

# **Econometrics**

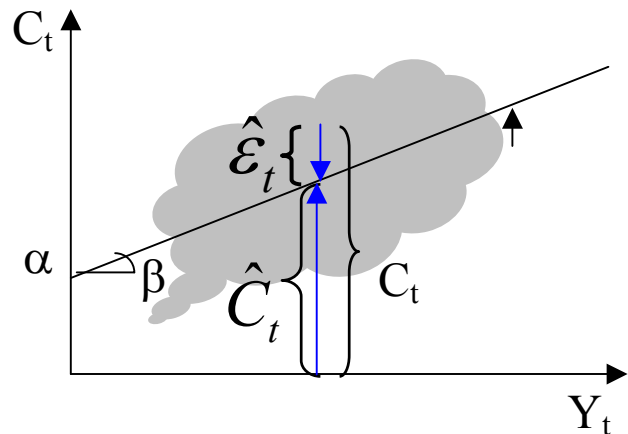
## **Lecture 7**

# 1. Economic vs Econometrics Theory

- It is given that the economic theory assumes a linear model between the dependent and independent explanatory variable, with some random (unexplained) deviations/errors/residuals:

$$C_t = \alpha + \beta Y_t + \varepsilon_t$$

$$\Rightarrow \varepsilon_t = (C_t - \alpha - \beta Y_t)$$



- We dislike errors.
- Our dissatisfaction from the errors increases very rapidly. If the error to either side doubles, our dissatisfaction increases more than double.

## 2. Ordinary Least Square (OLS) Estimators

- Therefore, we would like to impose some restrictions on our econometric model regarding the errors:

(I)  $\sum_t \varepsilon_t = 0$  - There is no systematic (aggregate) error.

(II)  $\text{Min } \sum_t \varepsilon_t^2$  - Penalize for larger errors to either side.

- It is easy to prove that: given our theoretical linear model, imposing the above restrictions provides us with the Best Linear Unbiased Estimators. In other words, the **OLS** estimators are **BLUE** (Gauss-Markov theorem).
- What can we infer from these restrictions? How can we use them in order to derive the estimators- the equations that estimate our coefficients:  $\alpha$  &  $\beta$ ? \*

- Before we proceed, notice:
  - (1) *Our sample includes  $n$  observations, and our regression runs  $m$  independent variables. When  $m > 1$  the regression is called multi-variable regression. We won't provide you with the estimators for multi-variable regression, but the intuition is the same.*
  - (2) *The hat (i.e.,  $\hat{c}$ ) denotes the estimator or the estimate. The former is the formula that we use to estimate the coefficient from the sample, and the latter is the value of the estimator after using the sample.*
  - (3) *The bar (i.e.,  $\bar{c}, \bar{y}$ ) denotes the mean of the variable*
  - (4) *When we use lower-case letters, then we refer to the deviation from the mean of that letter ( $c_i \equiv C_i - \bar{C}$ ).*
  - (5) *We want to minimize the vertical deviations from the regression line. That is different from minimizing the horizontal deviations from the regression line.*
  - (6) *The regression provides us with estimates for the correlation- its sign, magnitude and significance. However, further economics and econometric theory is needed for identifying the causality.*

$$(I) \sum_t \varepsilon_t = 0$$

$$(I.a) \quad \sum_t^n (C_t - \alpha - \beta Y_t) = 0$$

$$\sum_t^n C_t - \sum_t^n \alpha - \beta \sum_t^n Y_t = 0$$

$$\frac{\sum_t^N C_t}{N} - \frac{\sum_t^N \alpha}{N} - \beta \frac{\sum_t^N Y_t}{N} = 0$$

$$\bar{C} - \alpha - \beta \bar{Y} = 0$$

$$\boxed{\alpha = \bar{C} - \beta \bar{Y}} \quad - \text{The constant coefficients}$$

$$(I.b) \quad \frac{\sum_t^N \varepsilon_t}{N} = 0$$


$$\boxed{\bar{\varepsilon} = 0}$$

## (II) Min $\sum_t \varepsilon_t^2$

(II.a)  $\varepsilon_t = C_t - \alpha - \beta Y_t$  - By the model specification (see page 1)

$0 = \bar{C} - \alpha - \beta \bar{Y}$  - Our 1<sup>st</sup> conclusion from 1<sup>st</sup> restriction (see page 2)

$\Rightarrow \varepsilon_t = (C_t - \bar{C}) - \beta (Y_t - \bar{Y}) \equiv c_t - \beta y_t$

**$\varepsilon_t = c_t - \beta y_t$**  

(II.b)  $\text{Min } \sum_t \varepsilon_t^2 = \text{Min } \sum_t (c_t - \beta y_t)^2$   
 $= \text{Min } \sum_t (c_t^2 - 2\beta c_t y_t + \beta^2 y_t^2)$

- $\beta$  minimizes the Sum of Squared Errors (SSE) if the first derivative equals to zero (FOC):

