Lectures 7 & 8

Multiple Linear Regression

Oscar Volpe

10/18/2021 & 10/20/2021

Motivation

Suppose we have *i.i.d.* data about Y and explanatory variables $X_1, \dots, X_k$. Given $\{Y_i, X_{i,1}, \dots, X_{i,k}\}_{i=1}^n$, we write down a linear model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k} + U_i \\ &= X_i' \beta + U_i, \end{aligned}$$

where $X_i = (1, X_{i,1}, \dots, X_{i,k})'$ and $\beta = (\beta_0, \beta_1, \dots, \beta_k)'$.

- Using vectors allows us to write this model more compactly.

We can draw conclusions from this model under different assumptions.

- We may want to predict Y_i using multiple explanatory variables.
- We may want to characterize differences in $E(Y|X_1, \dots, X_k)$.
- We may want to give the β_j's a *causal* interpretation.

Multicollinearity

Throughout our analysis, we assume that no X_j can be written as a linear combination of the other explanatory variables $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$.

- Why? Write $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$, where $X_1 = c + dX_2$. It is impossible to make changes to X_1 without making changes to X_2.
- More generally, this issue is known as *perfect multicollinearity*.

Definition (Perfect Collinearity)

A matrix $\mathbf{X}$ is *perfectly collinear* if $P(c'\mathbf{X} = 0) = 1$ for some $c \neq 0$.

Theorem (Existence of $E(\mathbf{X}\mathbf{X}')^{-1}$)

$E(\mathbf{X}\mathbf{X}')$ is invertible if and only if there is no perfect collinearity in $\mathbf{X}$.

- As we will soon see, our least squares coefficients are undefined unless $E(\mathbf{X}\mathbf{X}')^{-1}$ exists, i.e. unless there is no perfect multicollinearity.

Interpretation 1: Best Linear Predictor

Write $U = Y - X'\beta$. To find the best linear predictor, we minimize:

$$MSE(b) = \min_{b \in \mathbb{R}^{k+1}} E[(Y - X'b)^2]$$

The solution (call it β) must satisfy the first-order condition:

$$\textit{FOC:} \quad -2E[X(Y - X'\beta)] = 0 \quad \Longrightarrow E(XU) = 0$$

If this condition holds, then we say $X'\beta$ is BLP$(Y|X)$.

- **Q1.** Is $E(U|X) = 0$?
- **Q2.** Is $\text{Cov}(X, U) = 0$?
- **Q3.** Is $E(U) = 0$?

Solving for the BLP

After minimizing $MSE(b)$, the solution to the least squares problem is:

$$\beta = E(XX')^{-1}E(XY)$$

What assumptions do we need for this equation to hold?

(1) $E(XU) = 0$ (implied by the FOC)

(2) $E(XX')$ must be invertible, i.e. no perfect collinearity in X.

Note that multicollinearity was not an issue for simple linear regression.

- Why? Because we only had one explanatory variable.
- Multicollinearity can be a big issue when estimating linear models with several variables (*example:* dealing with dummy variables).

Interpretation 2: Linear Conditional Expectation

Assume that $E(Y|X) = X'\beta$. Note that $U = Y - E(Y|X)$, because:

$$Y = X'\beta + U = E(Y|X) + U$$

From the properties of conditional expectation, we know that:

$$\begin{aligned}
(a) \quad & E(U|X) = E[Y - E(Y|X)|X] = E(Y|X) - E(Y|X) = 0 \\
(b) \quad & E(U) = E[E(U|X)] = E(0) = 0 \\
(c) \quad & E(XU) = E[E(XU|X)] = E[XE(U|X)] = 0 \\
(d) \quad & \text{Cov}(X, U) = E(XU) - E(X)E(U) = 0
\end{aligned}$$

The Law of Iterated Expectations gives us infinite moment restrictions of the form $E(f(X)U) = 0$, from which we can construct estimators of β.

$$E(f(X)[Y - X'\beta]) = 0 \quad \implies \beta = E(f(X)X')^{-1}E(f(X)Y)$$

Interpretation 3: Causal Model

Assume $Y = g(X, U)$, where X are observed (and U are unobserved) determinants of Y. The effect of X_j on Y, holding X_{-j} and U fixed, is given by $\partial g / \partial X_j$. We make the *assumption* that:

$$g(X, u) = X'\beta + U, \quad \text{so: } \frac{\partial g(X, U)}{\partial X_j} = \beta_j$$

Here, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$ has a causal interpretation. As long as there is a constant in the model, we can normalize U and β_0 so that: $E(U) = 0$.

- **Q1.** Is $E(U|X) = 0$?
- **Q2.** Is $E(XU) = 0$?

