Econometrics Discussion Section 2: panel data

John Green

Spring 2024

Linearity assumption

- We talk a lot about the OLS assumptions: conditional mean 0 of the error, finite 4th moments, no multicolinearity ...
- Lurking under the hood: assumption the relationship is linear
- This is a very strong assumption: think about relationship between earnings and wages
- So we may try to relax the assumption of linearity and estimate a more flexible form; but we will focus on models which still fit into the framework of OLS

Polynomial function

- If relationship between Y and X is not linear, we can try to approximate it by adding polynomials of X into the regression:
  - $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_3 X^n + u$
- OLS works the same way! Just with new variables which are powers of X
- Difficult to interpret coefficients
- Question: How many factors should we had?

Log approximation

- To a first approximation, $\log(1+x) \approx x$ for small x (though be careful)
  - This means we can think about a change in $\log(x)$ as a percentage change in x
- Different ways to introduce logs into $Y = X\beta + u$. How should we interpret:
  - log-linear
  - linear-log
  - log-log

Log approximation

- To a first approximation, $\log(1+x) \approx x$ for small x (though be careful)
  - This means we can think about a change in $\log(x)$ as a percentage change in x
- Different ways to introduce logs into $Y = X\beta + u$. How should we interpret:
  - **log-linear**: a change of z in X is associated with a $\beta z\%$ change in Y
  - **linear-log**: a change of $z\%$ in X is associated with a $\beta.0z$ change in Y
  - **log-log**: a change of $z\%$ in X is associated with a $\beta z\%$ change in Y
- Other (actual) nonlinear forms are possible too

Non-linear least squares

- These options (polynomials, logs) are still linear in the parameters
- But we can also estimate models which are nonlinear in the parameters using the non-linear least squares (NLLS) method
- Example:
  - Logistic curve: $Y_i = \frac{1}{1+\exp(-(\beta_0+\beta_1 X_i))} + u_i$
  - Negative exponential: $Y_i = \beta_0 \left[1 - e^{-\beta_1(X_i-\beta_2)}\right] + u$
- Still minimize the sum of squared residuals to find parameters, only now we need to use *numerical methods* (eg gradient descent) rather than the closed form solution of OLS

Validity

- *Internal validity* means our study holds for the sample we have
- *External validity* means we can generalize to the population
- External validity is hard, but even internal validity can fail:
  1. OVB
  2. Wrong functional form (eg linear when we should have a polynomial)
  3. Errors-in-variables/measurement error
  4. Sample selection
  5. Simultaneous causality/endogeneity
- We have already talked about the first two; let's talk about the other three

Measurement error

- Say the model is $Y_i = \beta_0 + \beta_1 X_i + u_i$
- But instead of observing X_i, we observe $X_i^* = X_i + v_i$, ie a noisy signal
- Then we estimate

$$Y_i = \beta_0 + \beta_1 X_i^* + u_i = \beta 0 + \beta_1 (X_i + v_i) + u_i$$

$$\rightarrow \beta 0 + \beta_1 X_i + (\beta_1 v_i + u_i) = \beta 0 + \beta_1 X_i + u_i^*$$

- This introduces measurement error bias: $\hat{\beta}_1 \rightarrow_p \beta_1 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_V^2}$
- Notice direction of bias is constant!

Sample selection

- If data are missing non-randomly due to the value of the dependent variable, we have a problem
- Example: we only observe wages for people who are employed
  - https://tinyurl.com/polyecon
- Solutions:
  - Filter your sample
  - RCT
  - Model the selection bias

Simultaneous Causality

- Most of the time, $X \rightarrow Y$ *and* $Y \rightarrow X$; think about supply and demand
- This will introduce endogeneity into our model
- So, we need to somehow control for the endogeneity; can use an instrumental variable (IV) or an RCT (which is really a type of IV)

Panel data

- Panel data has many advantages; in particular it let's us control for OVB so long as the omitted variable is constant over time
- We can *demean* the data: subtract the mean of each variable from each observation
- We can *difference* the data: subtract the previous observation from each observation (only if $T - 2$)
- We can add in $N - 1$ dummy variables, to control for each unit
- But we need to be careful! If we have measurement error, differencing will amplify it

First differences

- Suppose we have two periods of data: $\{X_{it}, Y_{it}\}_{i<=N,t<=2}$
- Our model is $Y_{it} = \beta_0 + \beta_1 X_{it} + \alpha_i + u_{it}$
- So take

$$\begin{aligned} Y_{i1} - Y_{i2} &= \beta_0 + \beta_1 X_{i1} + \alpha_i + u_{i1} - \beta_0 - \beta_1 X_{i2} - \alpha_i - u_{i2} \\ &= \beta_1 (X_{i1} - X_{i2}) - \tilde{u}_i \end{aligned}$$

- So we regress change in Y on change in X and get a consistent estimate of β_1

Time fixed effects

- Some things might vary across time, but not across units (national interest rate)
- We can de-mean based on *time* rather than unit
- Include $T - 1$ binary dummy variables
- Or, take the difference (if $T = 2$); then we include a slope in the difference equation

Assumptions for panel least squares

- 3 and 4 still the same: finite 44^{th} moment (outliers are rare) and no multicolinearity
- Conditional mean is different, need to condition on all periods of unit i plus the fixed effect: $E\left(u_{it} \mid X_{i1}, \ldots, X_{iT}, \alpha_i\right) = 0$
- $\left(X_{i1}, \ldots, X_{iT}, u_{i1}, \ldots, u_{iT}\right), i = 1, \ldots, n$ are i.i.d. across i
- Also need to watch out for *autocorrelation*, when $corr(u_{i,t}, u_{i,t+s}) \neq 0$ for some s
  - This will often be the case!
  - If we ignore, our standard errors will be too small; just like for heteroskedasticity
  - So, *cluster* standard errors by unit

Linear probability model

- Say we have some outcome $D \in \{0, 1\}$ and some regressor X
- We might try to model $D_i = \beta_0 + \beta_1 X_i + u_i$, where $\hat{D}_i$ is the probability that $D_i = 1$
- Good: β is easy to interpret
- Also good: inference is straightforward, R^2 works normally, etc.
- Bad: $\hat{D}_i$ can be < 0 or > 1, which does not make sense
- Also bad: change in probability is linear (can we relax this?)

Non-linear probability model

- Probit: $Pr(Y = 1 \mid X_1, X_2) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$
  - $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ is the z-score, so a change in X_1 shifts the z-score by β_1
- logit: $Pr(Y = 1 \mid X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$
- Both valid choices, doesn't really matter which you use
- Estimate numerically, just like for the non-linear least squares: solve $\min_{b_0,b_1} \sum_{i=1}^{n} [Y_i - \Phi(b_0 + b_1 X_i)]^2$

Maximum Likelihood Estimation (MLE)

- In practice, we use the *maximum-likelihood estimator*: instead of *minimizing* a sum of squared errors, we will choose parameters to *maximize* the "likelihood function"
- This estimate will be consistent, efficient, and normally distributed
- The specific form of the likelihood function depends on the problem you are solving, but intuition is familiar
  - With a t-test, we knew the distribution under H_0 and asked: given this distribution, how likely was it we would observe our data?
  - MLE turns that around, and asks: given our data, what is the most likely distribution that it comes from?
  - Find the parameters for that distribution
- We can evaluate the fit from a logit or a probit using the fraction correctly predicted or a psueudo-R^2