

# Multivariate Regression

AECN 896-002

# Outline

1. Introduction
2. FWL theorem
3. Small Sample Property

# Multivariate Regression: Introduction

---

# Univariate vs Multivariate Regression Models

## Univariate

The most important assumption  $E[u|x] = 0$  (zero conditional mean) is almost always violated (unless your data comes from randomized experiments) because all the other variables are sitting in the error term, which can be correlated with  $x$ .

## Multivariate

More independent variables mean less factors left in the error term, which makes the endogeneity problem **less** severe

## Bi-variate vs. Uni-variate

$$\text{Bi-variate } wage = \beta_0 + \beta_1 educ + \beta_2 exper + u_2$$

$$\text{Uni-variate } wage = \beta_0 + \beta_1 educ + u_1 (= u_2 + \beta_2 exper)$$

## What's different?

- **bi-variate**: able to measure the effect of education on wage, holding experience fixed because experience is modeled explicitly ( We say *exper* is controlled for. )
- **uni-variate**:  $\hat{\beta}_1$  is biased unless experience is uncorrelated with education because experience was in error term

### Another example

The impact of per student spending (`expend`) on standardized test score (`avgscore`) at the high school level

$$avgscore = \beta_0 + \beta_1 expend + u_1 (= u_2 + \beta_2 avginc)$$

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u_2$$

# Model with two independent variables

More generally,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- $\beta_0$ : intercept
- $\beta_1$ : measure the change in  $y$  with respect to  $x_1$ , holding other factors fixed
- $\beta_2$ : measure the change in  $y$  with respect to  $x_2$ , holding other factors fixed

# The Crucial Condition (Assumption) for Unbiasedness of the OLS Estimator

## Uni-variate

For  $y = \beta_0 + \beta_1 x + u$ ,

$$E[u|x] = 0$$

## Bi-variate

For  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ ,

- Mathematically:  $E[u|x_1, x_2] = 0$
- Verbally: for any values of  $x_1$  and  $x_2$ , the expected value of the unobservables is zero



### Mean independence condition: example

In the following wage model,

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

Mean independence condition is

$$E[u|educ, exper] = 0$$

**Verbally:** this condition would be satisfied if innate ability of students is on average unrelated to education level and experience.

# The model with $k$ independent variables

## Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

## Mean independence assumption?

$\beta_{OLS}$  (OLS estimators of  $\beta$ s) is unbiased if,

$$E[u|x_1, x_2, \dots, x_k] = 0$$

**Verbally:** this condition would be satisfied if the error term is uncorrelated with any of the independent variables,  $x_1, x_2, \dots, x_k$ .

# Deriving OLS estimators

## OLS

Find the combination of  $\beta$ s that minimizes the sum of squared residuals

## So,

Denoting the collection of  $\hat{\beta}$ s as  $\hat{\theta} (= \{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k\})$ ,

$$\text{Min}_{\theta} \sum_{i=1}^n \left[ y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \dots + \hat{\beta}_k x_{k,i}) \right]^2$$

Find the FOCs by partially differentiating the objective function (sum of squared residuals) wrt each of  $\hat{\theta}(= \{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k\})$ ,

$$\sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \dots + \beta_k x_{k,i})) = 0 \quad (\hat{\beta}_0)$$

$$\sum_{i=1}^n x_{i,1} \left[ y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \dots + \beta_k x_{k,i}) \right] = 0 \quad (\hat{\beta}_1)$$

$$\sum_{i=1}^n x_{i,2} \left[ y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \dots + \beta_k x_{k,i}) \right] = 0 \quad (\hat{\beta}_2)$$

$\vdots$

$$\sum_{i=1}^n x_{i,k} \left[ y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \dots + \beta_k x_{k,i}) \right] = 0 \quad (\hat{\beta}_k)$$

Or more succinctly,

$$\sum_{i=1}^n \hat{u}_i = 0 \quad (\hat{\beta}_0)$$

$$\sum_{i=1}^n x_{i,1} \hat{u}_i = 0 \quad (\hat{\beta}_1)$$

$$\sum_{i=1}^n x_{i,2} \hat{u}_i = 0 \quad (\hat{\beta}_2)$$

$\vdots$

$$\sum_{i=1}^n x_{i,k} \hat{u}_i = 0 \quad (\hat{\beta}_k)$$

# Implementation of multivariate OLS

## R code: Implementation in R

```
#--- load the fixest package ---#
library(fixest)

#--- generate data ---#
N <- 100 # sample size
x1 <- rnorm(N) # independent variable
x2 <- rnorm(N) # independent variable
u <- rnorm(N) # error
y <- 1 + x1 + x2 + u # dependent variable
data <- data.frame(y = y, x1 = x1, x2 = x2)

#--- OLS ---#
reg <- feols(y ~ x1 + x2, data = data)

## print the results
reg
```

```
## OLS estimation, Dep. Var.: y
## Observations: 100
## Standard-errors: IID
##           Estimate Std. Error  t value   Pr(>|t|)
## (Intercept) 0.813314   0.108173   7.51865 2.7875e-11 ***
## x1           1.067599   0.100276  10.64664 < 2.2e-16 ***
## x2           1.029790   0.112995   9.11362 1.1000e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Presenting regression results

When you are asked to present regression results in assignments or your final paper, use the `msummary()` function from the `modelsummary` package.

## Example

```
## load the package (isntall it if you have no  
library(modelsummary)  
  
