14.32 Recitation 6

Nina Wang

MIT Department of Economics

Lectures 11-12

Table of Contents

Table of Contents

Interaction terms

$$Y_i = \alpha + \beta_1 X_i + \beta_2 D_i + \beta_3 X_i \times D_i + e_i$$

- We use interaction terms when an effect of X on Y depends on the level of another variable.

Interactions: 2 binary variables

$$\ln(Wage_i) = \beta_0 + \beta_1 College_i + \beta_2 Female_i + \beta_3 College_i \times Female_i + e_i$$

Interactions: 2 binary variables

$$\ln(Wage_i) = \beta_0 + \beta_1 College_i + \beta_2 Female_i + \beta_3 College_i \times Female_i + e_i$$

- Interpreting coefficients:
  - $E(\ln(Wage)|Female = 0, College = 0) = \beta_0$
  - $E(\ln(Wage)|Female = 0, College = 1) = \beta_0 + \beta_1$
  - $E(\ln(Wage)|Female = 1, College = 0) = \beta_0 + \beta_2$
  - $E(\ln(Wage)|Female = 1, College = 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$

Interactions: 2 binary variables

$$\ln(Wage_i) = \beta_0 + \beta_1 College_i + \beta_2 Female_i + \beta_3 College_i \times Female_i + e_i$$

- Interpreting coefficients:
  - $E(\ln(Wage)|Female = 0, College = 0) = \beta_0$
  - $E(\ln(Wage)|Female = 0, College = 1) = \beta_0 + \beta_1$
  - $E(\ln(Wage)|Female = 1, College = 0) = \beta_0 + \beta_2$
  - $E(\ln(Wage)|Female = 1, College = 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$
- Thus, we have
  - The effect of graduating college for men is β_1
  - The effect of graduating college for women is $\beta_1 + \beta_3$
  - The effect of being female for non-college grads is β_2
  - The effect of being female for college grads is $\beta_2 + \beta_3$

Interactions: 1 binary 1 continuous

$$\ln(Wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Female_i + \beta_3 Educ_i \times Female_i + e_i$$

Interactions: 1 binary 1 continuous

$$\ln(Wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Female_i + \beta_3 Educ_i \times Female_i + e_i$$

- Interpreting coefficients:
  - $E(\ln(Wage)|Female = 0, Educ = x) = \beta_0 + \beta_1 x$
  - $E(\ln(Wage)|Female = 1, Educ = x) = \beta_0 + \beta_2 + (\beta_1 + \beta_3)x$

Interactions: 1 binary 1 continuous

$$\ln(Wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Female_i + \beta_3 Educ_i \times Female_i + e_i$$

- Interpreting coefficients:
  - $E(\ln(Wage)|Female = 0, Educ = x) = \beta_0 + \beta_1 x$
  - $E(\ln(Wage)|Female = 1, Educ = x) = \beta_0 + \beta_2 + (\beta_1 + \beta_3)x$
- Here, β_3 is the difference in returns to education between males and females.

Interactions: 2 continuous variables

$$\ln(Wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Exp_i + \beta_3 Educ_i \times Educ_i + e_i$$

Interactions: 2 continuous variables

$$\ln(Wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Exp_i + \beta_3 Educ_i \times Educ_i + e_i$$

- Interpreting coefficients:

$$\text{Returns to education: } \frac{\partial \ln(Wage)}{\partial Educ} = \beta_1 + \beta_3 Exp$$

$$\text{Returns to experience: } \frac{\partial \ln(Wage)}{\partial Exp} = \beta_2 + \beta_3 Educ$$

Review: Nonlinear Regressions

- Polynomial Regressions
  - Use when we suspect the relationship is nonlinear
  - Can test fit by using T or F test on higher order coefficients (F test when we're comparing two or more higher order terms against linear specification)
- Log Regressions
  - Linear-log, Log-linear, Log-log (know how to interpret!)
  - We use when we suspect relationship follows log relationship, or if distribution of regressor is heavily skewed.
  - Can test fit against linear specification by using log squared.

Review: Nonlinear Regressions

- Interaction terms
  - We use interaction terms when an effect of X on Y depends on the level of another variable.
  - Can test fit by testing difference in effect of Y between levels of X by testing coefficients.

