# ECON-UN 3412 - Econometrics Recitation 4

Matthew Alampay Davis
October 12, 2021

## 1 Bad controls

The following is how we've so far been taught to think about omitted variable bias and control variables:

1. We want to know how $X$ affects $Y$ as measured by the coefficient $\beta$ on $X$ in a linear regression
2. We're worried that some other variable $W$ might be correlated with both $X$ and $Y$. If it is, then $W$ is a potential "confounder" or "omitted variable" which could introduce bias to the regression of $Y$ on $X$ if it remains in the error term
3. If we have data on $W$, then we can include it as a control variable in a multiple regression to debias our estimator. This has the effect of "controlling" for $W$ so that we can calculate the effect on $X$ on $Y$, holding $W$ constant

If we are still concerned about potential confounders, we could try to include more controls to avoid other sources of omitted variable bias. My note last week told us to be cautious: yes, you can always add more controls, but it can also come at the cost of identifying variation in $X$, leading to large standard errors.

This makes sense so far, but isn't the end of the story: there's another reason to be cautious about mindlessly adding control variables. In particular, **adding an additional variable $Z$ as a control can be a bad idea if $Z$ is itself an outcome of variable $X$.** That is to say, if $X$ causally affects $Z$, then including $Z$ as a control is probably a bad idea and $Z$ may be considered a "bad control." The result is that the coefficient on $X$ can no longer be interpreted causally.

A real example might help clarify concepts (see Recitation 4 slides):

In 2017, a lawsuit was filed against Google accusing the company of gender pay discrimination against its female employees. In response, Google conducted its own statistical analysis and concluded that the pay gap actually did not exist and in fact found that in one job category—"Level 4 Software Engineer"—the women were actually paid more than men.

I dug into Google's presentation of their analysis and found that the model they used was essentially the same type of multiple regression we learned last week. In Google's words:

- *"We conducted OLS regressions to check for pay equity in each job group and job level... The OLS method allows us to account for factors that should influence pay and look for unexplained differences in total compensation across demographic groups... Specifically, we looked for pay differences based on gender... our analyses covered every job group"*

So we can imagine a simplified version of their model looking something like the following:

$$\text{Pay} = \beta_0 + \beta_1 \text{Female} + \beta_2 \text{Job level} + \beta_3 \text{Tenure} + u_i \quad (1)$$

In Google's framing, statistical evidence of a pay gap hinges on whether the OLS estimate $\hat{\beta}_1^{\text{OLS}}$ is statistically significant. Presumably, Google ran this exercise and found that it was not. If so, they would not be wrong to conclude that conditional on job level and tenure, there is no evidence of gender pay discrimination.

But that conditioning—"controlling for job level and tenure"—is potentially a candidate for what we're talking about: bad controls. If there are systemic barriers for women to attaining a particular job level or tenure with a company, it would not be captured by this model. That is to say, while it may be true that holding a particular job classification constant or fixed, the pay for employees within that category does not meaningfully vary by gender, it might be the case that women are systematically prevented from attaining that job classification to begin with.

This is indeed one of the claims made by those filing the lawsuit:

- *"Kelly Ellis, a plaintiff in the case, claimed she experienced this when she was first hired at Google. She felt she had enough experience to be placed at a higher responsibility level and didn't initially understand how the classification system worked at the company. She soon came to understand that male colleagues with similar education had been assigned more responsibilities and higher pay from the beginning."*
- *"Heidi Lamar, another plaintiff in the case claims she experienced the same patterns in her position in the Google childcare facilities. She said Google management blamed it on her work quality or the performance in her initial job interview, but investigations as part of the lawsuit found this wasn't the case. In this discovery process, during which the legal team searched for evidence of discrimination, an evaluation of her interview showed she received high marks."*
- *"The discovery process revealed a number of systemic discriminations, including that 49% of people hired as Level 2 software engineers were women but that percentage dropped for higher level positions – 22% for Level 3, 14.2% for Level 4, and 7.2% for Level 5."*

All this to say that $\hat{\beta}_1$ in Google's regression would not really capture the causal effect of "Female" on "Pay" if being female also has a causal effect on the control variable "Job level." We might consider this to be some version of selection bias. I think it's unfortunate that the Stock and Watson textbook does not really discuss this issue from what I can tell given how often its exercises involve testing for gender or racial outcome gaps.

## 2 Breaking perfect multicollinearity

Whenever you run a multiple regression where some subset of your regressors are perfectly multicollinear with one another, R and Stata will do one of two things to break the perfect multicollinearity problems: drop one of the collinear regressors or it keep them and drop the intercept/constant term. It might be helpful to show that these two methods are equivalent to one another even though they result in different coefficient estimates.

For simplicity, let's assume that we have two dummy variables $\text{Rich}_i$ and $\text{NotRich}_i$. For multicollinearity to apply, they must be exhaustive and mutually exclusive; in the two-dummy case, this means $\text{Rich}_i = 1 - \text{NotRich}_i$, i.e. every individual $i$ has a value of $1$ for one and $0$ for the other. Obviously, they contain the exact same information so including both leads to multicollinearity.

