

CAPITOLUL 5

Verificarea ipotezelor clasice ale modelului de regresie

1. Introducere

Modelarea econometrică se realizează în anumite condiții sau cu respectarea unui set de restricții care se numesc *ipoteze ale modelului de regresie*. Calitatea estimării parametrilor modelului de regresie depinde de îndeplinirea a două clase de ipoteze: ipoteze asupra componentei aleatoare sau asupra variabilei eroare și ipoteze asupra componentei deterministe sau asupra variabilelor independente.

- ipoteze asupra *componentei aleatoare* sau asupra *variabilei eroare*:
 - $M(\varepsilon_i) = 0$, media erorilor este nulă;
 - $V(\varepsilon_i) = \sigma^2$, ipoteza de homoscedasticitate;
 - $\varepsilon_i \sim N(0, \sigma^2)$, ipoteza de normalitate;
 - $cov(\varepsilon_i, \varepsilon_j) = 0$, ipoteza de necorelare sau de independență a erorilor.
- ipoteze asupra *componentei deterministe* sau asupra *variabilelor independente*:
 - variabilele independente sunt non-aleatoare, deterministe;
 - variabilele independente și variabila eroare sunt necorelate, $cov(X_j, \varepsilon) = 0$;
 - variabilele independente sunt necoliniare (cazul regresiei liniare multiple).

Nerespectarea acestor ipoteze:

- determină modificarea proprietăților estimatorilor parametrilor modelului de regresie;
- ridică probleme importante în realizarea demersului cercetării econometrice.

Proprietăților estimatorilor parametrilor modelului de regresie:

- estimatorii parametrilor sunt *nedeplasați*: $M(\hat{\beta}_j) = \beta_j$;
- estimatorii parametrilor urmează *o lege de repartiție normală*: $\hat{\beta}_j \sim N(\beta_j, \sigma_{\hat{\beta}_j}^2)$;
- estimatorii parametrilor sunt *eficienți*: $V(\hat{\beta}_j) = \min$;
- estimatorii parametrilor sunt *convergenți*: $\hat{\beta}_j \xrightarrow[p]{n \in N} \beta_j$.

2. Ipoteze cu privire la erori (variabila reziduală sau componenta aleatoare)

2.1. Media erorilor este zero: $M(\varepsilon_i) = 0$

a. Definirea ipotezei

Potrivit acestei ipoteze, restricția modelării econometrice este ca toți ceilalți factori neincluși în model și reprezentați de variabila reziduală, precum și erorile determinate de metoda statistică să nu afecteze sistematic media variabilei dependente Y .

Ipoteza $M(\varepsilon_i) = 0$ este echivalentă cu condiția: $M(Y/X) = \beta_0 + \beta_1 X$.

b. Efectele încălcării ipotezei

Dacă această ipoteză este încălcată, atunci se modifică proprietățile estimatorilor parametrilor modelului de regresie. Există două situații: când media variabilei reziduale este constantă și când aceasta nu este constantă.

- $M(\varepsilon_i) = \mu = \text{cst}$ - parametrul β_0 este estimat *nedeplasat* de estimatorul $\hat{\beta}_0^*$:
 $M(\hat{\beta}_0^*) = \beta_0 + \mu$
- $M(\varepsilon_i) = \mu_i$ - parametrul β_1 este estimat *nedeplasat* de estimatorul $\hat{\beta}_1$.

În concluzie, dacă ipoteza $M(\varepsilon_i) = 0$ este încălcată, estimarea parametrilor modelului se realizează cu o eroare sistematică. Este vorba de o deplasare de care suferă fie estimarea parametrului β_1 , fie estimarea parametrului β_0 .

c. Testarea ipotezei

(1) Formularea ipotezelor:

- ipoteza nulă $H_0: M(\varepsilon_i) = 0$ (media erorilor este nulă; media erorilor nu este semnificativă sau nu este reprezentativă; erorile nu afectează sistematic media variabilei dependente Y)
- ipoteza alternativă $H_1: M(\varepsilon_i) \neq 0$ (media erorilor este diferită semnificativ de zero; erorile nu au media nulă; media erorilor este semnificativă; erorile afectează sistematic media variabilei dependente Y)

(2) Alegerea pragului de semnificație: $\alpha = 0,05$

(3) Alegerea statisticii test:

$$t = \frac{\hat{M}(\varepsilon_i) - M(\varepsilon_i)}{\sqrt{\hat{V}(\hat{M}(\varepsilon_i))}} \sim t(n-1)$$

(4) Citirea valorii teoretice a statisticii test:

$$t_{teoretic} = t_{\alpha/2; n-1} = t_{0,025; 473} = 1,96$$

(5) Calcularea statisticii test:

$$t_{calc} = \frac{M(e_i)}{s/\sqrt{n}} = \frac{M(e_i)}{s_{\hat{M}(e_i)}}$$
$$t_{calc} = \frac{M(e_i)}{s/\sqrt{n}} = \frac{0}{12819/\sqrt{474}} = 0$$
$$t_{calc} = \frac{M(e_i)}{s_{\hat{M}(e_i)}} = \frac{0}{588,84} = 0$$

(6) Regula de decizie:

În funcție de statistica t :

- dacă $|t_{calc}| \leq t_{\alpha/2; n-1}$, atunci nu se respinge (se acceptă) ipoteza nulă (H_0), în condițiile unui risc α
- dacă $|t_{calc}| > t_{\alpha/2; n-1}$, atunci se respinge ipoteza nulă (H_0)

În funcție de Sig (probabilitatea asociată statisticii test):

- dacă $Sig \geq \alpha$, atunci se acceptă ipoteza nulă (H_0), în condițiile unui risc α
- dacă $Sig < \alpha$, atunci se respinge ipoteza nulă (H_0)

(7) Luarea deciziei:

- dacă nu se respinge ipoteza H_0 , se îndeplinește ipoteza cu privire la media erorilor, adică erorile au media nulă sau media erorilor nu este reprezentativă.
- dacă se respinge ipoteza H_0 , ipoteza cu privire la media erorilor nu este îndeplinită, adică media erorilor diferă semnificativ de zero, este reprezentativă.

$$|t_{calc}| = 0 < t_{\alpha/2; n-1} = 1,96 \Rightarrow \text{nu se respinge (se acceptă) ipoteza nulă } (H_0)$$

$$Sig = 1,000 > \alpha = 0,05 \Rightarrow \text{nu se respinge (se acceptă) ipoteza nulă } (H_0)$$

(8) Interpretarea deciziei luate:

În condițiile unui risc de 5%, se poate garanta că media erorilor nu este semnificativă sau nu este reprezentativă sau că erorile nu afectează sistematic media variabilei dependente Y . Cu alte cuvinte, ipoteza media erorilor este nulă ($M(\varepsilon_i) = 0$) este îndeplinită.

