Lab 6 (QMSS5015 Data Analysis)

Submitted by: Gideon Tay
My UNI: gt2528
Contact me at: gideon.tay@columbia.edu

Overview: for this lab, we will explore the factors affecting the Total early-stage Entrepreneurial Activity (TEA) in a country.

Import libraries and load in the data

For this lab, I will use the data from Global Entrepreneurship Monitor's Adult Population Survey (APS). Download their 2020 cross-sectional national level data 'GEM 2020 APS Global National Level Data' as a 'sav' file.

Moreover, supplement this dataset with University of Gothenburg's Quality of Government (QoG) Institute's basic cross-sectional dataset which includes data on countries from around 2020. Download the dataset as a csv from the website.

Since our key dependent variable of interest, the TEA, is found in the GEM dataset, we will left join the QoG dataset to the GEM dataset on country names. We do not need additional data for countries without TEA data.

In [3]:
```
# Import libraries needed for this lab assignment
import pandas as pd
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf

# Load in the data from the sav and csv stored in the local directory
df1 = pd.read_spss('GEM 2020 APS Global National Level Data_7April2021.sav')
df2 = pd.read_csv('qog_bas_cs_jan24.csv')

# Left join the GEM and QoG datasets
df = pd.merge(df1, df2, left_on='country', right_on='cname', how='left')

# Print the shape of the resulting dataset
print('Merged dataframe shape:', df.shape)
print('No. of countries in merged dataset:', df.shape[0])
print('No. of variables in merged dataset:', df.shape[1])

# View the first 5 rows of the data
df.head(5)
```

```
Merged dataframe shape: (43, 574)
No. of countries in merged dataset: 43
No. of variables in merged dataset: 574
```

Out[3]:

	country	country_name	ctryalp	REGION	WBinc	WBincREV	Bstart20	Bjobst
0	United States	United States	United States	Europe & N America	High	High	16.849614	7.6023
1	Russia	Russia	Russia	Europe & N America	Upper Middle	Middle	10.667370	4.6389
2	Egypt	Egypt	Egypt	Midde East & Africa	Lower Middle	Low	24.651582	18.1569
3	Greece	Greece	Greece	Europe & N America	High	High	7.818968	2.9826
4	Netherlands	Netherlands	Netherlands	Europe & N America	High	High	12.806618	10.1971

5 rows × 574 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?

Dependent variable (`TEA20`): the Total early-stage Entrepreneurial Activity (TEA) index. It is defined as the percentage of the 18-64 population who is either an active full or part owner of a nascent business which has not yet paid salaries for over 3 months, or owner-manager of a new business which has paid salaries for between 3 and 42 months.

Independent variables (X): we think the following variables are associated with TEA

- **Property rights (`wef_pr`)**: In your country, to what extent are property rights, including financial assets, protected? [1 = not at all; 7 = to a great extent]
- **Informal investor activity (`BUSANGVL`)**: the percentage of respondents who were informal investors in the last 3 years (and provided value of investment).
- **Perceived opportunities (`Opport20`)**: the percentage of 18-64 population who think that in the next 6 months there will be good opportunities for starting a business in the area where they live

Expectation: I believe that high TEA should be positively associated with a stronger property rights (`wef_pr`), higher informal investor activity (`BUSANGVL`) , and higher levels of perceived opportunities (`Opport20`).

Let's first drop all rows (countries) with incomplete data (without our variables of interest), since those rows are not suitable for analysis:

In [ ]:
```
# Drop rows with incomplete data
sub = df.dropna(subset = ['TEA20', 'wef_pr', 'BUSANGVL', 'Opport20'])

# Print the shape of the resulting dataset
print('Dataframe shape after dropping countries with incomplete data:',
      f'{sub.shape}')
print('No. of countries in dataset:', sub.shape[0])
print('No. of variables in dataset:', sub.shape[1])
```

```
Dataframe shape after dropping countries with incomplete data: (34, 574)
No. of countries in dataset: 34
No. of variables in dataset: 574
```

The TEA data in theory ranges from 0 to 100 since it is a percentage. Let's find the range of TEA for countries in our dataset:

In [10]:
```
# Find min and max TEA in our dataset
maxTEA = sub["TEA20"].max()
minTEA = sub["TEA20"].min()
medianTEA = sub["TEA20"].median()
print(f"Maximum TEA: {maxTEA}")
print(f"Minimum TEA: {minTEA}")
print(f"Median TEA: {medianTEA}")
```

```
Maximum TEA: 49.62224395590331
Minimum TEA: 1.9243090378454775
Median TEA: 11.96330017280575
```

