Lab1_TimeSeries_GideonTay

October 1, 2024

1 Lab 1 (QMSS5016 Time Series, Panel Data & Forecasting)

Submitted by: Gideon Tay
My UNI: gt2528
Contact me at: gideon.tay@columbia.edu

1.1 Question 2. Conduct an unpooled regression comparison across time-periods. Compare at least two time periods, running regressions on each and comparing coefficients. Explain your results.

Overview: For this lab, we will explore whether factors like education, gender, race etc. are good predictors of people's political views in terms of the extent to which they identify as liberal or conservative. Moreover, we would compare how the relevance, direction, and magnitude of these predictors differ in two separate time periods: 1980-1990 and 2010-2020.

1.1.1 Import all necessary libraries for this lab

[1]:

```
# Libraries for data analysis
import pandas as pd # also used to load in data, but used primarily for analysis
import numpy as np
import statsmodels.formula.api as smf
from scipy.stats import norm

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

1.1.2 Load in General Social Survey data

We load in data for the following columns: 'id', 'age', 'year', 'sex', 'incom16', 'born', 'race', 'educ', 'attend', and 'polviews'. We only load in their numeric labels.

[2]:

```
# 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'
```

```python
# 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 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()

    # 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 3a: Load only the selected columns into a pandas DataFrame with numeric labels
            columns_to_load = ['id', 'age', 'year', 'sex', 'incom16',
                               'born', 'race', 'educ', 'attend', 'polviews']
            print('Loading selected columns from Stata file...')
            df_numeric = pd.read_stata(
                stata_file_stream,
                columns=columns_to_load,
                convert_categoricals=False)
            print('Data with numeric labels loaded successfully!')

# Step 3: Display the first few rows of the final DataFrame
```

```
df_numeric.head()
```

```
  0%|          | 0.00/81.9M [00:00<?, ?iB/s]

Download completed!
Loading selected columns from Stata file…
Data with numeric labels loaded successfully!
```

```
[2]:    id   age  year  sex  incom16  born  race  educ  attend  polviews
     0   1  23.0  1972  2.0      3.0   NaN   1.0  16.0     2.0       NaN
     1   2  70.0  1972  1.0      4.0   NaN   1.0  10.0     7.0       NaN
     2   3  48.0  1972  2.0      3.0   NaN   1.0  12.0     4.0       NaN
     3   4  27.0  1972  2.0      3.0   NaN   1.0  17.0     0.0       NaN
     4   5  61.0  1972  2.0      2.0   NaN   1.0  12.0     0.0       NaN
```

1.1.3 Clean the data

After loading in the data, we have to clean it. We remove rows containing missing values, or non-standard entries such as: “Inapplicable” (code: -100), “No answer” (code: -99), “Do not Know/Cannot Choose” (code: -98), and “Skipped on Web” (code: -97). These numeric codes for non-standard entries are consistent across all columns.

We also remove rows with “Lived in institution” (code: 7) in the “incom16” column, as it breaks away from the ordinal pattern of 1 representing “far below average” and 5 representing “far above average” family income when the respondent was 16.

```
[3]: # Produce a copy of the dataframe for cleaning and analysis
     df = df_numeric.copy()

     # Define the non-standard codes to be treated as missing values
     non_standard_codes = [-100, -99, -98, -97]

     # Replace the non-standard codes with NaN
     df.replace(non_standard_codes, np.nan, inplace=True)

     # Drop missing values
     df.dropna(inplace=True)

     # Remove rows with '7' (Lived in institution) in the incom16 column.
     df = df[df['incom16'] != 7]
```

1.1.4 Recode the data

Next, we recode the data so a meaningful regression can be conducted on it. For example, it makes no sense for male = 1 and female = 2 as with the current ‘sex’ column. Instead, we recode it to a dummy variable where male = 1 and female = 0. We similarly produce dummy variables for race and birth origin (born in or outside of the U.S.)

[4]:

```python
# Create a dummy variable for 'Male' where 1 = Male, 0 = Female
df['male_dummy'] = df['sex'].apply(lambda x: 1 if x == 1 else 0)

# Create a dummy variable where 1 = Born in US, 0 = Not born in US
df['born_dummy'] = df['born'].apply(lambda x: 1 if x == 1 else 0)

