ECONOMETRICS-I Lecture Notes | Chapters 2 & 3

TWO VARIABLE REGRESSION MODEL

We have two Random Variables:

X: Explanatory (Independent) variable and

Y: Dependent Variable.

We want to estimate Y using X.

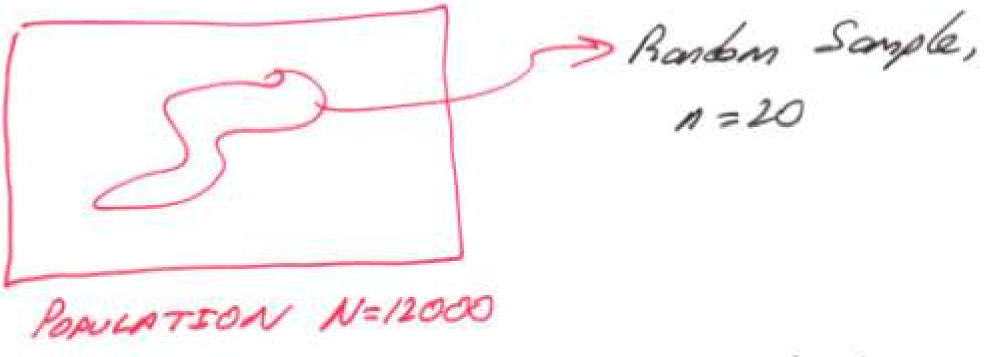
Basic Concepts:

Example Let X_i: Weekly studying hours of a student

Y_i: Cummulative GPA

Let's say, our study is "How efficient ~~does~~ the students at Bilkent University work?" So, our population is all the students at Bilkent, for example $N=12000$ students.

We take a random sample of size $n=20$



POPULATION $N=12000$

X_i: Weekly studying hours of a student (WSH)

Y_i: Cummulative GPA (Cum GPA)

Alp Giray Özen | 0533 549 91 08 | alp@lecturemania.com | www.lecturemania.com

Population Regression MODEL:

$$y_i = \beta_1 + \beta_2 X_i + u_i$$

Sample Regression MODEL

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 \cdot X_i + \hat{u}_i$$

	Population Parameters (Unknown Constants)	Sample Statistics (Known Variables)
Slope:	β_2	$\hat{\beta}_2$
Intercept:	β_1	$\hat{\beta}_1$
Residual term:	u_i	$\hat{u}_i$

PRF: Population Regression FUNCTION:

$$E(y_i / X_i) = \beta_1 + \beta_2 \cdot X_i$$

SRF: Sample Regression FUNCTION

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 \cdot X_i$$

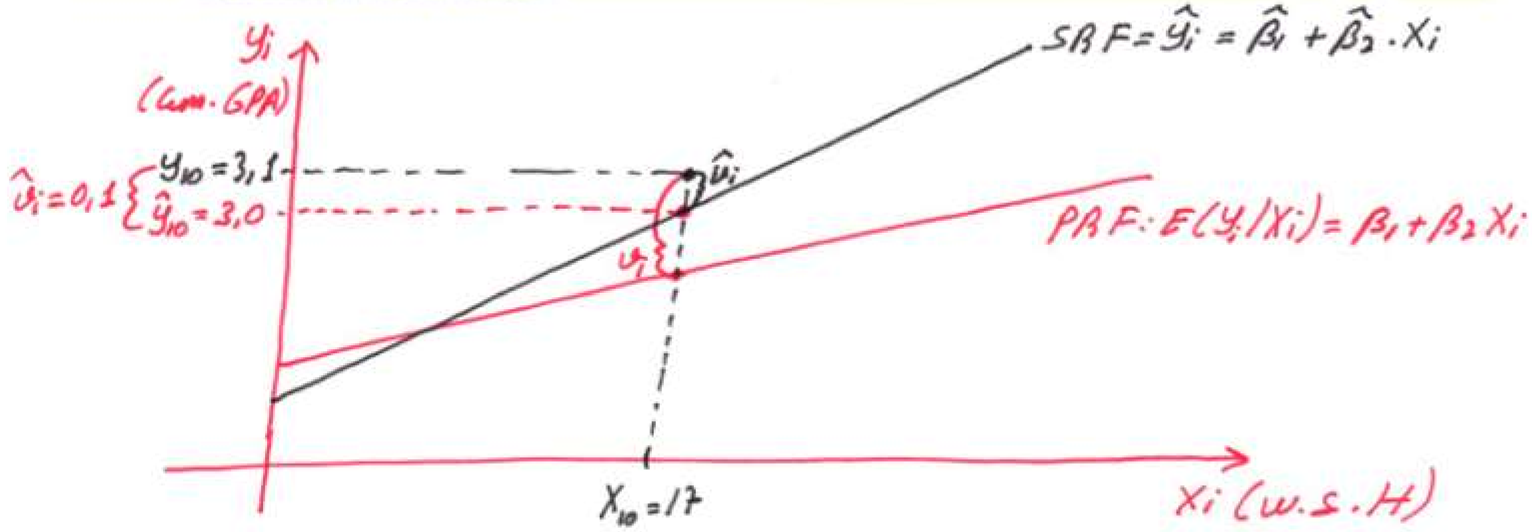
let the sample data is;

i	1	2	3	----	10	----	20
X_i	7	9	9	----	17	------	25
y_i	1,8	2,3	2,1	----	3,1	---	3,7

