PARTEA 2

METODE DE SELECȚIE A VARIABILELOR ÎN MODELE REGRESIONALE

INTRODUCERE IN SUBIECT

Cu toată varietatea existentă de algoritmi **Data Mining**, aproape toți se confruntă cu o dificultate comună– o chestiune de selecție semnificativă a ***caracteristicilor/factorilor/variabilelor*** de intrare pentru modelul cercetat (în literatura de specialitate străină această problemă este cunoscută ca *feature selection*).

Reducerea numărului de ***caracteristici/factori/variabile independente*** are menirea de a reduce/simplifica dimensiunea modelului nu numai pentru a elimina din el toate ***semnificațiile neimportante***, ce nu poartă în sine o anumită "pondere/valoare" utilă pentru analiza informației, dar și, astfel, facilitarea la procesul de simplificare a modelului, inclusiv și prin eliminarea excesului de ***semnificații neimportante***.

Suprapunerea de informații în cadrul excesului de *semnificații* nu doar îmbunătățește calitatea modelului, dar și, uneori, dimpotrivă, agravează comportamentul lui (de exemplu, în cazul multicoliniarității).

Este evident, una dintre căile posibile de ieșire din această problemă ar putea fi construirea unui model pe toate combinațiile posibile de seturi de ***caracteristici/factori/variabile independente*** de intrare cu selecția ulterioară a opțiunii care va avea cea mai bună abilitate descriptivă a caracteristicii **Y** ***rezultate/dependente*** și în același timp conține un ***număr minim de variabile independente***. O astfel de soluție este posibilă, doar numai dacă există un număr mic de seturi de ***caracteristici/factori/variabile independente*** de intrare, candidați pentru includerea în model. În cazul relativ mare a listei de potențiale seturi de ***caracteristici/factori/variabile independente*** de intrare, aplicarea acestei tehnici, este destul de dificilă, deoarece numărul de modele care va fi necesar a fi construite și verificate, va fi extrem de mare și, în general, egal cu **(2n – 1)** variante (așa-numitul "blestem" al dimensiunii). Având în vedere acest lucru, ar urma să apelăm la anumite metode de selecție in raport cu cei mai importanți ***factori/caracteristice/variabile independente***, care ar fi mult mai eficiente.

În acest context, există mai multe metode de a rezolva această problemă. În cazul regresiei acestea sunt:

- **metoda Forward Selection,**
- **metoda Backward Elimination,**
- **metoda Stepwise,**
- **metoda Best Subsets.**

F-TEST PRIVAT

Ne vom opri doar la prima metodă în această lucrare de laborator, avînd ca bază regresia liniara multiplă. Înainte însă de a face acest lucru, vom precăuta succint un citeriu, foarte important pentru procesul de selecție.

Înainte de a începe să facem cunoștință cu prima metodă, ***Forward Selection***, de selecție a variabilelor pentru modelul regresiei liniare multiple, vom precăuta esența unui criteriu nimit **F-test privat,** ce stă de fapt la baza primelor 3 metode.

Criteriul este proiectat pentru a evalua oportunitatea de a introduce mai multe variabile independente în dependența liniară a modelului de regresie multiplă, ecuație care, după cum se știe, are forma:

$$Y=\beta_0+\beta_1X_1+\cdots+\beta_nX_n+\varepsilon,$$

unde

Y –variabila dependentă,
$X_1,X_2,\ldots,X_n$ – variabile independente,
$\beta_0,\beta_1,\ldots,\beta n$ – parametrii modelului,
ε – o mărime aleatoare.

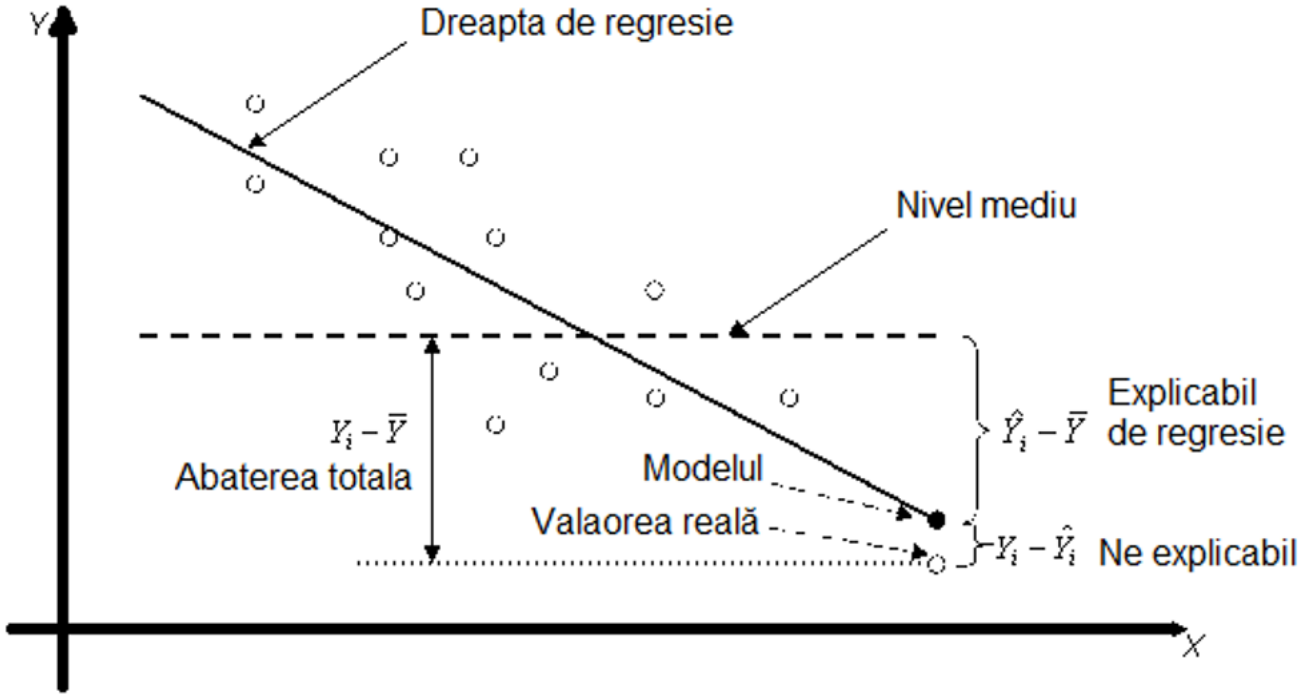
Ideea acestui criteriu se bazează în primul rând pe noțiunea de bază, cum este cea de **suma pătratelor regresiei**, adică definită de mărimea **SSR (***abaterea patratică a valorii pentru **Y**, $\widehat{Y}_\iota$ - obținută prin modelul regresional, în raport cu valaorea medie a observatiilor Y_i, $\bar{Y}$* **):**