Observație:

Tabelele pe baza cărora se poate calcula valoarea statisticii test Student sau se poate lua decizia cu privire la această ipoteză sunt: *Residual Statistic*, *One Sample Statistics*, *One Sample Test*, *One Sample Kolmogorov-Smirnov Test*, *Descriptive Statistics*.

Aplicație:

În studiul legăturii dintre variabila dependentă, *Salariul* (\$), și variabila independentă, *Nivelul de educație* (ani), înregistrate pentru un eșantion de 474 de persoane, s-au obținut următoarele rezultate:

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	-18331,2	2821,912		-6,496	,000
Educational Level (years)	3909,907	204,547	,661	19,115	,000

a. Dependent Variable: Current Salary

Să se testeze dacă media erorilor este zero.

Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	\$12,948.08	\$63,776.86	\$34,419.57	\$11,279.480	474
Residual	-\$21,567.422	\$79,042.953	\$.000	\$12,819.966	474
Std. Predicted Value	-1,904	2,603	,000	1,000	474
Std. Residual	-1,681	6,159	,000	,999	474

a. Dependent Variable: Current Salary

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Unstandardized Residual	474	,0000000	12819,96640	588,8406

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Unstandardized Residual	,000	473	1,000	,00000000	-1157,07	1157,067

One-Sample Kolmogorov-Smirnov Test

		Unstandardized Residual
N		474
Normal Parameters ^{a,b}	Mean	,0000000
	Std. Deviation	12819,96639730
Most Extreme Differences	Absolute	,110
	Positive	,110
	Negative	-,069
Kolmogorov-Smirnov Z		2,400
Asymp. Sig. (2-tailed)		,000

a. Test distribution is Normal.

b. Calculated from data.

Descriptive Statistics

	N	Minimu	Maximu	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Unstandardized Residual	474	-21567	79043,0	,000000	12819,97	1,764	,112	5,798	,224
Valid N (listwise)	474								

2.2. Ipoteza de homoscedasticitate: $V(\varepsilon_i) = \sigma^2$

a. Definirea ipotezei

În cazul a două variabile X , Y , între care există o legătură liniară, regresia este o medie condiționată definită pe repartiția bidimensională (X, Y) și pe repartițiile condiționate de forma: $Y | X = x_i$.

Regresia liniară este dată prin relația:

$$M(Y | X = x_i) = f(x_i) = \beta_0 + \beta_1 x_i.$$

La nivelul fiecărei repartiții condiționate, se definesc variabilele reziduale:

$$\varepsilon_i = y_i - M(Y | X = x_i) = y_i - \beta_0 - \beta_1 x_i.$$

Erorile astfel definite sunt homoscedastice dacă varianțele acestora sunt egale și constante: $V(\varepsilon_i) = \sigma^2 = \text{constantă}$.

b. Efectele încălcării ipotezei

Dacă ipoteza de homoscedasticitate este încălcată, *modelul de regresie* se numește *heteroscedastic*. Efectul încălcării ipotezei de homoscedasticitate este *pierderea eficienței* estimatorilor parametrilor modelului de regresie (estimează parametrul cu o varianță mai mare).

c. Testarea ipotezei

c.1. Procedee grafice: reprezentarea distribuției erorilor și aprecierea varianței acesteia (*Scatter plot*)

c.2. Testul Glejser (cazul regresiei liniare simple): presupune testarea semnificației parametrului α_1 din modelul de regresie estimat construit pe baza variabilei reziduale în valoare absolută ($|e|$) ca variabilă dependentă și variabila independentă (X). Ideea de bază a acestui test este că varianțele erorilor σ_i^2 ar putea fi explicate prin valorile variabilei independente.

Testarea homoscedasticității cu ajutorul testului Glejser presupune parcurgerea următorului demers:

- se construiește modelul de regresie $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ și se estimează valorile:
 $y_{x_i} = b_0 + b_1 x_i$;
- pentru modelul propus, se determină erorile estimate:
 $e_i = y_i - y_{x_i} = y_i - b_0 - b_1 x_i$;
- se construiește un model de regresie pe baza erorilor estimate în valoare absolută și variabila independentă aleasă ca posibilă sursă a heteroscedasticității. Un exemplu este reprezentat de modelul liniar de forma:
 $|e_i| = \alpha_0 + \alpha_1 x_i + u_i$
- se testează modelul din etapa anterioară: dacă parametrul α_1 este semnificativ, atunci modelul inițial este heteroscedastic. În caz contrar, modelul este homoscedastic.

(1) Formularea ipotezelor:

- $H_0: V(\varepsilon_i) = \sigma^2$ (varianța sau dispersia erorilor este constantă, adică erorile sunt homoscedastice sau au aceeași varianță; modelul de regresie este homoscedastic)
- $H_1: V(\varepsilon_i) \neq \sigma^2$ (varianța sau dispersia erorilor nu este constantă, adică erorile sunt heteroscedastice sau nu au aceeași varianță; modelul de regresie este heteroscedastic)

(1*) Formularea ipotezelor intermediare:

- $H_0: \alpha_1 = 0$ (parametrul α_1 nu este semnificativ, adică între erori și variabila independentă nu există o legătură liniară semnificativă sau variabila independentă nu explică semnificativ liniar variabila reziduală)
- $H_1: \alpha_1 \neq 0$ (parametrul α_1 este semnificativ, adică între erori și variabila independentă există o legătură liniară semnificativă sau variabila independentă explică semnificativ liniar variabila reziduală)