let $\boxed{\bar{y} = 2,5}$

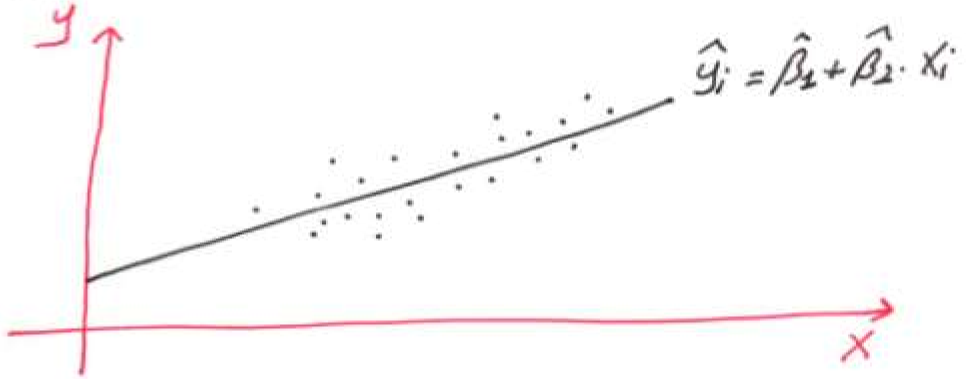
Also let; $\hat{y}_i = 1,3 + 0,1 . X$ (i.e. $\hat{\beta}_1 = 1,3$; $\hat{\beta}_2 = 0,1$)

$$\hat{y}_{10} = 1,3 + 0,1 . 17 = 3,0$$



Ordinary least Square (OLS) estimators;

Let, we have plotted the data on X-Y plane. Given the data, what is the best regression function we can draw? We want our line to be the closest one to the sample points. OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ minimizes the "Sum of Squared Residuals":



$$y_i = \hat{y}_i + \hat{u}_i$$

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 \cdot X_i$$

$$\underbrace{RSS}_{\text{Residual Sum of Squares}} = RSS(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 \cdot X_i)^2$$

We will minimize $RSS(\hat{\beta}_1, \hat{\beta}_2)$ with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$. Remember from Calculus, we take the partial derivatives and equate them to 0.

Alp Giray Özen | 0533 549 91 08 | alp@lecturemania.com | www.lecturemania.com

$$RSS(\hat{\beta}_1, \hat{\beta}_2) = \sum (y_i - \hat{\beta}_1 - \hat{\beta}_2 \cdot x_i)^2$$

① $$\frac{\partial RSS(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_1} = -2 \cdot \sum (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

② $$\frac{\partial RSS(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_2} = -2 \sum x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

* Remember the following properties of $\sum$

(i) $\sum_{i=1}^{n} a \cdot x_i = a \cdot \sum_{i=1}^{n} x_i$: Constant term goes out of summation

(ii) $\sum_{i=1}^{n} a = n \cdot a$: If there's no index, we sum n times a : $n \cdot a$

(i.e. $\sum_{i=1}^{3} 5 = 5 + 5 + 5 = 3 \cdot 5$)

Then, the equations become,

Normal Equations

① $\sum y_i - n \cdot \hat{\beta}_1 - \hat{\beta}_2 \cdot \sum x_i = 0$

② $\sum y_i \cdot x_i - \hat{\beta}_1 \cdot \sum x_i - \hat{\beta}_2 \cdot \sum x_i^2 = 0$

① $\sum y_i = n \cdot \hat{\beta}_1 + \hat{\beta}_2 \cdot \sum x_i$

② $\sum y_i \cdot x_i = \hat{\beta}_1 \cdot \sum x_i + \hat{\beta}_2 \cdot \sum x_i^2$

① $\Rightarrow \hat{\beta}_1 = \dfrac{\sum y_i - \hat{\beta}_2 \cdot \sum x_i}{n} \Rightarrow \boxed{\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \cdot \bar{x}}$

② $\Rightarrow \sum y_i x_i = \dfrac{\sum y_i - \hat{\beta}_2 \cdot \sum x_i}{n} \cdot \sum x_i + \hat{\beta}_2 \cdot \sum x_i^2$

$n \cdot \sum y_i x_i = \sum y_i \sum x_i - \hat{\beta}_2 \cdot (\sum x_i)^2 + n \cdot \hat{\beta}_2 \cdot \sum x_i^2$

$\hat{\beta}_2 \left((\sum x_i)^2 - n \cdot \sum x_i^2 \right) = \sum y_i \sum x_i - n \cdot \sum y_i x_i$

$$\boxed{\hat{\beta}_2 = \frac{n \cdot \sum y_i x_i - \sum y_i \sum x_i}{n \cdot \sum x_i^2 - (\sum x_i)^2}}$$

Alp Giray Özen | 0533 549 91 08 | alp@lecturemania.com | www.lecturemania.com

Problem Statement

Last year, five randomly selected students took a math aptitude test before they began their statistics course. The Statistics Department has three questions.

a. What linear regression equation best predicts statistics performance, based on math aptitude scores?

b. If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?

c. How well does the regression equation fit the data?

Answer a)

	Student	x_i	y_i	xi*xi	yi*yi	xi*yi
	1	95	85	9025	7225	8075
	2	85	95	7225	9025	8075
	3	80	70	6400	4900	5600
	4	70	65	4900	4225	4550
	5	60	70	3600	4900	4200
Sum		390	385	31150	30275	30500
Mean		78	77			

$\bar{x} = 78$ $\quad \bar{y} = 77$

$n = 5;\ \sum x_i = 390;\ \sum y_i = 385;\ \sum x_i^2 = 31150;\ \sum y_i^2 = 30275;\ \sum x_i y_i = 30500$

$$\hat{\beta}_2 = \frac{5 \cdot 30500 - 385 \cdot 390}{5 \cdot 31150 - 390^2} = 0{,}644;\quad \hat{\beta}_1 = 77 - 0{,}644 \cdot 78 = 26{,}78$$

$$\hat{y} = 26{,}78 + 0{,}644 \cdot X$$

b) $(\hat{y} \mid X = 80) = 26{,}78 + 0{,}644 \cdot 80 = 78{,}3$

c) R^2: Coefficient of Determination: We'll see it later.

