DS4E_hw4_template

November 29, 2022

1 DS4E: Homework 4

[2]:
```
# 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

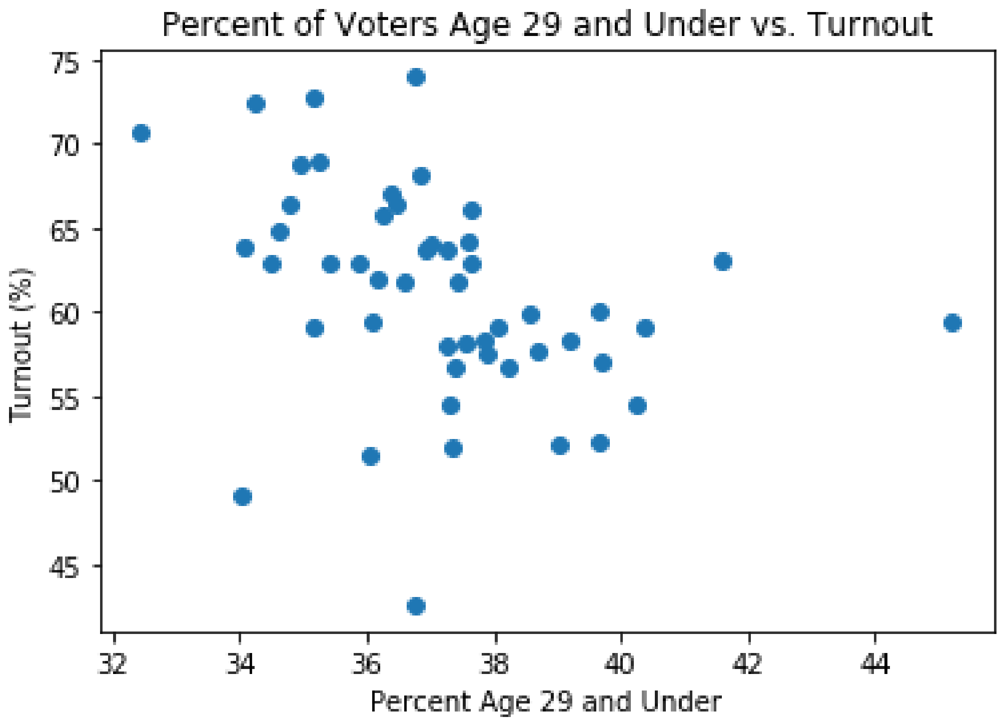
1(a)

[3]:
```
election_data = pd.read_csv('election_2016.csv')
election_data.head()
```

[3]:
```
        state stateid      cvap   turnout  age29andunder_pct  \
0     Alabama      AL   3639505  58.342192          37.864079
1     Arizona      AZ   4613575  56.978720          39.687833
2    Arkansas      AR   2175330  51.941361          37.333503
3  California      CA  24582600  57.689565          38.681232
4    Colorado      CO   3824445  72.696038          35.154120

