CURS: ANALIZA ȘI VIZUALIZAREA DATELOR

# Regresia liniară simplă și multiplă

*Realizat de st.gr. SD-231, Dediuc Anastasia*

# Cuprins



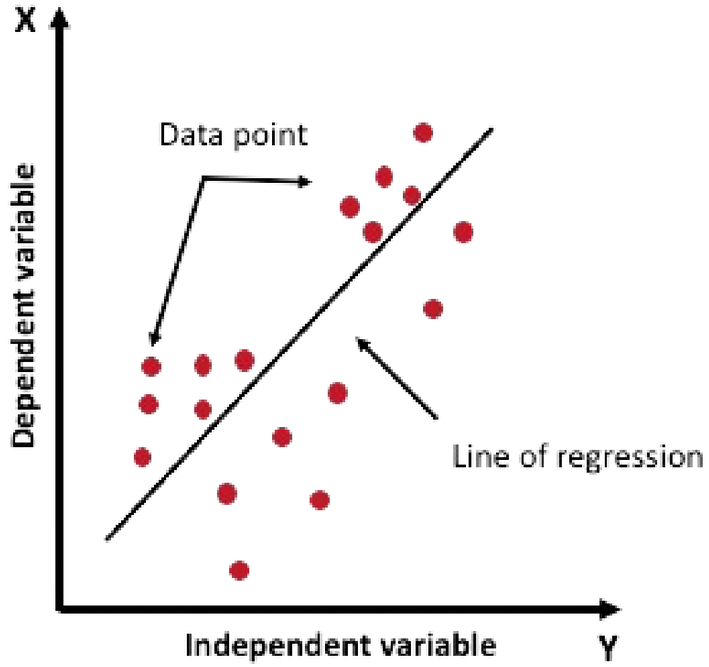
# Introducere

**Regresia liniară** - o tehnică statistică de modelare folosită pentru a descrie relația dintre o variabilă dependentă (țintă) și una sau mai multe variabile independente (predictori), asumând că această relație este aproximativ liniară. În esență, modelul de regresie liniară estimează o linie (sau hiperplan) care aproximează cel mai bine legătura dintre variabile, permițându-ne să prezicem valoarea variabilei dependente pe baza valorilor cunoscute ale variabilelor independente.


**Aplicații practice:**

- *Economie și finanțe:* estimarea vânzărilor în funcție de bugetul de publicitate, previzionarea prețului unei case pe baza suprafeței și a numărului de camere etc.
- *Științe sociale:* evaluarea relației dintre nivelul de educație și venit, corelarea scorurilor la teste cu orele de studiu.
- *Inginerie și științe naturale:* anticiparea consumului de combustibil al unei mașini în funcție de greutate, putere și alți parametri, sau estimarea creșterii unei plante pe baza cantității de îngrășământ și lumină primită.

Aceste exemple ilustrează că regresia liniară are un rol dublu: (1) descriptiv, pentru a înțelege și cuantifica relația dintre variabile (de ex., cât de mult scade consumul odată cu creșterea greutății mașinii?), și (2) predictiv, pentru a prezice rezultate viitoare pe baza modelelor obținute. Prin urmare, stă la baza multor analize statistice și machine learning, fiind un bun punct de plecare în modelarea datelor.



