Regression and Causality II

Pablo Balán

GOV 1263, February 8 2018

Some slides draw on slides by Joshua Angrist (MIT), Gary King (Harvard) and Teppei Yamamoto (MIT)
Housekeeping

- **Pablo’s OH**: Thursday 10-11 and 7-7:40
- **Dan’s OH**: Monday 2-3 and Wednesday 2-4
- **Assignment #1**: Due next Thursday, February 22 9:00 to both Pablo and Dan (and bring print copy to your section)
The menu

- Bad control
- How to survive the assignments
Concepts review

- Elite capture: the opposite of state autonomy
- Social capital: norms of reciprocity + networks of civic engagement (Robert Putnam)
- Corruption: ???
- Accountability: politicians respond to voter preferences (to some extent)
What do you think about the amount/level of math in your section?

23 responses

- 47.8%: Way too much math!
- 13%: A bit too much math
- 26.1%: The right amount of math
- 8.7%: Not enough math
- 8.7%: "Baby stats" is so disappointing!
In your perspective, does the math in section make concepts more clear or more confusing?

23 responses
So far, what's the best about section?

5 responses

- Good learning about econometrics
- Willingness to ensure people understand concepts.
- That he cares that people are with him!
- Aside from learning new things, Pablo's memes
- Pablo is good at what he does - getting to peel back the math is great.
So far, what's the best about section?
5 responses

<table>
<thead>
<tr>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good learning about econometrics</td>
</tr>
<tr>
<td>Willingness to ensure people understand concepts.</td>
</tr>
<tr>
<td>That he cares that people are with him!</td>
</tr>
<tr>
<td>Aside from learning new things, Pablo's memes</td>
</tr>
<tr>
<td>Pablo is good at what he does - getting to peel back the math is great.</td>
</tr>
</tbody>
</table>

YOU MADE MY DAY!
Sometimes, it appears to me the Math makes the concepts a bit more confusing, for example when discussing selection bias.

Slightly more engaging

I really think there could be more math

Notation should be explained far more thoroughly. It's far too confusing.

Slides before section please

explaining all variable in the equations. Selection bias equation instead of showing symbols try and label each symbol in order to clarify.

I'm not quite sure what I am supposed to be taking away and implementing myself - right now I am simply not understanding the math, just because I don't know how to read it - am I supposed to not bother - or stop the whole class in its track until I understand - or if you could maybe send out some videos about basic understanding ;)

For technical section, slides on regression and difference and means could go a little faster. Slower on the novel things but faster on stuff someone in a technical section would most definitely have already covered like regression.
BAD CONTROL
Bad control

- Bad control is subtle, very subtle
- More isn’t always better
- Can generate bias (aka post-treatment bias)
- Example: you want to study the effect of schooling on income, but you control for **occupation**
- Denote $W$ an indicator for a white collar job. $W = 1$ if white collar. $W = 0$ otherwise
- Problem: White collar occupation $(W)$ is affected by college $(C)$
- Lecture example: ethnicity affects knowledge about a candidate
Bad control

- Assume college is randomly assigned
- Potential income is independent of college: \( \{ Y_1, Y_0 \perp\!\!\!\!\!\!\!\perp C \} \)
- Potential occupation is independent of college: \( \{ W_1, W_0 \perp\!\!\!\!\!\!\!\perp C \} \)
- Realized income is a function of college: \( Y = CY_1 + (1 - C)Y_0 \)
- Realized occupation is a function of college: \( W = CW_1 + (1 - C)W_0 \)
- The causal effect of college on income:
  \[
  \mathbb{E}[Y|C = 1] - \mathbb{E}[Y|C = 0] = (\text{R.A.}) \mathbb{E}[Y_1 - Y_0]
  \]
- The causal effect of college on occupation:
  \[
  \mathbb{E}[W|C = 1] - \mathbb{E}[W|C = 0] = (\text{R.A.}) \mathbb{E}[W_1 - W_0]
  \]
Bad control

