Introduction to Multivariate Regression & Program Evaluation

HED 612

Lecture 10

Where are we going....

- This Lecture
  - Intro to multivariate regression
  - Intro to interpreting published multivariate regression results
  - Reading for next lecture:
    - Cabrera, N. L., Milem, J. F., Jaquette, O., & Marx, R. (2014). Missing the (student achievement) forest for all the (political) trees: Empiricism and the Mexican American Studies controversy in Tucson. American Educational Research Journal, 51(6), 1084-1118.
    - Powers, J. M. (2004). High-Stakes Accountability and Equity: Using Evidence From California's Public Schools Accountability Act to Address the Issues in Williams v. State of California. American Educational Research Journal, 41(4), 763–795.
  - Homework #10 posted!
- Next Lecture
  - Other OLS assumptions
  - Graphing multivariate regression results
  - Creating publication quality tables
- Next Next Lecture
  - Introduction to non-linear relationships between X and Y
  - Mini lesson on what each section of manuscrupt should accomplish!

Introduction to Multivariate Regression

Population Regression Model

- Same as in "simple" (univariate) regression!
- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_k X_{ki} + u_i$
- Where:
  - Y_i = observation i of dependent variable
  - X_{1i} = observation i of the first regressor, X_1
  - X_{2i} = observation i of the second regressor, X_2
  - X_{ki} = observation i of the Kth regressor, X_k
  - β_1 = population average effect of Y for one-unit increase in X_1
  - β_2 = population average effect of Y for one-unit increase in X_2
  - β_k = population average effect of Y for one-unit increase in X_k
  - β_0 = average value of Y when the value of all independent variables $(X_1, X_2 \ldots X_k)$ are equal to zero
  - u_i = all other variables that affect the valye of Y_i but are not included in the model

Things we do in multiple regression

1. Estimation

- Choose estimates for $\beta_0, \beta_1, \beta_2, ...\beta_k$ by selecting those that minimuze the sum of squared errors (i.e., make the best prediction of Y), yielding an OLS line
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ...\hat{\beta}_k X_{ki}$

2. Measures of model fit (e.g., R^2, SER)

- But formulas change slightly to account for degrees of freedom!
- Once you introduce multiple independent variables, use adjusted R-squared
- Adjusted R-squared
  - Adjusted for the number of predictors in the model
  - Every independent variable we add to the model will increase our "normal" R-squared; but doesn't necessarily mean it's a better fit!
  - Adjusted R squared increases only if new variable improves the model more than would be expected by chance!

3. Prediction

- Once you estimate OLS regression line, we can calculate predicted values for observations with particular values of all independent variables
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ...\hat{\beta}_k X_{ki}$

4. Hypothesis testing and confidence intervals about β_1

- Same as before but forumals for $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$ change slightly, but R calculates this for us!

Conditional Independence Assumption

- Assume students choose to participate in MAS
- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$
- Where: Y=graduation, X1= 0/1 MAS, X2= previous academic achievement, X3= SES
- **Conditional independence assumption**:
  - Once we include control variables, there are no omitted variables, Z, that satisfy *both* of these two conditions:
    1. (1) Z affects value of Y *and*
    2. (2) Z has a relationship with X
- If the conditional independent assumption is true:
  - Once we include relevant control variables, there are no omitted variables that affect Y and have a relationship with X
  - MAIN POINT: if we satisfy the conditional independence assumption through control variables, then multiple regression is just as good as randomized assignment experiment!

Multiavariate regression in Program Evaluation vs Social Science

- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$
- Program evaluation research or "econometrics"
  - We are only interested in estimating β_1 [the causal effect of X1 on Y]
  - The only reason we include other variables in the model besides X1 is to eliminate omitted variable bias
  - Therefore, we include all control variables that satisfy *both* conditions of omitted variable bias
  - once we include control variables, and no other variables satisfy both conditions, then we satisfy the conditional independence assumption and we can estimate a causal effect!
- Traditional social science statistics [most of my research!]
  - Purpose of multiple regression is to add new variable to your model (e.g., X_3) to see the effect of X_3 on Y
  - Can lead to sloppy research if you're not careful!
  - We "throw" everything and the kitchen sink into a model and see what's interesting!

Multivariate regression in R

- Research question: What is the effect of student teacher ratio on student reading test scores?
- Simple regression
  - $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$
  - Where: Y= reading test scores and X_1 = student teacher ratio
  - Interpretation of $\hat{\beta_1}$: The average effect of a one-unit increase in X_1 is associated with a $\hat{\beta_1}$ change in Y
- Multivariate regression
  - $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$
  - Where: Y= student test scores, X_1 = student teacher ratio, X_2 = % ELL
  - Interpretation of $\hat{\beta_1}$: The average effect of a one-unit increase in X_1 is associated with a $\hat{\beta_1}$ change in Y, holding the value of X_2 constant

What does “holding constant” mean?

- RQ: What is the effect of student teacher ratio on reading test scores?
  - $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$
  - Where: Y= student test scores, X_1 = student teacher ratio, X_2 = % ELL
- Setup:
  - We think student test scores go down if there’s a greater percentage of ELL students in the classroom
    - First condition of omitted variable bias (Z affects Y)
  - We think there is a negative relationship between percentage of ELL students in the classroom and student-teacher ratio
    - Second condition of omitted variable bias (Z has a relationship with X)
- Problem:
  - We think student teacher ratio and percentage of ELL move together
  - We want to know the relationship between reading scores and student teacher ratio when “percent ELL” is not allowed to move!
- “Holding the value of X_2 constant”
  - Means to estimate the relationship betwen X_1 and Y when we don’t allow the value of X_2 to vary
  - Said differenly: We analyze the relationship between student teacher ration (X_1) and reading test scores (Y) for applicants that have the same value of percent ELL (X_2) [calculus: partial derivatives!]