   age65andolder_pct  median_hh_inc  lesscollege_pct
0          16.930066      38.834925        83.080870
1          18.951752      44.166533        80.589436
2          18.258998      37.503720        84.499622
3          15.962776      58.091241        73.988558
4          17.294236      52.243594        69.555890
```

[4]:
```
age29andunder_pct = election_data[['age29andunder_pct']]
turnout = election_data[['turnout']]
plt.scatter(age29andunder_pct, turnout)
plt.title('Percent of Voters Age 29 and Under vs. Turnout')
plt.xlabel('Percent Age 29 and Under')
plt.ylabel('Turnout (%)')
```

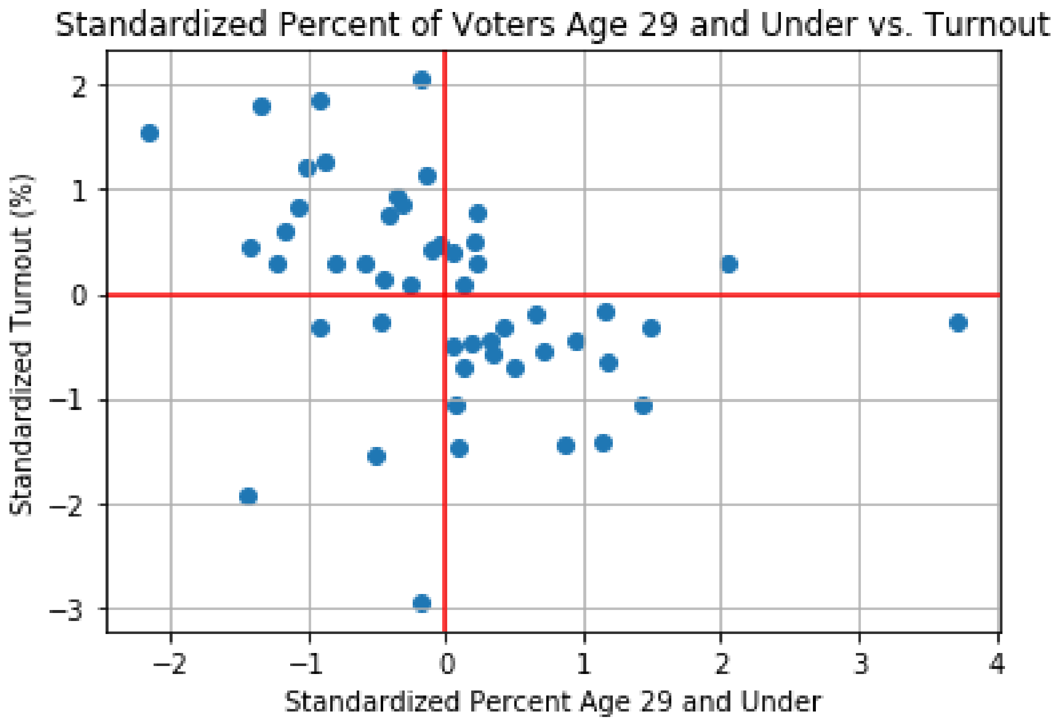
[4]: `Text(0, 0.5, 'Turnout (%)')`



1(b)

[5]:
```
def standardize(x):
    return(x - np.mean(x))/np.std(x)

standardized_age = standardize(age29andunder_pct)
standardized_turnout = standardize(turnout)

plt.scatter(standardized_age, standardized_turnout)
plt.axvline(0, color = 'red')
plt.axhline(0, color = 'red')
plt.grid()

plt.title('Standardized Percent of Voters Age 29 and Under vs. Turnout')
plt.xlabel('Standardized Percent Age 29 and Under')
plt.ylabel('Standardized Turnout (%)')
```

[5]:
```
Text(0, 0.5, 'Standardized Turnout (%)')
```



From the plot, I observed that the relationship between the two variables is negative, as it trends downwards not upwards (decreasing as we move to the right). The strength between the two variables doesn't seem to be very strong (low); I would assume that the correlation is around a -0.4, as there are a good amount of outliers, and the points are scattered around more than perfectly linearly associated. Since correlation is the measure of linear association, it is clear that the correlation isn't very high, as the points are scattered about pretty thouroughly, and since correlation is a measure of the relationship and strength, we can assume that the strength isn't great, or is closer to 0 than 1.

1(c)

[6]:
```
def correlation(x, y):
    sum = 0
    for observation in range(len(x)):
        sum = sum + (x.iloc[observation, 0] * y.iloc[observation, 0])
    corr = sum/len(x)
    return (corr)

