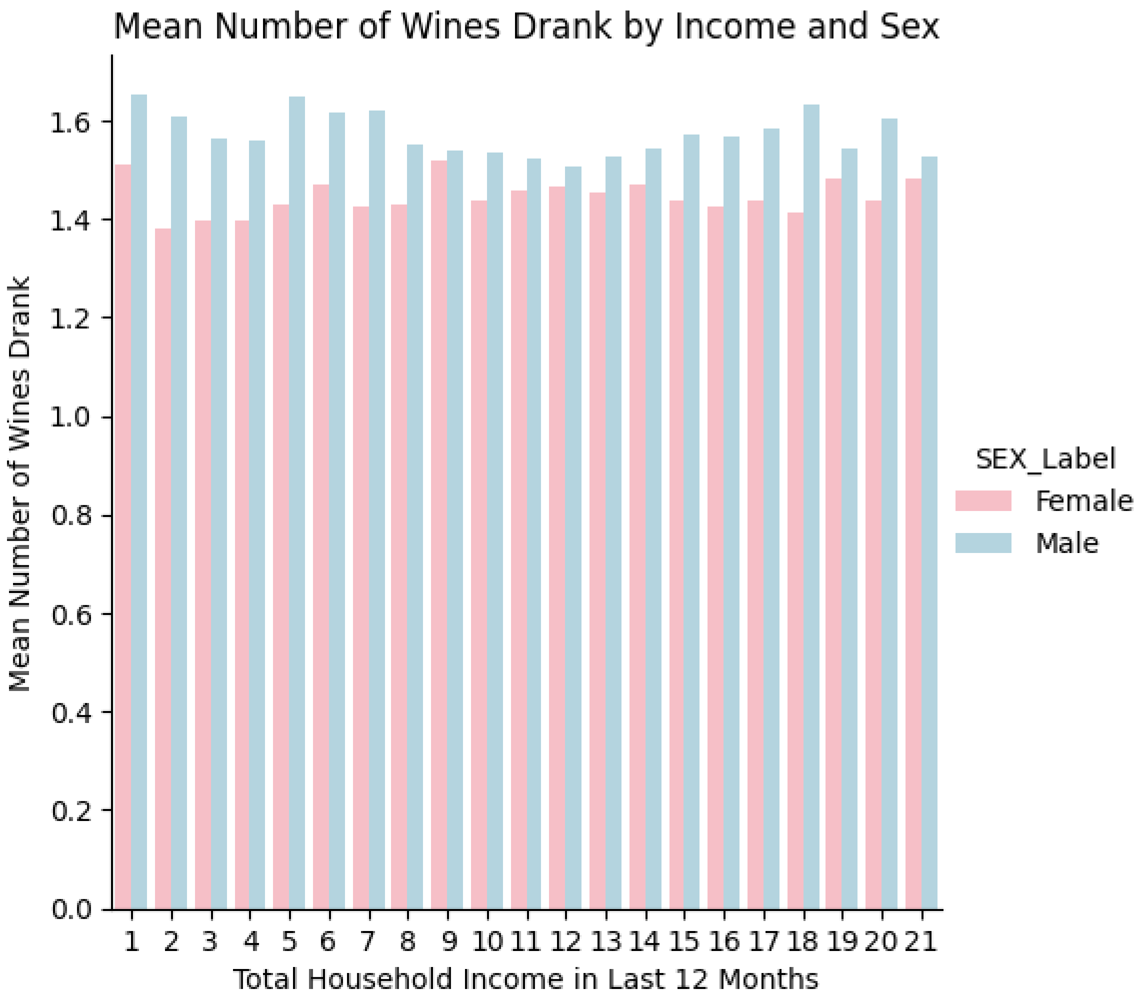
In [24]:

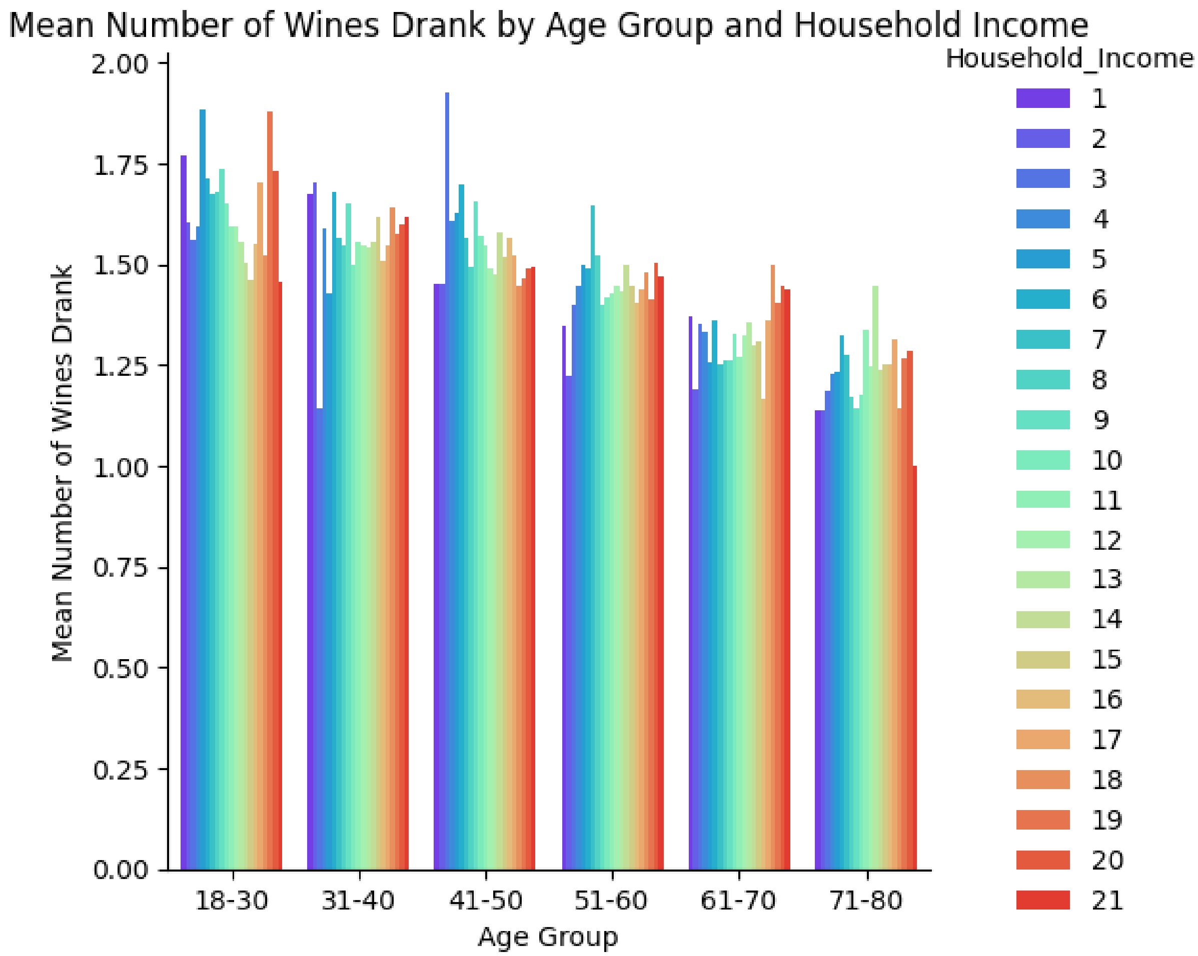
```
import numpy as np
import pandas as pd
import statsmodels.api as sm
import seaborn as sns
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt


#######################
# Data Pre-Processing #
#######################

# Bug fix for display formats to avoid run time errors
pd.set_option('display.float_format', lambda x:'%.2f'%x)

# read the data as pandas data frame
data = pd.read_csv('nesarc_pds.csv', low_memory=False)

# Create a list of the keys I would like to keep
keys_to_keep = [
    "S1Q12B",  # TOTAL HOUSEHOLD INCOME IN LAST 12 MONTHS: CAT
    "AGE",     # Age
    "SEX",     # Sex
    "S2AQ6B",  # Frequency of wine consumption in last 12 months
    "S2AQ6D"   # Number of wines drank when drinking in last 12 months
]

# Filter the data frame to keep only the specified columns
data = data[keys_to_keep]

# Setting variables to numeric
data['S1Q12B'] = pd.to_numeric(data['S1Q12B'], errors='coerce') # Household Income
data['S2AQ6B'] = pd.to_numeric(data['S2AQ6B'], errors='coerce') # Wine drinking frequ
data['S2AQ6D'] = pd.to_numeric(data['S2AQ6D'], errors='coerce') # Wine drinking amour
data['AGE'] = pd.to_numeric(data['AGE'], errors='coerce') # Age
data['SEX'] = pd.to_numeric(data['SEX'], errors='coerce') # Sex

# Replace all '99' values by 'NaN', since they do not benefit the evaluation
data['S2AQ6B'] = data['S2AQ6B'].replace(99, np.nan)