Table of Contents

Types of Bias

- Wrong functional form
- OVB
- Errors-in-variable bias (aka measurement error)
- Selection bias
- Simultaneous causality bias

Types of Bias

- Wrong functional form
  - Arises if the functional form is incorrect – for example, an interaction term is incorrectly omitted; then inferences on causal effects will be biased.

Types of Bias

- Wrong functional form
  - Arises if the functional form is incorrect – for example, an interaction term is incorrectly omitted; then inferences on causal effects will be biased.
- OVB
  - Arises if an omitted variable both (i) is a determinant of Y and (ii) is correlated with at least one included regressor.

Types of Bias

- Wrong functional form
  - Arises if the functional form is incorrect – for example, an interaction term is incorrectly omitted; then inferences on causal effects will be biased.
- OVB
  - Arises if an omitted variable both (i) is a determinant of Y and (ii) is correlated with at least one included regressor.
- Errors-in-variable bias aka measurement error
  - Arises if regressor of interest is collected with systematic error (non-random and not constant).

Types of Bias

- Wrong functional form
  - Arises if the functional form is incorrect – for example, an interaction term is incorrectly omitted; then inferences on causal effects will be biased.
- OVB
  - Arises if an omitted variable both (i) is a determinant of Y and (ii) is correlated with at least one included regressor.
- Errors-in-variable bias aka measurement error
  - Arises if regressor of interest is collected with systematic error (non-random and not constant).
- Selection bias
  - Arises when a selection process (i) influences the availability of data and (ii) that process is related to the dependent variable.

Types of Bias

- Wrong functional form
    - Arises if the functional form is incorrect – for example, an interaction term is incorrectly omitted; then inferences on causal effects will be biased.
- OVB
    - Arises if an omitted variable both (i) is a determinant of Y and (ii) is correlated with at least one included regressor.
- Errors-in-variable bias aka measurement error
    - Arises if regressor of interest is collected with systematic error (non-random and not constant).
- Selection bias
    - Arises when a selection process (i) influences the availability of data and (ii) that process is related to the dependent variable.
- Simultaneous causality bias
    - Arises if dependent variable influences regressor.

Chooseing between regression specifications

- General guidelines:
  - Think about which variables need to be included to prevent OVB from affecting the coeffcient of interest.
  - If the specifcations are nested, compare them with formal t or F tests.
  - If the specifcations are not nested, it may be helpful to compare R^2 values.
  - Consider which specifcation is easier to understand or interpret. All else equal, simpler is better.

Midterm Review

14.32, Fall 2021

October 22, 2021

Key Probability and Stats Concepts

Won't review them today, but should be comfortable with them:

- Random variables
  - Continuous vs. discrete
  - PDFs and CDFs
  - Definitions and properties of expectations, variance
- Convergence concepts
  - Convergence in probability
  - Convergence in distribution
  - LLN and CLT
- Hypothesis testing
  - One-sided vs. two-sided alternative hypotheses
  - Type 1 vs. Type 2 errors; size vs. power
  - Simple tests we can do by hand, e.g. difference of two means

Properties of Estimators

Desirable properties of an estimator $\hat{\theta}$ when the true value is θ:

$$\text{Consistency: } \hat{\theta} \xrightarrow{p} \theta$$

$$\text{Unbiasedness: } E[\hat{\theta}] = \theta$$

$$\text{Efficiency: } \hat{\theta} = \underset{E[\tilde{\theta}]=\theta}{\operatorname{argmin}} E[(\tilde{\theta} - \theta)^2]$$

OLS: Assumptions

Consider the standard multivariate regression setup:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_k X_{i,k} + e$$

The following assumptions are sufficient to guarantee the consistency, unbiasedness, and asymptotic normality of the OLS estimator:

- (Y_i, X_i) are iid observations
- Exogeneity: $E[e|X] = 0$
- No perfect multicollinearity among regressors
- Large outliers are unlikely: $E[X_{i,j}^4] < \infty, E[e_i^4] < \infty$

Note that we didn't say anything about homoskedasticity.

