Econometrics Assignment 3

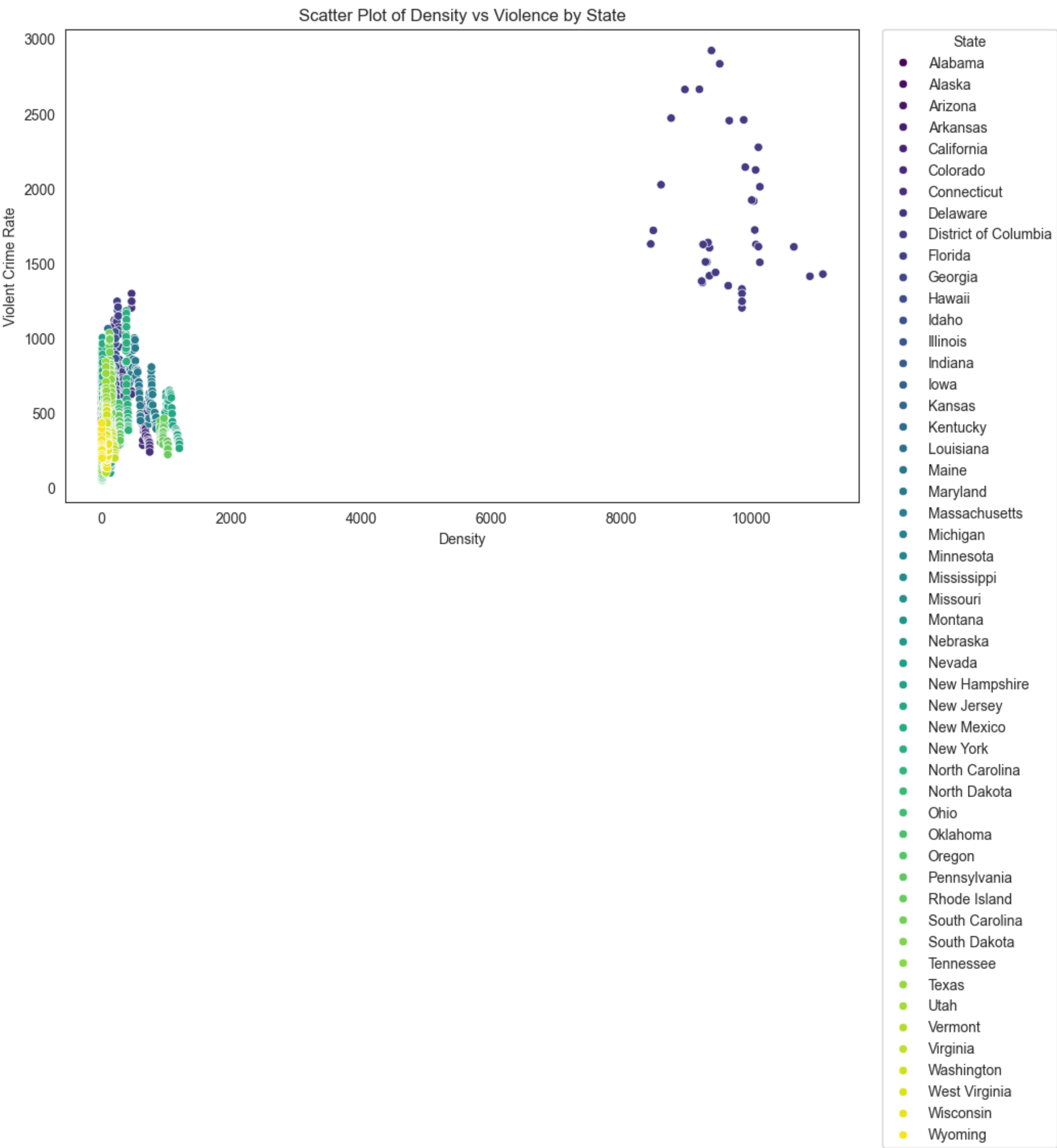
Question 1.1

In [ ]:
```
import numpy as np
import scipy.stats as sp
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import statsmodels.formula.api as smf
import statsmodels.api as sm

crime = pd.read_csv("C:/Users/shale/Downloads/assignment5_part1.csv")
crime.head()
```