Note that, $y_3 = 70$ and $\hat{y}_3 = 78{,}3$ (since $x_3 = 80$)

Then, for example, $\hat{u}_3 = 78{,}3 - 70 = 8{,}3$

we have overestimated this student's statistics performance.

* The following identities for $\hat{\beta}_2$ are important because we'll use them in some proofs and derivations.

$$\hat{\beta}_2 = \frac{n.\sum y_i X_i - \sum y_i \sum X_i}{n.\sum X_i^2 - (\sum X_i)^2}$$

dividing the numerator and denominator by n, we have

$$\hat{\beta}_2 = \frac{\sum y_i X_i - \frac{\sum y_i \sum X_i}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$

Numerator: $\sum (X_i - \bar{X}).(y_i - \bar{y}) = \sum X_i y_i - \sum X_i \bar{y} - \sum y_i \bar{X} + n.\bar{X}\bar{y}$

$= \sum X_i y_i - \bar{y}.\sum X_i - \bar{X}.\sum y_i + n.\bar{X}\bar{y}$

$= \sum X_i y_i - \frac{\sum y_i \sum X_i}{n} - \cancel{\frac{\sum X_i \sum y_i}{n}} + \cancel{n.\frac{\sum X_i \sum y_i}{n^2}} = \sum X_i y_i - \frac{\sum X_i \sum y_i}{n}$

So; $\sum y_i X_i - \frac{\sum y_i \sum X_i}{n} = \sum (X_i - \bar{X}).(y_i - \bar{y})$

Denominator; $\sum X_i^2 - \frac{(\sum X_i)^2}{n} = \sum (X_i - \bar{X})^2$ → simply put X_i instead of y_i

We have the notations (which are called "Deviation from the Mean")

$x_i = X_i - \bar{X}$ and $y_i = y_i - \bar{y}$

so; $\sum (X_i - \bar{X})(y_i - \bar{y}) = \sum x_i y_i$ and $\sum (X_i - \bar{X})^2 = \sum x_i^2$

Also Note that,

$\sum x_i y_i = \sum (X_i - \bar{X})(y_i - \bar{y}) = \sum (X_i - \bar{X}).y_i - \sum (X_i - \bar{X}).\bar{y}$

$= \sum \underbrace{(X_i - \bar{X})}_{= x_i}.y_i - \bar{y}.\underbrace{{}^{*}\sum (X_i - \bar{X})}_{=0} = \sum x_i y_i$

$*\ \sum(X_i-\bar{X}) = \sum X_i - n.\bar{X} = \sum X_i - n.\frac{\sum X_i}{n} = \sum X_i - \sum X_i = 0$

Likewise;

$\sum x_i y_i = \sum (X_i-\bar{X})(y_i-\bar{y}) = \sum X_i(y_i-\bar{y}) - \bar{X}.\sum(y_i-\bar{y}) = \sum X_i y_i$

so; $\sum y_i X_i - \frac{\sum y_i \sum X_i}{n} = \sum x_i y_i = \sum X_i y_i = \sum x_i y_i$ and

$$\boxed{\begin{gathered}\hat{\beta}_2 = \frac{n.\sum y_i X_i - \sum y_i \sum X_i}{n.\sum X_i^2 - (\sum X_i)^2} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum X_i y_i}{\sum x_i^2} = \frac{\sum x_i y_i}{\sum x_i^2} \\ \hat{\beta}_1 = \bar{y} - \hat{\beta}_2.\bar{X} \\ \text{SRF: } \hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2.\bar{X}\end{gathered}}$$

Questions about properties of SRF:

Q1) Show that mean value of residuals is 0 $(\bar{\hat{u}}_i = 0)$

Answer: By Normal Equation ①:

$$-2\sum(y_i - \hat{\beta}_1 - \hat{\beta}_2.X_i) = 0$$

$$\sum(y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)) = \sum(y_i - \hat{y}_i) = \sum \hat{u}_i = 0$$

$$\bar{\hat{u}}_i = \frac{\sum \hat{u}_i}{n} = 0$$

Q2) Show that mean value of estimated y_i's is equal to the mean value of actual y_i's $(\bar{\hat{y}} = \bar{y})$

Answer: $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2.X_i = \bar{y} - \hat{\beta}_2.\bar{X} + \hat{\beta}_2.X_i = \bar{y} + \hat{\beta}_2(X_i - \bar{X})$

$$\sum \hat{y}_i = \overbrace{\sum \bar{y}}^{=n.\bar{y}} + \hat{\beta}_2.\overbrace{\sum(X_i-\bar{X})}^{=0} = n.\bar{y}$$

$$\bar{y} = \frac{\sum \hat{y}_i}{n} = \bar{\hat{y}}_i$$

Alp Giray Özen | 0533 549 91 08 | alp@lecturemania.com | www.lecturemania.com

Q3) Show that SRF passes through mean values of y and X (through $(\bar{X}, \bar{y})$)

Answer: $y_i = \hat{\beta}_1 + \hat{\beta}_2 . X_i + \hat{u}_i$

$$\sum y_i = n.\hat{\beta}_1 + \hat{\beta}_2 . \sum X_i + \underbrace{\sum \hat{u}_i}_{=0}$$

$$\frac{\sum y_i}{n} = \hat{\beta}_1 + \hat{\beta}_2 \cdot \frac{\sum X_i}{n} \Rightarrow \bar{y} = \hat{\beta}_1 + \hat{\beta}_2 . \bar{X}$$

3.10. Suppose you run the following regression:

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{u}_i$$

where, as usual, y_i and x_i are deviations from their respective mean values. What will be the value of $\hat{\beta}_1$? Why? Will $\hat{\beta}_2$ be the same as that obtained from Eq. (3.1.6)? Why?

