In [1]:
```
import pandas as pd
import numpy as np
```

In [3]:
```
data=pd.read_csv("C:/Users/USER/Desktop/Kisii University 2023 students projects/Pauline group/survey lung cance
```

In [20]:
```
data.head
```

Out[20]:
```
<bound method NDFrame.head of     GENDER  AGE  SMOKING  YELLOW_FINGERS  ANXIETY  PEER_PRESSURE  \
0        M   69        1               2        2              1
1        M   74        2               1        1              1
2        F   59        1               1        1              2
3        M   63        2               2        2              1
4        F   63        1               2        1              1
..     ...  ...      ...             ...      ...            ...
304      F   56        1               1        1              2
305      M   70        2               1        1              1
306      M   58        2               1        1              1
307      M   67        2               1        2              1
308      M   62        1               1        1              2

     CHRONIC DISEASE  FATIGUE   ALLERGY   WHEEZING  ALCOHOL CONSUMING  \
0                  1         2         1         2                  2
1                  2         2         2         1                  1
2                  1         2         1         2                  1
3                  1         1         1         1                  2
4                  1         1         1         2                  1
..               ...       ...       ...       ...                ...
304                2         2         1         1                  2
305                1         2         2         2                  2
306                1         1         2         2                  2
307                1         2         2         1                  2
308                1         2         2         2                  2

     COUGHING  SHORTNESS OF BREATH  SWALLOWING DIFFICULTY  CHEST PAIN  \
0           2                    2                      2           2
1           1                    2                      2           2
2           2                    2                      1           2
3           1                    1                      2           2
4           2                    2                      1           1
..        ...                  ...                    ...         ...
304         2                    2                      2           1
305         2                    2                      1           2
306         2                    1                      1           2
307         2                    2                      1           2
308         1                    1                      2           1

     LUNG_CANCER
0              1
1              1
2              0
3              0
4              0
..           ...
304            1
305            1
306            1
307            1
308            1

[309 rows x 16 columns]>
```

In [6]:
```
print(data.columns)
```

```
Index(['GENDER', 'AGE', 'SMOKING', 'YELLOW_FINGERS', 'ANXIETY',
       'PEER_PRESSURE', 'CHRONIC DISEASE', 'FATIGUE ', 'ALLERGY ', 'WHEEZING',
       'ALCOHOL CONSUMING', 'COUGHING', 'SHORTNESS OF BREATH',
       'SWALLOWING DIFFICULTY', 'CHEST PAIN', 'LUNG_CANCER'],
      dtype='object')
```

In [7]:
```
# Define the independent variables
X_diet = data[['AGE', 'SMOKING', 'YELLOW_FINGERS', 'ANXIETY', 'PEER_PRESSURE', 'CHRONIC DISEASE', 'FATIGUE ', '

# Define the dependent variable
y_cardio = data['LUNG_CANCER']
```

In [13]:
```
from sklearn.preprocessing import LabelEncoder

# Initialize label encoder
label_encoder = LabelEncoder()

# Encode the 'LUNG_CANCER' column
data['LUNG_CANCER'] = label_encoder.fit_transform(data['LUNG_CANCER'])
```

In [16]:
```
data.columns
```

Out[16]:
```
Index(['GENDER', 'AGE', 'SMOKING', 'YELLOW_FINGERS', 'ANXIETY',
       'PEER_PRESSURE', 'CHRONIC DISEASE', 'FATIGUE ', 'ALLERGY ', 'WHEEZING',
       'ALCOHOL CONSUMING', 'COUGHING', 'SHORTNESS OF BREATH',
       'SWALLOWING DIFFICULTY', 'CHEST PAIN', 'LUNG_CANCER'],
      dtype='object')
```

