CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

AACSB ACCREDITED
EFMD EQUIS ACCREDITED
EFMD EPAS ACCREDITED

MSF Chula

Financial Econometrics

Lecture III:
Linear Regression with Multiple Regressors

Narapong Srivisal, Ph.D.

Outline

Multivariate Linear Regression Model

- Review of the 3 Interpretations
- Solving for the Parameters
- Multicollinearity
- Non-linearity in Linear Regression Model

The OLS Estimator

- The OLS Estimator
- Unbiasedness and Consistency
- Goodness of Fit

Linear Regression with Multiple Regressors

- Model: let $(Y, X_1, \dots, X_k, U)$ be a random vector and $\beta_0, \beta_1, \dots, \beta_k$ be parameters such that

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + U$$

regressand

- Define $X = (1, X_1, \dots, X_k)$ and $\beta = (\beta_0, \beta_1, \dots, \beta_k)$, then we can write this as

same

$$Y = X'\beta + U$$

$\beta' \underline{x}$, $\beta \underline{x}$, $\underline{x} \beta$

- Similar to the bivariate model, we can interpret the linear relationship in three ways:
  - Linear Conditional Expectation
  - Best Linear Prediction
  - Causal Relationship

1st interpretation

Linear Conditional Expectation

- Assume that $E[Y|X_1, X_2, \dots, X_k] = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ and define

$$U = Y - (\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) = Y - E[Y|X_1, X_2, \dots, X_k]$$

- Or in matrix notation: $E[Y|X] = X'\beta$ and $U = Y - E[Y|X]$
- As before, this interpretation implies that
  - mean of U in each group X = 0
  - $E[U|X] = E[U|X_1, \dots, X_k] = 0$
  - $E[E[U|X]] = E[U]$, $E[0] = 0$
  - $E[U] = 0$
  - mean independent
  - $E[XU] = E[X] \cdot E[U] + cov(x,y)$, $= 0$, $= 0$ because $E[U|X] = 0 = E[U] \Rightarrow cov(U,X)$ must be 0
  - $E[XU] = 0$, or $E[X_j U] = 0$ for $1 \le j \le k$
  - $Corr[X_j, U] = 0$
- In this case, β provides a convenient way of summarizing the conditional expectation, and X_j does NOT cause Y in this interpretation.

2nd interpretation

Best Linear Prediction of $E[Y|X]$

- The "best" linear prediction of $E[Y|X]$ means that we pick $(\beta_0, \beta_1, \dots, \beta_k)$ to minimize

$$E[(E[Y|X_1, \dots, X_k] - (\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k))^2]$$

- Similar to the bivariate case, we have the condition (in matrix notation) $E[XU] = 0$, where X is a vector of $k+1$ entries, and U is a scalar. So, this condition is a system of $k+1$ equations:

$$E[U] = 0, E[X_1 U] = 0, \dots, E[X_k U] = 0$$

$\Rightarrow \mathrm{cov}(U, X) = 0$
↳ for regressor

- Just like the bivariate model, in this case, $(\beta_0, \beta_1, \dots, \beta_k)$ provides a convenient way of summarizing **the "best" *linear*** function of X to predict Y, although $E[Y|X]$ may not be linear.
- We can think of the regressors as a set of indicators or signals for Y.
- Again, X_j does NOT cause Y under this interpretation.

Causal Model

- For this case, we assume, based on some sort of "theoretical models," that Y is a linear function of observed variables $X_1, \dots, X_k$ and unobserved determinant U

$$Y = g(X_1, \dots X_k, U) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + U = X'\beta + U$$

- X causes Y in this interpretation, and each β_j describes the change in Y when X_j changes by one unit, given other factors constant (including other regressors $X_{j'}$, for $j' \neq j$).
- Terminology: people use ***ceteris paribus*** to mean **holding all other factors fixed**
- Similar to the bivariate model, we can always normalize β_0 to have $E[U] = 0$.
- In this case, it's not necessary that $E[X_j U], E[U|X_j]$, or $E[U|X]$ is zero.

