# DS4E_hw4_template

November 22, 2022

## 1 DS4E: Homework 4

[1]:

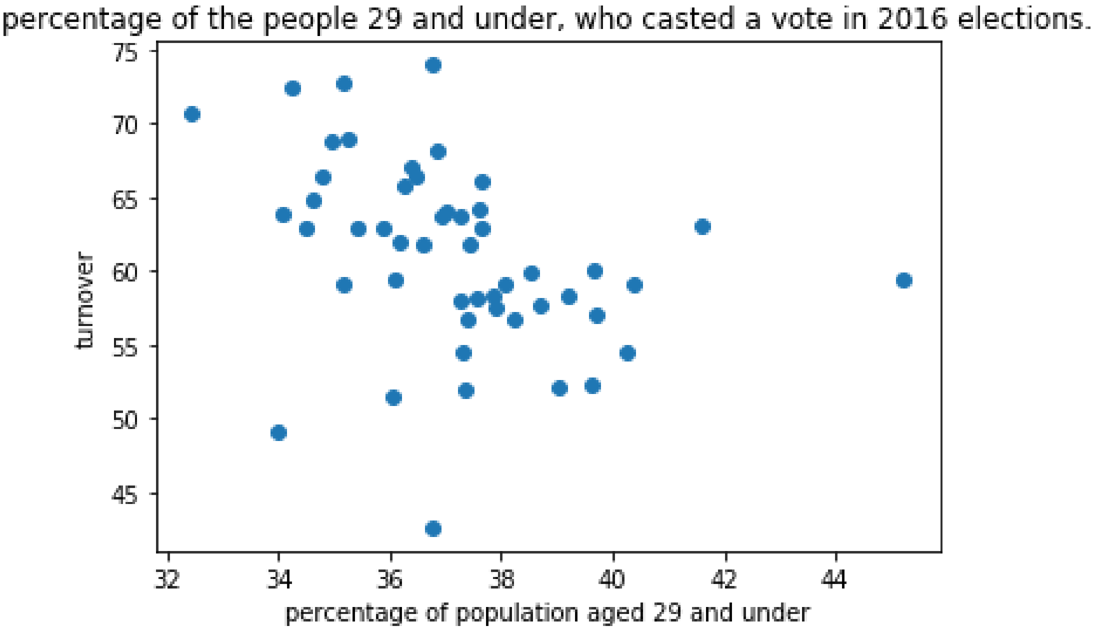
```
# import libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import statsmodels.formula.api as smf
```

### 1.1 Question 1

**1(a)**

[2]:

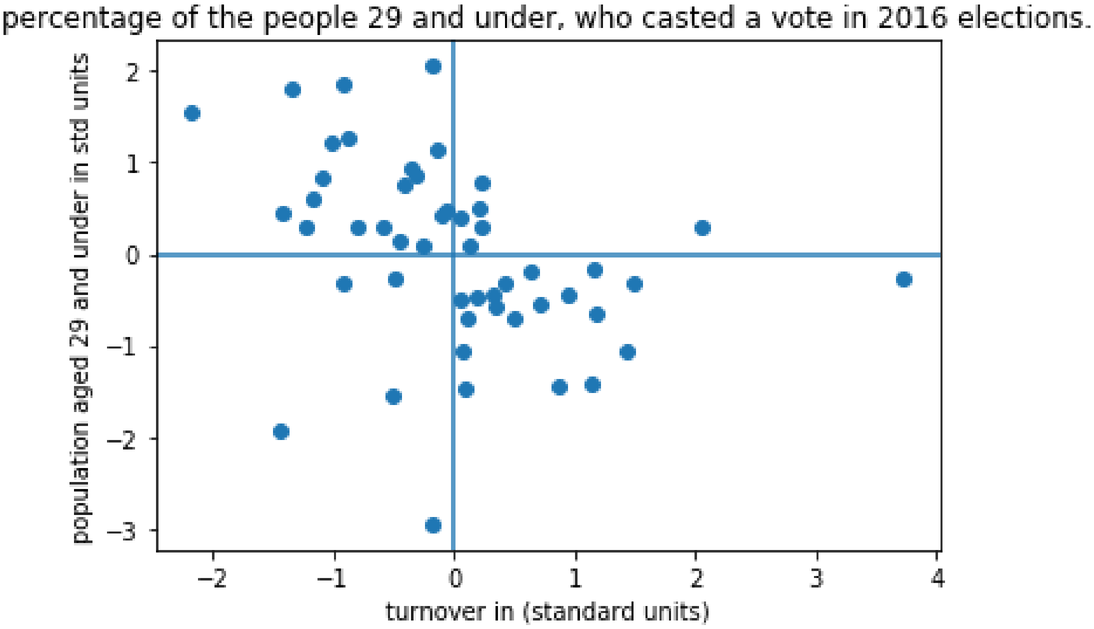
```
data = pd.read_csv('election_2016.csv')
plt.scatter(data['age29andunder_pct'], data['turnout'])
plt.xlabel("percentage of population aged 29 and under")
plt.ylabel("turnover")
plt.title("percentage of the people 29 and under, who casted a vote in 2016␣
  ↪elections.")
plt.show()
```