3.10) Deviation form:

① $y_i = \hat{\beta}_1 + \hat{\beta}_2 . X_i + \hat{u}_i$

② $\bar{y} = \hat{\beta}_1 + \hat{\beta}_2 . \bar{X}$

①-② $\Rightarrow$ $y_i - \bar{y} = \hat{\beta}_2 (X_i - \bar{X}) + \hat{u}_i$

$$\boxed{y_i = \hat{\beta}_2 . x_i + \hat{u}_i} \Rightarrow \boxed{\hat{y}_i = \hat{\beta}_2 . x_i}$$

Sample Regression Model — SRF

So, $\hat{\beta}_1 = 0$ and $\hat{\beta}_2$ have the same formula. Note that Deviation Form passes through origin $(0;0)$ because $\bar{y}_i = 0$ and $\bar{x}_i = 0$ (Since $\sum(y_i - \bar{y}) = \sum(X_i - \bar{X}) = 0$)

3.9. Consider the following formulations of the two-variable PRF:

Model I: $Y_i = \beta_1 + \beta_2 X_i + u_i$

Model II: $Y_i = \alpha_1 + \alpha_2 (X_i - \bar{X}) + u_i$

a. Find the estimators of β_1 and α_1. Are they identical? Are their variances identical?
b. Find the estimators of β_2 and α_2. Are they identical? Are their variances identical?
c. What is the advantage, if any, of model II over model I?

3.9) Model I: $\hat{\beta}_2 = \dfrac{\sum x_i y_i}{\sum x_i^2}$ and $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \cdot \bar{X}$

Model II: $$\hat{\alpha}_2 = \frac{n \cdot \sum y_i (X_i - \bar{X}) - \sum y_i \overbrace{\sum (X_i - \bar{X})}^{=0}}{\sum \left[(X_i - \bar{X}) - \underbrace{(\overline{X_i - \bar{X}})}_{=0} \right]^2}$$

$$= \frac{n \cdot \sum y_i X_i - n \cdot \sum y_i \cdot \bar{X}}{\sum x_i^2} = \frac{n \cdot \sum y_i X_i - \sum y_i \sum X_i}{\sum x_i^2} = \frac{\sum y_i x_i}{\sum x_i^2} = \hat{\beta}_2$$

$$\hat{\alpha}_1 = \bar{y} - \hat{\beta}_2 \cdot (\overline{X_i - \bar{X}}) = \bar{y} \neq \hat{\beta}_1$$

Q4) Show that $\sum \hat{y}_i \hat{u}_i = 0$

Answer: $$\sum \hat{y}_i \hat{u}_i = \sum \hat{\beta}_2 x_i \underbrace{\hat{u}_i}_{= y_i - \hat{\beta}_2 x_i} = \hat{\beta}_2 \sum x_i (y_i - \hat{\beta}_2 x_i)$$

$$= \hat{\beta}_2 \underbrace{\sum x_i y_i} - \hat{\beta}_2^2 \cdot \sum x_i^2 = \hat{\beta}_2 \cdot \hat{\beta}_2 \cdot \sum x_i^2 - \hat{\beta}_2^2 \cdot \sum x_i^2 = 0$$

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \Rightarrow \sum x_i y_i = \hat{\beta}_2 \cdot \sum x_i^2$$

Q5) Show that $\sum \hat{u}_i X_i = 0$

Alp Giray Özen | 0533 549 91 08 | alp@lecturemania.com | www.lecturemania.com

Answer: From Normal Equations ② ;

$$-2.\sum X_i(y_i-\hat\beta_1-\hat\beta_2X_i)=0$$

$$\sum X_i(y_i-\underbrace{(\hat\beta_1+\hat\beta_2X_i)}_{\hat y_i})=0$$

$$\sum X_i\underbrace{(y_i-\hat y_i)}_{=\hat u_i}=\sum X_i\cdot\hat u_i=0$$

The Classical Linear Regression Model (CLRM):
Model Assumptions

* Values taken by the regressor X are considered fixed in repeated samples. More technically, X is assumed to be "nonstochastic"

Example: let; X_i: Years lived abroad
y_i: Grade of foreign language exam.

Data:

X_i	y_i
2	65
2	68
2	62
2	71

$y_i/X_i=2$

X_i	y_i
3	70
3	78
3	69
3	75
3	72
3	77

$y_i/X_i=3$

X_i	y_i
4	84
4	75
4	83
4	86
4	80

$y_i/X_i=4$

$$\boxed{y_i=\beta_1+\beta_2X_i+u_i}$$

β_1, β_2 are model parameters. Since X_i is assumed nonstochastic, y_i is a random variable changing through the random variable: u_i

* Remember the following:

If Y and W are Random Variables;