# Create dummy variables for 'black' and 'other' where white is the base category
# Note that numeric codes are 1 for white, 2 for black, and 3 for other
df['black_dummy'] = df['race'].apply(lambda x: 1 if x == 2 else 0)
df['other_dummy'] = df['race'].apply(lambda x: 1 if x == 3 else 0)
```

1.1.5 Explaining each variable we picked

Before we begin analysis, here is an explanation of each variable of interest after recoding. Note that numbers in paranthesis indicate the numerical code of a corresponding response. The dependent variable we are interest in is "polviews" while the other variables would be tested as potential predictors of "polviews".

- **age**: respondent's age in years from 18 to 89 years or older. Numerical code values range from 18 to 89.
- **year**: year of the GSS survey.
- **male_dummy**: respondent is male (1) or female (0).
- **born_dummy**: respondent is born in this country (1) or not born in this country (0).
- **black_dummy**: respondent is black (1) or not black (0).
- **other_dummy**: respondent is neither white nor black (1) or respondent is white or black (0).
- **incom16**: Thinking about the time when you were 16 years old, compared with American families in general then, would you say your family income was– far below average (1), below average (2), average (3), above average (4), or far above average (5)?
- **educ**: highest year of school that the respondent finished and got credit for. Responses range from no formal schooling (0) to 8 or more years of college (20).
- **attend**: How often do you attend religious services? Responses range from never (0) to several times a week (8).
- **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 (1) to extremely conservative (7). Where would you place yourself on this scale?

1.1.6 Run a regression for the first time period: 1980-1990

Why this time period: We chose this time period since GSS collected data for our variables of interest for this time period. Note that some of our variables only had data starting 1974 as the question was not included in earlier GSS surveys. Hence, studying the 1970 to 1980 period for example would be suboptimal as there would be significant missing data. We thus chose the next earliest decade of 1980-1990.

[5]:

```python
# Running the first regression for the 1980-1990 period
polviews80 = smf.ols('polviews ~ age + year + male_dummy + born_dummy + '
                     'black_dummy + other_dummy + incom16 + educ + attend',
                     data=df.query('year >= 1980 & year <= 1990')).fit()

# Print the summary of the model
print(polviews80.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               polviews   R-squared:                       0.045
Model:                            OLS   Adj. R-squared:                  0.044
Method:                 Least Squares   F-statistic:                     72.31
Date:                Tue, 01 Oct 2024   Prob (F-statistic):          3.83e-131
Time:                        13:22:26   Log-Likelihood:                -23176.
No. Observations:               13789   AIC:                         4.637e+04
Df Residuals:                   13779   BIC:                         4.645e+04
Df Model:                           9
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept       8.6151      7.285      1.183      0.237      -5.665      22.895
age             0.0059      0.001      8.766      0.000       0.005       0.007
year           -0.0025      0.004     -0.684      0.494      -0.010       0.005
male_dummy      0.1248      0.023      5.528      0.000       0.081       0.169
born_dummy      0.1110      0.048      2.299      0.022       0.016       0.206
black_dummy    -0.4500      0.032    -13.864      0.000      -0.514      -0.386
other_dummy    -0.1062      0.072     -1.479      0.139      -0.247       0.035
incom16         0.0100      0.014      0.734      0.463      -0.017       0.037
educ           -0.0153      0.004     -3.885      0.000      -0.023      -0.008
attend          0.0719      0.004     16.731      0.000       0.063       0.080
==============================================================================
Omnibus:                       33.803   Durbin-Watson:                   1.945
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               28.749
Skew:                          -0.052   Prob(JB):                     5.72e-07
Kurtosis:                       2.802   Cond. No.                     1.31e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.31e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Analysis of results: Most variables appear to be statistically significant ($p < 0.05$) predictors of political views, holding the other variables constant. Collectively, the predictors explain about 4.5% of the variance in political views (R2 = 0.045).

In particular, people who are older, male (compared to females), born in the country, and who attend religious services often tend to be more conservative (positive coefficients). Meanwhile, people who are black (compared to whites) and who have completed more years of education tend to be more liberal (negative coefficients).

