Econometrics

Chapter 4 Multiple Regression: Estimation

In Choi

Sogang University

The model and examples

- Model

$$y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + u_t$$

  $\Rightarrow$ Several explanatory variables for y.
- β_0 is still the intercept
- β_1 to β_k all called slope parameters
- u_t is still the error term (or disturbance)
- Interpretation of β_j :

Impact of variable x_{jt} on y_t while other variables are being held constant.

The model and examples

Example

Earnings and education

$$\text{earnings}_i = \beta_0 + \beta_1 \text{education}_i + u_i$$

Here earnings_i is the $i-th$ individual's hourly earning and education_i the $i-th$ individual's number of years in school.
There may be omitted variables such as job experience, job experience^2, gender, marital status, etc. A more appropriate model is

$$\begin{aligned} \text{earnings}_i &= \beta_0 + \beta_1 \text{education}_i + \beta_2 \text{job experience}_i + \beta_3 \text{job experience}_i^2 \\ &\quad + \beta_4 \text{gender}_i + \beta_5 \text{marital status}_i + u_i \end{aligned}$$

The model and examples

Example

(continued) For gender and marital status, use dummy variables. That is,

$$\text{gender}_i = \begin{cases} 1 \text{ if male} \\ 0 \text{ if female} \end{cases}$$

$$\text{marital status}_i = \begin{cases} 1 \text{ if married} \\ 0 \text{ if single} \end{cases}$$

See, for example, Ashenfelter and Krueger, American Economic Review, 1994, 73–85.

The model and examples

Example

Class attendance and test scores

$$\begin{aligned} \text{score}_i &= \beta_0 + \beta_1 \left(\text{fraction of lectures attended}\right)_i \\ &\quad + \beta_2 \left(\text{fraction of problem sets completed}\right)_i + \varepsilon_i \end{aligned}$$

See Romer, Journal of Economic Perspectives, 1993.

Conditional expectation

Definition

For two continuous random variables, X and Y, we say that the conditional distribution of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$$

where $f(x,y)$ is the joint distribution of X and Y and $f_X(x)$ is the marginal distribution of X.

Remark

$f_{Y|X}(y|x)$ is a function of x and possibly a different probability distribution for each x.

Remark

When we wish to describe the entire family of distribution we use the phrase "the distribution of $Y \mid X$".

Conditional expectation

Remark

If X and Y are independent,

$$f_{Y|X}(y|x) = f_Y(y)$$

Definition

A conditional mean is the mean of the conditional distribution and is defined by

$$E[Y|X=x] = \begin{cases} \int_y y f_{Y|X}(y|x)\,dy & \text{if } y \text{ is continuous} \\ \sum_y y f_{Y|X}(y|x) & \text{if } y \text{ is discrete} \end{cases}$$

Example

Define the joint pdf of (X, Y) by

$$\begin{aligned} f(0,10) &= f(0,20) = \frac{2}{18}, \\ f(1,10) &= f(1,30) = \frac{3}{18}, \\ f(1,20) &= \frac{4}{18}, \\ f(2,30) &= \frac{4}{18}. \end{aligned}$$

The marginal pdf's are

$$\begin{aligned} f_X(0) &= f(0,10) + f(0,20) = \frac{4}{18} \\ f_X(1) &= f(1,10) + f(1,30) + f(1,20) = \frac{10}{18} \\ f_X(2) &= f(2,30) = \frac{4}{18}. \end{aligned}$$

Example

(continued) The conditional probability distribution of Y given that $X = 0$ is

$$f_{Y|X}(10 \mid 0) = \frac{f(0,10)}{f_X(0)} = \frac{2/18}{4/18} = \frac{1}{2}$$
$$f_{Y|X}(20 \mid 0) = \frac{f(0,20)}{f_X(0)} = \frac{2/18}{4/18} = \frac{1}{2}.$$

The conditional mean given that $X = 0$ is

$$E(Y \mid X = 0) = 10 \times \frac{1}{2} + 20 \times \frac{1}{2} = 25.$$

In addition, $E(Y \mid X)$ is a random variable that takes different values depending on the value of X. (Try to tabulate its distribution!)

