Proiect Modele de Regresie 2025

Țăranu Daria Cristiana-**grupa 321**
Lungu Anne-Marie-**grupa 321**

Tema: Selecția modelelor în regresie – aspecte teoretice și exemple în R

1. Introducere

In acest proiect ne propunem sa analizam ce factori determina popularitatea unui film. Pentru aceasta analiza, am folosit setul de date "movies.csv", care contine informatii despre 750 de filme de top, inclusiv: titlul filmului, anul lansarii, genul, scorul IMDb, scorul Metacritic, numarul de voturi primite si incasarile brute. Aceste date ne permit sa investigam daca variabile precum genul filmului, durata, anul lansarii sau scorul Metacritic pot fi folosite ca predictori ai succesului comercial sau critic al unui film.

Lucram cu urmatoarele date:

- **Series_Title**: titlul filmului
- **Released_Year**: anul in care filmul a fost lansat
- **Genre**: genul filmului
- **IMDB_Rating**: scorul IMDb (evaluarea publicului pe o scara de la 1 la 10)
- **Meta_score**: scorul Metacritic (evaluarea criticilor profesionisti pe o scara de la 0 la 100)
- **No_of_Votes**: numarul total de voturi primite pe IMDb – indicator al popularitatii
- **Gross**: incasarile brute ale filmului (venituri comerciale in dolari SUA)

Analiza pe setul de date ales:

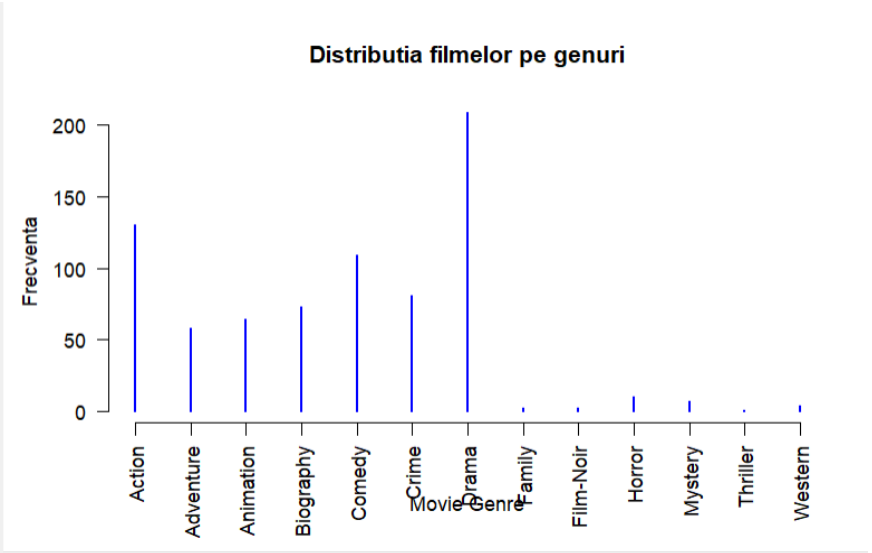
Cititm setul de date:

```
library(readr)
movies <- read_csv("C:/Users/Asus/Desktop/examen mr/Proiect/datain/movies.csv")
movies <- na.omit(movies)
View(movies)
summary(movies)
```

```
Series_Title       Released_Year         Genre             IMDB_Rating      Meta_score      No_of_Votes          Gross
Length:750         Length:750         Length:750         Min.   :7.600   Min.   : 28.00   Min.   :  25198   Min.   :     1305
Class :character   Class :character   Class :character   1st Qu.:7.700   1st Qu.: 70.00   1st Qu.:  88547   1st Qu.:  5014812
Mode  :character   Mode  :character   Mode  :character   Median :7.900   Median : 78.00   Median : 219734   Median : 31900000
                                                         Mean   :7.935   Mean   : 77.46   Mean   : 342133   Mean   : 74952069
                                                         3rd Qu.:8.100   3rd Qu.: 86.00   3rd Qu.: 481219   3rd Qu.: 98091571
                                                         Max.   :9.300   Max.   :100.00   Max.   :2343110   Max.   :936662225
```

Am facut distributia filmelor pe genuri:

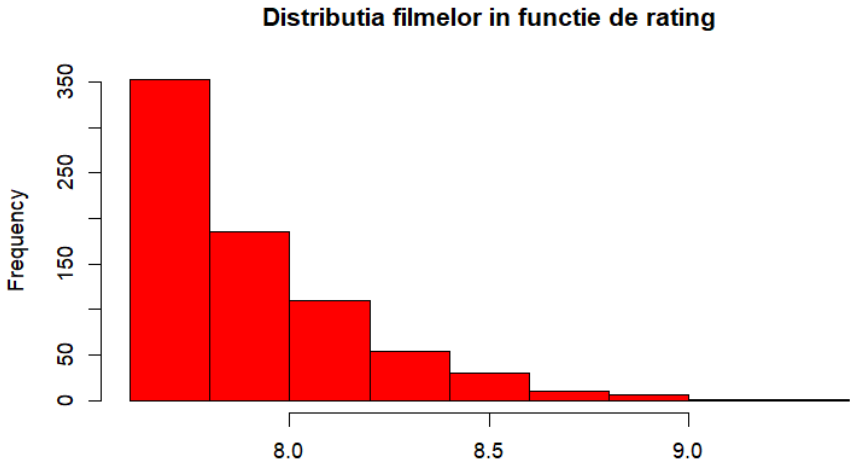
```
#numaram de cate ori apare fiecare gen de film in setul nostru de date
genre_counts <- table(movies$Genre)
#plotam un grafic in care vedem frecventa fiecarui gen
plot(genre_counts,
     col = "blue",
     las = 2,
     xlab = "Movie Genre",
     ylab = "Frecventa",
     main = "Distributia filmelor pe genuri")
```



Se remarca o predominanta clara a genului drama, care inregistreaza cea mai mare frecventa din setul de date, depasind pragul de 80 de aparitii. Acest lucru sugereaza ca filmele dramatice sunt cele mai reprezentate in esantionul analizat.

Distributia filmelor in functie de rating:

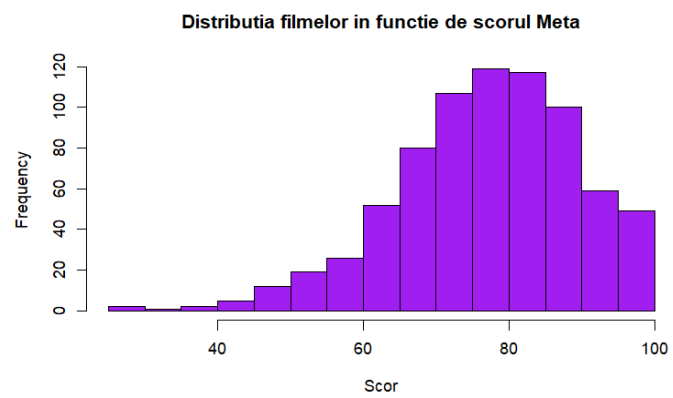
```
hist(movies$IMDB_Rating,
     col = "red",
     xlab = "rating",
     main = "Distributia filmelor in functie de rating")
```



Graficul arata distributia ratingurilor IMDb pentru filmele din set. Cele mai multe filme au scoruri intre 7.5 si 8.0, cu un varf in jur de 7.8. Pe masura ce ratingul creste, numarul filmelor scade. Distributia este asimetrica spre dreapta, ceea ce arata ca ratingurile foarte mari sunt mai rare.

Distributia filmelor in functie de scorul Meta:

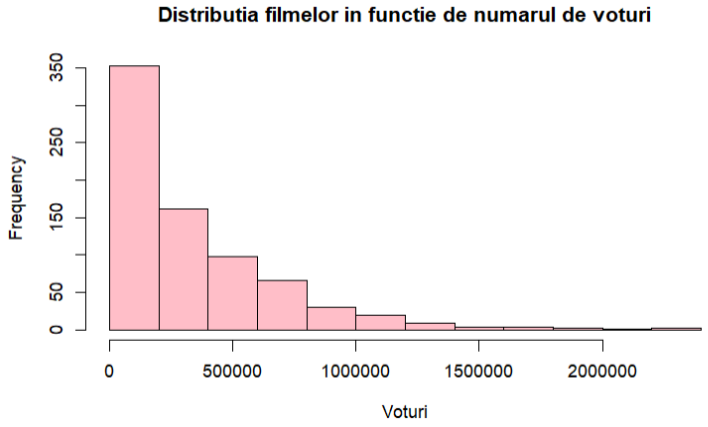
```
hist(movies$Meta_score,
     col = "purple",
     xlab = "Scor",
     main = "Distributia filmelor in functie de scorul Meta")
```



Se remarca o concentratie ridicata a filmelor cu scoruri Metacritic intre 70 si 90, cu un varf in jurul valorii de 80. Acest lucru sugereaza ca majoritatea filmelor din setul analizat au fost bine primite de critici, iar scorurile extreme (foarte mici sau foarte mari) sunt mai rare.