# Regresia liniară simplă

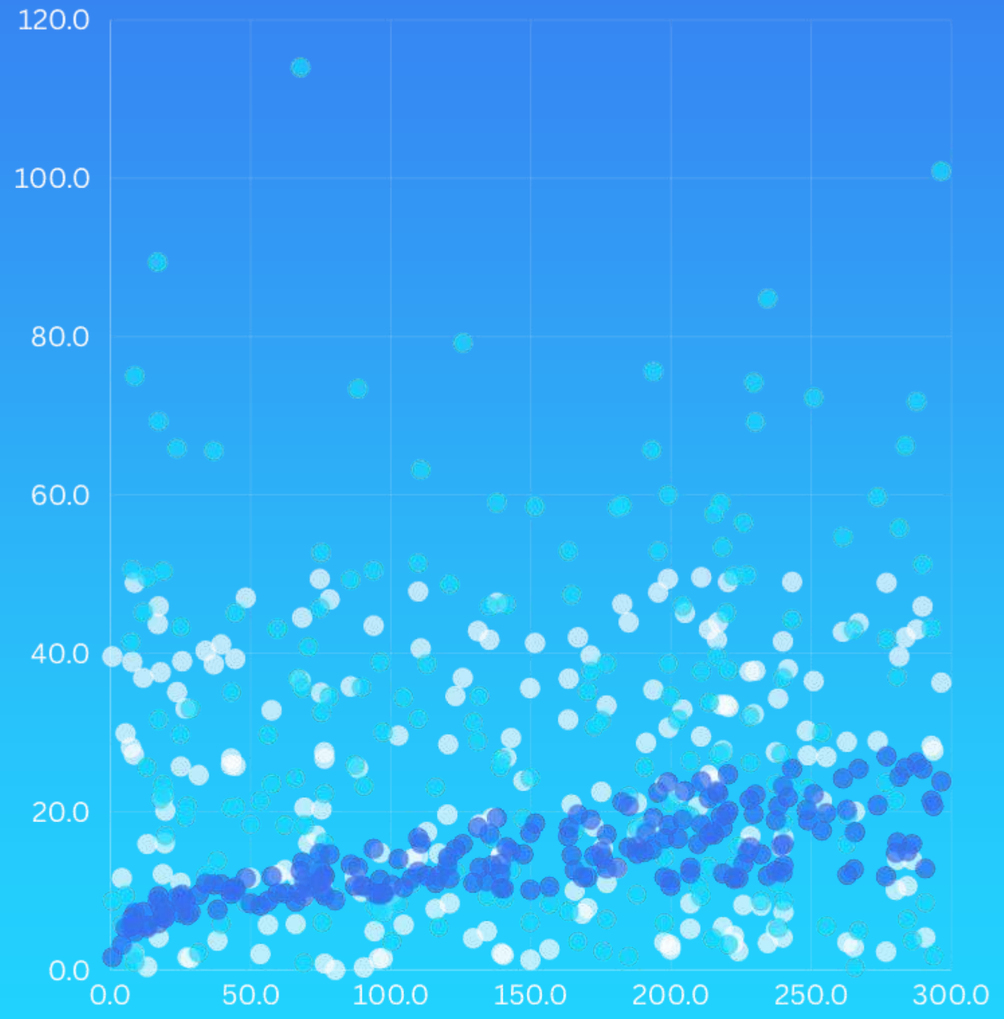
**Regresia liniară simplă** este o metodă statistică folosită pentru a analiza relația dintre două variabile, una independentă (predictor) și una dependentă (răspuns). Scopul este de a determina în ce măsură schimbările din variabila independentă influențează variabila dependentă și de a putea face predicții bazate pe această relație.

✅ *Când folosim regresia liniară simplă?*

- Când avem o singură variabilă care influențează rezultatul (ex: doar TV).
- Când vrem să înțelegem impactul unui factor asupra unei valori măsurabile.
- Când vrem să facem predicții bazate pe această relație.

⚠️ *Limitări ale regresiei liniare simple:*

- Nu ia în calcul alți factori care pot influența vânzările (ex: reclame pe radio, ziare, promoții).
- Relația trebuie să fie aproximativ liniară – dacă datele nu urmează o tendință dreaptă, modelul poate fi inexact.
- Nu explică cauzalitatea – doar pentru că două variabile sunt corelate, nu înseamnă că una o cauzează pe cealaltă.



# Regresia liniară multiplă

**Regresia liniară multiplă** este o extensie a regresiei liniare simple, utilizată pentru a analiza relația dintre o variabilă dependentă și două sau mai multe variabile independente.

*Când folosim regresia liniară multiplă?*
✓ Când avem mai mulți factori care influențează un rezultat.
✓ Când vrem să optimizăm investițiile – ex: ce canal publicitar aduce cele mai mari vânzări?
✓ Când vrem predicții mai precise comparativ cu regresia simplă.

*Limitări ale regresiei liniare multiple:*
◈ Multicolinearitate – Dacă predictorii sunt corelați între ei (ex: TV și Radio), modelul poate deveni instabil.
◈ Presupune o relație liniară – Dacă relația dintre variabile nu este liniară, modelul nu va funcționa bine.
◈ Poate include factori irelevanți – Dacă adăugăm prea multe variabile, modelul poate deveni mai puțin precis.