Out[ ]:

	state	year	stateid	vio	mur	rap	aga	bur
0	Alabama	1977	1	414.444444	14.200542	25.176152	278.265583	1135.528455
1	Alabama	1978	1	419.080706	13.335115	25.494388	281.159808	1229.315874
2	Alabama	1979	1	413.319183	13.159989	27.513929	263.146723	1287.264526
3	Alabama	1980	1	448.534313	13.181522	29.988611	273.238195	1526.674066
4	Alabama	1981	1	470.454545	11.874362	26.072523	306.052094	1450.740552

In [ ]:
```
plt.figure(figsize=(10, 6))
scatter_plot = sns.scatterplot(x='density', y='vio', hue='state', data=crime, pa
scatter_plot.legend(loc='upper right', bbox_to_anchor=(1.25, 1), borderaxespad=0
plt.title('Scatter Plot of Density vs Violence by State')
plt.xlabel('Density')
plt.ylabel('Violent Crime Rate')
plt.show()

highest_density_zscore = crime['density'].mean() + 3 * crime['density'].std()
highest_vio_zscore = crime['vio'].mean() + 3 * crime['vio'].std()

outlier_states_density = crime[crime['density'] > highest_density_zscore]['state
outlier_states_vio = crime[crime['vio'] > highest_vio_zscore]['state'].unique()

outlier_states = set(outlier_states_density) | set(outlier_states_vio)

outlier_states_str = ', '.join(outlier_states)

print("The outlier state is the " + outlier_states_str)
```



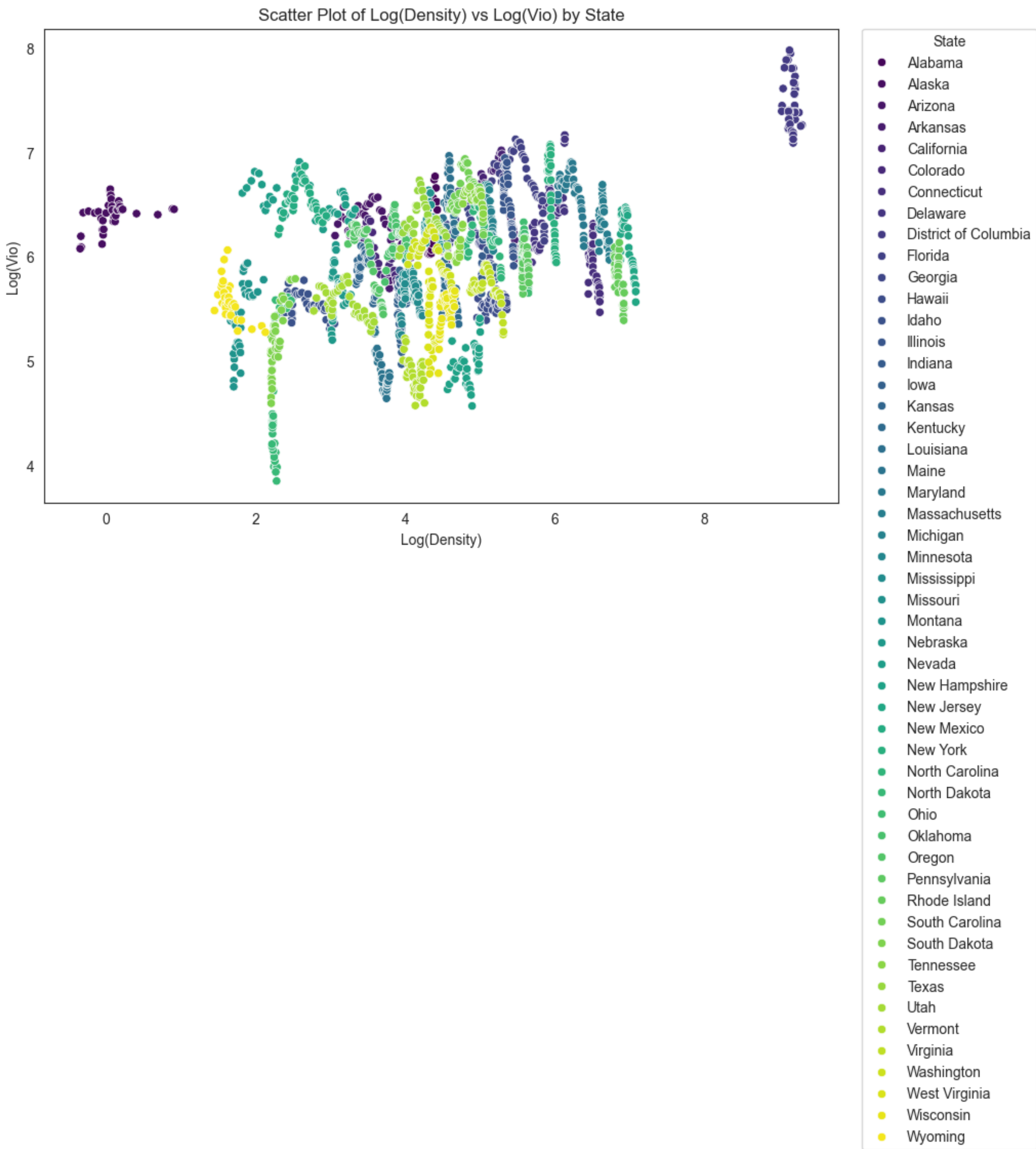

```
The outlier state is the District of Columbia
```