### 1(b)

[3]:
```
def standard_units(num):
    return (num - np.mean(num))/np.std(num)
turnout = standard_units(data['turnout'])
age_29 = standard_units(data['age29andunder_pct'])

plt.scatter(age_29, turnout)
plt.xlabel("turnover in (standard units)")
plt.ylabel("population aged 29 and under in std units")
plt.axhline(0)
plt.axvline(0)
plt.title("percentage of the people 29 and under, who casted a vote in 2016␣
  ↪elections.")
plt.show()
```



There is a weak negative relationship.

### 1(c)

[4]:
```
def correlation(turnout,age_29):
    multiple_corr =turnout*age_29
    sum_corr=sum(multiple_corr)
    corr=((1/50)*sum_corr)
    return corr
```

```
turnout = standard_units(data['turnout'])
age_29 = standard_units(data['age29andunder_pct'])
print("the correlation is:",correlation(turnout,age_29))
```

```
the correlation is: -0.35687306231856225
```

### 1(d)

[5]:
```
corr_matrix = data[['turnout','age29andunder_pct']]
corr_matrix.corr()
```

[5]:
```
                    turnout  age29andunder_pct
turnout            1.000000          -0.356873
age29andunder_pct -0.356873           1.000000
```

The correlation between x and itself is always one, so the correlation coefficient between turnout and turnout is one and same for age29 and under as well. The correleation between turnout and age, and age and turnout is the same which is shown to be -0.356873.

### 1(e)

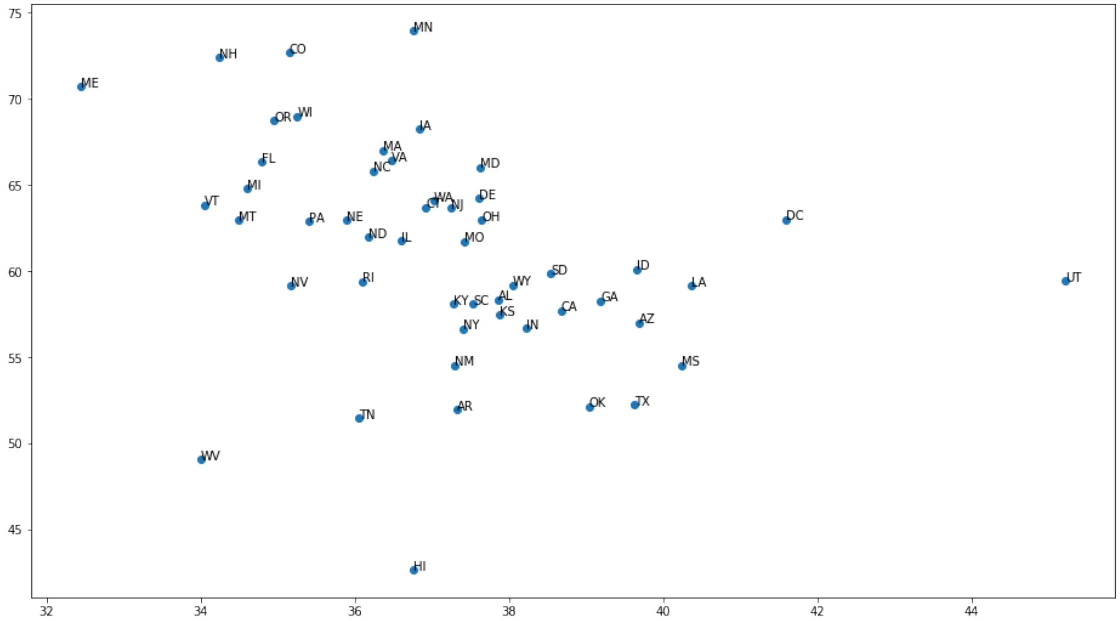
this is an example of ecological fallacy where we are using group data to infer about individuals, which in this case is that yount people are less interested in voting, which could be a wrong asumption.

