Lab 1 (QMSS5062 Social network analysis)

Submitted by: Gideon Tay
My UNI: gt2528

Overview: In this lab, we will explore how racial diversity and homophily in a person's social network may be associated with their attitudes on laws against inter-racial marriages.

Import all necessary libraries for this lab

In [1]:
```python
# Libraries for data analysis
import pandas as pd # also used to load data but primarily used in analysis
import numpy as np
import statsmodels.formula.api as smf
import statsmodels.api as sm
import matplotlib.pyplot as plt

# Libraries to load in data
import requests
import zipfile
import io
from tqdm.notebook import tqdm
```

Load in General Social Survey (GSS) data

We load in data for the following variables: 'id', 'racmar', 'year', 'race', 'race1','race2', 'race3', 'race4', 'race5', 'raclive', 'age', 'educ', 'polviews'. We only loaded the numeric labels for these variables. Information on these variables would be provided later. You may view the GSS codebook here.

In [2]:
```python
# Step 1: Download the ZIP file with progress bar
url = 'https://gss.norc.org/content/dam/gss/get-the-data/documents/stata/GSS_stata.zip'

# 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 total file size
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()
```

```python
    # 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
        with z.open(stata_file) as stata_file_stream:
            # Step 3: Load only the selected columns into a pandas DataFrame with numeric labels
            columns_to_load = [
                'id', 'racmar', 'year', 'race', 'race1','race2', 'race3',
                'race4', 'race5', 'raclive', 'age', 'educ', 'polviews']
            df = pd.read_stata(stata_file_stream, columns=columns_to_load, convert_categoricals=False
            print("Data loaded successfully!")

# Step 4: Display the first few rows of the final DataFrame
df.head()
```

100% ███████████████████████████████ 42.0M/42.0M [00:03<00:00, 12.3MiB/s]

```
Download completed!
Data loaded successfully!
```

Out[2]:

	id	racmar	year	race	race1	race2	race3	race4	race5	raclive	age	educ	polviews
0	1	2.0	1972	1.0	NaN	NaN	NaN	NaN	NaN	1.0	23.0	16.0	NaN
1	2	1.0	1972	1.0	NaN	NaN	NaN	NaN	NaN	1.0	70.0	10.0	NaN
2	3	1.0	1972	1.0	NaN	NaN	NaN	NaN	NaN	1.0	48.0	12.0	NaN
3	4	1.0	1972	1.0	NaN	NaN	NaN	NaN	NaN	1.0	27.0	17.0	NaN
4	5	2.0	1972	1.0	NaN	NaN	NaN	NaN	NaN	1.0	61.0	12.0	NaN

1- Develop a hypothesis about how some ego-network measure (e.g., degree/size, density, diversity, average-level of alters, homophily, structural holes, or brokerage) may be related to some other variable of interest.

I am interested in how the characteristics of people's social circles affect their attitudes.

- I hypothesize that people with more racial diversity in their social circle will tend to believe that there should not be laws against marriages between blacks and whites.
- I hypothesize that people with lower racial homophily in their social circle will tend to believe that there should not be laws against marriages between blacks and whites.

2- Explain why you think these two variables should be related.

People with more racial diversity in their social circle have close relationships with people of a variety of races. Since they socialize across races, they are probably more accepting of interracial marriages. At the very least, they are likely to believe that there should not be laws against marriages between blacks and whites.

This is a rather low bar to pass in terms of acceptance toward interracial marriage. After all, laws against interracial marriage seem relatively extreme and could be viewed as government over-reach, even if people are uncomfortable with close relatives marrying someone of another race.

Moreover, if someone's social circle is racially diverse, it is likely that the person has more exposure to people of other races in their daily lives (e.g. in the neighbourhood). With greater exposure to people of other races, people may have a better understanding and acceptance of other races socially and culturally. This in turn may lead to greater acceptance of interracial marriage and thus opposition against laws that ban them.

It is good to know whether a racially diverse social circle contributes to opposition against anti-interracial marriage laws beyond the correlation between racially diverse social circles and exposure to people of other races.