$\sum_t (0 - 2 c_t y_t + 2 \beta y_t^2) = 0$

$\sum_t 0 - 2 \sum_t c_t y_t + 2 \beta \sum_t y_t^2 = 0$

$\beta = \sum_t c_t y_t / \sum_t y_t^2 \Rightarrow \mathbf{\beta = SS_{cy} / SS_{yy}}$

Sum of Squares

Covariance( $C_t, Y_t$ )

Variance( $Y_t$ ) =  $\text{STD}(Y_t)^2$

$\beta = [\sum_t c_t y_t / n] / [\sum_t y_t^2 / n] \Rightarrow \mathbf{\beta = \sigma_{cy} / \sigma_y^2}$

Correlation( $C_t, Y_t$ )  $\leq 1$

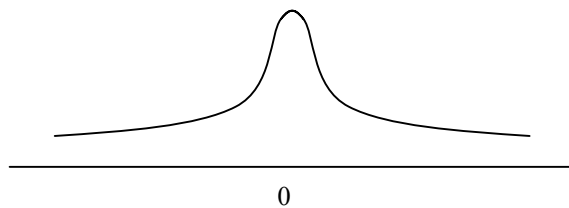
$\beta = [\sigma_{cy} / \sigma_y \sigma_c] [\sigma_c / \sigma_y] \Rightarrow \mathbf{\beta = \rho_{cy} * [\sigma_c / \sigma_y]}$

*When  $\rho_{cy}, \beta < 0$  (note, both have to have the sign), then we have a negative correlation between  $c$  and  $y$ , and visa versa.*

### 3. Confidence

- Since we try to fit a linear model to a **sample** from the **population**, our estimates based on the sample might be different from the true values (based on the population).
- By how much can our estimate be different from the actual?
- We can be  $(1-\lambda)\%$  confident that the true value is different from our estimate by no more than (*the confidence interval*):

$$\beta = \hat{\beta} \pm \hat{S}_{\hat{\beta}} t_{(n-m-1, 1-\lambda/2)} \quad , \text{ where } \hat{S}_{\hat{\beta}} = \sqrt{\frac{\sum_t^n \hat{\varepsilon}_t^2}{\sum_t^n \hat{y}_t^2}}$$



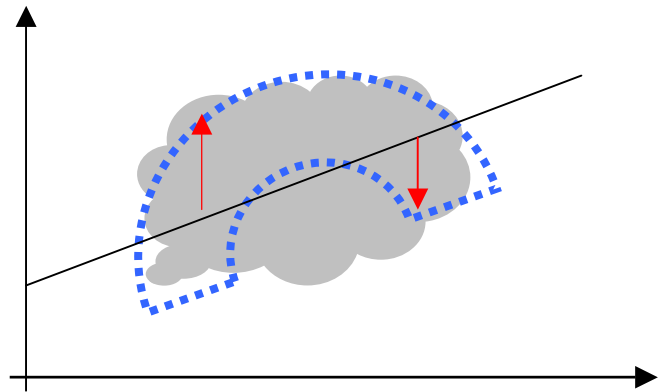
- Therefore, if *t-statistic*  $\equiv \left| \hat{\beta} / \hat{S}_{\hat{\beta}} \right| < t_{(n-m-1, 1-\lambda/2)}$ , then we don't have a **statistically significant** linear relation between  $C_t$  and  $Y_t$ . [ $m-1 \equiv df \equiv \text{degree of freedom}$ ].
- For (the common) 95% confidence interval, the *t-statistic* should be greater than 2.

## 4. Fitness:

- Though we could find the best curve that fits the data, but should it really be linear?
- Evaluating the fitness of the linear regression (how much accurate/powerful the fitted line in explaining the data) is provided by the following criterion:

$$R^2 = \frac{\sum_t \hat{c}_t}{\sum_t c_t} = 1 - \frac{\sum_t \hat{\varepsilon}_t}{\sum_t c_t}$$

$$\Rightarrow 0 \leq R^2 \leq 1$$



- Is there a different between “correlation” and “causality”?



## 5. Summary:

- The theoretical economic model assumes linear relation in levels:

$$C_t = \alpha + \beta Y_t + \varepsilon_t \quad \Rightarrow \quad \varepsilon_t = (C_t - \alpha - \beta Y_t)$$

- The OLS econometric model imposes the following:

$$(I) \sum_t \varepsilon_t = 0 \quad (II) \text{Min } \sum_t \varepsilon_t^2$$

- Gauss-Markov theorem: **OLS** estimators are **BLUE**.

