Lecture 7: More on Multiple Linear Regression

Pierre Biscaye

Fall 2022

Assumptions for MLR

- Just as with simple regressions, need a set of assumptions for consistency ($E[\hat{\beta_1}] = \beta_1$) in multiple regressions
- MLR1: In the population, $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i$
  - As before, can accommodate different functions of the x variables
- MLR2: Our observations (with variables $x_1, x_2, ..., y$) were sampled at random from the population

MLR3

- MLR3 is a bit more complicated than SLR3
- MLR3: *no perfect collinearity*
- In the sample, no independent variables are constant, and there are no exact linear relationships between independent variables
- Example of an exact linear relationship: $x_1 = 0.2 * x_2 + 0.8 * x_3$
    - Quadratics and other non-linear transformations do not involve linear relationships

Why is perfect collinearity a problem?

$$x_{1i} = 0.2 * x_{2i} + 0.8 * x_{3i} \quad (1)$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i \quad (2)$$

Substitute (1) into (2) and rearrange

$$y_i = \beta_0 + (\beta_2 + 0.2\beta_1) x_{2i} + (\beta_3 + 0.8\beta_1) x_{3i} + u_i \quad (3)$$

- Infinitely many solutions to this problem!
- Suppose true coefficient on x_2 is 5 and true coefficient on x_3 is 0.5
  - $\beta_2 = 5, \beta_1 = 0, \beta_3 = 0.5$ solve it
  - But so do $\beta_2 = 4, \beta_1 = 5, \beta_3 = -3.5$
  - And $\beta_2 = -10, \beta_1 = 75, \beta_3 = -59.5$

What makes multicollinearity a problem?

Partialling out interpration of "holding all else constant"

- If $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$
- And $x_{1i} = \delta_0 + \delta_1 x_{2i} + r_i$, we had

$$\hat{\beta_1} = \frac{\widehat{cov(r, y)}}{\widehat{var(r)}} = \frac{\sum_i \hat{r}_i y_i}{\sum_i \hat{r_i^2}} \tag{4}$$

- But if $x_{1i} = \delta_0 + \delta_1 x_{2i}$ (or, $r_i = 0$ for all i)
- Then $\hat{r}_i = 0$ for all $i \Rightarrow$ can't estimate β_1 because can't divide by 0
- You cannot estimate the effect of x_1 holding x_2 constant because x_2 perfectly explains x_1; there is nothing in x_1 to "partial out"

Dealing with multicollinearity

To Jupyter!

- With perfect collinearity (the extreme of multicollinearity), infinitely many $\hat{\beta_k}$ will solve the equations
- Solution: drop one of the x variables
- *Much* harder to detect if variables are almost perfectly multicollinear
- Makes $\hat{\beta}$'s much more variable; solution not straightforward
- Takeaway: think carefully about relationships between your X variables
  - It is ok if two X variables are correlated, but if they are too closely related it will cause problems in your estimation if you keep both

Equivalent of SLR4: MLR4

- Instead of $E[u|x] = 0$
- We now need $E[u|x_1, x_2, ..., x_k] = 0$
- Or, conditional on all of our explanatory variables, the error term has an expected value of zero
  - The error term is not correlated with *any* of the X variables
- In other words, we have successfully controlled for the determinants of Y that are correlated with our X variables

MLR4

- MLR4: $E[u|x_1, x_2, ..., x_k] = 0$
- MLR 4 is both stronger and weaker than SLR 4
  - Stronger: need u uncorrelated with *every* x
  - Weaker: have controlled for many x's: less remains in u
- Example:

$$ln(wage_i) = \beta_0 + \beta_1 Educ_i + u_i$$
$$ln(wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Exper_i + \beta_3 Gender_i + \beta_4 Urban_i + u_i$$

MLR4 Practice

Suppose we're interested in the impact of having more children on weekly hours of work, and we estimate

$$Hours_i = \beta_0 + \beta_1 NumKids_i + u_i \tag{5}$$

- Any violations of MLR4 we might be concerned about?
- Small group discussion

MLR4 practice

Suppose we're interested in the impact of having more children on weekly hours of work, and we estimate

$$Hours_i = \beta_0 + \beta_1 NumKids_i + u_i \tag{6}$$

- Any violations of MLR4 we might be concerned about?
- What about when we estimate

$$\begin{aligned} Hours_i = &\beta_0 + \beta_1 NumKids_i + \beta_2 Country_i + \beta_3 Urban_i &(7)\\ &+ \beta_4 SpouseIncome_i + \cdots + \beta_k x_{ki} + u_i &(8)\end{aligned}$$