Distributia filmelor in functie de numarul de voturi:

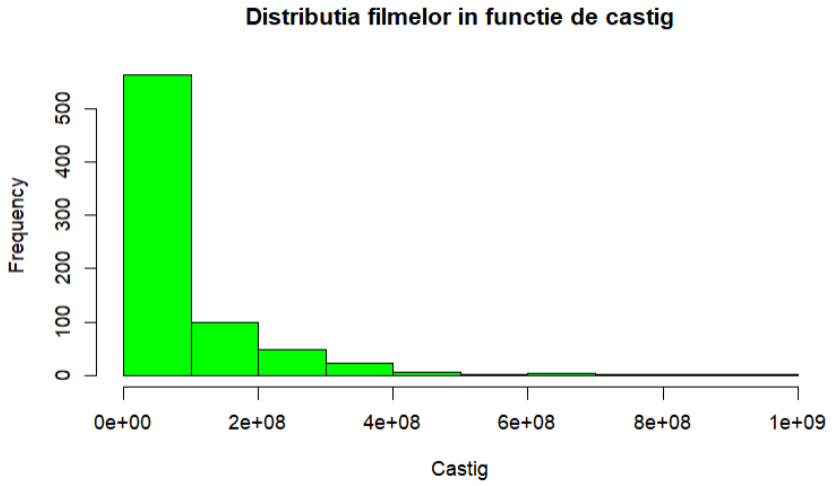
```
hist(movies$No_of_Votes,
     col = "pink",
     xlab = "Voturi",
     main = "Distributia filmelor in functie de numarul de voturi")
```



Se remarca o distributie puternic asimetrica spre dreapta, majoritatea filmelor avand un numar de voturi mai mic de 500.000. Filmele cu foarte multe voturi sunt rare, ceea ce sugereaza ca doar cateva titluri au reusit sa atraga un interes larg din partea publicului.

Distributia filmelor in functie de castig:

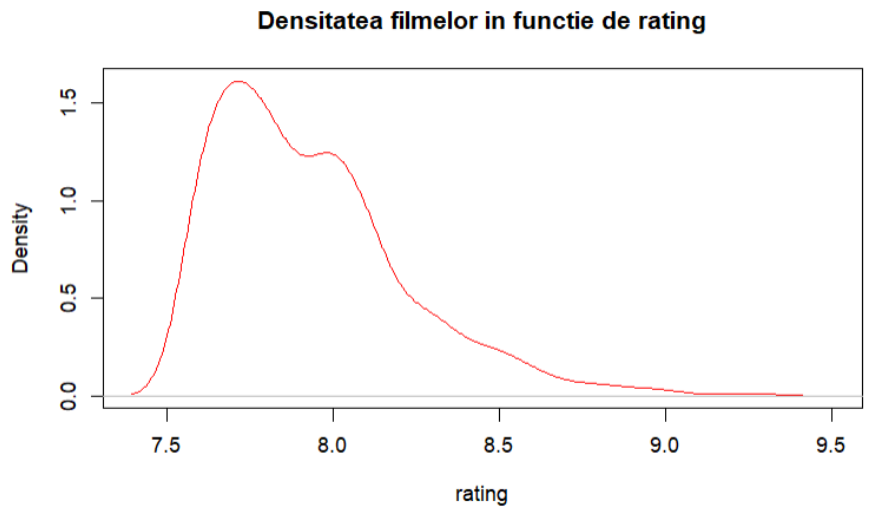
```
hist(movies$Gross,
     col = "green",
     xlab = "Castig",
     main = "Distributia filmelor in functie de castig")
```



Se remarca o distributie extrem de asimetrica spre dreapta, majoritatea filmelor avand incasari brute relativ mici. Cele mai multe filme au castiguri sub 200 de milioane USD, iar doar un numar redus depasesc praguri mari de box office. Aceasta distributie sugereaza ca succesul comercial major este rar si concentrat in jurul catorva titluri.

Densitatea in functie de rating:

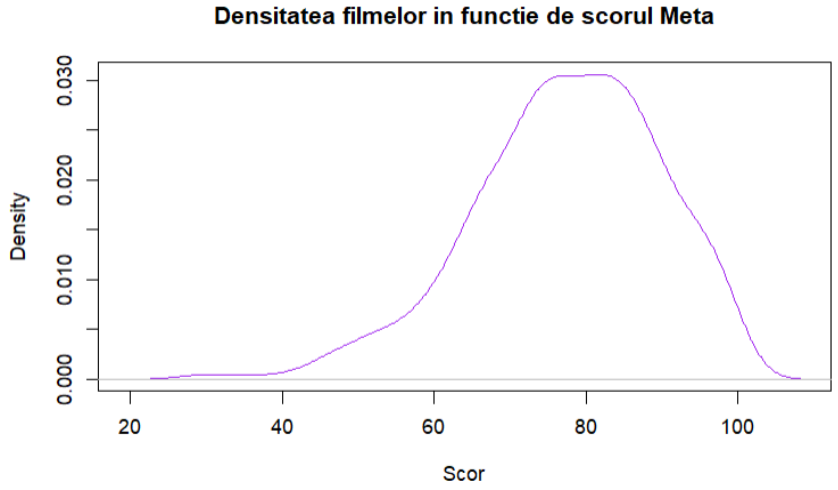
```
#densitatile fiecarora
dens_imd <- density(movies$IMDB_Rating)
plot(dens_imd,
     col = "red",
     xlab = "rating",
     main = "Densitatea filmelor in functie de rating")
```



Curba de densitate arata ca cele mai multe filme au ratinguri IMDb in jurul valorii de 7.8, cu doua varfuri evidente. Distributia este usor asimetrica spre dreapta, ceea ce inseamna ca ratingurile mari sunt mai rare. Aceasta reprezentare evidentiaza tendinta generala a datasetului de a include filme cu scoruri bune, dar nu extreme.

Densitatea filmelor in functie de scorul Meta:

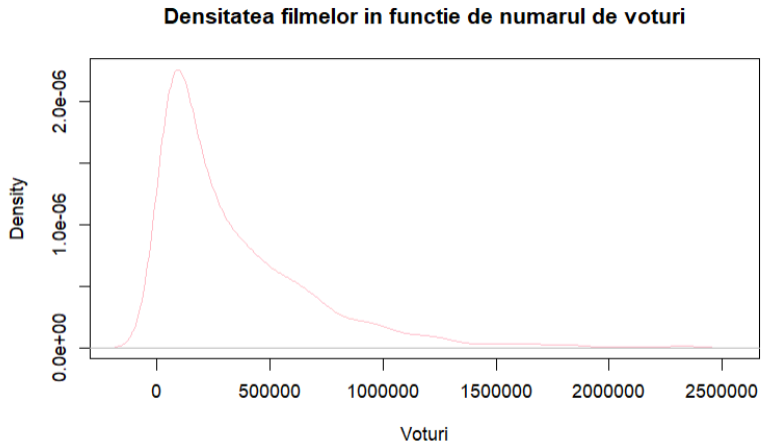
```
dens_meta <- density(movies$Meta_score)
plot(dens_meta,
     col = "purple",
    xlab = "Scor",
    main = "Densitatea filmelor in functie de scorul Meta")
```



Curba de densitate pentru scorul Metacritic arata o distributie aproape simetrica, centrata in jurul valorii de 80. Acest lucru sugereaza ca majoritatea filmelor din set au fost evaluate pozitiv de critici, iar extremele – scoruri foarte mici sau foarte mari – sunt rare.

Densitatea filmelor in functie de numarul de voturi:

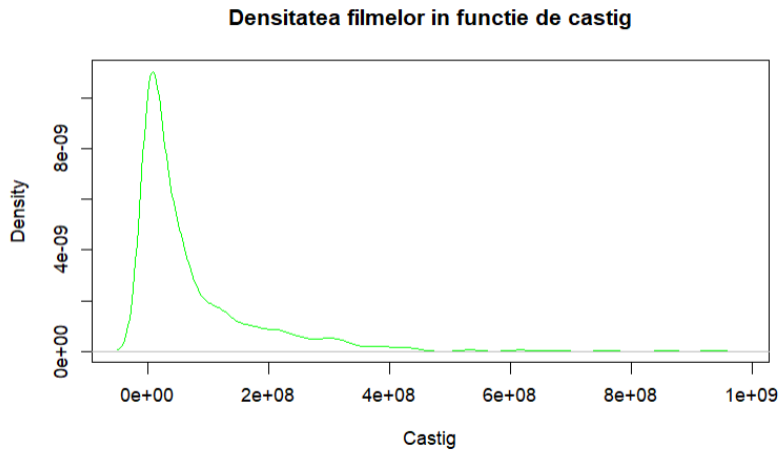
```
dens_vot <- density(movies$No_of_Votes)
plot(dens_vot,
     col = "pink",
     xlab = "Voturi",
     main = "Densitatea filmelor in functie de numarul de voturi")
```



Curba de densitate arata ca majoritatea filmelor au un numar relativ mic de voturi pe IMDb, cu un varf accentuat sub 500.000. Pe masura ce numarul de voturi creste, densitatea scade rapid, indicand ca doar cateva filme au reusit sa atraga un interes masiv din partea publicului. Distributia este vizibil asimetrica spre dreapta.

