ECON 251 Discussion

Multiple Linear Regression

Elird Haxhiu

Fall 2022

Outline

1. Mincer (1972) regression framework
2. Why use multiple (vs simple) linear regression?
3. Omitted variable bias (OVB) formula
4. Multiple linear regression (MLR) model assumptions
5. Ordinary least squares (OLS) estimator
6. Main theorems on bias, consistency, and efficiency
7. Hypothesis tests: t and F
8. Confidence intervals

Mincer (1972) regression framework

- $\log Y_i \geq 0$ denotes log earnings (outcome)
- $S_i \in \{0,1\}$ is whether i finished college (treatment)
- U_i is unobserved error term (ex: ability)
- simple linear population regression function (PRF) $\quad \log Y_i = \alpha + \beta \cdot S_i + U_i$
- $\beta \approx$ Mincer (1972) returns to college

- $\log Y_i \geq 0$ denotes log earnings (outcome)
- $S_i \in \{0,1\}$ is whether i finished college (treatment)
- U_i is unobserved error term (ex: ability)
- simple linear population regression function (PRF) $\log Y_i = \alpha + \beta \cdot S_i + U_i$
- $\beta \approx$ Mincer (1972) returns to college

- Potential outcomes + treatment effects $\text{ATE} \coloneqq E[\log Y_i(1) - \log Y_i(0)]$

- Need independence to *identify* ATE with simple comparison, which is given by

$$S_i \perp \log Y_i(1), \log Y_i(0)$$
$$\Leftrightarrow E[U_i | S_i] = E[U_i] = 0$$

$$\hat{\beta}^{OLS}$$

- $\log Y_i \geq 0$ denotes log earnings (outcome)
- $S_i \in \{0,1\}$ is whether i finished college (treatment)
- U_i is unobserved error term (ex: ability)
- simple linear population regression function (PRF) $\quad \log Y_i = \alpha + \beta \cdot S_i + U_i$
- $\beta \approx$ Mincer (1972) returns to college

- Potential outcomes + treatment effects $\quad \text{ATE} \coloneqq E[\log Y_i(1) - \log Y_i(0)]$

- Need independence to *identify* ATE with simple comparison, which is given by

$$S_i \perp \log Y_i(1), \log Y_i(0)$$
$$\Leftrightarrow E[U_i | S_i] = E[U_i] = 0$$

$$\begin{aligned}
\hat{\beta}^{OLS} &= \hat{E}[\log Y_i \,|S_i = 1] - \hat{E}[\log Y_i \,|S_i = 0] \\
&= \overline{\log Y_1} - \overline{\log Y_0} \\
&= \frac{1}{N_1}\sum_{i|S_i=1} \log Y_i - \frac{1}{N_0}\sum_{i|S_i=0} \log Y_i
\end{aligned}$$

Why use multiple (vs simple) linear regression?

- Multiple regression: we can get closer to satisfying the hypothetical (but necessary, and luckily also sufficient assumption known as) random assignment/independence by conditioning on some observable characteristics X_i (provocative example: IQ test score)

Why use multiple (vs simple) linear regression?

- Multiple regression: we can get closer to satisfying the hypothetical (but necessary, and luckily also sufficient assumption known as) random assignment/independence by conditioning on some observable characteristics X_i (provocative example: IQ test score)

$$\log Y_i = \alpha + \beta \cdot S_i + \delta \cdot X_i + U_i$$

$$S_i \perp \log Y_i(1), \log Y_i(0) \mid X_i$$
$$\Leftrightarrow E[U_i \mid S_i, X_i] = 0$$

- Inclusion of X_i allows us to "control" for any reasons why there may not be truly random assignment of treatment (in simple PRF)

Omitted variable bias (OVB)

"True" model	$\log Y_i = \alpha + \beta \cdot S_i + \delta \cdot X_i + U_i$	$\text{Cov}(S_i, U_i) = 0$
Our model	$\log Y_i = a + b \cdot S_i + E_i$	
Auxiliary model	$X_i = c + \boldsymbol{\gamma} \cdot S_i + \eta_i$	