(2) Alegerea pragului de semnificație: $\alpha = 0,05$

(3) Alegerea statisticii test:

$$t = \frac{\hat{\alpha}_1 - \alpha_1}{\hat{\sigma}_{\hat{\alpha}_1}} \sim t(n-2)$$

(4) Citirea valorii teoretice a statisticii test:

$$t_{teoretic} = t_{\alpha/2; n-2} = t_{0,025; 472} = 1,96$$

(5) Calcularea statisticii test:

$$t_{calc} = \frac{a_1}{s_{\hat{a}_1}} = \frac{a_1}{std. error} = \frac{821,842}{135,194} = 6,079$$

(6) Regula de decizie:

În funcție de statistica t :

- dacă $|t_{calc}| \leq t_{\alpha/2; n-2}$, atunci se acceptă ipoteza nulă (H_0), în condițiile unui risc α
- dacă $|t_{calc}| > t_{\alpha/2; n-2}$, atunci se respinge ipoteza nulă (H_0)

În funcție de Sig (probabilitatea asociată statisticii test):

- dacă $Sig \geq \alpha$ atunci se acceptă ipoteza nulă (H_0), în condițiile unui risc α
- dacă $Sig < \alpha$, atunci se respinge ipoteza nulă (H_0)

(7) Luarea deciziei:

- dacă se acceptă ipoteza H_0 , parametrul α_1 nu este semnificativ sau nu diferă semnificativ de zero, ceea ce înseamnă că varianța sau dispersia erorilor este constantă, adică se îndeplinește ipoteza de homoscedasticitate a erorilor.

- dacă se respinge ipoteza H_0 , parametrul α_1 este semnificativ, ceea ce înseamnă că varianța sau dispersia erorilor nu este constantă, adică nu se îndeplinește ipoteza de homoscedasticitate a erorilor; erorile sunt heteroscedastice.

$$|t_{calc}| = 6,079 > t_{\alpha/2; n-2} = 1,96 \Rightarrow \text{se respinge ipoteza nulă } (H_0)$$

$$Sig = 0,000 < \alpha = 0,05 \Rightarrow \text{respinge ipoteza nulă } (H_0)$$

(8) Interpretarea deciziei luate:

În condițiile unui risc de 5%, se poate garanta că parametrul α_1 este semnificativ (adică între erori și variabila independentă există o legătură liniară semnificativă sau variabila independentă explică semnificativ liniar variabila reziduală), ceea ce înseamnă că varianța sau dispersia erorilor nu este constantă, adică erorile sunt heteroscedastice sau nu au aceeași varianță (modelul de regresie este heteroscedastic). Cu alte cuvinte, ipoteza de homoscedasticitate a erorilor ($V(\varepsilon_i) = \sigma^2$) nu este îndeplinită.

Observație:

Tabelele pe baza cărora se poate calcula testul sau se poate lua decizia cu privire la această ipoteză sunt: *Coefficients*.

Aplicație:

Se studiază legătura dintre variabila dependentă, *Salariul* (\$), și variabila independentă, *Nivelul de educație* (ani), înregistrate pentru un eșantion de 474 de persoane. În vederea verificării ipotezei de homoscedasticitate a erorilor, se aplică testul Glejser iar rezultatele obținute sunt prezentate în tabelul de mai jos:

Coefficients ^a					
Model		Unstandardized Coefficients		Standardized Coefficients	
		B	Std. Error	Beta	
1	(Constant)	-1773,944	1865,129		-,951
	Educational Level (years)	821,842	135,194	,269	6,079
					,000

a. Dependent Variable: abs

c.2. Testul Breusch-Pagan-Godfrey (cazul regresiei liniare multiple): ideea de bază a acestui test este aceea de a construi un model de regresie care are ca variabilă dependentă variabila eroare estimată la pătrat, iar ca variabile independente pot fi utilizate variabilele din modelul de regresie de bază pentru care se testează heteroscedasticitatea.

Astfel, pentru un model de forma $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, testarea homoscedasticității presupune următorul demers:

- se construiește modelul de regresie de bază și se estimează parametrii acestuia;
- se estimează erorile pe baza datelor de la nivelul unui eșantion reprezentativ;

- se construiește modelul auxiliar de regresie:

$$e_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + u_i$$
- pe baza estimației coeficientului de determinație din acest model (R_a^2), se construiește un test Chi-pătrat cu $k - 1$ grade de libertate, unde k este numărul de parametri din modelul auxiliar;
- valoarea calculată a testului este de forma:

$$\chi_{calc}^2 = n \cdot R_a^2;$$
- decizia de a accepta ipoteza de homoscedasticitate se ia pe baza comparării valorii teoretice din tabele ($\chi_{\alpha, k-1}^2$) cu cea calculată în etapa anterioară.

(1) Formularea ipotezelor:

- $H_0: V(\varepsilon_i) = \sigma^2$ (varianța sau dispersia erorilor este constantă, adică erorile sunt homoscedastice sau au aceeași varianță; modelul de regresie este homoscedastic)
- $H_1: V(\varepsilon_i) \neq \sigma^2$ (varianța sau dispersia erorilor nu este constantă, adică erorile sunt heteroscedastice sau nu au aceeași varianță; modelul de regresie este heteroscedastic)

(2) Alegerea pragului de semnificație: $\alpha = 0,05$

(3) Regula de decizie:

În funcție de *Sig* (probabilitatea asociată statisticii test):

- dacă $Sig \geq \alpha$ atunci se acceptă ipoteza nulă (H_0), în condițiile unui risc α
- dacă $Sig < \alpha$, atunci se respinge ipoteza nulă (H_0)

(4) Luarea deciziei:

- dacă se acceptă ipoteza H_0 , varianța sau dispersia erorilor este constantă, adică se îndeplinește ipoteza de homoscedasticitate a erorilor.
- dacă se respinge ipoteza H_0 , nu se îndeplinește ipoteza de homoscedasticitate a erorilor; erorile sunt heteroscedastice.

$Sig = 0,000 < \alpha = 0,05$ se respinge ipoteza nulă (H_0)

(5) Interpretarea deciziei luate:

În condițiile unui risc de 5%, se poate garanta că varianța sau dispersia erorilor nu este constantă, adică erorile sunt heteroscedastice sau nu au aceeași varianță, modelul de regresie este heteroscedastic (modelul de regresie auxiliar este semnificativ statistic – specificul testului Breusch-Pagan-Godfrey).