Ceteris paribus → all of x constant

Pure effect / impact of x_1 on y

$$y = \beta_0 + \underset{\hat{\beta}}{\beta_1} x_1 + \underbrace{\beta_2 x_2 + \dots + \beta_k x_k}_{} = 0$$

↳ may have effect on y

" CONTROLS "

↳ should put enough controls

↳ read past literatures to justify your work

Solving for the Parameters $\boldsymbol{\beta_0, \dots, \beta_k}$

- We assume the following conditions
  1. $Y = X'\beta + U = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + U$
  2. $E[U] = 0$
  3. $E[X_j U] = 0$ for $1 \leq j \leq k$
  4. $E[XX'] < \infty$ and no perfect collinearity in X
- Solving for β: $E[XU] = 0 \;\Rightarrow 0 = E[X(Y - X'\beta)]$

$$\Rightarrow 0 = E[XY] - E[XX']\beta$$

$$\Rightarrow \boldsymbol{\beta = E[XX']^{-1}E[XY]}$$

- Notice that this is similar to the bivariate case:

$$\beta_1 = \frac{Cov[X_1, Y]}{Var[X_1]} = \left(E[X_1^2] - E[X_1]^2\right)^{-1}(E[X_1 Y] - E[X]E[Y])$$

- Notice that we want $E[XX]^{-1}$ to exist. This is true if the columns of the matrix $E[XX']$ are linearly independent, which is an implication of no perfect collinearity or no multicollinearity in X.

Problem ⇒ we don't have β to approximate ; β is not unique
it can be many value

Multicollinearity

- **Multicollinearity** or **Perfect collinearity** in X means that one of the entries of X can be expressed as linear combination of the others, e.g.

linear function of other x

$$X_1 = c_0 + c_2X_2 + c_3X_3 + \cdots + c_kX_k, \text{ or}$$

$$0 = (c_0 - X_1 + c_2X_2 + c_3X_3 + \cdots + c_kX_k)$$

- This implies that for any number $\boldsymbol{M}$,

$$Y = (\beta_0 + \boldsymbol{M}c_0) + (\beta_1 - \boldsymbol{M})X_1 + (\beta_2 + \boldsymbol{M}c_2)X_2 + \cdots + (\beta_k + \boldsymbol{M}c_k)X_k + U$$

- Then with multicollinearity, the linear regression model is not identified because each of the coefficients is not unique. So, each of them is not a fixed value to be estimated; it is not a parameter.

- If we have just one regressor, multicollinearity means that X is a constant and $Var[X] = 0$. Thus, assumption 4, $E[XX'] < \infty$, is the same as $0 < Var[X] < \infty$ for the bivariate case.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + 0$$

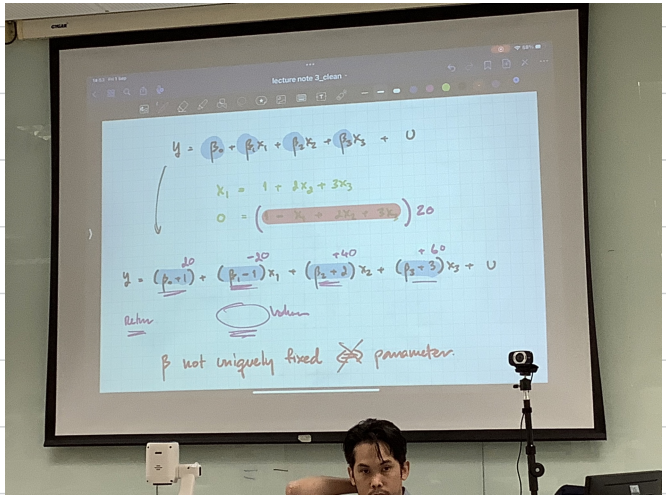
$$x_1 = 1 + 2x_2 + 3x_3$$

$$0 = (1 - x + 2x_2 + 3x_3)\,20$$

$$y = (\overset{20}{\beta_0 + 1}) + (\overset{-20}{\beta_1 - 1})\,x_1 + (\overset{+40}{\beta_2 + 2})\,x_2 + (\overset{+60}{\beta_3 + 3})\,x_3 + u$$