Omitted variable bias (OVB)

"True" model $\log Y_i = \alpha + \beta \cdot S_i + \delta \cdot X_i + U_i$ $\quad \mathrm{Cov}(S_i, U_i) = 0$

Our model $\log Y_i = a + b \cdot S_i + E_i$

Auxiliary model $X_i = c + \boldsymbol{\gamma} \cdot S_i + \eta_i$

By (naively) assuming $\mathrm{Cov}(S_i, E_i) = 0$ in our model, the population value of slope is

$$b = \frac{\mathrm{Cov}(S_i, \log Y_i)}{\mathrm{Var}(S_i)}$$

$$= \beta + \delta \cdot \boldsymbol{\gamma}$$

Omitted variable bias (OVB)

“True” model $\log Y_i = \alpha + \beta \cdot S_i + \delta \cdot X_i + U_i$ $\mathrm{Cov}(S_i, U_i) = 0$

Our model $\log Y_i = a + b \cdot S_i + E_i$

Auxiliary model $X_i = c + \boldsymbol{\gamma} \cdot S_i + \eta_i$

By (naively) assuming $\mathrm{Cov}(S_i, E_i) = 0$ in our model, the population value of slope is

$$b = \frac{\mathrm{Cov}(S_i, \log Y_i)}{\mathrm{Var}(S_i)} = \frac{\mathrm{Cov}(S_i, \alpha + \beta \cdot S_i + \delta \cdot X_i + U_i)}{\mathrm{Var}(S_i)}$$

Omitted variable bias (OVB)

"True" model	$\log Y_i = \alpha + \beta \cdot S_i + \delta \cdot X_i + U_i$	$\mathrm{Cov}(S_i, U_i) = 0$
Our model	$\log Y_i = a + b \cdot S_i + E_i$	
Auxiliary model	$X_i = c + \boldsymbol{\gamma} \cdot S_i + \eta_i$	

By (naively) assuming $\mathrm{Cov}(S_i, E_i) = 0$ in our model, the population value of slope is

$$b = \frac{\mathrm{Cov}(S_i, \log Y_i)}{\mathrm{Var}(S_i)} = \frac{\mathrm{Cov}(S_i, \alpha + \beta \cdot S_i + \delta \cdot X_i + U_i)}{\mathrm{Var}(S_i)}$$

$$= \frac{\mathrm{Cov}(S_i, \alpha) + \mathrm{Cov}(S_i, \beta S_i) + \mathrm{Cov}(S_i, \delta X_i) + \mathrm{Cov}(S_i, U_i)}{\mathrm{Var}(S_i)}$$

Omitted variable bias (OVB)

"True" model $\quad \log Y_i = \alpha + \beta \cdot S_i + \delta \cdot X_i + U_i \quad \mathrm{Cov}(S_i, U_i) = 0$

Our model $\quad \log Y_i = a + b \cdot S_i + E_i$

Auxiliary model $\quad X_i = c + \boldsymbol{\gamma} \cdot S_i + \eta_i$

By (naively) assuming $\mathrm{Cov}(S_i, E_i) = 0$ in our model, the population value of slope is

$$b = \frac{\mathrm{Cov}(S_i, \log Y_i)}{\mathrm{Var}(S_i)} = \frac{\mathrm{Cov}(S_i, \alpha + \beta \cdot S_i + \delta \cdot X_i + U_i)}{\mathrm{Var}(S_i)}$$

$$= \frac{\mathrm{Cov}(S_i, \alpha) + \mathrm{Cov}(S_i, \beta S_i) + \mathrm{Cov}(S_i, \delta X_i) + \mathrm{Cov}(S_i, U_i)}{\mathrm{Var}(S_i)}$$

$$= \frac{\beta \cdot \mathrm{Var}(S_i) + \delta \cdot \mathrm{Cov}(S_i, X_i)}{\mathrm{Var}(S_i)}$$