The scatterplot shows many vertical clusters, which indicates a large variety of the rate of violent crime for a given state density. Apart from the outlier state, the states seem to reach a maximum of around 1500 incidents per 100,000 members of the population. It is difficult to observe any other particular patterns or relationships with this scatterplot.

The state that stands out relative to the rest is the District of Columbia, which has a higher rate of violent crime and density relative to the other states.

In [ ]:
```
crime['log_density'] = np.log(crime['density'])
crime['log_vio'] = np.log(crime['vio'])

# Scatter plot of log(density) vs log(vio) with color varying by state
plt.figure(figsize=(10, 6))
scatter_plot = sns.scatterplot(x='log_density', y='log_vio', hue='state', data=c

# Adjust the legend size
scatter_plot.legend(loc='upper right', bbox_to_anchor=(1.25, 1), borderaxespad=0

plt.title('Scatter Plot of Log(Density) vs Log(Vio) by State')
```

```
plt.xlabel('Log(Density)')
plt.ylabel('Log(Vio)')
plt.show()
```



Transforming the variables is advantageous in the visualization and further analysis, as it allows us to observe nonlinear relationships between variables. In this case, the graph above shows much clearer patterns and less clusters in the data.

Question 1.2

In [ ]:
```
crime_pd = crime.set_index(['state','year'])
import linearmodels as lm
crime_pool = lm.PooledOLS.from_formula('log_vio~shall + 1',
                                        data = crime_pd).fit()
print(crime_pool)
```

```
                          PooledOLS Estimation Summary
================================================================================
Dep. Variable:                log_vio   R-squared:                        0.0687
Estimator:                  PooledOLS   R-squared (Between):              0.1172
No. Observations:                1938   R-squared (Within):              -0.2159
Date:                Tue, Nov 28 2023   R-squared (Overall):              0.0687
Time:                        19:08:49   Log-likelihood                   -1672.9
Cov. Estimator:            Unadjusted
                                        F-statistic:                      142.85
Entities:                          51   P-value                           0.0000
Avg Obs:                       38.000   Distribution:                  F(1,1936)
Min Obs:                       38.000
Max Obs:                       38.000   F-statistic (robust):             142.85
                                        P-value                           0.0000
Time periods:                      38   Distribution:                  F(1,1936)
Avg Obs:                       51.000
Min Obs:                       51.000
Max Obs:                       51.000

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
Intercept      6.1214     0.0177     346.31     0.0000      6.0868      6.1561
shall         -0.3149     0.0264    -11.952     0.0000     -0.3666     -0.2633
==============================================================================
```

