In [1]:

```
import pandas as pd
import statsmodels.api as sm

# Reading the data which was saved to an excel sheet after pre-processing
df_cleaned = pd.read_excel('finally_clean_data_for_plotting.xlsx')

# Remove the 99 category (="Unknown") from the data, since it does not benefit the ar
df_cleaned = df_cleaned[~(df_cleaned['noofwines'] == 99)]
df_cleaned = df_cleaned[~(df_cleaned['howoftenwine'] == 99)]

print(df_cleaned)
```

```
       age  sex  householdincome  howoftenwine  noofwines  \
0       34    2               12            10          1
1       84    2                7             6          1
2       29    2               13            10          1
3       68    2                6             5          1
4       54    2               11             9          1
...    ...  ...              ...           ...        ...
14556   18    2                1             9          1
14557   18    1                1            10          1
14558   51    1                6             6          1
14559   21    1                1            10          2
14560   18    2                1            10          1

                      wine_frequency              wine_amount  \
0      1 or 2 times in the last year     One glass/ container
1                2 to 3 times a month     One glass/ container
2      1 or 2 times in the last year     One glass/ container
3                          Once a week     One glass/ container
4      3 to 6 times in the last year     One glass/ container
...                              ...                      ...
14556  3 to 6 times in the last year     One glass/ container
14557  1 or 2 times in the last year     One glass/ container
14558            2 to 3 times a month     One glass/ container
14559  1 or 2 times in the last year  Two glasses/ containers
14560  1 or 2 times in the last year     One glass/ container

          income_category
0      $50,000 to $59,999
1      $20,000 to $24,999
2      $60,000 to $69,999
3      $15,000 to $19,999
4      $40,000 to $49,999
...                   ...
14556    Less than $5,000
14557    Less than $5,000
14558  $15,000 to $19,999
14559    Less than $5,000
14560    Less than $5,000

[14522 rows x 8 columns]
```

I will again focus on the data, I prepared for my research question. I chose the household income variable to do the categorization. As per definition of the assignment description, I will put one of the categories in the household income category to 0.

In [2]:
```
# Replace one category with 0, as per definition of the assignment description: "Less
df_cleaned['householdincome'] = df_cleaned['householdincome'].replace(1, 0)
```

Now, the dependent and independent variables are defined. Then proceed, like in the example.

In [3]:
```
# Define the dependent variable (Y). I chose the number of consumed wine per occasion
Y = df_cleaned['noofwines']

# Define the independent variables (X). The others are age, sex and household income
X = df_cleaned[['householdincome']]
```

In [4]:
```
# Fit the regression model, using the
model = sm.OLS(Y, X).fit()

# Print the summary of the regression
print(model.summary())
```

```
                                 OLS Regression Results
================================================================================
==
Dep. Variable:              noofwines   R-squared (uncentered):                   0.6
67
Model:                            OLS   Adj. R-squared (uncentered):              0.6
67
Method:                 Least Squares   F-statistic:                          2.911e
+04
Date:                Fri, 28 Mar 2025   Prob (F-statistic):
0.00
Time:                        08:15:12   Log-Likelihood:                         -2019
6.
No. Observations:               14522   AIC:                                  4.039e
+04
Df Residuals:                   14521   BIC:                                  4.040e
+04
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
householdincome     0.1110      0.001    170.615      0.000       0.110       0.112
==============================================================================
Omnibus:                     5857.186   Durbin-Watson:                   1.898
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            49481.335
Skew:                           1.716   Prob(JB):                         0.00
Kurtosis:                      11.367   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a
constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly spec
ified.
```

Interpretation of the model result:

Dependent Variable: Wine Consumption Amount (noofwines): The R-squared of 0.667 indicates that approximately 66.7% of the variance in the noofwines variable is explained by the independent variables in the model. This is a high value indicating a good fit.

The F-statistic value of 2.911e+04, which is a high value, suggests that at least one of the predictors is significantly related to the dependent variable.

The p-value Prob (F-statistic) = 0.00 is low and indicates that the model is statistically significant.

Summing these findings up, we can state that the predictor variable is significantly related to the dependent variable.

The coefficient for household income is 0.1110, indicating that for each unit increase in household income, the number of wines consumed increases by approximately 0.1110, holding other variables constant. As for all the other variables, this effect is statistically significant since the p-value = 0.000.

Conclusion:

The model indicates that the household income is a significant predictor of the number of wines consumed. The high R-squared value suggests that the model explain much of the variability in the number of wines consumed.