- $Var(Y) = E[(Y - E(Y))^2] = E(Y^2) - E^2(Y)$
- $Cov(Y, W) = E[(Y - E(Y))(W - E(W))] = E(Y.W) - E(Y).E(W)$

so; $Cov(Y, Y) = Var(Y)$

- If Y and W are INDEPENDENT,

$$Cov(Y, W) = 0 \Rightarrow E(Y.W) = E(Y).E(W)$$

* Assumptions about Residual Term: u_i

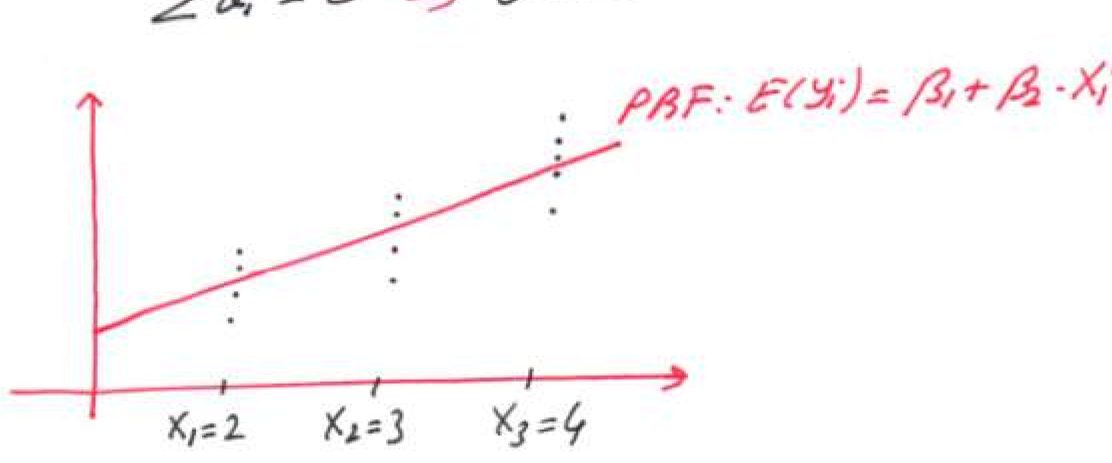
(i) $E(u_i / X_i) = 0$. Given the value of X, the mean or expected value of residual term is 0.

Note that, $\bar{u} = 0$ for each X_i value of our example.

To illustrate; $(\hat{Y} | X_i = 2) = \frac{65 + 68 + 62 + 71}{4} = 66,5$

$u_i = Y_i - \hat{Y}_i \Rightarrow \hat{u}_1 = 1,5$; $\hat{u}_2 = -1,5$; $\hat{u}_3 = 4,5$; $\hat{u}_4 = -4,5$

$$\sum \hat{u}_i = 0 \Rightarrow \bar{\hat{u}} = 0$$



(ii) $Var(u_i/X_i) = E\left[(u_i - E(u_i/X_i))^2\right] = E(u_i^2/X_i) = \sigma^2$

Given the value of X, the variance of u_i is the same for all observations. This is called "homoscedasticity"

To illustrate, we assume that the variances of grades for people whose $X=2$; $X=3$ or $X=4$ are the same.

Homoscedasticity: $Var(u_i/X_i) = \sigma^2$

Heteroscedasticity: $Var(u_i/X_i) = \sigma_i^2$

(iii) Residuals are UNCORRELATED:

$Cov(u_i, u_j \overbrace{/X_i, X_j}^{\text{this is not written for short.}}) = E\{(u_i - E(u_i)).(u_j - E(u_j))\}$

$= E(u_i . u_j) = 0 \quad (i \neq j)$

Given any two values, we assume that the correlation between any two u_i and u_j $(i \neq j)$ is zero

(iv) Residuals and X_i are UNCORRELATED;

$Cov(u_i, X_i) = E\left[(u_i - E(u_i)).(X_i - E(X_i))\right]$

$= E\left[u_i (X_i - E(X_i))\right] = E(u_i X_i) - \underbrace{E(u_i)}_{=0} E(X_i)$

$= E(u_i X_i) = 0$

* Note that, for Y, W random and a, b, c constants:

- $E(aW + bY + c) = a\,E(W) + b\,E(Y) + c$ and
- $Var(aW + bY + c) = Var(aW + bY) = a^2 Var(W) + b^2 Var(Y) + 2ab\,Cov(W, Y)$

Alp Giray Özen | 0533 549 91 08 | alp@lecturemania.com | www.lecturemania.com

3.1. Given the assumptions in column 1 of the table, show that the assumptions in column 2 are equivalent to them.

ASSUMPTIONS OF THE CLASSICAL MODEL

	(1)	(2)
a)	$E(u_i \mid X_i) = 0$	$E(Y_i \mid X_i) = \beta_2 + \beta_2 X$
b)	$\text{cov}(u_i, u_j) = 0 \; i \neq j$	$\text{cov}(Y_i, Y_j) = 0 \; i \neq j$
c)	$\text{var}(u_i \mid X_i) = \sigma^2$	$\text{var}(Y_i \mid X_i) = \sigma^2$

3.1) a) $E(y_i / X_i) = E(\beta_1 + \beta_2 X_i + u_i / X_i)$

$$= \beta_1 + \beta_2 X_i + \underbrace{E(u_i / X_i)}_{=0} = \beta_1 + \beta_2 X_i$$

b) $Cov(y_i, y_j) = Cov(\beta_1 + \beta_2 X_i + u_i, \beta_1 + \beta_2 X_j + u_j)$

$$= E\left[(\beta_1 + \beta_2 X_i + u_i - \underbrace{E(\beta_1 + \beta_2 X_i + u_i)}_{=\beta_1 + \beta_2 X_i})(\beta_1 + \beta_2 X_j + u_j - \underbrace{E(\beta_1 + \beta_2 X_j + u_j)}_{=\beta_1 + \beta_2 X_j}) \right]$$

$$= E(u_i \cdot u_j) = Cov(u_i, u_j) = 0$$

c) $Var(y_i / X_i) = Var(\underbrace{\beta_1 + \beta_2 X_i}_{\text{constant}} + u_i / X_i) = Var(u_i / X_i) = \sigma^2$

Gauss-Markov THEOREM

The least square estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are BLUE: Best, linear Unbiased Estimators.

