Facultatea de Matematică şi Informatică,
Universitatea din Bucureşti

Proiect Modele de regresii

MODELE DE REGRESIE LOGISTICĂ

- simplă și multiplă -
- aspecte teoretice și exemple în R -

Prof. Alexandru Amărioarei

Studenți:
Mihalache Diana, grupa 321
Băjan Ionica-Mariana, grupa 322

Matematici Aplicate,
Semestrul II, anul 2024

1. Introducere

1.1 Motivația

Motivația proiectului este de a identifica și explora relația dintre dimensiunea orașului și performanța academică a copiilor și tinerilor. Prin utilizarea modelelor de regresie logistică și analiza datelor disponibile, proiectul își propune să identifice și să evalueze factorii care ar putea explica de ce copiii și tinerii din orașele mai mici obțin rezultate academice superioare celor din orașele mai mari.

Proiectul este structurat în 4 capitole, în primul prezentându-se setul de date, și o scurtă analiză exploratorie a acestuia. În capitolul al doilea, sunt prezentate noțiunile teoretice ale regresiilor logistice simple și multiple. În ultimul capitol, sunt evidențiate noțiunile teoretice prin intermediul exemplelor, folosindu-se limbajul R, și sunt interpretate rezultatele finale.

1.2 Setul de date

Această lucrare urmărește prezentarea teoretică a modelelor de regresie logistică, simplă și multiplă. Conceptele vor fi exemplificate cu ajutorul setului de date, din partea Biroului Național de Statistică din UK, folosit în lansarea: "Why do children and young people in smaller towns do better academically than those in larger towns?", publicat în iulie 2023.

Setul de date conține observații structurate în 1104 de rânduri și 31 de coloane (variabile) și reprezintă un studiu în care se analizează diferite aspecte legate de nivelul de educație, activitățile la vârsta de 19 ani și clasificările socio-economice ale diferitelor orașe și localități. Se examinează, printre altele, proporția de elevi care au atins anumite niveluri de calificare la vârste specifice, activitățile de învățare și angajare la vârsta de 19 ani, precum și clasificările legate de venituri și educație ale diferitelor zone geografice.

1.3. Analiza exploratorie

Avem următoarele variabile:

TOWN11CD	Codul orașului
TOWN11NM	Numele orașului
POPULATION_2011	Masura populatiei rezidente obișnuite în oraș
SIZE_FLAG	Categoria de mărime a zonei construite
RGN11NM	Nume englezesc al regiunii

COASTAL	Variabilă folosită pentru a descrie orașele ca fiind de coastă sau nu
COASTAL DETAILED	Orașele de coastă împărțite în funcție de dimensiune și de orașe de pe litoral și alte orașe de coastă (fără litoral)
TTWA11CD	Cod zona de călătorie la locul de muncă
TTWA11NM	Numele zonei de călătorie la locul de muncă
TTWA CLASSIFICATION	Clasificarea deplasării la locul de muncă
JOB DENSITY FLAG	Variabilă utilizată pentru a descrie orașele ca fiind active, rezidențiale sau mixte.
INCOME FLAG	Variabilă utilizată pentru a descrie orașele ca fiind cu lipsuri de venituri mai mici
UNIVERSITY FLAG	Variabilă utilizată pentru a descrie dacă orașul are o universitate
Level4Qual_residents35-64_2011	Proporția rezidenților orașului cu vârsta cuprinsă între 35-64 de ani cu o calificare de Nivel 4 sau mai mare.
KS4_2012-2013_counts	Numărul de elevi din oraș
Key Stage 2 attainment_school year 2007 to 2008	Proporția elevilor care au atins nivelul 4 sau mai mult (nivelul așteptat) în etapa cheie 2 la engleză și matematică în anul școlar 2007-2008