## 1.2 Question 2

### 2(a)

[19]:
```
ids=data['stateid']
turnout_var = data['turnout']
age_var = data['age29andunder_pct']
plt.figure(figsize=(16,9))
plt.scatter(age_var,turnout_var)
for i, txt in enumerate(data['stateid']):
    plt.annotate(txt, (age_var[i], turnout_var[i]))
plt.show()

#used this websource to get an understanding of how to show lables on the␣
 ↪scatter points. https://www.delftstack.com/howto/matplotlib/
 ↪matplotlib-label-scatter-plot-points/
```



### 2(b)

q(2bi):Utah<br>
Q2B(ii): Minnesota<br>
Q2b(iii):Maine<br>
q2b(iv): Hawaii

### 2(c)

[7]:
```
regex_table = smf.ols('turnout_var ~ age_var', data=data).fit()
regex_table.summary()
# used the codebook to get an idea on how to display the regression table.
```

[7]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:            turnout_var   R-squared:                       0.127
Model:                            OLS   Adj. R-squared:                  0.109
Method:                 Least Squares   F-statistic:                     7.005
Date:                Tue, 22 Nov 2022   Prob (F-statistic):             0.0110
Time:                        17:51:13   Log-Likelihood:                -159.09
No. Observations:                  50   AIC:                             322.2
Df Residuals:                      48   BIC:                             326.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
```

```
------------------------------------------------------------------------------
Intercept     99.2005     14.422      6.879      0.000      70.204     128.197
age_var       -1.0259      0.388     -2.647      0.011      -1.805      -0.247
==============================================================================
Omnibus:                        9.802   Durbin-Watson:                   2.127
Prob(Omnibus):                  0.007   Jarque-Bera (JB):                9.875
Skew:                          -0.813   Prob(JB):                      0.00717
Kurtosis:                       4.447   Cond. No.                         638.
==============================================================================

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

### 2(d)

the estimated coeffecient for the intercept is 99.2005 and the estimated coeffecient for the slope is -1.0259

### 2(e)

The values tell us that there is a negative relationship as the coefficient of slope is negative, so increasing one will decrease the other one, also at 99.2005 turnout, there would be no voters aged 29 and under. The value of slope tells us the rate of change of turnover, which means, a 1 increase in slope will decrease the turnout by 1.0259.

### 2(f)

the value of p is 0.011 which is less than 0.05, and 95% confidence interval is between -1.805 to -0.247, and our coeffcient is between this value, which means that it is statistically significant at the 95% level.

### 2(g)

Since the value of r-squared is 0.127, it shows that the model explains 12.7% of the variance.

