

REGRESIA LINIARĂ MULTIPLĂ

TEMATICA C6

1. Estimarea indicatorilor de corelație
2. Raportul de determinație ajustat
3. Testarea indicatorilor de corelație
4. Testarea influenței marginale a unei variabile
5. Utilitatea modelului de regresie cu variabile standardizate

Exemple



7. Estimarea indicatorilor de corelatie

Pentru un *model de regresie liniară multiplă*, pot fi determinați următorii coeficienți de corelație:

- *coeficienți de corelație simplă* - între variabila dependentă și fiecare variabilă independentă (*coeficienți bivariați* r_{YX});
- *coeficienți de corelație parțială* ($r_{YX_1.X_2}$; $r_{YX_2.X_1}\dots$)
- *coeficientul de corelație multiplă* (r) și *coeficientul de determinație multiplă* (R^2) .

Coeficienți de corelație bivariată (r_{YXi})

Pentru un model liniar de forma:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

există trei *coeficienți de corelație bivariată*:

$$r_{y1} = \frac{n \sum_i x_{1i} y_i - \sum_i x_{1i} \sum_i y_i}{\sqrt{[n \sum_i x_{1i}^2 - (\sum_i x_{1i})^2][n \sum_i y_i^2 - (\sum_i y_i)^2]}}$$

$$r_{y2} = \frac{n \sum_i x_{2i} y_i - \sum_i x_{2i} \sum_i y_i}{\sqrt{[n \sum_i x_{2i}^2 - (\sum_i x_{2i})^2][n \sum_i y_i^2 - (\sum_i y_i)^2]}}$$

$$r_{12} = \frac{n \sum_i x_{1i} x_{2i} - \sum_i x_{1i} \sum_i x_{2i}}{\sqrt{[n \sum_i x_{1i}^2 - (\sum_i x_{1i})^2][n \sum_i x_{2i}^2 - (\sum_i x_{2i})^2]}}$$

Coeficienți de corelație si de determinatie parțială

... și trei coeficienți de corelație parțială calculați cu ajutorul coeficienților de corelație bivariată:

$$r_{y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}} \quad r_{12.y} = \frac{r_{12} - r_{y1}r_{y2}}{\sqrt{(1-r_{y1}^2)(1-r_{y2}^2)}} \quad r_{y2.1} = \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{(1-r_{y1}^2)(1-r_{12}^2)}}$$

Corelația parțială ($r_{y1.2}$) măsoară dependența parțială dintre două variabile (y și x_1), considerând influența celeilalte variabile din model (x_2) constantă, atât în raport cu variabila dependentă (r_{y2}) cât și în raport cu variabila independentă (r_{12}).

Coeficientul de determinație parțială ($r_{y1.2}^2$) măsoară cât la sută din variația lui Y a rămas neexplicată din modelul dintre Y și X_2 și este explicată prin introducerea în model a variabilei X_1 .

În funcție de **numărul variabilelor a căror influență este izolată**, coeficienții de corelație parțială pot fi **de ordinul întâi** (pentru o variabilă izolată), de **ordinul doi** (pentru două variabile) etc.

Coeficientul de corelație multiplă (r)

Coeficientul de corelație multiplă (r) se calculează numai pentru **modelele multiple liniare** și se exprimă cu ajutorul coeficienților de corelație simplă dintre variabilele perechi.

Astfel, în cazul corelației dintre o *variabilă rezultativă* Y și două *variabile independente* X_1, X_2 , la nivelul unui eșantion, **coeficientul de corelație multiplă**, notat cu r , se calculează după relația:

$$r = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}} \Leftrightarrow$$
$$r = \sqrt{r_{y1}^2 + (1 - r_{y1}^2)r_{y2.1}} \Leftrightarrow r = \sqrt{r_{y2}^2 + (1 - r_{y2}^2)r_{y1.2}}$$

Coeficientul de corelație multiplă (r) este un indicator care masoara intensitatea legaturii dintre variabila dependenta si toate variabilele independente cuprinse in model.