Objective 1: Investigate the Relationship Between Diet and Cardiovascular Health

In [18]:
```
import statsmodels.api as sm
import pandas as pd

# Assuming 'data' is your DataFrame with the provided column names
# Define the independent variables
X_diet = data[['AGE', 'SMOKING', 'YELLOW_FINGERS', 'ANXIETY', 'PEER_PRESSURE',
               'CHRONIC DISEASE', 'FATIGUE ', 'ALLERGY ', 'WHEEZING',
               'ALCOHOL CONSUMING', 'COUGHING', 'SHORTNESS OF BREATH',
               'SWALLOWING DIFFICULTY', 'CHEST PAIN']]

# Define the dependent variable
y_cardio = data['LUNG_CANCER']

# Add a constant term to the independent variables
X_diet = sm.add_constant(X_diet)

# Fit the multiple linear regression model for Objective 1
model_obj1 = sm.OLS(y_cardio, X_diet).fit()

# Print the summary of the linear regression model for Objective 1
print(model_obj1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            LUNG_CANCER   R-squared:                       0.400
Model:                            OLS   Adj. R-squared:                  0.371
Method:                 Least Squares   F-statistic:                     13.98
Date:                Tue, 05 Dec 2023   Prob (F-statistic):           1.83e-25
Time:                        22:59:20   Log-Likelihood:                -18.971
No. Observations:                 309   AIC:                             67.94
Df Residuals:                     294   BIC:                             123.9
Df Model:                          14
Covariance Type:            nonrobust
============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                     -1.2119      0.194     -6.258      0.000      -1.593      -0.831
AGE                        0.0014      0.002      0.758      0.449      -0.002       0.005
SMOKING                    0.0732      0.032      2.270      0.024       0.010       0.137
YELLOW_FINGERS             0.1197      0.041      2.945      0.003       0.040       0.200
ANXIETY                    0.0768      0.043      1.800      0.073      -0.007       0.161
PEER_PRESSURE              0.0880      0.035      2.485      0.014       0.018       0.158
CHRONIC DISEASE            0.0912      0.032      2.865      0.004       0.029       0.154
FATIGUE                    0.1503      0.039      3.891      0.000       0.074       0.226
ALLERGY                    0.1488      0.033      4.453      0.000       0.083       0.215
WHEEZING                   0.0600      0.035      1.714      0.088      -0.009       0.129
ALCOHOL CONSUMING          0.1945      0.038      5.125      0.000       0.120       0.269
COUGHING                   0.1058      0.038      2.791      0.006       0.031       0.180
SHORTNESS OF BREATH        0.0445      0.039      1.138      0.256      -0.032       0.121
SWALLOWING DIFFICULTY      0.0997      0.038      2.609      0.010       0.024       0.175
CHEST PAIN                 0.0279      0.033      0.836      0.404      -0.038       0.093
==============================================================================
Omnibus:                       30.983   Durbin-Watson:                   1.742
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               39.039
Skew:                          -0.743   Prob(JB):                     3.33e-09
Kurtosis:                       3.907   Cond. No.                         826.
==============================================================================

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

Objective 2: Examine the Influence of Physical Activity on Cardiovascular Health

In [19]:
```
import statsmodels.api as sm

# Assuming 'data' is your DataFrame with the provided column names
# Define the independent variables for physical activity
X_physical_activity = data[['AGE', 'SMOKING', 'YELLOW_FINGERS', 'ANXIETY', 'PEER_PRESSURE',
                            'CHRONIC DISEASE', 'FATIGUE ', 'ALLERGY ', 'WHEEZING',
                            'ALCOHOL CONSUMING', 'COUGHING', 'SHORTNESS OF BREATH',
                            'SWALLOWING DIFFICULTY', 'CHEST PAIN']]