data['S2AQ6D'] = data['S2AQ6D'].replace(99, np.nan)

# Drop rows where there are NaN values in columns 'S2AQ6B' or 'S2AQ6D'
data = data.dropna(subset=['S2AQ6B', 'S2AQ6D'])

# Renaming of the columns for more obvious interpretation
data = data.rename(columns={'S1Q12B': 'Household_Income', 'S2AQ6B': 'Winefrequency',


#######################
# Regression Analysis #
#######################

# OLS Regression for Wineamount ~ Household_Income + AGE + SEX
regression_model = smf.ols('Wineamount ~ Household_Income + AGE + SEX', data=data).fi

# Print the result
print(regression_model.summary())
```

```python
##############################################
# Plotting of data for better interpretation #
##############################################

# Listwise deletion for calculating means for regression model observations
data_cleaned = data[['Wineamount', 'Household_Income', 'AGE', 'SEX']].dropna()

# Group means and standard deviations
mean_values = data_cleaned.groupby(['Household_Income', 'SEX']).mean()['Wineamount']
std_values = data_cleaned.groupby(['Household_Income', 'SEX']).std()['Wineamount']

# Map numeric values to labels for the 'SEX' variable
data_cleaned['SEX_Label'] = data_cleaned['SEX'].map({1: 'Male', 2: 'Female'})

# Define a custom color palette for the SEX variable
custom_palette = {'Male': (173/255, 216/255, 230/255, 1),  # Light Blue