Recode TEA to binary variable: Given this TEA range, let's recode TEA into a binary variable so low TEA (0 to 12) is 0 and high TEA (>12 to 100) is 1. Based on this definition, let's view the number of high and low TEA countries:

In [12]:
```
# Recode high and low TEA
conditions = [(sub['TEA20'] <= 12) , (sub['TEA20'] > 12)]
choices = [0, 1]
sub['high_TEA'] = np.select(conditions, choices, default=np.nan)

# Display number of high and low TEA countries
pd.crosstab(index=sub["high_TEA"], columns="count")
```

```
C:\Users\gideo\AppData\Local\Temp\ipykernel_11872\712408503.py:4: SettingWithCopyWar
ning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  sub['high_TEA'] = np.select(conditions, choices, default=np.nan)
```

Out[12]:

high_TEA	col_0 count
0.0	17
1.0	17

Run model: Now, let's run a multiple linear probability model...

In [14]:
```
lm1 = smf.ols(formula = 'high_TEA ~ wef_pr + BUSANGVL + Opport20', data = sub).fit(
print (lm1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               high_TEA   R-squared:                       0.399
Model:                           OLS   Adj. R-squared:                  0.339
Method:                Least Squares   F-statistic:                     6.635
Date:               Fri, 06 Dec 2024   Prob (F-statistic):            0.00143
Time:                       10:57:02   Log-Likelihood:                -16.026
No. Observations:                 34   AIC:                             40.05
Df Residuals:                     30   BIC:                             46.16
Df Model:                          3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.9370      0.537      1.745      0.091      -0.160       2.034
wef_pr        -0.1964      0.095     -2.073      0.047      -0.390      -0.003
BUSANGVL       0.0582      0.018      3.271      0.003       0.022       0.095
Opport20       0.0041      0.004      1.053      0.301      -0.004       0.012
==============================================================================
Omnibus:                        7.844   Durbin-Watson:                   1.747
Prob(Omnibus):                  0.020   Jarque-Bera (JB):                3.286
Skew:                           0.476   Prob(JB):                        0.193
Kurtosis:                       1.812   Cond. No.                         430.
==============================================================================

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

Result interpretation:

- **Property rights (`wef_pr`)** is negatively associated with high TEA, and the association is statistically significant (p-value < 0.05).
  - This is directly opposite of our initial expectation. Property rights is negatively rather than positively associated with high TEA.
  - Our initial expectation is based on a hypothesized causal effect. One possible explanation for this result is that property rights and TEA are related by a confounding variable which produces this unintuitive result.
- **Informal investor activity (`BUSANGVL`)** is positively associated with high TEA, and the association is statistically significant (p-value < 0.01).

    - This is in line with our expectation.
- **Perceived opportunities (`Opport20`)** has a very small coefficient and is not statistically significant (p-value > 0.1).
    - Unlike our expectation, perceived opportunities appears to not be associated with performance of the economy, after controlling for property rights and informal investor activity.
    - It is still possible that our expectation holds true in a model of just TEA against perceived opportunities, where we do not control for property rights or informal investor activity.

2. Run a multiple (binary) logistic model. (It can be the same as the above LPM or a new model.) If it is 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?

Let's run the logit on the same independent and dependent variables as in question (1):

In [15]:
```
logit1 = sm.formula.logit(formula = 'high_TEA ~ wef_pr + BUSANGVL + Opport20', data
print (logit1.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.362068
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:               high_TEA   No. Observations:                   34
Model:                          Logit   Df Residuals:                       30
Method:                           MLE   Df Model:                            3
Date:                Fri, 06 Dec 2024   Pseudo R-squ.:                  0.4776
Time:                        11:05:07   Log-Likelihood:                -12.310
converged:                       True   LL-Null:                       -23.567
Covariance Type:            nonrobust   LLR p-value:                 5.100e-05
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.7149      3.870      0.443      0.658      -5.870       9.300
wef_pr        -1.7454      0.938     -1.860      0.063      -3.585       0.094
BUSANGVL       1.1476      0.578      1.987      0.047       0.015       2.280
Opport20       0.0458      0.036      1.269      0.205      -0.025       0.116
==============================================================================

Possibly complete quasi-separation: A fraction 0.12 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

Result interpretation:

- **Property rights (`wef_pr`)** is still negatively associated with high TEA, but the association is no longer statistically significant (p-value > 0.05).