A similar argument holds for racial homophily. People with low racial homophily in their social circle have close relationships with people of other races, though not necessarily a variety of other races. It could simply be close relationships with many people of a single, different race. Nevertheless, this should also reflect greater acceptance of other races and thus greater acceptance of interracial marriages and opposition against anti-interracial marriage laws.

3- Tell me about your variables. What is your dependent variable? What are your independent variables? How are they coded? How are they recoded? How are the calculated, if appropriate?

Dependent variable `rracmar` :

- Original variable: `racmar`

  - > Do you think there should be laws against marriages between (Negroes/Blacks/African-Americans) and whites?

  - Possible responses: Yes (1), No (2)
  - Missing values: Don't know (8), No answer (9), Not applicable (0)
- In 1972-1977, `racmar` was asked of non-blacks only. On the 1978-1996 surveys those questions were asked of all respondents. As I will show later, we will only analyze data for the survey year 1985. As such, `racmar` was asked of all respondents (source).

- For ease of analysis, we recode `racmar` into `rracmar` , a binary variable where Yes (1) and No (0). Missing values are coded as NaN.

Let's code this:

```
In [3]:   # Recode rracmar from racmar
          df['rracmar'] = df['racmar'].replace({
              1: 1, 2: 0, 0: np.nan, 8: np.nan, 9: np.nan
              })
```

Independent variable `rhhi`

- Our key variable of interest is racial diversity in a person's social circle.

- In the 1985 and 2005 social networks topical module, respondents were asked:

  - > From time to time, most people discuss important matters with other people. Looking back over the last six months - who are the people with whom you discussed matters important to you? Just tell me their first names or initials. IF LESS THAN 5 NAMES MENTIONED, PROBE: Anyone else? ONLY RECORD FIRST 5 NAMES

- The variables `race1` , `race2` , `race3` , `race4` , `race5` capture the race of the (up to) 5 people mentioned and recorded.

  - Possible responses: Asian (1), Black (2), Hispanic (3), White (4), Other (5)
  - Missing values: Don't know (8), No answer (9), Not applicable (0)
- The Herfindahl–Hirschman Index (HHI) can be used as a measure racial diversity. While initially designed as an indicator to measure firm concentration in an industry, it is a good measure for diversity too:

- $$HHI = \sum_{i=0}^{N} p_i^2$$
  - where p_i the proportion of individuals from race i in the social circle, and n is the number of different races represented
  - Its value ranges from 0 to 1
- High HHI values indicate high concentration of one large group, while small HHI values indicates a more equitable distribution of somewhat smaller groups. Hence, smaller HHI reflects greater diversity.

- We create `rhhi`, a variable ranging from 0 to 1, where higher values reflect greater racial diversity in a person's social circle

  - `rhhi` $= 1 - HHI$

Let's code this:

```python
In [4]:  # List of race columns
         race_cols = ['race1', 'race2', 'race3', 'race4', 'race5']

         # Replace missing values with NaN for easier processing
         df[race_cols] = df[race_cols].replace({0: np.nan, 8: np.nan, 9: np.nan})

         # Function to calculate (rhhi = 1 - HHI) for each respondent
         def calculate_rhhi(row):
             # Drop NaN values and get the count of each unique race
             race_counts = row.dropna().value_counts(normalize=True)

             # If all race1 to race5 values are NaN, return NaN
             if race_counts.empty:
                 return np.nan

             # Calculate the HHI: sum of squared proportions of each race
             hhi = (race_counts ** 2).sum()

             return 1 - hhi

         # Apply the function to each row and create a new column 'HHI' for diversity
         df['rhhi'] = df[race_cols].apply(calculate_rhhi, axis=1)
```

Alternate independent variable `racehomo`

- Homophily is a concept related to but different from diversity. We shall code a variable capturing racial homophily in a person's social circle which we will use later in *Question 5*.

- Homophily refers to ego's tendency to have alters who share ego's attributes. In this case, we want to measure to what extent a person's social circle only includes people of the same race.

- Unlike our diversity measure which increases with a greater variety of races in the social circle, homophily only cares about whether the race of others in the social circle are the same as the ego's.

