Vanderbilt University Political Science - Stats I
Fall 2024 - Prof. Jim Bisbee

16 Lecture 16

16.1 Estimating the error variance

- You'll recall that in the univariate context, we encountered a roadblock when we wrote $VAR(\overline{Y}) = \frac{\sigma_Y^2}{n}$. That is that we rarely know σ_Y^2. Well, when we write $VAR\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{SST_x}$, we have the same problem. We rarely have reason to know σ^2 in the OLS context, either.

- What did we do in the univariate case? We estimated σ_Y^2 with $S_U^2 = \frac{\sum_i (y_i - \overline{y})^2}{n-1}$. You'll recall that this was the empirical variance of y adjusted for the number of degrees of freedom (one) used in generating the estimate.

- Well, we'll do a similar thing here. We will estimate σ^2 with

$$\widehat{\sigma}^2 = \frac{\sum \widehat{u}_i^2}{(n-2)} = \frac{SSR}{(n-2)}.$$

- A proof that $E(\widehat{\sigma}^2) = \sigma^2$ may be found on p.57 of Wooldridge. The intuition here is that we have the variance of the residuals, again adjusted by the number of degrees of freedom (two) – since we've already generated estimates $\left(\widehat{\beta}_0 \text{ and } \widehat{\beta}_1\right)$ of two parameters using the two first order conditions for deriving the OLS estimators, which required that:

$$\sum \widehat{u}_i = 0 \text{ and } \sum \widehat{u}_i x_i = 0.$$

- The way to think about this (or any degrees of freedom scenario) is: how many pieces of data are free to vary once we've made our estimate? Here, if we know $n-2$ of the residuals, we can always calculate the other two residuals via the formulas above. They are not free to vary. We therefore lose two degrees of freedom, resulting in a total of $n-2$ degrees of freedom in our estimate of σ^2.

- Thus our unbiased estimators of $VAR\left(\widehat{\beta}_1\right)$ and $VAR\left(\widehat{\beta}_0\right)$ are:

$$\begin{aligned}
\widehat{VAR\left(\widehat{\beta}_1\right)} &= \frac{\widehat{\sigma}^2}{SST_x} = \frac{\frac{SSR}{(n-2)}}{SST_x} \\
\widehat{VAR\left(\widehat{\beta}_0\right)} &= \frac{\widehat{\sigma}^2 \frac{\sum_i x_i^2}{n}}{SST_x} = \frac{\frac{SSR}{(n-2)} \frac{\sum_i x_i^2}{n}}{SST_x}.
\end{aligned}$$

- $\widehat{\sigma}^2$, our estimate of σ^2, plays another important role, because

$$\sqrt{\widehat{\sigma}^2} = \widehat{\sigma} \xrightarrow{p} \sigma.$$

- Thus $\widehat{\sigma}$ is an interesting quantity in and of itself. It is expressed in units of y, which means that it tells us:
  - empirically, how far off the typical fitted value of y is away from the observed value; and
  - theoretically, the extent to which unexplained factors are affecting the value of y.
- It is a very informative statistic that gets much less attention than it deserves.
- Terminology:
  - Wooldridge calls $\widehat{\sigma}$ the Standard Error of the Regression (SER).
  - In Stata's regression output, $\widehat{\sigma}$ is displayed as "Root MSE," which stands for the root of the mean squared error of the regression.
  - I call $\widehat{\sigma}$ the standard error of the estimate, or SEE.
  - And sometimes you'll just see it displayed as $\widehat{\sigma}$.
- [NEXT YEAR: RELATIONSHIP BETWEEN R^2 AND $\widehat{\sigma}$.]

16.2 Hypothesis tests about β_1

- For now, we'll hold off on a discussion of how to conduct hypothesis tests on β_1. It will be more efficient to turn to it once we encounter multiple regression in the next lecture.

16.3 Controlling for a variable

- We are about to move on to multivariate regression.
- But before we do that, let's motivate the notion of controlling for a variable, and noticing how this does and does not compare to multiple regression.
- As we conduct research on political phenomena, we are often interested in what is known as the *ceteris paribus*–that is, the "all things being equal"–relationship between X and Y. [Draw diagram on board.]
  - That is, we are interested in the (often counterfactual case) of what the relationship between X and Y would look like if all other aspects of our units were the same.
    - ∗ We often call those other aspects variables Z.
- [NEXT YEAR: WHY IS THIS A PROBLEM? BECAUSE IF Z IS CORRELATED WITH BOTH X AND Y, THEN THE BIVARIATE RELATIONSHIP BETWEEN X AND Y MAY LEAD US TO IMPROPER CONCLUSIONS ABOUT THE CETERIS PARIBUS RELATIONSHIP BETWEEN X AND Y.
  - MAYBE INCLUDE EXAMPLES WITH CORRELATIONS?
  - Sometimes we do this because we are interested in the effect of X on Y, and we want to be sure that it is not due to Z.
  - But often, we're simply interested in the relationship between X and Y, holding everything else constant.
- Let's get specific about the terminology used here:
  - In this context, Z is called the potential **confound.**
  - If Z confounds the relationship between X and Y, it *renders the relationship spurious*.
    - ∗ That is, it leads us to improper conclusions about the *ceteris paribus*–that is, the "all things being equal"–relationship between X and Y.
  - Let's think a bit about potential confounds that may render a relationship spurious:ftbpF3.2655in2.4561

- To determine whether Z renders the relationship between X and Y spurious, we:
  * ″control for Z″
  * ″condition on Z″
  * ″hold Z constant.″
- All three of these phrases typically mean the same thing.
- But there are several different ways to do this. Ideally, we would do exactly what ″holding Z constant″ suggests: divide our units by categories of Z and examine the relationship between X and Y within each category of Z.
  * If the relationship persists after controlling for Z, we say that it is not spurious.
  * If it no longer persists, we say that Z is a confound rendering the relationship between X and Y spurious.
- In practice, we usually do something much less careful.
- Handout: controlling for a variable.