Lab 4 example

Allan Tan

Initial code to load in data, etc.

```
import zipfile
import pandas as pd
import requests
import io
from tqdm.notebook import tqdm

from __future__ import division
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
import os
import matplotlib.pyplot as plt
from scipy.stats import skew, kurtosis
import seaborn as sns

# Step 1: Download the ZIP file with progress bar
url = 'https://web.archive.org/web/20250601100319/https://gss.norc.org

# Make a streaming request to get the content in chunks
response = requests.get(url, stream=True)
total_size = int(response.headers.get('content-length', 0))  # Get the
block_size = 1024  # 1 Kilobyte

# Progress bar for downloading
tqdm_bar = tqdm(total=total_size, unit='iB', unit_scale=True)
content = io.BytesIO()

# Download the file in chunks with progress bar
for data in response.iter_content(block_size):
    tqdm_bar.update(len(data))
    content.write(data)

tqdm_bar.close()
```

```
    # Check if the download is successful
    if total_size != 0 and tqdm_bar.n != total_size:
        print("Error in downloading the file.")
    else:
        print("Download completed!")

    # Step 2: Extract the ZIP file in memory and display progress
    with zipfile.ZipFile(content) as z:
        # List all files in the zip
        file_list = z.namelist()

        # Filter for the .dta file (assuming there is only one)
        stata_files = [file for file in file_list if file.endswith('.dta'

        # If there is a Stata file, proceed to extract and read it
        if stata_files:
            stata_file = stata_files[0]  # Take the first .dta file

            # Step 3: Load the dataset without 'hhtype'
            with z.open(stata_file) as stata_file_stream:
                # First, read the file to get all column names
                print("Loading dataset to determine columns...")
                df_columns = pd.read_stata(stata_file_stream, convert_cat

                # Get all column names and exclude 'hhtype'
                all_columns = df_columns.variable_labels().keys()
                columns_to_load = [col for col in all_columns if col != '

            # Reload the file to load only the selected columns
            with z.open(stata_file) as stata_file_stream:
                print("Loading dataset with numeric labels excluding 'hht
                df_numeric = pd.read_stata(stata_file_stream, columns=col
                print("Data with numeric labels loaded successfully!")

            # Reload the file again to load only the selected columns with
            with z.open(stata_file) as stata_file_stream:
                print("Loading dataset with string (categorical) labels e
                df_categorical = pd.read_stata(stata_file_stream, columns

                # Step to rename categorical columns with a 'z' prefix
                df_categorical = df_categorical.rename(columns={col: f'z{
                print("Categorical columns renamed with 'z' prefix.")

    # Step 4: Concatenate both numeric and categorical dataframes
    df = pd.concat([df_numeric, df_categorical], axis=1)
```

```
# The final dataframe is now called `df` and contains all variables e:

# Step 5: Display the first few rows of the final DataFrame
df.head()
```

100% 1.69M/1.69M [00:00<00:00, 6.87MiB/s]

```
Download completed!
Loading dataset to determine columns...
Loading dataset with numeric labels excluding 'hhtype'...
Data with numeric labels loaded successfully!
Loading dataset with string (categorical) labels excluding 'hhtype'...
Categorical columns renamed with 'z' prefix.
```

	year	id	wrkstat	hrs1	hrs2	evwork	wrkslf	wrkgovt	occ80	prestg8
0	2006	1	1.0	35.0	NaN	NaN	2.0	2.0	95.0	66.
1	2006	2	1.0	40.0	NaN	NaN	2.0	2.0	243.0	44.
2	2006	3	5.0	NaN	NaN	1.0	2.0	2.0	715.0	29.
3	2006	4	2.0	24.0	NaN	NaN	2.0	2.0	313.0	46.
4	2006	5	6.0	NaN	NaN	2.0	NaN	NaN	NaN	Nal

5 rows × 2646 columns

CODEBOOK: The GSS 2006 data can be looked at here:
https://www.thearda.com/data-archive?tab=2&fid=GSS2006

Question #1- Run a simple regression, with at least two Xs in it (one X should be continuous-ish and the other should be a binary (0 vs 1), and interpret your results. Did the results fit your expectations? Why? Why not?