                  'Female': (255/255, 182/255, 193/255, 1)}  # Light Red

# Bivariate bar graph for Household_Income and Wineamount
sns.catplot(x = "Household_Income", y = "Wineamount", hue = "SEX_Label", data = data_
plt.xlabel('Total Household Income in Last 12 Months')
plt.ylabel('Mean Number of Wines Drank')
plt.title('Mean Number of Wines Drank by Income and Sex')
plt.show()

# Create age bins
data_cleaned['AGE_Bin'] = pd.cut(data_cleaned['AGE'], bins = [18, 30, 40, 50, 60, 70,

# Create a rainbow color palette based on the number of unique Household Income categ
rainbow_palette = sns.color_palette("rainbow", n_colors=data_cleaned['Household_Incor

# Bivariate bar graph for AGE_Bin and Wineamount, colored by Household_Income
sns.catplot(x = "AGE_Bin", y = "Wineamount", hue = "Household_Income", data = data_cl
plt.xlabel('Age Group')
plt.ylabel('Mean Number of Wines Drank')
plt.title('Mean Number of Wines Drank by Age Group and Household Income')
plt.show()
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Wineamount   R-squared:                       0.030
Model:                            OLS   Adj. R-squared:                  0.029
Method:                 Least Squares   F-statistic:                     147.1
Date:                Tue, 01 Apr 2025   Prob (F-statistic):           6.14e-94
Time:                        13:53:51   Log-Likelihood:                -16759.
