Nhu Vo

Lab #5

```
from __future__ import division
import pandas as pd
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
import os
import matplotlib.pyplot as plt
```

Use the 2006 GSS again.

```
from google.colab import files
uploaded = files.upload()

import io
g = pd.read_csv(io.BytesIO(uploaded['GSS.2006.csv']))
```

Choose Files GSS.2006.csv

- **GSS.2006.csv**(text/csv) - 8232424 bytes, last modified: 6/7/2023 - 100% done

```
Saving GSS.2006.csv to GSS.2006 (1).csv
```

```
g["happy"]
```

```
0       2.0
1       1.0
2       2.0
3       1.0
4       2.0
       ...
4505    2.0
4506    2.0
4507    NaN
4508    3.0
4509    3.0
Name: happy, Length: 4510, dtype: float64
```

```
g.head()
```

	vpsu	vstrat	adults	ballot	dateintv	famgen	form	formwt	gender1	hompop	..
0	1	1957	1	3	316	2	1	1	2	3	.
1	1	1957	2	2	630	1	2	1	2	2	.
2	1	1957	2	2	314	2	1	1	2	2	.
3	1	1957	1	1	313	1	2	1	2	1	.
4	1	1957	3	1	322	2	2	1	2	3	.

5 rows × 1261 columns

1. Run a multiple linear probability model (have at least 2 Xs in the model). Tell me how you think your independent variables will affect your dependent variable. Interpret your results. Were your expectations correct? Why or why not?

The question is what are some things that predict how appy you are

I want to see how my selected independent variables affect my dependent variables:

Indepdent Variable:

1) SOCFREND: Would you use this card and tell me which answer comes closest to how often you do the following things: spend a social evening with friends who live outside your neighborhood?

1) Almost everyday

2) Once or twice a week

3) Several times a month

4) About once a month

5) Several times a year

6) About once a year

7) Never

2) SATJOB: On the whole, how satisfied are you with the work you do? Would you say you are very satisfied, moderately satisfied, a little dissatisfied or very dissatisfied?

1) Very satisfied

2) Moderately satisfied

3) A little dissatisfied

4) Very dissatisfied

3) HRS1: Hours worked in the past week (HRS1) - recoded (HRS1R)

1) 1-24 hours

2) 25-39 hours

3) 40-49 hours

4) 50-59 hours

5) 60+ hours

4) CHILDS: Number of children

0) 0

1) 1

2) 2

3) 3-4

4) 5+

5) SPRTPRSN: To what extent do you consider yourself a spiritual person?

1) Very spiritual

2) Moderately spiritual

3) Slightly spiritual

4) Not spiritual

Dependent Variable:

HAPPY: Taken all together, how would you say things are these days? Would you say that you are very happy, pretty happy or not too happy?

1) Very happy

2) Pretty happy

3) Not too happy

```
#reverse order of "socfrend" making higher score --> higher frequency INSTEAD of lower score --> higher frequency
g["rsocfrend"] = 3-g["socfrend"]


#reverse order of "sprtprsn" making higher score --> higher spirituality INSTEAD of lower score --> higher spirituality
g["rsprtprsn"] = 5-g["sprtprsn"]


#reverse order of satjob making higher score = higher satisfaction because originally, health scale is lower score --> more satisf
g["rsatjob"] = 5-g["satjob"]


#reverse order of "happy" making higher score --> more happy INSTEAD of lower score --> happier
# 1) Not too happy
# 2) Pretty happy
# 3) Very happy
g["rhappy"] = 4-g["happy"]


# get rid of all missings; necessary for predictions later ##

g.dropna(subset = ["rhappy", "rsocfrend", "rsatjob", "hrs1", "childs", "rsprtprsn"], inplace=True)



lm1 = smf.ols(formula = 'rhappy ~ rsocfrend + rsatjob + hrs1 + rsprtprsn + childs', data = g).fit()
print (lm1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 rhappy   R-squared:                       0.077
Model:                            OLS   Adj. R-squared:                  0.073
Method:                 Least Squares   F-statistic:                     19.18
Date:                Tue, 27 Jun 2023   Prob (F-statistic):           2.28e-18
Time:                        01:10:15   Log-Likelihood:                -979.69
No. Observations:                1157   AIC:                             1971.
Df Residuals:                    1151   BIC:                             2002.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.3711      0.099     13.909      0.000       1.178       1.565
rsocfrend      0.0119      0.011      1.101      0.271      -0.009       0.033
rsatjob        0.1765      0.021      8.494      0.000       0.136       0.217
hrs1           0.0012      0.001      1.055      0.292      -0.001       0.003
rsprtprsn      0.0691      0.018      3.803      0.000       0.033       0.105
childs         0.0133      0.011      1.178      0.239      -0.009       0.036
==============================================================================
Omnibus:                        9.609   Durbin-Watson:                   1.856
Prob(Omnibus):                  0.008   Jarque-Bera (JB):                6.731
Skew:                          -0.030   Prob(JB):                       0.0345
Kurtosis:                       2.631   Cond. No.                         268.
==============================================================================

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

Intercept = 1.3711

The intercept coefficient = 1.3711 represents the expected value of happy when all the independent variables are zero. In other words, it indicates the baseline happiness level.

rsocfrend's coefficient = 0.0119 = 1.2%

For every category increase in socfrends, an individual is 1.2 percentage points more likely to be happier, net of other independent variables. However, this coefficient is not statistically significant (p-value = 0.850), indicating that the relationship between socfrend and happy is not statistically significant.

rsatjob's coefficient = 0.1765 = 17.7%

For every category increase in satjob, an individual is 17.7 percentage points more likely to be happier, net of other independent variables. This coefficient is statistically significant (p-value = 0), indicating that higher job satisfaction is associated with a higher probability of being happy. ***

hrs1's coefficient = 0.0012 = 0.12%

For every category increase in hrs1, an individual is 0.12 percentage points more likely to be happier, net of other independent variables. However, this coefficient is not statistically significant (p-value = 0.292), indicating that the relationship between hrs1 and happy is not statistically significant.

rsprtprsn's coefficient = 0.0691 = 6.9%

For every category increase in sprtprsn, an individual is 6.9 percentage points more likely to be happier, net of other independent variables. This coefficient is statistically significant (p-value = 0), indicating that being more spiritual is associated with a higher probability of being happy.***

childs' coefficient = 0.0133 = 1.3%

For every category increase in childs, an individual is 1.3 percentage points more likely to be happier, net of other independent variables. However, this coefficient is not statistically significant (p-value = 0.239s), indicating that the relationship between childs and happy is not statistically significant.

Only satjob and sprtprsn are statistically significant regarding its correlation with level of happiness.

2. Run a multiple (binary) logistic model. (It can be the same as the above LPM or a new model.) If it's a new model, tell me how you think your independent variables will affect your dependent variable. Interpret your results in the logit scale. Were your expectations correct? Why or why not?