The coefficient of the 'shall' variable is -0.3149. It has a p-value that is very close to zero, and is therefore statistically significant. The interpretation of this coefficient is that an increase in 0.1 in shall will result in a 3.149% decrease in the violent crime rate on average ceteris paribus. We interpret using a 0.1 or 10% change in shall as it would make less sense to interpret the change as a result of an increase by a factor of one as the variable is defined as a fraction.

Shall-issue concealed carry laws compell local authorities to issue a gun permit if the citizen passes basic requirements. Therefore, this model would support the idea that shall carry laws, which lead to an increase to more guns leads to decreased violent crime rates.

Question 1.3

In [ ]:
```
model = lm.PanelOLS.from_formula("log_vio ~ shall + EntityEffects + TimeEffects"
data = crime_pd)
crime_fte = model.fit(cov_type='clustered', cluster_entity=True)
print(crime_fte)
```

```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:                log_vio   R-squared:                        0.0015
Estimator:                   PanelOLS   R-squared (Between):              0.0041
No. Observations:                1938   R-squared (Within):              -0.0074
Date:                Tue, Nov 28 2023   R-squared (Overall):              0.0041
Time:                        19:08:49   Log-likelihood                    385.70
Cov. Estimator:             Clustered
                                        F-statistic:                      2.8245
Entities:                          51   P-value                           0.0930
Avg Obs:                       38.000   Distribution:                  F(1,1849)
Min Obs:                       38.000
Max Obs:                       38.000   F-statistic (robust):             0.2939
                                        P-value                           0.5878
Time periods:                      38   Distribution:                  F(1,1849)
Avg Obs:                       51.000
Min Obs:                       51.000
Max Obs:                       51.000

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
shall          0.0282     0.0519     0.5421     0.5878     -0.0737      0.1300
==============================================================================

F-test for Poolability: 156.59
P-value: 0.0000
Distribution: F(87,1849)

Included effects: Entity, Time
```

The coefficient of shall in the model is 0.0282, which indicates that states implementing shall-issue laws over a longer period of time results in a higher violent crime rate, on average ceteris paribus. However, the coefficient has a p-value of 0.5878 and is not statistically significant. The coefficient of shall changes by about 6.6 standard deviations relative to the pooled model, which is a significant change.

The change is most likely as a consequence of the fact that the previous model was a pooled regression, which assumes that the relationship between variables are constant across entities and time, and this model is a fixed effects regression, which accounts for individual and time effects. In this case, the aforementioned individual and time effects may have made the coefficient no longer statistically significant because of the resultant change to the coefficient of the shall variable in the fixed effects regression.

Question 1.4

In [ ]:
```python
model = lm.PanelOLS.from_formula("log_vio ~ shall + EntityEffects + TimeEffects
                                 data = crime_pd)
crime_fte = model.fit(cov_type='clustered', cluster_entity=True)
print(crime_fte)
```

```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:                log_vio   R-squared:                        0.1351
Estimator:                   PanelOLS   R-squared (Between):             -0.0126
No. Observations:                1938   R-squared (Within):              -0.0634
Date:                Tue, Nov 28 2023   R-squared (Overall):             -0.0127
Time:                        19:08:50   Log-likelihood                    524.87
Cov. Estimator:             Clustered
                                        F-statistic:                      31.954
Entities:                          51   P-value                           0.0000
Avg Obs:                       38.000   Distribution:                  F(9,1841)
Min Obs:                       38.000
Max Obs:                       38.000   F-statistic (robust):             3.8307
                                        P-value                           0.0001
Time periods:                      38   Distribution:                  F(9,1841)
Avg Obs:                       51.000
Min Obs:                       51.000
Max Obs:                       51.000

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
shall         -0.0317     0.0501    -0.6319     0.5276     -0.1300      0.0666
rpcpi      -1.627e-05  2.029e-05    -0.8018     0.4228  -5.607e-05   2.353e-05
rpcui         -0.0014     0.0004    -3.0973     0.0020     -0.0023     -0.0005
rpcim          0.0003     0.0005     0.5671     0.5707     -0.0007      0.0012
density       -0.0004     0.0002    -1.7746     0.0761     -0.0007   3.714e-05
pbm1019        0.0628     0.1098     0.5722     0.5672     -0.1524      0.2781
pbm2029        0.2200     0.1094     2.0117     0.0444      0.0055      0.4345
pwm1019       -0.0858     0.0412    -2.0824     0.0374     -0.1665     -0.0050
pwm2029        0.1099     0.0442     2.4859     0.0130      0.0232      0.1966
==============================================================================

F-test for Poolability: 106.23
P-value: 0.0000
Distribution: F(87,1841)

Included effects: Entity, Time
```

