5304 Econometrics Autumn 2023[1]

Problem set 2: Group 15

Part 1: OLS and bad controls in practice

Trump voting and unemployment

1. Clean the data and generate variables in exactly the same way as you did in PS1.

2. Regress the Trump voter dummy on not working in the last week. Interpret the coefficient.

   The OLS estimate of the coefficient on the binary variable for not working in the last week is 0.018 with standard error 0.0136 and corresponding t-statistic of 1.34. We therefore do not reject the null of $\hat{\beta}_{notworking} = 0$ at any conventional significance level. If the effect was statistically significant, we would interpret it as the probability of voting for Trump being 1.8 percentage points higher for people that have not worked during the last week.

3. Add age as a control variable. Why do you think the coefficient on not working changes?

   The coefficient on not working changes from 0.018 to -0.037 and statistically significant at the 5% significance level when controlling for age. We use the OVB formula for interpretation. The estimated coefficient of the effect of age on Trump is statistically significant and positive, i.e. the probability of voting for Trump is estimated to be higher for older people. By assuming a positive relationship between age and not working (retirement/other health reasons make older people are less likely to work) we would expect a positively biased estimate in the short regression when age is omitted. This is in line with the results: $0.018 > -0.037$.

4. In PS1 you looked at the relationship between Trump voting and age, which turned out not to be linear. Report specifications that control for 1) a quadratic in age and 2) quintiles of age. Interpret what you find in terms of the functional form of age.

   When age is controlled for in quadratic form, the probability of voting for Trump is estimated as follows:

   $$Trump_i = 0.0952043 - 0.0244552notworking_i + 0.0103045age_i - 0.0000661age_i^2.$$

   In terms of the functional form of age both coefficients are be statistically significant and the resulting effect of age was estimated to have a concave shape. Age is estimated to increase the probability of voting (linear coefficient positive) for Trump but with a diminishing rate (quadratic coefficient negative).When controlling for quintiles of age, the probability of voting for Trump is estimated as:
   $Trump_i = 0.0267175notworking_i + 0.3468473ageQuintile1_i + 0.3987228agequintile2_i + 0.4776708agequintile3_i + 0.4753438agequintile4_i + 0.5022264agequintile5_i$,
   where $agequintileK_i$ indicates whether individual i belongs to age quintile K.
   In terms of the functional form of age all coefficients are statistically significant. Age is again positively associated with trump voting, specifically, the first two quintile coefficient confidence intervals both include 37%, while the last three all include 47% probability of voting for Trump. Note that the 95% CI for the not working effect contains zero for both the quadratic and quintile age model specifications.

[1] 42624@student.hhs.se, 42613@student.hhs.se, 42632@student.hhs.se, @student.hhs.se

5. How many percentage points more likely are the second oldest quintile (4th) of respondents to vote Trump compared to the second youngest (2nd), conditional on the not working dummy? Test the hypothesis that the share of Trump voters in these age groups are equal and report the p-value (use robust standard errors). Do the same for the 5th versus the 1st age quintiles (for very small p-values, interpret the value as virtually zero).

   The second oldest quintile of respondents is estimated to be 7.7 percentage points more likely to vote for Trump compared to the second youngest when controlling for people not working. When testing the null hypothesis $H_0 : \gamma_2 = \gamma_4$ for the model

$$Trump_i = \beta notWorking_i + \gamma_1 ageQuintile1_i + \gamma_2 ageQuintile2_i +$$

$$+\gamma_3 ageQuintile3_i + \gamma_4 ageQuintile4_i + \gamma_5 ageQuintile5_i + \epsilon_i$$

   the P-value is 0.0005. Hence we reject the null hypothesis at any conventional significance level and have a strong evidence to support the hypothesis that coefficients for the voters in age quintiles 2 and 4 are not equal.

   The oldest quintile of respondents is estimated to be 15.5 percentage points more likely to vote for Trump compared to the youngest when controlling for people not working. When testing the null hypothesis $H_0 : \gamma_1 = \gamma_5$ for the model

$$Trump_i = \beta notworking_i + \gamma_1 agequintile1_i + \gamma_2 agequintile2_i +$$

$$+\gamma_3 agequintile3_i + \gamma_4 agequintile4_i + \gamma_5 agequintile5_i + \epsilon_i$$

   the P-value is virtually zero. Hence we can reject the null hypothesis and have even stronger reason to believe that the voters in age quintiles 1 and 5 are not equal.

6. Choose one of the specifications including a non-linear function of age and include the dummy for being white. Write down the regression specification, and run the regression in Stata. Does the estimated coefficient on not working change, and if so, what does the change imply for the correlation between being white and not working (conditional on age)? Comment on the sign of the correlation and interpreting what this means.