Return = Volume

β not uniquely fixed $\Leftrightarrow$ parameter



GPAX ⇒ MALE : average = 3.52

FEMALE : average = 3.60

M < 1 = Male, 0 = Female

F < 1 = Female, 0 = Male

regress GPAX on dummy variable "M"

1) $GPAX = \beta_0 + \beta_1 M + U$

if M = 1 ; $GPAX = (\beta_0 + \beta_1) + U$

$E[GPAX \mid male] = \beta_0 + \beta_1 = 3.52 \Rightarrow \beta_1 = -0.08$

difference btw mean of male and female's GPAX

if M = 0 ; $GPAX = \beta_0 + U$

$E[GPAX \mid female] = \beta_0 = 3.6$

2) $GPAX = \beta_0 + \beta_1 F + U$

$\beta_1 = 3.6 - 3.52 = 0.08$

β_0 : $E[GPAX \mid male] = 3.52$

3) $GPAX = \beta_0 M + \beta_1 F + U$

Male : M = 1, F = 0 } $GPAX = \beta_0 + U$

β_0 : $E[GPAX \mid male] = 3.52$

Female : M = 0, F = 1 } $GPAX = \beta_1 + U$

β_1 : $E[GPAX \mid female] = 3.60$

4) $GPAX = \beta_0 + \beta_1 M + \beta_2 F + U$

Male : M = 1, F = 0 } $GPAX = \underbrace{\beta_0 + \beta_1} + U$

$E[GPAX \mid male] = \beta_0 + \beta_1 = 3.52$

Female : M = 0, F = 1 } $GPAX = \underbrace{\beta_0 + \beta_2} + U$

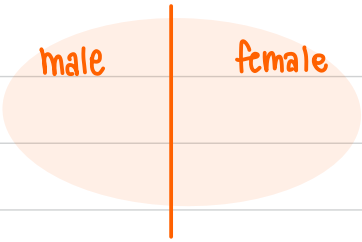
$E[GPAX \mid Female] = \beta_0 + \beta_2 = 3.6$

3 unknown variables from 2 equations

$\beta_0, \beta_1, \beta_2$ are not unique / fixed, can be any value

⇓

multicollinearly problem



	M	+	F	= 1
male	1	+	0	= 1
female	0	+	1	= 1

$\underbrace{M = 1 + F}_{\text{can be solved by drop 1 dummy variable}}$ ⇒ linear combination ⇒ multicollinearly problem!

Multicollinearity problem

$\Downarrow$ regressors are linear combination

β are not unique

$\Downarrow$

~~parameter~~

Solution $\Rightarrow$ dropping some X or constant term

- E.g. consider the following regression model

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 exp + \beta_3 male + \beta_4 female + U$$

where $male$ is a dummy variable taking value 1 if the subject is male, and $female$ is another dummy variable taking value 1 if the subject is female. The regressors in this case has multicollinearity, since

linear combination

$$male = 1 - female$$

special case of multicolinearity

This is sometimes referred to as the **dummy variable trap**: when you have dummies for all groups and constant term includes dummies for the same thing.

not linear function

$x_3 = x_2^2$

A | B | C | D

$1 = x^A + x^B + x^C + x^D$

dummy for A

- E.g. consider the model

$$\ln(wage) = \beta_0 + \beta_1 \overset{x_1}{educ} + \beta_2 \overset{x_2}{exp} + \beta_3 \underset{x_3}{exp^2} + \beta_4 female + U$$

There is no perfect collinearity in this model, although exp^2 is a function of exp, because it is not a linear relationship.

GPAX

A	B	C	D	F
3.9	3.78	3.5	2.9	2.5

based group

group that don't have dummy is captured by constant term

Don't have dummy variable trap because the function doesn't include all dummies for each regressor

1) $GPAX = \overset{2.5}{\beta_0} + \underbrace{\beta_1 A}_{1.4 \text{ from } 3.9-2.5} + \overset{1.28}{\beta_2 B} + \overset{1}{\beta_3 C} + \overset{0.4}{\beta_4 D} + U$