Theorem

- Suppose MLR1-MLR4 are all true
- Then $E[\hat{\beta}_j] = \beta_j$ for all j: consistency

Omitted variable bias

- Suppose $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$
- But we estimate $y_i = \tilde{\beta_0} + \tilde{\beta_1} x_{1i}$
- Violates MLR4
- Why might we do this?

$$ln(wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Ability_i + u_i \tag{9}$$

- There might be some relevant variables we don't observe/measure

Misspecified models with omitted variables

- When the true model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
- And we estimate $y = \tilde{\beta_0} + \tilde{\beta_1} x_1$
- We get an estimate $\tilde{\beta_1} = \frac{\widehat{cov(x_1,y)}}{\widehat{var(x_1)}}$
- But if we estimate the true model, and consider $x_1 = \gamma_0 + \gamma_1 x_2 + r$ (partialling out)
- We know that $\hat{\beta_1} = \frac{\widehat{cov(\hat{r},y)}}{\widehat{var(\hat{r})}}$
- What is the difference between $\hat{\beta_1}$ and $\tilde{\beta_1}$?

Omitted variables bias

The SLR estimates

$$\begin{aligned} E[y|x] &= \beta_0 + \beta_1 x_1 + E[v|x_1] \\ &= \beta_0 + \beta_1 x_1 + \beta_2 E[x_2|x_1] + E[u|x_1] \end{aligned}$$

Suppose

$$E[x_2|x_1] = \delta_0 + \delta_1 x_1 \tag{10}$$

Then

$$E[y|x_1] = \beta_0 + \beta_2 \delta_0 + (\beta_1 + \beta_2 \delta_1) x_1 + E[u|x_1] \tag{11}$$

- Then in this framework, $E[\tilde{\beta_1}] = E[\hat{\beta_1}] + E[\hat{\beta_2}]\delta_1$, where $\hat{\beta}$s are what we would estimate from the MLR model

Omitted variable bias

- In this framework, $E[\tilde{\beta_1}] = E[\hat{\beta_1}] + E[\hat{\beta_2}]\delta_1$
- Thus, the *Omitted Variables Bias* is $E[\tilde{\beta_1} - \hat{\beta_1}] = E[\hat{\beta_2}]\delta_1$
- If the MLR satisfies SLR1-SLR4, then OVB is $E[\tilde{\beta_1}] - \beta_1 = \beta_2\delta_1$
- We often will have to accept *some* omitted variables bias
  - Usually don't or can't measure all relevant variables
  - Can almost always think of possible omitted variables
- When does it matter?

When does omitting a variable matter?

- $E[\tilde{\beta_1} - \hat{\beta_1}] = E[\hat{\beta_2}]\delta_1$
- When is $\tilde{\beta_1} \approx \hat{\beta_1}$?
  - When $\delta_1 \approx 0$ ($cov(x_1, x_2) \approx 0$)
  - Or $\beta_2 \approx 0$ ($cov(x_2, y) \approx 0$)
- If $\beta_2, \delta_1 >> 0$ or $\beta_2, \delta_1 << 0$ then $\tilde{\beta_1} >> \hat{\beta_1}$: upward bias
- If $\beta_2 > 0, \delta_1 < 0$ or $\beta_2 < 0, \delta_1 > 0$ then $\tilde{\beta_1} < \hat{\beta_1}$: downward bias
- We often won't be able to estimate β_2, δ_1...
  - If we could estimate them, then we could just add x_2 to the regression
  - Need to then think about likely sign and magnitude based on theory and prior beliefs

Thought exercise

Suppose we're interested in the impact of class attendance (number of lectures attended in the semester) and class performance (final grade out of 100), and we estimate

$$Grade_i = \beta_0 + \beta_1 Attend_i + u_i \tag{12}$$

- Think of a potential omitted variable Z
- Think of what sign you would expect for $\delta_1 = cov(Attend, Z)$
- Think of what sign you would expect for $\beta_2 \propto cov(Z, Grade)$
- What is the expected sign of the bias for $be\tilde{t}a_1$ in the SLR relative to $\hat{\beta_1}$ in the MLR that includes Z?
- Small group discussion

Thought exercise

Suppose we're interested in the impact of class attendance (number of lectures attended in the semester) and class performance (final grade out of 100), and we estimate

$$Grade_i = \beta_0 + \beta_1 Attend_i + u_i \tag{13}$$

- Consider $Z = PriorMetrics$, an indicator for having taken a different econometrics course previously
- What sign you would expect for $\delta_1 = cov(Attend, Z)$?
- What sign you would expect for $\beta_2 \propto cov(Z, Grade)$?
- What is the expected sign of the bias for $\tilde{beta}_1$ in the SLR relative to $\hat{\beta_1}$ in the MLR that includes Z?
- What if $Z = FoundAnswers$, an indicator for finding my solutions the day of the final exam?

