPLIFE  2.664843  4.900671  3.613795

In [15]:
```

Console 1/A

```
Optimization terminated successfully.
         Current function value: 0.633241
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:            NICOTINEDEP   No. Observations:                 1320
Model:                          Logit   Df Residuals:                     1317
Method:                           MLE   Df Model:                            2
Date:                Sun, 08 Nov 2015   Pseudo R-squ.:                 0.05600
Time:                        14:30:41   Log-Likelihood:                -835.88
converged:                       True   LL-Null:                       -885.46
                                        LLR p-value:                 2.930e-22
================================================================================
                   coef    std err          z      P>|z|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------
Intercept        0.0826      0.066      1.243      0.214        -0.048     0.213
PANIC            0.3554      0.183      1.941      0.052        -0.003     0.714
MAJORDEPLIFE     1.2848      0.155      8.266      0.000         0.980     1.589
================================================================================
Odds Ratios
Intercept       1.086115
PANIC           1.426815
MAJORDEPLIFE    3.613795
dtype: float64
              Lower CI  Upper CI        OR
Intercept     0.953426  1.237270  1.086115
PANIC         0.996509  2.042933  1.426815
MAJORDEPLIFE  2.664843  4.900671  3.613795

In [15]: 
```


Binary Outcome Variable:
0 = No Outcome
1 = Outcome Occurred