Densitatea filmelor in functie de castig:

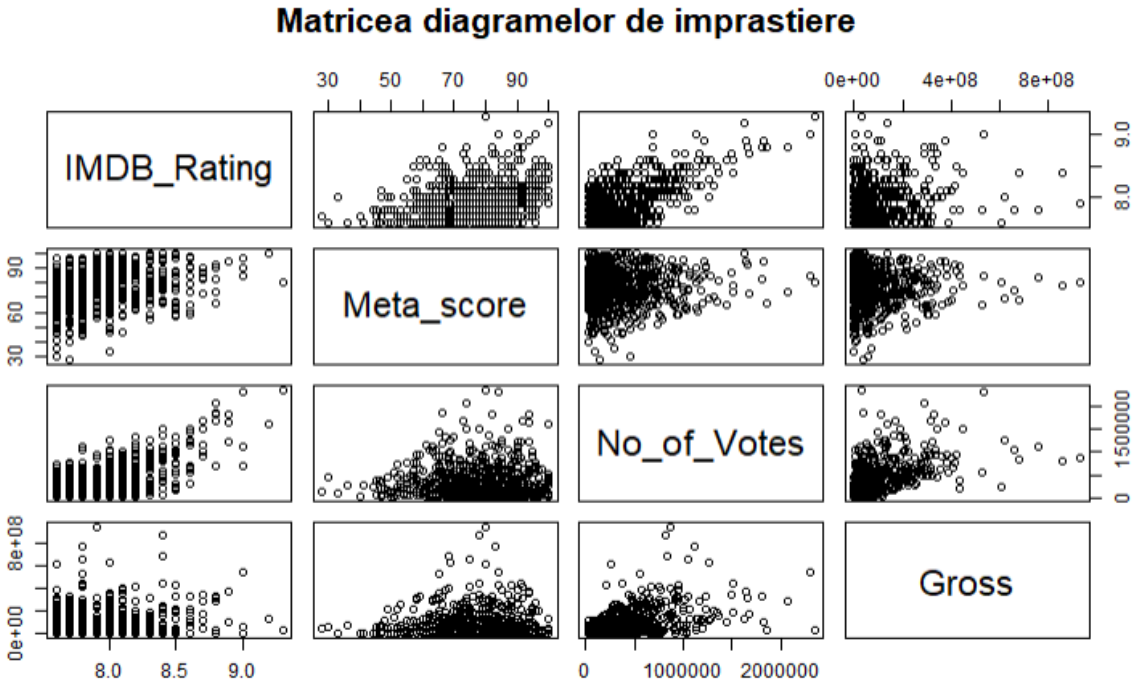
```
dens_castig <- density(movies$Gross)
plot(dens_castig,
     col = "green",
     xlab = "Castig",
     main = "Densitatea filmelor in functie de castig")
```



Curba de densitate pentru castig arata o distributie puternic asimetrica spre dreapta. Majoritatea filmelor au incasari reduse, iar frecventa scade rapid pe masura ce castigurile cresc. Doar un numar mic de filme au generat venituri foarte mari, ceea ce reflecta inegalitatea in succesul comercial.

Matricea diagramelor de imprastiere permite observarea relatiilor dintre variabilele explicative si variabila tinta, precum si posibile tipare liniare sau valori extreme.

```
# Selectam doar coloanele numerice
numeric_cols <- sapply(movies, is.numeric)
#facem matricea de imprastiere
pairs(movies[, numeric_cols], main = "Matricea diagramelor de imprastiere")
```



Analizand graficul, se observa o corelatie pozitiva intre IMDB_Rating si Meta_score, ceea ce sugereaza ca filmele apreciate de critici sunt, in general, bine evaluate si de public. De asemenea, exista o asociere clara intre No_of_Votes si Gross, indicand ca filmele cu incasari mari tind sa primeasca si un numar mare de voturi. Celelalte relatii dintre variabile par mai dispersate si nu prezinta o legatura liniara evidenta, insa acestea pot fi investigate in continuare in analiza de regresie.

Modelul de regresie:

O relatie de regresie se defineste ca o functie de parametrii $x_1,...,x_n$ unde $x_1,...,x_n$ sunt criterile dupa care ne uitam.

$y(x_1,...,x_n)$ = $f(x_1,...,x_n)$ + $\varepsilon(x_1,...,x_n)$

↓ ↓ ↓

[variabila de raspuns] [componenta sistematica a modelului] [componenta aleatoare]

$x_1,...,x_n$ - variabile aleatoare

Masura de risc are rolul de a minimiza abaterea modelului fata de observaiile reale și este definită prin:

$R(X,Y)=\mathbf{E}[L(X,Y)]$

L(X,Y) - este functia de pierdere, care evalueaza diferenta dintre valoarea estimata de model si valoarea reala a observatiei.

Modelul de regresie liniara simpla:

$y= \mathbf{E}[Y|X] + \varepsilon$

↓

f(X)

Presupunem ca avem observațiile $\{(x_i, y_i)\}$, i = 1, 2,…, n. Daca reprezentam aceste puncte intr-un sistem de axe xOy si observăm o relație aproape liniara intre y și x, atunci putem presupune existenta unei relatii de forma:

$y = \beta_0 + \beta_1 x$

In practica, relatia nu este perfect liniara, motiv pentru care introducem un termen de eroare ε, rezultand modelul:

$y = \beta_0 + \beta_1 x + \varepsilon$, cu $\varepsilon \sim N(0, \sigma^2)$.

Ipotezele modelului : $\mathbf{E}[\varepsilon_i] = 0$ si $Cov(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$

In populatie avem : Vrem sa gasim

$$\text{argmin } \mathbf{E}[(Y - \beta_0 - \beta_1 X)^2]$$

Notam MSE $(\beta_0, \beta_1) = \mathbf{E}[(Y - \beta_0 - \beta_1 X)^2]$

$= \mathbf{E}[Y^2] - 2\beta_0\mathbf{E}[Y] - 2\beta_1\mathbf{E}[XY] + 2\beta_0\beta_1\mathbf{E}[X] + \beta_0^2 + \beta_1^2\mathbf{E}[X^2]$

Derivam MSE in functie de β_0 si β_1 si egalam derivatele cu zero , de unde obtinem ca :

- $\beta_0 = \mathbf{E}[Y] - \beta_1\mathbf{E}[X]$
- $\beta_1 * (\mathbf{E}[X^2] - \mathbf{E}[X]^2) + \mathbf{E}[X]\mathbf{E}[Y] - \mathbf{E}[XY] = 0 \Rightarrow$

$\beta_1 = \frac{Cov(X,Y)}{Var(X)}$

In esantion : $y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

In esantion, dorim sa gasim β_0 si β_1 pentru care $RSS(\beta_0, \beta_1)$ sa fie minima, unde:

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - [(\beta_0 + \beta_1 x_i)^2]]$$

La fel ca mai devreme, cautam punctele critice si avem ca:

$$\widehat{\beta_0} = \overline{y_n} - \widehat{\beta_1 x_n}$$

$$\widehat{\beta_1}=\frac{\sum_{i=1}^{n}[(x_i-\overline{x_n})(y_i-\overline{y_n})]}{\sum_{i=1}^{n}(x_i-\overline{x_n})^2}$$

$\widehat{\beta_0}$ si $\widehat{\beta_1}$ mai au urmatoarele formule:

$$\widehat{\beta_0}=\beta_0+\sum_{i=1}^{n}d_i\varepsilon_i$$

$$\widehat{\beta_1}=\beta_1+\sum_{i=1}^{n}\omega_i\varepsilon_i$$

$\widehat{\beta_0}$ si $\widehat{\beta_1}$ sunt estimatori nedeplasati si variantele lor au urmatoarele formule :

$$\text{Var}(\widehat{\beta_1})=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x_n})^2}$$

$$\text{Var}(\widehat{\beta_0})=\frac{\sigma^2}{n}+\frac{\sigma^2\overline{x_n}^2}{\sum_{i=1}^{n}(x_i-\overline{x_n})^2},$$

iar $\text{Cov}(\widehat{\beta_0},\ \widehat{\beta_1})=\text{Cov}(\overline{y_n}-\widehat{\beta_1}\overline{x_n},\ \widehat{\beta_1})=-\frac{\sigma^2\overline{x_n}}{\sum_{i=1}^{n}(x_i-\overline{x_n})^2}$