$E[GPAX \mid F] = \beta_0 = 2.5$

2) $GPAX = \overset{3.9}{\beta_0} + \beta_1 B + \beta_2 C + \beta_3 D + \beta_4 F + U$

captures diff btw B and based group

3) $GPAX = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + U$

D & F is based group

$2.9 + 2.5 = 5.4$

$y = \beta_0 + \beta_1 \underline{MALE} + \beta_2 MALE^2 + \dots$

dummy

$1 \longrightarrow 1$

$0 \longrightarrow 0$

even it is not linear combination

$MALE = (MALE)^2$ ✗ because MALE is dummy variable

- E.g. consider the following regression model

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 exp + \beta_3 age + U$$

In this case, there may be multicollinearity if the everybody starts school at the same age, say 6 years old, and they work every year that they are not in school, which implies

$$age = 6 + edu + exp$$

- If the data are not perfectly collinear, but nearly collinear, then it is possible to solve for β, but the standard errors of the OLS estimates will be very large.

age = 6 + edu + exp everyone has same life cycle form

↳ Have to drop some regressors until there's no linear combination

Non-linearity in Linear Regression

- Using a non-linear function of random variables such as log or polynomials as we saw in the bivariate case can still be applied in this case.
- For more than one regressor, we can also have nonlinearity by using an **interaction** (multiplication) of regressors as another regressor.

  random variable mutiply with another
- E.g. consider the following model in which $educ$ and exp denote years of education and working experience respectively.

$$wage = \beta_0 + \beta_1 educ + \beta_2 exp + \beta_3 \boldsymbol{exp \cdot educ} + U$$

- In this case, the effect of additional experience on wage varies with education:

$$\frac{\partial wage}{\partial exp} = \boldsymbol{\beta_2 + \beta_3 educ}$$

- If $\beta_3 > 0$, return on experience is increasing with years of education.
- If $\beta_3 < 0$, return to experience is decreasing with years of education.
- If $\beta_3 = 0$, education has no impact on the return on experience, and experience has no impact on the return on education as well.

$$\ln wage = \beta_0 + \beta_1 edu + \beta_2 exp + \beta_3 exp \cdot edu + u$$

What is effect of edu on wage?

$100\% \left(\frac{d\,wage}{wage} \right)$

$\underset{diff}{edu \uparrow} \Rightarrow \underbrace{(\beta_1 + \beta_3 exp)}_{\text{wage increases by}} 100\% \; dedu$

What is effect of exp on wage?

$exp \uparrow \Rightarrow (\beta_2 + \beta_3 edu)\, 100\%$

$$\ln[wage] = \beta_0 + \beta_1 exp + \beta_2 \underbrace{MALE}_{\text{dummy}} \cdot exp + u$$

1 male
0 female

dummy variable

different is the effect of exp between male and female

based group

impact of exp : $(\beta_1 + \beta_2 MALE)\, 100\%$

male : $(\beta_1 + \beta_2)\, 100\%$

female : $\beta_1 \; 100\%$

1 graduted from MSF
0 no

$$wage = \beta_0 + \beta_1 male + \beta_2 MSF + \beta_3 edu + \underbrace{\beta_4 male \cdot edu + \beta_5 MSF \cdot edu + \beta_6 male \cdot MSF \cdot edu}_{\text{effect of edu on wage for sure!}} + u$$

Effect of edu on wage : ?

diff by wage : $\beta_3 + \beta_4 male + \beta_5 MSF + \beta_6 MSF \cdot male$

* 2 dummies ⇒ 4 groups

	0 male	male 1
MSF : 0	female, not MSF β_3	male, not MSF $\beta_3 + \beta_4$
MSF : 1	female, MSF $\beta_3 + \beta_5$	male, MSF $\beta_3 + \beta_4 + \beta_5 + \beta_6$

$\beta_1 + \beta_2 male$

$y = \beta_0 + \beta_1 x + \beta_2 x \cdot male + \ldots$

	m	f
β_1	$\beta_1 + \beta_2$	β_1

What is the meaning of

we have to consider this, there're more than 1 dummy variable

1) β_4 : For non-MSF graduate, how much the effect of education on wage for male is higher than for female

2) $\beta_6 - \beta_5$: very complicate

3) $\beta_6 + \beta_4$: For MSF graduate, how much the effect of education on wage for male is higher than for female

- E.g. Consider the following model of the determinants of the number of financial products utilization (FQ) of households in rural areas:

$$FQ = \beta_0 + \beta_1 entp + \beta_2 income + U$$

where $entp$ is a dummy variable for a household having a business or household enterprise, and $income$ is the annual income.