What does “holding constant” mean? Another example....

- RQ: What is the relationship between years of education(X1) on income(Y), after controlling for years of work experience (X2)?
- General interpretation of $\hat{\beta}_1$:
  - The average effect of a one-unit increase in X1 is a $\hat{\beta}_1$ unit increase in Y, holding the value of X2 constant
- Interpretation of $\hat{\beta}_1$, applied to example
  - The effect of having one additional year of education (X1) on income (Y), when we don't allow value of "years of experience" (X2) to change
  - Maybe people w/ more education have fewer years of experience
- Said differently: analyze the effect of increasing years of education on income for people who have same years of experience

Interpreting $\hat{\beta}_1$ for continuous X

- RQ: What is the effect of student teacher ratio (X1) on average district reading test scores (X2)?
  - $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$
  - Where:
    - Y= reading test scores
    - X_1 = average district student teacher ratio, X_2 = 0/1 majority ELL district, X_3 = avg district income ($000s)
  - **IMPORTANT NOTE**: All independent variables should be at the same "level"; here it is district!
- General interpretation of $\hat{\beta}_1$ for continuous X
  - The avaerage effect of a one unit increase in X_1 is a $\hat{\beta}_1$ unit change in Y, holding the values of X_2 and X_3 constant
  - OR The avaerage effect of a one unit increase in X_1 is a $\hat{\beta}_1$ unit change in Y, after controlling for X_2 and X_3
  - The avaerage effect of a one unit increase in X_1 is a $\hat{\beta}_1$ unit change in Y, holding the values of covariates constant
- Run regression in R!
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$
- $\hat{Y}_i = 646.2 - 0.8X_{1i} - 23.5X_{2i} + 1.7X_{3i}$
- Specific example interpretation [run regression in R]
  - The average effect of a one-unit increase in average district student teacher ratio (i.e., one additional student per teacher) is a 0.8 point decrease in average district reading score, holding the values of majority ELL and district average income constant

Interpreting $\hat{\beta}_1$ for categorical X

- RQ: What is the effect of student teacher ratio (X1) on average district reading test scores (X2)?
  - $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$
  - Where:
    - Y= reading test scores
    - X_1 = 0/1 majority ELL, X_2 = avg district student teacher ratio X_3 = avg district income ($000s)
  - **Stylistic Note**: Your main independent variable of interest should always be X_1
- General interpretation of $\hat{\beta}_1$ for categorical X
  - Being [non-reference group] as opposed to [reference group] is associated with a $\hat{\beta}_1$ unit change in Y, holding the values of X_2 and X_3 constant
- Run regression in R [same coef values as previous model but in diff order]
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$
  - $\hat{Y}_i = 646.2 - 23.5 X_{1i} - 0.8 X_{2i} + 1.7 X_{3i}$
- Specific interpretation
  - Reference group is the zero value of my dummy ELL var= non-ELL majority district; Non-Reference group is the one value of my dummy ELL var = majority ELL district
  - Being a majority ELL district as opposed to a non-majority ELL district is associated with a 23 point decrease in average district reading scores, holding values of average student-teacher ratio and district average income constant

Prediction still works the same way!

- RQ: What is the effect of student teacher ratio (X1) on average district reading test scores (X2)?
  - $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$
  - Where:
    - Y= reading test scores
    - $X_1 =$ 0/1 majority ELL, $X_2 =$ avg district student teacher ratio $X_3 =$ avg district income ($000s)
- Run regression in R [same coef values as previous model but in diff order]
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$
  - $\hat{Y}_i = 646.2 - 23.5X_{1i} - 0.8X_{2i} + 1.7X_{3i}$
- What's the predicted average reading score for a district that is a non-ELL majority district, has a student teacher ratio of 25, and average district income of $22,000?
  - $(Y|X_1 = 0, X_2 = 25, X_3 = 22) =$ 646.2 - (23.5 * 0) - (0.8 * 25) + (1.7 * 22)
  - $(Y|X_1 = 0, X_2 = 25, X_3 = 22) =$ 646.2 - (0) - (20) + (37.4)
  - $(Y|X_1 = 0, X_2 = 25, X_3 = 22) =$ 663.6

How to read regression results in academic journals

- Cabrera et al (2014)
  - RQ: What is the effect of participating in MAS on high school graduation?
  - Use program evaluation framework; but their model is a logistic regression because their Y= 0/1 graduated and X= 0/1 MAS participation
- Powers (2004)
  - RQ: what is relationship between school resource variables and school-level academic performance index (API)
  - Don't frame article as "program evaluation" but it is! Y= School's academic performance index score X vars= school resource variables
- Regression results are pretty standardized across all fields and journals!
  - Regression tables usually show the coefficient and standard error (usually in parantheses) for each independent variable
  - Columns are individual models!
    - Usually you start with a simple regression model that only includes your main independent variable of interest: "model 1"
    - Then you add controls; sometimes done in groupings
    - Sometimes models in seperate columns also indicate various samples!