Aplicație:

Se studiază legătura dintre variabila dependentă, *Salariul* (\$), și variabilele independente, *Nivelul de educație* (ani) și *Experiența în muncă* (ani), înregistrate pentru un eșantion de 474 de

persoane. În vederea verificării ipotezei de homoscedasticitate a erorilor, se aplică testul Breusch-Pagan-Godfrey, iar rezultatele obținute sunt prezentate în tabelul de mai jos:

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	147,119	2	73,559	74,003	,001(a)
	Residual	468,201	471	0,994		
	Total	615,420	473			

a Predictors: (Constant), Education level, Experience

b Dependent Variable: **Residuals-squared**

2.3. Ipoteza de normalitate a erorilor: $\varepsilon_i \sim N(0, \sigma^2)$

a. Definirea ipotezei

Erorile ε_i urmează o lege normală de medie 0 și varianță σ^2 : $\varepsilon_i \sim N(0, \sigma^2)$

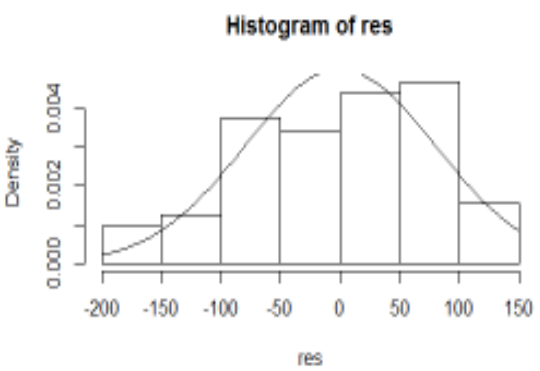
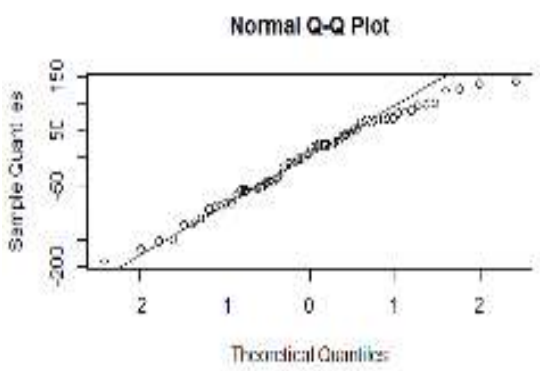
b. Efectele încălcării ipotezei

Ipoteza de normalitate a erorilor este importantă pentru stabilirea proprietăților estimatorilor parametrilor modelului de regresie:

- dacă $\varepsilon_i \sim N(0, \sigma^2)$, atunci estimatorii parametrilor modelului de regresie urmează, de asemenea, o lege normală: $\hat{\beta}_j \sim N(\beta_j, \sigma_{\hat{\beta}_j}^2)$;
- dacă ipoteza de normalitate este încălcată, atunci estimatorii construiți pe baza metodei celor mai mici pătrate nu urmează o lege de repartiție normală, având doar proprietăți asimptotice, adică necesită eșantioane sau seturi mari de date.

c. Testarea ipotezei

c.1. Procedee grafice: histograma (curba frecvențelor); Box-Plot; Q-Q Plot; P-P Plot.

- pe baza <i>histogramei erorilor</i>	- pe baza <i>diagramei QQ-Plot</i>
 <p>The histogram shows the distribution of residuals. The x-axis is labeled 'res' and ranges from -200 to 150. The y-axis is labeled 'Density' and ranges from 0.000 to 0.004. The bars show a distribution that is skewed to the left, with a peak around 50. A normal distribution curve is overlaid, which does not fit the data well, indicating a violation of the normality assumption.</p>	 <p>The Normal Q-Q Plot shows the sample quantiles on the y-axis (ranging from -200 to 150) against the theoretical quantiles on the x-axis (ranging from -2 to 2). The points follow a straight line for most of the range but show a distinct S-shaped deviation at the tails, which is characteristic of non-normal residuals.</p>
Interpretare: Distribuția erorilor este asimetrică la stânga (după asimetrie) și platicurtică (după boltire). Putem considera că repartiția erorilor nu respectă condiția de normalitate.	Interpretare: Erorile modelului de regresie dintre cele două variabile analizate par să se abată de la forma repartiției normale. Evident, o decizie clară cu privire la ipoteza de normalitate presupune realizarea unui test.

c.2. Testul Kolmogorov-Smirnov:

(1) Formularea ipotezelor:

- $H_0: \varepsilon_i \sim N(0, \sigma^2)$ (ipoteza de normalitate: erorile urmează o lege de repartiție normală sau erorile sunt normal distribuite sau distribuția erorilor **nu diferă semnificativ** de distribuția normală)
- $H_1: \varepsilon_i \neq N(0, \sigma^2)$ (erorile nu urmează o lege de repartiție normală sau erorile nu sunt normal distribuite sau distribuția erorilor **diferă semnificativ** de distribuția normală)

(2) Alegerea pragului de semnificație: $\alpha = 0,05$

(3) Regula de decizie:

În funcție de Sig (probabilitatea asociată statisticii test):

- dacă $Sig \geq \alpha$ atunci nu se respinge (se acceptă) ipoteza nulă (H_0), în condițiile unui risc α
- dacă $Sig < \alpha$, atunci se respinge ipoteza nulă (H_0)

(4) Luarea deciziei:

- dacă se acceptă ipoteza H_0 , este îndeplinită ipoteza de normalitate a erorilor, adică erorile sunt normal distribuite;
- dacă se respinge ipoteza H_0 , erorile nu urmează o lege de repartiție normală.

$$Sig = 0,000 < \alpha = 0,05 \Rightarrow \text{se respinge ipoteza nulă } (H_0)$$

(6) Interpretarea deciziei luate:

În condițiile unui risc de 5%, se poate garanta că erorile nu urmează o lege de repartiție normală sau erorile nu sunt normal distribuite sau distribuția erorilor diferă semnificativ de distribuția normală. Cu alte cuvinte, ipoteza de normalitate a erorilor ($\varepsilon_i \sim N(0, \sigma^2)$) nu este îndeplinită.