  - For example, the homophily measure of a white person with 2 whites, 1 black, and 1 asian in his social circle is identical to another white person with 2 white and 2 blacks in his social circle.
  - However, the racial diversity of the first person's social circle is higher than the second person's.
- The variable `race` captures the respondent's race

  - Possible responses: White (1), Black (2), Other (3)
  - Unlike `race1` to `race5`, asians and hispanics are grouped into 'other' in this variable. Hence, we can only calculate racial homophilly for whites and blacks.
- We measure racial homophily `racehomo` with the Krackhardt and Stern E-I Index.

  - *racehomo = (E - I) / (E + I)*

- Where *E* (external) is number of ties to people of a different race, and *I* (internal) is number of ties to people of same race
  - Varies between -1 (homophily) and +1 (heterophily)
- As explained above, we will only calculate `racehomo` for blacks and whites. `racehomo` is NaN for other rows.

Let's code this:

In [5]:

```python
# Recode respondent's race to match the social contacts' race coding
df['rrace'] = df['race'].replace({1: 4, 2: 2})  # White→4, Black→2

# Compute in-group (I) and out-group (E) counts only for respondents who are White (1) or Black (2)
df['I'] = df[race_cols].apply(lambda row: sum(row.dropna() == df.at[row.name, 'rrace'])
                              if df.at[row.name, 'race'] in [1, 2] else np.nan, axis=1)
df['E'] = df[race_cols].apply(lambda row: sum(row.dropna().isin([1, 2, 3, 4, 5]) & (row != df.at[row.
                              if df.at[row.name, 'race'] in [1, 2] else np.nan, axis=1)

# Compute the E-I Index only for White and Black respondents
df['racehomo'] = (df['E'] - df['I']) / (df['E'] + df['I'])

# Handle division by zero (E + I = 0)
df['racehomo'] = df['racehomo'].replace([np.inf, -np.inf], np.nan)
```

Other variables to include:

There are other variables that could affect views toward laws against marriages between blacks and whites, besides racial diversity and homophily of a person's social circle. These may be added to our models later:

- `white`

  - Recoded from `race` : the respondent's race
    - Possible responses: White (1), Black (2), Other (3)
  - Recoded values: White (1), Black (0), Other (3)
  - Since other races (3) will be dropped later, this will become a binary variable for white people.
- `rraclive`

  - Recoded from `raclive`

  - Are there any ("whites" for black respondents, "blacks" for non-black respondents) living in this neighborhood now?

    - Possible responses: Yes (1), No (2)
    - Missing values: Don't know (8), No answer (9), Not applicable (0)
  - Recoded values: Yes (1), No (0)
- `age` : respondent's age

  - Possible responses: ranges from 1 to 8, each representing a decade (e.g. 1 represents 10-19 years old). 8 captures 80 or over.
  - Missing values: Don't know, No answer (9)
- `educ` : respondent's education

  - Possible responses: ranges from 0 to 20, higher values indicating higher education levels
  - Missing values: Don't know (98), No answer (99)
- `polviews` :

  - We hear a lot of talk these days about liberals and conservatives. I'm going to show you a seven-point scale on which the political views that people might hold are arranged from

> extremely liberal--point 1--to extremely conservative--point 7. Where would you place yourself on this scale?