One important predictor of how people rate their health is their age, since health tends to decline gradually as people get older. Gender may also play a role, as women often report slightly worse health or more health limitations than men. It is possible that these two factors together help explain variation in self-rated health. Based on prior research and general expectations, I would expect age to have a positive relationship with the health score (indicating worse health at older ages), and I would expect women to report slightly worse health than men on average.

```
def ifelse_num_float(var, condition, yes, no):
    var = pd.Series(var)
    cond = pd.Series(condition)
    out = pd.Series(np.nan, index=var.index, dtype='float')
    mask = var.notna()
    out.loc[mask & cond] = float(yes)
    out.loc[mask & ~cond] = float(no)
    return out
```

```
df['female'] = ifelse_num_float(df['sex'], df['sex'] == 2, 1, 0)
```

```
print(df['age'].value_counts(dropna=False))
```

```
age
47.0    110
48.0    109
36.0    105
44.0    100
42.0     99
       ...
85.0     18
84.0     17
86.0     17
88.0     11
87.0      7
Name: count, Length: 73, dtype: int64
```

```
# let's check that original variable is recoded properly
pd.crosstab(df['sex'], df['female'])
```

female	0.0	1.0
sex		
1	2003	0
2	0	2507

Basic descriptive statistics show that the sample is mostly middle-aged on average, and women make up about half of respondents. Health scores range from excellent to poor, and the averages are consistent with typical GSS patterns. Overall, the variables look well-behaved and suitable for regression analysis.

```
df[['health', 'age', 'female', 'sibs']].describe()
```

	health	age	female	sibs
count	3516.000000	4492.000000	4510.000000	2988.000000
mean	2.033561	47.141585	0.555876	3.211847
std	0.835901	16.894264	0.496923	1.939571
min	1.000000	18.000000	0.000000	0.000000
25%	1.000000	34.000000	0.000000	2.000000
50%	2.000000	46.000000	1.000000	3.000000
75%	3.000000	59.000000	1.000000	5.000000
max	4.000000	89.000000	1.000000	6.000000

A quick correlation matrix showed that health is positively correlated with age ($r \approx 0.20$), meaning older respondents tend to rate their health worse. The correlation between gender and health is small ($r \approx 0.04$), indicating only a slight difference between men and women on average. The remaining correlations are weak, which is expected given that these demographic variables are only loosely related. Overall, the correlations align with theoretical expectations and support the patterns observed in the descriptive statistics.

```
df[['health', 'age', 'female', 'sibs']].corr()
```

	health	age	female	sibs
health	1.000000	0.197461	0.038101	0.141230
age	0.197461	1.000000	0.034091	0.090697
female	0.038101	0.034091	1.000000	0.019406
sibs	0.141230	0.090697	0.019406	1.000000

```
result = smf.ols(formula='health ~ age + female', data=df).fit()
print(result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 health   R-squared:
Model:                            OLS   Adj. R-squared:
Method:                 Least Squares   F-statistic:
Date:                Tue, 18 Nov 2025   Prob (F-statistic):
Time:                        19:27:44   Log-Likelihood:
No. Observations:                3504   AIC:
Df Residuals:                    3501   BIC:
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025
------------------------------------------------------------------------------
Intercept      1.5482      0.043     35.686      0.000       1.463
age            0.0097      0.001     11.823      0.000       0.008
female         0.0457      0.028      1.637      0.102      -0.009
==============================================================================
Omnibus:                      134.016   Durbin-Watson:
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
Skew:                           0.470   Prob(JB):
Kurtosis:                       2.710   Cond. No.
==============================================================================

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

Results from the overall model show that age is a strong and significant predictor of self-rated health. Each additional year of age is associated with about a 0.0097-point increase in the health score, meaning older respondents report slightly worse health. The coefficient for female is positive but not statistically significant, suggesting only a small and inconclusive gender difference in self-rated health. The R-squared value (0.04) indicates that the model explains a modest amount of variation, which is typical for survey-based health measures.

Question 2. Now, run to separate regressions by subgroup (based on the binary X from above). Explain why you would expect different slopes between these two models. Which slope should be bigger or a different sign than the other slope? Explain your results. Did it work out? Yes? No?

Next, I ran separate regressions for men (female=0) and women (female=1) to see whether the effect of age on health differs by gender. I expected the slope for women to be larger because prior research shows that women often report health problems more frequently than men, especially as they age.

The results confirm this expectation. The coefficient for age is larger among women than among men, indicating that aging is associated with worse self-rated health more strongly for female respondents. This suggests that gender moderates the relationship between age and health.