Din:

$$y_i=\ \beta_0+\beta_1x_i+\ \varepsilon_i$$

$$\widehat{y_i}=\widehat{\beta_1}x_i+\widehat{\beta_0}$$

$$\Rightarrow(y_i-\widehat{y_i})=\ (\beta_0-\widehat{\beta_0})\ +(\beta_1-\widehat{\beta_1})x_i+\varepsilon_i\Rightarrow$$

$$\mathbf{E}[\widehat{\varepsilon_i}]=0$$

iar $\widehat{\varepsilon_i} = (y_i - \widehat{y_i})$ se numesc valori reziduale

Un lucru foarte important in regresia liniara este sa intelegem cum se poate imparti variatia valorilor lui Y in doua bucati si anume: o parte pe care o explica modelul nostru si o parte care ramane neexplicata. Are urmatoarea formula:

Var(Y)=Var(**E**[Y|X])+**E**[Var(Y|X)]

Notam:

SST= $\sum_{i=1}^{n} (y_i - \overline{y_n})^2$ → suma abaterilor patratice totale

SSR= $\sum_{i=1}^{n} (\widehat{y_i} - \overline{y_n})^2$ → suma abaterilor patratice explicate de modelul de regresie

RSS= $\sum_{i=1}^{n} (y_i - \widehat{y_i})^2$ → suma abaterilor patratice reziduale

$$SST=SSR+RSS$$

R^2este coeficientul de determinare si masoara proportia din variatia variabilei raspuns explicata pe modelul de regresie si are formula:

$$R^2 = \frac{SSR}{SST}$$

si $R^2 = r_{xy}^2 = r_{y\widehat{y}}^2$ (unde r_{xy}= coef. de rel. empirica)

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}[(x_i - \overline{x_n})(y_i - \overline{y_n})]}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y_n})}}$$

Exemple: Luam ca variabila tinta IMDB_Rating

1) In functie de Meta_score:

```
model_meta <- lm(IMDB_Rating~Meta_score,data=movies)
summary((model_meta))
```

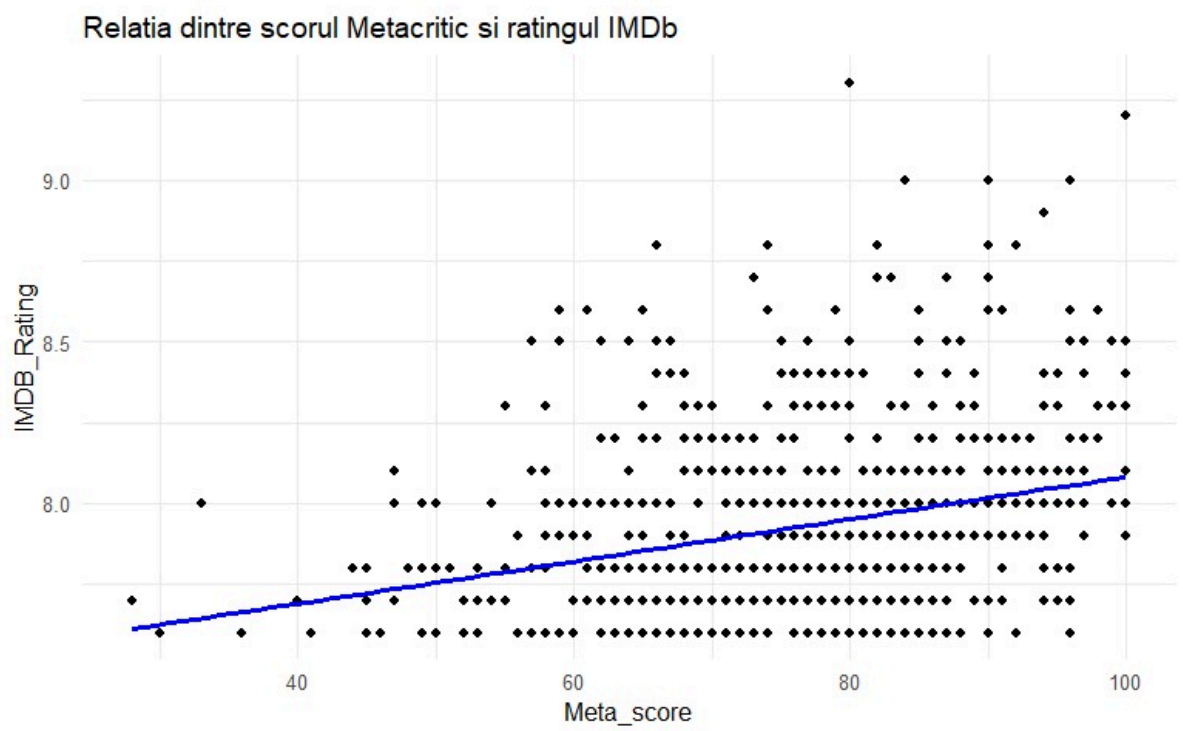
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.45608 -0.19908 -0.05796  0.13691  1.34859

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.4281133  0.0636741  116.66  < 2e-16 ***
Meta_score  0.0065413  0.0008116    8.06 3.01e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2775 on 748 degrees of freedom
Multiple R-squared:  0.07991,   Adjusted R-squared:  0.07868
F-statistic: 64.97 on 1 and 748 DF,  p-value: 3.013e-15
```

Rezultatul arata ca exista o relatie pozitiva si semnificativa intre cele doua variabile. Coeficientul estimat pentru Meta_score este 0.0065, ceea ce inseamna ca, in medie, o crestere cu 1 punct a scorului Metacritic este asociata cu o crestere de aproximativ 0.0065 puncte a ratingului IMDb.

Valoarea p asociata coeficientului este extrem de mica ($p < 0.001$), ceea ce indica faptul ca efectul este semnificativ statistic.

Totusi, valoarea R-squared este 0.0799, ceea ce inseamna ca doar aproximativ 8% din variatia ratingului IMDb este explicata de scorul Metacritic. Asadar, desi relatia este semnificativa, scorul criticilor nu este un predictor foarte puternic luat singur.

```r
# Reprezentare grafica a relatiei dintre Meta_score si IMDB_Rating
ggplot(movies, aes(x = Meta_score, y = IMDB_Rating)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  theme_minimal() +
  labs(title = "Relatia dintre scorul Metacritic si ratingul IMDb",
       x = "Meta_score",
       y = "IMDB_Rating")
```



Graficul evidentiaza relatia dintre Meta_score si IMDB_Rating pentru filmele din setul analizat. Fiecare punct din grafic reprezinta un film, pozitionat in functie de cele doua variabile.

Se observa o tendinta generala pozitiva, ilustrata prin linia de regresie albastra. Aceasta indica faptul ca, in medie, filmele care au obtinut un scor mai mare din partea criticilor tind sa fie mai bine evaluate si de public. Cu toate acestea, punctele sunt raspandite in jurul liniei, ceea ce sugereaza ca relatia nu este una foarte puternica.

Aceasta dispersie arata ca exista si alte variabile care influenteaza ratingul IMDb si care nu sunt surprinse doar de scorul criticilor. Totusi, chiar si cu aceasta variabilitate, relatia este semnificativa statistic, ceea ce valideaza influenta partiala a scorului Metacritic asupra perceptiei publicului.

2) In functie de No_of_Votes:

```
model_vot <- lm(IMDB_Rating~No_of_Votes,data=movies)
summary(model_vot)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.53246 -0.16939 -0.01547  0.16727  0.88931