In this model, shall has a coefficient of -0.0317, which indicates that shall-issue gun laws reduce violent crime rates on average ceteris paribus. However, it has a p-value of 0.5276, which means it is not statistically significant. This coefficient is not significantly different from the previous fixed effects regression, as it is around 1.2 standard deviations different from the previous coefficient.

However, this model does show an opposite effect of shall-issue laws relative to the previous model. This may have been caused by the inclusion of new variables into this model, indicating that the previous model was suffering from omitted variable bias. It is significantly different from the first model's coefficient, with about 5.6 standard errors of difference. This mostly is because both this model and the previous are fixed effects regressions, whereas the first model is a pooled regression.

Question 2.1

In [ ]:
```
unem = pd.read_csv("C:/Users/shale/Downloads/assignment5_part2.csv")
```

```
unem.head()
```

Out[ ]:

	age	race	earnwke	employed	unemployed	married	union	ne_states	so_states
0	53	1	NaN	1	0	1	0	0	0
1	39	1	NaN	1	0	1	0	0	0
2	41	1	500.0	1	0	1	0	0	1
3	27	1	520.0	1	0	1	0	0	1
4	29	3	615.0	1	0	0	0	0	1

5 rows × 21 columns

In [ ]:

```
unem = unem[~((unem['employed'] == 0) & (unem['unemployed'] == 0))]

total_participants = unem.shape[0]
percent_employed = unem['employed'].sum() / total_participants
percent_unemployed = unem['unemployed'].sum() / total_participants

data = {'Employed': [percent_employed],
        'Unemployed': [percent_unemployed]}
df = pd.DataFrame(data)
print(df)
```

```
   Employed  Unemployed
0  0.951979    0.048021
```

This indicates that 95.1979% of people who were employed in April 2008 were still employed in April 2009, and 4.8021% became unemployed.