Omitted variables in multiple regression

- What if $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i$
- But we estimate $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_{k-1} x_{k-1i} + e_i$?
- With multiple regression, *all* of our $\hat{\beta}$'s will be biased
- Bias becomes harder to describe or sign.
  - Intuition: suppose x_k correlated with x_2 but not x_1. Then $\hat{\beta_2}$ clearly biased.
  - But then if x_1 is correlated with x_2, will estimate $\hat{\beta_1}$ based on the wrong $\hat{\beta_2}$ (remember we are trying to "hold all else constant")
  - For this class: most important to be aware of the issue

Dealing with OVB in practice

- If there is an omitted variable you have data on, can include it in the model
- If you don't have data on your omitted variable, need to think about implications for bias
- We can still often get some useful approximations of $\hat{\beta}$s of interest if we are willing to tolerate a risk of some bias
  - Often don't have a choice
  - Need to think critically though about possible sources of bias
  - For people to believe you have a causal estimate will need strong arguments about why any bias is likely minimal

R^2 in multiple regression

- We can still calculate R^2 as a measure of goodness-of-fit

$$R^2 = 1 - \frac{SSR}{SST_y} = 1 - \frac{\sum \hat{u}_i^2}{\sum (y_i - \bar{y})^2} \tag{14}$$

- R^2 can only increase when you add more variables
- R outputs the "Adjusted R^2" with regression results which corrects for this
  - $\bar{R}^2 = 1 - (\frac{SSR}{SST_y})\frac{n-1}{n-k} = 1 - (1 - R^2)\frac{n-1}{n-k}$
  - n is sample size and k is number of explanatory variables
  - Penalizes you for adding more variables to the data, particularly if they have limited (additional) explanatory power
  - Will primarily consider $\bar{R}^2$ when comparing models going forward

To Jupyter!

Variability in MLR estimators

- We need concept of R^2 in MLR setting to calculate variance of $\hat{\beta}$s
- MLR5 (Homoskedasticity): The variance of the error term is unrelated to all of the model x's. That is, $var(u|x_1, ..., x_k) = \sigma_u^2$.
- Under MLR1-MLR5:

$$var(\hat{\beta}_j) = \frac{\sigma_u^2}{SST_j(1 - R_j^2)} \tag{15}$$

$$SST_j = \sum_i (x_{ji} - \bar{x}_j)^2 \tag{16}$$

- R_j^2 is the R^2 from a regression of x_j on $x_1, ..., x_{j-1}, x_{j+1}, ..., x_k$

MLR $\hat{\beta}$ variance

$$var(\hat{\beta}_j) = \frac{\sigma_u^2}{SST_j(1 - R_j^2)} \tag{17}$$

- is smaller when σ_u^2 is smaller
- is smaller when there is more variance in x_j
- is smaller when R_j^2 is smaller

R_j^2

- Remember "partialing out" in MLR.
  - We are only using unexplained variation in x_j to estimate $\hat{\beta}_j$
- Consider perfect multicollinearity: suppose $x_j = \delta_0 + \delta 1 x_1 + ... + \delta_{j-1} x_{j-1} + \delta_{j+1} x_{j+1} + ... + \delta_k x_k$
- MLR3 fails, $R_j^2 = 1$, and $var(\hat{\beta}_j) \to \infty$
- With near perfect multicollinearity: $x_j = \delta_0 + \delta 1 x_1 + ... + \delta_{j-1} x_{j-1} + \delta_{j+1} x_{j+1} + ... + \delta_k x_k + u$ but u are small
- Then $\hat{\beta}_j$ can be estimated, but is quite variable.
  - Hard to isolate precise effect of x_j when it is closely related to other variable we want to hold constant

To Jupyter!

Under MLR1-MLR5, we have an expression for $var(\hat{\beta}_j)$

- Still don't observe σ_u^2

$$\hat{\sigma_u^2} = \frac{\sum_i \hat{u_i}^2}{n-k-1} \tag{18}$$

- $n-k-1$ is the *degrees of freedom*

$$std.error(\hat{\beta}_j) = \sqrt{\frac{\sum_i \hat{u_i}^2}{(n-k-1)\sum_i (x_{ji} - \bar{x_j}^2)(1-R_j^2)}} \tag{19}$$

- Compare to SLR

$$std.error(\hat{\beta_1}) = \sqrt{\frac{\sum_i \hat{u_i}^2}{(n-2)\sum_i (x_{1i} - \bar{x_1}^2)}} \tag{20}$$