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.762e+00  1.165e-02  666.45   <2e-16 ***
No_of_Votes 5.058e-07  2.377e-08   21.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2283 on 748 degrees of freedom
Multiple R-squared:  0.3771,    Adjusted R-squared:  0.3763
F-statistic: 452.9 on 1 and 748 DF,  p-value: < 2.2e-16
```

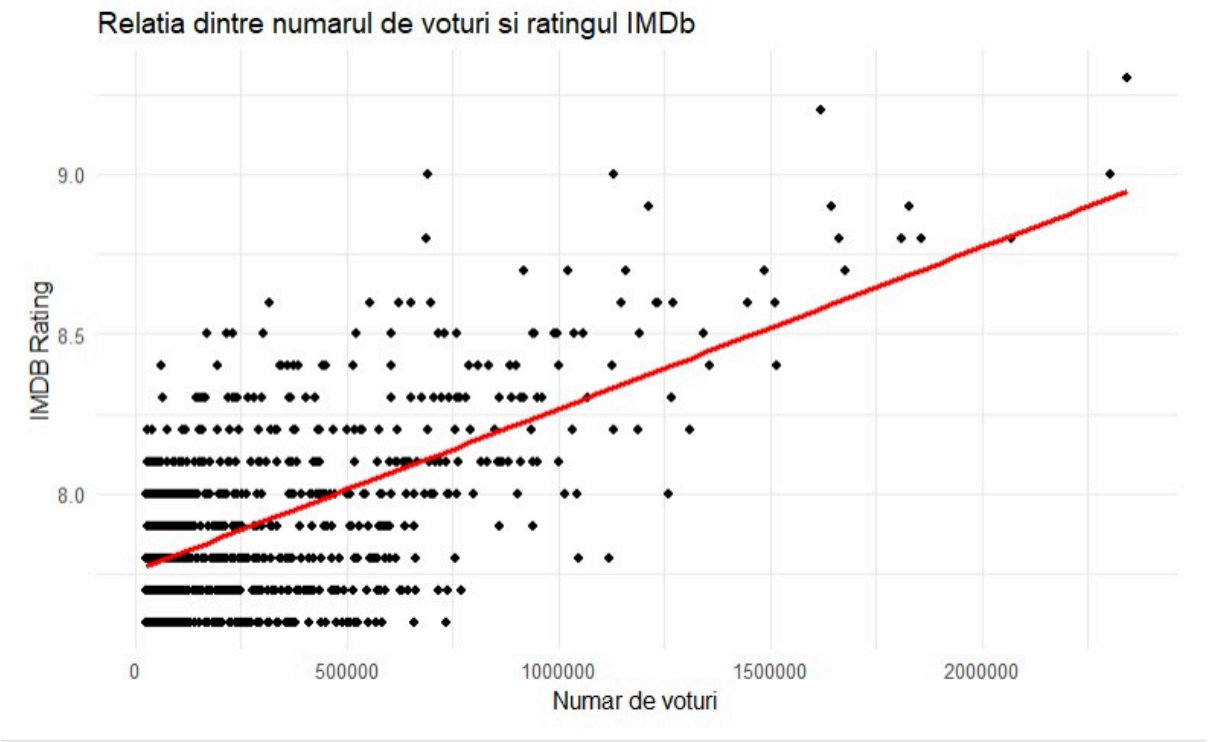
Rezultatul arata ca exista o relatie pozitiva si semnificativa intre No_of_Votes si IMDB_Rating. Coeficientul estimat pentru No_of_Votes este foarte mic, 0.0000005058, dar acest lucru se datoreaza faptului ca variabila are valori foarte mari (zeci sau sute de mii).

Interpretarea practica este ca o crestere cu 100.000 de voturi este asociata, in medie, cu o crestere de aproximativ 0.0505 puncte in ratingul IMDb.

Valoarea p asociata coeficientului este extrem de mica ($p < 0.001$), ceea ce indica faptul ca efectul este semnificativ din punct de vedere statistic.

Valoarea R-squared este 0.3771, ceea ce inseamna ca aproximativ 37.7% din variatia ratingului IMDb este explicata de numarul de voturi. Comparativ cu modelul precedent (cu Meta_score), acesta ofera o capacitate predictiva semnificativ mai buna.

```r
ggplot(movies, aes(x = No_of_Votes, y = IMDB_Rating)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  theme_minimal() +
  labs(title = "Relatia dintre numarul de voturi si ratingul IMDb",
       x = "Numar de voturi",
       y = "IMDB Rating")
```



Graficul evidentiaza relatia dintre No_of_Votes si IMDB_Rating pentru filmele din setul analizat. Fiecare punct din grafic reprezinta un film, pozitionat in functie de cele doua variabile.

Linia de regresie rosie arata o tendinta pozitiva, ceea ce sugereaza ca filmele cu un numar mai mare de voturi tind, in medie, sa aiba un rating mai ridicat pe IMDb. Cu alte cuvinte, popularitatea unui film, masurata prin numarul de voturi, este corelata cu evaluarea sa de catre public.

Dispersia punctelor este insa destul de mare, mai ales pentru valorile mici ale numarului de voturi, ceea ce indica o variabilitate considerabila in rating. Chiar daca relatia este semnificativa statistic, nu toate filmele cu multe voturi au neaparat un scor ridicat si invers.

Aceasta vizualizare confirma ca numarul de voturi este un bun predictor pentru ratingul IMDb, dar nu singurul factor care il influenteaza.

3) A treia data vedem in functie de Gross :

```
model_castig <- lm(IMDB_Rating~Gross,data=movies)
summary(model_castig)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.50884 -0.21778 -0.02657  0.16924  1.38040

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.910e+00  1.256e-02 629.627  < 2e-16 ***
Gross       3.261e-10  9.251e-11   3.525 0.000449 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2869 on 748 degrees of freedom
Multiple R-squared:  0.01634,   Adjusted R-squared:  0.01503
F-statistic: 12.43 on 1 and 748 DF,  p-value: 0.0004485
```

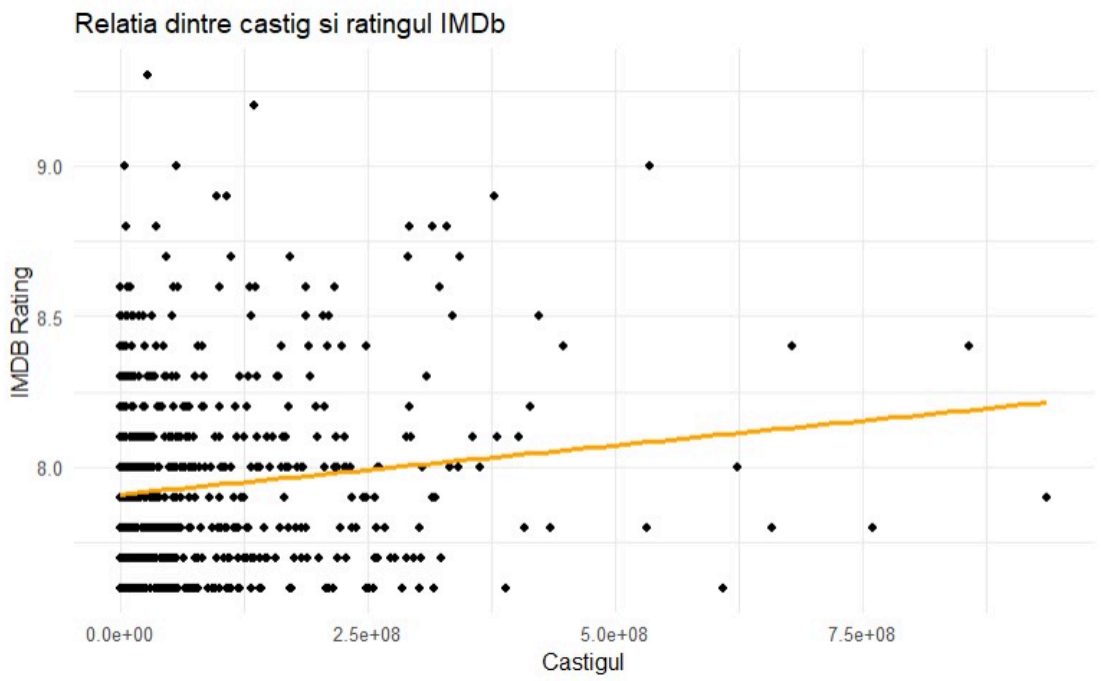
Rezultatul arata ca exista o relatie pozitiva si semnificativa intre incasarile brute ale filmului (Gross) si ratingul IMDb (IMDB_Rating). Coeficientul estimat este 0.00000003261, o valoare foarte mica, din cauza ca variabila Gross are valori foarte mari (zeci sau sute de milioane).

Interpretarea practica este ca o crestere cu 100 de milioane USD a incasarilor brute este asociata, in medie, cu o crestere de aproximativ:

Asta inseamna ca, desi relatia este semnificativa statistic ($p < 0.001$), efectul real este foarte slab.

Valoarea R-squared este 0.0163, ceea ce inseamna ca doar aproximativ 1.6% din variatia ratingului IMDb este explicata de Gross. Prin urmare, incasarile comerciale nu reprezinta un bun predictor pentru ratingul IMDb in acest model simplu.