- The effect of college, only for white collar workers \((W = 1)\):

\[
\mathbb{E}[Y|W = 1, C = 1] - \mathbb{E}[Y|W = 1, C = 0]
\]

- Potential outcomes (by the “linking equation”):

\[
\mathbb{E}[Y_1|W_1 = 1, C = 1] - \mathbb{E}[Y_0|W_0 = 1, C = 0]
\]

- Remember: \(\{Y_1, Y_0 \perp\!\!\!\!\!\!\perp C\}\) and \(\{W_1, W_0 \perp\!\!\!\!\!\!\perp C\}\)

- Independence means we can get rid of \(C\)

\[
\mathbb{E}[Y_1|W_1 = 1] - \mathbb{E}[Y_0|W_0 = 1]
\]
Bad control

\[ \mathbb{E}[Y_1|W_1 = 1] - \mathbb{E}[Y_0|W_0 = 1] \]

- Same old trick: add zero

\[ \mathbb{E}[Y_1|W_1 = 1] - \mathbb{E}[Y_0|W_0 = 1] + \mathbb{E}[Y_0|W_1 = 1] - \mathbb{E}[Y_0|W_1 = 1] \]

- Rearranging, you get your good old friend:

\[ \underbrace{\mathbb{E}[Y_1|W_1 = 1] - \mathbb{E}[Y_0|W_1 = 1]}_{\text{Causal Effect (TOT)}} + \underbrace{\mathbb{E}[Y_0|W_1 = 1] - \mathbb{E}[Y_0|W_0 = 1]}_{\text{Selection Bias}} \]
Interpretation of bias term

\[ E[Y_1 | W_1 = 1] - E[Y_0 | W_1 = 1] + E[Y_0 | W_1 = 1] - E[Y_0 | W_0 = 1] \]

- **First term**: Expected potential non-college earnings, given that potential white collar status associated with college education is equal to 1.

- **Second term**: Expected potential non-college earnings, given that potential white collar status associated with non-college education is equal to 1

If, despite no college education, you get a white collar job, you are probably special (have a high \( Y_0 \)).
### How bad control creates selection bias

<table>
<thead>
<tr>
<th>Type of worker</th>
<th>Potential occupation</th>
<th>Potential earnings</th>
<th>Average earnings by occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without college</td>
<td>With college</td>
<td>Without college</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>Always Blue (AB)</td>
<td>Blue</td>
<td>Blue</td>
<td>1,000</td>
</tr>
<tr>
<td>Blue White (BW)</td>
<td>Blue</td>
<td>White</td>
<td>2,000</td>
</tr>
<tr>
<td>Always White (AW)</td>
<td>White</td>
<td>White</td>
<td>3,000</td>
</tr>
</tbody>
</table>
Graphic intuition (from lecture)

- How do results change if we account for differences in knowledge?

\[
\begin{align*}
&+ \left( \begin{array}{c}
Y \\
Y \\
Y \\
N
\end{array} \right) - \left( \begin{array}{c}
Y
\end{array} \right) \\
&+ \left( \begin{array}{c}
N \\
N \\
N \\
N
\end{array} \right) - \left( \begin{array}{c}
N \\
N \\
N
\end{array} \right)
\end{align*}
\]

Effect of ethnicity accounting for differences in knowledge

- Differences in knowledge are likely to be a result (endogenous to) ethnicity.
  - Voters are more likely to know about co-ethnic candidates.
- We cannot control for part of the role of ethnicity (information about a candidate) that we are trying to assess!
Controlling for some variables can make things even worse.
You don’t want to control for variables that are affected by the treatment.
It will bias your result (multicollinearity will just decrease precision).
No easy choice

Regression:
"when you fix one bug, you introduce several newer bugs."
SURVIVING THE ASSIGNMENTS
General tips

- The research question is the most important thing
- Think big: who’s mind are you gonna change about what?
- Think: be precise
- Keep it short: cut. Once you’re done, cut even more
- Go straight to the point: this is not rocket science, but it’s not comparative literature