# Define the dependent variable
y_cardio = data['LUNG_CANCER']

# Add a constant term to the independent variables
X_physical_activity = sm.add_constant(X_physical_activity)

# Fit the multiple linear regression model for Objective 2
model_obj2 = sm.OLS(y_cardio, X_physical_activity).fit()
```

```
# Print the summary of the linear regression model for Objective 2
print(model_obj2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            LUNG_CANCER   R-squared:                       0.400
Model:                            OLS   Adj. R-squared:                  0.371
Method:                 Least Squares   F-statistic:                     13.98
Date:                Tue, 05 Dec 2023   Prob (F-statistic):           1.83e-25
Time:                        23:00:11   Log-Likelihood:                -18.971
No. Observations:                 309   AIC:                             67.94
Df Residuals:                     294   BIC:                             123.9
Df Model:                          14
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           -1.2119      0.194     -6.258      0.000      -1.593      -0.831
AGE                              0.0014      0.002      0.758      0.449      -0.002       0.005
SMOKING                          0.0732      0.032      2.270      0.024       0.010       0.137
YELLOW_FINGERS                   0.1197      0.041      2.945      0.003       0.040       0.200
ANXIETY                          0.0768      0.043      1.800      0.073      -0.007       0.161
PEER_PRESSURE                    0.0880      0.035      2.485      0.014       0.018       0.158
CHRONIC DISEASE                  0.0912      0.032      2.865      0.004       0.029       0.154
FATIGUE                          0.1503      0.039      3.891      0.000       0.074       0.226
ALLERGY                          0.1488      0.033      4.453      0.000       0.083       0.215
WHEEZING                         0.0600      0.035      1.714      0.088      -0.009       0.129
ALCOHOL CONSUMING                0.1945      0.038      5.125      0.000       0.120       0.269
COUGHING                         0.1058      0.038      2.791      0.006       0.031       0.180
SHORTNESS OF BREATH              0.0445      0.039      1.138      0.256      -0.032       0.121
SWALLOWING DIFFICULTY            0.0997      0.038      2.609      0.010       0.024       0.175
CHEST PAIN                       0.0279      0.033      0.836      0.404      -0.038       0.093
==============================================================================
Omnibus:                       30.983   Durbin-Watson:                   1.742
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               39.039
Skew:                          -0.743   Prob(JB):                     3.33e-09
Kurtosis:                       3.907   Cond. No.                         826.
==============================================================================

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

Objective 3: Assess the Impact of Stress Levels on Cardiovascular Health

In [21]:

```
import statsmodels.api as sm

# Assuming 'data' is your DataFrame with the provided column names
# Define the independent variables for stress levels
X_stress = data[['AGE', 'SMOKING', 'YELLOW_FINGERS', 'ANXIETY', 'PEER_PRESSURE',
                 'CHRONIC DISEASE', 'FATIGUE ', 'ALLERGY ', 'WHEEZING',
                 'ALCOHOL CONSUMING', 'COUGHING', 'SHORTNESS OF BREATH',
                 'SWALLOWING DIFFICULTY', 'CHEST PAIN']]

# Define the dependent variable
y_cardio = data['LUNG_CANCER']

# Add a constant term to the independent variables
X_stress = sm.add_constant(X_stress)

# Fit the multiple linear regression model for Objective 3
model_obj3 = sm.OLS(y_cardio, X_stress).fit()

# Print the summary of the linear regression model for Objective 3
print(model_obj3.summary())
```

OLS Regression Results

Dep. Variable:	LUNG_CANCER	R-squared:	0.400
Model:	OLS	Adj. R-squared:	0.371
Method:	Least Squares	F-statistic:	13.98
Date:	Tue, 05 Dec 2023	Prob (F-statistic):	1.83e-25
Time:	23:06:40	Log-Likelihood:	-18.971
No. Observations:	309	AIC:	67.94
Df Residuals:	294	BIC:	123.9
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.2119	0.194	-6.258	0.000	-1.593	-0.831
AGE	0.0014	0.002	0.758	0.449	-0.002	0.005
SMOKING	0.0732	0.032	2.270	0.024	0.010	0.137
YELLOW_FINGERS	0.1197	0.041	2.945	0.003	0.040	0.200
ANXIETY	0.0768	0.043	1.800	0.073	-0.007	0.161
PEER_PRESSURE	0.0880	0.035	2.485	0.014	0.018	0.158
CHRONIC DISEASE	0.0912	0.032	2.865	0.004	0.029	0.154
FATIGUE	0.1503	0.039	3.891	0.000	0.074	0.226
ALLERGY	0.1488	0.033	4.453	0.000	0.083	0.215
WHEEZING	0.0600	0.035	1.714	0.088	-0.009	0.129
ALCOHOL CONSUMING	0.1945	0.038	5.125	0.000	0.120	0.269
COUGHING	0.1058	0.038	2.791	0.006	0.031	0.180
SHORTNESS OF BREATH	0.0445	0.039	1.138	0.256	-0.032	0.121
SWALLOWING DIFFICULTY	0.0997	0.038	2.609	0.010	0.024	0.175
CHEST PAIN	0.0279	0.033	0.836	0.404	-0.038	0.093

Omnibus:	30.983	Durbin-Watson:	1.742
Prob(Omnibus):	0.000	Jarque-Bera (JB):	39.039
Skew:	-0.743	Prob(JB):	3.33e-09
Kurtosis:	3.907	Cond. No.	826.

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