Question 2.2

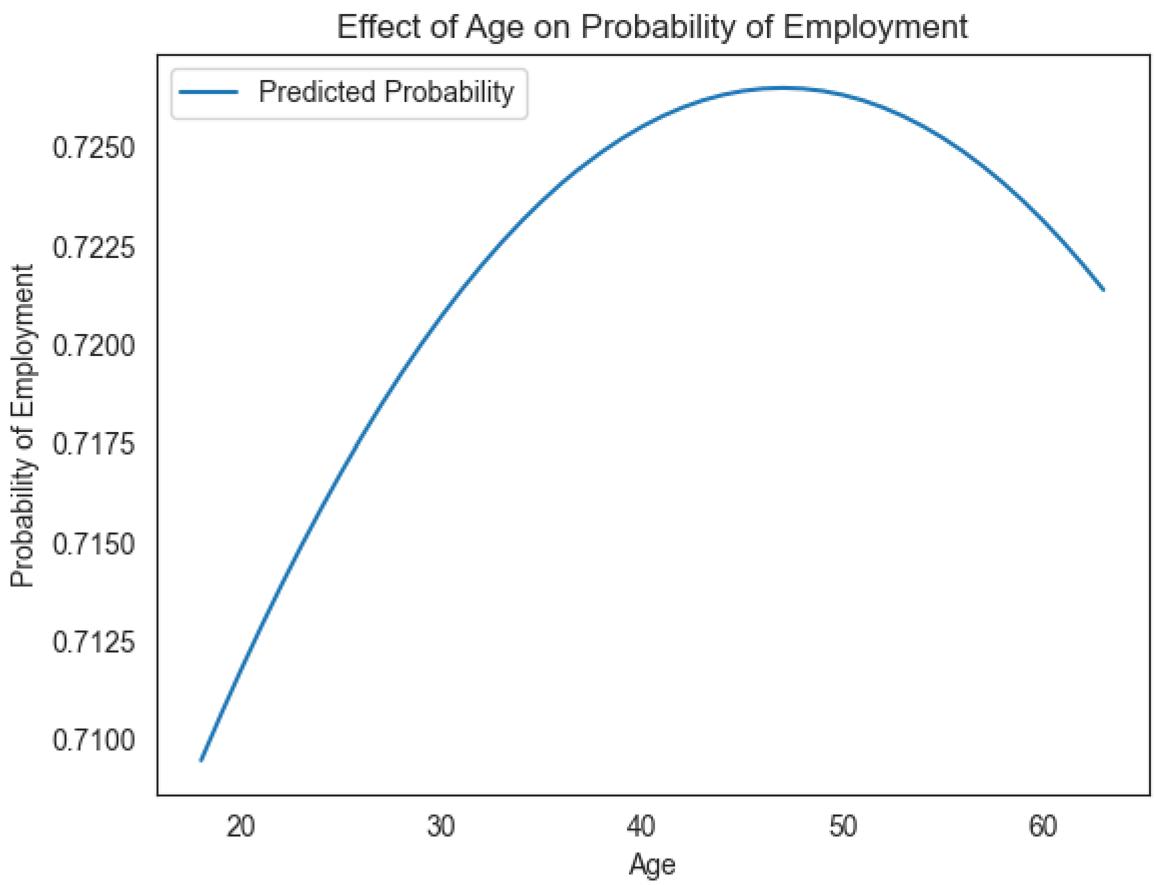
In [ ]:

```
# Regression equation
model = 'employed ~ age + I(age**2)'

# OLS regression
fit_ols = sm.OLS.from_formula(model, data=unem).fit()

# Heteroskedasticity-robust standard errors
fit_robust = fit_ols.get_robustcov_results(cov_type='HC1')

# Display results
print(fit_robust.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               employed   R-squared:                       0.005
Model:                            OLS   Adj. R-squared:                  0.005
Method:                 Least Squares   F-statistic:                     8.479
Date:                Tue, 28 Nov 2023   Prob (F-statistic):           0.000211
Time:                        19:08:50   Log-Likelihood:                 629.34
No. Observations:                4977   AIC:                            -1253.
Df Residuals:                    4974   BIC:                            -1233.
Df Model:                           2
Covariance Type:                  HC1
================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept       0.7560      0.049     15.524      0.000       0.661       0.852
age             0.0094      0.002      3.982      0.000       0.005       0.014
I(age ** 2)    -0.0001   2.74e-05     -3.817      0.000      -0.000   -5.08e-05
==============================================================================
Omnibus:                     4146.258   Durbin-Watson:                   1.962
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            65846.244
Skew:                          -4.195   Prob(JB):                         0.00
Kurtosis:                      18.720   Cond. No.                     2.77e+04
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 2.77e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [ ]:

```
# Sample coefficients based on your provided results
coeff_intercept = 0.7560
coeff_age = 0.0094
coeff_age_squared = -0.0001

# Generate a range of ages for prediction
ages = np.arange(unem['age'].min(), unem['age'].max() + 1, 1)

# Calculate the predicted probabilities based on the logistic function
log_odds = coeff_intercept + coeff_age * ages + coeff_age_squared * ages**2
predicted_probabilities = 1 / (1 + np.exp(-log_odds))

