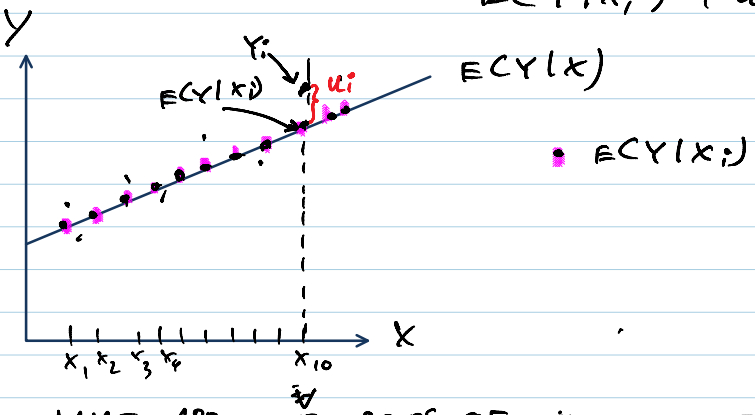
WEEK 5 (7, 9 FEB 2012)

THE SIGNIFICANCE OF THE STOCHASTIC DISTURBANCE TERM

LAST CLASS : $u_i = Y_i - E(Y|X_i)$

OR $Y_i = \underbrace{E(Y|X_i)}_{\text{SYSTEMATIC}} + \underbrace{u_i}_{\text{NON-SYSTEMATIC}}$



Q: WHAT ARE THE ROLES OF THE ERROR TERM?

A: REASONS ARE MANY:

(1) VAGUENESS OF THEORY : EVEN THOUGH INCOME IS A CRUCIAL VARIABLE INFLUENCING CONSUMPTION BEHAVIOR, BUT THERE ARE SOME OTHER VARIABLES THAT MIGHT BE IMPORTANT AND WE IGNORE OR OMMIT THEM. SO WE USE u_i AS "A SURROGATE" (= SUBSTITUTE) FOR ALL THOSE VARIABLES THAT ARE OMMITED FROM THE MODEL BUT THAT COLLECTIVELY AFFECT Y.

(2) UNAVAILABILITY OF DATA : YOU CANNOT INCLUDE SOME VARIABLES THAT ARE IMPORTANT, SIMPLY BECAUSE THE CONSTRAINT OVER DATA AVAILABILITY,

EX: FAMILY WEALTH IS GENERALLY UNAVAILABLE OR COSTLY TO GET A PRECISE DATA SET.

(3) CORE VARIABLES Vs. PERIPHERAL VARIABLES'

(Y) X_1, X_2, X_3, X_4 (SEX), X_5 (RELIGION)

X_1 ↓ INCOME; X_2 ↓ # children; X_3 ↓ family wealth

THEY MIGHT INFLUENCE Y JOINTLY AND/OR YOU BELIEVE THAT INFLUENCES OF ALL VARIABLES ON Y ARE NOT SO BIG.

SO R

NOT SO BIG.

SO $B_{\text{INCLUDING THEM}} < C_{\text{EXCLUDING THEM}}$

SO, WE INCORPORATE ALL VARIABLES INTO u_i

④ "INTRINSIC RANDOMNESS" IN HUMAN BEHAVIOR

⇒ Y HAS "INTRINSIC RANDOMNESS" THAT CANNOT BE EXPLAINED NO MATTER HOW HARD WE TRY. THE u_i TERM REFLECTS THIS INTRINSIC RANDOMNESS.

⑤ POOR PROXY VARIABLES.

TAKE MILTON FRIEDMAN'S MODEL AS AN EXAMPLE:

PERMANENT CONSUMPTION (Y^P) IS A FUNCTION OF PERMANENT INCOME (X^P)

PROXY ? → Y^P → USES CURRENT CONSUMPTION

X^P → USES CURRENT INCOME

⇓

ERRORS OF MEASUREMENT

AS A RESULT, u_i REPRESENTS THIS MEASUREMENT ERRORS.

