Multivariate Regression

Taro Mieno

AECN 896-003: Applied Econometrics

Univariate regression model

Drawback

The most important assumption $E[u|x]$ is almost always violated (unless you data comes from randomized experiments)

Multivariate regression model

Improvement over univariate regression model

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

Two independent variables

Bi-variate vs. uni-variate

$$\begin{aligned} wage =& \beta_0 + \beta_1 educ + \beta_2 exper + u_2 \\ wage =& \beta_0 + \beta_1 educ + u_1 (= u_2 + \beta_2 exper) \end{aligned}$$

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.)

univariate : $\hat{\beta}_1$ is biased unless experience is uncorrelated with education because experience was in error term

Two independent variables

Another example

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

$$
\begin{aligned}
avgscore =& \beta_0 + \beta_1 expend + u_1 (= u_2 + \beta_2 avginc) \\
avgscore =& \beta_0 + \beta_1 expend + \beta_2 avginc + u_2
\end{aligned}
$$

Two independent variables

More generally,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{1}$$

β_0 : intercept

β_1 : measure the change in y with respect to x_1, holding other factors fixed

β_2 : measure the change in y with respect to x_1, holding other factors fixed

The Crucial Condition (Assumption) for Unbiasedness of OLS

Uni-variate

$E[u|x] = 0$

Bi-variate

- 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

In the following wage model,

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

Mean independence condition is,

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

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

Multivariate regression model (in general)

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

The model with k independent variables

Multivariate regression model (in general)

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

Mean independence assumption?

β_{OLS} (OLS estimators of βs) is unbiased if,

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

Deriving OLS estimators

OLS

Find the combination of βs that minimizes the sum of squared residuals

So,

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

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

Deriving OLS estimators

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

$$\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \cdots + \beta_k x_{k,i}) = 0 \quad (\beta_0)$$

$$\sum_{i=1}^{n} x_{i,1}\left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \cdots + \beta_k x_{k,i})\right] = 0 \quad (\beta_1)$$

$$\sum_{i=1}^{n} x_{i,2}\left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \cdots + \beta_k x_{k,i})\right] = 0 \quad (\beta_2)$$

$$\vdots$$

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

Deriving OLS estimators

Or more succinctly,

$$\begin{aligned}
\sum_{i=1}^{n} \hat{u}_i =& 0 \quad (\beta_0) \\
\sum_{i=1}^{n} x_{i,1}\hat{u}_i =& 0 \quad (\beta_1) \\
\sum_{i=1}^{n} x_{i,2}\hat{u}_i =& 0 \quad (\beta_2) \\
&\vdots \\
\sum_{i=1}^{n} x_{i,k}\hat{u}_i =& 0 \quad (\beta_k)
\end{aligned}$$

Implementation of multivariate OLS

R code: Implementation in R

```r
#--- 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 <- lm(y~x1+x2,data=data)
reg_sum <- summary(reg) # get summary
reg_sum$coef # print out coef estimates
```

```
            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 1.010631 0.09123388 11.07736 6.448002e-19
x1          1.085433 0.07576744 14.32586 1.125846e-25
x2          1.101331 0.09299542 11.84285 1.510965e-20
```

A partialling out interpretation

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 β_1.

A partialling out interpretation

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 β_0, β_1, β_2, β_3 at the same time (just like you normally do)

Method 2: 3-step

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

A partialling out interpretation

Frisch-–Waugh-–Lovell theorem

Methods 1 and 2 produces the same coefficient estimate on x_1

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

A partialling out interpretation: Demonstration using R

R code: FWL theorem

```r
#--- data generation  ---#
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)

#--- method 1 (one-step) ---#
beta_m1 <- lm(y~x1+x2,data=data)$coef['x1'] # OLS

#--- method 3 (three-step) ---#
y_res <- lm(y~x2,data=data)$residuals # OLS
x1_res <- lm(x1~x2,data=data)$residuals # OLS
beta_m2 <- lm(y~x,data=data.frame(y=y_res,x=x1_res))$coef['x']
```

What does this mean?

Method 2: 3-step

1. $y_i = \gamma_0 + \gamma_2 x_2 + \gamma_3 x_3 + v_1$
2. $x_i = \delta_0 + \delta_2 x_2 + \delta_3 x_3 + v_2$
3. $\hat{v}_1 = \alpha_1 \hat{v}_2 + v_3$

Partialling out

- $\hat{\beta}_1$ is the impact of x_1 with the impacts of the other variables netted out!!
- including other variables (take them out from the error term)

Unbiasedness of OLS Estimators

Unbiasedness

OLS estimators of multivariate models are unbiased under certain conditions

Unbiasedness of OLS Estimators

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)

Unbiasedness of OLS Estimators

Zero Conditional Mean

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

Unbiasedness of OLS estimators

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

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

Endogeneity ($E[u|x_1, x_2, \ldots, x_k] \neq 0$)

What could cause endogeneity problem?

- functional form misspecification

$$
\begin{aligned}
wage =& \beta_0 + \beta_1 log(x_1) + \beta_2 x_2 + u_1 \quad \text{(true)} \\
wage =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u_2 (= log(x_1) - x_1) \quad \text{(yours)}
\end{aligned}
$$

- omission of variables that are correlated with any of $x_1, x_2, \ldots, x_k$ (more on this soon)
- other sources of enfogeneity later

Variance of the OLS estimators

Homoeskedasticity

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

Variance

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

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

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 σ^2

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

uni-variate regression

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

multi-variate regression (k independent variables with intercept)

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

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