OLS Estimator: Derivation

Set up the estimation problem in matrix form:

$$Y = X\beta + e,$$
$$\hat{\beta} = \underset{\beta}{\text{argmin}}(Y - X\beta)'(Y - X\beta),$$

where Y is a $n\,x\,1$ vector, X is a $n\,x\,k$ matrix with a vector of ones as its first column, and e is a $n\,x\,1$ vector.

First-order condition gives:

$$X'(Y - X\hat{\beta}) = 0$$

Interpretation: the OLS residuals are orthogonal to each column of X.

Solving for $\hat{\beta}$ gives:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

OLS Estimator: Interpreting Coefficients

What do the OLS coefficients represent?

$$E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$
$$\implies \frac{\partial E[Y|X]}{\partial X_j} = \beta_j$$

OLS coefficient β_j gives the change in the conditional expectation of Y that results from a 1-unit increase in X_j, holding all other independent variables constant.

OLS Estimator: Consistency and Unbiasedness

$\hat{\beta}$ is consistent and unbiased:

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(X\beta + e) \\ &= \beta + (X'X)^{-1}X'e \\ &\xrightarrow{p} \beta,\end{aligned}$$

because $E[e|X] = 0$ and $X'e \xrightarrow{p} 0$ by the law of large numbers.

$$\begin{aligned}E[\hat{\beta}] &= \beta + E[(X'X)^{-1}X'e] \\ &= \beta + E[E[(X'X)^{-1}X'e|X]] \\ &= \beta + E[(X'X)^{-1}E[X'e|X]] \\ &= \beta,\end{aligned}$$

because $E[e|X] = 0$.

OLS Estimator: Homoskedasticity vs. Heteroskedasticity

If the data are homoskedastic ($E[e_i^2]$ does not depend on X), then the expression for the OLS estimator's asymptotic variance simplifies:

$$\begin{aligned}
\Sigma_X^{-1} Q \Sigma_X^{-1} &= E[X_i X_i']^{-1} E[e_i^2 X_i X_i'] E[X_i X_i']^{-1} \\
&= E[e_i^2] E[X_i X_i']^{-1} E[X_i X_i'] E[X_i X_i']^{-1} \\
&= E[e_i^2] E[X_i X_i']^{-1}
\end{aligned}$$

Heteroskedasticity-robust standard errors don't make this simplifying assumption, and instead plug in an estimate of $Q = E[e_i^2 X_i X_i']$.

Takeaway: heteroskedasticity only affects standard errors, not coefficients. Heteroskedasticity-robust standard errors allow the variance of e to depend on X.

OLS: Hypothesis Testing

Let $V \equiv \Sigma_X^{-1} Q \Sigma_X^{-1}$ be the covariance matrix of the OLS estimator that we derived above. Three types of hypothesis tests we discussed:

- Single coefficient

$$H_0 : \beta_k = b, H_1 : \beta_k \neq b$$

$$t = \frac{\sqrt{n}(\beta_k - b)}{\sqrt{V_{kk}}} \xrightarrow{d} N(0,1)$$

- Single linear restriction on multiple coefficients

$$H_0 : a'\beta = b, H_1 : a'\beta \neq b$$

$$t = \frac{\sqrt{n}(a'\beta - b)}{\sqrt{a' V_{kk} a}} \xrightarrow{d} N(0,1)$$

OLS: Hypothesis Testing

- In practice:

- Test the null that a single coefficient equals zero by dividing the coefficient estimate by its standard error.
- Test the null that a group of coefficients are all zero by looking at reported F statistic.

R^2 and Goodness of Fit

Recall the definition of R^2:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Interpretation: fraction of variation in Y that can be explained by our OLS model, with $\hat{Y}_i = X_i'\hat{\beta}$.

Ranges between 0 and 1, with higher values indicating better fit.

Assessing the overall fit of the OLS model is different than assessing the importance of individual independent variables! In economics, we often have significant coefficients but small R^2 values.