# Prezentarea datasetului

Setul de date *„Advertisement & Sales Data For Analysis”* conține 200 de observații despre impactul publicității pe TV, Radio și Newspaper asupra vânzărilor.

| ID | TV | Radio | Newspaper | Sales |
|---|---|---|---|---|
| Identificator unic pentru fiecare înregistrare | Buget alocat pentru reclame TV (mii dolari) | Buget alocat pentru reclame Radio (mii dolari) | Buget alocat pentru reclame în Ziare (mii dolari) | Vânzările produsului (mii unități) |

# Aplicarea practică în R

```
> # Încărcare librării necesare
> library(ggplot2)
> library(dplyr)
>
> # Instalare și încărcare librărie pentru corelații
> if (!require(corrplot)) install.packages("corrplot", dependencies=TRUE)
> library(corrplot)
>
> # Citirea datasetului
> file_path <- "~/Downloads/Advertising And Sales.csv"
> data <- read.csv(file_path)
>
> # Explorarea inițială a datasetului
> str(data)
'data.frame':   200 obs. of  5 variables:
 $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ TV       : num  230.1 44.5 17.2 151.5 180.8 ...
 $ Radio    : num  37.8 39.3 45.9 41.3 12.8 48.9 32.8 19.6 2.1 2.6 ...
 $ Newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
 $ Sales    : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
> summary(data)
       ID              TV             Radio         Newspaper          Sales
 Min.   :  1.00   Min.   :  0.70   Min.   : 0.00   Min.   :  0.30   Min.   : 1.60
 1st Qu.: 50.75   1st Qu.: 74.38   1st Qu.:10.07   1st Qu.: 12.75   1st Qu.:10.40
 Median :100.50   Median :149.75   Median :22.90   Median : 25.75   Median :12.90
 Mean   :100.50   Mean   :147.03   Mean   :23.29   Mean   : 30.55   Mean   :14.04
 3rd Qu.:150.25   3rd Qu.:218.82   3rd Qu.:36.52   3rd Qu.: 45.10   3rd Qu.:17.40
 Max.   :200.00   Max.   :296.40   Max.   :49.60   Max.   :114.00   Max.   :27.00
```

```
> # Vizualizare distributie variabile
> par(mfrow = c(1,3))
> hist(data$TV, main = "TV Advertising", col = "lightblue", xlab = "TV Budget (in $1000)")
> hist(data$Radio, main = "Radio Advertising", col = "lightgreen", xlab = "Radio Budget (in $1000)")
> hist(data$Newspaper, main = "Newspaper Advertising", col = "lightpink", xlab = "Newspaper Budget (in
$1000)")
>
> # Scatter plots pentru fiecare variabila fata de Sales
> par(mfrow = c(1,3))
> plot(data$TV, data$Sales, main = "TV vs Sales", xlab = "TV Budget", ylab = "Sales", col = "skyblue", p
ch = 19)
> plot(data$Radio, data$Sales, main = "Radio vs Sales", xlab = "Radio Budget", ylab = "Sales", col = "pa
legreen", pch = 19)
> plot(data$Newspaper, data$Sales, main = "Newspaper vs Sales", xlab = "Newspaper Budget", ylab = "Sale
s", col = "lightcoral", pch = 19)
>
> # Corelatii intre variabile
> cor_matrix <- cor(data[,2:5])
> corrplot(cor_matrix, method = "color", type = "upper", tl.col = "black")
>
> # Regresie liniara simpla
> model_tv <- lm(Sales ~ TV, data = data)
> summary(model_tv)

Call:
lm(formula = Sales ~ TV, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.398 -1.917 -0.207  2.073  7.198

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.05273    0.45604   15.46   <2e-16 ***
TV           0.04751    0.00268   17.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.247 on 198 degrees of freedom
Multiple R-squared:  0.6135,    Adjusted R-squared:  0.6115
F-statistic: 314.3 on 1 and 198 DF,  p-value: < 2.2e-16
```