No. Observations:               14522   AIC:                         3.353e+04
Df Residuals:                   14518   BIC:                         3.356e+04
Df Model:                           3
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept            2.0207      0.033     60.554      0.000       1.955       2.086
Household_Income    -0.0013      0.001     -0.980      0.327      -0.004       0.001
AGE                 -0.0075      0.000    -19.407      0.000      -0.008      -0.007
SEX                 -0.1083      0.013     -8.333      0.000      -0.134      -0.083
==============================================================================
```

```
Omnibus:                     9018.341   Durbin-Watson:                   2.016
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           164889.819
Skew:                           2.670   Prob(JB):                         0.00
Kurtosis:                      18.620   Cond. No.                         270.
==============================================================================

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





Interpretation of the results:

The R-squared value of the model is 0.030, which indicates that only about 3 % of the variance in 'Wineamount' is explained by the model. Hence, the model might not be a strong predictor of wine consumption based on the included variables.

The F-statistic is 147.1 with a p-value of 6.14e-94 , indicating that the overall model is statistically significant.

Coefficients:

- Household_Income: The coefficient for Household_Income is -0.0013, with a p-value of 0.327. This indicates that there is no statistically significant relationship between household income and the number of wines drank, as the p-value is greater than 0.05.
- AGE: The coefficient for AGE is -0.0075, with a p-value of 0.000. This suggests that for each additional year of age, the number of wines drank decreases by approximately 0.0075, and this relationship is statistically significant.
- SEX: The coefficient for SEX is -0.1083, with a p-value of 0.000. This indicates that being female (assuming SEX is coded as 1 for male and 2 for female) is associated with drinking approximately 0.1083 fewer wines compared to males, and this relationship is statistically

significant.

Group Means and Standard Deviations

Mean Wineamount by Household Income and SEX: The mean values show the average number of wines consumed for each combination of household income and sex. For example, for household income category 1, males (SEX = 1) drink an average of 1.65 wines, while females (SEX = 2) drink an average of 1.51 wines. This pattern continues across different income categories, showing variations in wine consumption based on both income and sex. Concluding these findings, there is a neglectable tendency of wine drinking amount, to change with household income. Further, males seem to drink slightly more in average than females.

Standard Deviation of Wineamount: The standard deviation values indicate the variability in wine consumption within each group. For instance, for household income category 1, males have a standard deviation of 1.19, while females have a lower standard deviation of 0.76, suggesting that male wine consumption is more variable than female consumption in this income category.

Conclusion:

The regression analysis indicates that age and sex are significant predictors of wine consumption, while household income does not appear to have a significant effect.

Assessment of relevance for initial hypotheses:

My initial hypotheses were:

H1: "Drinking of wine increases with increasing income"

H2: "The amount of consumed wine is independent of age"

Based on the model output, H1 cannot be confirmed. Further, H2 can also not be confirmed, since the model output states, that the age is decreasing with increasing age (Coefficient -0.0075 with a P-Value of 0.00, which is smaller than 0.05), which is also visible from the plot.