Conditional expectation

1. **Law of Iterated Expectations**

$$E[Y] = E\left[E\left[Y|X\right]\right]$$

This theorem holds because

$$\begin{aligned} EY &= \int\int yf(x,y)dxdy \\ &= \int \left[\int yf(y \mid x)dy\right] f_X(x)dx \\ &= \int \left[E(Y \mid x)\right] f_X(x)dx \\ &= E(Y \mid X). \end{aligned}$$

2.

$$E\left[g(Y)f(X)|X\right] = f(X)E\left[g(Y)|X\right]$$

Trivially, $E\left[g(Y)h(X)|X=x\right] = h(x)E\left[g(Y)|X=x\right]$. But this holds for any $x \in R$. Thus, the stated result holds. See, e.g., Ash (1972) for a formal proof.

Assumptions

1. $\{x_{it}\}$ is a sequence of random variables.
2. x_{it} is not linearly related to x_{jt} for any i and $j(\neq i)$. (No redundant information in regressors)
3. Zero conditional mean of the disturbance

$$E\left(u_t | x_{11}, x_{12}, \ldots x_{k(n-1)}, x_{kn}\right) = 0 \text{ for all } t.$$

Whatever values $x_{11}, x_{12}, \ldots x_{k(n-1)}, x_{kn}$ take, the mean of u_t is zero. This assumption implies

$$E\left(u_t\right) = 0$$

(this follows from the law of iterated expectation) and

$$Cov\left(u_t, x_{jt}\right) = 0 \text{ for any } t \text{ and } j.$$

Assumptions

4. Spherical disturbances

$$\begin{cases} Var\left(u_t | x_{11}, x_{12}, \ldots x_{k(n-1)}, x_{kn}\right) = \sigma^2 \text{ for all } t = 1, 2, \cdots, n \\ Cov\left(u_t u_s | x_{11}, x_{12}, \ldots x_{k(n-1)}, x_{kn}\right) = 0 \text{ for all } s \neq t. \end{cases}$$

The assumption of common variance for u_t is called homoskedasticity.

Least squares estimation

- The objective function for the least squares estimation is

$$S(\beta_0, ..., \beta_k) = \sum_{t=1}^{n} (y_t - \beta_0 - \beta_1 x_{1t} - \cdots - \beta_k x_{kt})^2 .$$

We need to minimize this function. The first–order conditions for the minimization are

$$\begin{aligned}
\frac{\partial S(\beta_0, ..., \beta_k)}{\partial \beta_0} &= -2 \sum_{t=1}^{n} (y_t - \beta_0 - \beta_1 x_{1t} - \cdots - \beta_k x_{kt}) = 0 \\
&\vdots \\
\frac{\partial SS(\beta_0, ..., \beta_k)}{\partial \beta_k} &= -2 \sum_{t=1}^{n} x_{kt} (y_t - \beta_0 - \beta_1 x_{1t} - \cdots - \beta_k x_{kt}) = 0.
\end{aligned}$$

- The first–order conditions can be written as

$$\begin{gathered}\sum_{t=1}^{n}\left(y_t-\beta_0-\beta_1 x_{1t}-\cdots-\beta_k x_{kt}\right)=0\\ \sum_{t=1}^{n} x_{1t}\left(y_t-\beta_0-\beta_1 x_{1t}-\cdots-\beta_k x_{kt}\right)=0\\ \vdots\\ \sum_{t=1}^{n} x_{kt}\left(y_t-\beta_0-\beta_1 x_{1t}-\cdots-\beta_k x_{kt}\right)=0\end{gathered}$$

or

$$\begin{gathered}\sum_{t=1}^{n} y_t=n\beta_0+\beta_1\sum_{t=1}^{n} x_{1t}+\cdots+\beta_k\sum_{t=1}^{n} x_{kt}\\ \sum_{t=1}^{n} x_{1t}y_t=\beta_0\sum_{t=1}^{n} x_{1t}+\beta_1\sum_{t=1}^{n} x_{1t}^2+\cdots+\beta_k\sum_{t=1}^{n} x_{1t}x_{kt}\\ \vdots\\ \sum_{t=1}^{n} x_{kt}y_t=\beta_0\sum_{t=1}^{n} x_{kt}+\beta_1\sum_{t=1}^{n} x_{kt}x_{1t}^2+\cdots+\beta_k\sum_{t=1}^{n} x_{kt}^2\end{gathered}$$