Raportul de determinație și raportul de corelație multiplă

Parametri

$$\eta^2 = \frac{\sum_i (y_{x_i} - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{V_E}{V_T} = 1 - \frac{V_R}{V_T} \Rightarrow \eta = \sqrt{\eta^2}$$

Estimatori

$$\hat{\eta}^2 = \frac{\hat{V}_E}{\hat{V}_T} = 1 - \frac{\hat{V}_R}{\hat{V}_T} = 1 - \frac{\sum_i \varepsilon_i^2}{\sum_i (y_i - \bar{y})^2} \Rightarrow \hat{\eta} = \sqrt{\hat{\eta}^2}$$

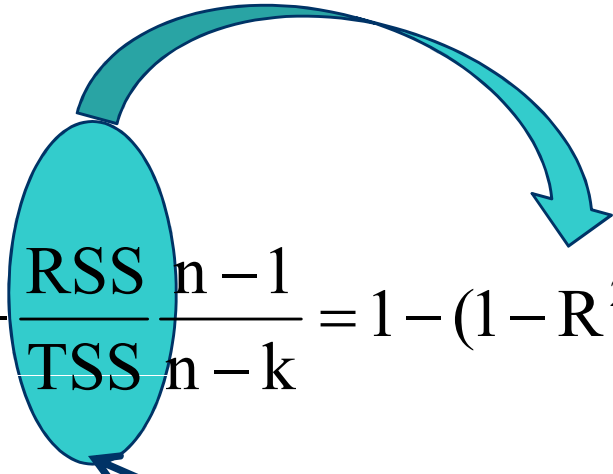
Estimațiile

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2} \Rightarrow R = \sqrt{R^2}$$



Estimatorul ajustat a raportului de determinatie

Adjusted R square

$$\bar{R}^2 = 1 - \frac{\text{RMS}}{\text{TMS}} = 1 - \frac{\frac{\text{RSS}}{n-k}}{\frac{\text{TSS}}{n-1}} = 1 - \frac{\text{RSS}}{\text{TSS}} \frac{n-1}{n-k} = 1 - (1 - R^2) \frac{n-1}{n-k}$$


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

Se observă că $\bar{R}^2 < R^2$ pentru $k > 1$.



8. Testarea indicatorilor de corelație

Raportul de determinație și ***raportul de corelație*** se testează cu ***testul F*** după algoritmul prezentat la modelul liniar simplu, ținând cont de faptul că **$k=p+1$** reprezintă numărul parametrilor din noul model (**$F_{th} = F_{\alpha, k-1, n-k}$**)

Coeficienții de corelație se testează cu ajutorul ***testului t*** după algoritmul prezentat la modelul liniar simplu, ținând cont de faptul că **$k=p+1$** reprezintă numărul parametrilor din noul model (**$t_{th} = t_{\alpha/2, n-k}$**)

Testarea influenței marginale a unei variabile independente, nou **introduse** in model, asupra variabilei dependente
(Metoda intrarilor)

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \text{ (old)} \xrightarrow{+x_{2i}} y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \text{ (new)}$$

1. Formularea ipotezelor

H₀: variabila independentă **nou introdusă în model** nu are o influență semnificativă asupra variației variabilei dependente

H₁: variabila independentă **nou introdusă în model** are o influență semnificativă asupra variației variabilei dependente

2. Fixarea pragului de semnificație $\alpha=0,05$

3. Alegerea statisticii test

$$F_{\text{calc}} = \frac{\hat{V}_{E_{\text{new}}} - \hat{V}_{E_{\text{old}}}}{\hat{V}_{R_{\text{new}}}} \cdot \frac{n - k_{\text{new}}}{k_{\text{new}} - k_{\text{old}}} = \frac{\hat{\eta}_{\text{new}}^2 - \hat{\eta}_{\text{old}}^2}{1 - \hat{\eta}_{\text{new}}^2} \cdot \frac{n - k_{\text{new}}}{k_{\text{new}} - k_{\text{old}}}$$