Observație:

Tabelele pe baza cărora se poate lua decizia cu privire la această ipoteză sunt: **One Sample Kolmogorov-Smirnov Test**.

Aplicație:

Se studiază legătura dintre variabila dependentă, *Salariul* (\$), și variabila independentă, *Nivelul de educație* (ani), înregistrate pentru un eșantion de 474 de persoane. În vederea verificării ipotezei de normalitate a erorilor, se aplică testul Kolmogorov-Smirnov, iar rezultatele obținute sunt prezentate în tabelul de mai jos:

One-Sample Kolmogorov-Smirnov Test

		Unstandardized Residual
N		474
Normal Parameters ^{a,b}	Mean	,0000000
	Std. Deviation	12819,96639730
Most Extreme Differences	Absolute	,110
	Positive	,110
	Negative	-,069
Kolmogorov-Smirnov Z		2,400
Asymp. Sig. (2-tailed)		,000

a. Test distribution is Normal.

b. Calculated from data.

c.2. Testul Jarque-Bera: se bazează pe verificarea simultană a proprietăților de asimetrie (coeficientul de asimetrie - *Skewness* (**Sw**)) și boltire (coeficientul de boltire - *Kurtosis* (**K**)) a seriei reziduurilor (erorilor)

(1) Formularea ipotezelor:

- $H_0: \varepsilon_i \sim N(0, \sigma^2)$ (ipoteza de normalitate: erorile urmează o lege de repartiție normală sau erorile sunt normal distribuite sau distribuția erorilor **nu diferă semnificativ** de distribuția normală)
- $H_1: \varepsilon_i \neq N(0, \sigma^2)$ (erorile nu urmează o lege normală sau erorile nu sunt normal distribuite sau distribuția erorilor **diferă semnificativ** de distribuția normală)

(2) Alegerea pragului de semnificație: $\alpha = 0,05$

(3) Alegerea statisticii test:

$$JB = \frac{n}{6} \left(\widehat{Sw}^2 + \frac{\widehat{K}^2}{4} \right) \sim \chi^2(2)$$

(4) Citirea valorii teoretice a statisticii test:

$$\chi_{teoretic}^2 = \chi_{\alpha;2}^2 = \chi_{0,05;2}^2 = 5,991$$

(5) Calcularea statisticii test:

$$JB_{calc} = \frac{n}{6} \left(sw^2 + \frac{k^2}{4} \right) = \frac{474}{6} \left(1,764^2 + \frac{5,798^2}{4} \right) = 908,84$$

unde sw este estimația coeficientul de asimetrie Sw și k este estimația coeficientul de boltire K .

(6) Regula de decizie:

În funcție de valoarea calculată a statisticii JB :

- dacă $JB_{calc} \leq \chi^2_{\alpha;2}$, atunci se acceptă ipoteza nulă (H_0), în condițiile unui risc α
- dacă $JB_{calc} > \chi^2_{\alpha;2}$, atunci se respinge ipoteza nulă (H_0)

(7) Luarea deciziei:

- dacă se acceptă ipoteza H_0 , este îndeplinită ipoteza de normalitate a erorilor, adică erorile sunt normal distribuite;
- dacă se respinge ipoteza H_0 , erorile nu urmează o lege de repartiție normală, adică ipoteza de normalitate a erorilor este încălcată.

$$JB_{calc} = 908,84 > \chi^2_{\alpha;2} = 5,991 \Rightarrow \text{atunci se respinge ipoteza nulă } (H_0)$$

(8) Interpretarea deciziei luate:

În condițiile unui risc de 5%, se poate garanta că erorile nu urmează o lege de repartiție normală sau erorile nu sunt normal distribuite sau distribuția erorilor diferă semnificativ de distribuția normală. Cu alte cuvinte, ipoteza de normalitate a erorilor ($\varepsilon_i \sim N(0, \sigma^2)$) nu este îndeplinită.

Observație:

Tabelele pe baza cărora se poate calcula testul: tabelul *Descriptive Statistics* în care apar indicatorii *Skewness* și *Kurtosis*.

Aplicație:

Se studiază legătura dintre variabila dependentă, *Salariul* (\$), și variabila independentă, *Nivelul de educație* (ani), înregistrate pentru un eșantion de 474 de persoane. În vederea verificării ipotezei de normalitate a erorilor, se aplică testul Jarque-Bera, iar rezultatele obținute sunt prezentate în tabelul de mai jos:

Descriptive Statistics									
	N	Minimu	Maximu	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Unstandardized Residual	474	-21567	79043,0	,000000	12819,97	1,764	,112	5,798	,224
Valid N (listwise)	474								

Interpretarea teoretică a coeficientului de asimetrie *Skewness*:

- $sw = 0 \Rightarrow$ distribuție simetrică
- $sw > 0 \Rightarrow$ distribuție asimetrică la dreapta
- $sw < 0 \Rightarrow$ distribuție asimetrică la stânga

Interpretarea teoretică a coeficientului de boltire *Kurtosis*:

- $k = 0 \Rightarrow$ distribuție mezocurtică sau normală
- $k > 0 \Rightarrow$ distribuție leptocurtică
- $k < 0 \Rightarrow$ distribuție platicurtică

Interpretarea coeficienților de asimetrie și boltire pe baza rezultatelor obținute în tabelul de mai sus: distribuția erorilor estimate este asimetrică la dreaptă ($sw > 0$) și leptocurtică ($k > 0$).

2.4. Ipoteza de necorelare sau de independență a erorilor: $cov(\varepsilon_i, \varepsilon_j) = 0$

a. Definirea ipotezei

Ipoteza de necorelare sau de independență a erorilor se referă la lipsa unei corelații între variabilele reziduale sau la faptul că eroarea asociată unei valori a variabilei dependente nu este influențată de eroarea asociată altei valori a variabilei dependente.

În condițiile în care ipoteza de independență a erorilor nu este verificată, modelul de regresie înregistrează o **autocorelare** a erorilor sau o **corelație serială**. Autocorelarea sau corelația serială presupune existența unei autocorelări între erorile ε_i , altfel spus: $cov(\varepsilon_i, \varepsilon_j) \neq 0$ sau $M(\varepsilon_i \cdot \varepsilon_j) \neq 0$.