- Possible responses: ranges from 1 to 7, from extremely liberal to extremely conservative
- Missing values: Don't know (8), No answer (9), Not applicable (0)

```python
In [6]:  # Other variables to include in regression excluding age
         vars_excl_age = ['race', 'raclive', 'educ', 'polviews']

         # Replace missing values with NaN for easier processing
         df[vars_excl_age] = df[vars_excl_age].replace({
             0: np.nan, 8: np.nan, 9: np.nan, 98: np.nan, 99: np.nan, -1: np.nan
         })
         df['age'] = df['age'].replace({9: np.nan}) # 8 is meaningful for age

         # Recode white from race, white will be a binary as 3 is dropped later
         df['white'] = df['race'].replace({1: 1, 2: 0})

         # Recode rraclive from raclive to a dummy variable
         df['rraclive'] = df['raclive'].replace({1: 1, 2: 0})

         # Define key variables and drop rows with missing values
         key_vars = ['rracmar', 'rhhi', 'racehomo', 'rraclive', 'white', 'educ', 'polviews', 'age', 'year']
         df_clean = df[key_vars].dropna()  # Ensures that all models use the same complete data

         # Display summary of cleaned and recoded data
         df_summary = df_clean.describe(include='all')
         print(df_summary)
```

```
              rracmar         rhhi     racehomo     rraclive        white  \
count    1118.000000  1118.000000  1118.000000  1118.000000  1118.000000
mean        0.231664     0.027283    -0.896064     0.504472     0.915921
std         0.422084     0.104901     0.384299     0.500204     0.277630
min         0.000000     0.000000    -1.000000     0.000000     0.000000
25%         0.000000     0.000000    -1.000000     0.000000     1.000000
50%         0.000000     0.000000    -1.000000     1.000000     1.000000
75%         0.000000     0.000000    -1.000000     1.000000     1.000000
max         1.000000     0.720000     1.000000     1.000000     1.000000

                educ     polviews          age    year
count    1118.000000  1118.000000  1118.000000  1118.0
mean       13.176208     4.161002    42.847048  1985.0
std         2.806229     1.362307    16.472294     0.0
min         1.000000     1.000000    18.000000  1985.0
25%        12.000000     3.000000    29.000000  1985.0
50%        12.000000     4.000000    39.000000  1985.0
75%        15.000000     5.000000    55.000000  1985.0
max        20.000000     7.000000    89.000000  1985.0
```

Notes from descriptive summary

- `rracmar` : About 23.2% (mean = 0.232) of the respondents think that there should be laws against marriages between blacks and whites. Note that this data sample is from 1985, where interracial marriages were much more taboo compared to today.

- `rhhi` : There is rather low racial diversity in social circles on average (mean = 0.027). Up to the 75th percentile of people have social circles with no racial diversity at all (75th percentile = 0). The small variation (sd = 0.105) may hinder our ability to find statistically signficant relationships later.

- `racehomo` : There is high racial homophily in social circles (mean = -0.896). Up to the 75th percentile of people have social circles with no one of another race at all (75th percentile = -1).

- `white` : there are only two races of people in the dataset (unique = 2). Since we could only code `racehomo` for blacks and whites, only they will be analyzed. However, other races are included in the computation of racial diversity of social circles `rhhi` .

- `year` : It is clear that all 1118 observations that will be used for the subsequent models are from surveys taken in 1985 (minimum and maximum `year` is 1985).

4- Present your initial results from your first few models. What do they indicated about your hypothesis?

Model 1: `rhhi`

Since `rracmar` is a binary variable, we will use logistic regression models, fitted with maximum likelihood estimators. We start with a simple model with only `rhhi` and an intercept.

```
In [7]:  # Model 1: rracmar against rhhi
         X = sm.add_constant(df_clean[['rhhi']])   # Add intercept
         logit_model = sm.Logit(df_clean['rracmar'], X).fit()

         print(logit_model.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.536644
         Iterations 6
                         Logit Regression Results
==============================================================================
Dep. Variable:                rracmar   No. Observations:                 1118
Model:                          Logit   Df Residuals:                     1116
Method:                           MLE   Df Model:                            1
Date:                Tue, 18 Feb 2025   Pseudo R-squ.:                 0.008563
Time:                        15:27:17   Log-Likelihood:                -599.97
converged:                       True   LL-Null:                       -605.15
Covariance Type:            nonrobust   LLR p-value:                  0.001285
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.1450      0.072    -15.842      0.000      -1.287      -1.003
rhhi          -2.7619      1.001     -2.760      0.006      -4.723      -0.801
==============================================================================
```

Interpretation of Model 1:

- The coefficient on `rhhi` is negative and statistically signficant (p-value < 0.01). This suggests that higher social circle racial diversity is associated with the belief that there should not be laws against marriages between blacks and whites. This is line with our hypothesis.

- Specifically, going from zero to full social circle racial diversity (increase of 1 unit of `rhhi`) lowers the log odds of thinking that there should be laws against marriages between blacks and whites by 2.7619.

  - Note that full racial diversity `rhhi` =1 is impossible. Given the 5-person social circle limitation of the survey design. The maximum racial diversity achievable is $5 \times (\frac{1}{5})^2 = 0.8$, where the respondent has 5 friends of different races. Even without the 5-person limit, you can only approach, but never obtain, `rhhi` =1.