```
female1 = smf.ols('health ~ age', data=df, subset=df['female']==1).fi
print(female1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 health   R-squared:
Model:                            OLS   Adj. R-squared:
Method:                 Least Squares   F-statistic:
Date:                Tue, 18 Nov 2025   Prob (F-statistic):           5
Time:                        19:32:49   Log-Likelihood:
No. Observations:                1941   AIC:
Df Residuals:                    1939   BIC:
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025
------------------------------------------------------------------------------
Intercept      1.6256      0.057     28.696      0.000       1.515
age            0.0091      0.001      8.170      0.000       0.007
==============================================================================
Omnibus:                       81.096   Durbin-Watson:
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
Skew:                           0.476   Prob(JB):                     8
Kurtosis:                       2.650   Cond. No.
==============================================================================

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

Among women, the slope is about 0.0091, indicating that health worsens slightly with each additional year of age.

```
female0 = smf.ols('health ~ age', data=df, subset=df['female']==0).fi
print(female0.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 health   R-squared:
Model:                            OLS   Adj. R-squared:
Method:                 Least Squares   F-statistic:
Date:                Tue, 18 Nov 2025   Prob (F-statistic):
Time:                        19:33:02   Log-Likelihood:
No. Observations:                1563   AIC:
Df Residuals:                    1561   BIC:
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025
------------------------------------------------------------------------------
Intercept      1.5060      0.060     25.027      0.000       1.388
age            0.0106      0.001      8.683      0.000       0.008
==============================================================================
Omnibus:                       53.812   Durbin-Watson:
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
Skew:                           0.456   Prob(JB):
Kurtosis:                       2.774   Cond. No.
==============================================================================

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

Next, I ran separate regressions for men (female = 0) and women (female = 1) to see whether the effect of age on health differs by gender. I expected the slope for women to be larger because prior research shows that women often report health problems more frequently than men, especially as they age.

The results only partly match this expectation. The age coefficient is slightly larger among men (0.0106) than among women (0.0091), but the difference is small. This suggests that age is associated with worse self-rated health in both groups, and the strength of this relationship is fairly similar for men and women.

Question #3- Now, run a full model, with an interaction term added to that model, so you can test for whether the earlier slope differences might be statistically significantly different. What did you find from a statistical standpoint?

```
health_int = smf.ols('health ~ age * female', data=df).fit()
print(health_int.summary())
```

```
                            OLS Regression Results
================================================================================
Dep. Variable:                 health   R-squared:
Model:                            OLS   Adj. R-squared:
Method:                 Least Squares   F-statistic:
Date:                Tue, 18 Nov 2025   Prob (F-statistic):
Time:                        19:40:46   Log-Likelihood:
No. Observations:                3504   AIC:
Df Residuals:                    3500   BIC:
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                 coef    std err          t      P>|t|      [0.025
--------------------------------------------------------------------------------
Intercept      1.5060      0.062     24.169      0.000       1.384
age            0.0106      0.001      8.385      0.000       0.008
female         0.1196      0.083      1.437      0.151      -0.044
age:female    -0.0016      0.002     -0.943      0.346      -0.005
================================================================================
Omnibus:                      133.712   Durbin-Watson:
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
Skew:                           0.469   Prob(JB):
Kurtosis:                       2.708   Cond. No.
================================================================================

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

To test whether the difference in slopes is statistically significant, I included an interaction term between age and female. The interaction coefficient was −0.0016 with a p-value of 0.346. Because this coefficient is not statistically significant, there is no evidence that the effect of age on self-rated health differs between men and women.

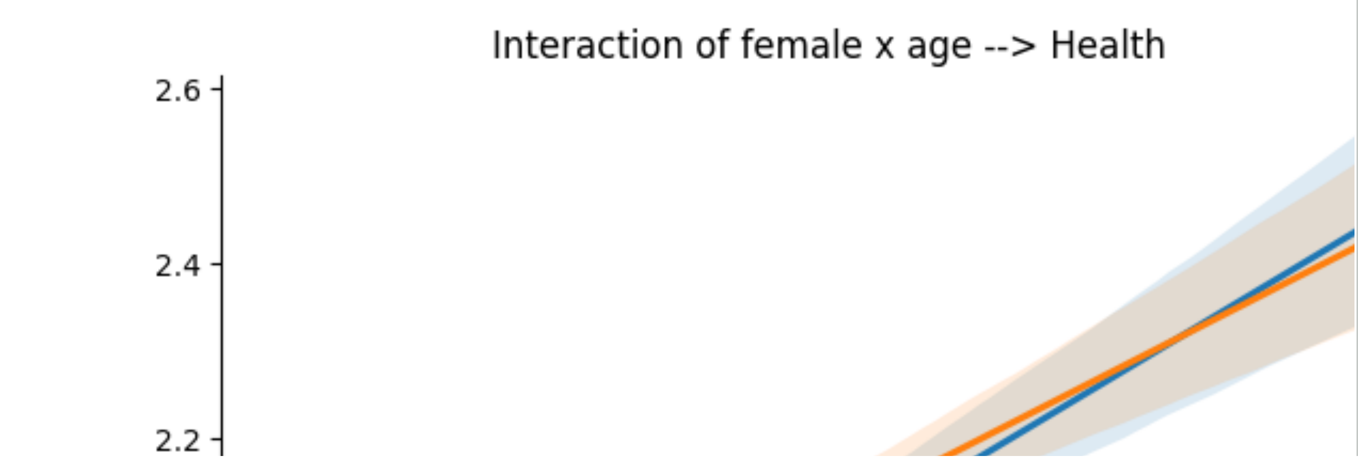
In other words, although the separate subgroup regressions showed slightly different slopes for men and women, these differences are not statistically meaningful from a formal standpoint. Gender does not significantly moderate the relationship between age and health in this sample.