```
q2_g = g.copy()
```

Recoding the 'happy' variable so that the "Very Happy" category is 1, and the other two categories are assigned 0.

```
#replace happy variable so that 1 = 1, and more than 1 = 0
q2_g['rhappy'].replace({1: 0, 2: 0, 3: 1}, inplace=True)
q2_g['rhappy']
```

```
1       1.0
3       1.0
5       0.0
9       1.0
13      0.0
       ...
4482    0.0
4486    1.0
4492    1.0
4500    1.0
4501    0.0
Name: rhappy, Length: 1157, dtype: float64
```

```
logit1 = sm.formula.logit(formula = "rhappy ~ rsocfrend + rsatjob + hrs1 + rsprtprsn + childs", data = q2_g).fit()
print (logit1.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.588672
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                 rhappy   No. Observations:                 1157
Model:                          Logit   Df Residuals:                     1151
Method:                           MLE   Df Model:                            5
Date:                Tue, 27 Jun 2023   Pseudo R-squ.:                0.04300
Time:                        01:10:30   Log-Likelihood:                -681.09
```

```
converged:                       True   LL-Null:                      -711.70
Covariance Type:            nonrobust   LLR p-value:                6.841e-12
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.7068      0.438     -8.465      0.000      -4.565      -2.849
rsocfrend     -0.0129      0.043     -0.301      0.763      -0.097       0.071
rsatjob        0.5497      0.094      5.863      0.000       0.366       0.733
hrs1           0.0009      0.005      0.194      0.846      -0.008       0.010
rsprtprsn      0.3023      0.074      4.076      0.000       0.157       0.448
childs         0.0518      0.044      1.174      0.241      -0.035       0.138
==============================================================================
```

Intercept = -3.7068

This is the log-odds of being happy when all other independent variables are set to zero.

socfrend = -0.0129

for every category increase in socfrend, an individual decrease their logit by -0.013 of being happy, net of all the other independent variables. This effect is not statistically significant

satjob = 0.5497

for every category increase in satjob, an individual increase their logit by 0.5497 of being happy, net of all the other independent variables. The positive coefficient suggests that higher job satisfaction significantly increases the likelihood of being happy, as confirmed by the statistically significant p-value (< 0.000).

hrs1 = 0.0009

for every category increase in hrs1, an individual increase their logit by 0.0009 of being happy, net of all the other independent variables. This effect is not statistically significant

sprtprsn = 0.3023

for every category increase in hrs1, an individual increase their logit by 0.3023 of being happy, net of all the other independent variables. The positive coefficient suggests that being more spiritual significantly increases the likelihood of being happy, as confirmed by the statistically significant p-value (< 0.000).

childs = 0.3023

for every category increase in childs, an individual increase their logit by 0.3023 of being happy, net of all the other independent variables. This effect is not statistically significant

3. Get odds ratios from your logit model in Question 2 and interpret some of them.

```
np.exp(logit1.params)
```

```
Intercept    0.024555
rsocfrend    0.987218
rsatjob      1.732742
hrs1         1.000876
rsprtprsn    1.352947
childs       1.053158
dtype: float64
```

rsocfrend:

The odds ratio of 0.987218 indicates that a one-unit increase in rsocfrend is associated with a 1.278% (100 * (1 - 0.987218)) decrease in the odds of being happy, holding other variables constant.

rsatjob:

The odds ratio of 1.732742 suggests that a one-unit increase in rsatjob is associated with a 73.27% (100 * (1.732742 - 1)) increase in the odds of being happy, holding other variables constant.

hrs1:

The odds ratio of 1.000876 indicates that a one-unit increase in hrs1 is associated with a negligible 0.0876% (100 * (1.000876 - 1)) increase in the odds of being happy, holding other variables constant. Since the increase is very small, it suggests that hrs1 has a minimal impact on the odds of being happy.

rsprtprsn:

The odds ratio of 1.352947 implies that a one-unit increase in rsprtprsn is associated with a 35.29% (100 * (1.352947 - 1)) increase in the odds of being happy, holding other variables constant.

childs:

The odds ratio of 1.053158 suggests that a one-unit increase in the number of children (childs) is associated with a 5.32% (100 * (1.053158 - 1)) increase in the odds of being happy, holding other variables constant.