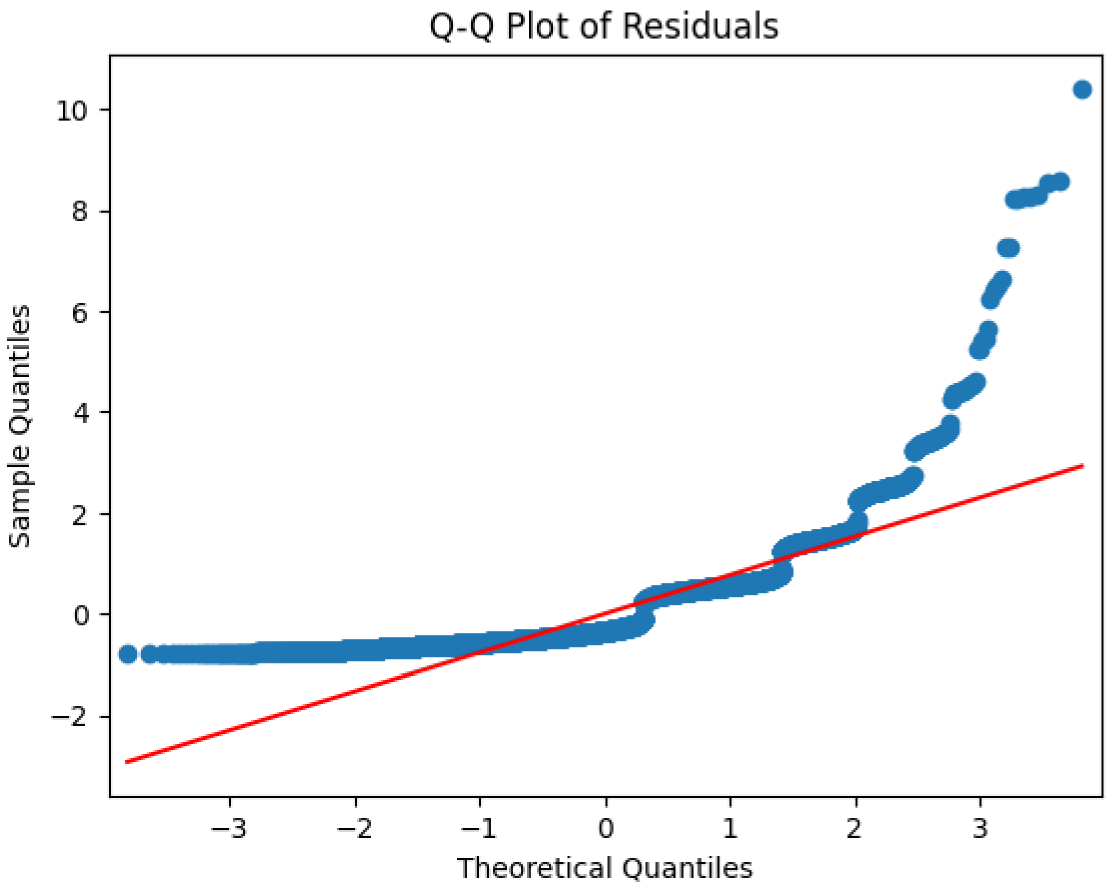
Note: For the interpretation of the household income, please refer to the codebook:

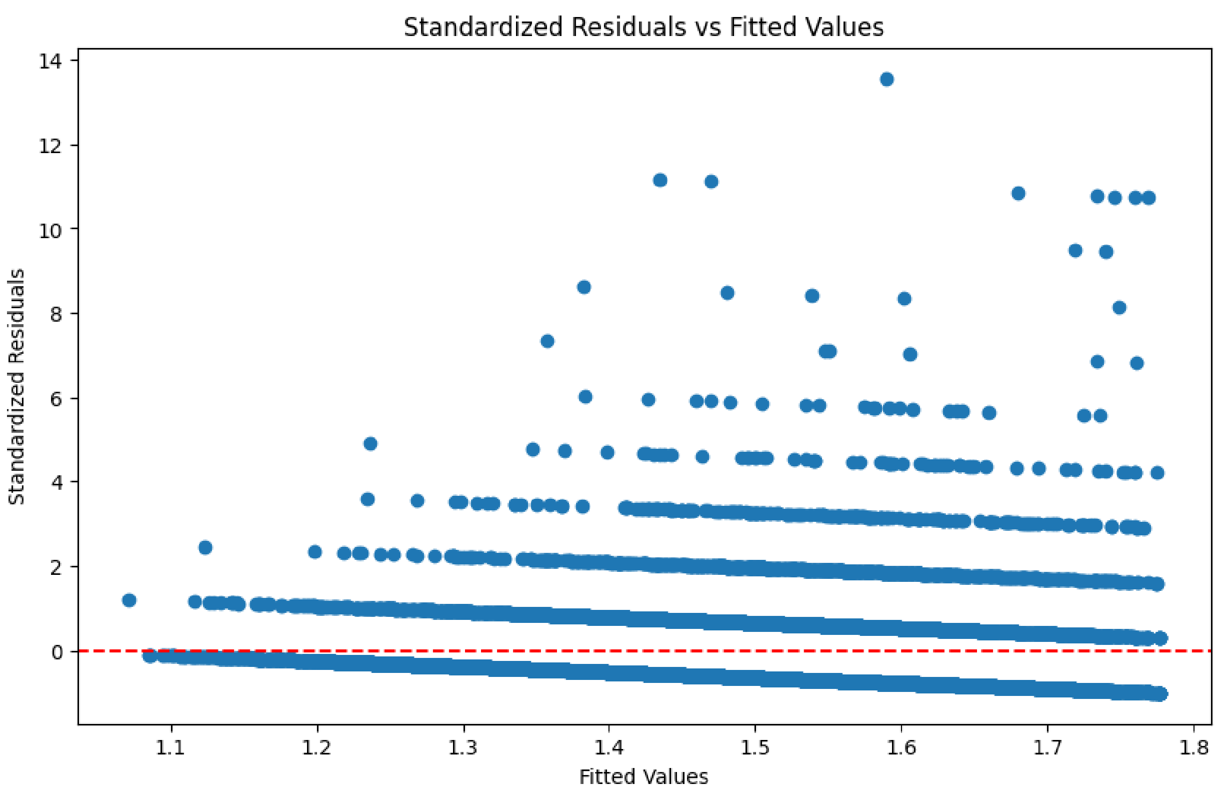
```
1. Less than $5,000
2. $5,000 to $7,999
3. $8,000 to $9,999
4. $10,000 to $12,999
5. $13,000 to $14,999
6. $15,000 to $19,999
7. $20,000 to $24,999
8. $25,000 to $29,999
9. $30,000 to $34,999
10. $35,000 to $39,999
11. $40,000 to $49,999
12. $50,000 to $59,999
13. $60,000 to $69,999
```

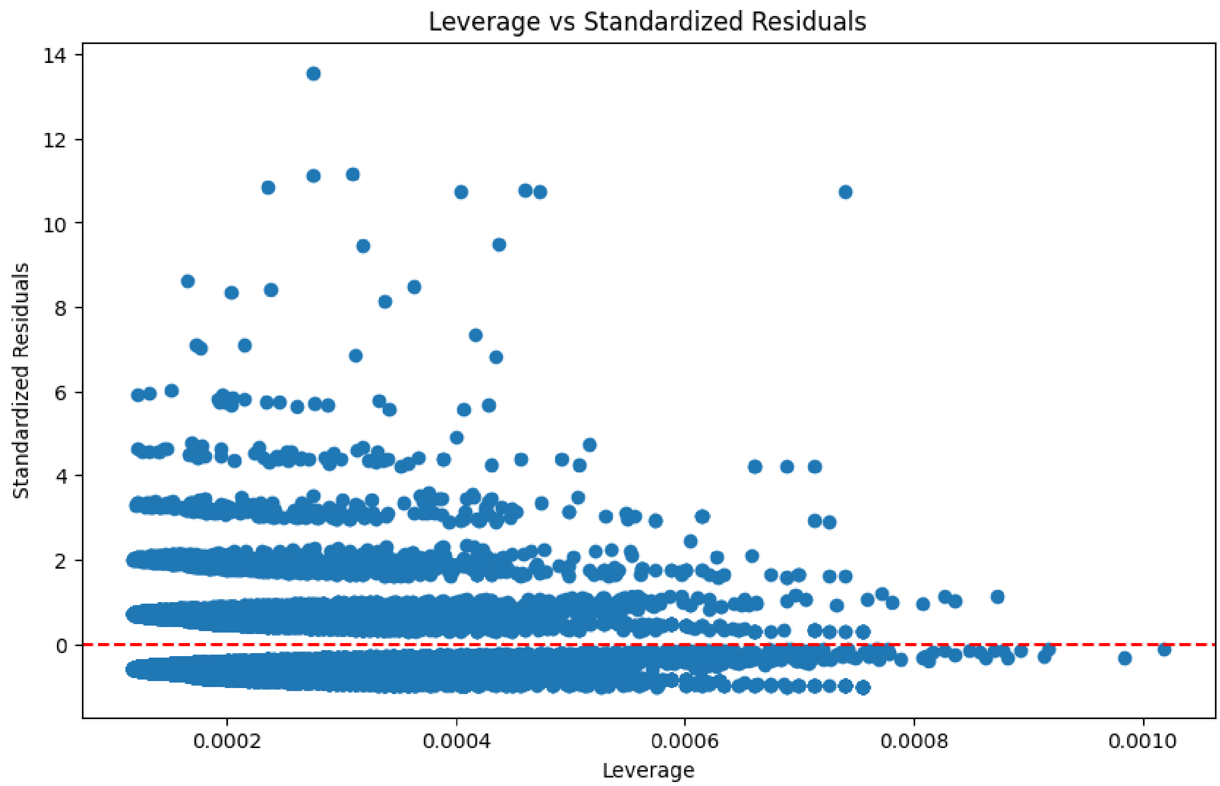
14. $70,000 to $79,999
15. $80,000 to $89,999
16. $90,000 to $99,999
17. $100,000 to $109,999
18. $110,000 to $119,999
19. $120,000 to $149,999
20. $150,000 to 199,999
    21. $200,000 or more

In [25]:

```
# After the regression analysis
#############################################
# Regression Diagnostic Plots #
#############################################

# Q-Q Plot
sm.qqplot(regression_model.resid, line = 's')
plt.title('Q-Q Plot of Residuals')
plt.show()