$$SSR = \sum_{i=1}^{i=n}(\widehat{Y}_i - \bar{Y})^2,$$

unde

$\widehat{Y}_\iota$ **– valoarea, obținută în baza, modelului regresional,**
$\bar{Y}$ **– valaorea medie a observațiilor Y,**
n – volumul eșantionului.

Acest indicator caracterizează acea cota totală a variației (variabilitate) eficace a factorului rezultant **Y**, pe care am reușit să-l explicăm cu ajutorul regresiei. Vom ilustra acest fapt, pe exemplul celei mai simple regresii, ceo liniară, ce depinde de o singură variabilă /*vezi figura ce urmează*/ În această imagine este explicată divizarea variației generale în 2 componente – ***una explicabilă și altă neexplicabilă***.



Succint comentariu la imagine. Este evident că, dacă regresia nu ar include nici un factor, atunci modelul ar fi emis valori egale $\overline{Y}$.

În consecință, diferența de $Y_i - \overline{Y}$, (i=1,2,…,n) nu ar fi o abatere aleatoare ne explicabilă. Datorită faptului că apare/ se introduce variabila independentă X, evaluările obținute pe baza modelului, tind să fie cât mai aproape de valorile reale ale unei variabile aleatoare **Y**. În acest sens, fiecare mărime/abatere $Y_i - \overline{Y}$ poate fi descompusă în 2 componente:

$\hat{Y}_i - \overline{Y}$ – este explicată de regresie,

$Y_i - \hat{Y}_i$ – nu poate fi explicată de regresie.

Din cele prezentate mai sus derivă că modelul este cu atît mai bun cu cît , diferența $Y_i - \hat{Y}_i$, (i=1,2,…,n). este mai aproape de zero, adică partea ei ce nu poate fi exeplicată, tinde la zero.

Acum, să presupunem că, pe baza variabilelor $X_1, X_2, \ldots, X_k$ a fost construit modelul regresional, pentru care cota de variabilitate (adică de schimbare), explicată de dependența liniară, a constituit valoarea $SSR^{initial.}$.

Să presupunem că dorim să introducem în modelul regresional un nou ***factor/caracteristică/variabilă independentă de intrare*** $\mathbf{X^{extra}}$.

Suma patratelor regresiei **SSR**, construit pe variabilele independente $\mathbf{X_1, X_2, \ldots X_k}$ și $\mathbf{X^{extra}}$, îl vom nota prin $\mathbf{SSR^{full}}$.

Este evident, că

$$\mathbf{SSR^{full} \geq SSR^{initial}}.$$

Evaluăm cu cît a crescut express posibilitatea modelului de a explica procesul cercetat cu ajutorul lui la introducerea unei noi variabile $\mathbf{X^{extra}}$:

$$\mathbf{SSR^{extra} = SSR^{full} - SSR^{initial}}$$

Prin urmare, mărimea $\mathbf{SSR^{extra}}$ poate fi caracterizată drept contribuție a variabilei noi $\mathbf{X^{extra}}$ introduse în model si deci $\mathbf{SSR^{extra}}$ prezintă explicația totală a variabilității eficace a elementului **Y.** În mod evident, cu cît mai mare este această valoare, cu atât mai complex/mare este această contribuție. Natural atunci să ne întrebăm:

"Cum ar urma să fie ales pragul/(valoarea maximă) pentru $\mathbf{SSR^{extra}}$ *pentru a o recunoaște că este destul de mare și, în consecință, a decide cu privire la importanța elementului* $\boldsymbol{X^{extra}}$ pentru model *?"*.

Răspuns la această întrebare îl poate oferi așa-numitul **F-test privat**. De fapt, acest criteriu are menirea de a verifica următoarea ipoteză:

H_0: contribuția $\mathbf{SSR^{extra}}$, cauzată de introducerea variabilei independente $\mathbf{X^{extra}}$ nu este suficient de mare, nu este semnificativă, și deci această variabilă nu ar trebui să fie inclusă în modelul regresional cercetat.

contra alternativă

H_0: contribuția $\mathbf{SSR^{extra}}$ cauzată de introducerea variabilei independente $\mathbf{X^{extra}}$, este una semnificativă, și deci această variabilă ar trebui să fie inclusă în modelul regresional cercetat.

Pentru a verifica aceste ipoteze, trebuie să trecem de la acest indicator $\mathbf{SSR^{extra}}$ la o altă statistică care este determinată de următoarea formulă:

$$\gamma = \frac{\mathbf{SSR^{extra}}}{\mathbf{MSE^{full}}} \qquad \mathbf{(*)}$$

în cazul în care $\mathbf{MSE^{full}}$ reprezintă suma pătratelor erorilor **SSE** (modelul este construit pe variabilile independente $\mathbf{X_1, X_2, \ldots X_k}$ și $\mathbf{X^{extra}}$), atribuită la un singur grad de libertate $\mathbf{df^{sse}}$.