- In this model, the effect of income on financial product utilization is simply β_2 and does not depend on whether the households conducting business.
- But it may be possible that household enterprises generally have to involve more transactions than the others, and more income from enterprise profit would even require more financial product utilization than simply higher salary income.
- Thereby, it is possible that the impact of income on financial product utilization is different between household enterprises and other households.
- We can tackle this problem in two ways:

Interaction with a dummy variable:

$$FQ_i = \beta_0 + \beta_1 \boldsymbol{entp_i} + \beta_2 income_i + \beta_3 \boldsymbol{entp_i}.income_i + U_i$$

- For household enterprises, $entp = 1$, therefore

$$FQ_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) income_i + U_i$$

  and the effect of income is $\beta_2 + \beta_3$

- For the other households, $entp = 0$, the effect of income is simply β_2 as

$$FQ_i = \beta_0 + \beta_2 income_i + U_i$$

- Notice that the coefficient β_3 of the interaction with a dummy variable captures the difference in effects between the two groups defined by the dummy variable.

Separate Regressions for Two Groups:

$FQ_i = \alpha_0 + \alpha_1 income_i + U_i$ if household i is an enterprise

$FQ_i = \gamma_0 + \gamma_1 income_i + U_i$ if household i is not an enterprise

- Note that the two methods give identical results:

$$\alpha_0 = \beta_0 + \beta_1, \alpha_1 = \beta_2 + \beta_3, \gamma_0 = \beta_0, \text{ and } \gamma_1 = \beta_2$$

The OLS Estimator

- Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be an iid sample from the distribution of (X, Y), where each X_i is a vector of one and k regressors $(1, x_1, \dots, x_k)$. Suppose that all the four assumptions stated before are true
- Then, we know the value of the parameters $\beta = E[XX']^{-1}E[XY]$
- So, a natural estimator of β is

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i{}'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right)$$

- Recall that β solves the system of equation $E[XU] = 0$. Similarly, $\hat{\beta}$ solves the sample counterpart

$$\frac{1}{n}\sum_{i=1}^{n} X_i \widehat{U}_i = \frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - X_i'\hat{\beta}\right) = 0$$

- which can be thought of the FOC of the following problem:

$$\min_{\hat{\beta}} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i'\hat{\beta}\right)^2 = \boxed{\min_{\hat{\beta}_0,\ldots,\hat{\beta}_k} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \cdots - \hat{\beta}_k x_k\right)^2}$$

- For this reason, the estimator $\hat{\beta}$ is called the ordinary least squares (OLS) of the parameter β

- As before, $\hat{Y}_i = X_i\hat{\beta}$ is the predicted value of Y_i, called the fitted value the ith observation

- $\hat{U}_i = Y_i - \hat{Y}_i$ is the residual for the ith observation

- Recall the immediate properties of OLS by construction (FOC) is

$$\frac{1}{n}\sum_{i=1}^{n}\hat{U}_i = \frac{1}{n}\sum_{i=1}^{n}x_{1i}\hat{U}_i = \frac{1}{n}\sum_{i=1}^{n}x_{2i}\hat{U}_i = \cdots = \frac{1}{n}\sum_{i=1}^{n}x_{ki}\hat{U}_i = 0$$

Properties of OLS: Bias?

- Recall that the OLS estimator is

$$\begin{aligned}\hat{\beta} &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right)\\ &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i (X_i'\beta + U_i)\right)\\ &= \beta + \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i U_i\right)\end{aligned}$$

- Hence,

$$\begin{aligned}E[\hat{\beta}|X_1,\dots,X_n] &= \beta + E\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i U_i\right)\middle| X_1,\dots,X_n\right]\\ &= \beta + \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i E[U_i|X_1,\dots X_n]\right)\\ &= \beta + \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \underbrace{\boldsymbol{E[U_i|X_i]}}_{=0}\right) = \beta\end{aligned}$$

OLS is unbiased

- Again, we need the assumption that $E[U|X] = 0$
- In words, **OLS will be unbiased if the error term is mean independent of all the regressors.**

Properties of OLS: Consistent?