Autocorelarea erorilor poate fi cauzată de:

- neincluderea în modelul de regresie a uneia sau a mai multor variabile explicative (independente) importante;
- modelul de regresie nu este corect specificat;
- sistematizarea și pregătirea datelor pentru prelucrare;
- inerția fenomenelor în timp și decalajul, în cazul seriilor de timp.

În condițiile încălcării ipotezei, se poate considera că între erori există o relație de forma: $\varepsilon_i = \rho \varepsilon_{i-1} + u_i$, unde u_i reprezintă o variabilă pur aleatoare care respectă ipotezele modelului clasic de regresie.

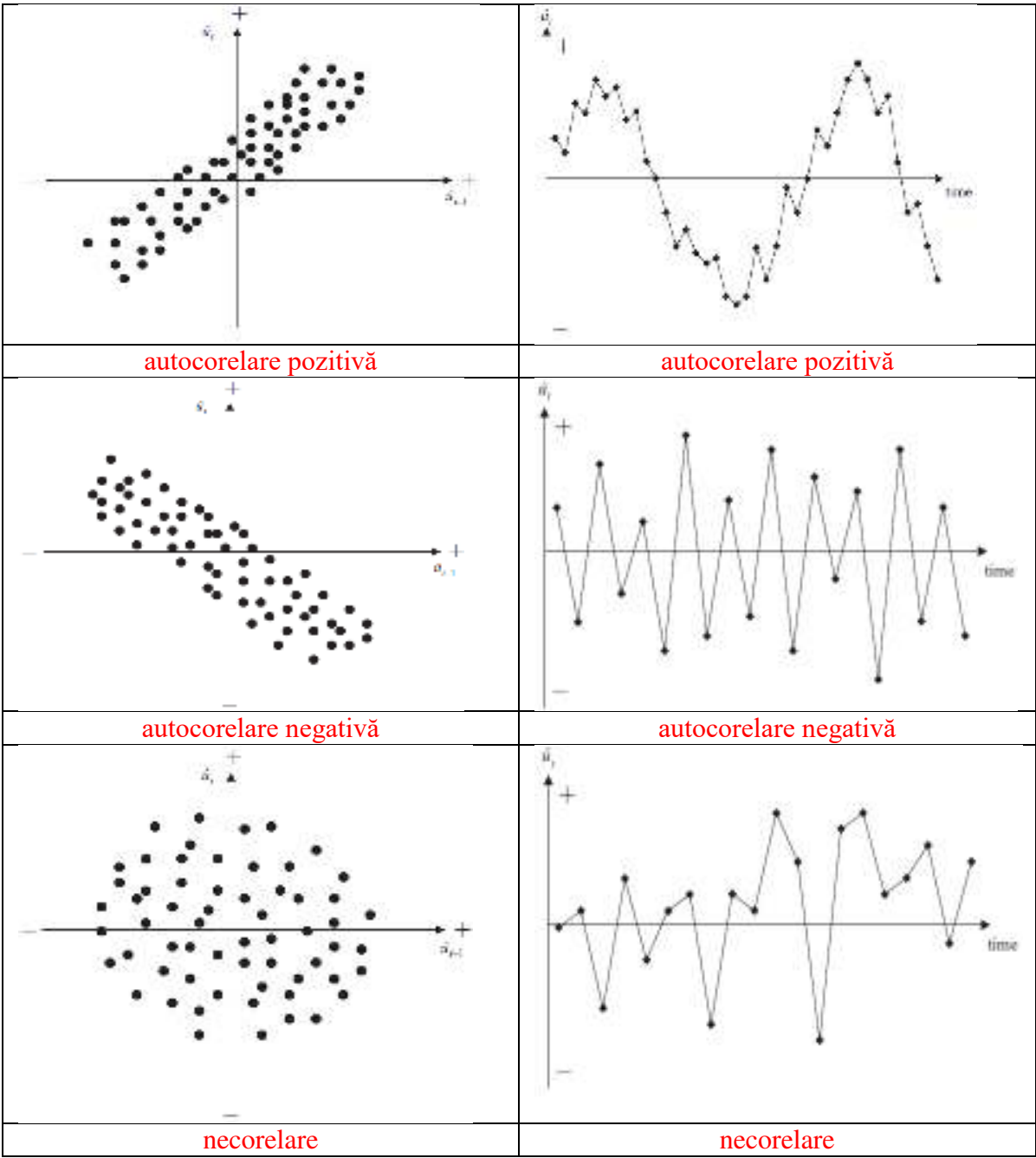
b. Efectele încălcării ipotezei

În condițiile existenței autocorelării erorilor, este afectată calitatea estimațiilor obținute prin metoda celor mai mici pătrate.

Se poate demonstra că, prin aplicarea metodei celor mai mici pătrate, pentru parametrul β_0 , se obține un **estimator neeficient**.

c. Testarea ipotezei

c.1. Procedee grafice



c.2. Testul Durbin-Watson

(1) Formularea ipotezelor:

- $H_0: cov(\varepsilon_i, \varepsilon_j) = 0$ (erorile sunt necorelate sau independente; nu există autocorelare a erorilor)
- $H_1: cov(\varepsilon_i, \varepsilon_j) \neq 0$ (erorile sunt corelate sau dependente; există autocorelare a erorilor)

(1*) Formularea ipotezelor intermediare:

- $H_0: \rho = 0$ (coeficientul de autocorelare a erorilor nu este semnificativ, ceea ce înseamnă că între erori nu există o legătură semnificativă)
- $H_1: \rho \neq 0$ (coeficientul de autocorelare a erorilor este semnificativ, ceea ce înseamnă că între erori există o legătură semnificativă)

(2) Alegerea pragului de semnificație: α

(3) Alegerea statisticii test:

$$DW = d = 2(1 - \hat{\rho}) \sim DW(d_L, d_U)$$

Deoarece $(-1 \leq \hat{\rho} \leq 1)$, valorile DW sunt date de intervalul $(0 \leq d \leq 4)$:

- dacă $\hat{\rho} = 1 \Rightarrow d = 0$, există *autocorelare pozitivă maximă* a erorilor;
- dacă $\hat{\rho} = -1 \Rightarrow d = 4$, există *autocorelare negativă maximă* a erorilor;
- dacă $\hat{\rho} = 0 \Rightarrow d = 2$, *nu există autocorelare* sau erorile nu sunt autocorelate.