```
ggplot(movies, aes(x = Gross, y = IMDB_Rating)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "orange") +
  theme_minimal() +
  labs(title = "Relatia dintre castig si ratingul IMDb",
       x = "Castigul",
       y = "IMDB Rating")
```



Graficul evidentiaza relatia dintre Gross si IMDB_Rating. Fiecare punct din grafic reprezinta un film, pozitionat in functie de valoarea incasarilor si a ratingului primit din partea publicului.

Linia de regresie portocalie are o panta usor pozitiva, ceea ce sugereaza o tendinta slaba de crestere a ratingului odata cu marirea incasarilor. Cu alte cuvinte, exista o usoara asociere intre succesul financiar si aprecierea din partea publicului.

Totusi, punctele sunt foarte raspandite in jurul liniei, iar distributia este vizibil concentrata in zona filmelor cu incasari mici. Acest lucru indica faptul ca relatia este foarte slaba, iar in general, incasarile nu sunt un bun predictor pentru ratingul IMDb.

Desi semnificativa din punct de vedere statistic, aceasta nu este un predictor bun pentru rating, asa cum se poate observa si din forma aproape orizontala a liniei de regresie.

Modelul de regresie liniara multipla:

$y = f(x_1,...,x_p) + \varepsilon$

$\|$

$\mathbf{E}[Y | x_1,...,x_p]$

Regresia liniară multiplă are ca scop modelarea relației dintre o variabilă X și cel puțin doi predictori. Considerăm cazul general cu p variabile explicative, notate $X_1,...,X_p$.

Modelul regresiei presupune că variabila de interes Y poate fi exprimată ca o combinație liniară a predictorilor, plus o componentă aleatoare ε. Relația are forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

unde $\beta_0,\beta_1,\ldots,\beta_p$ sunt parametrii regresiei, iar $\varepsilon \sim N(0,\sigma^2)$ reprezinta eroarea aleatoare asociata predictiei. Fie datele observate:

$\{(x_{i1},...,x_{ip},y_i)\}$, i=1,2,...,n

Pentru fiecare observatie i, modelul poate fi scris astfel:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

Ipotezele modelului : $\mathbf{E}[\varepsilon] = 0$ si $Var(\varepsilon)=\sigma^2 I_n$, rang(x)=p+1

Vrem sa obtinem un estimator folosind MSE:

$$\widehat{\beta} = \text{argmin}\frac{1}{n}\sum_{i=1}^{n} y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})$$

In populatie:

$$\text{argmin } \mathbf{E}[Y-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip})]$$

In esantion:

$$\text{argmin}\frac{1}{n}\sum_{i=1}^{n} y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \Rightarrow$$

$$\widehat{\beta}=\text{argmin}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 = \text{argmin}(Y - X\beta)^T(Y - X\beta)$$

Daca modelul de regresie este valid si ipotezele sunt adevarate, atunci se obtine estimatorul $\widehat{\beta} = (X^T X)^{-1}(X^T Y)$

Exemple:

1) In functie de Meta_score si No_of_Votes

```
model1 <- lm(IMDB_Rating~Meta_score+No_of_Votes,data=movies)
summary(model1)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.56837 -0.15133 -0.02171  0.14869  0.77156

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.267e+00  4.962e-02  146.46   <2e-16 ***
Meta_score  6.395e-03  6.259e-04   10.22   <2e-16 ***
No_of_Votes 5.035e-07  2.228e-08   22.60   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.214 on 747 degrees of freedom
Multiple R-squared:  0.4535,    Adjusted R-squared:  0.4521
F-statistic:   310 on 2 and 747 DF,  p-value: < 2.2e-16
```

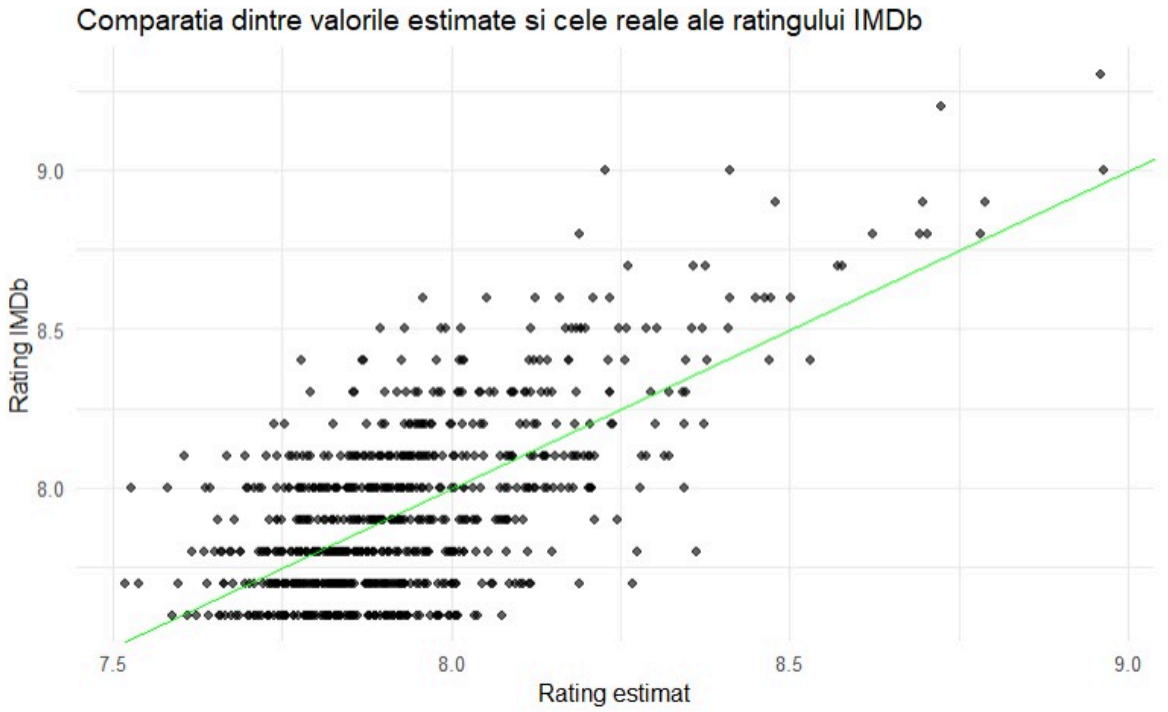
Modelul include doua variabile explicative: Meta_score si No_of_Votes. Ambele coeficiente sunt pozitive si semnificative din punct de vedere statistic ($p < 0.001$), ceea ce arata ca atat aprecierea criticilor, cat si popularitatea filmului au o influenta reala asupra ratingului IMDb.

Coeficientul pentru Meta_score este aproximativ 0.0064, ceea ce inseamna ca o crestere cu 1 punct in scorul Metacritic este asociata cu o crestere medie de 0.0064 puncte in ratingul IMDb, tinand constant numarul de voturi. Coeficientul pentru No_of_Votes este foarte mic (≈ 0.0000005035), dar semnificativ, si indica o crestere lenta, dar constanta, a ratingului pe masura ce un film primeste mai multe voturi.

Valoarea R-squared este 0.4535, ceea ce inseamna ca aproximativ 45.4% din variatia ratingului IMDb este explicata de acest model. Aceasta performanta este considerabil mai buna decat cea a modelelor simple construite anterior.

In concluzie, modelul combinat ofera o estimare mult mai buna a ratingului IMDb decat fiecare variabila luata separat si arata ca atat perceptia criticilor, cat si nivelul de expunere al filmului (masurat prin voturi) contribuie la evaluarea finala a publicului.

```r
#Cream o noua coloana cu valorile estimate de model
movies$predicted1 <- predict(model1)
#Vizualizam valorile estimate vs cele reale
ggplot(movies, aes(x = predicted1, y = IMDB_Rating)) +
  geom_point() +
  geom_abline( method = "lm", se = FALSE, color = "green") +
  theme_minimal() +
  labs(title = "Comparatia dintre valorile estimate si cele reale ale ratingului IMDb",
       x = "Rating estimat",
       y = "Rating IMDb")
```



Graficul ilustreaza comparatia dintre valorile estimate de modelul de regresie multipla (Meta_score si No_of_Votes) si valorile reale ale ratingului IMDb pentru fiecare film din setul de date. Fiecare punct din grafic reprezinta un film, pozitionat in functie de valoarea estimata (axa X) si cea reala (axa Y).

Linia verde diagonala reprezinta situatia ideala in care modelul prezice perfect ratingul – adica valoarea estimata este exact egala cu cea reala.

Distributia punctelor in jurul acestei linii arata ca modelul functioneaza destul de bine. Majoritatea punctelor sunt grupate aproape de linie, ceea ce inseamna ca diferentele dintre valorile estimate si cele reale sunt in general mici. Totusi, exista si puncte mai indepartate, ceea ce sugereaza ca pentru unele filme predictia este mai putin precisa.

In ansamblu, graficul confirma rezultatele numerice ale modelului: un R-squared de aproximativ 0.45, ceea ce indica o capacitate buna de predictie. Vizual, modelul reuseste sa surprinda tendinta generala a ratingului IMDb in functie de popularitatea si aprecierea critica a filmului.

2) In functie de Meta_score si Gross

```
model2 <- lm(IMDB_Rating~Meta_score+Gross,data=movies)
summary(model2)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.53764 -0.20132 -0.05216  0.13471  1.36458