- Estimators:

$$\hat{\beta} = \sum_t c_t y_t / \sum_t y_t^2 \Rightarrow \hat{\beta} = \hat{S}_{cy} / \hat{S}_{yy}$$

Sum of Squares

$$\hat{\beta} = [\sum_t c_t y_t / n] / [\sum_t y_t^2 / n] \Rightarrow \hat{\beta} = \sigma_{CY} / \sigma_y^2$$

Covariance( $C_t, Y_t$ )

Variance( $Y_t$ ) =  $\text{STD}(Y_t)^2$

$$\hat{\beta} = (\hat{\sigma}_{CY} / \hat{\sigma}_y \hat{\sigma}_c) \times (\hat{\sigma}_c / \hat{\sigma}_y) \Rightarrow \hat{\beta} = \hat{\rho}_{CY} \times (\hat{\sigma}_c / \hat{\sigma}_y)$$

Correlation( $C_t, Y_t$ )  $\leq 1$

$$\hat{\alpha} = \bar{C} - \hat{\beta} \bar{Y}$$

- $(1-\lambda)\%$  confidence interval:  $\beta = \hat{\beta} \pm \hat{S}_{\hat{\beta}} t_{(n-m-1, 1-\lambda/2)}$ , where

$$\hat{S}_{\hat{\beta}} = \sqrt{\frac{\sum_t \hat{\varepsilon}_t^2}{\sum_t \hat{y}_t^2}}$$

If  $t$ -statistic  $\equiv \left| \hat{\beta} / \hat{S}_{\hat{\beta}} \right| < t_{(n-m-1, 1-\lambda/2)}$ , then there is

NO statistically significant linear relation between  $C_t$  and  $Y_t$ .

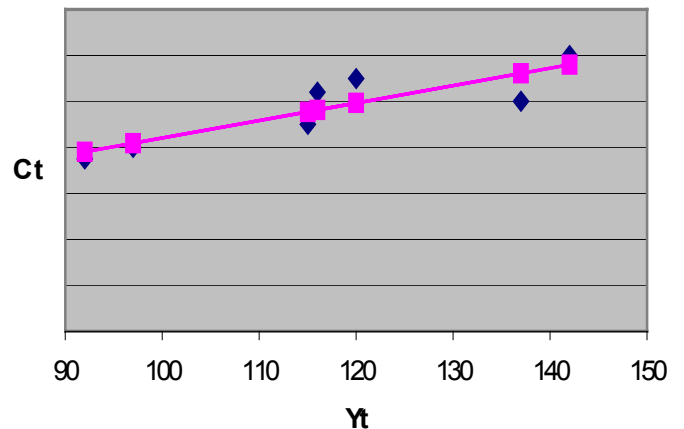
- Linear Fitness:  $R^2 = \frac{\sum_t \hat{c}_t}{\sum_t c_t} = 1 - \frac{\sum_t \hat{\varepsilon}_t^2}{\sum_t c_t^2} \Rightarrow 0 \leq R^2 \leq 1$

## 6. A numerical example:

Observation #	$C_t$	$Y_t$	$C_t$	$Y_t$	$C_t^2$	$Y_t^2$	$C_t * Y_t$	$\hat{C}_t$	$\hat{\varepsilon}_t$	$\hat{\varepsilon}_t^2$
1	100	137	3	20	9	400	60	112	-12	149
2	90	115	(7)	(2)	49	4	14	95	-5	30
3	75	92	(22)	(25)	484	625	550	78	-3	9
4	110	120	13	3	169	9	39	99	11	115
5	104	116	7	(1)	49	1	(7)	96	8	60
6	120	142	23	25	529	625	575	116	4	16
7	<u>80</u>	<u>97</u>	<u>(17)</u>	<u>(20)</u>	<u>289</u>	<u>400</u>	<u>340</u>	<u>82</u>	<u>-2</u>	<u>3</u>
<b>Sum</b>	679	819			1578	2064	1571	679	0	382
<b>Average</b>	97	117			225	295	224	97	0	55

Therefore, the estimates are (so add ^ for all):

$\sigma_c^2$	=	225		$S_{cc}$	=	1578
$\sigma_y^2$	=	295		$S_{yy}$	=	2064
$\sigma_{yc}$	=	224		$S_{yc}$	=	1571
$\rho_{cy}$	=	0.93		$S_b$	=	0.43
$\beta$	=	0.76	<div style="font-size: 2em; margin-left: 10px;">}</div> <div style="margin-left: 10px;"> <math>t_{(n-m-1, 1-\lambda/2)} = 2.571</math> </div>	$t$ -statistic = 1.77		
$\alpha$	=	7.95		95% Confidence Interval? Is it significant?		
$R^2$	=	0.76				



## 7. Caveats

- **(I) Spurious Regression:** Due to the trend nature of many macroeconomic time series data, one should be warrant that a spurious strong linear correlation might be found from running a linear regression, whether or not there is really any regression at work. [*Can the Somalian population explain the increase in the US GDP since both have been increasing?*]. Therefore, note that the independent variable should help us to estimate the deviation of the dependent variable from *its drift* (which is just its mean if the variable is stationary).
- **(II) Missing variables:** notice that if the regression does not include a variable that has significant effect on the dependant variable, and this missing variable is correlated with the included independent variable, then the effect of the missing variable will show up via the included independent variable, and therefore, it's coefficient will be biased. *Example & Proof!*
- **(III) Miss specification:** it is important not only to pick up the right independent variables, but also the right formation of relationship. In other words, econometrist has to choose the right function. For example, he has to decide whether it is linear in levels or in logs!
- **Correlation  $\nleftrightarrow$  Causality**, and flipping the variables is not just a linear transformation of the coefficients.

## 8. $t$ -distribution (from Green)