Model 2: `rhhi` , `rraclive`

Remember that we wondered if high racial diversity in social circles contribute to people's opposition toward laws against interracial marriage above and beyond the correlation between racially diverse social circles and exposure to people of other races.

In other words, people with racially diverse social circles tend to be exposed to people of other races more than people with less diverse social circles. It is possible that simple exposure to people of other races is sufficient to produce opposition toward laws against interracial marriage.

Model 2 tests this by including `rraclive`, a binary variable that indicates if there are "whites" (for black respondents) or "blacks" (for non-black respondents) living in their neighborhood. This is thus a dummy for exposure to people of another race.

Let's run the regression:

```
In [8]:  # Model 2: rracmar against rhhi and rraclive
         X = sm.add_constant(df_clean[['rhhi', 'rraclive']])
         logit_model = sm.Logit(df_clean['rracmar'], X).fit()

         print(logit_model.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.535447
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                rracmar   No. Observations:                 1118
Model:                          Logit   Df Residuals:                     1115
Method:                           MLE   Df Model:                            2
Date:                Tue, 18 Feb 2025   Pseudo R-squ.:                 0.01077
Time:                        15:27:17   Log-Likelihood:                -598.63
converged:                       True   LL-Null:                       -605.15
Covariance Type:            nonrobust   LLR p-value:                  0.001474
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.0339      0.098    -10.564      0.000      -1.226      -0.842
rhhi          -2.6443      1.003     -2.635      0.008      -4.611      -0.678
rraclive      -0.2339      0.143     -1.633      0.102      -0.514       0.047
==============================================================================
```

Interpretation of Model 2:

- The coefficient on `rhhi` *remains* negative and statistically signficant (p-value < 0.01). This suggests that positive effect of social circle racial diversity on opposition toward laws against marriages between blacks and whites *persists* even when including a proxy variable for exposure to people of other races.

- The coefficient on `rhhi` decreases in magnitude slightly but does not change much. This suggests that `rraclive` is not a significant channel through which `rhhi` acts on `rracmar`.

- The coefficient on `rraclive` is directionally in line with expectations. People who have other races living in their neighbourhood tend to oppose laws against marriages between blacks and whites, though this effect is statistically insignficant (p-value > 0.1).

Model 1 and 2 both support our initial hypothesis.

5- Consider alternate specifications of your variables (i.e., recodings of various kinds). Consider interactions among your variables.

Model 3: `racehomo`, `rraclive`

Instead of racial diversity in a social circle, let us consider an alternative but related independent variable, racial homophily. Perhaps, having multiple people of a race different from one's own is sufficient to impact attitudes on interracial marriage, and racial variety adds no further value.

In other words, it is possible that a white person with 1 asian and 1 black friend would have essentially the same attitudes on interracial marriage compared to a white person with 2 asian friends. The additional racial variety in the first person's social group does not move the needle for their attitudes on interracial marriage.

Model 3 tests this, as we study `racehomo`, our measure of racial homophily/ heterophily. Higher values are associated with greater racial heterophily and less homophily in the social circle.

We do not include `racehomo` and `rhhi` in the same model, as they measure extremely similar things and are thus highly correlated, which leads to problems with imperfect multicollinearity.

Let's run the regression:

```
In [9]:  # Model 3: rracmar against racehomo, rraclive
         X = sm.add_constant(df_clean[['racehomo', 'rraclive']])
         logit_model = sm.Logit(df_clean['rracmar'], X).fit()

         print(logit_model.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.534228
         Iterations 7
                         Logit Regression Results
==============================================================================
Dep. Variable:                rracmar   No. Observations:                1118
Model:                          Logit   Df Residuals:                    1115
Method:                           MLE   Df Model:                           2
Date:                Tue, 18 Feb 2025   Pseudo R-squ.:                 0.01303
Time:                        15:27:17   Log-Likelihood:                -597.27
converged:                       True   LL-Null:                       -605.15
Covariance Type:            nonrobust   LLR p-value:                 0.0003773
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.8869      0.312     -6.043      0.000      -2.499      -1.275
racehomo      -0.8727      0.308     -2.838      0.005      -1.476      -0.270
rraclive      -0.2517      0.143     -1.759      0.079      -0.532       0.029
==============================================================================
```

Interpretation of Model 3:

- The coefficient on `racehomo` is negative and statistically signficant (p-value < 0.01). This suggests that lower social circle racial homophily (higher racial heterophily) is associated with the belief that there should not be laws against marriages between blacks and whites. This is line with our hypothesis.