4. Calcularea statisticii test

$$F_{\text{calc}} = \frac{ESS_{\text{new}} - ESS_{\text{old}}}{RSS_{\text{new}}} \cdot \frac{n - k_{\text{new}}}{k_{\text{new}} - k_{\text{old}}} = \frac{R^2_{\text{new}} - R^2_{\text{old}}}{1 - R^2_{\text{new}}} \cdot \frac{n - k_{\text{new}}}{k_{\text{new}} - k_{\text{old}}}$$

5. Criterii de decizie:

Dacă $F_{\text{calc}} \leq F_{\alpha, k_{\text{new}} - k_{\text{old}}, n - k_{\text{new}}}$ \Rightarrow nu respingem H₀ (AH₀) cu o prob. de 1- α .

Dacă $F_{\text{calc}} > F_{\alpha, k_{\text{new}} - k_{\text{old}}, n - k_{\text{new}}}$ \Rightarrow se respinge H₀ (RH₀) cu un risc asumat α .

Testarea influenței marginale a unei variabile
independente, **exclude** din model, asupra variabilei
dependente
(Metoda ieseirilor)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \text{ (old)} \xrightarrow{-x_{2i}} y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \text{ (new)}$$

1. Formularea ipotezelor

H₀: variabila independentă **nou scoasa din model** nu are o influență semnificativă asupra variației variabilei dependente

H₁: variabila independentă **nou scoasă din model** are o influență semnificativă asupra variației variabilei dependente

2. Fixarea pragului de semnificație $\alpha=0,05$

3. Alegerea statisticii test

$$F_{\text{calc}} = \frac{\hat{V}_{E_{\text{old}}} - \hat{V}_{E_{\text{new}}}}{\hat{V}_{R_{\text{old}}}} \cdot \frac{n - k_{\text{old}}}{k_{\text{old}} - k_{\text{new}}} = \frac{\hat{\eta}_{\text{old}}^2 - \hat{\eta}_{\text{new}}^2}{1 - \hat{\eta}_{\text{old}}^2} \cdot \frac{n - k_{\text{old}}}{k_{\text{old}} - k_{\text{new}}}$$

4. Calcularea statisticii test

$$F_{\text{calc}} = \frac{\text{ESS}_{\text{old}} - \text{ESS}_{\text{new}}}{\text{RSS}_{\text{old}}} \cdot \frac{n - k_{\text{old}}}{k_{\text{old}} - k_{\text{new}}} = \frac{R_{\text{old}}^2 - R_{\text{new}}^2}{1 - R_{\text{old}}^2} \cdot \frac{n - k_{\text{old}}}{k_{\text{old}} - k_{\text{new}}}$$

5. Criterii de decizie:

Dacă $F_{\text{calc}} \leq F_{\alpha, k_{\text{new}} - k_{\text{old}}, n - k_{\text{new}}}$ => nu respingem H₀ (AH₀) cu o prob. de 1- α .

Dacă $F_{\text{calc}} > F_{\alpha, k_{\text{new}} - k_{\text{old}}, n - k_{\text{new}}}$ => se respinge H₀ (RH₀) cu un risc asumat α .



Utilitatea modelului de regresie cu variabile standardizate

Modelul liniar multiplu cu variabile standardizate permite compararea coeficienților de regresie din model; fiecare coeficient arătând impactul partial al variației cu o unitate a variabilei independente standardizate asupra variabilei dependente standardizate.

Valoarea coeficientilor de regresie din modelul standardizat se interpreteaza ca **abateri standard pentru variabila dependenta**.

Aceasta este o modalitate de **ierarhizare a variabilelor dependente** în funcție de importanța lor în model.

Cel mai mare coeficient in valoare absoluta indica **cea mai mare influenta asupra variabilei dependente**, iar semnul coeficientului arata sensul acestei influente.