# Plot the effect of age on the probability of employment
plt.plot(ages, predicted_probabilities, label='Predicted Probability')
plt.xlabel('Age')
plt.ylabel('Probability of Employment')
plt.title('Effect of Age on Probability of Employment')
plt.legend()
plt.show()
```



This graph indicates that there is a nonlinear effect of age on the probability of being employed, as the initially positive effect of age decreases and eventually becomes negative. The regression results above also show that both age and age squared are statistically significant, further validating this claim.

The effect of increasing age by 1 depends on the initial age of the person, as there is a diminishing effect. In general, the effect can be represented as:

employed = 0.0094 - 0.0002 * age

Where 'age' represents the age of the person in question, on average ceteris paribus.

Question 2.3

In [ ]:
```
# Logit model for employed conditional on specified variables
logit = sm.Logit.from_formula('employed ~ age + I(age**2) + earnwke + C(race) +
logit_result = logit.fit(cov_type="HC1")

# Display regression results
print(logit_result.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.196305
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:               employed   No. Observations:                 4407
Model:                          Logit   Df Residuals:                     4391
Method:                           MLE   Df Model:                           15
Date:                Tue, 28 Nov 2023   Pseudo R-squ.:                 0.04201
Time:                        19:08:50   Log-Likelihood:                -865.11
converged:                       True   LL-Null:                       -903.05
Covariance Type:                  HC1   LLR p-value:                 3.938e-10
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept        0.1808   4.34e+07   4.16e-09      1.000   -8.51e+07    8.51e+07
C(race)[T.2]    -0.3775      1.065     -0.354      0.723      -2.465       1.710
C(race)[T.3]     0.0113      2.229      0.005      0.996      -4.358       4.381
age              0.1193      0.303      0.394      0.694      -0.474       0.713
I(age ** 2)     -0.0014      0.004     -0.400      0.689      -0.008       0.006
earnwke          0.0001      0.000      0.600      0.549      -0.000       0.001
married          0.3088      0.156      1.984      0.047       0.004       0.614
female           0.4326      0.172      2.516      0.012       0.096       0.770
ne_states        0.2087   7.76e+06   2.69e-08      1.000   -1.52e+07    1.52e+07
so_states        0.0395   3.54e+06   1.11e-08      1.000   -6.95e+06    6.95e+06
ce_states        0.1611   3.91e+06   4.12e-08      1.000   -7.67e+06    7.67e+06
we_states       -0.2285   3.18e+06  -7.18e-08      1.000   -6.24e+06    6.24e+06
educ_lths       -0.7656    2.2e+07  -3.49e-08      1.000    -4.3e+07     4.3e+07
educ_hs         -0.2633        nan        nan        nan         nan         nan
educ_somecol     0.1723   3.51e+07   4.91e-09      1.000   -6.87e+07    6.87e+07
educ_aa          0.2772   3.43e+07   8.07e-09      1.000   -6.73e+07    6.73e+07
educ_bac         0.1543        nan        nan        nan         nan         nan
educ_adv         0.6059        nan        nan        nan         nan         nan
================================================================================
```

The variables that are most relevant to the probability of being employed are whether the person is married and their gender. This is because these are the only variables in the model that are statistically significant at the 5% level.

As this is a logistic regression, the actual values of the coefficients cannot be interpreted directly, but rather we can claim that there is a positive effect on the probability of being employed if the person is married (relative to being single) or if the person is female (relative to being male) on average ceteris paribus.

Question 2.4

In [ ]:
```python
# Logit model for employed conditional on specified variables
logit = sm.Logit.from_formula("unemployed ~ age + I(age ** 2) + earnwke + race +
logit_result = logit.fit(cov_type="HC1")