# Aplicarea practică în R

```
> model_radio <- lm(Sales ~ Radio, data = data)
> summary(model_radio)

Call:
lm(formula = Sales ~ Radio, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7232  -2.1459   0.8084   2.7740   8.1920

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.34715    0.56247  16.618   <2e-16 ***
Radio        0.20141    0.02037   9.887   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.273 on 198 degrees of freedom
Multiple R-squared:  0.3305,    Adjusted R-squared:  0.3272
F-statistic: 97.76 on 1 and 198 DF,  p-value: < 2.2e-16

>
> model_newspaper <- lm(Sales ~ Newspaper, data = data)
> summary(model_newspaper)

Call:
lm(formula = Sales ~ Newspaper, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2520  -3.3879  -0.8452   3.4851  12.7487

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.37992    0.62067  19.946  < 2e-16 ***
Newspaper    0.05427    0.01656   3.278  0.00124 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.086 on 198 degrees of freedom
Multiple R-squared:  0.05147,   Adjusted R-squared:  0.04668
F-statistic: 10.74 on 1 and 198 DF,  p-value: 0.001235
```

```
> # Adaugare liniilor de regresie pe grafice
> par(mfrow = c(1,3))
> plot(data$TV, data$Sales, main = "TV vs Sales", xlab = "TV Budget", ylab = "Sales", col = "skyblue", p
ch = 19)
> abline(model_tv, col = "lightblue", lwd = 2)
>
> plot(data$Radio, data$Sales, main = "Radio vs Sales", xlab = "Radio Budget", ylab = "Sales", col = "pa
legreen", pch = 19)
> abline(model_radio, col = "lightgreen", lwd = 2)
>
> plot(data$Newspaper, data$Sales, main = "Newspaper vs Sales", xlab = "Newspaper Budget", ylab = "Sale
s", col = "lightcoral", pch = 19)
> abline(model_newspaper, col = "lightpink", lwd = 2)
>
> # Regresie liniara multipla
> model_multi <- lm(Sales ~ TV + Radio + Newspaper, data = data)
> summary(model_multi)

Call:
lm(formula = Sales ~ TV + Radio + Newspaper, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8335 -0.8662  0.2411  1.1927  3.4411

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.003556   0.313296   9.587   <2e-16 ***
TV           0.045686   0.001402  32.583   <2e-16 ***
Radio        0.187110   0.008649  21.634   <2e-16 ***
Newspaper   -0.001330   0.005905  -0.225    0.822
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.695 on 196 degrees of freedom
Multiple R-squared:  0.8958,    Adjusted R-squared:  0.8942
F-statistic: 561.4 on 3 and 196 DF,  p-value: < 2.2e-16

>
> # Comparare modele
> r_squared <- c(summary(model_tv)$r.squared, summary(model_radio)$r.squared, summary(model_newspaper)
$r.squared, summary(model_multi)$r.squared)
> names(r_squared) <- c("TV", "Radio", "Newspaper", "Multiple")
> print(r_squared)
       TV     Radio  Newspaper   Multiple
0.61348037 0.33054333 0.05147076 0.89575901
```

```
> # Calcul RMSE pentru fiecare model
> rmse <- function(model) {
+     sqrt(mean(model$residuals^2))
+ }
> rmse_values <- c(rmse(model_tv), rmse(model_radio), rmse(model_newspaper), rmse(model_mult
i))
> names(rmse_values) <- c("TV", "Radio", "Newspaper", "Multiple")
> print(rmse_values)
       TV     Radio Newspaper  Multiple
 3.230613  4.251679  5.060863  1.677716
>
> # Grafic predicții vs valori reale
> predictions <- predict(model_multi)
> plot(data$Sales, predictions, main = "Predicții vs Vânzări Reale", xlab = "Vânzări Reale",
ylab = "Predicții Vânzări", col = "plum", pch = 19)
> abline(0, 1, col = "lavender", lwd = 2)
```

# Aplicarea practică în R

**Histogramă pentru distribuția bugetelor publicitare (TV, Radio, Newspaper)**

*Observații:*

cheltuielile pentru TV și Radio sunt mai distribuite, în timp ce bugetele pentru Newspaper sunt mai concentrate în valori mici. Rezultă că, majoritatea companiilor preferă să investească mai mult în TV, urmat de Radio, iar cel mai puțin în Newspaper.

**Scatter plot: Bugetele publicitare vs. Vânzări**

*Observații:*

TV vs. Vânzări → Se observă o relație clară de creștere (mai mult buget → mai multe vânzări).

Radio vs. Vânzări → De asemenea, există o tendință pozitivă, dar mai dispersată.

Newspaper vs. Vânzări → Pare mai aleatoriu, ceea ce sugerează o relație mai slabă.

Publicitatea TV influențează cel mai mult vânzările, urmată de Radio.

Newspaper pare să aibă cel mai mic impact asupra vânzărilor.

**Regresie liniară pentru fiecare canal de publicitate**

Linia de regresie arată tendința generală a relației dintre cheltuieli și vânzări.

*Observații:*

TV are cea mai puternică relație liniară cu vânzările.

Radio are o relație pozitivă, dar mai variabilă.

Newspaper arată o relație slabă, ceea ce sugerează că investițiile în ziare nu sunt eficiente pentru creșterea vânzărilor.