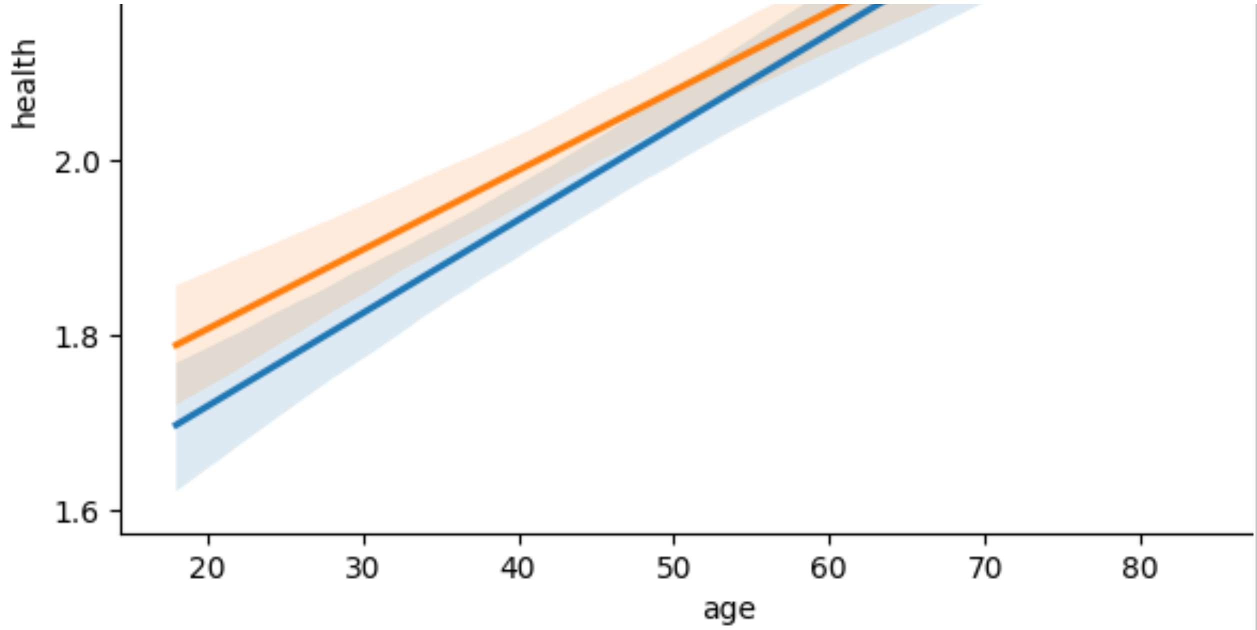
Question #4- Plot the relationship found in the interaction.

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

sns.lmplot(
    x='age',
    y='health',
    hue='female',
    data=df,
    scatter=False,
    height=5,
    aspect=1.3
)

plt.title('Interaction of female x age --> Health')
plt.show()
```





The plot shows two regression lines: one for women (female = 1) and one for men (female = 0). The lines are almost parallel, which visually confirms the statistical finding from the interaction model: the age × female interaction is not significant. This suggests that age affects self-rated health similarly for both men and women.

In conclusion, age is a clear predictor of self-rated health—older respondents report worse health. Women also report slightly poorer health than men. Although the slopes differed somewhat in the subgroup models, the interaction test showed that these differences are not statistically significant. Overall, age matters more than gender in predicting health in this sample.