Lecture 16: More on Categorical Variables

Pierre Biscaye

Fall 2022

Agenda

1. Hypothesis testing with categorical variables
2. Categorical variables and policy analysis
3. Bad controls
4. Interactions with categorical variables
5. Chow tests

Hypothesis testing with the interactive model: practice

$$wage_i = \delta_0 + \delta_1 married_i + \delta_2 female_i + \delta_3 female_i * married_i + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + u_i$$

1. How to test if the impact of marriage on wages is different for females than non-females?
2. How to test if married females earn a different amount than single females?
3. How to test if females earn a different amount than non-females?

Hypothesis testing with the interactive model: practice

$$wage_i = \delta_0 + \delta_1 married_i + \delta_2 female_i + \delta_3 female_i * married_i + \\ \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + u_i$$

1. How to test if the impact of marriage on wages is different for females than non-females?
   - $H_0 : \delta_3 = 0$
2. How to test if married females earn a different amount than single females?
   - $H_0 : \delta_1 + \delta_3 = 0 \; (\delta_1 + \delta_2 + \delta_3 = \delta_2)$
3. How to test if females earn a different amount than non-females?
   - $H_0 : \delta_2 = 0 \; and \; \delta_3 = 0$

Hypothesis testing with the categorical model: practice

$$\begin{aligned} wage_i = \gamma_0 + \gamma_1 marrmale_i + \gamma_2 singfem_i + \gamma_3 marrfem_i \\ + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + u_i \end{aligned}$$

Excluded category: single non-females

1. How to test if the impact of marriage on wages is different for females than non-females?

2. How to test if females earn a different amount than non-females?

Hypothesis testing with the categorical model: practice

$$wage_i = \gamma_0 + \gamma_1 marrmale_i + \gamma_2 singfem_i + \gamma_3 marrfem_i$$
$$+ \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + u_i$$

Excluded category: single non-females

1. How to test if the impact of marriage on wages is different for females than non-females?
   - $H_0 : \gamma_3 - \gamma_2 = \gamma_1$
2. How to test if females earn a different amount than non-females?
   - $H_0 : \gamma_2 = 0$ *and* $\gamma_3 - \gamma_1 = 0$

To Jupyter!

Categorical variables and policy analysis

It is often useful to test whether groups experience different outcomes. Examples:

1. Test for discrimination: do differences in mean wages between females and non-females (the female wage penalty) reflects discrimination or other underlying differences.
   - If underlying differences (e.g. in education or childcare responsibilities), target those. If discrimination, enact anti-discrimination policy.
2. Test for program impacts: do mean outcomes differ between treatment and control units?
   - If yes, program has an impact! May be relevant for policy.
3. Test for heterogeneity: does the impact of a treatment differ by level of education?
   - If yes, implementation should account for this.

Example: the female wage penalty

- May want to test whether differences in mean wages between females and non-females (the female wage penalty) reflects discrimination or other underlying differences.
- Could estimate

$$log(wage_i) = \beta_0 + \beta_1 female_i + \beta_2 educ_i + \beta_3 exper_i + \beta_4 tenure_i + u_i \quad (1)$$

- Possible test for gendered wage discrimination is $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.
- In policy analysis, will often want to go farther.

What about potential omitted variables?

- Wage differences by sex could be due to differences in other factors: omitted variable concerns.
- If so, could have $\hat{\delta}_1 \neq 0$ even when $\delta_1 = 0$.
- Solution: test the sensitivity of $\hat{\delta}_1$ to additional controls.
- For example, what if wage differences result from different preferred occupations?

$$\begin{aligned} log(wage_i) = \beta_0 + \beta_1 female_i + \beta_2 educ_i + \beta_3 exper_i + \beta_4 tenure_i \\ + \beta_5 profocc_i + \beta_6 services_i + u_i \end{aligned} \tag{2}$$

To Jupyter!

Concern about "bad controls"

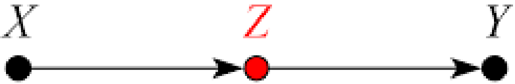
- Occupational choice matters for wages, but doesn't account for a large share of the female wage penalty.
  - β_1 changes from -0.30 without occupation controls to -0.27.
  - Could say occupational sorting accounts for around 10% of the gender wage gap.

Concern about "bad controls"

- Occupational choice matters for wages, but doesn't account for a large share of the female wage penalty.
  - β_1 changes from -0.30 without occupation controls to -0.27.
  - Could say occupational sorting accounts for around 10% of the gender wage gap.
- But what if occupational sorting is not an independent choice? What if discrimination leads women to choose certain occupations?
- In this case we have a concern about "bad controls."
  - If part of the effect of discrimination is through job choice, then including those controls prevents us from estimating the full effect of discrimination.
  - They are "bad controls" in the sense that *Female* has a causal effect on them.
- Including controls removes possible sources of omitted variable bias, but also removes certain pathways through which *Female* might affect wages.

Bad controls

On the board: illustrating "bad control" relationships.