Omitted variable bias (OVB)

"True" model $\log Y_i = \alpha + \beta \cdot S_i + \delta \cdot X_i + U_i$ $\quad \mathrm{Cov}(S_i, U_i) = 0$

Our model $\log Y_i = a + b \cdot S_i + E_i$

Auxiliary model $X_i = c + \boldsymbol{\gamma} \cdot S_i + \eta_i$

By (naively) assuming $\mathrm{Cov}(S_i, E_i) = 0$ in our model, the population value of slope is

$$b = \frac{\mathrm{Cov}(S_i, \log Y_i)}{\mathrm{Var}(S_i)} = \frac{\mathrm{Cov}(S_i, \alpha + \beta \cdot S_i + \delta \cdot X_i + U_i)}{\mathrm{Var}(S_i)}$$

$$= \frac{\mathrm{Cov}(S_i, \alpha) + \mathrm{Cov}(S_i, \beta S_i) + \mathrm{Cov}(S_i, \delta X_i) + \mathrm{Cov}(S_i, U_i)}{\mathrm{Var}(S_i)}$$

$$= \frac{\beta \cdot \mathrm{Var}(S_i) + \delta \cdot \mathrm{Cov}(S_i, X_i)}{\mathrm{Var}(S_i)} = \beta + \delta \cdot \frac{\mathrm{Cov}(S_i, X_i)}{\mathrm{Var}(S_i)}$$

Omitted variable bias (OVB)

"True" model $\log Y_i = \alpha + \beta \cdot S_i + \delta \cdot X_i + U_i$ $\quad \mathrm{Cov}(S_i, U_i) = 0$

Our model $\log Y_i = a + b \cdot S_i + E_i$

Auxiliary model $X_i = c + \boldsymbol{\gamma} \cdot S_i + \eta_i$

By (naively) assuming $\mathrm{Cov}(S_i, E_i) = 0$ in our model, the population value of slope is

$$b = \frac{\mathrm{Cov}(S_i, \log Y_i)}{\mathrm{Var}(S_i)} = \frac{\mathrm{Cov}(S_i, \alpha + \beta \cdot S_i + \delta \cdot X_i + U_i)}{\mathrm{Var}(S_i)}$$

$$= \frac{\mathrm{Cov}(S_i, \alpha) + \mathrm{Cov}(S_i, \beta S_i) + \mathrm{Cov}(S_i, \delta X_i) + \mathrm{Cov}(S_i, U_i)}{\mathrm{Var}(S_i)}$$

$$= \frac{\beta \cdot \mathrm{Var}(S_i) + \delta \cdot \mathrm{Cov}(S_i, X_i)}{\mathrm{Var}(S_i)} = \beta + \delta \cdot \frac{\mathrm{Cov}(S_i, X_i)}{\mathrm{Var}(S_i)} = \beta + \delta \cdot \boldsymbol{\gamma}$$

Assumptions

- MLR1 (linear outcome model) $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i$
- MLR2 (random sampling) $\{Y_i, X_{i1}, \dots, X_{ik}\}_{i=1}^{N}$ is random draw
- MLR3 (no collinearity) no X_{ij} linear function of any other X_{il}

Assumptions

- MLR1 (linear outcome model) $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i$
- MLR2 (random sampling) $\{Y_i, X_{i1}, \ldots, X_{ik}\}_{i=1}^{N}$ is random draw
- MLR3 (no collinearity) no X_{ij} linear function of any other X_{il}

- MLR4 (independence) $E[U_i | X_{i1}, \ldots, X_{ik}] = 0$

Assumptions

- MLR1 (linear outcome model) $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i$
- MLR2 (random sampling) $\{Y_i, X_{i1}, \dots, X_{ik}\}_{i=1}^{N}$ is random draw
- MLR3 (no collinearity) no X_{ij} linear function of any other X_{il}

