Programarea calculatoarelor şi limbaje de programare I

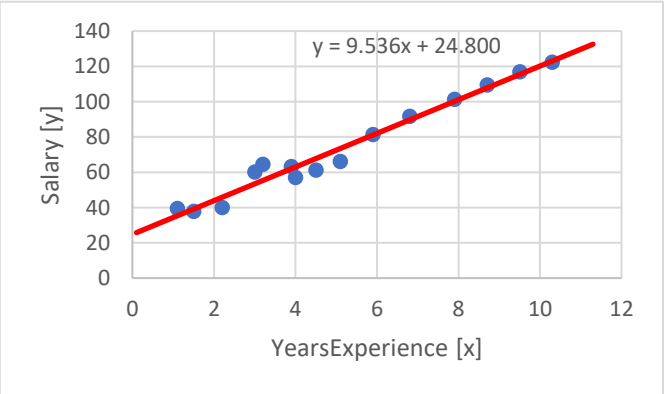
Tema 2

Termen de predare: săptămâna 11-15 ianuarie 2021
Punctaj: 10% din nota finală

Regresia este o metodă din domeniul statisticii care estimează relația dintre o variabilă dependentă (denumită și explicație, efect, rezultat) Y și una sau mai multe variabile independente (denumite și factori, predictori, atribute) $X_1, X_2, \dots, X_k$. Dacă valorile observate pentru aceste variabile sunt $\left(y_i, (x_{1i}, x_{2i}, \dots, x_{ki})\right), i = 1, \dots, n$, atunci funcția de regresie este acea funcție $Y = f(X_1, X_2, \dots, X_n)$ care aproximează cel mai bine setul de date observate. Dacă funcția f este liniară, atunci obținem o ***regresie liniară***. Dacă folosim o singură variabilă independentă X cu valorile observate $x_i, i = 1, \dots, n$, atunci reprezentarea grafică a funcției de regresie liniară este o dreaptă $\bar{y} = w_1 x + w_2$ care se mai numește și ***dreaptă de regresie***.

În exemplul de mai jos avem un set de date numerice care reprezintă salariul mediu exprimat în mii de USD raportat la numărul de ani de experiență. Variabila independentă este `YearsExperience`, iar variabila dependentă este `Salary`. Parametrii $w_1 = 9.536$ și $w_2 = 24.8$ ai dreptei de regresie $\bar{y} = 9.536x + 24.8$ minimizează eroarea cumulată pătratică E dintre valoarea y_i a salariului și valoarea estimată $\bar{y}_i$: $E = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2$.

YearsExperience [x]		Salary [y]	
x_1	1.1	y_1	39.343
x_2	1.5	y_2	37.731
x_3	2.2	y_3	39.891
x_4	3	y_4	60.15
x_5	3.2	y_5	64.445
x_6	3.9	y_6	63.218
x_7	4	y_7	56.957
x_8	4.5	y_8	61.111
x_9	5.1	y_9	66.029
x_{10}	5.9	y_{10}	81.363
x_{11}	6.8	y_{11}	91.738
x_{12}	7.9	y_{12}	101.302
x_{13}	8.7	y_{13}	109.431
x_{14}	9.5	y_{14}	116.969
x_{15}	10.3	y_{15}	122.391



Scrieți o aplicație care calculează parametrii w_1 și w_2 ai dreptei de regresie astfel:

$$w_2 = \frac{\sum_{i=1}^{n}(x_i - x_{med})\,(y_i - y_{med})}{\sum_{i=1}^{n}(x_i - x_{med})^2}$$

$$w_1 = y_{med} - w_2 x_{med}$$

unde x_{med} și y_{med} sunt mediile valorilor de pe coloanele `YearsExperience`, respectiv `Salary`.

Datele vor fi citite din fișierul Salary_Data.csv, iar rezultatele vor fi afișate pe ecran.