## run regression  
reg_results <- feols(speed ~ dist, data = cars)  
  
## report regression table  
msummary(  
  reg_results,  
  # keep these options as they are  
  stars = TRUE,  
  gof_omit = "IC|Log|Adj|F|Pseudo|Within"  
)
```

Model 1	
(Intercept)	8.284***
	(0.874)
dist	0.166***
	(0.017)
Num.Obs.	50
R2	0.651
Std. errors	IID
* p < 0.1, ** p < 0.05, *** p < 0.01	

# Frisch–Waugh–Lovell Theorem

---



Consider the following simple model,

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

Suppose you are interested in estimating only  $\beta_1$ .

Let's consider the following two methods,

### Method 1: Regular OLS

Regress  $y$  on  $x_1$ ,  $x_2$ , and  $x_3$  with an intercept to estimate  $\beta_0, \beta_1, \beta_2, \beta_3$  at the same time (just like you normally do)

### Method 2: 3-step

- regress  $y$  on  $x_2$  and  $x_3$  with an intercept and get residuals, which we call  $\hat{u}_y$
- regress  $x_1$  on  $x_2$  and  $x_3$  with an intercept and get residuals, which we call  $\hat{u}_{x_1}$
- regress  $\hat{u}_y$  on  $\hat{u}_{x_1}$  ( $\hat{u}_y = \alpha_1 \hat{u}_{x_1} + v_3$ )

### Frisch-Waugh-Lovell theorem

Methods 1 and 2 produces the same coefficient estimate on  $x_1$

$$\hat{\beta}_1 = \hat{\alpha}_1$$

# Partialing out Interpretation from Method 2

## Step 1

Regress  $y$  on  $x_2$  and  $x_3$  with an intercept and get residuals, which we call  $\hat{u}_y$

- $\hat{u}_y$  is void of the impact of  $x_2$  and  $x_3$  on  $y$

## Step 2

Regress  $x_1$  on  $x_2$  and  $x_3$  with an intercept and get residuals, which we call  $\hat{u}_{x_1}$

- $\hat{u}_{x_1}$  is void of the impact of  $x_2$  and  $x_3$  on  $x_1$

## Step 3

Regress  $\hat{u}_y$  on  $\hat{u}_{x_1}$ , which produces an estimate of  $\beta_1$  that is identical to that you can get from regressing  $y$  on  $x_1$ ,  $x_2$ , and  $x_3$

# Interpretation

- Regressing  $y$  on all explanatory variables ( $x_1$ ,  $x_2$ , and  $x_3$ ) in a multivariate regression is as if you are looking at the impact of a single explanatory variable with the effects of all the other effects partiled out
- In other words, including variables beyond your variable of interest lets you **control for (remove the effect of)** other variables, avoiding confusing the impact of the variable of interest with the impact of other variables.

# Small Sample Properties of OLS Estimators

---

## Unbiasedness of OLS Estimator

OLS estimators of multivariate models are unbiased under **certain** conditions

### Condition 1

Your model is correct (Assumption  $MLR.1$ )

### Condition 2

Random sampling (Assumption  $MLR.2$ )

### Conditions 3

No perfect collinearity (Assumption  $MLR.3$ )

# Perfect Collinearity

## No Perfect Collinearity

Any variable cannot be a linear function of the other variables

## Example (silly)

$$wage = \beta_0 + \beta_1 educ + \beta_2(3 \times educ) + u$$

( More on this later when we talk about dummy variables)



## Zero Conditional Mean

$$E[u|x_1, x_2, \dots, x_k] = 0 \quad (\text{Assumption MLR.4})$$

## Unbiasedness of OLS estimators

If all the conditions  $MLR.1 \sim MLR.4$  are satisfied, OLS estimators are unbiased.

$$E[\hat{\beta}_j] = \beta_j \quad \forall j = 0, 1, \dots, k$$

## Endogeneity (Definition)

$$E[u|x_1, x_2, \dots, x_k] = f(x_1, x_2, \dots, x_k) \neq 0$$

## What could cause endogeneity problem?

- functional form misspecification

$$wage = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + u_1 \quad (\text{true})$$

$$wage = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u_2 (= \log(x_1) - x_1) \quad (\text{yours})$$

- omission of variables that are correlated with any of  $x_1, x_2, \dots, x_k$  ( [more on this soon](#) )
- [other sources of endogeneity later](#)

# Variance of the OLS estimators

## Homoeskedasticity

$$\text{Var}(u|x_1, \dots, x_k) = \sigma^2 \quad (\text{Assumption MLR.5})$$

## Variance

Under conditions *MLR.1* through *MLR.5*, conditional on the sample values of the independent variables,

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)},$$

where  $SST_j = \sum_{i=1}^n (x_{ji} - \bar{x}_j)^2$  and  $R_j^2$  is the R-squared from regressing  $x_j$  on all other independent variables including an intercept. ( [We will revisit this equation](#) )

# Estimating $\sigma^2$

Just like uni-variate regression, you need to estimate  $\sigma^2$  if you want to estimate the variance (and standard deviation) of the OLS estimators.

## uni-variate regression

$$\hat{\sigma}^2 = \sum_{i=1}^N \frac{\hat{u}_i^2}{n - 2}$$

## multi-variate regression

A model with  $k$  independent variables with intercept.

$$\hat{\sigma}^2 = \sum_{i=1}^N \frac{\hat{u}_i^2}{n - (k + 1)}$$

You solved  $k + 1$  simultaneous equations to get  $\hat{\beta}_j$  ( $j = 0, \dots, k$ ). So, once you know the value of  $n - k - 1$  of the residuals, you know the rest.