   We have choosed the following regression specification:

$$Trump_i = \beta_0 + \beta_1 notworking_i + \beta_2 white + \beta_3 age + \beta_4 age^2 + \epsilon_i$$

   The estimated coefficient on the not working variable remained not statistically significant (p-value of 0.49). Focusing on the change in the point estimate, it increased from -0.0245 to -0.0106 by including the white dummy when compared to controlling only for age with the quadratic specification model. Since the effect associated with being white on the probabilitiy of voting for Trump was estimated to be statistically significant and positive ($\hat{\beta}_2 = 0.2417$), the OVB formula implies negative sign of the correlation between the not working a white indicator variables. Suggesting that white people are less likely to not have worked during the last week.

Trump voting and unemployment

1. Regress Trump voting on the higher education dummy. Why is the estimated coefficient unlikely to have a causal interpretation? Mention several possible sources of bias.

First, the relationship between Trump voting and higher education may be influenced by unobservable variables not included in the model. For example, cultural or regional factors that are correlated with both Trump voting and educational attainment could lead to bias in the coefficient estimate. Other factors that are correlated with both Trump voting and educational attainment, such as income, race, or urban vs. rural residence, can introduce bias if not properly controlled for.

Further, higher education is not randomly assigned to individuals, so individuals who choose to pursue higher education may have different characteristics than those who do not. This self-selection can lead to endogeneity, as individuals' preferences and abilities can influence both their educational choices and their political preferences.

2. Let's disregard the problems with omitted variable bias you have just brought up. A friend claims that an interesting link between higher education and Trump voting is that education makes people more interested in factual policy, which makes them less likely to vote for Trump no matter their political views. However, higher education may also affect people's political views directly due to them being in a particular social milieu. Your friend asks you to isolate the 'factual policy' channel between education and Trump voting, holding any potential effect via political ideology constant. Regress Trump voting on high education and a dummy variable that equals one if feeling towards conservatives exceeds 50 and zero otherwise, as a proxy for being conservative.

Bachelor or higher education	-0.0981***
	(0.0112)
Feeling mostly towards conservatives	0.571***
	(0.0113)
Observations	5363
R-squared	0.357

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3. Why is the specification estimated in the previous question problematic (again, disregarding the problems with omitted factors that you have already discussed)? Explain.

Using a proxy as a control can weaken the causal interpretation of the relationship. In the context our question, if the proxy variable (feeling towards conservatives) is affected by education itself, it becomes an endogenous variable, which complicates the interpretation of causality. The Conditional Independence Assumption says $\{Y_{0i}, Y_{1i}\} \perp D_i|X$, i.e. potential outcomes $\{Y_{0i}, Y_{1i}\}$ are independent of the treatment variable D_i, conditional on some other factor X. So for the conditional independence assumption to hold in this case, higher education should not affect both feeling towards conservatives and voting for Trump, and feeling towards conservatives should not affect both higher education and voting for Trump. However, it is reasonable to assume that higher education not only affects Trump voting behavior, but also people's feeling towards conservatives, which violates the Conditional Independence Assumption.

Moreover, measurement error in the variables used in the model, including self-reported conservative feelings or educational attainment, can lead to violations of the Conditional Independence Assumption.

Part 2: Interpreting published results (Born et al. (2022))

Table 1: Differences in willingness to lead across gender and team composition

	Dependent variable: Willingness to lead (1-10)			
	(1)	(2)	(3)	(4)
Male	1.633***	1.584***	1.778***	1.417*
	(0.261)	(0.262)	(0.361)	(0.600)
Male-majority team			-1.386***	-1.355***
			(0.403)	(0.402)
Male X Male-majority team			0.713	0.706
			(0.593)	(0.588)
Relative performance first task		-0.269*		-0.319
(1=best, 4=worst)		(0.117)		(0.171)
Male X Relative performance first task				0.127
				(0.203)
Constant	5.633***	6.303***	6.005***	6.792***
	(0.174)	(0.345)	(0.188)	(0.465)
N	580	580	580	580
***F*-test:** Male-majority team +			-0.673	-0.650
'Male X Male-majority team':			(*F*=3.586)	(*F*=3.380)

Note: OLS regressions. Standard errors clustered at the team level. The coefficient for *Male-majority team* indicates the treatment effect on women, while the final row shows the treatment effect on men. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

1. What is the gender gap in willingness to lead the group exercise among men and women of equal performance in the individual (first) task, unconditional on team gender composition?

   Among men and women of equal performance in the individual (first) task, men are 1.584 more willing on a scale from 1-10 to lead the group exercise.

   Let Y be the outcome (i.e., dependent) variable on the willingness to lead. Using Column (2) specification: $Y_i = \beta_0 + \beta_1 Male_i + \beta_4 FirstTask_i + \epsilon_i$

   Since $\mathbb{E}[\epsilon_i] = 0$ by Gauss-Markov assumptions, then crudely speaking partially differentiating Y_i with respect to $Male_i$ (i.e. since we are holding relative performance on the first task constant) yields β_1, which is the coefficient for the "Male" variable.