So, we have for $\hat{\beta}_2$ (this is the important one)

(i) $\hat{\beta}_2$ is linear function of random observations y_i (Note that we assume X_i nonstochastic $\Rightarrow$ NOT Random)

(ii) $\hat{\beta}_2$ is unbiased (Namely, $E(\hat{\beta}_2) = \beta_2$)

(iii) $\hat{\beta}_2$ has minimum variance (is best) among all unbiased linear estimators.

PROOFS : (Especially (i) and (ii) are important)

$$(i)\ \hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i\, y_i}{\sum x_i^2} = \sum \underbrace{\boxed{\frac{x_i}{\sum x_i^2}}}_{=k_i} \cdot y_i = \sum k_i\, y_i$$

$$\boxed{\hat{\beta}_2 = \sum k_i \cdot y_i} \quad \text{where} \quad k_i = \frac{x_i}{\sum x_i^2}$$

↳ So, $\hat{\beta}_2$ is a linear function of y_i

* For the weights k_i, we'll use the following identities:

- $\boxed{\sum k_i = 0}$

Because, $\sum k_i = \sum \frac{x_i}{\sum x_i^2} = \frac{1}{\sum x_i^2} \cdot \underbrace{\sum x_i}_{=0} = 0$

- $\boxed{\sum k_i^2 = \frac{1}{\sum x_i^2}}$

Because, $\sum k_i^2 = \sum \left(\frac{x_i}{\sum x_i}\right)^2 = \sum \frac{x_i^2}{(\sum x_i)^2} = \frac{1}{(\sum x_i)^2} \cdot \sum x_i^2 = \frac{1}{\sum x_i^2}$

- $\boxed{\sum k_i x_i = \sum k_i X_i = 1}$

Because; $\sum k_i x_i = \sum \frac{x_i}{\sum x_i^2} \cdot x_i = \frac{1}{\sum x_i^2} \cdot \sum x_i^2 = 1$

$$\sum k_i X_i = \sum \frac{x_i}{\sum x_i^2} \cdot X_i = \frac{1}{\sum x_i^2} \cdot \underbrace{{}^{*}\sum x_i X_i}_{=\sum x_i^2} = 1$$

$$^{*}\ \sum x_i X_i = \sum (X_i - \bar{X}) \cdot X_i = \sum X_i^2 - \bar{X} \cdot \sum X_i$$

$$= \sum X_i^2 - \frac{\sum X_i}{n} \cdot \sum X_i = \sum X_i^2 - \frac{(\sum X_i)^2}{n} = \sum (X_i - \bar{X})^2 = \sum x_i^2$$

(ii) $\hat{\beta}_2 = \sum k_i Y_i$ and $Y_i = \beta_1 + \beta_2 X_i + u_i$

$$\hat{\beta}_2 = \sum k_i (\beta_1 + \beta_2 X_i + u_i) = \beta_1 \cdot \underbrace{\sum k_i}_{=0} + \beta_2 \cdot \underbrace{\sum k_i X_i}_{=1} + \sum k_i u_i$$

$$\hat{\beta}_2 = \beta_2 + \sum k_i u_i$$

$$E(\hat{\beta}_2) = E[\beta_2 + \sum k_i u_i] = \beta_2 + E(\sum k_i u_i) = \beta_2 + \sum k_i \underbrace{(E(u_i))}_{=0}$$

$$\boxed{E(\hat{\beta}_2) = \beta_2}$$

(iii) $\hat{\beta}_2 = \sum k_i Y_i$ and $Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2}$: we'll show this later.

We want to show $Var(\hat{\beta}_2)$ is minimum among all unbiased linear estimators β_2^*.

let $\beta_2^* = \sum w_i \cdot Y_i$

- $E(\beta_2^*) = E(\sum w_i Y_i) = \sum w_i (E(Y_i)) = \sum w_i (\beta_1 + \beta_2 X_i)$

$$E(\beta_2^*) = \beta_1 \cdot \sum w_i + \beta_2 \cdot \sum w_i X_i$$

since β_2^* is unbiased, we have;

$$\sum w_i = 0 \text{ and } \sum w_i X_i = 1 \ (= \sum w_i x_i)$$

- $Var(\beta_2^*) = Var[\sum w_i Y_i] = \sum w_i^2 Var(Y_i) = \sum w_i^2 Var(u_i)$

$$= \sum w_i^2 \cdot \sigma^2 = \sigma^2 \cdot \sum w_i^2$$

$$\sum w_i^2 = \sum \Big(\underbrace{w_i - \frac{x_i}{\sum x_i^2}}_{a} + \underbrace{\frac{x_i}{\sum x_i^2}}_{b} \Big)^2 = \sum \Big(w_i - \frac{x_i}{\sum x_i^2} \Big)^2 + \underbrace{\sum \Big(\frac{x_i}{\sum x_i^2} \Big)^2}_{= \frac{1}{\sum x_i^2}} + 2 \sum a \cdot b$$

$$\sum ab = \sum\left(w_i - \frac{x_i}{\sum x_i^2}\right)\left(\frac{x_i}{\sum x_i^2}\right) = \frac{1}{\sum x_i^2}\underbrace{\sum w_i x_i}_{=1} - \frac{\cancel{\sum x_i^2}}{\left(\sum x_i^2\right)^{\cancel{2}}} = 0$$

So; $Var(\beta_2^*) = \sigma^2 \cdot \left[\left(w_i - \frac{x_i}{\sum x_i^2}\right)^2 + \frac{1}{\sum x_i^2}\right]$

But $Var(\beta_2^*)$ is minimum when $w_i = \frac{x_i}{\sum x_i^2} = k_i$.

So, $\hat{\beta}_2 = \sum k_i y_i$ is BLUE.