**Predicții vs. Vânzări reale (Model de regresie multiplă)**

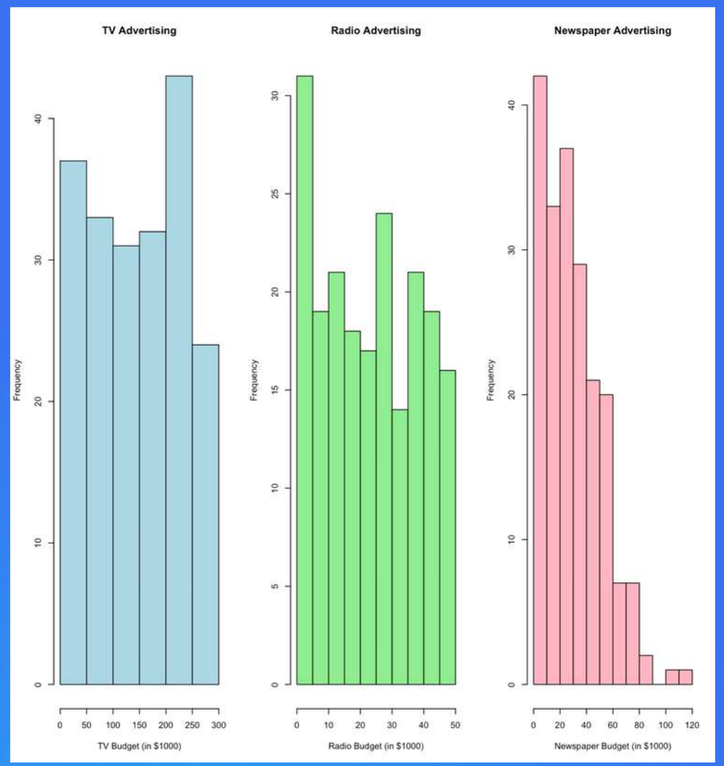
*Observații:*

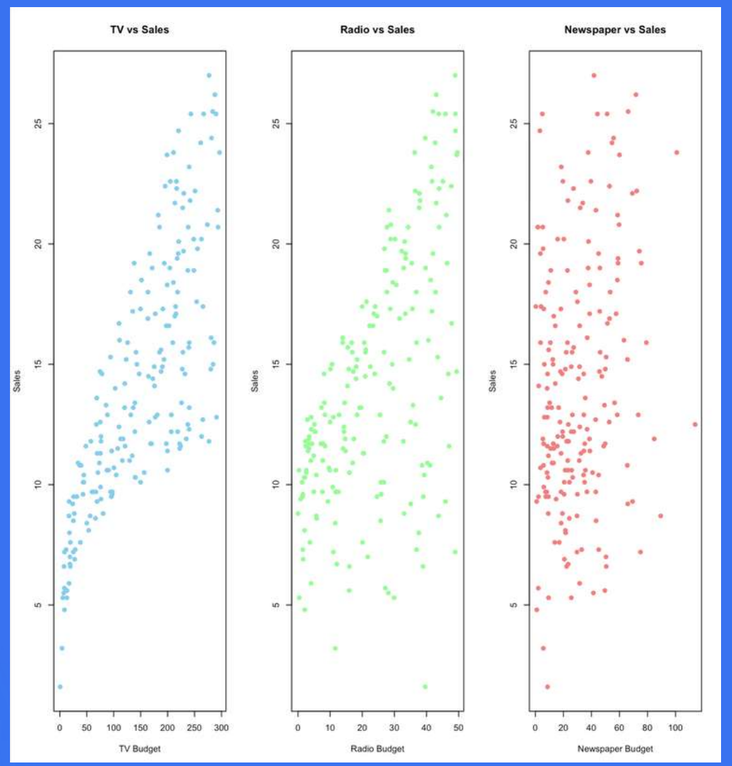
Acest grafic compară vânzările reale cu valorile prezise de modelul de regresie.

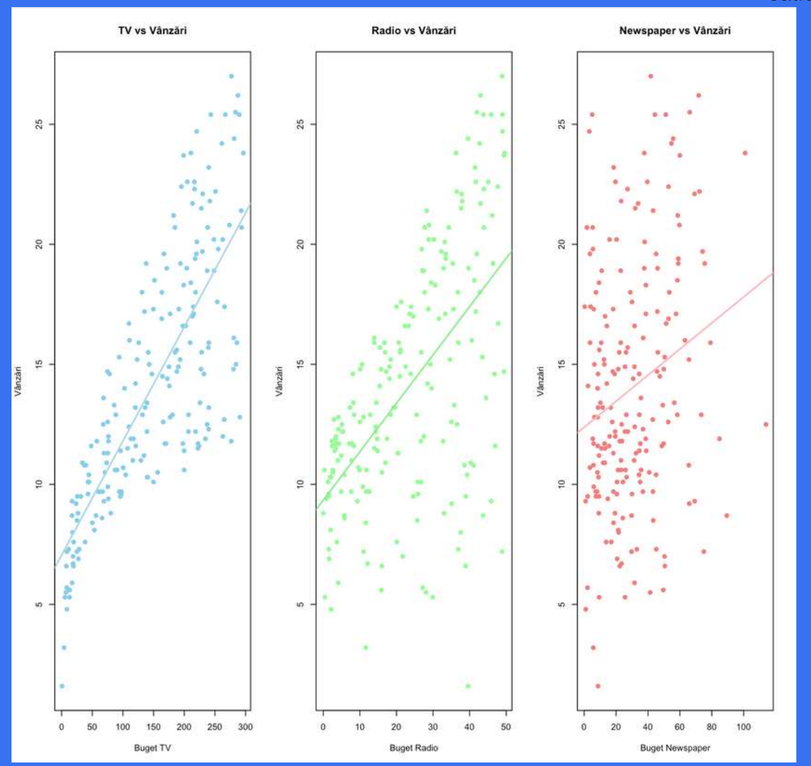
Dacă modelul ar fi perfect, toate punctele ar fi pe linia diagonală.