Decomposing the Coefficient Vector

Suppose you want to solve for β_1 in the multiple regression model:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_k X_{i,k} + U_i$$

Let $X_{i,-1} = (1, X_{i,2}, X_{i,3}, \ldots, X_{i,k})'$ and $\beta_{-1} = (\beta_0, \beta_2, \beta_3, \ldots, \beta_k)'$. Then:

$$Y_i = X_i'\beta + U_i = \begin{bmatrix} X_{i,1} & X_{i,-1}' \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_{-1} \end{bmatrix} + U_i$$

Under our BLP assumptions, we know that $E(X_i U_i) = 0$, which gives:

$$\beta = E(X_i X_i')^{-1} E(X_i Y_i),$$

or, equivalently, you decompose β in the following way:

$$\begin{bmatrix} \beta_1 \\ \beta_{-1} \end{bmatrix} = E\left(\begin{bmatrix} X_{i,1}^2 & X_{i,1} X_{i,-1}' \\ X_{i,-1} X_{i,1} & X_{i,-1} X_{i,-1}' \end{bmatrix} \right)^{-1} E\left(\begin{bmatrix} X_{i,1} Y_i \\ X_{i,-1} Y_i \end{bmatrix} \right)$$

Another Approach

Alternatively, we can solve for β_1 by taking three steps:

(1) regress Y_i on $X_{i,-1}$ to get "residuals" $\tilde{Y}_i = Y_i - \text{BLP}(Y_i|X_{i,-1})$

(2) regress $X_{i,1}$ on $X_{i,-1}$ to get "residuals" $\tilde{X}_{i,1} = X_{i,1} - \text{BLP}(X_{i,1}|X_{i,-1})$

(3) regress $\tilde{Y}$ on $\tilde{X}_{i,1}$, and the coefficient on $\tilde{X}_{i,1}$ equals β_1

Intuition: β_1 characterizes the relationship between $X_{i,1}$ and Y_i after controlling for the rest of the regressors $X_{i,-1} = (1, X_{i,2}, X_{i,3}, \ldots, X_{i,k})'$.

Consider the linear regression model $\tilde{Y}_i = \tilde{\beta}_1 \tilde{X}_{i,1} + \tilde{U}$, where $\tilde{\beta}_1$ equals:

$$\tilde{\beta}_1 = E(\tilde{X}_{i,1}\tilde{X}_{i,1}')^{-1} E(\tilde{X}_{i,1}\tilde{Y}_i),$$

and $E(\tilde{X}_{i,1}\tilde{U}) = 0$. Then $\tilde{\beta}_1$ will be equal to β_1.

Example: Simple Linear Regression

Consider the simple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + U$$

Define $X = (1, X_1)'$ and $\beta = (\beta_0, \beta_1)'$. Under our BLP assumptions:

$$\beta = E(XX')^{-1}E(XY)$$

If we want to solve for β_1 alone, consider the model $\tilde{Y} = \tilde{\beta}_1 \tilde{X}_1 + \tilde{U}$.

$$\beta_1 = \tilde{\beta}_1 = \frac{E(\tilde{X}_1 \tilde{Y})}{E(\tilde{X}_1^2)} = \frac{E([X_1 - E(X_1)][Y - E(Y)])}{E([X_1 - E(X_1)])} = \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)}$$

We derived this same expression for β_1 before! We now have a way to generalize this process for regression models with multiple variables.

Omitting One Variable

Let $k = 2$, so that $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$. Suppose that you estimate:

$$Y = \beta_0^* + \beta_1^* X_1 + U^*,$$

where you maintain the BLP assumptions: $E(U^*) = 0$ and $E(X_1 U^*) = 0$.

$$\beta_1^* = \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)} = \beta_1 + \beta_2 \frac{\text{Cov}(X_1, X_2)}{\text{Var}(X_1)}$$

In general, it is not true that $\beta_1^* = \beta_1$.

- If we "control" for X_2 in the model, we change the coefficient on X_1. The two exceptions to this are if $\text{Cov}(X_1, X_2) = 0$ and/or $\beta_2 = 0$.
- Omitted variable bias can be a *huge* issue for causal inference.
  - Why? Suppose Y = earnings, X_1 = education level, X_2 = SES. We cannot interpret β_1^* as the "effect" of education on earnings given SES.
  - Alternatively, let X_2 = "ability". We may not be able to measure it!