2. How much more are men willing to lead than women when assigned to male-majority teams, unconditional on performance in the individual (first) task?

   Men are 2.491 more willing to lead than women when assigned to male-majority teams, unconditional on performance in the individual (first) task.

   Using Column (3) specification, letting Maj represent the dummy variable for being in a male-majority team: $Y_i = \beta_0 + \beta_1 Male_i + \beta_2 Maj_i + \beta_3 MaleMaj_i + \beta_4 FirstTask_i + \epsilon_i$

   Conditional on being on a Male-majority team,

   $Y_i|_{Maj=1} = \beta_0 + \beta_1 Male_i + \beta 2(1) + \beta_3 Male_i(1) + \beta_4 FirstTask_i + \epsilon_i$

   such that partially differentiating with respect to $Male_i$ yields $\beta_1 + \beta_3$.

3. Conditioning on the full set of variables available, are men who perform the worst on the individual task more willing to lead than the women who perform the best? Is this true for both male- and female-majority teams?

   Let k be the probability of being randomly assigned to a male-majority team. We will later assume that for the creation of a balanced experiment, the natural value for $k = 1/2$.

   Using the full specification in Column (4):

   $Y_i = \beta_0 + \beta_1 Male_i + \beta_2 Maj_i + \beta_3 MaleMaj_i + \beta_4 FirstTask_i + \beta_5 MaleFirstTask_i + \epsilon_i$

   For the worst-performing males:

   $Y_i|_{Male=1,FirstTask=4} = \beta_0 + \beta_1(1) + \beta_2 k + \beta_3(1)k + \beta_4(4) + \beta_5(1)(4) = \beta_0 + \beta_1 + k\beta_2 + k\beta_3 + 4\beta_4 + 4\beta_5$

   While for the best-performing females:

   $Y_i|_{Male=0,FirstTask=1} = \beta_0 + \beta_1(0) + \beta_2 k + \beta_3(0)k + \beta_4(1) + \beta_5(0)(1) = \beta_0 + k\beta_2 + \beta_4$

   Such that their difference

   $Y_i|_{Male=1,FirstTask=4} - Y_i|_{Male=0,FirstTask=1} = \beta_1 + k\beta_3 + 3\beta_4 + 4\beta_5$

   Conditioning on the full set of variables available, men who perform the worst on the individual task are 1.321 more willing to lead than the women who perform the best.
   To replicate the set up but focusing on only **male-majority** teams, we now have:

   For the worst-performing males:

   $Y_i|_{Male=1,FirstTask=4,Maj=1} = \beta_0 + \beta_1(1) + \beta_2(1) + \beta_3(1)(1) + \beta_4(4) + \beta_5(1)(4) = \beta_0 + \beta_1 + \beta_2 + \beta_3 + 4\beta_4 + 4\beta_5$

   While for the best-performing females:

   $Y_i|_{Male=0,FirstTask=1,Maj=1} = \beta_0 + \beta_1(0) + \beta_2(1) + \beta_3(0)(1) + \beta_4(1) + \beta_5(0)(1) = \beta_0 + \beta_2 + \beta_4$

   Such that their difference

   $Y_i|_{Male=1,FirstTask=4,Maj=1} - Y_i|_{Male=0,FirstTask=1,Maj=1} = \beta_1 + \beta_3 + 3\beta_4 + 4\beta_5$

   Given this, men who perform the worst on the individual task and are on a Male-majority team are 1.674 more willing to lead than the women who perform the best and are on a Male-majority team.

   To replicate the set up but focusing on only **female-majority** teams, we now have:

   For the worst-performing males:

   $Y_i|_{Male=1,FirstTask=4,Maj=0} = \beta_0 + \beta_1(1) + \beta_2(0) + \beta_3(1)(0) + \beta_4(4) + \beta_5(1)(4) = \beta_0 + \beta_1 + 4\beta_4 + 4\beta 5$

   While for the best-performing females:

   $Y_i|_{Male=0,FirstTask=1,Maj=0} = \beta_0 + \beta_1(0) + \beta_2(0) + \beta_3(0)(0) + \beta_4(1) + \beta_5(0)(1) = \beta_0 + \beta_4$

   Such that their difference

   $Y_i|_{Male=1,FirstTask=4,Maj=0} - Y_i|_{Male=0,FirstTask=1,Maj=0} = \beta_1 + 3\beta_4 + 4\beta_5$

   Given this, men who perform the worst on the individual task and are on a Female-majority team are 0.968 more willing to lead than the women who perform the best and are on a Female-majority team.

References

Born, A., E. Ranehill, and A. Sandberg (2022): "Gender and Willingness to Lead: Does the Gender Composition of Teams Matter?" *The review of economics and statistics*, 104, 259–275.