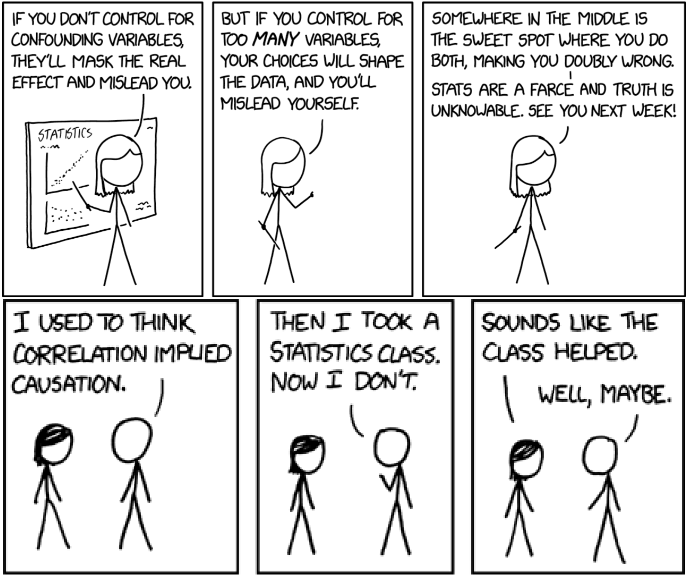
(11)

Goal: block spurious paths between X and Y, but don't perturb causal paths between them.

What to do about bad controls?

- Suppose we want to identify the effect of discrimination in the female wage penalty.
- Run a regression *without* controls.
  - Estimate the effect of being female on wages across all potential pathways.
  - Discuss concerns about likely sources of OVB and directions of bias.
- Then run regression(s) *with* controls.
  - Removes some possible sources of OVB.
  - Identifies residual effect of being female through pathways other than what you've controlled for. Discuss if any controls could be shutting off some mechanisms/pathways.
  - Discuss MLR interpretation of *Female* coefficient.
- Takeaway: be careful about how independent variables might be related and how interpretation changes with controls.
  - Be especially careful of whether one variable causally affects another.

Break



Source: xkcd

Deeper policy analysis: *why* do we observe this wage penalty

- Earlier, we looked at the marriage premium for wages.

$$
\begin{aligned}
log(wage_i) = \delta_0 + \delta_1 married_i + \delta_2 female_i + \delta_3 female_i * married_i + \\
\beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + u_i
\end{aligned} \quad (3)
$$

- We find $\hat{\delta_1} = 0.292$ $(p = 0.09)$, $\hat{\delta_2} = -0.097$ $(p < 0.01)$, and $\hat{\delta_3} = -0.316$ $(p < 0.01)$.
- How do we interpret δ_3?

Deeper policy analysis: *why* do we observe this wage penalty

- Earlier, we looked at the marriage premium for wages.

$$log(wage_i) = \delta_0 + \delta_1 married_i + \delta_2 female_i + \delta_3 female_i * married_i + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + u_i \quad (3)$$

- We find $\hat{\delta_1} = 0.292$ $(p = 0.09)$, $\hat{\delta_2} = -0.097$ $(p < 0.01)$, and $\hat{\delta_3} = -0.316$ $(p < 0.01)$.
- How do we interpret δ_3?
- For women, the marriage "premium" increases hourly wages by 31.6% less than for men (making it negative overall), all else equal.
- Why? Research suggests marriage is associated with stability and responsibility for men, but concerns about availability (due to pregnancy, childcare, and the possibility of dropping out of the workforce) for women.
- This seems to explain a large share of the female wage penalty.

Other explanations for wage penalty

- We can also evaluate whether females and non-females have different returns to other attributes.
- What if females and non-females face different returns to education?
- How do we separate these?

Interaction terms help us explore differences by sex

$$log(wage_i) = \beta_0 + \delta_0 female_i + \beta_2 educ_i + \delta_1 female_i * educ_i + u_i \quad (4)$$

- $\Delta log(wage) = (\beta_2 + \delta_1 female)\Delta educ$
- For non-females

$$\frac{\Delta log(wage)}{\Delta educ} = \beta_2 \quad (5)$$

- While for females, the slope is different by δ_1

$$\frac{\Delta log(wage)}{\Delta educ} = \beta_2 + \delta_1 \quad (6)$$

- Graph on board

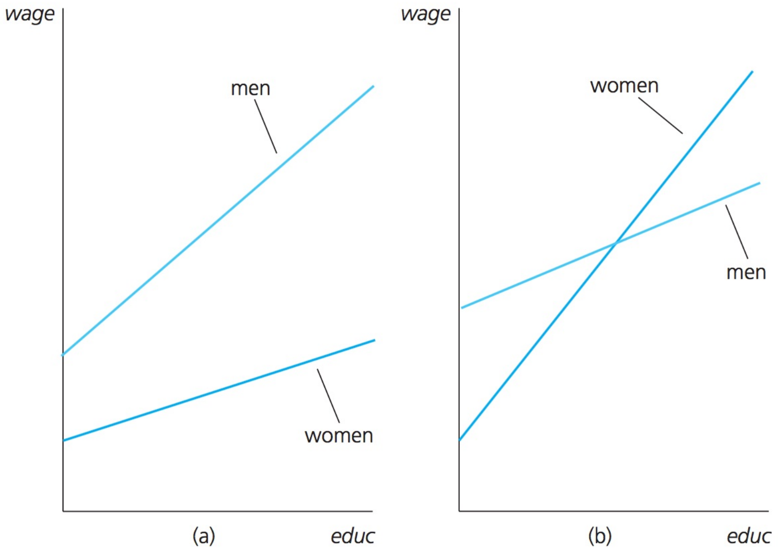
Interpretations on gender gaps

$$log(wage_i) = \beta_0 + \delta_0 female_i + \beta_2 educ_i + \delta_1 female_i * educ_i + u_i \quad (7)$$