Valoarea $\mathbf{MSE^{full}}$ poate fi determinată prin formula:

$$\mathbf{MSE^{full}} = \frac{SSE}{df^{sse}} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-k-2}$$

unde

$\mathbf{Y_i}$ – valoarea reală a variabilei rezultante,
$\hat{Y}_i$ – evaluarea, obținute pe baza modelului de regresie,
$\mathbf{n}$ – volumul eșantionului,
$\mathbf{k}$ – numărul de variabile a modelului inițial (fără $\mathbf{X^{extra}}$).

Este demonstrat, că statistica, formată în conformitate cu regula de mai sus **(*),** în cazul indeplinirii ipotezei $\mathbf{H_0}$ este redristribuită după legea lui **Fisher (F-distribution, Distributia F).**

Astfel, în acest proces, totul se reduce la verificare ipotezei $\mathbf{H_0}$. Algoritmul de verificare a ipotezei $\mathbf{H_0}$ este după cum urmează:

1. Stabilim nivelul de semnificație $\boldsymbol{\alpha}$, de exemplu **0,01** sau **0,05**.
   Această valoare caracterizează riscul de a lua o decizie greșită.

2. În ***tabelele speciale de distribuție a lui Fisher (fisier atasat)***, determinăm (**0,99** sau **0,95)** $\boldsymbol{\alpha}$-rata procentuală a punctului de distribuție a lui **Fisher,** $\mathbf{K\alpha,}$ cu gradele de libertate **d1=1** și **d2=n−k−2.** Această valoare va fi o valoare de graniță pentru statistică $\boldsymbol{\gamma}$ din **(*)**. */d1=1 fiindca mereu în model se include cite o variabila independentă/*
   Comparăm constată $\mathbf{K\alpha,}$ determinată mai sus, (**0,99** sau **0,95)** $\boldsymbol{\alpha}$-rata procentuală a punctului de distribuție a lui **Fisher,** cu valoarea statistică $\boldsymbol{\gamma}$ din **(*)** ***(pentru cazul concret al modelului construit pe variabilile independente*** $\mathbf{X_1, X_2, \ldots X_k}$ și $\mathbf{X^{extra}}$).
   a. Dacă, $\boldsymbol{\gamma > K\alpha,}$ atunci se trage concluzia că variabila $\mathbf{X^{extra}}$ este una important, eficientă și, în consecință, ***trebuie să fie inclusă*** în modelul regresiei liniare cercetat (*se oferă preferință ipotezei* $\boldsymbol{H_1}$ *, precum că cu probabilitatea* $\boldsymbol{\alpha}$ *am putea face această alegere una gresită*).

b. Dacă $\gamma \leq K\alpha$, atunci se trage concluzia că variabila X^{extra} este una neimportantă, ineficientă și, în consecință, ***nu trebuie să fie inclusă*** în model regresiei liniare cercetat (*adică se oferă preferință ipotezei* H_0 *, ceea ce inseamnă că, cu o probabilitate de* $1-\alpha$ *se acceptă faptul că datele nu contravin experimentului*).

Notă. Întrebarea de verificare a ipotezelor H_0 și H_1 , pe baza **F-criteriului**, mai poate fi precăutată si din alt punct de vedere, și anume, ne definind un anumit nivel de semnificatie α, să găsim probabilitatea ca o variabilă aleatoare γ va lua o valoare mai mare decât valoarea calculată a criteriului.

Deoarece legea de distribuție pentru această statistică este cunoscută, este simplu de a face acest lucru. Valoarea corespunzătoare a aceste probabilități poate fi determinată după cum urmează:

$$\alpha_{real}=1-pF(\gamma,d_1,d_2),$$

unde

$pF(\gamma,d_1,d_2)$ – valoarea funcției de distribuție Fisher în punctual/pentru valaorea γ;
d_1,d_2 – numărul de grade de libertate.

Acum, dacă valoarea identificată α_{real} se va dovedi a fi suficient de mică, atunci ar trebui să se ia decizia cu privire la includerea variabilei X^{extra} în model. În caz contrar, această idee trebuie să se respingă.

Și în final, acum când este clară metodologia de luare a unei decizii privind ***includerea separată a unei variabile independente într-un model*** (sau, dimpotrivă, de inutilitate a acestui act), putem trece la prezentarea unei metode concret de selecție a ***caracteristicilor/factorilor/variabilelot independente*** importante pentru un model regresional de cercetare.

METODA DE SELECȚIE ÎNAINTE (FORWARD SELECTION)

Acest algoritm presupune următoarele etape:

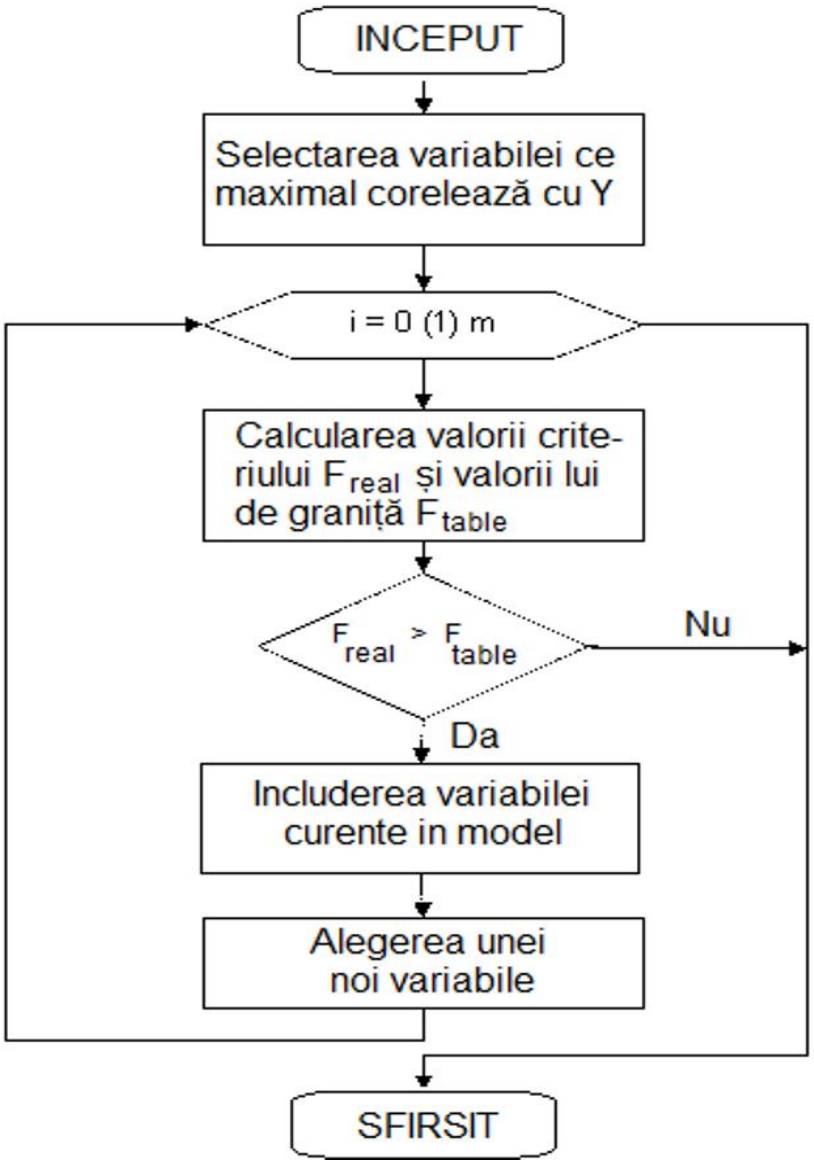
1. Dintr-o listă a ***caracteristicilor/factorilor/variabilelor independente*** posibile de intrare, **este selectată aceea** care are cea mai mare corelație cu **Y,** după aceasta modelul, ***care conține doar o singură variabilă independentă***, este testat la *semnificație/importanță* cu ajutorul ***F-criteriului privat*** stabilit mai sus. Dacă *importanța/semnificația* modelului nu se confirmă, atunci algoritmul se termină aici, pentru lipsa variabilelor de intrare semnificative. În caz contrar, această variabilă este introdusă în model și se trece la următorul punct din algoritm. De remarcat că, în acest caz, verificarea de ansamblu a *semnificației/importanței* modelului, în general, va fi echivalentă cu verificarea *semnificației/importanței* variabilei independente selectate, deoarece la această etapă, modelul nu conține alte variabile de intrare.

2. Asupra tuturor celorlalte variabile rămase, pe baza formulei (*) se calculează valoarea statistică a parametrului γ care este, **raportul *creșterii sumei pătratelor de regresie*** $\mathbf{SSR^{extra}}$, realizată prin introducerea în modelul corespunzător a unei variabile externe/adiționale (în comparație cu valoarea $\mathbf{SSR^{initial}}$ calculate doar pe baza variabilelor deja introduse), ***la mărimea*** $\mathbf{MSE^{full}}$**.**

3. Dintre toate variabilele-candidate pentru a fi incluse în model, este aleasă aceea, care ***are cea mai mare valoare de criteriu*** γ calculat la etapa a 2-a.

4. Se efectuează verificarea *semnificației/importanței* variabilei independente alese la etapa a 3-ia. Dacă *semnificația/importanța* acesteia este confirmată, atunci ea se include în model și se trece la pasul 2 (dar deja cu o nouă variabilă independentă în cadrul modelului). În caz contrar, algoritmul se oprește.

Procedura de selecție variabilei independente prin ***metoda Forward Selection*** se poate prezenta în formă de schemă bloc după cum urmează.



Schema Bloc a metodei Forward Selection

UN EXEMPLU DE UTILIZARE A METODEI FORWARD SELECTION / DE SELECȚIE INAINTE

Vom analiza metoda descrisă mai sus, de selecție a variabilelor în modelul de regresie liniară multiplă printr-un exemplu concret. În calitate de ***caracteristici/factori/variabileindependente*** vom precăuta următorii indicatori din sectorul bancar, prezentati in **Tabelul 1**.

Tabelul 1 – Variabilele utilizate pentru dezvoltarea unui model de regresie liniară multiplă

Factorul	Notare	Tipul
Numărul de întîrzieri cu plățile la bancă	**Y**	Variabila dependentă
Vechimea stajiului de lucru la ultimul loc de muncă	**X_1**	**Variabila independentă**
Termenul de împrumut	**X_2**	**Variabila independentă**
Suma împrumutului	**X_3**	**Variabilă independentă**

Tabelul 2 conține informații despre 10 debitori pentru fiecare dintre variabilele declarate.

Debitor - persoană (fizică sau juridică) care are datorie în bani sau în alte bunuri (Dex)

№ Debitor	Întîrzieri, Nr. (Y)	Stagiu, ani (X_1)	Termenul împr., luni. (X_2)	Suma imprum., lei. (X_3)
1	0	7,5	12	170 000
2	0	4,5	12	120 000
3	0	6,5	12	85 000
4	1	2,5	12	160 000
5	1	3,5	24	105 000
6	0	6,5	12	90 000
7	3	2	24	80 000
8	2	3,5	24	395 000
9	2	6	36	150 000
10	4	2	60	70 000

METODA DE SELECȚIE DIRECT (FORWARD SELECTION) A VARIABILELOR PEBTRU MODELUL DE REGRESIE LINIARĂ MULTIPLĂ

Vom precăuta metoda de mai sus pornind de la datele prezentate în Tabelul 2, pentru care formulăm **SARCINĂ utilizînd regresia liniară multiplă din DM, să identificăm un model prin intemediul căruia sa putem prognoza Y in dependență de factorii precăutați X_1,X_2,X_3.**

Conform recomandărilor de mai sus a metodei precăutate, la primul pas al algoritmului umează să includem în model unul dintre factorii care este cel mai puternic corelat cu rezultatul.