However, 3 variables 'year', 'other_dummy' and 'incom16' are not statistically significant predictors:

- **year**: This makes it seem like there is no clear unidirectional movement in political views toward being more liberal or conservative over the 1980-1990 period. However, since the question is phrased in a relative sense on how a respondent views his extent of conservativeness/ liberalness to other people in society, it is perfectly possible that society has collectively become more conservative or liberal over the time period, but each person's perception of their relative conservativeness/ liberalness have not changed on average over the years. Hence, this must be interpreted with caution.
- **other_dummy**: since 'white' is the omitted category for the race dummy variables, this suggests that the 'other' race group (non-black and non-white) is not systematically more conservative or liberal compared to whites. However, this does not preclude the possibility of variation within the 'other' race group (e.g. hypothetically if asian americans are more conservative and hispanics are more liberal, compared to whites, and the effects cancel out so there is no net effect).
- **incom16**: this variable is a good proxy for the socio-economic background of respondents' upbringing. I initially hypothesized that those who grew up in poorer backgrounds would tend to be more liberal as they would more likely be in favour of greater government support to the less fortunate and a stronger social net funded by higher taxes at higher income brackets. However, this hypothesis appears to be incorrect for this time period.

1.1.7 Run a regression for the second time period: 2010-2020

Why this time period: Again, we chose this time period to maximise data availability across our variables of interest. Variable incom16's data was not collected in the 2000 GSS survey, so studying the 2000-2010 would lead to significant missing data, and we would only have been able to study 2002-2010. We thus studied 2010-2020.

[6]:

```python
# Running the first regression for the 2000-2010 period
polviews10 = smf.ols('polviews ~ age + year + male_dummy + born_dummy + '
                     'black_dummy + other_dummy + incom16 + educ + attend',
                     data=df.query('year >= 2010 & year <= 2020')).fit()

# Print the summary of the model
print(polviews10.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               polviews   R-squared:                       0.093
Model:                            OLS   Adj. R-squared:                  0.092
Method:                 Least Squares   F-statistic:                     125.7
Date:                Tue, 01 Oct 2024   Prob (F-statistic):          4.44e-226
Time:                        13:22:26   Log-Likelihood:                -19367.
```

```
No. Observations:               11038   AIC:                         3.875e+04
Df Residuals:                   11028   BIC:                         3.883e+04
Df Model:                           9
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept      -2.7144      9.789     -0.277      0.782     -21.904      16.475
age             0.0041      0.001      5.159      0.000       0.003       0.006
year            0.0033      0.005      0.687      0.492      -0.006       0.013
male_dummy      0.1743      0.027      6.464      0.000       0.121       0.227
born_dummy      0.1717      0.044      3.867      0.000       0.085       0.259
black_dummy    -0.4641      0.038    -12.181      0.000      -0.539      -0.389
other_dummy    -0.2663      0.050     -5.291      0.000      -0.365      -0.168
incom16         0.0040      0.015      0.266      0.790      -0.025       0.033
educ           -0.0538      0.005    -11.619      0.000      -0.063      -0.045
attend          0.1360      0.005     27.777      0.000       0.126       0.146
==============================================================================
Omnibus:                       87.020   Durbin-Watson:                   1.933
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               70.000
Skew:                          -0.121   Prob(JB):                     6.30e-16
Kurtosis:                       2.694   Cond. No.                     1.48e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.48e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Analysis of results: Again, most variables appear to be statistically significant ($p < 0.05$) predictors of political views, holding the other variables constant. This time, the predictors explain about 9.3% of the variance in political views (R2 = 0.093), much higher compared to the 1980-1990 period.

As with the previous regression, people who are older, male (compared to females), born in the country, and who attend religious services still tend to be more conservative (positive coefficients). Meanwhile, people who are black or another race (compared to whites) and who have completed more years of education tend to be more liberal (negative coefficients).

As with the previous regression, 'year' and 'incom16' are not statistically significant predictors of political views. However, for the 2010-2020 regression, 'other_dummy' is now statistically signficant, whearas it was not significant in the 1980-1990 regression.

Let us interpret some of the coefficients below:

- **age**: a respondent who is a year older is on average 0.0041 points more conservative on the 7-point political view scale, compared to a resondent who is one year younger, all other variables held constant.
- **male_dummy**: a male respondent is on average 0.1743 points more conservative on the 7-

point political view scale, compared to a female respondent, all other variables held constant.

- **born_dummy**: a respondent who is born in the country is on average 0.1717 points more conservative on the 7-point political view scale, compared to a respondent who is born outside the country, all other variables held constant.
- **black_dummy**: a black respondent is on average 0.4641 points less conservative on the 7-point political view scale, compared to a white respondent, all other variables held constant.
- **other_dummy**: a non-white and non-black respondent is on average 0.2663 points less conservative on the 7-point political view scale, compared to a white respondent, all other variables held constant

1.1.8 Compare coefficients between the 2 time periods

We create a comparison table of 1980-1990 and 2010-2020 coefficients. We also calculate a Z-score and p-value to test if the coefficients for each variable differs between the 2 periods in a statistically significant manner.

[7]:

```
# Extracting the coefficients and std. errors and renaming columns
p80 = (
    polviews80.summary2() # provides a detailed summary of polviews80 regression
    .tables[1] # table of coefficients, std errors etc. vs table[0] with R^2␣
 ↪etc.
    .reset_index()
    .rename(columns={ # rename so we can distinguish between regressions later
        'Coef.': 'coef80',
        'Std.Err.': 'se80'
    }))
p10 = (
    polviews10.summary2()
    .tables[1]
    .reset_index()
    .rename(columns={
        'Coef.': 'coef10',
        'Std.Err.': 'se10'
    }))

# Merge the two regression results by on the index column (variable name)
df_merged = pd.merge(
    p80[['index', 'coef80', 'se80']],
    p10[['index', 'coef10', 'se10']],
    on='index')

# Apply the Z formula to compare coefficients
df_merged['b1minusb2'] = df_merged['coef80'] - df_merged['coef10']
df_merged['denom'] = (df_merged['se80']**2 + df_merged['se10']**2)**0.5
df_merged['Z'] = df_merged['b1minusb2'] / df_merged['denom']

# Calculate p-values and set them to 4 decimal places
df_merged['pvalue'] = 2 * (1 - norm.cdf(abs(df_merged['Z'])))
```

```
df_merged['pvalue'] = df_merged['pvalue'].apply(lambda x: f"{x:.4f}")


# Selecting the relevant columns
df_compare = df_merged[['index', 'coef80', 'coef10', 'Z', 'pvalue']]

# Display the result
print(df_compare)
```

```
         index    coef80    coef10         Z  pvalue
0    Intercept  8.615129 -2.714366  0.928439  0.3532
1          age  0.005890  0.004052  1.777721  0.0754
2         year -0.002513  0.003339 -0.960568  0.3368
3   male_dummy  0.124839  0.174263 -1.405372  0.1599
4   born_dummy  0.110987  0.171661 -0.925116  0.3549
5  black_dummy -0.449995 -0.464108  0.281974  0.7780
6  other_dummy -0.106202 -0.266293  1.825516  0.0679
7      incom16  0.009959  0.003963  0.297623  0.7660
8         educ -0.015252 -0.053795  6.349855  0.0000
9       attend  0.071862  0.135980 -9.845164  0.0000
```

Analysis of results: Most variables appear to have essentially the same effect in the 1980-1990 and the 2010-2020 period as observed by the large p-values ($p > 0.05$). Whilst the magnitude of some coefficients have changed, and in the case of 'year', even changed directions (from negative to positive), the changes are mostly not statistically signficant.

Only the effect of two variables appear to have changed significantly: 'educ' and 'attend' ($p < 0.0001$).

- The magnitude of the effect of education on political views have increased by about 3.5 times while that of religious service attendance have increased by about 1.9 times.
- The direction of the effects in both cases have been maintained: more educated people tend to be more liberal while people who attend religious services more often tend to be more conservative.

This suggests that political views in the U.S. have become increasingly split on education and religious lines. This could explain the increase in R2 from 0.045 to 0.093 between the two time periods, as 'educ' and 'attend' become stronger predictors and explain more of the variance in political views in 2010-2020 compared to 1980-1990.

At the same time, age, gender, whether one was born in the country, and race continue to predict political views in a similar manner, even 30 years later.