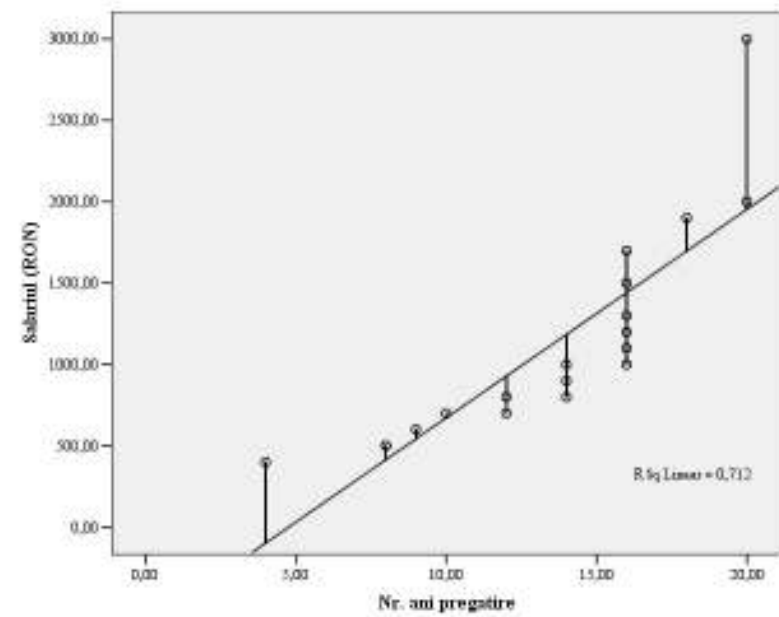
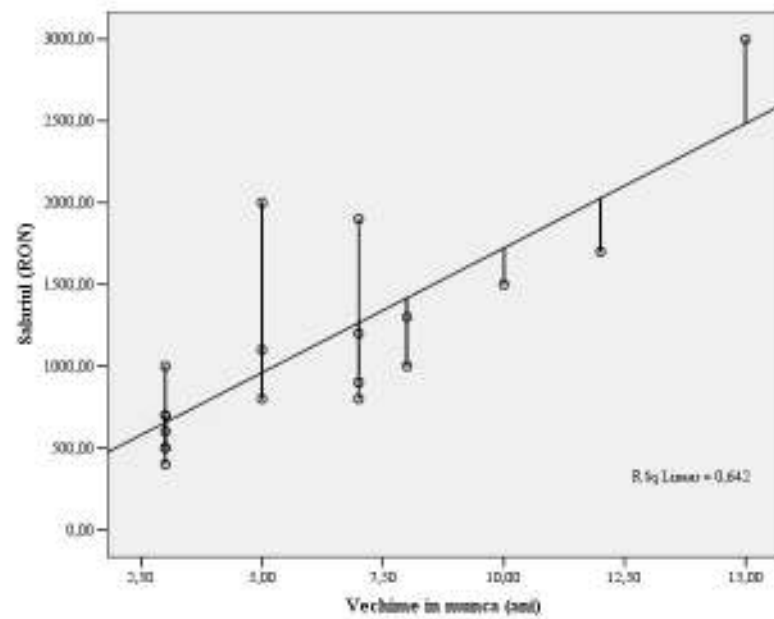
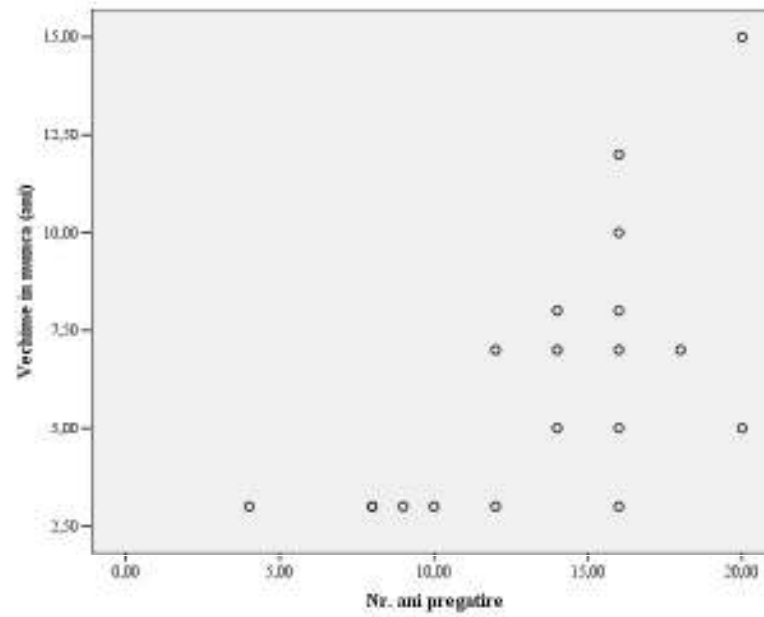
EXEMPLUL 1:

Var. dependenta:

Salariul

Var. independente (predictori):

Nr. de ani de pregatire, Vechimea in munca



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,910 ^a	,829	,807	285,65322	,829	38,718	2	16	,000

a. Predictors: (Constant), Nr. ani pregatire, Vechime in munca (ani)

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6318646	2	3159323,188	38,718	,000 ^a
	Residual	1305564	16	81597,759		
	Total	7624211	18			

a. Predictors: (Constant), Nr. ani pregatire, Vechime in munca (ani)

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-545,101	224,894		-2,424	,028
	Vechime in munca (ani)	84,315	25,550	,443	3,300	,005
	Nr. ani pregatire	85,298	20,394	,562	4,182	,001

a. Dependent Variable: Salariul (RON)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,910 ^a	,829	,807	285,65322

a. Predictors: (Constant), Nr. ani pregatire, Vechime in munca (ani)

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6318646	2	3159323,188	38,718	,000 ^a
	Residual	1305564	16	81597,759		
	Total	7624211	18			

a. Predictors: (Constant), Nr. ani pregatire, Vechime in munca (ani)

b. Dependent Variable: Salariul (RON)

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	-545,101	224,894		-2,424	,028			
	Vechime in munca (ani)	84,315	25,550	,443	3,300	,005	,801	,636	,341
	Nr. ani pregatire	85,298	20,394	,562	4,182	,001	,844	,723	,433

a. Dependent Variable: Salariul (RON)

Corelatii bivariate și partiale

Correlations

		Salariul (RON)	Vechime in munca (ani)	Nr. ani pregatire
Salariul (RON)	Pearson Correlation	1	.801**	.844**
	Sig. (2-tailed)		.000	.000
	N	19	19	19
Vechime in munca (ani)	Pearson Correlation	.801**	1	.637**
	Sig. (2-tailed)	.000		.003
	N	19	19	19
Nr. ani pregatire	Pearson Correlation	.844**	.637**	1
	Sig. (2-tailed)	.000	.003	
	N	19	19	19

** . Correlation is significant at the 0.01 level (2-tailed).

Correlations

Control Variables			Salariul (RON)	Vechime in munca (ani)
Nr. ani pregatire	Salariul (RON)	Correlation	1.000	.636
		Significance (2-tailed)	.	.005
		df	0	16
	Vechime in munca (ani)	Correlation	.636	1.000
		Significance (2-tailed)	.005	.
		df	16	0

Correlations

Control Variables			Vechime in munca (ani)	Nr. ani pregatire
Salariul (RON)	Vechime in munca (ani)	Correlation	1.000	-.120
		Significance (2-tailed)	.	.635
		df	0	16
	Nr. ani pregatire	Correlation	-.120	1.000
		Significance (2-tailed)	.635	.
		df	16	0

Correlations

Control Variables			Nr. ani pregatire	Salariul (RON)
Vechime in munca (ani)	Nr. ani pregatire	Correlation	1.000	.723
		Significance (2-tailed)	.	.001
		df	0	16
	Salariul (RON)	Correlation	.723	1.000
		Significance (2-tailed)	.001	.
		df	16	0

EXEMPLUL 2:

Var. dependenta:

Greutate(kg)

Var. independente:

Inaltime (cm), Consum zilnic paine(g)