The solution of these equations are the OLS estimators of β_0, β_1 and β_k. These are denoted by

$$\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k.$$

Least squares estimation

- Let

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & & & \\ 1 & x_{1n} & & x_{kn} \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

and

$$\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}.$$

Then, the equations are written as

$$X'y = (X'X)\beta.$$

Least squares estimation

- The solution of this equation is

$$\hat{\beta} = (X'X)^{-1}X'y.$$

This is the celebrated formula of the OLS estimator.

Least squares estimation

- Let

$$\hat{u}_t = y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{1t} - \cdots - \hat{\beta}_k x_{kt}.$$

  We call these residuals. We may write this as

$$\begin{aligned} y &= \hat{\beta}_0 + \hat{\beta}_1 x_{1t} + \cdots + \hat{\beta}_k x_{kt} + \hat{u}_t \\ &= \hat{y}_t + \hat{u}_t. \end{aligned}$$

- The residuals satisfy the $k+1$ equations

$$\begin{aligned} \sum_{t=1}^{n} \hat{u}_t &= 0 \\ \sum_{t=1}^{n} x_{1t}\hat{u}_t &= 0 \\ &\vdots \\ \sum_{t=1}^{n} x_{kt}\hat{u}_t &= 0. \end{aligned}$$

Partial regression

1. Regress y_t on $1, x_{1t}, ..., x_{(j-1)t}, x_{(j+1)t}, ..., x_{kt}$ and let the resulting residual be y_t^{dx}.
2. Regress x_{jt} on $1, x_{1t}, ..., x_{(j-1)t}, x_{(j+1)t}, ..., x_{kt}$ and let the resulting residual be x_{jt}^{dx}.
3. Regress y_t^{dx} on x_{jt}^{dx} and get the OLS estimator $\hat{\gamma}_j$.
4. $\hat{\gamma}_j = \hat{\beta}_j$.

y_t^{dx} can be interpreted as the part of y_t that is devoid of the effects of $x_{1t}, ..., x_{(j-1)t}, x_{(j+1)t}, ..., x_{kt}$. The same interpretation can be given to x_{jt}^{dx}. Thus, $\hat{\beta}_j$ measures the impact of x_{jt} on the part of y_t that cannot be explained by $x_{1t}, ..., x_{(j-1)t}, x_{(j+1)t}, ..., x_{kt}$. Or $\hat{\beta}_j$ measures the impact of x_{jt} on y_t while other variables are being held constant.

Multicollinearity

- If the regressors are perfectly or nearly correlated, the variance of the LSE becomes high to the extent that the regression results look unreliable. This is called the multicollinearity problem. When the regression model is subject to multicollinearity,

1. Small change in the data $\longrightarrow$ wide swings in the parameter estimates
2. High standard errors and insignificant individual coefficient estimates even though the coefficients estimates are jointly significant
3. Wrong coefficients signs or implausible magnitudes.

Goodness-of-fit measures

- R^2 (Coefficient of determination)

$$\begin{aligned} \text{TSS} &= \sum_{t=1}^{n} (y_t - \bar{y})^2 : \text{ Total Sum of Squares} \\ \text{ESS} &= \sum_{t=1}^{n} (\hat{y}_t - \bar{y})^2 : \text{ Explained Sum of Squares} \\ \text{RSS} &= \sum_{t=1}^{n} \hat{u}_t^2 : \text{ Residual Sum of Squares} \end{aligned}$$

$$\Rightarrow \quad R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

(i) R^2 always increases by an addition of extra explanatory variables.
(ii) R^2 is useful as a measure for model selection only when the candidate models have the same number of explanatory variables.

Goodness-of-fit measures

- Theil's $\bar{R}^2$ (adjusted R^2)

$$\bar{R}^2 = 1 - \frac{\sum_{t=1}^{n} \widehat{u}_t^2 / (n-k-1)}{\sum_{t=1}^{n} (y_t - \bar{y})^2 / (n-1)} = 1 - \frac{n-1}{n-k-1}\left(1 - R^2\right)$$

  $\bar{R}^2$ will fall (rise) when the variable x is deleted from the regression if the t–ratio associated with this variable is greater (less) than 1.