print(correlation(standardized_age, standardized_turnout))
```

```
-0.35687306231856225
```

The correlation between the standardized variables is -0.35687.

[7]:
```
#checking results
results = smf.ols('standardized_turnout ~ standardized_age',␣
↪data=election_data).fit()
results.summary()
```

[7]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
================================================================================
Dep. Variable:     standardized_turnout   R-squared:                       0.127
Model:                              OLS   Adj. R-squared:                  0.109
Method:                   Least Squares   F-statistic:                     7.005
Date:                  Tue, 29 Nov 2022   Prob (F-statistic):             0.0110
Time:                          19:03:14   Log-Likelihood:                -67.541
No. Observations:                    50   AIC:                             139.1
Df Residuals:                        48   BIC:                             142.9
Df Model:                             1
Covariance Type:              nonrobust
================================================================================
====
                       coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
----
Intercept        -1.301e-15      0.135  -9.65e-15      1.000      -0.271
0.271
standardized_age    -0.3569      0.135     -2.647      0.011      -0.628
-0.086
==============================================================================
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.                         1.00
==============================================================================

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

1(d)

[8]:
```
frames = [standardized_age, standardized_turnout]
standardized_data = pd.concat(frames, axis = 1)
standardized_data.corr()
```

[8]:
```
                   age29andunder_pct   turnout
age29andunder_pct           1.000000 -0.356873
turnout                    -0.356873  1.000000
```

What is the value of the diagonal elements of the matrix, or the numbers that go from top left to bottom right, are both 1.0000. This is because in both these slots, the same exact variable is being compared. In the top left, Age 29 and Under Percent is being compared with Age 29 and Under Percent, and in the bottom right, Turnout is being compared with Turnout. Therefore, the correlation is a perfect 1, because the values on the x and y axis are both the same, creating perfect linearity/linear association (points would be in a straight line on a plot).

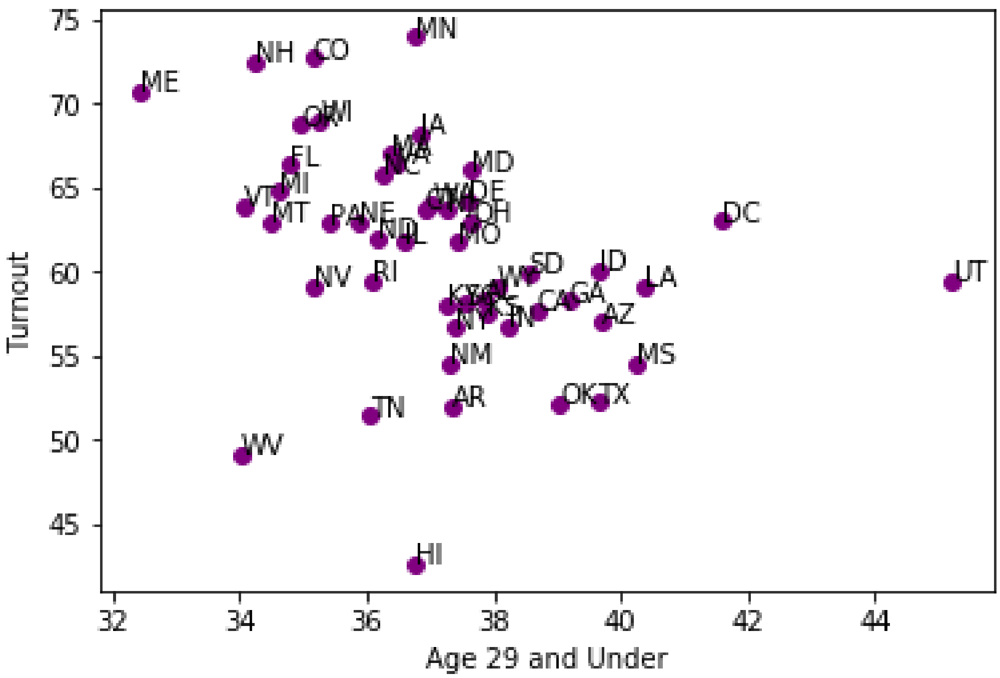
1(e)

Correlation is a measure of linear association between two variables, it does not prove causation. Just because there is correlation between two variables, does not mean they directly cause each other; there are many other variables that could be involved, altering the results. Correlation can be used to describe the relationship between two variables, and is sometimes used to predict the relationship between variables, but only loosely; it isn't enough to infer about variables. It is kind of similar to the ecological fallacy in a way, in that the kid is Furthermore, the correlation in this case is low; so even if the data was very telling, there isn't a strong linear relationship between the variables; so, the reader's inference wouldn't even make sense, as the relationship between percent of voters under the age of 29 and voter turnout is low.

1.2 Question 2

2(a)

[9]:
```python
stateid_data = election_data[['stateid']]