# Standardized Residuals Plot
standardized_residuals = regression_model.get_influence().resid_studentized_internal
fitted_values = regression_model.fittedvalues

plt.figure(figsize = (10, 6))
plt.scatter(fitted_values, standardized_residuals)
plt.axhline(0, color = 'red', linestyle = '--')
plt.xlabel('Fitted Values')
plt.ylabel('Standardized Residuals')
plt.title('Standardized Residuals vs Fitted Values')
plt.show()

# Leverage Plot
influence = regression_model.get_influence()
leverage = influence.hat_matrix_diag

plt.figure(figsize = (10, 6))
plt.scatter(leverage, standardized_residuals)
plt.axhline(0, color = 'red', linestyle = '--')
plt.xlabel('Leverage')
plt.ylabel('Standardized Residuals')
plt.title('Leverage vs Standardized Residuals')
plt.show()
```


Q-Q Plot of Residuals
Sample Quantiles
Theoretical Quantiles
10
8
6
4
2
0
−2
−3
−2
−1
0
1
2
3


Standardized Residuals vs Fitted Values
Standardized Residuals
Fitted Values
14
12
10
8
6
4
2
0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8



Interpretation of the q-q plot, standardized residuals and leverage plot:

- q-q-plot:
  - While the model fits well for a central range of values (between -1 and 2), it struggles with extreme values on both ends (negative and positive). This could indicate potential issues with the model's assumptions, such as non-normality of residuals or outliers/ influential points.
- standardized residuals:
  - In a well-fitted linear regression model, one would expect the residuals to be randomly scattered around zero without any patterns. The fact that they are forming parallel lines suggests that as the fitted values increase, the residuals tend to decrease in a systematic way. Again, outliers might be a problem here.
- leverage plot:
  - A lot of points clustered together at the lower end of the x-axis can be observed, which means that there are many data points with low values of x. This can be explained by the small range of data and the high amount of data for low values.