    - Unlike our expectation, property rights appears to not be associated with performance of the economy, after controlling for property rights and informal investor activity.
    - The directionality of the coefficient (negative) is still against our expectation.
- **Informal investor activity (`BUSANGVL`)** is still positively associated with high TEA, and the association is still statistically significant (p-value < 0.01).
    - This is in line with our expectation.
- **Perceived opportunities (`Opport20`)** is still not statistically significant (p-value > 0.1).
    - Unlike our expectation, perceived opportunities appears to not be associated with performance of the economy, after controlling for property rights and informal investor activity.
    - It is still possible that our expectation holds true in a model of just TEA against perceived opportunities, where we do not control for property rights or informal investor activity.

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

In [ ]:
```
# Get odd ratios by taking exponent of logit model coefficients
np.exp(logit1.params)
```

Out[ ]:
```
Intercept    5.556194
wef_pr       0.174567
BUSANGVL     3.150574
Opport20     1.046815
dtype: float64
```

In [ ]:
```
# Find percentage decrease for wef_pr
1-0.174567
```

Out[ ]:
```
0.825433
```

Interpretation:

- For each one-unit increase in property rights (on 1 to 7 scale), the odds of the country being considered as having high TEA decrease by 82.5%.
- For each 1% increase in informal investor activity, the odds of the country being considered as having high TEA increases by 115.1%
- For each 1% increase in perceived opportunities, the odds of the economy being considered as having high TEA increase by 4.7%.

4. Get predicted probabilities from your logit model in Question 2 for some constellations of X values and interpret the results.

First, let's define a function which gives us probability from logit value. We also create variables for the intercept and logit coefficient values for each dependent variable in our model:

In [24]:
```
# Define a function which gives us probability from logit value
def logit2prob (logit):
    odds = np.exp(logit)
    prob = odds / (1 + odds)
    return(prob)

intercept = logit1.params.Intercept
b_wef_pr = logit1.params.wef_pr
b_BUSANGVL = logit1.params.BUSANGVL
b_Opport20 = logit1.params.Opport20
```

Constellation 1

- For X values:
  - wef_pr = 7 (recall this is on a 1 to 7 scale)
  - BUSANGVL = 10 (recall this is a %)
  - Opport20 = 5 (recall this is a %)
- Predicted probability: 77% chance of economy being considered as having high TEA

In [ ]:
```
# X values
value_wef_pr = 7
value_BUSANGVL = 10
value_Opport20 = 5

# Calculate predicted probability
logits_exh = intercept + (value_wef_pr * b_wef_pr) + (value_BUSANGVL * b_BUSANGVL)
pred_prob = logit2prob(logits_exh)
print(f"Predicted probability: {round(pred_prob,2)}")
```

```
Predicted probability: 0.77
```

Constellation 2

- For X values:
  - wef_pr = 3 (recall this is on a 1 to 7 scale)
  - BUSANGVL = 5 (recall this is a %)
  - Opport20 = 35 (recall this is a %)
- Predicted probability: 98% chance of economy being considered as having high TEA

In [41]:
```
# X values
value_wef_pr = 3
value_BUSANGVL = 5
value_Opport20 = 35

# Calculate predicted probability
logits_exh = intercept + (value_wef_pr * b_wef_pr) + (value_BUSANGVL * b_BUSANGVL)
```

```
pred_prob = logit2prob(logits_exh)
print(f"Predicted probability: {round(pred_prob,2)}")
```

```
Predicted probability: 0.98
```

Constellation 3

- For X values:
  - wef_pr = 6 (recall this is on a 1 to 7 scale)
  - BUSANGVL = 9 (recall this is a %)
  - Opport20 = 20 (recall this is a %)
- Predicted probability: 92% chance of economy being considered as having high TEA

In [43]:
```
# X values
value_wef_pr = 6
value_BUSANGVL = 9
value_Opport20 = 20

# Calculate predicted probability
logits_exh = intercept + (value_wef_pr * b_wef_pr) + (value_BUSANGVL * b_BUSANGVL)
pred_prob = logit2prob(logits_exh)
print(f"Predicted probability: {round(pred_prob,2)}")
```

```
Predicted probability: 0.92
```