Variance of $\hat{\beta}_2$:

$$Var(\hat{\beta}_2) = E\left[\left(\hat{\beta}_2 - \underbrace{E(\hat{\beta}_2)}_{=\beta_2}\right)^2\right] = E\left[(\hat{\beta}_2 - \beta_2)^2\right]$$

$$\hat{\beta}_2 = \sum k_i y_i = \sum k_i(\beta_1 + \beta_2 X_i + u_i) = \beta_1 \cdot \underbrace{\sum k_i}_{=0} + \beta_2 \cdot \underbrace{\sum k_i X_i}_{=1} + \sum k_i u_i$$

$$\hat{\beta}_2 = \beta_2 + \sum k_i u_i$$

$$\hat{\beta}_2 - \beta_2 = \sum k_i u_i$$

$$Var(\hat{\beta}_2) = E\left[\left(\sum k_i u_i\right)^2\right]$$

$$= E\left[k_1^2 u_1^2 + k_2^2 u_2^2 + \dots + k_n^2 u_n^2 + 2k_1 k_2 u_1 u_2 + \dots + 2k_{n-1} k_n u_{n-1} u_n\right]$$

$$= k_1^2 \underbrace{E(u_1^2)}_{=\sigma^2} + k_2^2 \underbrace{E(u_2^2)}_{=\sigma^2} + \dots + k_n^2 \underbrace{E(u_n^2)}_{=\sigma^2} + 2k_1 k_2 \underbrace{E(u_1 u_2)}_{=0} + \dots + 2k_{n-1} k_n \underbrace{E(u_{n-1} u_n)}_{=0}$$

$$= \sigma^2 (k_1^2 + k_2^2 + \dots + k_n^2) = \sigma^2 \cdot \sum k_i^2 = \sigma^2 \cdot \frac{1}{\sum x_i^2}$$

$$\boxed{Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2}}$$

Covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = E\left[\left(\hat{\beta}_1 - \underbrace{E(\hat{\beta}_1)}_{=\beta_1}\right)\left(\hat{\beta}_2 - \underbrace{E(\hat{\beta}_2)}_{=\beta_2}\right)\right]$$

$$= E\left[(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2)\right]$$

we have; ① $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 . \bar{X}$

② $E(\hat{\beta}_1) = \bar{y} - E(\hat{\beta}_2).\bar{X}$

① - ② $\Rightarrow$ $\hat{\beta}_1 - \beta_1 = -\bar{X}(\hat{\beta}_2 - \beta_2)$

then, $Cov(\hat{\beta}_1, \hat{\beta}_2) = E\left[-\bar{X}(\hat{\beta}_2 - \beta_2)(\hat{\beta}_2 - \beta_2)\right] = -\bar{X}\underbrace{E\left[(\hat{\beta}_2 - \beta_2)^2\right]}_{= Var(\hat{\beta}_2)}$

$$= -\bar{X}. Var(\hat{\beta}_2) = -\bar{X}.\frac{\sigma^2}{\sum x_i^2}$$

$$\boxed{Cov(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\bar{X}.\sigma^2}{\sum x_i^2}}$$

* To summarize, we have the following:

SRF: $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 . X_i$

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \qquad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 . \bar{X}$$

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \qquad Var(\hat{\beta}_1) = \frac{\sum X_i^2}{n.\sum x_i^2}.\sigma^2$$

(Proof is NOT given)

we estimate σ^2 from the data by;

$$\boxed{\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}}$$

Alp Giray Özen | 0533 549 91 08 | alp@lecturemania.com | www.lecturemania.com

The Coefficient of determination: r^2

How well is our model? We measure the "Goodness of fit" of our model by r^2. r^2 answers, "How much (what percentage) of the total variation in Y can be explained by the model (for now, by X)?"

We'll see the logic under this fact. Consider;

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 \cdot X_i + \hat{u}_i$$

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

$$y_i = \hat{y}_i + \hat{u}_i \text{ : Deviation form}$$

$$\sum y_i^2 = \sum (\hat{y}_i + \hat{u}_i)^2 = \sum (\hat{y}_i^2 + \hat{u}_i^2 + 2\hat{y}_i \hat{u}_i)$$

$$= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 + 2 \underbrace{\sum \hat{y}_i \hat{u}_i}_{= \sum \hat{\beta}_2 x_i \hat{u}_i = \hat{\beta}_2 \underbrace{\sum x_i \hat{u}_i}_{=0} = 0}$$

