Lecture 6: Multiple Linear Regression

Pierre Biscaye

Fall 2022

Recap: Motivating Multiple Linear Regression

- We have an estimator $\hat{\beta_1}$
- if SLR1-SLR4 hold $E[\hat{\beta_1}] = \beta_1$
- If SLR5 also holds, $\widehat{var(\hat{\beta_1})} = \frac{SSR}{(n-2)\sum_i (x_i - \bar{X})^2}$
- The goal of increasing precision of our $\hat{\beta_1}$ estimate, and concerns about SLR4, motivate *multiple linear regression* (MLR)

Increasing precision

$$\widehat{var(\hat{\beta_1})} = \frac{SSR}{(n-2)\sum_i (x_i - \bar{X})^2} \quad (1)$$

- We care about how close $\hat{\beta_1}$ is to the true β_1, and the variance helps us estimate this
- Variance will be small (precision will increase) when SSR is small
- How to reduce $SSR = \sum_i \hat{u_i}^2$?
- Consider two models you can estimate:

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i$$
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

- Which will have the smaller SSR?

Concerns about SLR4

- SLR4: $E[u|x] = 0$
- Suppose we want to estimate the causal impact of education on wages, and the true model is

$$ln(wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Exper_i + \epsilon_i \tag{2}$$

- What happens if we instead estimate a simple regression model?

$$ln(wage_i) = \beta_0 + \beta_1 Educ_i + u_i \tag{3}$$

Omitted variables bias

We will estimate

$$\begin{aligned} E[ln(wage_i|Educ_i)] &= \beta_0 + \beta_1 Educ_i + E[u_i|Educ_i] \\ &= \beta_0 + \beta_1 Educ_i + \beta_2 E[exper_i|Educ_i] + E[\epsilon_i|Educ_i] \end{aligned}$$

Suppose

$$E[Exper|Educ_i] = \delta_0 + \delta_1 Educ_i \tag{4}$$

Then

$$E[ln(wage_i)|Educ_i] = \beta_0 + \beta_2\delta_0 + (\beta_1 + \beta_2\delta_1)Educ_i + E[\epsilon_i|Educ_i] \tag{5}$$

- Our line of best fit will find $\hat{\beta_1} \approx \beta_1 + \beta_2\delta_1$!
- This is what is called *omitted variables bias*: $E[\hat{\beta_1}] - \beta_1 = \beta_2\delta_1$ (in this case - more on this in future lecture)

Multiple regression

Suppose instead we estimate

$$ln(wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Exper_i + u_i \tag{6}$$

- Experience is no longer in u - no omitted variables bias from *Exper*
- Interpretations change: How does Education relate to wages *holding experience constant* (also referred to as "*ceteris paribus*")
- Or, compare two people with the same amount of experience. If one has one more year of education, how much more do they earn? β_1

Other uses of multiple regression

Polynomial relationships

- Consider Kuznets:

$$Gini_i = \beta_0 + \beta_1 GDP_i + \beta_2 GDP_i^2 + u_i \tag{7}$$

- We estimate a different relationship

$$\Delta Gini = (\beta_1 + 2\beta_2 GDP)\Delta GDP \tag{8}$$

- With polynomials, *ceteris paribus* interpretations involve estimating impacts *at a given level* of X (GDP in this case)

Interaction terms: preview of future lecture

$$ln(wage_i) = \beta_0 + \beta_1 Ed_i + \beta_2 Gender_i + \beta_3 Ed_i * Gender_i + u_i \tag{9}$$

Generalization: $k > 2$

- For any k, we can use the statistical model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + ... + \beta_k x_{ki} + u_i \tag{10}$$

- What changes is the interpretation
  - When we interpret β_1, it is the effect of x_1 holding $x_2, x_3, ..., x_k$ constant
- For example

$$ln(wage_i) = \beta_0 + \beta_1 Ed_i + \beta_2 Exper_i + \beta_3 Gender_i + u_i \tag{11}$$

How is MLR like SLR?