Measurement Error

Let $Y = \beta_0 + \beta_1 X_1 + U$, but we only observe $\hat{X}_1 = X_1 + V$. For simplicity, assume that $E(V) = E(X_1 V) = E(UV) = 0$. We estimate the model:

$$Y = \beta_0^* + \beta_1^* \hat{X}_1 + U^*$$

where you maintain the BLP assumptions: $E(U^*) = 0$ and $E(X_1 U^*) = 0$.

$$\beta_1^* = \frac{\text{Cov}(\hat{X}_1, Y)}{\text{Var}(\hat{X}_1)} = \frac{\text{Var}(X_1)}{\text{Var}(X_1) + \text{Var}(V)} \beta_1$$

The quantity $\frac{\text{Var}(X_1)}{\text{Var}(X_1)+\text{Var}(V)}$ is called "attenuation bias".

- *Note:* the *attenuation bias* is bounded between 0 and 1.
- Therefore, β_1^* will be smaller in magnitude than β_1.
- Again, this can become a *huge* issue when making causal inferences.

Powers of Regressors

Even if the relationship between Y and X is believed to be nonlinear, linear regression can still be useful. As an example, let $Y =$ wages and $X =$ age.

- We might think that wages rise when you are young and then fall as you transition toward retirement (i.e. wage-age profile is *concave*).
- *Strategy:* account for nonlinearities with a quadratic term X^2.

Suppose you write down the multiple regression model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + U$$

Our BLP assumptions require that $E(U) = E(XU) = E(X^2 U) = 0$.

- We could even add in cubic or quartic terms (e.g. X^3 or X^4).
- Conveniently for us, perfect multicollinearity is not an issue even though the regressors are deterministic functions of one another.

Categorical Variables

Due to issues surrounding perfect multicollinearity, we must be careful when dealing with categorical variables as regressors. For example, let:

$$X_1 = \mathbb{I}\{\text{didn't graduate high school}\}$$
$$X_2 = \mathbb{I}\{\text{graduated high school, but didn't graduate college}\}$$
$$X_3 = \mathbb{I}\{\text{graduated college, but no higher degrees}\}$$
$$X_4 = \mathbb{I}\{\text{higher degrees}\}$$

Since $X_4 = 1 - X_1 - X_2 - X_3$, we cannot put all four regressors in the model. Instead, we need to leave one of these variables (e.g. X_4) out:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U,$$

The BLP assumptions require: $E(U) = E(X_1 U) = E(X_2 U) = E(X_3 U) = 0$.

- Alternatively, we could regress Y on $X_1, \ldots, X_4$ without a constant.

Interaction Terms

Another type of nonlinear transformation of regressors is their product.

- *Example:* suppose that $X_1 = \mathbb{I}\{\text{female}\}$, $X_2 =$ avg. daily hours worked, and $Y =$ amount of TV watching. You believe that the relationship between work hours and TV watching differs by gender.
- *Strategy:* put an interaction term $X_1 X_2$ into the model.

Suppose you write down the multiple regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + U$$

Our BLP assumptions require: $E(U) = E(X_1 U) = E(X_2 U) = E(X_1 X_2 U) = 0$.

- We can also interact different categorical variables.
- Just as before, perfect multicollinearity is not an issue even though the third variable $X_1 X_2$ depends deterministically on X_1 and X_2.

Logarithms

It is common to take the natural log of the regressand and/or regressors.

- Why take a "log-transform"? Logarithms approximate proportional changes.
- Let x and $\tilde{x}$ be numbers with $\tilde{x} - x$ "small". Then $\frac{\tilde{x}-x}{x} \approx \log(\tilde{x}) - \log(x)$.

Example

Let $W =$ wages and $S =$ years of schooling. You consider the model:

$$\log(W) = \beta_0 + \beta_1 S + U$$

Suppose S increases by ΔS years. Then $\log(W)$ increases by $\beta_1 \Delta S$. In this case, the percentage change in wages is then given by:

$$100 \times \frac{W \exp(\beta_1 \Delta S) - W}{W} \approx 100 \times [\log(W \exp(\beta_1 \Delta S)) - \log(W)] \approx 100 \times \beta_1 \Delta S$$

Fixing U, an additional year of schooling S changes W by $100\beta_1\%$.

Level-Log and Log-Log Models

Other possible models relating Y to X and U are:

$$Y = \beta_0 + \beta_1 \log(X) + U \quad (1)$$
$$\log(Y) = \beta_0 + \beta_1 \log(X) + U \quad (2)$$

The first model (1) is called a *level-log* model.

- Holding U fixed, a 1% increase in X changes Y by $\beta_1/100$.

The second model (2) is called a *log-log* model.

- Holding U fixed, a 1% increase in X changes Y by $\beta_1\%$.

In practice, these log approximation interpretations are not too good.

- Approximations become better when we look at small changes in X.
- Used frequently economics when considering *elasticities* of wages.