⑥ WRONG FUNCTIONAL FORM:

$$Y_i = a + bX_i + u_i$$

$$Y_i = a + bX_i + cX_i^2 + u_i$$

EVEN IF WE HAVE THEORETICALLY CORRECT VARIABLES TO BE USED IN EXPLAINING Y, BUT QUITE OFTEN WE DON'T EXACTLY KNOW THE FUNCTIONAL RELATIONSHIP BET. Y AND X(S).

⑦ PRINCIPLE OF PARSIMONY

OCCAM'S RAZOR: KEEP THING AS SIMPLE AS POSSIBLE.

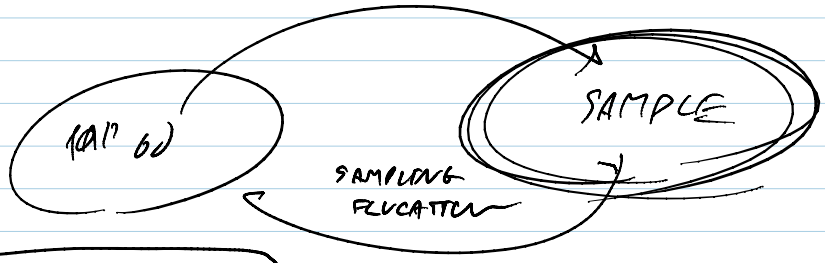
IF BEHAVIOR OF Y CAN BE "SUBSTANTIALLY"

EXPLAINED BY TWO OR THREE VARIABLES' + OUR THEORY IS NOT STRONG ENOUGH TO SUGGEST WHAT ELSE SHOULD BE PUT IN, WHY PUT MORE ?

ANOTHER INTERPRETATION: IF YOU HAVE TWO COMPETING THEORIES' OR MODELS THAT MAKE EXACTLY THE SAME PREDICTION, THE SIMPLER ONE IS THE BEST.

(SIMPLICITY IS A VIRTUE.)

NOW, THE SAMPLE REGRESSION FUNCTION (SRF)



UNKNOWN PARAMETERS

FROM PRF: $Y_i = a + bX_i + u_i$

$E(Y_i|X_i) = a + bX_i$

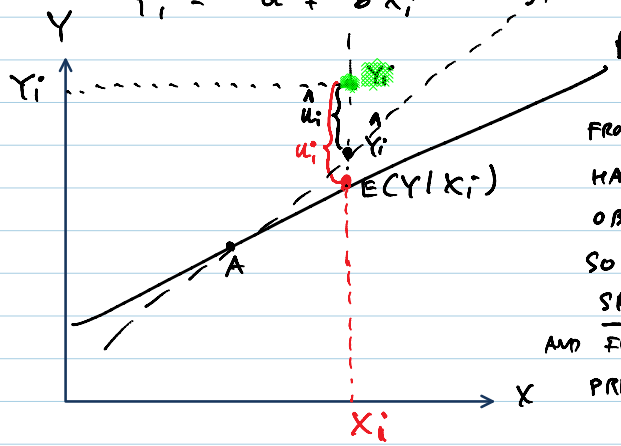
SRF: $Y_i = \hat{a} + \hat{b}X_i + \hat{u}_i$

$\hat{Y}_i = \hat{a} + \hat{b}X_i$

(TRUE) UNKNOWN PARAMETERS: a, b,

ESTIMATORS: $\hat{a}, \hat{b}$,

$\hat{Y}_i = \hat{a} + \hat{b}X_i$



SRF:

PRF: $E(Y|X_i) = a + bX_i$

FROM $X = X_i$, YOU HAVE ONE (SAMPLE) OBSERVATION, $Y = Y_i$,

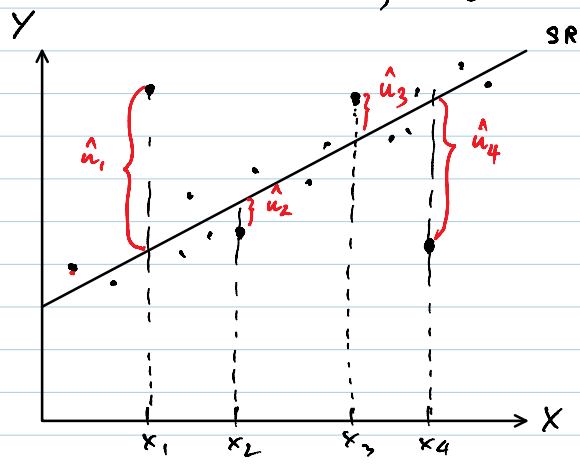
SO IN TERM OF SRF: $Y_i = \hat{Y}_i + \hat{u}_i$

AND IN TERM OF PRF: $Y_i = E(Y|X_i) + u_i$

(9 FEB 2012)

PRINCIPLES OF ORDINARY LEAST SQUARES'

READ: GUJARATI, CH. 3



SRF: $\hat{Y}_i = \hat{a} + \hat{b}X_i$

TWO-VARIABLE REGRESSION MODEL

$Y_i = a + bX_i + u_i$ (PRF)

(CREWS)

- a and b are parameters to be estimated
- Y = dependent variable
  X = independent variable

TASK: WE WANT TO ESTIMATE PRF ON THE BASIS OF SRF AS ACCURATELY AS POSSIBLE.

HOW TO ACCOMPLISH THIS?

- WE WANT TO ESTIMATE a AND b FROM SRF:

$$Y_i = \hat{a} + \hat{b}X_i + \hat{u}_i$$

$$\hat{u}_i = Y_i - \hat{Y}_i$$

$$= Y_i - \hat{a} + \hat{b}X_i$$

- WE WANT THE LINE TO BE AS CLOSE TO Y_i AS POSSIBLE.

- WE WANT TO FIND ESTIMATORS $\hat{a}$ AND $\hat{b}$ SUCH THAT THE SUM OF SQUARED ERRORS IS THE LEAST

LET $\hat{u}_i$ BE ESTIMATED ERRORS.

$$\therefore \quad \hat{u}_1 = Y_1 - \hat{Y}_1$$
$$\hat{u}_2 = Y_2 - \hat{Y}_2$$
$$\vdots$$
$$\hat{u}_n = Y_n - \hat{Y}_n$$

some are positives
some are negatives

$$\sum_{i=1}^{n} (\hat{u}_i) \quad \text{OR} \quad \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = 0$$

SUPPOSE $10, -2, +2, -10 \rightarrow \sum \hat{u}_i = 0$

$\hat{u}_1$ $\hat{u}_2$ $\hat{u}_3$ $\hat{u}_4$

$$\sum (\hat{u}_i)^2 \quad \text{OR} \quad \sum (Y_i - \hat{Y}_i)^2$$

$\Sigma(\hat{u}_i)$ OR $\Sigma(Y_i - \hat{Y}_i)$

- FIND $\hat{a}$, $\hat{b}$ (ESTIMATORS) SUCH THAT THE SUM OF SQUARED OF THE DIFFERENCE BETWEEN THE ACTUAL Y AND THE ESTIMATED Y $(\hat{Y}_i = \hat{a} + \hat{b}X_i)$ AT LEAST.

$\therefore \quad \hat{Y}_i = \hat{a} + \hat{b}X_i \Rightarrow \hat{u}_i = Y_i - \hat{Y}_i$

$= Y_i - (\hat{a} + \hat{b}X_i)$

$\hat{u}_i^2 = (Y_i - \hat{a} - \hat{b}X_i)^2$

$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i)^2$

NOTICE THAT $\sum_{i=1}^{n} \hat{u}_i^2 = f(\hat{a}, \hat{b})$ is a function of or depends on

MINIMIZE $\sum_{i=1}^{n} \hat{u}_i^2$ WITH RESPECT TO $\hat{a}$, $\hat{b}$

LET'S DENOTE $\hat{u}_i = e_i$ FOR CONVENIENCE.