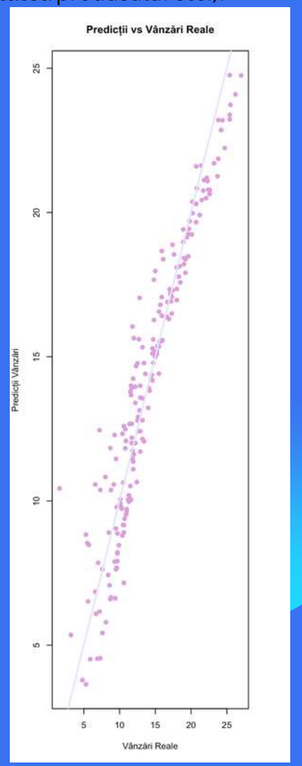
Modelul pare destul de precis, deoarece majoritatea punctelor sunt aproape de linia diagonală.

Totuși, există unele abateri, ceea ce înseamnă că alți factori ar putea influența vânzările (ex: sezonalitate, calitatea produsului etc.).



# Prezentarea datasetului

Setul de date *„AI, ML, Data Science Salary”* conține informații despre salariile din domeniile Inteligenței Artificiale, Machine Learning și Data Science pentru perioada 2020-2025. Acest set de date permite o analiză comparativă a factorilor care influențează salariile și poate fi utilizat pentru predicții și tendințe în domeniul tehnologiei.

| | | |
|---|---|---|
| **work_year**<br>Anul în care a fost plătit salariul | **experience_level**<br>Nivelul de experiență (Entry, Mid, Senior, Executive) | **employment_type**<br>Tipul de angajare (Full-time, Part-time, Contract, Freelance) |
| **job_title**<br>Denumirea postului ocupat | **salary**<br>Vânzările produsului (mii unități) | **salary_currency**<br>Moneda în care a fost plătit salariul |
| **salary_in_usd:**<br>Salariul exprimat în USD | **employee_residence**<br>Țara de reședință a angajatului | **remote_ratio**<br>Procentul muncii desfășurate la distanță, cu următoarele valori posibile:<br>• 0 – Fără muncă remote (mai puțin de 20%)<br>• 50 – Parțial remote / Hibrid<br>• 100 – Complet remote (mai mult de 80%) |
| **company_location**<br>Țara în care se află sediul principal sau filiala angajatoare | **company_size**<br>Numărul mediu de angajați din companie pe parcursul anului:<br>• S – Companie mică (mai puțin de 50 de angajați)<br>• M – Companie medie (între 50 și 250 de angajați)<br>• L – Companie mare (peste 250 de angajați) | |