- MLR4 (independence) $E[U_i | X_{i1}, \dots, X_{ik}] = 0$

- MLR5 (homoskedasticity) $\text{Var}(U_i | X_{i1}, \dots, X_{ik}) = \sigma^2$
- MLR6 (normality) $U_i \sim N(0, \sigma^2)$

$$\Rightarrow Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}, \sigma^2)$$

Ordinary Least Squares (OLS) Estimator

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i$$

$$\min_{\{\beta_0, \beta_1, \ldots, \beta_k\}} \frac{1}{N} \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})^2$$

$$\Rightarrow \hat{\beta}_j^{OLS}$$

Ordinary Least Squares (OLS) Estimator

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i$$

$$\min_{\{\beta_0,\beta_1,\ldots,\beta_k\}} \frac{1}{N}\sum_{i=1}^{N}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})^2$$

$$\Rightarrow \hat{\beta}_j^{OLS} = \frac{\widehat{\mathrm{Cov}}(\tilde{X}_{ij}, Y_i)}{\widehat{\mathrm{Var}}(\tilde{X}_{ij})} \qquad \forall j = \{0,1,\ldots,k\}$$

$$= \frac{\widehat{\mathrm{Cov}}\left(X_{ij} - \hat{\theta}_1 X_{i1} - \cdots - \hat{\theta}_k X_{ik}, Y_i\right)}{\widehat{\mathrm{Var}}\left(X_{ij} - \hat{\theta}_1 X_{i1} - \cdots - \hat{\theta}_k X_{ik}\right)}$$

by Frisch, Waugh, and Lovell (FWL) <3

OLS Results

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i$$

- T1 (unbiased) MLR1+2+3+4 $\Rightarrow$ $E\left[\widehat{\beta}_j^{\,OLS}\right] = \beta_j \quad \forall j = \{0,1,\dots,k\}$

OLS Results

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i$$

- T1 (unbiased) MLR1+2+3+4 $\Rightarrow$ $E\left[\hat{\beta}_j^{\ OLS}\right] = \beta_j \qquad \forall j = \{0,1,\ldots,k\}$

- T2 (efficient) MLR1+2+3+4+5 $\Rightarrow$ $E\left[\hat{\beta}_j^{\ OLS}\right] = \beta_j \qquad \forall j = \{0,1,\ldots,k\}$

(Gauss-Markov) $\mathrm{Var}\left[\hat{\beta}_j^{\ OLS}\right] \leq \mathrm{Var}\left[\hat{\beta}_j^{\ \text{other linear}}\right]$

OLS Results $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i$

- T1 (unbiased) MLR1+2+3+4 $\Rightarrow$ $E\left[\hat{\beta}_j^{OLS}\right] = \beta_j \quad \forall j = \{0,1,\ldots,k\}$

- T2 (efficient) MLR1+2+3+4+5 $\Rightarrow$ $E\left[\hat{\beta}_j^{OLS}\right] = \beta_j \quad \forall j = \{0,1,\ldots,k\}$

(Gauss-Markov)

$$\mathrm{Var}\left[\hat{\beta}_j^{OLS}\right] \leq \mathrm{Var}\left[\hat{\beta}_j^{\text{other linear}}\right]$$

$$\mathrm{Var}\left[\hat{\beta}_j^{OLS}\right] = \frac{\sigma^2}{\mathrm{Var}(X_{ij}) \cdot \left[1 - R^2_{\text{reg } X_j \text{ on all } X_k}\right]}$$

$$E[\hat{\sigma}^2] = E\left[\frac{1}{N-k-1} \sum_{i=1}^{N} \hat{U}_i^2\right] = \sigma^2$$

$$\mathrm{se}\left[\hat{\beta}_j^{OLS}\right] := \sqrt{\widehat{\mathrm{Var}}\left[\hat{\beta}_j^{OLS}\right]}$$

OLS Results $\quad Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i$

- T3 (efficient) MLR1+2+3+4+5+6 $\Rightarrow$ $\quad \widehat{\beta}_j^{OLS} \sim N\left(\beta_j, \mathrm{Var}\left[\beta_j\right]\right) \quad \forall j = \{0,1,\ldots,k\}$