- Specifically, an increase of 1 unit of `racehomo` lowers the log odds of thinking that there should be laws against marriages between blacks and whites by 0.8727.

- This suggests that racial variety adds no additional value in affecting attitudes in `rracmar`. Simply having more people of a race other than yours in your social circle is sufficient to explain attitudes towards laws against marriages between blacks and whites. `racehomo` contains less information than `rhhi`, and is still a statistically significant predictor of `rracmar`.

- Since `rraclive` is included in this model, it suggests that low racial homophily (high racial heterophly) in social circles contribute to people's opposition toward laws against interracial marriage above and beyond the correlation between racially heterophilous social circles and exposure to people of other races.

Model 4: `racehomo`, `rraclive`, controls

Given this finding, let us proceed with `racehomo` instead of `rhhi` in our subsequent models. Let us now consider more factors that we think might affect attitudes toward anti-interracial marriage laws and might be

correlated with social circle racial homophily:

- `white` : whether a person is white or black. It is possible that racial homophily is higher in one of these racial groups. Given historical context, whites may be more opposed to interracial marriage than blacks, on average, affecting `rracmar` .

- `age` : as integration policies were only recently implemented in the decades leading up to survey year 1985, it is possible that older people maintain disapproval to interracial marriages and generally have friends of the same race as they grew up in a more segregated era. Then, they may be more supportive of laws against interracial marriage, and have more racially homophilic social circles.

- `educ` : perhaps more educated individuals care less about race as they value the internal aspects of a person (e.g. intellect) more than external aspects of a person, and are thus more approving of interracial marriages.

- `polviews` : perhaps people with more liberal political views would support interracial marriage more and have more racially heterophilic social circles compared to people with more conservative political leanings. People with more liberal political views tend to be more socially liberal as well.

```
In [10]:  # Model 4: rracmar against rhhi, rraclive, controls
          X = sm.add_constant(df_clean[['racehomo', 'rraclive', 'white', 'age', 'educ', 'polviews']])
          logit_model = sm.Logit(df_clean['rracmar'], X).fit()

          print(logit_model.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.453972
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                rracmar   No. Observations:                 1118
Model:                          Logit   Df Residuals:                     1111
Method:                           MLE   Df Model:                            6
Date:                Tue, 18 Feb 2025   Pseudo R-squ.:                  0.1613
Time:                        15:27:17   Log-Likelihood:                -507.54
converged:                       True   LL-Null:                       -605.15
Covariance Type:            nonrobust   LLR p-value:                  1.976e-39
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.5369      0.752     -3.376      0.001      -4.010      -1.064
racehomo      -0.7849      0.299     -2.622      0.009      -1.372      -0.198
rraclive      -0.1013      0.158     -0.639      0.523      -0.412       0.209
white          2.5361      0.559      4.536      0.000       1.440       3.632
age            0.0308      0.005      6.557      0.000       0.022       0.040
educ          -0.2752      0.035     -7.803      0.000      -0.344      -0.206
polviews       0.0816      0.059      1.380      0.168      -0.034       0.198
==============================================================================
```

Interpretation of Model 4:

- The inclusion of other variables lowers the magnitude of `racehomo` 's coefficient slightly, but it is still statistically signficant (p-value < 0.01). The association of social circle racial homophily/ heterophily and attitudes toward anti-interracial marriage laws is therefore robust to the inclusion of these other variables.

- The other variables included in the model are directionally associated with `rracmar` as per our expectations:

  - Being white (rather than black) raises the log odds of thinking that there should be laws against marriages between blacks and whites by 2.5361, net of other factors. This is highly statistically significant (p-value < 0.001).

- Every additional year of a person's age raises the log odds of thinking that there should be laws against marriages between blacks and whites by 0.0308, net of other factors. This is highly statistically significant (p-value < 0.001).