election_data_shortened = pd.DataFrame(election_data, columns = ['stateid',␣
 ↪'turnout', 'age29andunder_pct'])

fig, ax = plt.subplots()

ax.
 ↪scatter(x=election_data_shortened['age29andunder_pct'],y=election_data_shortened['turnout']

ax.set_xlabel('Age 29 and Under')
ax.set_ylabel('Turnout')

for idx, row in election_data_shortened.iterrows():
    ax.annotate(row['stateid'], (row['age29andunder_pct'], row['turnout']) )

plt.show()
```



2(b)

i) Utah is the state with the highest percentage of voters aged 29 and under

ii) Minnesota is the state with the highest voter turnout

iii) Maine is the state with the lowest percentage of voters aged 29 and under

iv) Hawaii is the state with the lowest voter turnout

2(c)

[10]:
```
results = smf.ols('turnout ~ age29andunder_pct', data=election_data).fit()
results.summary()
```

[10]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                turnout   R-squared:                       0.127
Model:                            OLS   Adj. R-squared:                  0.109
Method:                 Least Squares   F-statistic:                     7.005
Date:                Tue, 29 Nov 2022   Prob (F-statistic):             0.0110
Time:                        19:03:16   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
age29andunder_pct     -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 coefficient for the age29andunder_pct variable is -1.0259. The estimated intercept for the age29andunder_pct variable is 99.2205.

2(e)

The estimated intercept tells us that when the percentage of voters aged under 29 is zero, the turnout is 99.2205.

The estimated slope tells us that for every one percent increase in age under 29 (in the state/pop), there is a -1.0259 decrease in voter turnout (in the state/pop). For every one percent increase in age under 29, the voter tunrout changes by minus 1.0259 percentage points.

2(f)

The coefficient on the independent variable is statistically significant at the 0.05 level. We can see this in the p-value of 0.011; since this p-level is less than 0.05, that indicates that it is statistically significant. Additionally, the confidence interval of -1.805 to -0.247 doesn't contain zero, which also indicates that it is statistically significant.

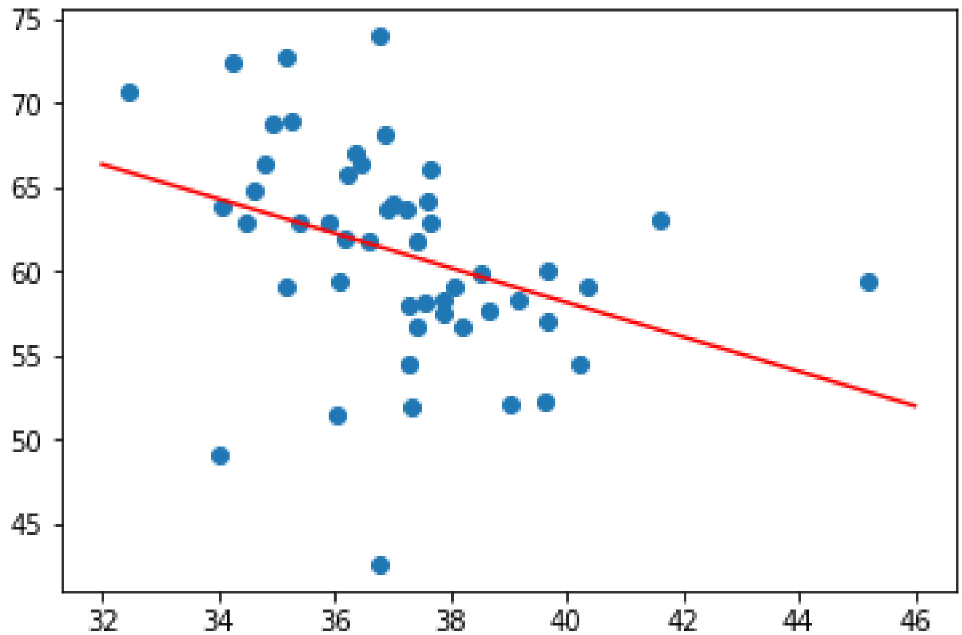
2(g)

R-squared is the percentage of the variation in y explained by the entire model. The percentage of the variance of the outcome explained in this model is 0.127. It is important to note that we are using the regular r-squared value, not the adjusted r-squared value, because we are only working with one independent variable, not two or more.

1.3 Question 3

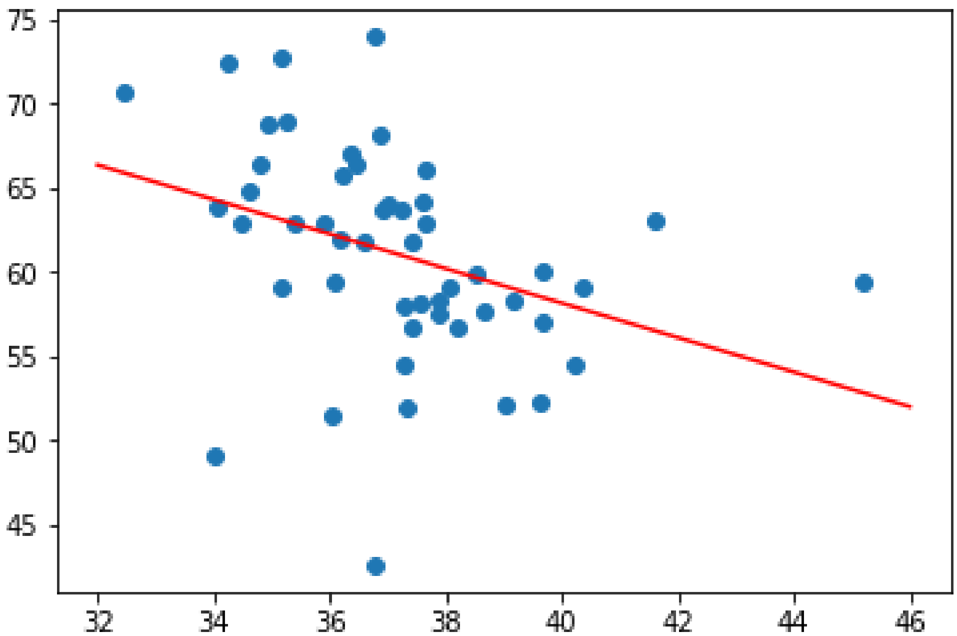
3(a)

[11]:
```
plt.scatter(age29andunder_pct, turnout)
m = -1.0259
b = 99.2005