Key Stage 4 attainment_school year 2012 to 2013	Proporția elevilor care au obținut 5 GCSE sau mai mult, inclusiv engleză și matematică, cu notele A*-C în anul școlar 2012-2013
Level 2 at age 18	Proporția din grupul cheie din etapa 4 a orașului în 2012/13 care a obținut calificări de nivel 2 la vârsta de 18 ani.
Level 3 at age 18	Proporția din grupul cheie din etapa 4 a orașului 2012/13 care a obținut calificări de nivel 3 la vârsta de 18 ani.
Activity at age 19: full-time higher education	studii superioare cu normă întreagă la vârsta de 19 ani.
Activity at age 19: sustained further education	Educație continuă susținută la 19 ani
Activity at age 19: apprenticeships	Proporție a orașului din etapa 4 din 2012/13 într-o ucenicie la vârsta de 19 ani.
Activity at age 19: employment with earnings above £0	Angajare cu câștiguri de peste 0 GBP la vârsta de 19 ani.
Activity at age 19: employment with earnings above £10,000	Angajare cu câștiguri de peste 10.000 GBP la vârsta de 19 ani.
Activity at age 19: out-of-work	Șomer la vârsta de 19 ani.

Highest level qualification achieved by age 22: less than level 1	Calificare de cel mai înalt nivel atins până la vârsta de 22 de ani,cu mai puțin de o calificare de nivel 1.
Highest level qualification achieved by age 22: level 1 to level 2	Cel mai înalt nivel de calificare atins până la vârsta de 22 de ani: o calificare de nivel 1 sau de nivel 2
Highest level qualification achieved by age 22: level 3 to level 5	Cel mai înalt nivel de calificare atins până la vârsta de 22 de ani: cu calificare de nivel 3, nivel 4 sau nivel 5 .
Highest level qualification achieved by age 22: level 6 or above	Cel mai înalt nivel de calificare atins până la vârsta de 22 de ani: cu calificare de nivel 6 sau mai mare.
Highest level qualification achieved b age 22: average score	Cel mai înalt nivel de calificare obținut la vârsta de 22 de ani: scor mediu oraș
EDUCATION SCORE	Scorul de educație al orașului

```
data <- read_csv("english_education.csv")
summary(data)
```