$$\boxed{\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{u}_i^2}$$

$$TSS = ESS + RSS$$

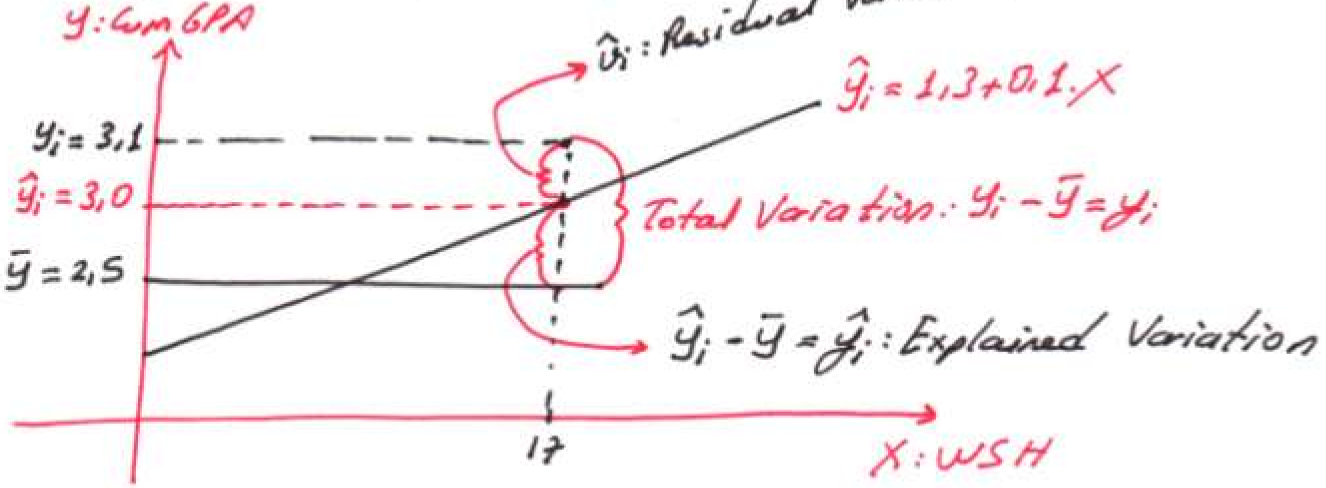
TSS: Total Sum of Squares
ESS: Explained Sum of Squares
RSS: Residual Sum of Squares.

$$TSS = \sum y_i^2$$

$$ESS = \sum \hat{y}_i^2 = \sum (\hat{\beta}_2 \cdot x_i)^2 = \hat{\beta}_2^2 \cdot \sum x_i^2$$

$$RSS = \sum \hat{u}_i^2 = TSS - ESS = \sum y_i^2 - \hat{\beta}_2^2 \cdot \sum x_i^2$$

Example : (WSH, Cum GPA)



The idea is as follows: If we do NOT know weekly studying hour of this specific student, our estimate about her Cum. GPA would be $2,5 = \bar{Y}$. In fact, her CumGPA is 3,1 because she is a hardworking student. With the information that she studies 17 hours a week, our estimate has upgraded to 3,0. However, we still can NOT explain the residual, 0,1.

Our explained proportion is;

$$\boxed{r^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}_2^2 \cdot \sum x_i^2}{\sum y_i^2} = \underbrace{\frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2 \sum y_i^2}}}$$

Because $\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$

Note that; $0 \leq r^2 \leq 1$

Also Note that, r : Sample Correlation Coefficient

$$-1 \leq r \leq 1$$

Example: let's turn back to "Problem Statement": page 5.

c) we have; $\hat{y} = \underbrace{26,78}_{=\hat{\beta}_1} + \underbrace{0,644}_{=\hat{\beta}_2}X$

$$TSS = \sum y_i^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} = 30275 - \frac{385^2}{5} = 630$$

$$ESS = \sum \hat{y}_i^2 = \hat{\beta}_2^2 \cdot \underbrace{\sum x_i^2} = 0,644^2 \cdot 730 = 303$$

$$\sum x_i^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n} = 31150 - \frac{390^2}{5} = 730$$

$$RSS = \sum \hat{u}_i^2 = TSS - ESS = 630 - 303 = 327$$

$$r^2 = \frac{ESS}{TSS} = \frac{303}{630} = 0,481$$

Then, 48,1% of the total variation in statistics performance (Y) can be explained by the variation in math aptitude scores (X)

d) what is the estimated error variance?

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} = \frac{RSS}{n-2} = \frac{327}{3} = 109$$

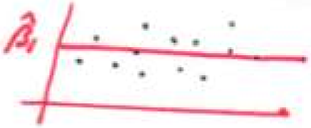
e) Estimate the ~~standard deviation~~ error of $\hat{\beta}_2$.

$$\hat{\sigma}_{\hat{\beta}_2}^2 = \frac{\hat{\sigma}^2}{\sum x_i^2} = \frac{109}{730} = 0,149$$

$$SE(\hat{\beta}_2) = \sqrt{\sigma_{\hat{\beta}_2}^2} = \sqrt{0,149} = 0,386$$

3.17. *Regression without any regressor.* Suppose you are given the model: $Y_i = \beta_1 + u_i$. Use OLS to find the estimator of β_1. What is its variance and the RSS? Does the estimated β_1 make intuitive sense? Now consider the two-variable model $Y_i = \beta_1 + \beta_2 X_i + u_i$. Is it worth adding X_i to the model? If not, why bother with regression analysis?

3.17) $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \cdot \bar{X} = \bar{Y}$



RSS = TSS because ESS = 0

If X explains a significant portion of Y, we add X to the model.

3.19. *The relationship between nominal exchange rate and relative prices.* From the annual observations from 1980 to 1994, the following regression results were obtained, where Y = exchange rate of the German mark to the U.S. dollar (GM/\$) and X = ratio of the U.S. consumer price index to the German consumer price index; that is, X represents the relative prices in the two countries:

$$\hat{Y}_t = 6.682 - 4.318 X_t \qquad r^2 = 0.528$$
$$\text{se} = (1.22)(1.333)$$

a. Interpret this regression. How would you interpret r^2?
b. Does the negative value of X_t make economic sense? What is the underlying economic theory?
c. Suppose we were to redefine X as the ratio of German CPI to the U.S. CPI. Would that change the sign of X? And why?

3.19) a) $r^2 = 0{,}528 \Rightarrow$ 52,8 % of the total variation in exchange rate of German mark to US dollar can be explained by the ratio of US CPI to German CPI.

$\hat{\beta}_2 = -4{,}318$ implies the estimated effect of a unit increase in X to Y. Namely, if Ratio of US CPI to German CPI increases by 1, the exchange Rate of German Mark to US Dollar is expected to decrease by 4,318.

Note that $SE(\hat{\beta}_1) = 1{,}22$ and $SE(\hat{\beta}_2) = 1{,}333$