$\frac{\partial \Sigma e_i^2}{\partial \hat{a}} = \frac{\partial \Sigma (Y_i - \hat{a} - \hat{b}X_i)^2}{\partial \hat{a}} = 2\Sigma(Y_i - \hat{a} - \hat{b}X_i)(-1) = 0$

$= -2\Sigma(Y_i - \hat{a} - \hat{b}X_i) = 0$

$\Downarrow \qquad \hat{Y}_i$

IT IMPLIES THAT $\Sigma(Y_i - \hat{a} - \hat{b}X_i) = 0$.

OR

$\Sigma(Y - \hat{Y}_i) = 0$

OR

$\Sigma e_i = 0$ ALWAYS.

$\frac{\partial \Sigma e_i^2}{\partial \hat{b}} = -2\Sigma(Y_i - \hat{a} - \hat{b}X_i)X_i = 0$

$\Downarrow$ ACTUALLY ESTIMATED Y

∂ b̂

⇓ ACTUAL ESTIMATED

IT IMPLIES THAT $\sum (Y_i - \hat{a} - \hat{b} X_i) X_i = 0$

$\sum (Y_i - \hat{Y}_i) X_i = 0$

$\sum e_i X_i = 0$

THE EXPLANATORY VARIABLE (X) AND ERRORS (e_i) MUST BE INDEPENDENT.

FROM $\sum (Y_i - \hat{a} - \hat{b} X_i) = 0 \Leftrightarrow \sum Y_i - \sum \hat{a} - \hat{b} \sum X_i = 0$

$\sum Y_i = \sum \hat{a} + \hat{b} \sum X_i$

$\sum Y_i = n\hat{a} + \hat{b} \sum X_i$

$n\hat{a} = \sum Y_i - \hat{b} \sum X_i$

$\hat{a} = \frac{\sum Y_i - \hat{b} \sum X_i}{n}$

$= \frac{\sum Y_i}{n} - \frac{\hat{b} \sum X_i}{n}$

$= \bar{Y} - \hat{b} \bar{X}$

$\therefore \quad \boxed{\hat{a} = \bar{Y} - \hat{b} \bar{X}}$

FROM $\sum (Y_i - \hat{a} - \hat{b} X_i) X_i = 0$

$\sum X_i Y_i - \sum \hat{a} X_i - \sum \hat{b} X_i^2 = 0$

$\sum X_i Y_i = \hat{a} \sum X_i + \hat{b} \sum X_i^2$

$\hat{b} \sum X_i^2 = \sum X_i Y_i - \hat{a} \sum X_i$

$\hat{b} \sum X_i^2 = \sum X_i Y_i - (\bar{Y} - \hat{b} \bar{X}) \sum X_i$

$= \sum X_i Y_i - \left(\frac{\sum Y_i}{n} - \hat{b} \frac{\sum X_i}{n} \right) \sum X_i$

$= \sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n} + \hat{b} \frac{(\sum X_i)^2}{n}$

$$\hat{b}\sum x_i^2 - \frac{\hat{b}(\sum x_i)^2}{n} = \sum x_i y_i - \frac{\sum x_i y_i}{n}$$

$$\frac{\hat{b}n\sum x_i^2 - \hat{b}(\sum x_i)^2}{n} = \frac{n\sum x_i y_i - \sum x_i y_i}{n}$$

$$\hat{b}n\sum x_i^2 - \hat{b}(\sum x_i)^2 = n\sum x_i y_i - \sum x_i y_i$$

$$\hat{b} = \frac{n\sum x_i y_i - \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

HOME WORK

SHOW THAT $\hat{b} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum x_i^2}$ (LOWER CASE LETTER — LOWER CASE LETTER)

WHERE

$x_i = X_i - \bar{X}$ (DEVIATION OF X_i FROM ITS MEAN)

AND $y_i = Y_i - \bar{Y}$ (DEVIATION OF Y_i FROM ITS MEAN)