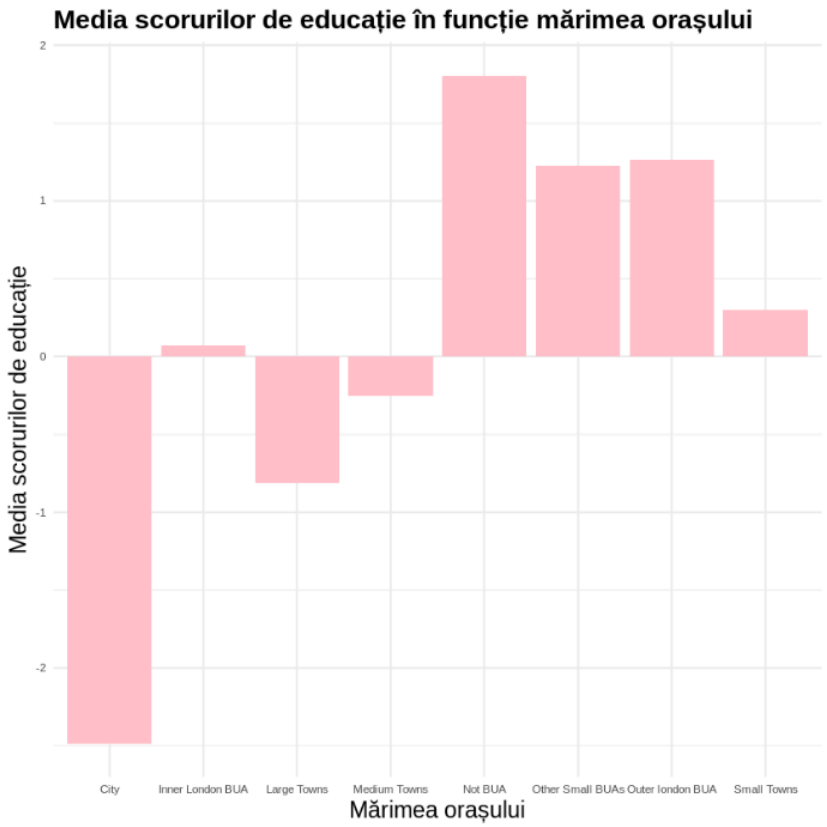
```
  town11cd           town11nm         population_2011    size_flag
Length:1104        Length:1104        Min.   :   5003   Length:1104
Class :character   Class :character   1st Qu.:   8076   Class :character
Mode  :character   Mode  :character   Median :  15436   Mode  :character
                                      Mean   :  33347
                                      3rd Qu.:  32722
                                      Max.   :1085810
                                      NA's   :4
  rgn11nm            coastal          coastal_detailed     ttwa11cd
Length:1104        Length:1104        Length:1104        Length:1104
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character


  ttwa11nm         ttwa_classification job_density_flag   income_flag
Length:1104        Length:1104         Length:1104        Length:1104
Class :character   Class :character    Class :character   Class :character
Mode  :character   Mode  :character    Mode  :character   Mode  :character


university_flag    level4qual_residents35_64_2011 ks4_2012_2013_counts
Length:1104        Length:1104                    Min.   :   14.0
Class :character   Class :character               1st Qu.:   92.0
Mode  :character   Mode  :character               Median :  172.5
                                                  Mean   :  511.6
                                                  3rd Qu.:  374.2
                                                  Max.   :59743.0

key_stage_2_attainment_school_year_2007_to_2008
Min.   :28.09
1st Qu.:68.69
Median :74.21
Mean   :74.07
3rd Qu.:80.02
Max.   :98.63

key_stage_4_attainment_school_year_2012_to_2013 level_2_at_age_18
Min.   :33.33                                   Min.   :56.00
1st Qu.:54.41                                   1st Qu.:79.40
Median :60.85                                   Median :83.65
Mean   :61.30                                   Mean   :83.62
3rd Qu.:67.61                                   3rd Qu.:88.40
Max.   :92.86                                   Max.   :99.35
```

```
level_3_at_age_18 activity_at_age_19_full_time_higher_education
Min.   :16.54     Min.   : 7.874
1st Qu.:41.41     1st Qu.:26.316
Median :48.52     Median :32.394
Mean   :49.39     Mean   :33.531
3rd Qu.:56.74     3rd Qu.:39.441
Max.   :85.71     Max.   :73.446
                  NA's   :1
activity_at_age_19_sustained_further_education
Min.   : 7.353
1st Qu.:17.161
Median :20.375
Mean   :20.769
3rd Qu.:23.915
Max.   :47.561
NA's   :58
activity_at_age_19_appprenticeships
Min.   : 2.997
1st Qu.:10.714
Median :13.131
Mean   :13.661
3rd Qu.:16.154
Max.   :39.024
NA's   :195

activity_at_age_19_employment_with_earnings_above_0
Min.   :28.96
1st Qu.:44.63
Median :49.29
Mean   :49.25
3rd Qu.:53.78
Max.   :71.43
NA's   :1
activity_at_age_19_employment_with_earnings_above_10_000
Min.   : 7.062
1st Qu.:20.000
Median :23.944
Mean   :24.264
3rd Qu.:28.070
Max.   :48.980
NA's   :27
activity_at_age_19_out_of_work
Min.   : 2.273
1st Qu.: 6.894
Median : 9.292
Mean   : 9.863
3rd Qu.:12.095
Max.   :24.638
NA's   :462
```

```
highest_level_qualification_achieved_by_age_22_less_than_level_1
Min.   :0.900
1st Qu.:2.000
Median :2.600
Mean   :2.763
3rd Qu.:3.300
Max.   :7.700
NA's   :859
highest_level_qualification_achieved_by_age_22_level_1_to_level_2
Min.   : 6.50
1st Qu.:21.18
Median :26.40
Mean   :26.73
3rd Qu.:31.90
Max.   :51.70
NA's   :48
highest_level_qualification_achieved_by_age_22_level_3_to_level_5
Min.   :23.00
1st Qu.:40.00
Median :43.90
Mean   :43.88
3rd Qu.:47.70
Max.   :66.10
NA's   :1

highest_level_qualification_achieved_by_age_22_level_6_or_above
Min.   :12.10
1st Qu.:23.50
Median :28.00
Mean   :29.55
3rd Qu.:34.60
Max.   :61.40
NA's   :231
highest_level_qualification_achieved_b_age_22_average_score education_score
Min.   :2.567                                               Min.   :-10.0280
1st Qu.:3.264                                               1st Qu.: -2.5707
Median :3.479                                               Median : -0.2824
Mean   :3.514                                               Mean   :  0.0000
3rd Qu.:3.725                                               3rd Qu.:  2.2265
Max.   :5.143                                               Max.   : 11.8715
```

