PLSC 597 - Homework 1

Maya Dalton (9/11/23)

Please find my Jupyter Notebook file and data file for this assignment here: https://github.com/mad6821/PLSC-597---Machine-Learning

Bauhr, M., & Charron, N. (2021). Will Women Executives Reduce Corruption? Marginalization and Network Inclusion. Comparative Political Studies, 54(7), 1292–1322. https://doi.org/10.1177/0010414020970218

With a wide plethora of research investigating the relationship between women's representation in government and corruption levels, Bauhr and Charron add to this literature by focusing on (1) local-level representation in the form of female mayors, (2) if women's representation diminishes corruption levels, and (3) if these effects last over time. The authors utilize a simple OLS regression model, along with a regression discontinuity design and a first-differences design to investigate the direction and magnitude of the relationship between women's representation and corruption in French municipalities. The authors find that women mayors, in fact, reduce corruption risks, with newly elected female mayors primarily driving this relationship. Interestingly, when women mayors are re-elected, they find negligible differences in corruption levels. The replication archive with the data and code for this article can be found here: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FOSMRN. The authors are very straightforward in their research design goals. They want to uncover causal effects of electing female mayors on corruption risks. To do so, they primarily utilize the regression discontinuity design (RDD), but also employ a simple OLS regression. While there is no discussion of why this OLS regression appears in their results, it seems to be utilized as a baseline or benchmark to compare with the RDD. Overall, their modeling goal is explanatory, as they apply statistical models to data for testing causal hypotheses about theoretical constructs of gender and corruption. They utilize a variety of covariates, such as economic development, population density, voter turnout, wages, number of parties competing in elections, rounds of elections, and incumbency.

If I were to propose a change to the regression model, I would remove some additional covariates. Specifically, I would remove the variable controlling for the level of income inequality in the municipality, measured by the wages of an average worker. Additionally, I would remove the control for the number of total and commercial only registered firms in each municipality, as well as the re-election variable. This deicsion is due to the fact that these variables not only are statistically significant, but they also do not have a strong effect on the dependent variable, as shown below in the regression table.

```
# Importing libraries, storing as shorthand
import numpy as np        #base package
import pandas as pd       #working with data frames
import matplotlib.pyplot as plt   #plotting library
from sklearn.model_selection import train_test_split  #standard machine learning library
from sklearn.metrics import mean_squared_error
from sklearn.linear_model import LinearRegression

# Path to the .dta file
file_path = "/storage/home/mad6821/work/Machine Learning/Homework/Bauhr and Charron CPS data rd estimat

# Read the .dta file into a pandas DataFrame
df = pd.read_stata(file_path)
df.fillna(df.mean(), inplace=True)

# Fit regression model
model = sm.OLS.from_formula("rev_sb_elec_periodw ~ fem_win08 + logPopDens + re_elect + year + turnout_r
result = model.fit()

# Print the summary of the regression results
print(result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     rev_sb_elec_periodw   R-squared:                       0.020
Model:                             OLS   Adj. R-squared:                  0.012
Method:                  Least Squares   F-statistic:                     2.419
Date:                 Mon, 04 Sep 2023   Prob (F-statistic):            0.00754
Time:                         21:07:00   Log-Likelihood:                 45.039
No. Observations:                 1195   AIC:                            -68.08
Df Residuals:                     1184   BIC:                            -12.13
Df Model:                           10
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.1092      0.093      1.172      0.241      -0.074       0.292
fem_win08     -0.0301      0.015     -2.072      0.039      -0.059      -0.002
logPopDens     0.0083      0.007      1.278      0.202      -0.004       0.021
re_elect       0.0007      0.015      0.048      0.962      -0.029       0.030
year           0.0457      0.015      3.137      0.002       0.017       0.074
turnout_r1     0.0013      0.001      1.277      0.202      -0.001       0.003
rounds         0.0093      0.015      0.620      0.535      -0.020       0.039
ave_highEd    -0.0020      0.001     -2.236      0.026      -0.004      -0.000
ave_t_wage  5.371e-07    1.29e-05      0.042      0.967   -2.47e-05    2.58e-05
no_firms    1.238e-05    1.69e-05      0.732      0.464   -2.08e-05    4.55e-05
com_firm    -6.98e-05      0.000     -0.478      0.632      -0.000       0.000
==============================================================================
Omnibus:                      360.813   Durbin-Watson:                   2.025
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              920.134
Skew:                           1.601   Prob(JB):                    1.57e-200
Kurtosis:                       5.868   Cond. No.                     3.37e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.37e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Therefore, I will present an assessment of the out-of-sample predictive performance of the same model as Bauhr and Charron's model, with income inequality (wages) removed performs better than that of the original model.

```python
# Split the data
train_data, test_data = train_test_split(df, test_size=0.3, random_state=1234)

# Independent variables
independent_vars = ['fem_win08', 'logPopDens', 'year', 'turnout_r1', 'rounds', 'ave_highEd']
covariates = ['ave_t_wage'] + ['no_firms'] + ['com_firm'] + ['re_elect']

# Model with all covariates
X_train_original = train_data[independent_vars + covariates]
Y_train = train_data['rev_sb_elec_periodw']

model_original = LinearRegression().fit(X_train_original, Y_train)

X_test_original = test_data[independent_vars + covariates]
```

```python
Y_test = test_data['rev_sb_elec_periodw']

predictions_original = model_original.predict(X_test_original)
mse_original = mean_squared_error(Y_test, predictions_original)

# Model without covariates
X_train_alt = train_data[independent_vars]

model_alt = LinearRegression().fit(X_train_alt, Y_train)

X_test_alt = test_data[independent_vars]

predictions_alt = model_alt.predict(X_test_alt)
mse_alt = mean_squared_error(Y_test, predictions_alt)

# Compare the MSE values
print("MSE with original mode:", mse_original)
print("MSE with removed covariates:", mse_alt)
```

```
MSE with original mode: 0.05949942073067781
MSE with removed covariates: 0.05940828578882852
```

As shown above, the mean squared error after removing covariates is smaller than the mean squared error of the original model. This shows that throwing in covariates for any potential cofounder is not always a good decision when deciding on a statistical model in our research. Wages as a measure of income inequality, the total number of firms, the number of commericial only firms, and a re-election variable did not provide much to the original regression model, and were not statistically significant, therefore the predicitive performance of a model without them is stronger.