$$\begin{aligned}
E[log(wage)|female = 1, educ] &= \beta_0 + \delta_0 + \beta_2 educ_i + \delta_1 educ_i \\
E[log(wage)|female = 0, educ] &= \beta_0 + \beta_2 educ_i \\
E[log(wage)|female = 1, educ] &- E[log(wage)|female = 0, educ] \\
&= \delta_0 + \delta_1 educ
\end{aligned}$$

- So the gender gap depends on education (if $\delta_1 \neq 0$).
    - Helpful to evaluate at the median.
    - In these data, $p50(Educ) = 12$.
    - So at the median value of education, the gender wage gap = $\delta_0 + 12 * \delta_1$

Graphs of equation (7.16): (a) $\delta_0 < 0$, $\delta_1 < 0$; (b) $\delta_0 < 0$, $\delta_1 > 0$.



Does the gender gap widen or close with education?

$$log(wage_i) = \beta_0 + \delta_0 female_i + \beta_2 educ_i + \delta_1 female_i * educ_i + u_i \quad (8)$$

- We've already shown $\delta_0 < 0$.
- Changes in gender gap with education depends on the sign of δ_1.

To Jupyter!

Interpretation

- In these data there is a large and robust difference in wages between genders.
- It doesn't appear strongly related to differences in the returns to education.
  - Note that there could be differences in the returns to education even if there were *no* average difference in wages.
- Can we test whether there are *any* important differences in returns between genders?

Testing for any differences 1: many interactions

$$log(wage_i) = \beta_0 + \delta_1 female_i + \beta_2 educ_i + \beta_3 exper_i + \beta_4 tenure_i \tag{9}$$
$$+\delta_2 female_i * educ_i + \delta_3 female_i * exper_i + \delta_4 female_i * tenure_i + u_i$$

- How to test whether there are no differences by gender?

Testing for any differences 1: many interactions

$$
\begin{aligned}
log(wage_i) = \beta_0 + \delta_1 female_i + \beta_2 educ_i + \beta_3 exper_i + \beta_4 tenure_i \\
+\delta_2 female_i * educ_i + \delta_3 female_i * exper_i + \delta_4 female_i * tenure_i + u_i
\end{aligned} \tag{9}
$$

- How to test whether there are no differences by gender?
- Run an F test for $H_0 : \delta_1, \delta_2, \delta_3, \delta_4 = 0$.
- Tests whether any of the parameters differs by gender.

Testing for any differences 2: Chow test

A way to do the same test without specifying all the interaction terms.

1. Take the full (P=pooled) sample. Estimate

$$log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + u_i \qquad (10)$$

- This is your "restricted" model: no differences by gender.

2. Run the same regression twice more.
   - First on the subsample $female_i = 0$ (F).
   - Then on the subsample $female_i = 1$ (NF).
   - These are your "unrestricted" regressions: allow the parameters to vary across models (by *female*).
3. Test whether these separate subsample analyses do a significantly better job explaining the variation in wages than the pooled model.
   - Observe that $SSR_U = SSR_F + SSR_N F$ (Chow's insight).

Chow test

- Run an F-test, modifying the test statistics as follows:

$$F = \frac{\frac{SSR_r - SSR_u}{q}}{\frac{SSR_u}{n-k-1}} = \frac{\frac{SSR_P - (SSR_F + SSR_{NF})}{k+1}}{\frac{SSR_F + SSR_{NF}}{n-2(k+1)}} \tag{11}$$

- The hypothesis that the βs are the same across subsamples involved $q = k + 1$ restrictions.
- The unrestricted model has $n - 2(k + 1)$ degrees of freedom because each subsample estimates $k + 1$ parameters.

To Jupyter!

Chow test

- Run an F-test, modifying the test statistics as follows:

$$F = \frac{\frac{SSR_r - SSR_u}{q}}{\frac{SSR_u}{n-k-1}} = \frac{\frac{SSR_P - (SSR_F + SSR_{NF})}{k+1}}{\frac{SSR_F + SSR_{NF}}{n-2(k+1)}} \quad (11)$$

- The hypothesis that the βs are the same across subsamples involved $q = k + 1$ restrictions.
- The unrestricted model has $n - 2(k + 1)$ degrees of freedom because each subsample estimates $k + 1$ parameters.

To Jupyter!

- Interpretation: $F = 18.8$, $p < 0.001$: strongly reject that there are no differences in returns between genders.
- Even though the interaction term is only significant for the interaction with experience at a 10% level.