Următoarea figură reprezintă un grafic ce oferă o imagine generală a modului în care scorurile de educație variază în funcție de mărimea orașului. Astfel putem concluziona că setul de date susține ipoteza conform căreia orașele mai mici au o realizare educațională mai bună.

```r
# media scorurilor de educație pentru fiecare dimensiune de oraș
education_scores_summary <- relevant_data %>%
  group_by(size_flag) %>%
  summarise(avg_education_score = mean(education_score))
# graficul
ggplot(education_scores_summary, aes(x = size_flag, y = avg_education_score)) +
  geom_bar(stat = "identity", fill = "pink") +
  labs(title = "Media scorurilor de educație în funcție mărimea orașului",
       x = "Mărimea orașului",
       y = "Media scorurilor de educație") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme_minimal() +
  theme(plot.title = element_text(size = 16, face = "bold"),
        axis.title = element_text(size = 14),
        axis.text = element_text(size = 7))
```



2. Metodologie

2.1. Regresia logistică simplă

Regresia logistică simplă este o metodă statistică utilizată pentru a analiza relația dintre o variabilă dependentă binară (răspuns) și o variabilă independentă (explicativă). Acest model este folosit în special atunci când variabila dependentă este de tip binar. În cazul nostru, pe baza setului de date ales și a ipotezei de la care am pornit, obținem ca variabilă răspuns: (1 sau 0) elevii depășesc un anumit prag educațional și astfel putem analiza cum elevii din orașele mai mici se descurcă mai bine din punct de vedere academic decât cei din orașele mai mari.

Scopul este de a găsi o relație între variabila independentă și probabilitatea ca variabila dependentă să fie într-una din cele două categorii, adică noi vrem să prezicem probabilitatea ca o observație să fie într-una dintre cele două categorii.

Formula regresiei logistice simple este:

$P(Y=1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1*X)}}$, unde:

- P(Y=1|X) reprezintă probabilitatea ca variabila dependentă Y să fie 1 dată valoarea variabilei independente X.
- β_0 si β_1 sunt coeficienții modelului.

Interpretarea coeficientilor:

- Coeficienții β_0 si β_1 reprezintă efectul variabilei independente X asupra logaritmului șansei (raportul șanselor) de a fi în categoria 1.
- Un coeficient pozitiv indică o creștere a șanselor de a fi în categoria 1, în timp ce un coeficient negativ indică o scădere a acestor șanse.

Metode de evaluare a performanței modelului:

- Devianța: Măsoară cât de bine se potrivește modelul datelor observate. O devianță mai mică indică o potrivire mai bună a modelului.
- AUC-ROC (AUC - Area Under the Receiver Operating Characteristic curve): Măsoară capacitatea modelului de a distinge între cele două categorii. Cu cât AUC-ROC este mai mare (aproape de 1), cu atât modelul este mai bun la clasificare.

2.2. Regresia logistică multiplă

Regresia logistică multiplă este o extensie a regresiei logistice simple, în care sunt implicate mai multe variabile independente pentru a modela relația cu o variabilă dependentă binară. Acest model este utilizat atunci când dorim să evaluăm impactul mai multor variabile independente asupra probabilității de a fi într-una dintre cele două categorii ale variabilei dependente (răspuns).