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.772 ^a	.596	.577	4.86988

a. Predictors: (Constant), PAINE/zi(g), INALTIMEA (CM)

b. Dependent Variable: GREUTATEA (KG)

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1471.049	2	735.525	31.014	.000 ^b
	Residual	996.062	42	23.716		
	Total	2467.111	44			

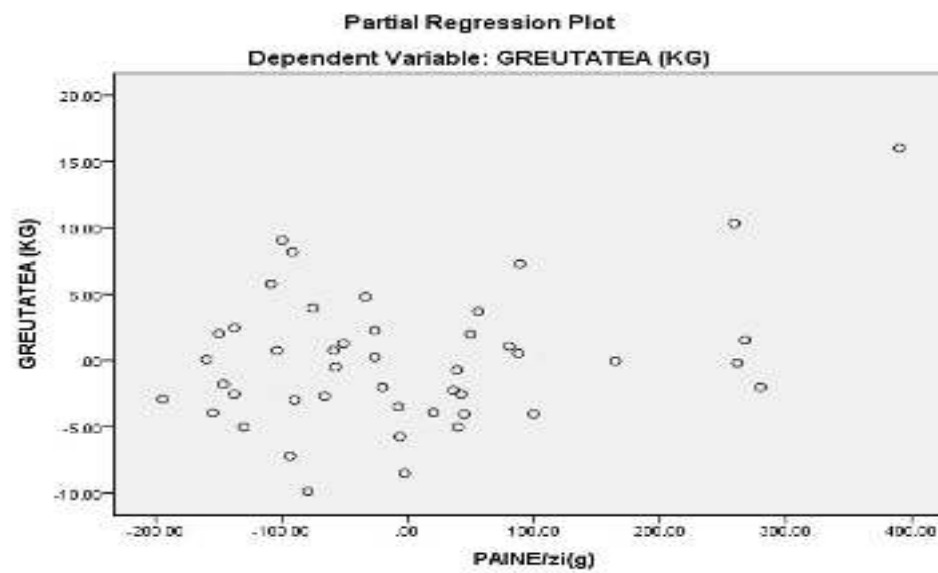
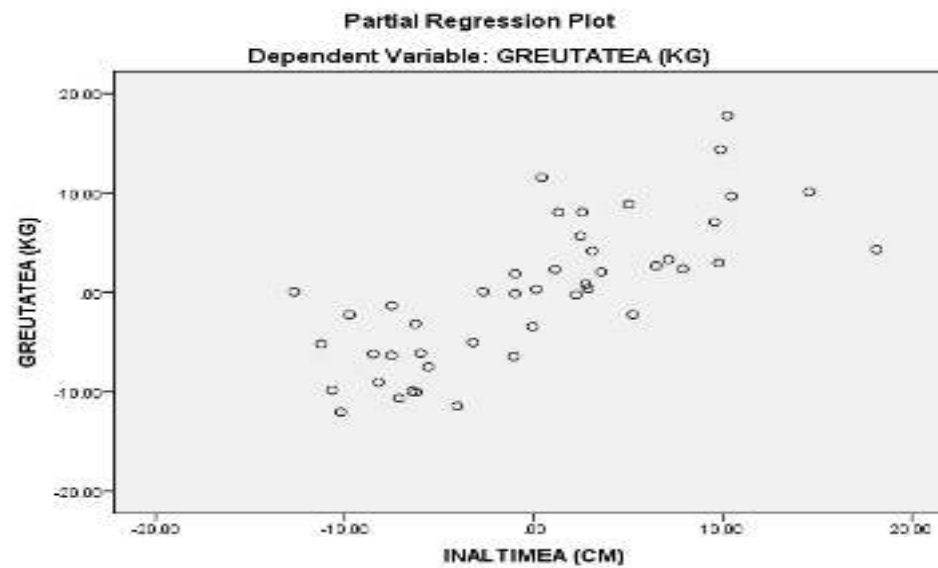
a. Dependent Variable: GREUTATEA (KG)

b. Predictors: (Constant), PAINE/zi(g), INALTIMEA (CM)