# Aplicarea practică în R

```
> # Încărcarea librăriilor necesare
> library(ggplot2)
> library(dplyr)
>
> # Citirea datasetului
> data <- read.csv("~/Downloads/salaries.csv", na.strings = c("", "NA", "NaN", "Inf", "-Inf"))
>
> # Explorarea inițială a datasetului
> str(data)
'data.frame':   88584 obs. of  11 variables:
 $ work_year         : int  2025 2025 2025 2025 2025 2025 2025 2025 2025 2025 ...
 $ experience_level  : chr  "MI" "SE" "SE" "SE" ...
 $ employment_type   : chr  "FT" "FT" "FT" "FT" ...
 $ job_title         : chr  "Customer Success Manager" "Engineer" "Engineer" "Applied Scientist" ...
 $ salary            : int  57000 165000 109000 294000 137600 82000 44000 149800 89700 200000 ...
 $ salary_currency   : chr  "EUR" "USD" "USD" "USD" ...
 $ salary_in_usd     : int  60000 165000 109000 294000 137600 82000 44000 149800 89700 200000 ...
 $ employee_residence: chr  "NL" "US" "US" "US" ...
 $ remote_ratio      : int  50 0 0 0 0 0 0 0 0 0 ...
 $ company_location  : chr  "NL" "US" "US" "US" ...
 $ company_size      : chr  "L" "M" "M" "M" ...
> summary(data)
   work_year    experience_level   employment_type     job_title
 Min.   :2020   Length:88584       Length:88584       Length:88584
 1st Qu.:2024   Class :character   Class :character   Class :character
 Median :2024   Mode  :character   Mode  :character   Mode  :character
 Mean   :2024
 3rd Qu.:2024
 Max.   :2025
     salary         salary_currency    salary_in_usd    employee_residence
 Min.   :   14000   Length:88584       Min.   : 15000   Length:88584
 1st Qu.:  106000   Class :character   1st Qu.:106097   Class :character
 Median :  147000   Mode  :character   Median :146307   Mode  :character
 Mean   :  161932                      Mean   :157568
 3rd Qu.:  199500                      3rd Qu.:198600
 Max.   :30400000                      Max.   :800000
  remote_ratio    company_location   company_size
 Min.   :  0.00   Length:88584       Length:88584
 1st Qu.:  0.00   Class :character   Class :character
 Median :  0.00   Mode  :character   Mode  :character
 Mean   : 21.29
 3rd Qu.:  0.00
 Max.   :100.00
```

```
> # Verificarea și corectarea numelor coloanelor pentru a evita erori
> colnames(data) <- gsub(" ", "_", colnames(data))
> colnames(data) <- gsub("\\.", "_", colnames(data)) # Înlocuirea punctelor dacă există
> print(colnames(data))
 [1] "work_year"          "experience_level"   "employment_type"
 [4] "job_title"          "salary"             "salary_currency"
 [7] "salary_in_usd"      "employee_residence" "remote_ratio"
[10] "company_location"   "company_size"
>
> # Eliminarea valorilor lipsă doar pentru variabilele relevante
> cols_to_check <- c("salary_in_usd", "experience_level", "remote_ratio", "company_size")
> data <- data[complete.cases(data[, cols_to_check]), ]
>
> # Transformarea variabilelor categorice în numerice pentru analiză
> data$experience_level <- as.numeric(as.factor(data$experience_level))
> data$remote_ratio <- as.numeric(data$remote_ratio)
> data$company_size <- as.numeric(as.factor(data$company_size))
>
> # Regresie liniară simplă: Salary ~ Experience Level
> model_simple <- lm(salary_in_usd ~ experience_level, data = data)
> summary(model_simple)

Call:
lm(formula = salary_in_usd ~ experience_level, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-155862  -49205  -11671   35795  688864

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      81805.7      902.5   90.64   <2e-16 ***
experience_level 22466.4      258.2   87.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70580 on 88582 degrees of freedom
Multiple R-squared:  0.07873,   Adjusted R-squared:  0.07871
F-statistic:  7570 on 1 and 88582 DF,  p-value: < 2.2e-16
```

```
> # Vizualizare regresie multiplă (predicții vs. valori reale) cu culori pastelate
> predictions <- predict(model_multiple)
> ggplot(data, aes(x = salary_in_usd, y = predictions)) +
+     geom_point(color = "#92C5DE", alpha = 0.7) +
+     geom_abline(slope = 1, intercept = 0, color = "#F4A582", linetype = "dashed") +
+     labs(title = "Regresie Liniară Multiplă: Predicții vs. Valori Reale",
+          x = "Salary Real",
+          y = "Salary Prezis")
```

# Aplicarea practică în R

Ilustrarea relației dintre nivelul de experiență al angajaților și salariul în USD.

