Econometrics Discussion Section 1: nonlinear methods and IV

John Green

Spring 2025

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 wages and education:
    - How much do we expect earnings to increase if we go from 8 years to 12 years of ed? What about 12 to 16?
- So we may try to relax the assumption of linearity
- Many such options, but we will focus on models which still fit into the framework of OLS: polynomials and logs

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, and $\frac{\partial Y}{\partial X}$ now depends on X
- How many factors should we include?
- What are the tradeoffs?

Example

- Silly example: $Y = \mathit{examscore}$ and $X = \mathit{coffee}$. Why might relationship be nonlinear?

Example

- Silly example: $Y = examscore$ and $X = coffee$. Why might relationship be nonlinear?
- Score should increase with coffee up to a point, when it may then decrease; model is:

$$examscore_i = \beta_0 + \beta_1 coffee_i + \beta_2 coffee_i^2 + u_i$$

- What sign should β_1 and β_2 have?

Example

- Silly example: $Y = examscore$ and $X = coffee$. Why might relationship be nonlinear?
- Score should increase with coffee up to a point, when it may then decrease; model is:

$$examscore_i = \beta_0 + \beta_1 coffee_i + \beta_2 coffee_i^2 + u_i$$

- What sign should β_1 and β_2 have?
- Expect $\beta_1 > 0$ and $\beta_2 < 0$
- So then what is the effect of an increase of 1 cup of coffee?

Example

- Silly example: $Y = \mathit{examscore}$ and $X = \mathit{coffee}$. Why might relationship be nonlinear?
- Score should increase with coffee up to a point, when it may then decrease; model is:

$$\mathit{examscore}_i = \beta_0 + \beta_1 \mathit{coffee}_i + \beta_2 \mathit{coffee}_i^2 + u_i$$

- What sign should β_1 and β_2 have?
- Expect $\beta_1 > 0$ and $\beta_2 < 0$
- So then what is the effect of an increase of 1 cup of coffee?

$$\frac{\partial \mathit{examscore}}{\partial \mathit{coffee}} = \beta_1 + 2\beta_2 \mathit{coffee}_i$$

- So it depends on how much coffee we've consumed!

Log approximation

- Logarithmic transformations are another very useful way to relax linearity
- 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
- 3 different ways to introduce logs into $Y = X\beta + u$. How should we interpret:
  - **log-linear**: a 1 unit change in X is associated with a $\beta\%$ change in Y
  - **linear-log**: a 1% change in X is associated with a β change in Y
  - **log-log**: a 1% change in X is associated with a $\beta\%$ change in Y *What concept from elements does this make you think of?*
- Other (actual) nonlinear forms are possible too, but we won't discuss these

Example

How do we interpret β_1 in these models?

- **log-linear**: $log(wage_i) = \beta_0 + \beta_1 educ_i$
- **linear-log**: $pollution_i = \beta_0 + \beta_1 log(distance_i)$
- **log-log**: $log(hours_i) = \beta_0 + \beta_1 log(wage_i)$

Simultaenous causality

- Focus on last example: $log(hours_i) = \beta_0 + \beta_1 log(wage_i)$
- What might be a problem with estimating this model?

Simultaenous causality

- Focus on last example: $log(hours_i) = \beta_0 + \beta_1 log(wage_i)$
- What might be a problem with estimating this model?
  - Hours may increase wages (better performance at work), and wages may increase hours (more incentive to work)
  - So we have simultaneity (and thus endogeneity): $hours_i$ and $wage_i$ are jointly determined, and regular OLS will be biased
- We can use an *instrumental variable* to get around this problem

Instrumental variables

- Shift to general setup: $Y_i = \beta_0 + \beta_1 X_i + u_i$ where X is endogenous
- An instrument is a variable Z which we can use as a "shifter" for X; Z gives us *variation* in X not related to u, which can be used to estimate β_1
- Z must satisfy two conditions:
  1. **Relevance**: Z must be correlated with X (otherwise it won't shift X)
  2. **Exogeneity**: Z must not be correlated with u (otherwise it won't give us variation in X that is uncorrelated with u)
- This is equivalent to saying that Z only affects Y through X (and not directly); *exclusion restriction*

Instrumental variables

- If assumptsions satisfied, then we can use a *two-stage least squares* (2SLS) estimator to estimate β_1
- First stage: regress X on Z and any other exogenous variables (e.g. W):

$$X_i = \pi_0 + \pi_1 Z_i + \pi_2 W_i + v_i$$

- Second stage: regress Y on the predicted values of X from the first stage:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \beta_2 W_i + u_i$$

- Can show that $\hat{\beta}_1^{2SLS}$ is consistent and unbiased estiamtor for β_1
- Easily generalized to multiple variables and even multiple instruments

Weak instruments

- First condition: Z must be **relevant**
- If relationship is weak, this causes some problems: IV estimates can be imprecise, and worse, testing procedures may fail; OLS with a bit of bias may be preferable
- How can we test?

Weak instruments

- First condition: Z must be **relevant**
- If relationship is weak, this causes some problems: IV estimates can be imprecise, and worse, testing procedures may fail; OLS with a bit of bias may be preferable
- How can we test?
- F-test on the first stage regression; rule of thumb, F¿10 (or F¿30 nowadays) is usually good
- Intuition: the model for X which includes Z should be substantially better than the model which does not include Z; if not, then Z is weak (doesn't tell us much about X)

Exogeneity

- Exogeneity is harder to test; can only do so if our model is overidentified (more instruments than endogeneous variables)
- In that case, can use a J-test
- Idea: instruments should not be correlated with residuals from second stage
- So, estimate 2SLS model, generate the residuals $\hat{u}_i$, then regress residuals on exogeneous variables and instruments
- Coefficients on instruments should be jointly 0
- Can only say whether or not the set of instruments is exogeneous; if we reject, cannot say which is endogeneous