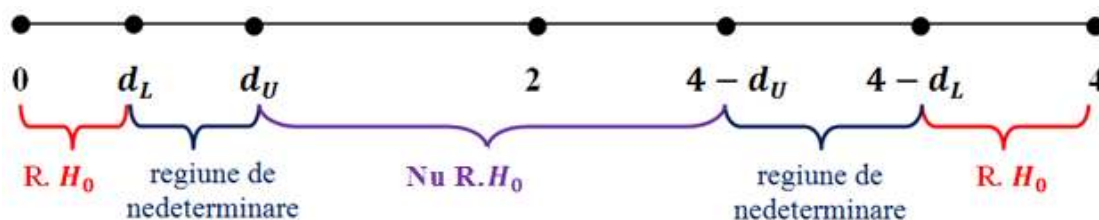
(4) Citirea valorii teoretice a statisticii test:

Se citesc valorile din tabela Durbin-Watson pentru d_L și d_U , ținând cont de numărul de parametrii din model (k), de volumul eșantionului n și de riscul asumat α .

Așadar, pentru $\alpha = 0,05$, $k = 2$ și $n = 474$ (≥ 200), din tabela Durbin-Watson, se citesc valorile: $d_L = 1,758$ și $d_U = 1,779$.

(5) Regula de decizie:

În funcție de aceste valori critice se determină următoarele intervale, care permit luarea deciziei de respingere sau acceptare a ipotezei nule:



- dacă $(0 < d_{calc} < d_L)$, se respinge ipoteza nulă (H_0), erorile înregistrează autocorelare pozitivă;

- dacă $(4 - d_L < d_{calc} < 4)$, se respinge ipoteza nulă (H_0), erorile înregistrează autocorelare negativă;
- $(d_L < d_{calc} < d_U)$ și $(4 - d_U < d_{calc} < 4 - d_L)$ sunt regiuni de nedeterminare, nu se poate decide asupra existenței autocorelării erorilor;
- dacă $(d_U < d_{calc} < 4 - d_U)$, se acceptă ipoteza nulă (H_0), erorile nu sunt autocorelate sau nu există autocorelare a erorilor.

(6) Luarea deciziei:

- dacă se acceptă ipoteza H_0 , coeficientul de autocorelare a erorilor nu este semnificativ, adică nu există autocorelare a erorilor; ipoteza de necorelare sau de independență a erorilor este verificată;
- dacă se respinge ipoteza H_0 , coeficientul de autocorelare a erorilor este semnificativ, adică erorile sunt autocorelate sau dependente; ipoteza de necorelare sau de independență a erorilor nu este verificată.

(7) Interpretarea deciziei:

$$d_{calc} = 1,863 \in (d_U, 4 - d_U)$$

$$d_{calc} = 1,863 \in (1,779; 2,221) \Rightarrow \text{Nu se respinge ipoteza nulă } H_0$$

În condițiile unui risc de 5%, se poate garanta că coeficientul de autocorelare a erorilor nu este semnificativ, adică nu există autocorelare a erorilor. Cu alte cuvinte, ipoteza de necorelare sau de independență a erorilor este verificată.

Observație:

Tabelele pe baza cărora se poate lua decizia cu privire la această ipoteză sunt: *Model Summary* - coloana *Durbin-Watson*.

Aplicație:

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,661 ^a	,436	,435	\$12,833.540	1,863

a. Predictors: (Constant), Educational Level (years)

b. Dependent Variable: Current Salary

c.3. Testul Runs:

- se bazează pe ideea că valorile variabilei reziduale se constituie în secvențe sau seturi de valori pozitive sau negative numite runs (notate cu k), care se succed într-o anumită ordine sau aleator.

- ipoteza de bază a acestui test este aceea că, în cazul lipsei autocorelării erorilor, succesiunea acestor seturi este aleatoare sau numărul acestora este distribuit normal.
- aplicarea testului presupune testarea ipotezei de normalitate a variabilei *numărul de runs (K)*.

(1) Formularea ipotezelor:

- $H_0: cov(\varepsilon_i, \varepsilon_j) = 0$ (erorile sunt necorelate sau independente; nu există autocorelare a erorilor)
- $H_1: cov(\varepsilon_i, \varepsilon_j) \neq 0$ (erorile sunt corelate sau dependente; există autocorelare a erorilor)

(1*) Formularea ipotezelor intermediare:

- H_0 : variabila K este distribuită normal (succesiunea runs-urilor este aleatoare sau numărul lor este distribuit normal)
- H_1 : variabila K nu este distribuită normal (succesiunea runs-urilor nu este aleatoare sau numărul acestora nu este distribuit normal)

(2) Alegerea pragului de semnificație: α

În funcție de statistica Sig (probabilitatea asociată statisticii test):

- dacă $Sig \geq \alpha$ atunci se acceptă ipoteza nulă (H_0), în condițiile unui risc α
- dacă $Sig < \alpha$, atunci se respinge ipoteza nulă (H_0)

(3) Luarea deciziei:

- dacă se acceptă ipoteza H_0 , variabila runs K este distribuită normal, adică nu există autocorelare a erorilor sau erorile nu sunt autocorelate, ci sunt independente, ceea ce înseamnă că este îndeplinită ipoteza de necorelare a erorilor de modelare.
- dacă se respinge ipoteza H_0 , variabila runs K nu este distribuită normal, adică există autocorelare a erorilor sau erorile sunt autocorelate sau dependente, ceea ce înseamnă că nu este îndeplinită ipoteza de necorelare a erorilor de modelare.

(4) Interpretarea deciziei luate:

$$Sig = 0,081 > \alpha = 0,05 \Rightarrow \text{Nu se respinge ipoteza nulă } (H_0)$$

În condițiile unui risc de 5%, se poate garanta că erorile sunt necorelate sau independente. Așadar, ipoteza de necorelare sau de independență a erorilor este verificată.