Formula regresiei logistice multiple este similară cu cea a regresiei logistice simple, dar include mai mulți coeficienți pentru fiecare variabilă independentă:

$$P(Y=1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1*X_1+\beta_2*X_2+.....+\beta_n*X_n)}}$$, unde:

- $X_1, X_2,.....X_n$ reprezintă variabilele independente.
- $\beta_0, \beta_1, \beta_2,.....\beta_n$ sunt coeficienții corespunzători.

Interpretarea coeficienților în contextul regresiei logistice multiple:

- Coeficienții $\beta_1, \beta_2,.....\beta_n$ reprezintă efectul fiecărei variabile independente asupra logaritmului șansei (raportul șanselor) de a fi în categoria 1, ținând cont de celelalte variabile independente din model.

Diagnosticarea coliniarității și alte probleme potențiale:

- Coliniaritatea este o problemă în regresia logistică multiplă atunci când două sau mai multe variabile independente sunt puternic corelate între ele, ceea ce poate duce la probleme în interpretarea coeficienților.
- Alte probleme potențiale includ suprapunerea datelor, distribuția incorectă a erorilor, sau dimensiunea inadecvată a eșantionului.

Descrierea procesului de modelare:

1. **Definirea problemei:**
   Identificarea relației dintre dimensiunea orașului și performanța academică a copiilor și tinerilor.

2. **Colectarea și pregătirea datelor:**
   Colectarea datelor relevante referitoare la performanța academică, caracteristicile socio-economice și dimensiunea orașului.

3. **Modelarea regresiei logistice:**
   Implementarea unui model de regresie logistică simplă pentru a analiza impactul dimensiunii orașului asupra performanței academice.
   Extinderea modelului la regresia logistică multiplă pentru a lua în considerare mai mulți factori de influență.

4. **Evaluarea și interpretarea rezultatelor:**
   Evaluarea semnificației statistice a coeficienților și a calității modelului.
   Interpretarea rezultatelor pentru a înțelege modul în care dimensiunea orașului afectează performanța academică.

3. Rezultate obținute

3.1. Regresie logistică simplă în R

Pentru a modela regresia logistică simplă, folosim datele din coloanele "town11nm", "size_flag", "education_score". Regresia simplă ilustrează modul în care probabilitatea de a obține un scor educațional mai ridicat (ca variabilă rezultat) este influențată de schimbările în mărimea populației în 2011, pentru diferite orașe sau regiuni, precum: Cheltenham, Edenbridge, Pembury, Inner London BUAs[1], Outer London BUAs etc.

- **town11nm** reprezintă coloana cu numele orașelor
- **size_flag** reprezintă coloana cu diferitele categorii din care fac parte diferite orașe în funcție de populația lor
- **education_score** reprezintă coloana cu scorul indice calculat folosind datele elevilor care frecventează școala obligatorie în respectivul oraș
- **population_2011** reprezintă coloana cu măsura populației rezidente obișnuite în orașe

```
# Încărcarea setului de date
data <- read_csv("english_education.csv")
# Vizualizarea primelor câteva rânduri ale datelor selectate
relevant_data <- data %>%
  select("town11nm", "size_flag", "education_score") %>%
  head()
```