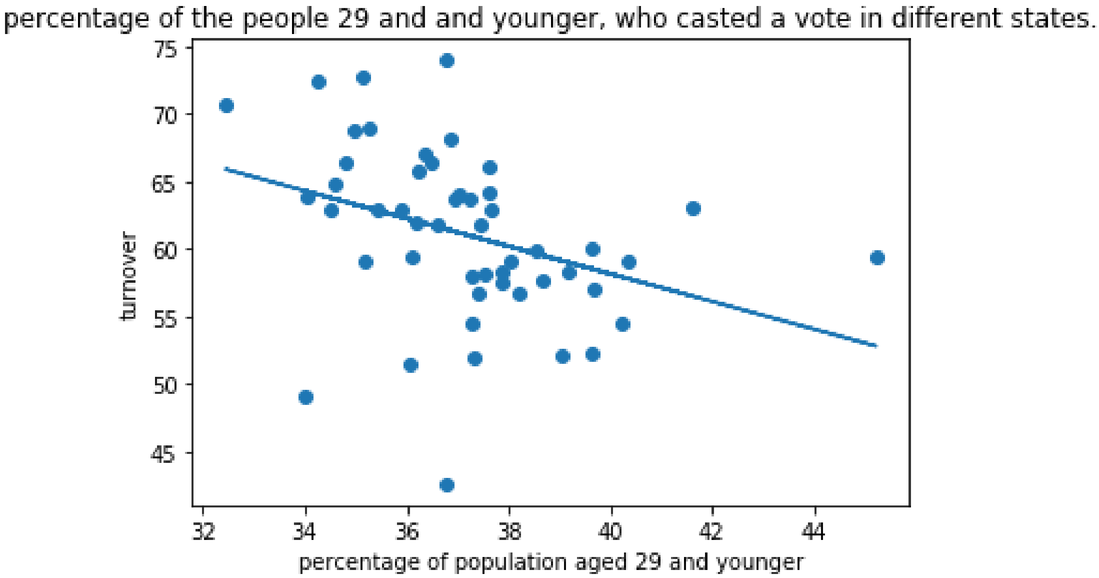
## 1.3 Question 3

### 3(a)

[25]:

```
turnout_var = data['turnout']
age_29 = data['age29andunder_pct']
plt.scatter(age_29,turnout_var)
arr_turnout = turnout_var.to_numpy()
arr_age = age_29.to_numpy()
m, b = np.polyfit(arr_age, arr_turnout, 1)
plt.ylabel("turnover")
plt.xlabel("percentage of population aged 29 and younger")
plt.title("percentage of the people 29 and and younger, who casted a vote in␣
  ↪different states.")
```

```
plt.plot(arr_age, m*arr_age+b)
plt.show()
#Citations: used this website to understand on how can I print the ols␣
 ↪regression line on the scatterplot. https://stackoverflow.com/questions/
 ↪42261976/how-to-plot-statsmodels-linear-regression-ols-cleanly
# also used this website to understand how to convert a dataframe to a␣
 ↪numpyarray. https://datatofish.com/dataframe-to-numpy-array/
```



**3(b)**

```
[26]: output= 99.2005+(40*(-1.0259))
      print("the turnout expected is ", output)
```

```
the turnout expected is  58.164500000000004
```

**3(c)**

```
[27]: def turnover(age):
          output= 99.2005+(age*(-1.0259))
          return output
      print("the expected turnover for new york would be",turnover(37.4068527902047) )
      print("the expected turnover for texas would be",turnover(39.6346134017231) )
      print("the expected turnover for West Virginia would be",turnover(34.
       ↪0080554689208) )
```

```
the expected turnover for new york would be 60.824809722529
the expected turnover for texas would be 58.539350111172276
the expected turnover for West Virginia would be 64.31163589443415
```

### 3(d)

[28]:

```
def difference(original,estimated):
    return original-estimated
print("the difference between turnout for new york is ", difference(56.
 ↪647399226289,60.824809722529))
print("the difference between turnout for texas is ", difference(52.
 ↪2146024885869,58.539350111172276))
print("the difference between turnout for West Virginia is ", difference(49.
 ↪0673061704778,64.31163589443415))

print("the largest residual is West Virginia")
```

```
the difference between turnout for new york is  -4.17741049624
the difference between turnout for texas is  -6.324747622585377
the difference between turnout for West Virginia is  -15.244329723956348
the largest residual is West Virginia
```

### 3(e)

[29]:

```
print("the percentage required is",(80-99.2005)/(-1.0259))
```

```
the percentage required is 18.71576177015304
```