Omitted Variable Bias

Consider the following model, with Y a $n \, x \, 1$ vector, X (the variable of interest) a $n \, x \, 1$ vector, and W a $n \, x \, k$ vector of other independent variables:

$$Y = X\beta + W\gamma + e$$

Suppose we omit the W variables and run the "short" regression:

$$Y = X\beta + u$$

Omitted Variable Bias

The OLS estimate from the short regression is:

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(X\beta + W\gamma + e) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'W\gamma + (X'X)^{-1}X'e \\ &\xrightarrow{p} \beta + (X'X)^{-1}X'W\gamma\end{aligned}$$

So the asymptotic omitted variable bias is:

$$\hat{\beta} - \beta = \overbrace{(X'X)^{-1}X'W}^{\text{OLS coeffs. of } W \text{ on } X} \underbrace{\gamma}_{\text{effect of } W \text{ on } Y}$$

So for OVB to be a problem, we need the omitted variables W to i) be correlated with X, and ii) have an effect on Y.

Omitted Variable Bias

Expanding the expression from above,

$$\begin{aligned}\hat{\beta} - \beta &= (X'X)^{-1}X'W\gamma \\ &= \sum_{j=1}^{k} \frac{\text{cov}(X, W_j)}{\text{var}(X)}\gamma_j\end{aligned}$$

So total OVB is equal to the sum of the OVBs caused by each omitted variable.

Recall the example from PS2, where the dependent variable was cost and the variable of interest was a dummy for private colleges. Omitting college size created positive bias, while omitting the liberal arts dummy created negative bias. The bias from omitting size was larger, so the net effect of omitting both was positive bias.

Nonlinear Regression

We can allow for nonlinearities within the OLS framework by including nonlinear transformations of the independent variables as regressors. Since the model is still linear in parameters, we can still solve it by OLS.

Simplest example: polynomials. Including an additional squared term allows us to model a quadratic relationship between X and Y:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + e$$
$$\frac{\partial Y}{\partial X} = \beta_1 + 2\beta_2 X$$

Another common case: taking log transformations of dependent and/or independent variables:

$$\log(Y) = \beta_0 + \beta_1 X + e$$
$$\frac{\partial \log(Y)}{\partial X} \approx \frac{\partial Y / Y}{\partial X} = \beta_1$$

Interpreting Log Transformations

Three possible cases:

- $\log(Y)$ regressed on X
  - 1-unit increase in X changes $E[Y]$ by $\beta * 100$ percent.
- Y regressed on $\log(X)$
  - 1-percent increase in X changes $E[Y]$ by $\beta/100$.
- $\log(Y)$ regressed on $\log(X)$
  - 1-percent increase in X changes $E[Y]$ by β percent. β is an elasticity.

Interpreting Interactions

Again, three possible cases:

- Dummy interacted with another dummy. Example:

$$\log(\text{wage}_i) = \alpha + \beta \text{college}_i + \gamma \text{female}_i + \theta \text{college}_i * \text{female}_i$$

  β is the return to college for men, and $\beta + \theta$ is the return to college for women. The interaction coefficient θ is a **difference-in-difference** estimator: the difference in the return to college between men and women. Including an interaction term allows the return to college to depend on gender.

Interpreting Interactions

- Dummy interacted with a continuous variable. Example:

$$\log(\text{wage}_i) = \alpha + \beta\text{educ}_i + \gamma\text{female}_i + \theta\text{educ}_i * \text{female}_i$$

β is the percentage increase in men's wages from an additional year of education, and $\beta + \theta$ is the percentage increase in women's wages from an additional year of education. The interaction coefficient θ is a **slope-shifter**: the difference in the wage-education slope between men and women.

Interpreting Interactions

- Continuous variable interacted with another continuous variable. Example:

$$\log(\text{wage}_i) = \alpha + \beta \text{educ}_i + \gamma \text{exp}_i + \theta \text{educ}_i * \text{exp}_i$$

  β is the percentage increase in wage from an additional year of education for a person with zero years of experience. γ is the percentage increase in wage from an additional year of experience for a person with zero years of education.

  θ is again a slope-shifter, but in a continuous sense. A one-year increase in experience raises the percentage return to an additional year of education by θ. Similarly, a one-year increase in education raises the percentage return to an additional year of experience by θ.

Choosing Between Regression Specifications

General guidlines:

- Think about which variables need to be included to prevent OVB from affecting the coefficient of interest.
- If the specifications are nested, compare them with formal t or F tests.
- If the specifications are not nested, it may be helpful to compare R^2 values.
- Consider which specification is easier to understand or interpret. All else equal, simpler is better.