x = np.linspace(32,46)
plt.plot(x, m*x+b, color = 'red')
plt.show()
```



3(b)

[12]:
```
plt.scatter(age29andunder_pct, turnout)
m = -1.0259
b = 99.2005

x = np.linspace(32,46)
plt.plot(x, m*x+b, color = 'red')
plt.show()

x = 40
prediction_line = (-1.0259*x)+99.2005
print(prediction_line)
```



58.164500000000004

3(c)

[13]:
```
m = -1.0259
b = 99.2005

def turnout_estimate(x):
    return (m*x)+b

#used .iloc[0,4] to retrieve the column; could have also done a row index,␣
 ↪while setting the column to ageunder29_pct,
#but i felt that this way was more elegant

#New York:
new_york_estimate = election_data[election_data['state']=='New York'].iloc[0, 4]
print('New York: ' + str(turnout_estimate(new_york_estimate)))

#Texas:
texas_estimate = election_data[election_data['state']=='Texas'].iloc[0, 4]
print('Texas: ' + str(turnout_estimate(texas_estimate)))

#West Virginia:
west_virginia_estimate = election_data[election_data['state']=='West Virginia'].
 ↪iloc[0, 4]
print('West Virginia: ' + str(turnout_estimate(west_virginia_estimate)))
```