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	-67.093	16.372		-4.098	.000			
	INALTIMEA (CM)	.729	.099	.726	7.388	.000	.741	.752	.724
	PAINE/zi(g)	.012	.006	.218	2.221	.032	.268	.324	.218

a. Dependent Variable: GREUTATEA (KG)



Correlations

Control Variables			INALTIMEA (CM)	GREUTATEA (KG)	PAINE/zi(g)
-none- ^a	INALTIMEA (CM)	Correlation	1.000	.741	.068
		Significance (2-tailed)	.	.000	.658
		df	0	43	43
	GREUTATEA (KG)	Correlation	.741	1.000	.268
		Significance (2-tailed)	.000	.	.076
		df	43	0	43
	PAINE/zi(g)	Correlation	.068	.268	1.000
		Significance (2-tailed)	.658	.076	
		df	43	43	0
PAINE/zi(g)	INALTIMEA (CM)	Correlation	1.000	.752	
		Significance (2-tailed)	.	.000	
		df	0	42	
	GREUTATEA (KG)	Correlation	.752	1.000	
		Significance (2-tailed)	.000	.	
		df	42	0	

a. Cells contain zero-order (Pearson) correlations.

Correlations

Control Variables			GREUTATEA (KG)	PAINE/zi(g)	INALTIMEA (CM)
-none- ^a	GREUTATEA (KG)	Correlation	1.000	.268	.741
		Significance (2-tailed)	.	.076	.000
		df	0	43	43
	PAINE/zi(g)	Correlation	.268	1.000	.068
		Significance (2-tailed)	.076	.	.658
		df	43	0	43
	INALTIMEA (CM)	Correlation	.741	.068	1.000
		Significance (2-tailed)	.000	.658	.
		df	43	43	0
INALTIMEA (CM)	GREUTATEA (KG)	Correlation	1.000	.324	
		Significance (2-tailed)	.	.032	
		df	0	42	
	PAINE/zi(g)	Correlation	.324	1.000	
		Significance (2-tailed)	.032	.	
		df	42	0	

a. Cells contain zero-order (Pearson) correlations.



EXEMPLUL 3

Y: Nota examen Econometrie

X1: Nota examen Matematica

X2: Nota examen Statistica

Model Summary^c

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.863 ^a	.745	.716	.8943	.745	26.241	1	9	.001
2	.898 ^b	.806	.757	.8276	.061	2.509	1	8	.152

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20.985	1	20.985	26.241	.001 ^b
	Residual	7.197	9	.800		
	Total	28.182	10			
2	Regression	22.703	2	11.351	16.575	.001 ^c
	Residual	5.479	8	.685		
	Total	28.182	10			

a. Dependent Variable: *ECONOMETRIE*

b. Predictors: (Constant), *MATE*

c. Predictors: (Constant), *MATE*, *STAT*

Coefficients ^a									
Model		Unstandardize d Coefficients		Standardize d Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero- order	Partial	Part
1	(Constant)	2.167	1.033		2.098	.065			
	MATE	.685	.134	.863	5.123	.001	.863	.863	.863
2	(Constant)	1.604	1.020		1.573	.154			
	MATE	.442	.197	.557	2.245	.055	.863	.622	.350
	STAT	.307	.194	.393	1.584	.152	.827	.489	.247

a. Dependent Variable: ECONOMETRIE

		Correlations		
		MATE (X1)	STAT (X2)	ECONOM (Y)
MATE (X1)	Pearson Correlation	1	.778**	.863**
	Sig. (2-tailed)		.005	.001
	N	11	11	11
STAT (X2)	Pearson Correlation	.778**	1	.827**
	Sig. (2-tailed)	.005		.002
	N	11	11	11
ECONOMETRIE (Y)	Pearson Correlation	.863**	.827**	1
	Sig. (2-tailed)	.001	.002	
	N	11	11	11

** . Correlation is significant at the 0.01 level (2-tailed).