Tabelul 3 – Corelația variabilelor de intrare

	Întîrzieri, Nr. (Y)	*Stagiu, ani (X1)*	*Termenul împr., luni. (X2)*	*Suma imprum., lei. (X3)*
Stagiu, ani (X1)	**-0,721369**	1		
Termenul împr., luni. (X2)	**0,8707703**	**-0,466074**	1	
Suma imprum., lei. (X3)	**0,0184673**	**-0,02203**	**-0,11082**	1

Obținut cu Pachetul Data Analysis

	Y
X1	-0,721
X2	0,871
X3	0,018

După cum putem vedea din Tabelul 3, X_2 se caracterizează prin cea mai mare forță de corelare liniară cu rezultatul Y. De aceea această variabilă ar trebui să fie prima supusă procedurii de verificare la includerea în model. Să precăutăm evaluarea valorilor setului de indicatori statistici obținuți pe baza modelului, prin care putem decide includerea în model doar a unei singure variabile independente X_2. Rezultatele corespunzătoare sunt rezumate în Tabelul 4.

Tabelul 4 – Datele inițiale și datele de calcul, necesare pentru verificarea semnificației variabilei X_2

nr, d/o	X_{2i}	Y_i	$\hat{Y}_i$	$\overline{Y}$	$(\hat{Y}_i-\overline{Y})^2$	$(Y_i-\hat{Y}_i)^2$
1	12	0	0,436	1,3	0,476	0,19
2	12	0	0,436	1,3	0,476	0,19
3	12	0	0,436	1,3	0,476	0,19
4	12	1	0,436	1,3	0,476	0,318
5	24	1	1,396	1,3	0,009	0,157
6	12	0	0,436	1,3	0,476	0,19
7	24	3	1,396	1,3	0,009	2,573
8	24	2	1,396	1,3	0,009	0,365
9	36	2	2,356	1,3	1,115	0,127
10	60	4	4,275	1,3	8,852	0,076
Suma	-	-	-	-	**13,724**	**4,376**

Acum, folosind datele din Tabelul 4, vom calcula valoarea ***F-criteriului privat***, corespunzător X_2.

$$\mathbf{MSE^{full}=\frac{SSE}{df^{sse}}=\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n-k-2}=\frac{4,376}{8}=0,547}$$

$$\mathbf{SSR^{extra}=SSR^{full}-SSR^{initial}}$$

$$\mathbf{Freal=\frac{SSR^{extra}}{MSE^{full}}=\frac{SSR^{full}-SSR^{initial}}{MSE^{full}}=\frac{SSR^{full}-0}{MSE^{full}}=\frac{13,724}{0,547}=25.09}$$

În tabelul de valori a **F-criteriului Fisher** la nivel de semnificatie **α=0,05** găsim valoarea limită pentru **Freal** (în cazul în care numărul de grade de libertate $\mathbf{d_1=1}$ și $\mathbf{d_2=8}$).

Informațiile necesare pentru obținerea informației cu privire la semnificația/importanța variabilei $\mathbf{X_2}$ pentru modelul de regresie liniaară multiplă cercetat (sau, dimpotrivă, de inutilitate) pot fi găsite în **Tabelul 5.**

Rezultatele analizei regresiei liniare pentru cazul în care in model este inclusă doar variabila X2 este prezentată mai jos, rezultate obținute cu ajutorul Pachetului Data Analysis

SUMMARY OUTPUT

Regression Statistics								
Multiple R	0,87077							
R Square	0,758241							
Adjusted R Square	0,728021							
Standard Error	0,739581							
Observations	10							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	**1**	**13,72416**	**13,72416**	25,09079755	0,001041			
Residual	**8**	**4,375839**	**0,54698**					
Total	**9**	**18,1**						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	**-0,52349**	0,432691	-1,20985	0,260871095	-1,52128	0,474297	-1,52128	0,474297
Termenul împr., luni. (X2)	**0,079978**	0,015967	5,009072	0,001040968	0,043159	0,116797	0,043159	0,116797
RESIDUAL OUTPUT				PROBABILITY OUTPUT				
Observation	*Predicted Întîrzieri, Nr. (Y)*	*Residuals*	*Standard Residuals*		*Percentile*	*Întîrzieri, Nr. (Y)*		
1	0,436242	-0,43624	-0,62563		5	0		
2	0,436242	-0,43624	-0,62563		15	0		
3	0,436242	-0,43624	-0,62563		25	0		
4	0,436242	0,563758	0,808507		35	0		
5	1,395973	-0,39597	-0,56788		45	1		
6	0,436242	-0,43624	-0,62563		55	1		
7	1,395973	1,604027	2,300394		65	2		
8	1,395973	0,604027	0,866257		75	2		
9	2,355705	-0,3557	-0,51013		85	3		
10	4,275168	-0,27517	-0,39463		95	4		

Tabelul 5 – Analiza variației pentru verificarea semnificației/importanței factorului X_2

Sursa	Numărul de grade de libertate (df)	Suma pătratelor (SS)	SS la un singur grad de libertate (MS)	F_{real}	F_{table}
Regresia liniară condiționată de introducerea în model a variabilei X_2	1	13,724	13,724	**25,091 >**	**5,318**
Erori	**8**	**4,376**	**0,547**		

După cum putem vedea din **Tabelul 5**, valoarea calculată **F-criteriu,** semnificativ este mai mare decât valoarea de prag din **tabelul F-criteriului Fisher Ftable**, ceea ce indică necesitatea de a include variabila X_2 în modelul regresional liniar multiplu (***Interesant!*** în acest caz, probabilitatea ca decizia de includere se va dovedi a fi greșită, este de **α=0,05, adică doar de 5%).**

La această etapă a metodei date, listă variabilelor potențiale ale modelului s-a redus cu **1**, datorită transferului variabilei X_2 ("Termenul de împrumut") în categoria variabile importante.