## 1.4 Question 4

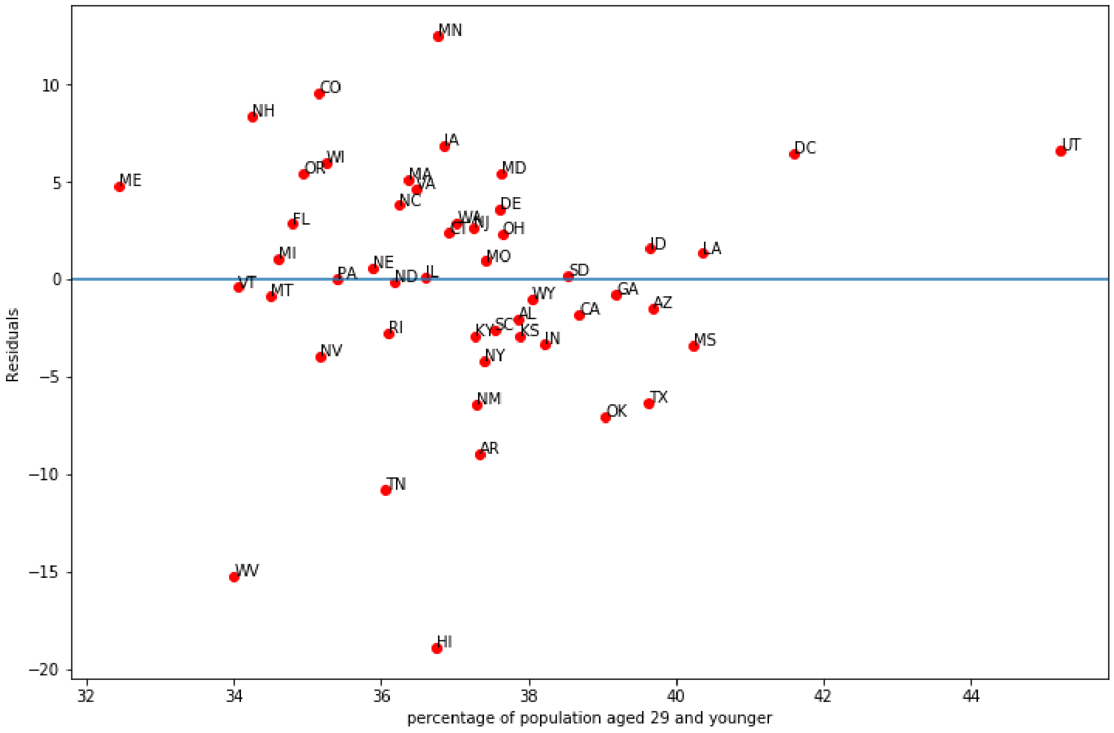
### 4(a)

[30]:

```
plt.figure(figsize=(12,8))
data["predicted_turnout"] = turnover(data["age29andunder_pct"])
data["residuals"]=data["turnout"]-data["predicted_turnout"]

plt.scatter(data['age29andunder_pct'], data['residuals'],color='red')
plt.xlabel('percentage of population aged 29 and younger')
plt.ylabel('Residuals')
plt.axhline(0)
for i, txt in enumerate(data['stateid']):
    plt.annotate(txt, (age_var[i], data['residuals'][i]))

plt.show()
```



between 35 and 40 percent of the age group, the values of the residuals are clustered near to the line and the residuals values are lower , where as we go above or below these percentage values, the residual values increase or decrease and thus overestimate and underestimation inreases or decsreses.

### 4(b)

```
[31]: print(turnover(data[data['stateid']=='PA']['age29andunder_pct'].values[0]))
print(data[['stateid', 'turnout']].loc[[37]].head())
print("the resideual value for pensalvania is ", 62.877945-62.86653772250857)
```

```
62.86818524119958
   stateid    turnout
37      PA  62.877945
the resideual value for pensalvania is  0.011407277491429113
```

### 4(c)

The largest positive residual is minnesota and the largest negative residual is Hawaii.

## 1.5 Question 5

### 5(a)

There is no informed consent as the people are unaware that a study is going on, and the people that are deciding to get the vaccine after the notification are being affected, the subjects in this case are being harmed.

**5(b)**

altough they are pushing for increased covid vaccination, there is a possibility that this vaccination can cause more harm than benefits, as there has been not any technical assement of the vaccine.

**5(c)**

There is not a fair distribution of risks and benefits of research as only men are being targeted, which can be a violation of the justice principle as it can harm them as well as the women will not benefit from the research and study, as women will not be not recieving the vaccine.

**5(d)**

The respect for law and public interest principle could be violated as there is no accountability, as it is very hard unclear who they can hold accountable. Furthermore, the company might not be following compialance anc confidentiality laws of the government as they do ask if a person is vaccinated or not, which might be a confidential information.

General Citations: Lecture notes and the textbook was used to understand some code, key concepts.