```
New York: 60.824809722528954
Texas: 58.53935011117226
West Virginia: 64.3116358944342
```

3(d)

[14]:
```python
election_data_simplified = pd.DataFrame(election_data, columns = ['state',␣
 ↪'turnout'])
new_york_actual = election_data_simplified[election_data_simplified['state'] ==␣
 ↪'New York']
new_york_actual = election_data[election_data['state']=='New York'].iloc[0, 3]
texas_actual = election_data[election_data['state']=='Texas'].iloc[0, 3]
west_virginia_actual = election_data[election_data['state']=='West Virginia'].
 ↪iloc[0, 3]

new_york_difference =  new_york_actual - new_york_estimate
texas_difference = texas_actual - texas_estimate
west_virginia_difference = west_virginia_actual - west_virginia_estimate

print('New York difference: ' + str(new_york_difference))
print('Texas difference: ' + str(texas_difference))
print('West Virgina difference: ' + str(west_virginia_difference))

print('')
print('New York has the largest difference between estimated and observed␣
 ↪turnout')
```

```
New York difference: 19.24054643608423
Texas difference: 12.579989086863776
West Virgina difference: 15.05925070155702

New York has the largest difference between estimated and observed turnout
```

3(e)

[16]:
```python
m = -1.0259
b = 99.2005

def turnout_estimate(y):
    x = (y-b)/m
    return x

print(turnout_estimate(80))
```

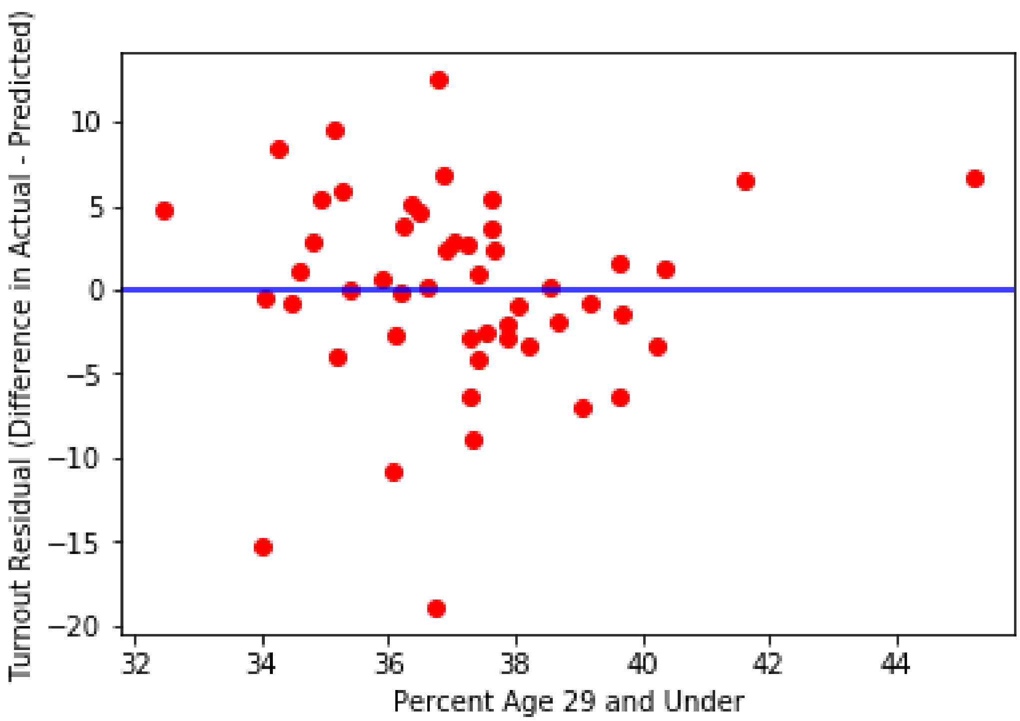
```
18.71576177015304
```

1.4 Question 4

4(a)

[16]:
```
residual_data = election_data
residual_data['predicted_turnout'] =␣
    ↪turnout_estimate(residual_data['age29andunder_pct'])
residual_data['turnout_residual'] = residual_data['turnout'] -␣
    ↪residual_data['predicted_turnout']
plt.scatter(residual_data['age29andunder_pct'],␣
    ↪residual_data['turnout_residual'], color = 'red')