Prin urmare, rămâne de rezolvat problema de includere pentru X_1 ("Vechimea stajului de lucru la ultimul loc de muncă ") și X_3 ("Suma împrumutului").

Rezultatele analizei regresiei liniare pentru cazul în care in model este inclusă variabila X_2 și variabila X_1 este prezentată mai jos, rezultate obținute cu ajutorul Pachetului Data Analysis sunt prezentate pe pagina următoare

Rezultatele analizei regresiei liniare pentru cazul în care in model este inclusă variabila X_2 și variabila X_3 este prezentată mai jos, rezultate obținute cu ajutorul Pachetului Data Analysis sunt prezentate după pagina următoare

SUMMARY OUTPUT							

Regression Statistics								
Multiple R	**0,94097**							
R Square	**0,885425**							
Adjusted R Square	**0,852689**							
Standard Error	**0,544296**							
Observations	**10**							
ANOVA								
	df	***SS***	***MS***	***F***	***Significance F***			
Regression	**2**	**16,02619**	**8,013097**	**27,04769**	**0,000509113**			
Residual	**7**	**2,073807**	**0,296258**					
Total	**9**	**18,1**						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	**1,112386**	0,667683	1,666039	0,139647	-0,46643402	2,691207	-0,46643	2,691207
Stagiu, ani (X1)	**-0,2792**	0,100162	-2,78754	0,027004	-0,51604946	-0,04236	-0,51605	-0,04236
Termenul împr., luni. (X2)	**0,062723**	0,013281	4,722603	0,002151	0,031317139	0,094128	0,031317	0,094128
RESIDUAL OUTPUT					PROBABILITY OUTPUT			
Observation	*Predicted Întîrzieri, Nr. (Y)*	*Residuals*	*Standard Residuals*		*Percentile*	*Întîrzieri, Nr. (Y)*		
1	-0,22898	0,228977	0,477012		5	0		
2	0,608636	-0,60864	-1,26793		15	0		
3	0,050227	-0,05023	-0,10463		25	0		
4	1,167045	-0,16705	-0,34799		35	0		
5	1,640511	-0,64051	-1,33433		45	1		
6	0,050227	-0,05023	-0,10463		55	1		
7	2,059318	0,940682	1,959656		65	2		
8	1,640511	0,359489	0,748897		75	2		
9	1,69517	0,30483	0,63503		85	3		
10	4,31733	-0,31733	-0,66107		95	4		
SUMMARY OUTPUT								

Regression Statistics								
Multiple R	**0,87842**							
R Square	**0,771622**							
Adjusted R Square	**0,706371**							
Standard Error	**0,768454**							
Observations	**10**							
ANOVA								
	df	***SS***	***MS***	***F***	***Significance F***			
Regression	**2**	**13,96635**	**6,983176**	**11,82545**	**0,005692356**			
Residual	**7**	**4,133648**	**0,590521**					
Total	**9**	**18,1**						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	**-0,7969**	0,61999	-1,28534	0,239562	-2,26294246	0,669146	-2,26294	0,669146
Termenul împr., luni. (X2)	**0,081162**	0,016693	4,862144	0,001831	0,041690331	0,120634	0,04169	0,120634
Suma imprum., lei. (X3)	**1,73E-06**	2,7E-06	0,640415	0,54229	-4,6553E-06	8,11E-06	-4,7E-06	8,11E-06
RESIDUAL OUTPUT					PROBABILITY OUTPUT			
Observation	*Predicted Întîrzieri, Nr. (Y)*	*Residuals*	*Standard Residuals*		*Percentile*	*Întîrzieri, Nr. (Y)*		
1	0,470998	-0,471	-0,69498		5	0		
2	0,384542	-0,38454	-0,56741		15	0		
3	0,324024	-0,32402	-0,47811		25	0		
4	0,453707	0,546293	0,806084		35	0		
5	1,332553	-0,33255	-0,4907		45	1		
6	0,332669	-0,33267	-0,49087		55	1		
7	1,289325	1,710675	2,524189		65	2		
8	1,833995	0,166005	0,244949		75	2		
9	2,38431	-0,38431	-0,56707		85	3		
10	4,193876	-0,19388	-0,28607		95	4		

Vom determina creșterea sumei pătratelor regresiei, a SSR^{extra} ,care poate fi observată atunci când succesiv includem în model variabilile X_1 și X_3 (*cu condiția ca variabila X_2 deja intră în compoziția modelului de regresie liniară*). Rezultatele calculului sunt prezentate în **Tabelul 6**.

Tabelul 6 – Calculul creșterii SSR prin includerea in model avariabilelor de intrare X_1 și X_3

Variabila analizată	SSR^{full}	$SSR^{initial}$	SSR^{extra}
X_1	16,026	13,724	2,302
X_3	13,966	13,724	0,242