So let's start with the regression equation that does not include the constant but that includes both of the dummy variables:

$$y_i = \beta_1 \text{Rich}_i + \beta_2 \text{NotRich}_i + e_i \tag{2}$$

Now subtract and add the term $\beta_2 \text{Rich}_i$ on the RHS:

$$\begin{aligned} y_i &= \beta_1 \text{Rich}_i - \beta_2 \text{Rich}_i + \beta_2 \text{Rich}_i + \beta_2 \text{NotRich}_i + e_i \\ &= (\beta_1 - \beta_2)\text{Rich}_i + \beta_2(\text{Rich}_i + \text{NotRich}_i) + e_i \end{aligned} \tag{3}$$

Now substitute in that $\text{Rich}_i = 1 - \text{NotRich}_i$:

$$\begin{aligned} y_i &= (\beta_1 - \beta_2)\text{Rich}_i + \beta_2(1 - \text{NotRich}_i + \text{NotRich}_i) + e_i \\ &= (\beta_1 - \beta_2)\text{Rich}_i + \beta_2 + e_i \end{aligned} \tag{4}$$

If we let $\gamma = \beta_1 - \beta_2$, we get

$$y_i = \gamma \text{Rich}_i + \beta_2 + e_i \tag{5}$$

This is exactly the regression with a constant included and one of the dummy variables excluded. You can now see also how the coefficients between the two regressions 2 and 5 are related:

- In regression 5, which excludes the second dummy variable but includes the constant, the coefficient on the constant (i.e. the intercept) is exactly the coefficient of that excluded dummy variable in regression 2, which includes both dummies but drops the intercept.
- The coefficient on the first dummy variable is exactly $\beta_1 - \beta_2$, the difference between the two effect sizes in the regression that excludes the constant. So by dropping the second dummy variable, the coefficients can be interpreted as the effect size of the first dummy relative to the effect size of the second dummy (which is, to emphasize, given by the constant since $\text{Rich}_i = 0$ corresponds to $\text{NotRich}_i = 1$).
- Using the $\gamma = \beta_1 - \beta_2$ relation allows us to recover the coefficients for one regression using the estimates from the other

The same holds in the more general case where we have three or more regressors that together are perfectly co-linear (e.g., if we had four regressors that were dummies for North, South, East, West).

## 3 Identifying variation and joint significance

In last week's notes on identifying variation, I noted that if two covariates have high enough correlation, then it becomes impossible to disentangle the two effects and we can not confidently estimate the effect of one covariate on the outcome variable. This is related to last week's discussion of testing for joint significance of multiple covariates: sometimes, we don't want to check whether a specific variable has a statistically significant effect, we want to test whether *at least one* of several variables jointly have a significant effect.

A couple of exam-type questions to think about on your own: if two covariates are jointly significant at a given level, is it necessarily true that at least one of them has a $t$-statistic that is also significant at that level? And conversely, if one covariate is individually significant at a given level, is it necessarily true that the joint significance test of that covariate and another covariate is jointly significant at that level?

## 4 Fooling R or Stata

This isn't adding anything to the lecture coverage, but I thought having this in writing could be useful. We begin with the multiple regression

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i \tag{6}$$

We're interested in testing the null hypothesis $H_0 : \beta_1 = \beta_2$. Note this is not a test of whether either of these are significantly different from 0, just whether they are significantly different from one another even if one of them is not significantly different from 0.

The point of the fooling R or Stata method is to produce regression output that executes a test of this hypothesis even though normal regression output usually only tests a specific kind of null hypothesis, whether individual coefficients are 0.

To the regression equation, we subtract and add the term $\beta_2 X_1$ and re-arrange:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i \\ &= \beta_0 + \beta_1 X_1 - \beta_2 X_1 + \beta_2 X_1 + \beta_2 X_2 + u_i \\ &= \beta_0 + (\beta_1 - \beta_2) X_1 + \beta_2 (X_1 + X_2) + u_i \\ &= \beta_0 + \lambda X_1 + \beta_2 W + u_i \end{aligned} \tag{7}$$

where $\lambda := \beta_1 - \beta_2$ and $W := X_1 + X_2$. We can run this regression by creating a new variable $W$ which is just the sum of $X_1$ and $X_2$ and running regressions that include $X_1$ and $W$ as regressors, omitting $X_2$. Obviously, $X_1$ is an element of both $X_1$ and $W$ so standard interpretation is invalid. But the $t$ statistic for the first coefficient is indeed a test of the null hypothesis $H_0 : \lambda = 0$, which by definition is equivalent to a test of whether $\beta_1 - \beta_2 = 0 \Leftrightarrow \beta_1 = \beta_2$, which we want.

In practice, R has a much easier way of directly testing this hypothesis or basically any other linear hypothesis using the *linearHypothesis* function, which we cover in recitation this week. But this method is still worth learning for the purposes of this course.