plt.xlabel('Percent Age 29 and Under')
plt.ylabel('Turnout Residual (Difference in Actual - Predicted)')

plt.axhline(0, color = 'blue')
```

[16]:
```
<matplotlib.lines.Line2D at 0x7f2725702d90>
```



Comment on any pattern you observe in the residuals. What does this suggest about where the model overestimates and/or underestimates turnout, given the percentage of a state's population aged 29 and under?

In the observed residuals, I noticed that the values tend to cluster around the 0 line, indicating that they are fairly accurate in their predictions due to the fact that the actual-predicted would be zero. However, even though there are a fair amount of residuals at 0, there are also many points scattered over and under the line, at a pretty equal rate, indicating that the model overestimates the turnout and underestimates the turnout at a fairly similar rate, given the percentage of a state's population aged 29 and under. It seems that the model tends to overestimate the turnout when there is a smaller percentage of people aged 29 and under, and underestimate as the percent age under 29 increases. The outliers reflect the opposite though as the two on the top right are overestimates at a higher 29 and under percent, while the ones in the bottom left are underestimates at a lower percent.

4(b)

[20]:
```
turnout_residual_data = pd.DataFrame(residual_data, columns = ['state',␣
↪'turnout_residual'])
pennsylvania_residual = residual_data[residual_data['state']=='Pennsylvania'].
↪iloc[0, 9]
print('Pennsylvania Residual is: ' + str(pennsylvania_residual))
```

```
Pennsylvania Residual is: 0.009759366098712974
```

4(c)

[20]:
```
turnout_residual_data = pd.DataFrame(residual_data, columns = ['state',␣
↪'turnout_residual'])
highest_residual = turnout_residual_data.max()
lowest_residual = turnout_residual_data.min()
print('State with highest residual: ' + str(highest_residual))
print('State with lowest residual: ' + str(lowest_residual))
```

```
State with highest residual: state               Wyoming
turnout_residual     12.501
dtype: object
State with lowest residual: state               Alabama
turnout_residual   -18.8826
dtype: object
```

1.5 Question 5

5(a)

The respect for persons principle could be violated by the proposed experiment due to the fact that they are not being treated as autonomous subjects. According the the respect for persons principle, the subjects must be given relevant information in a comprehensible format and voluntarily agree to be research subjects, and also know what they are agreeing to. This experiment says that "users will not be informed that an experiment, is taking place," therefore this is violating the respect for persons principle due to the fact that they don't know there is an experiment, and haven't agreed to it (getting notifications). This would harm the subjects because they could unwillingly get notifications and be observed/tracked when they don't want to be.

5(b)

The benefice principle is basically "do no harm," where it is the goal to maximize possible benefits and minimize possible harms. The benefice principle could be violated by the experiment if the risks and benefits weren't possibly assessed/balanced, and the risks outweighed the benefits. For example, if sending notifications about vaccinations to individuals does more harm, due to say, actually annoying and deterring individuals from getting the vaccine, this would do more harm than benefit in society. It is important for the researchers to thoroughly research and make sure the benefits of the study outweigh the risks.

5(c)

The justice principle means that there is a fair distribution of the risks and benefits of research. The justice principle could be violated by the experiment due to the fact that only men are being used as the population subjects, whereas the society as a whole, including men and women, are benefitting. Not only does this violate the principle due to the fact that both men and women are benefitting from the experiment when only men are being subjected to it, but it will likely lead to bias due to the fact that only men's tendencies are being recorded and not women's, so how could that accurately reflect the whole population?

5(d)

The respect for law and public interest principle could be harmed due to the fact that the researchers are collecting a huge amount of data from the population, including if they are currently vaccinated and if vaccination rates increase due to notifications. They have to make sure that this data collection is in compliance with the rules of the USA, and could run into issues if they don't comply. Furthermore, if they aren't transparent in their methods and results, they could run into issues; they need to be transparent in their methods with the subjects and with the society as they share their results.