# Display regression results
print(logit_result.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.196569
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:             unemployed   No. Observations:                 4407
Model:                          Logit   Df Residuals:                     4392
Method:                           MLE   Df Model:                           14
Date:                Tue, 28 Nov 2023   Pseudo R-squ.:                 0.04072
Time:                        19:08:50   Log-Likelihood:                -866.28
converged:                       True   LL-Null:                       -903.05
Covariance Type:                  HC1   LLR p-value:                 4.382e-10
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept       -0.2634    2.5e+13  -1.06e-14      1.000   -4.89e+13    4.89e+13
age             -0.1164      0.233     -0.499      0.617      -0.573       0.340
I(age ** 2)      0.0014      0.003      0.511      0.609      -0.004       0.007
earnwke         -0.0001      0.000     -0.634      0.526      -0.001       0.000
race             0.0836      0.114      0.732      0.464      -0.140       0.307
married         -0.3291      0.237     -1.386      0.166      -0.794       0.136
female          -0.4287      0.150     -2.863      0.004      -0.722      -0.135
ne_states       -0.2260    2.5e+13  -9.05e-15      1.000   -4.89e+13    4.89e+13
so_states       -0.0289    2.5e+13  -1.16e-15      1.000   -4.89e+13    4.89e+13
ce_states       -0.1788    2.5e+13  -7.16e-15      1.000   -4.89e+13    4.89e+13
we_states        0.1703    2.5e+13   6.82e-15      1.000   -4.89e+13    4.89e+13
educ_lths        0.7580        nan        nan        nan         nan         nan
educ_hs          0.2561   6.92e+06     3.7e-08      1.000   -1.36e+07    1.36e+07
educ_somecol    -0.1752   4.96e+06   -3.53e-08     1.000   -9.73e+06    9.73e+06
educ_aa         -0.2965   3.08e+06   -9.62e-08     1.000   -6.04e+06    6.04e+06
educ_bac        -0.1780   4.14e+06   -4.29e-08     1.000   -8.12e+06    8.12e+06
educ_adv        -0.6279   2.88e+06   -2.18e-07     1.000   -5.65e+06    5.65e+06
================================================================================
```

The only statistically significant variable at the 5% level is female. Again as this is a logistic regression, the interpretation of the effect would be that there is a negative effect on the probability of being unemployed if the person was female (relative to being male) on average ceteris paribus.

Question 2.5

In [ ]:
```
model = sm.Logit.from_formula("employed ~ age + I(age ** 2) + earnwke + race + m
results = model.fit(cov_type="HC1")

df = pd.DataFrame({
    'age': [35],
    'age_squared': [35 ** 2],
    'earnwke': [865],
    'race': [1],
    'married': [1],
    'female': [1],
    'ne_states': [1],
    'so_states': [0],
    'ce_states': [0],
    'we_states': [0],
    'educ_lths': [0],
    'educ_hs': [0],
    'educ_somecol': [0],
    'educ_aa': [0],
```

```
    'educ_bac': [1],
    'educ_adv': [0]})

predicted_prob = results.predict(df)
print(f"Predicted probability of being employed: {predicted_prob.iloc[0]}")
```

```
Optimization terminated successfully.
         Current function value: 0.196569
         Iterations 8
Predicted probability of being employed: 0.9781974451821999
```

In [ ]:
```
model = sm.Logit.from_formula("employed ~ age + I(age ** 2) + earnwke + race + m
results = model.fit(cov_type="HC1")

df = pd.DataFrame({
    'age': [35],
    'age_squared': [35 ** 2],
    'earnwke': [865],
    'race': [1],
    'married': [0],
    'female': [0],
    'ne_states': [1],
    'so_states': [0],
    'ce_states': [0],
    'we_states': [0],
    'educ_lths': [0],
    'educ_hs': [0],
    'educ_somecol': [0],
    'educ_aa': [0],
    'educ_bac': [1],
    'educ_adv': [0]})

predicted_prob = results.predict(df)
print(f"Predicted probability of being employed: {predicted_prob.iloc[0]}")
```

```
Optimization terminated successfully.
         Current function value: 0.196569
         Iterations 8
Predicted probability of being employed: 0.9546034784370414
```

The probability of the first individual being employed is around 97.82% on average ceteris paribus.

The probability of being employed changes to 95.54% on average ceteris paribus if the individual is changed to be an unmarried male. This would indicate that being female and being married has a positive effect on the probability of being employed in April 2009.