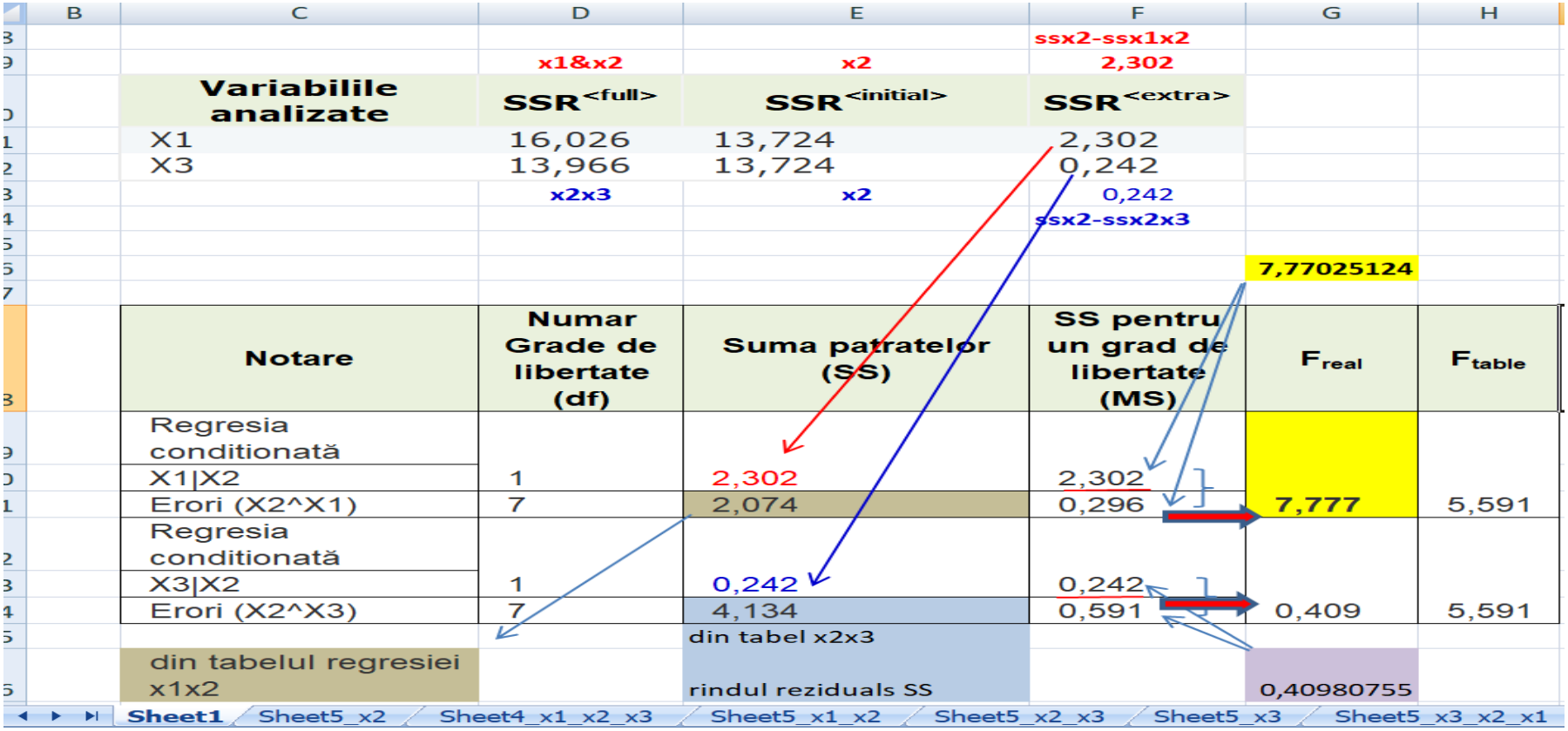
În tabelul de valori a **F-criteriului Fisher** la nivel de semnificatie **α=0,05** găsim valoarea limită pentru **Freal** (în cazul în care numărul de grade de libertate **d_1=1** și **d_2=7**)., pentru X_1 și X_3. Datele relevante sunt prezentate în **Tabelul 7**.

Tabelul 7 – Calculul, luind in considerare F-criteriu privat, pentru X_1 și X_3 atunci când sunt introduce succesiv în modelul regesional în care se află deja variabilei X_2.

Sursa	Numărul de grade de libertate (df)	Suma pătratelor (SS)	SS la un singur grad de libertate (MS)	F_{real}	F_{table}
Regresia liniară condiționată de întroducerea $X_1\|X_2$ – este in model	1	2,302	2,302	7,777 >	5,591
De Eroare(X_2^X_1)	7	2,074	0,296		
Regresia liniară condiționată de întroducerea $X_3\|X_2$ – este in model	1	0,242	0,242	0,409 <	5,591 ??
De Eroare(X_2^X_3)	7	4,134	0,591		

În **Tabelul 7** de prezentare a rezultatelor, **"$X_1|X_2$"** înseamnă că calculul este efectuat pentru variabila X_1 fără a ține cont de acea cota de variație, care a fost deja explicată anterior cînd a fost întrodusă în modelul regresional liniar variabila X_2.

Algoritmul de calcul pentru Tabelul 7



	Variabilile analizate	$SSR^{<full>}$	$SSR^{<initial>}$	$SSR^{<extra>}$
		x1&x2	x2	ssx2-ssx1x2
				2,302
	X1	16,026	13,724	2,302
	X3	13,966	13,724	0,242
		x2x3	x2	0,242
				ssx2-ssx2x3

7,77025124

Notare	Numar Grade de libertate (df)	Suma patratelor (SS)	SS pentru un grad de libertate (MS)	F_{real}	F_{table}
Regresia conditionată					
X1|X2	1	2,302	2,302		
Erori (X2^X1)	7	2,074	0,296	7,777	5,591
Regresia conditionată					
X3|X2	1	0,242	0,242		
Erori (X2^X3)	7	4,134	0,591	0,409	5,591
		din tabel x2x3			
din tabelul regresiei x1x2		rindul reziduals SS		0,40980755	

Sheet1 | Sheet5_x2 | Sheet4_x1_x2_x3 | Sheet5_x1_x2 | Sheet5_x2_x3 | Sheet5_x3 | Sheet5_x3_x2_x1

După cum putem vedea din **Tabelul 7**, cea mai mare valoare a ***F-criteriului privat***, corespunde variabilei $\mathbf{X_1}$. De aceea această variabilă trebuie să fie verificată în primul rând la posibilitatea includerii ei în modelul regresional liniar multiplu.