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.395e+00  6.364e-02 116.184  < 2e-16 ***
Meta_score  6.638e-03  8.042e-04   8.254 6.91e-16 ***
Gross       3.485e-10  8.866e-11   3.931 9.25e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2749 on 747 degrees of freedom
Multiple R-squared:  0.09856,   Adjusted R-squared:  0.09615
F-statistic: 40.84 on 2 and 747 DF,  p-value: < 2.2e-16
```

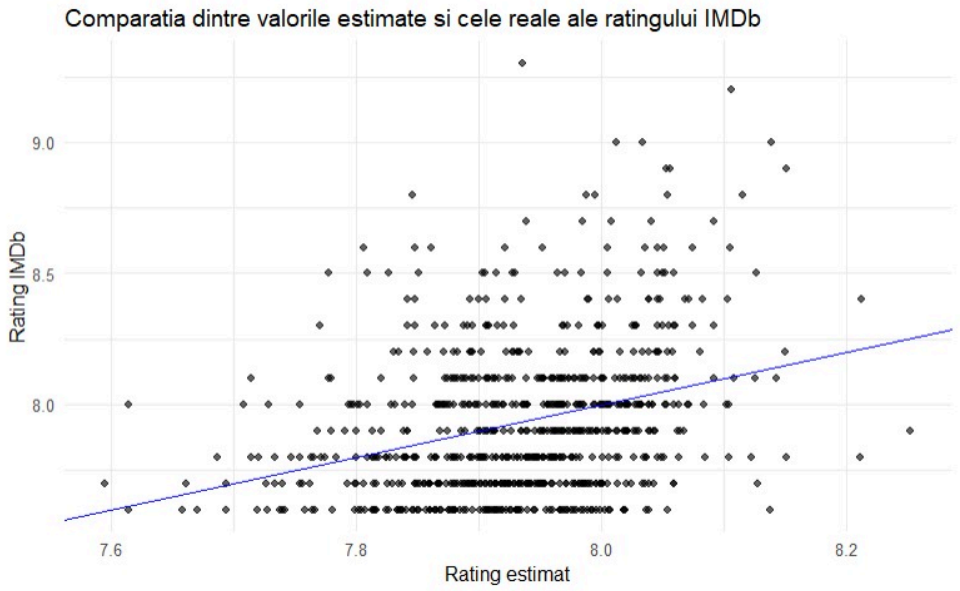
Modelul de regresie multipla foloseste ca variabile explicative Meta_score si incasarile filmului (Gross) pentru a prezice IMDB_Rating. Ambele variabile au coeficienti pozitivi si sunt semnificativi statistic ($p < 0.001$), ceea ce inseamna ca fiecare contribuie la explicarea ratingului IMDb.

Coeficientul pentru Meta_score este aproximativ 0.0066, ceea ce sugereaza ca o crestere cu 1 punct in scorul Metacritic este asociata cu o crestere medie de 0.0066 puncte in ratingul IMDb, presupunand ca incasarile raman constante.

Coeficientul pentru Gross este foarte mic (≈ 0.0000003485), dar semnificativ. Din punct de vedere practic, o crestere cu 100 de milioane USD in incasari este asociata cu o crestere de aproximativ 0.03485 puncte in ratingul IMDb.

Valoarea R-squared este 0.09856, ceea ce inseamna ca doar aproximativ 9.9% din variatia ratingului IMDb este explicata de acest model. Prin comparatie cu modelul precedent care includea No_of_Votes, performanta acestui model este mai slaba.

```r
movies$predicted2 <- predict(model2)
#Vizualizam valorile estimate vs rating
ggplot(movies, aes(x = predicted2, y = IMDB_Rating)) +
  geom_point() +
  geom_abline(,method = "lm", se = FALSE, color = "blue") +
  theme_minimal() +
  labs(title = "Comparatia dintre valorile estimate si cele reale ale ratingului IMDb",
       x = "Rating estimat",
       y = "Rating IMDb")
```



Graficul ilustreaza comparatia dintre valorile estimate de modelul de regresie multipla care foloseste Meta_score si Gross si valorile reale ale ratingului IMDb. Fiecare punct reprezinta un film, cu ratingul estimat pe axa X si ratingul observat pe axa Y.

In acest caz, distributia punctelor este mult mai dispersata comparativ cu modelul anterior (cu No_of_Votes). Multe puncte sunt grupate departe de linia diagonala, in special in zona ratingurilor mai scazute. Aceasta imprastiere sugereaza ca modelul 2 nu reuseste sa aproximeze foarte bine valorile reale ale ratingului IMDb.

Aceasta observatie este in concordanta cu rezultatele numerice ale modelului, care are un R-squared de doar 0.098, ceea ce indica o putere explicativa redusa. Chiar daca ambele variabile (Meta_score si Gross) sunt semnificative statistic, predictiile generate de acest model sunt considerabil mai slabe decat cele obtinute in modelul anterior (Meta_score + No_of_Votes).

3)In functie de No_of_Votes si Gross

```
model3 <- lm(IMDB_Rating~No_of_Votes+Gross,data=movies)
summary(model3)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.48085 -0.16063 -0.01398  0.15463  0.78470

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.772e+00  1.108e-02 701.462   <2e-16 ***
No_of_Votes  6.470e-07  2.705e-08  23.916   <2e-16 ***
Gross       -7.868e-10  8.378e-11  -9.392   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2161 on 747 degrees of freedom
Multiple R-squared:  0.4429,    Adjusted R-squared:  0.4414
F-statistic: 296.9 on 2 and 747 DF,  p-value: < 2.2e-16
```

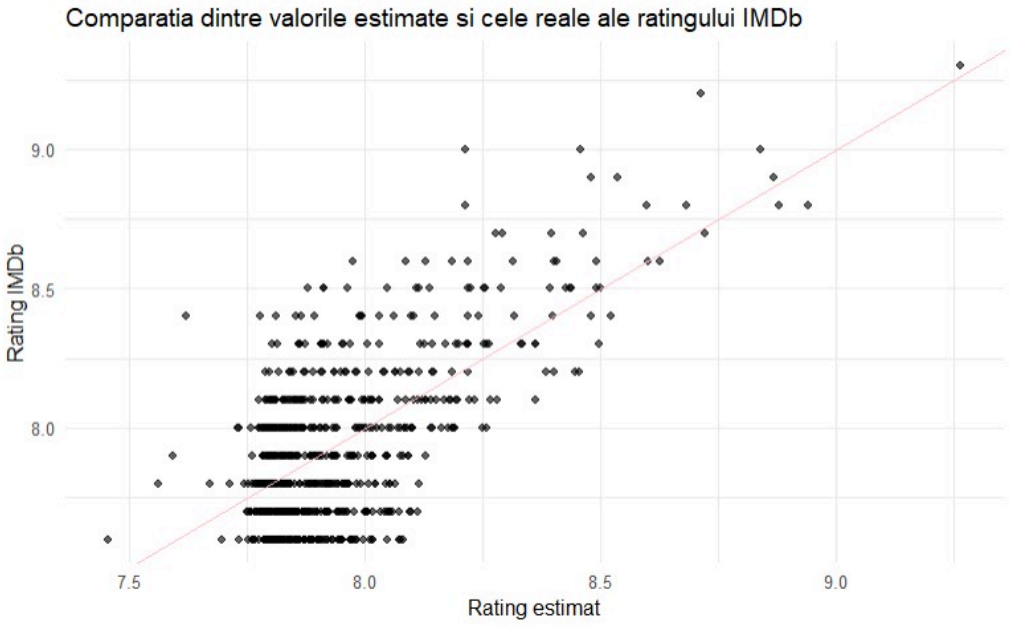
Modelul 3 analizeaza influenta No_of_Votes si a incasarilor (Gross) asupra IMDB_Rating. Ambele variabile sunt semnificative statistic ($p<0.001$), ceea ce inseamna ca ele au un impact real asupra variabilei tinta.

Coeficientul pentru No_of_Votes este pozitiv (≈ 0.000000647), indicand ca un numar mai mare de voturi este asociat cu un rating mai mare. Practic, o crestere cu 100.000 de voturi este asociata cu o crestere de aproximativ 0.065 puncte in ratingul IMDb.

In mod interesant, coeficientul pentru Gross este negativ (≈ -0.0000000787), ceea ce inseamna ca, atunci cand se tine cont si de numarul de voturi, o crestere a incasarilor este asociata cu o scadere usoara a ratingului. Acest rezultat poate reflecta faptul ca unele filme comerciale cu incasari mari nu sunt neaparat bine cotate de public.

Valoarea R-squared este 0.4429, foarte apropiata de cea a modelului 1 (Meta_score + No_of_Votes, care avea 0.4535). Asta inseamna ca acest model explica aproximativ 44.3% din variatia ratingului IMDb, ceea ce il face un model performant.