Assume that our sample dataset is iid, $E[Y^4] < \infty, E\left[X_j^4\right] < \infty \; for \; 1 \leq j \leq k$, and there is no perfect collinearity, then we know that WLLN implies

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \xrightarrow{P} E[XX']$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i \xrightarrow{P} E[XY']$$

- Then, by the CMT,

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) \xrightarrow{P} E[XX']^{-1}E[XY]$$

- Since $E[XX']^{-1}E[X\boldsymbol{Y}] = E[XX']^{-1}E[X(\boldsymbol{X'\beta + U})] = \beta + E[XX']^{-1}\boldsymbol{E[XU]}$, we will have a consistent $\hat{\beta}$ OLS when $E[XU] = 0$.
- In words, **OLS will be consistent if each of the regressors is uncorrelated with the error term**.

- Recall that the first interpretation (Conditional Expectation) yields the properties $E[U|X] = 0$ and $E[XU] = 0$.
- Recall that the second interpretation (Best Linear Prediction) yields the property $E[XU] = 0$.
- For the ceteris-paribus causal effect, we can neither prove mathematically for $E[U|X] = 0$ nor $E[XU] = 0$.
- Therefore, **the OLS estimator is always unbiased under the first interpretation and always consistent under the first as well as the second interpretations**.
- For the causal model, we need to justify by some theories that the regressors are not correlated with the unobserved determinants.
- One way to help achieve that is to control for all possible factors that can affect Y by including them as regressors in the model to be estimated.

MSF Chula

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

AACSB ACCREDITED · EFMD EQUIS ACCREDITED · EFMD EPAS ACCREDITED

Research Question

Correlation vs Causation

software cannot help!

Model Construction

- No Multicollinearity
- Functional form: Non-linear, Interaction?
  ln as % change or level
- Control variables

Hard

Estimation

Easy if estimator is consistent / unbiased

Goodness of Fit: SER

- Goodness of fit tells us how well the model fit with the data we have.
- An easy measure of goodness of fit is the dispersion of dependent variable Y_i around the regression line $X_i'\beta$, or the dispersion of the error term U_i.
- A natural measure of the dispersion is standard deviation, but since we don't observe U_i, we need to use $\widehat{U}_i$ instead.
- An unbiased estimator of the standard deviation of the error term, estimated by $\widehat{U}_i$ is called the **standard error of the regression (SER)**, given by

want $\hat{U}_i^2$ to be small

$$\uparrow SER = \sqrt{\frac{1}{\underbrace{n}_{\text{\# of observations}} - \underbrace{k}_{\text{\# of } \beta \text{ in equation}} - 1} \sum_{i=1}^{n} \widehat{U}_i^2} \uparrow$$

- The lower the value, the less disperse and, thus, the better fit is the model.

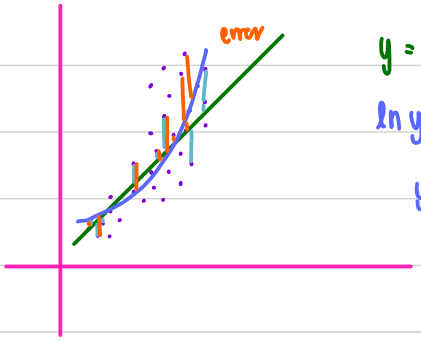
SER better fit

- One drawback of SER is that it has the same unit as the regressand and does not have a standard scale to use for comparison.