- Information criteria: goodness-of fit measure + penalty term for the # of parameters

$$AIC\,(k) = \ln \frac{\sum_{t=1}^{n} \widehat{u}_t^2}{n} + \frac{2(k+1)}{n} \text{ (Akaike's information criteria)}$$

$$BIC\,(k) = \ln \frac{\sum_{t=1}^{n} \widehat{u}_t^2}{n} + \frac{(k+1)\ln n}{n} \text{ (Bayesian information criteria)}$$

  The smaller, the better.

Unbiasedness

- For all $j = 0, 1, .., k$,

$$E(\hat{\beta}_j) = \beta_j.$$

Illustration Consider the simple regression

$$y_t = \beta_0 + \beta_1 x_t + u_t.$$

The least squares estimator of $\hat{\beta}_1$ can be written as

$$\hat{\beta}_1 = \beta_1 + \sum_{t=1}^{n} w_t u_t, \;\; w_t = (x_t - \bar{x}) / \sum_{t=1}^{n} (x_t - \bar{x})^2.$$

Thus

$$E(\hat{\beta}_1 \mid x_1, ..., x_n) = \beta_1 + \sum_{t=1}^{n} E(w_t u_t \mid x_1, ..., x_n)$$

$$= \beta_1 + \sum_{t=1}^{n} w_t E(u_t \mid x_1, ..., x_n) = \beta_1.$$

Using the law of iterated expectations, we obtain $E(\hat{\beta}_1) = \beta_1$.

Variance of the OLS estimator

$$Var(\hat{\beta}_j \mid all\ x) = \frac{\sigma^2}{\sum_{t=1}^{n}(x_{jt} - \bar{x}_j)^2(1 - R_j^2)}$$

where R_j^2 is the R^2 from regressing x_j on all other $x's$.

- The error variance: a larger σ^2 implies a larger variance for the OLS estimators
- The total sample variation: a larger $\sum_{t=1}^{n}(x_{jt} - \bar{x}_j)^2$ implies a smaller variance for the estimators
- Linear relationships among the independent variables: a larger R_j^2 implies a larger variance for the estimators
- σ^2 is estimated by

$$s^2 = \frac{1}{n-k-1}\sum_{t=1}^{n}\hat{u}_t^2.$$

Variance of the OLS estimator

- Let $\hat{u} = (\hat{u}_1, ..., \hat{u}_n)'$. Then, letting $P_X = X(X'X)^{-1}X'$ and $M_X = I - P_X$, we have
  $\hat{u} = y - X\hat{\beta} = y - X(X'X)^{-1}X'y = (I - P_X)y = M_X u$, because $M_X X = 0$.
- Moreover, $\hat{u}'\hat{u} = u' M_X u$, because $M_X' = M_X$ and $M_X M_X = M_X$.
- Thus

$$\begin{aligned} E\left[\hat{u}'\hat{u}|X\right] &= E\left[u' M_X u|X\right] \\ &= E\left[tr\left(u' M_X u\right)|X\right] \\ &= E\left[tr\left(M_X u\ u'\right)|X\right] \\ &= tr\left(M_X E\left(u\ u'|X\right)\right) \\ &= tr\left(M_X \sigma^2 I\right) \\ &= \sigma^2 tr\left(M_X\right). \end{aligned}$$

Variance of the OLS estimator

But

$$\begin{aligned} tr\left(M\right) &= tr\left[I_n - X\left(X'X\right)^{-1}X'\right] \\ &= tr\left(I_n\right) - tr\left(\left(X'X\right)^{-1}X'X\right) \\ &= tr\left(I_n\right) - tr\left(I_{k+1}\right) \\ &= n-k-1. \end{aligned}$$

These give

$$E(s^2) = EE\left[\hat{u}'\hat{u}/(n-k-1)|X\right] = \sigma^2.$$

The Gauss-Markov Theorem

- Given our assumptions it can be shown that OLS is "BLUE" (Best Linear Unbiased Estimator)

Thus, if the assumptions hold, use OLS.