(Classical)

OLS Results

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i$$

- T3 (efficient) MLR1+2+3+4+5+6 $\Rightarrow$ $\quad \widehat{\beta_j}^{OLS} \sim N\left(\beta_j, \mathrm{Var}\left[\beta_j\right]\right) \quad \forall j = \{0,1,\ldots,k\}$

(Classical)

$$\frac{\widehat{\beta_j}^{OLS} - \beta_j}{\mathrm{sd}[\beta_j]} \sim N(0,1)$$

$$\frac{\widehat{\beta_j}^{OLS} - \beta_j}{\mathrm{se}[\beta_j]} \sim t(N-k-1)$$

$$\mathrm{se}\left[\widehat{\beta_j}^{OLS}\right] \coloneqq \sqrt{\widehat{\mathrm{Var}}\left[\widehat{\beta_j}^{OLS}\right]}$$

t and F tests $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i$

- Individual hypothesis test about slope parameter (t-test)

$$H_0: \beta_j = 0$$

$$t_{\widehat{\beta_j}^{OLS}} := \frac{\widehat{\beta_j}^{OLS} - 0}{\mathrm{se}\left[\widehat{\beta_j}^{OLS}\right]} \sim t(N-k-1)$$

t and F tests

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i$$

- Individual hypothesis test about slope parameter (t-test)

$$H_0 : \beta_j = 0$$

$$t_{\widehat{\beta_j}^{OLS}} := \frac{\widehat{\beta_j}^{OLS} - 0}{\mathrm{se}\left[\widehat{\beta_j}^{OLS}\right]} \sim t(N-k-1)$$

- Joint hypothesis test about entire linear model

$$SSR_U := \sum_{i=1}^{N} \widehat{U}_i^{\,2} = \sum_{i=1}^{N} \left(Y_i - \widehat{Y}_i\right)^2$$

$$SSR_R := \sum_{i=1}^{N} \left(Y_i - \widehat{\beta}_0\right)^2$$

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$F := \frac{\frac{SSR_R - SSR_U}{k}}{\frac{SSR_R}{N-k-1}} \sim F(k, N-k-1)$$

Confidence Intervals

$$P\left[\widehat{\beta}_j^{OLS} - c_{\alpha} \cdot \mathrm{se}\left[\widehat{\beta}_j^{OLS}\right] \leq \beta_j \leq \widehat{\beta}_j^{OLS} + c_{\alpha} \cdot \mathrm{se}\left[\widehat{\beta}_j^{OLS}\right]\right] = 1 - \alpha$$

significance level (rate we tolerate Type 1 errors) $\alpha \in \{0.01, 0.05, 0.1\}$

critical value associated w/ α in distribution $c_{\alpha} \approx 1.96$ if 5%

estimated standard error $\mathrm{se}\left[\widehat{\beta}_j^{OLS}\right] \coloneqq \sqrt{\widehat{\mathrm{Var}}\left[\widehat{\beta}_j^{OLS}\right]}$

interpretation = ???

Confidence Intervals

$$P\left[\widehat{\beta}_j^{OLS} - c_\alpha \cdot \text{se}\left[\widehat{\beta}_j^{OLS}\right] \leq \beta_j \leq \widehat{\beta}_j^{OLS} + c_\alpha \cdot \text{se}\left[\widehat{\beta}_j^{OLS}\right]\right] = 1 - \alpha$$

significance level (rate we tolerate Type 1 errors) $\alpha \in \{0.01, 0.05, 0.1\}$

critical value associated w/ α in distribution $c_\alpha \approx 1.96$ if 5%

estimated standard error $\text{se}\left[\widehat{\beta}_j^{OLS}\right] \coloneqq \sqrt{\widehat{\text{Var}}\left[\widehat{\beta}_j^{OLS}\right]}$

interpretation = this procedure to estimate bounds will cover true β_j parameter 95% of the time (over many hypothetical repeated samples)