```r
movies$predicted3 <- predict(model3)
#Vizualizam valorile estimate vs rating
ggplot(movies, aes(x = predicted3, y = IMDB_Rating)) +
  geom_point() +
  geom_abline(method = "lm", se = FALSE, color = "pink") +
  theme_minimal() +
  labs(title = "Comparatia dintre valorile estimate si cele reale ale ratingului IMDb",
       x = "Rating estimat",
       y = "Rating IMDb")
```



Graficul afiseaza comparatia dintre valorile estimate de modelul 3 si valorile reale ale ratingului IMDb pentru fiecare film.

Distributia punctelor arata ca modelul surprinde bine tendinta generala a ratingului IMDb. Majoritatea observatiilor sunt aliniate in apropierea liniei diagonale, ceea ce semnaleaza o buna potrivire intre estimari si valori reale. Totusi, exista si o zona compacta de puncte sub ratingul 8.0 unde predictiile sunt mai putin precise.

Aceasta vizualizare este in acord cu performanta numerica a modelului (R-squared ≈ 0.44), ceea ce arata ca No_of_Votes si Gross explica o parte considerabila din variatia ratingului IMDb. Totusi, modelul nu este perfect si poate fi imbunatatit prin adaugarea altor predictori.

4) In functie de Meta_score, No_of_Votes si Gross

```r
model4 <- lm(IMDB_Rating~Meta_score+No_of_Votes+Gross,data=movies)
summary(model4)
```

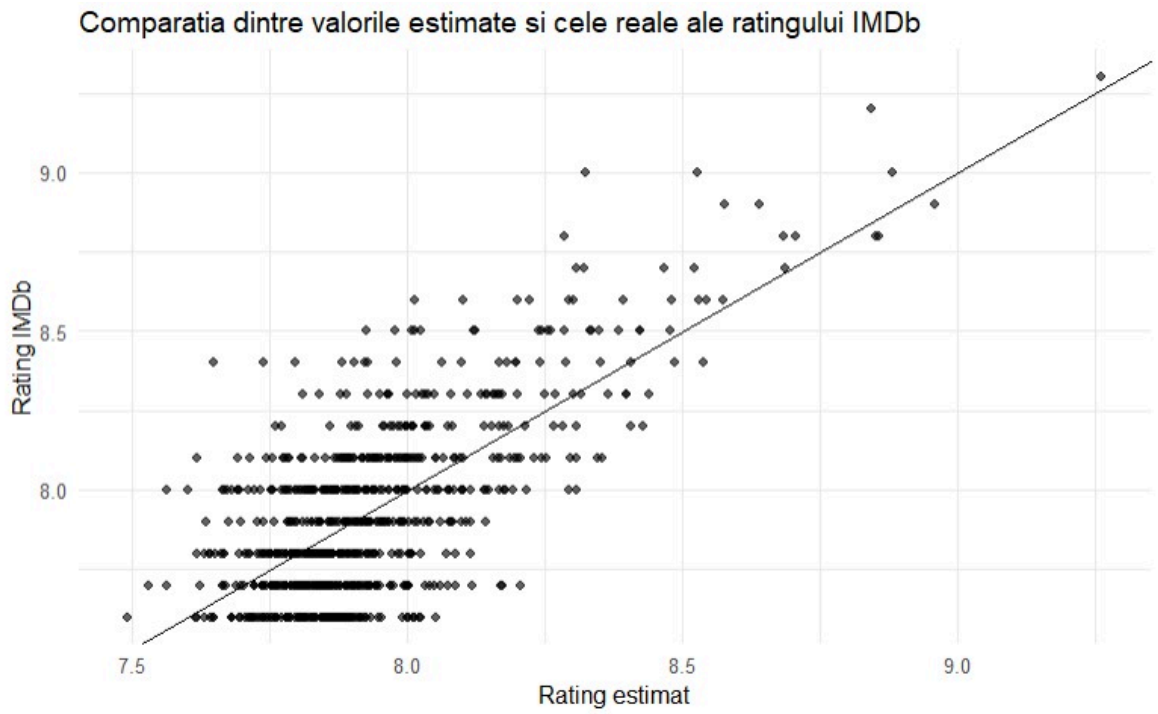
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.50665 -0.14062 -0.02458  0.13393  0.75178

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.296e+00  4.695e-02 155.394   <2e-16 ***
Meta_score   6.148e-03  5.916e-04  10.392   <2e-16 ***
No_of_Votes  6.383e-07  2.531e-08  25.217   <2e-16 ***
Gross       -7.512e-10  7.843e-11  -9.578   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2021 on 746 degrees of freedom
Multiple R-squared:  0.5134,    Adjusted R-squared:  0.5114
F-statistic: 262.3 on 3 and 746 DF,  p-value: < 2.2e-16
```

Modelul 4 foloseste toate cele trei variabile explicative – Meta_score, No_of_Votes si Gross – pentru a prezice scorul IMDb al unui film. Toti coeficientii estimati sunt pozitivi sau negativi semnificativi din punct de vedere statistic, cu valori p extrem de mici, ceea ce inseamna ca fiecare variabila contribuie semnificativ la model.

Coeficientul pentru Meta_score este de aproximativ 0.00615. Acesta sugereaza ca, atunci cand celelalte variabile sunt mentinute constante, o crestere cu un punct in scorul Metacritic determina in medie o crestere de 0.0061 puncte in ratingul IMDb. Numarul de voturi are un coeficient estimat de aproximativ 0.000000683, ceea ce inseamna ca o crestere cu 100.000 de voturi duce, in medie, la o crestere de 0.0683 in scorul IMDb. Variabila Gross, care masoara incasarile brute, are un coeficient negativ de aproximativ -0.0000000751. Acest rezultat arata ca, atunci cand sunt controlate celelalte doua variabile, o crestere a incasarilor este asociata cu o usoara scadere a ratingului IMDb. Aceasta relatie inversa poate reflecta faptul ca succesul comercial nu garanteaza neaparat o apreciere crescuta din partea publicului.

Modelul obtinut are o valoare R-squared de aproximativ 0.513, ceea ce inseamna ca explica peste 51% din variabilitatea ratingului IMDb. Acesta este cel mai performant model dintre toate cele testate, oferind o capacitate predictiva superioara in comparatie cu modelele anterioare care foloseau doar una sau doua dintre aceste variabile.

```r
movies$predicted4 <- predict(model4)
#Vizualizam valorile estimate vs rating
ggplot(movies, aes(x = predicted4, y = IMDB_Rating)) +
  geom_point() +
  geom_abline(method = "lm", se = FALSE, color = "black") +
  theme_minimal() +
  labs(title = "Comparatia dintre valorile estimate si cele reale ale ratingului IMDb",
       x = "Rating estimat",
       y = "Rating IMDb")
```



Graficul arata comparatia dintre valorile prezise de modelul 4 si ratingurile reale IMDb pentru filmele analizate.

Se observa ca majoritatea punctelor sunt concentrate de-a lungul acestei linii, ceea ce indica o potrivire buna intre predictiile modelului si valorile reale. Fata de modelele anterioare, dispersia este mai redusa si punctele tind sa se alinieze mai strans, ceea ce confirma si performanta ridicata a modelului (R-squared ≈ 0.51).

Exista totusi o aglomerare de filme in zona ratingurilor mai mici (sub 8), unde modelul nu reuseste intotdeauna sa faca predictii foarte precise. Insa, per ansamblu, graficul arata clar ca modelul 4 este cel mai performant dintre toate cele analizate.

Concluzii:

Scopul acestui proiect a fost sa analizam care sunt factorii care influenteaza ratingul IMDb al unui film si daca putem construi modele de regresie capabile sa prezica acest rating pe baza unor variabile precum scorul criticilor (Meta_score), numarul de voturi primite (No_of_Votes) si incasarile brute (Gross).

Analiza exploratorie a aratat ca cele mai frecvente genuri sunt drama, comedie si actiune, iar majoritatea filmelor din setul de date au ratinguri IMDb cuprinse intre 7.5 si 8.5.

Am construit mai multe modele de regresie liniara, atat simple, cat si multiple. Modelele simple au aratat ca No_of_Votes este cel mai puternic predictor individual al ratingului IMDb, explicand aproximativ 37.7% din variatie. Scorul criticilor si incasarile au avut contributii semnificative, dar mai slabe luate individual.

Modelele multiple au confirmat ca predictia este imbunatatita atunci cand combinam variabile. Modelul final, care foloseste toate cele trei predictori (Meta_score, No_of_Votes si Gross), a reusit sa explice aproximativ 51.3% din variatia ratingului IMDb. Acesta este cel mai performant model testat, avand coeficienti semnificativi statistic si o potrivire vizuala buna intre valorile estimate si cele reale.

In concluzie, acest proiect ne-a oferit o perspectiva clara asupra modului in care anumite caracteristici ale filmelor se coreleaza cu scorurile primite pe IMDb. Rezultatele obtinute pot servi ca baza pentru intelegerea mai profunda a factorilor care contribuie la evaluarea unui film, oferind indicii utile pentru viitoare studii sau decizii in industrie, dar fara pretentia unei previziuni exacte.