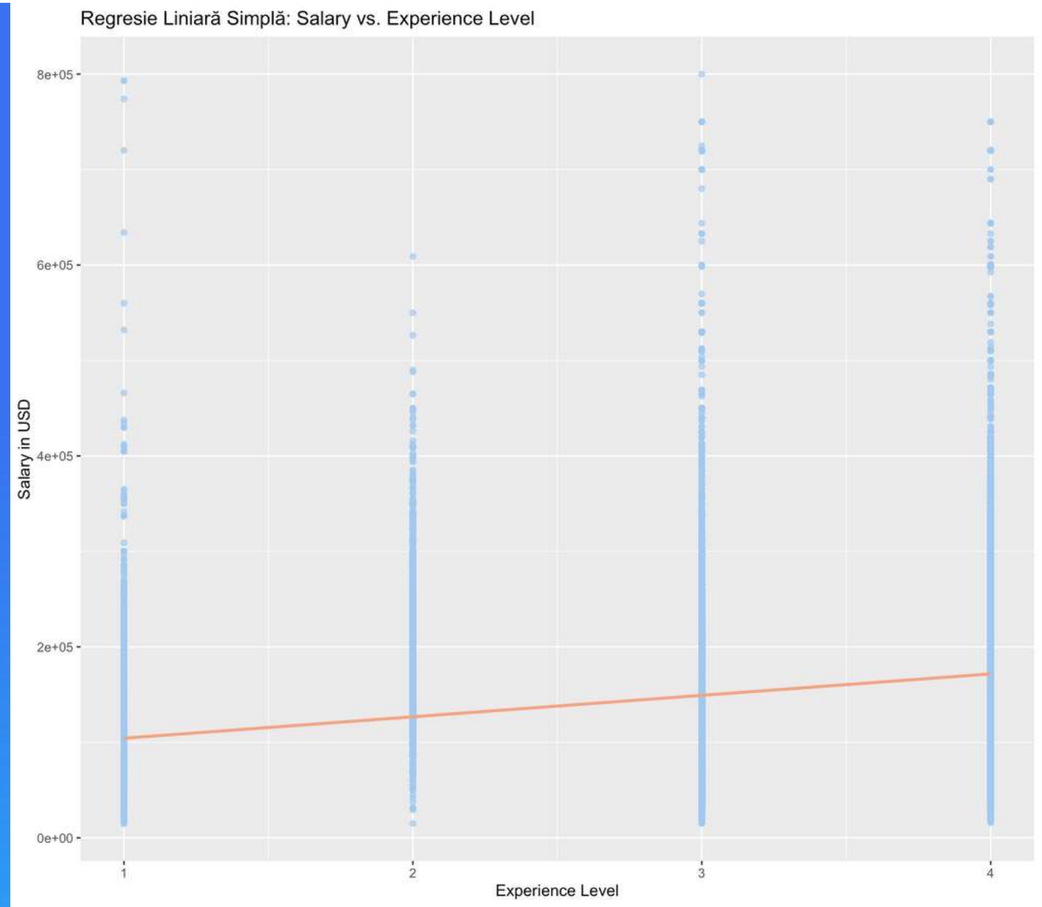
*Observații:*

O tendință ascendentă ușoară, ceea ce sugerează că salariile cresc odată cu experiența.

Totuși, variabilitatea salariilor este foarte mare pentru fiecare nivel de experiență.

Există multe valori extreme (outliers), ceea ce indică că salariile pot varia considerabil în funcție de alte factori.

Experiența are un impact pozitiv asupra salariului, dar nu explică complet variația acestuia.




Compararea salariilor reale din dataset cu valorile prezise de modelul de regresie multiplă.

*Observații:*

Modelul reușește să surprindă o tendință generală, dar există o mare variabilitate între salariile reale și cele prezise.

Punctele sunt destul de dispersate, ceea ce sugerează că există alți factori importanți care influențează salariile și care nu sunt incluși în acest model.

Regresia multiplă oferă predicții mai precise decât regresia simplă, dar nu explică în totalitate variația salariilor.

Alte variabile, cum ar fi job title, industry, sau competențele tehnice, ar putea îmbunătăți precizia modelului.

# Compararea modelelor

| Criteriu | Regresie Liniară Simplă | Regresie Liniară Multiplă |
|---|---|---|
| Număr de variabile independente | O singură variabilă independentă | Două sau mai multe variabile independente |
| Complexitate | Scăzută - Model ușor de înțeles | Ridicată - Necesită mai multă analiză și calcul |
| Interpretabilitate | Foarte interpretabilă - Ușor de vizualizat relația dintre două variabile | Mai dificilă - Interacțiunile dintre variabile pot fi mai greu de interpretat |
| Precizie în predicții | Scăzută - Nu ia în considerare alți factori | Mai precisă - Include mai mulți factori care influențează variabila dependentă |
| Utilitate practică | Utilizată când avem o relație clară între două variabile | Utilizată când mai mulți factori influențează rezultatul |
| Exemplu de aplicare | Predicția salariului doar pe baza experienței | Predicția salariului în funcție de experiență, remote work și mărimea companiei |

# Concluzii

Regresia liniară este un instrument puternic pentru analiza relațiilor dintre variabile și realizarea de predicții. Regresia liniară simplă este utilă când există o relație clară între două variabile, fiind ușor de interpretat, dar limitată în precizie. Pe de altă parte, regresia liniară multiplă oferă predicții mai exacte prin includerea mai multor factori, însă este mai complexă și necesită o interpretare atentă. Alegerea modelului potrivit depinde de obiectivul analizei și de datele disponibile.