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + ... + \beta_k x_{ki} + u_i \qquad (12)$$

- Still calculate $\hat{\beta_k}$ by minimizing squared residuals (OLS)

$$min_{\beta_0, \beta_1, ..., \beta_k} \sum_i (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - ... - \beta_k x_{ki})^2 \qquad (13)$$

Similar FOCs

$$\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - ... - \hat{\beta}_k x_{ki}) = 0 \tag{14}$$

$$\sum_i x_{1i}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - ... - \hat{\beta}_k x_{ki}) = 0 \tag{15}$$

$$\cdots = 0 \tag{16}$$

$$\sum_i x_{ki}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - ... - \hat{\beta}_k x_{ki}) = 0 \tag{17}$$

What else is the same

- With k equations and k unknowns, we can't calculate estimators easily by hand
  - Easy for computers, though
- Can manually calculate predicted values and residuals

$$\hat{y}_i = \hat{\beta_0} + \hat{\beta_1} x_{1i} + \hat{\beta_2} x_{2i} + ... \hat{\beta}_k x_{ki} \tag{18}$$

$$\hat{u}_i = y_i - \hat{\beta_0} - \hat{\beta_1} x_{1i} - \hat{\beta_2} x_{2i} - ... - \hat{\beta}_k x_{ki} \tag{19}$$

Estimating MLR βs example: wages and education

$$ln(wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 X_{2i} + \cdots + u_i$$

- Let's go back to the question of the causal impact of education on wages
- How does controlling for other variables affect our estimated β_1?
- Keep in mind what we said about omitted variables bias
- If we leave out X_2 and we suppose $E[X_2|Educ] = \delta_0 + \delta_1 Educ$, omitted variables bias will be $E[\hat{\beta_1}] - \beta_1 = \beta_2\delta_1$

To Jupyter!

- What variables to include in a regression? Depends on desired interpretations.

What does it mean to hold x_2 constant?

- Last time, we said that in the multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

- we could interpret $\hat{\beta_1}$ as the effect of x_1 on y, *holding* x_2 *constant*.
- What does "holding constant" mean?

A mathematical interpretation

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \tag{20}$$

- Suppose we consider regressing
$$x_{1i} = \delta_0 + \delta_1 x_{2i} + r_i \tag{21}$$

- We could estimate $\hat{\delta_0}$, $\hat{\delta_1}$
- We could then predict $\hat{x_{1i}} = \hat{\delta_0} + \hat{\delta_1} x_{2i}$ and $\hat{r_i} = x_{1i} - \hat{\delta_0} - \hat{\delta_1} x_{2i}$
- What is the interpretation of $\hat{r_i}$?

Characterizing $\hat{\beta}_1$

- If $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$
- And $x_{1i} = \delta_0 + \delta_1 x_{2i} + r_i$, then (recalling $E[\hat{r}] = 0$ by construction)

$$\hat{\beta_1} = \frac{\widehat{cov(r,y)}}{\widehat{var(r)}} = \frac{\sum_i \hat{r}_i y_i}{\sum_i \hat{r_i^2}} \tag{22}$$

- This is the "partialling-out" interpretation
- Compare to the SLR formula

$$\hat{\beta_1} = \frac{\widehat{cov(x,y)}}{\widehat{var(x)}} \tag{23}$$

Example with wages and education

$$ln(wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Exper_i + u_i \quad (24)$$

$$Educ_i = \delta_0 + \delta_1 Exper_i + r_i \quad (25)$$

To Jupyter!

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
- An exact linear relationship: $x_1 = 0.2 * x_2 + 0.8 * x_3$
- Why would this be a problem?

MLR3 - example

$$x_{1i} = 0.2 * x_{2i} + 0.8 * x_{3i} \tag{26}$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i \tag{27}$$

$$y_i = \beta_0 + (\beta_2 + 0.2\beta_1) x_{2i} + (\beta_3 + 0.8\beta_1) x_{3i} + u_i \tag{28}$$

- Infinitely many solutions to this problem!

Dealing with multicollinearity

- 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

To Jupyter!

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 valye of zero: the error term is not correlated with any of 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:

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

Theorem

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