Observație:

Tabelele pe baza cărora se poate lua decizia cu privire la această ipoteză sunt: *Runs Test*

Aplicație:

Runs Test

	Unstandardized Residual
Test Value ^a	-2502,56250
Cases < Test Value	237
Cases >= Test Value	237
Total Cases	474
Number of Runs	219
Z	-1,747
Asymp. Sig. (2-tailed)	,081

a. Median

3. Ipoteze cu privire la variabilele independente (componenta deterministă)

3.1. Variabilele independente sunt nestocastice sau deterministe

- **definirea ipotezei:** este legată de gradul de omogenitate a variabilelor independente. Deoarece în relațiile varianțelor estimatorilor apare varianța variabilelor independente, este important ca această varianță să fie posibil de calculat, să fie finită și diferită de zero.

3.2. Variabilele independente și eroare sunt necorelate: $cov(X_i, \varepsilon_i) = 0$

- **definirea ipotezei:** este respectată dacă este îndeplinită condiția ca variabilele independente să fie variabile deterministe sau nealeatoare.

3.3. Ipoteza de necoliniaritate a variabilelor independente

a. Definierea coliniarității (multicoliniarității)

Multicoliniaritatea poate fi definită ca o legătură liniară funcțională existentă între două sau mai multe variabile independente ale unui model de regresie de forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

a.1. multicoliniaritate perfectă

- apare atunci când între variabilele independente X_1, X_2, \dots, X_p există o legătură liniară perfectă, funcțională:

$$\lambda_0 + \lambda_1 X_1 + \lambda_2 X_2 + \dots + \lambda_p X_p = 0$$

a.2. multicoliniaritate imperfectă

- poate fi definită ca o relație liniară puternică existentă între două sau mai multe variabile independente:

$$\lambda_0 + \lambda_1 X_1 + \lambda_2 X_2 + \dots + \lambda_p X_p + \varepsilon = 0$$

b. Efectele coliniarității sau efectele încălcării ipotezei de necoliniaritate

Dacă pentru un model de regresie multiplă variabilele independente sunt coliniare, varianța estimatorilor parametrilor modelului de regresie crește, adică estimatorii își pierd proprietatea de eficiență.

b.1. coliniaritate perfectă:

- varianța estimatorilor parametrilor modelului este infinită, ceea ce înseamnă că parametrii pentru aceste variabile nu pot fi estimați.

b.2. coliniaritate imperfectă:

- varianțele pentru estimatorii parametrilor modelului sunt mari.

c. Testarea ipotezei:

c.1. Procedee grafice: Scatter plot

c.2. Depistarea coliniarității

- un prim indiciu pentru existența coliniarității poate fi următorul: dacă între variabilele independente există o legătură de tip liniar, cel mai probabil, coeficientul de determinație pentru acest model va avea o valoare ridicată, însă testul Student pentru fiecare parametru al variabilelor coliniare nu va fi semnificativ statistic. În consecință, se poate testa coliniaritatea prin testarea coeficienților de regresie, iar indiciul este existența unui coeficient de determinație mare. În condițiile în care parametrii modelului de regresie sunt nesemnificativi, se poate decide că modelul admite fenomenul de coliniaritate.
- o altă metodă de testare a coliniarității este testarea parametrilor *modelelor de regresie auxiliară* construite ca modele de regresie liniară doar pe baza variabilelor independente. Dacă parametrii acestor modele sunt semnificativi, atunci variabilele independente sunt coliniare.
- pe baza modelelor de regresie auxiliare, se pot construi doi indicatori cu ajutorul cărora se poate detecta existența coliniarității: *Tolerance* și *VIF (Variance Inflation Factor)*.

c.3. Factorul varianței crescute (VIF)

- relația de calcul pentru $VIF_j = \frac{1}{1-R_j^2}$
- interpretarea lui VIF_j :
 - dacă legăturile dintre variabilele independente sunt puternice, atunci R_j^2 se apropie de 1, iar raportul VIF_j este infinit;
 - dacă între variabilele independente nu există corelație ($R_j^2 = 0$), valoarea raportului VIF_j este egală cu 1.
- în practică, o valoare $VIF_j > 10$ indică prezența coliniarității.
- tabelul în care găsim valorile pentru VIF_j : *Coefficients* - coloana *VIF*

c.4. Factorul toleranță (TOL)

- relația de calcul pentru $TOL_j = \frac{1}{VIF_j}$
- interpretarea lui TOL_j :
 - dacă $TOL_j = 0$, există coliniaritate perfectă;
 - dacă $TOL_j \leq 0,1$, variabilele independente induc fenomenul de coliniaritate;
 - dacă $0,1 < TOL_j \leq 1$, variabilele independente nu sunt coliniare sau nu există coliniaritate.
- tabelul în care găsim valorile pentru TOL_j : *Coefficients* - coloana *TOL*

c.5. Raportul de determinație al modelului auxiliar (R_j^2)

- este calculată pe baza lui $VIF_j \Rightarrow R_j^2 = \frac{VIF_j - 1}{VIF_j}$ sau pe baza lui $TOL_j \Rightarrow R_j^2 = 1 - TOL_j$
- interpretarea lui R_j^2 : arată cât la sută din variația variabilei independente X_j este explicată de variația simultană a celorlalte variabile independente.

Aplicație:

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-20978,3	3087,258		-6,795	,000		
	Educational Level (years)	4020,343	210,650	,679	19,085	,000	,936	1,068
	Previous Experience (months)	12,071	5,810	,074	2,078	,038	,936	1,068

a. Dependent Variable: Current Salary

d. Corectarea coliniarității

Metodele de corecție a coliniarității trebuie să țină cont de tipul de coliniaritate dintre variabile, de numărul de variabile din model și de informațiile suplimentare despre fenomenul studiat. În literatura de specialitate, se întâlnesc mai multe metode de corectare a coliniarității.

Cea mai facilă metodă este eliminarea variabilei care introduce coliniaritatea la nivelul modelului de regresie. În această situație însă, există riscul eliminării din model a unei variabile importante pentru explicarea fenomenului studiat.

O altă metodă este construirea unui model de regresie cu variabile transformate prin diverse funcții sau operatori (de exemplu, prin operatorul decalaj, diferență), iar în acest mod, se poate elimina dependența liniară dintre variabilele factoriale.