- Every additional year of education lowers the log odds of thinking that there should be laws against marriages between blacks and whites by 0.2752, net of other factors. This is highly statistically significant (p-value < 0.001).

- Every 1-point increase of a person's conservativeness on a 7-point liberal-conservative scale raises the log odds of thinking that there should be laws against marriages between blacks and whites by 0.0816, net of other factors. However, this effect is not statistically significant (p-value > 0.1).

Model 5: `racehomo`, `rraclive`, interaction term, controls

Model 5 considers if there is an interaction between the effect of social circle racial homopily and age. The intuition here being that older people's views toward interracial marriage may be more deep-seated due to the environment in which they grew up in and the possibility that odler folk face greater inertia in changing their views. Then, we will expect the interaction to have a positive coefficient, indicating that age reduces the effect that lower racial homophily (greater racial heterophily) has on attitudes toward anti-interracial marriage laws.

In [11]:
```python
# Model 5: rracmar against rhhi, rraclive, controls, interaction term
X = sm.add_constant(df_clean[['racehomo', 'rraclive', 'white', 'age', 'educ', 'polviews']])
X['racehomo:age'] = df_clean['racehomo'] * df_clean['age']  # Add interaction term
logit_model = sm.Logit(df_clean['rracmar'], X).fit()

print(logit_model.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.450646
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:                 rracmar   No. Observations:                 1118
Model:                           Logit   Df Residuals:                     1110
Method:                            MLE   Df Model:                            7
Date:                 Tue, 18 Feb 2025   Pseudo R-squ.:                  0.1674
Time:                         15:27:17   Log-Likelihood:                -503.82
converged:                        True   LL-Null:                       -605.15
Covariance Type:             nonrobust   LLR p-value:                 3.144e-40
==============================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const           -5.4091      1.688     -3.204      0.001      -8.718      -2.100
racehomo        -3.8553      1.583     -2.436      0.015      -6.957      -0.753
rraclive        -0.1046      0.159     -0.658      0.510      -0.416       0.207
white            2.5460      0.558      4.566      0.000       1.453       3.639
age              0.0889      0.028      3.160      0.002       0.034       0.144
educ            -0.2786      0.036     -7.842      0.000      -0.348      -0.209
polviews         0.0695      0.059      1.172      0.241      -0.047       0.186
racehomo:age     0.0604      0.028      2.129      0.033       0.005       0.116
==============================================================================
```

Interpretation of Model 5:

- As expected, the coefficient on the interaction term is positive, and it is statistically significant (p-value < 0.05). This suggests that the effect of racial homophily on attitudes toward anti-interracial marriage laws is lower for older people.

- The coefficient on `racehomo` is still negative and statistically significnat (p-value < 0.05). An increase of 1 unit of `racehomo` lowers the log odds of thinking that there should be laws against marriages between

blacks and whites by 3.8553, net of other factors. This is much larger than our prior estimates. However, every additional year of a person's age reduces this effect's magnitude by 0.0604.

- In short, having a social circle with lower racial homophily affects the `rracemar` attitudes of the young much more than the old.

6- And give your best conclusion as to whether your initial hypothesis held up and if not, why not.

I have given my conclusions throughout. In summary, the answer appears to be: yes, people with more racial diversity and lower racial homophily in their social circle will tend to believe that there should not be laws against marriages between blacks and whites.

- This association is robust to the inclusion of other factors like exposure to people of the other race (other race lives in the neighbourhood),race, age, education, and political views.

- Note however, that the association we find is still descriptive (hence the use of 'tend to' in the hypothesis, rather than 'causes' or 'results in'). Whilst including potential confounding variables is a step toward countering omitted variable bias, we do not argue that we have necessarily found a causal link.

  - It is possible that more racially diverse and less homophilic social circles changes attitudes on laws against interracial marriage (the causal possibility).
  - However, it is also possible that people have some other underlying reason to care little about race as a factor in socializing or marriage. Then, they will tend to have more racially diverse and less homophilic social circles, and they will also tend to oppose laws against interracial marriage. The association may therefore arise from some other unobserved confounding personal characteritic.