Ca și anterior, un semn de semnificație pentru variabila dată, este faptul depășirii **Freal** calculat pentru valorile lui $\mathbf{X_1}$ a **Ftable** din tabelul de valori a **F-criteriului Fisher** la nivel de semnificatie $\boldsymbol{\alpha}$**=0,05**. În cazul nostru, evident, se realizează inegalitatea:

Freal(=7,777)>Ftable(=5,591)

În acest sens, ipoteza zero cu privire la lipsa de semnificație pentru această variabilă **X_1**, ar trebui să fie respinsă, ca un ace contravene datelor experimentale (în acest caz, probabilitatea de eroare este de **α=0,05**). Prin urmare, putem concluziona:
variabila **X_1** *("*Vechimea stajului de lucru la ultimul loc de muncă *") este esențială pentru modelul regresional liniar multiplu, și trebuie să fie inclusă în modelul regresional impreună cu* ***X_2***.

Printre candidați pentru includerea în model a rămas doar o singură variabilă **X_3.** Toate celelalte au trecut cu succes testul de semnificație.

Supunem analizei variabila **X_3** similar cazurilor de mai sus, utilizînd testul, **F-criteriul privat**, în presupunerea că, variabilile **X_1** și X_2 au intrat deja în model. Rezultatele de calcul pot fi găsite în **Tabelul 8.**

Rezultatele analizei regresiei liniare pentru cazul în care in model sunt incluse déjà variabilile X_1 și X_2, și dorim să cunoastem dacă putem include cu semnificație si variabila X_3 este prezentată mai jos, rezultate obținute cu ajutorul Pachetului Data Analysis sunt prezentate pe pagina următoare

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,944901
R Square	0,892837
Adjusted R Square	0,839256
Standard Error	0,568572
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	16,16036	5,386786	16,66323	0,002581408
Residual	6	1,939643	0,323274		
Total	9	18,1			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	**0,874967**	0,788844	1,109176	0,309822	-1,05526568	2,805199	-1,05527	2,805199
Suma imprum., lei. (X3)	**1,29E-06**	2E-06	0,644217	0,543262	-3,6139E-06	6,2E-06	-3,6E-06	6,2E-06
Stagiu, ani (X1)	**-0,27354**	0,104998	-2,60515	0,040382	-0,53045816	-0,01662	-0,53046	-0,01662
Termenul împr., luni. (X2)	**0,063958**	0,014006	4,566594	0,003824	0,029687308	0,098228	0,029687	0,098228

RESIDUAL OUTPUT

Observation	*Predicted Întîrzieri, Nr. (Y)*	*Residuals*	*Standard Residuals*
1	-0,18951	0,189513	0,408226
2	0,566522	-0,56652	-1,22033
3	-0,02575	0,025753	0,055475
4	1,165255	-0,16525	-0,35597
5	1,588178	-0,58818	-1,26698
6	-0,0193	0,019296	0,041565
7	1,966196	1,033804	2,226889
8	1,96271	0,03729	0,080326
9	1,729945	0,270055	0,581718
10	4,255756	-0,25576	-0,55092

PROBABILITY OUTPUT

Percentile	*Întîrzieri, Nr. (Y)*
5	0
15	0
25	0
35	0
45	1
55	1
65	2
75	2
85	3
95	4

Tabelul 8 – Calculul, luind in considerare F-criteriu privat, pentru variabila X_3 când în model sunt déjà introduse X_2 și X_1

Sursa	Numărul de grade de libertate (df)	Suma pătratelor (SS)	SS la un singur grad de libertate (MS)	F_{real}	F_{table}
Regresia liniară condiționată de întroducerea $X_3\|(X_1\wedge X_2)$ – este in model	1	0,134	0,134	0,415 < ?	5,987
De Eroare ($X_1\wedge X_2\wedge X_3$)	6	1,94	0,323		

Algoritmul de calcul pentru Tabelul 8

Notare	Numar Grade de libertate (df)	Suma patratelor (SS)	SS pentru un grad de libertate (MS)	F_{real}	F_{table}
Regresia conditionată				0,415	5,987
X3|(X1^X2)	1	0,134	0,134		
Erori (X1^X2^X3)	6	1,94	~~0,323~~		
		din tabel x3x2x1 rindul reziduals SS			

Ca și anterior, caracterul semnificativ al contribuției variabilei X_3 în capacitatea modelului de regresie este determinată de mărimea **Freal.** Această valoare pentru variabila X_3 nu a reușit să depășească nivelul dorit din **Ftable,** și de aceea variabila X_3= "Suma împrumutului" nu se consideră una semnificativă și deci nu trebuie de inclus în modelul regresiei liniare multiple. Altfel spus, ar trebui să fie acceptată ipoteza H_0 care cere a nu include variabila respectivă în modelul regresional cu probabilitatea $1-\alpha$ corectitudinea acestei decizii.

Prin urmare, în tabelul următor sunt prezentate variabilile incluse în modelul regresional liniar multiplu Tabelul 9.

Tabelul 9 –Selecția variabilelor în modelul regresional liniar multiplu efectuat cu procedura Forward Selection

Factorul	Notare	Rezultatul verificării
Vechimea stajului de lucru la ultimul loc de muncă	**X_1**	**Include în model**
Termenul de împrumut	**X_2**	**Include în model**
Suma împrumutului	X_3	Nu se includ în model

CONCLUZIE

Metoda de selecție a variabilelor nu se limitează doar la modelul liniar multiplu de regresie, dar, de asemenea, poate avea o gamă mai largă de aplicare, în special, pentru a o utilizate în regresie logistică. În acest caz este necesar doar de a alege un alt criteriu de verificare variabilă în funcție de relevanță, decât **F-test privat** descris mai sus. O alternativă în acest caz, pote deveni *testul de multiplicatori Lagrange* sau *Testul raportului de probabilitate/*