Which model line fit data more

Goodness of fit

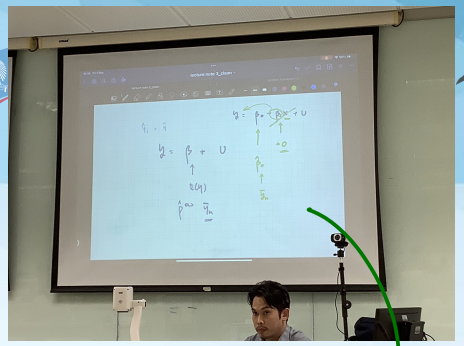
Population		Sample
Model	←	Estimate



$$y = \beta_0 + \beta_1 x$$

$$\ln y = \beta_0 + \beta_1 x$$

$$y = e^{\beta_0 + \beta_1 x}$$



Goodness of Fit: R^2

popular

- This is to measure how much variation in Y_i is explained by the variation in the fitted value $\hat{Y}_i$. Formally,

only $\hat{Y}_i = \bar{Y}_n$ for all observation

$(0, 0)$ x cannot explain anything of y

observations

$$R^2 = \frac{\sum_{i=1}^{n}\left[\left(\hat{Y}_i - \bar{Y}_n\right)^2\right]}{\sum_{i=1}^{n}\left[\left(Y_i - \bar{Y}_n\right)^2\right]}$$

same for all models from same population & always positive

same mean

data y of each observation

- The denominator is called **Explained Sum of Squares (ESS),** and the numerator is called **Total Sum of Squares (TSS)**
- Define **Residual sum of squares** as **RSS** $= \sum_{i=1}^{n} \hat{U}_i^2$
- It can be proved that $TSS = RSS + ESS$. Therefore,

worse fit

best fit

$$\boldsymbol{R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \in [0, 1]}$$

- The higher R^2, the better fit is the model, since $\sum_{i=1}^{n} \hat{U}_i^2$ is lower.

$$R^2 = 1 - \frac{\sum \hat{u}_i^2}{TSS}$$

OLS → $\min \sum \hat{u}^2$

prefer ← higher R^2

① $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$

② $Y = \beta_0 + \beta_1 X_1 + \underbrace{0}_{\text{fix } 0} \cdot X_2 + U$

R^2 always recommend more regressors

Goodness of Fit: $\bar{R}^2$

- A drawback of R^2 is that it is non-decreasing with an additional regressor because the OLS pick $\hat{\beta} = (1, \hat{\beta}_1, \dots \hat{\beta}_k, \hat{\beta}_{k+1})$ to minimize the RSS, and it can pick $\hat{\beta}_{k+1} = 0$ if adding X_{k+1} does not help.
- So, when comparing goodness of fit of multivariate models, econometricians prefer to use the **adjusted R^2** or $\bar{R}^2$, defined as

fixed

decrease not much if additional regressor is not meaningful

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{RSS}{TSS} = 1 - \frac{SER}{\hat{\sigma}_Y^2}$$

regressor

- Notice that when we add more regressors, i.e. k increases, the value in the parenthesis goes up, while RSS declines. Thus, the adjusted R^2 can decrease if adding more regressors does not help much in explaining the dependent variable.
- $\bar{R}^2$ is always less than R^2; so, $\bar{R}^2 \leq 1$, but it can be negative.

* First, try to link with theory first

⇓

make sense ?

Review

Software cannot help you
write down equation that answer your question

Model construction (Population)

1) Correlation vs. Causation
   - Correlation: predict, indicator, approximate, signal
   - Causation: effect, impact
2) Controls < we need to add it to model to know pure effect of x >
3) functional form < level, ln, ... >
   - ln or level
   - interaction terms (multiplication)
   - Dummies
4) Goodness of fit → to choose model

   * อันไหน make sense ไหม
   - SER → lower the better
   - R^2 — [0,1]: always encourage more x → but x may not meaningful
   - $\bar{R}^2$ — $(-\infty, 1]$ → keep adding non-meaningful x will make it becomes more negative

   the higher, the better fit

! Multicollinearity problem!

↳ x linear combination of other regressors

↳ cannot be parameter

β not unique

drop some x until there's no multicollinearity problem

Estimation (Sample)

- Data (observed)
- Estimator
  - consistent : $cov(u,x) = 0$
  - or
  - Unbiased : $E(u|x) = 0$

  ↳ OLS

Inference (use the estimation to infer back to population)

Population

$y = \beta_0 + \beta_1 x + \ldots + u$

$\beta_1 \overset{?}{=} 0$

Sample

OLS + data

$\Downarrow$

$\hat{\beta}_1 = a$ may not equal to 0

want to answer

Inference :

make conclusion about parameter based on $\hat{\beta}$