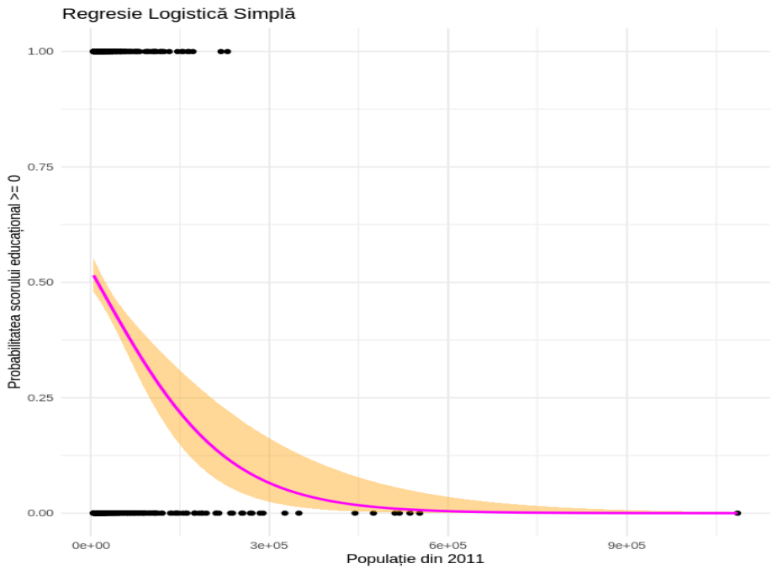
town11nm	size_flag	education_score
<chr>	<chr>	<dbl>
Carlton in Lindrick BUA	Small Towns	-0.5337504
Dorchester (West Dorset) BUA	Small Towns	1.9520187
Ely BUA	Small Towns	-1.0441280
Market Weighton BUA	Small Towns	-1.2492619
Downham Market BUA	Small Towns	-1.1690785
Penrith BUA	Small Towns	0.8451311

Figura următoare prezintă graficul regresiei logistice simple. Curba scade brusc pe măsură ce dimensiunea populației crește, începând de la o probabilitate aproape de 1 (indicând realizarea aproape sigură a pragului de educație superioară) când populația este

[1] BUA = built-up area

foarte mică. Pe măsură ce populația crește, probabilitatea scade semnificativ, apropiindu-se de zero pentru populații foarte mari. Zona umbrită în jurul curbei reprezintă intervalul de încredere al predicțiilor, oferind o reprezentare vizuală a incertitudinii în predicțiile modelului pentru diferite dimensiuni ale populației. Lățimea acestei zone indică variabilitatea predicției; zonele mai înguste sugerează o mai mare încredere în estimările modelului la acele puncte.

```
# convertirea variabilei education_score în binară
selected_data <- data %>%
  mutate(education_binary = ifelse(education_score >= 0, 1, 0))
# variabila independentă(population_2011)
logit_model_s <- glm(education_binary ~ population_2011,
                     data = selected_data,
                     family = binomial)
ggplot(selected_data, aes(x = population_2011, y = education_binary)) +
       geom_point() +  # punctele pentru fiecare observație
       geom_smooth(method="glm", method.args=list(family="binomial"),
                   se=TRUE, col="magenta", fill="orange") +
# curba de regresie logistică simplă
       labs(title = "Regresie Logistică Simplă",
            x = "Populație din 2011",
            y = "Probabilitatea scorului educațional >= 0") +
       theme_minimal()
```




```
summary(logit_model_s)# rezumatul modelului
```

```
Call:
glm(formula = education_binary ~ population_2011, family = binomial,
    data = selected_data)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.070e-01  7.866e-02   1.360    0.174
population_2011 -9.231e-06  1.887e-06  -4.893 9.91e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1517.6  on 1099  degrees of freedom
Residual deviance: 1481.0  on 1098  degrees of freedom
  (4 observations deleted due to missingness)
AIC: 1485

Number of Fisher Scoring iterations: 5
```

Aceste rezultate sunt pentru regresia logistică simplă efectuată pentru variabila 'population_2011' în cadrul setului de date selectat. Interpretarea rezultatelor este următoarea:

- **population_2011**: Coeficientul pentru variabila 'population_2011' este -9.231e-06. Acest coeficient indică schimbarea în logaritmul șanselor de a avea o scorare educațională mai mare sau egală cu pragul pentru o unitate de schimbare în populația din 2011. Deoarece coeficientul este negativ, acest lucru sugerează că o creștere în populația din 2011 este asociată cu o scădere în șansele de a avea o scorare educațională mai mare sau egală cu pragul ales.

- **Coeficientul pentru 'population_2011'** este semnificativ diferit de zero, deoarece valoarea p asociată este foarte mică (9.91e-07), ceea ce indică o asociere semnificativă între populația din 2011 și probabilitatea de a avea o scorare educațională mai mare sau egală cu pragul ales.

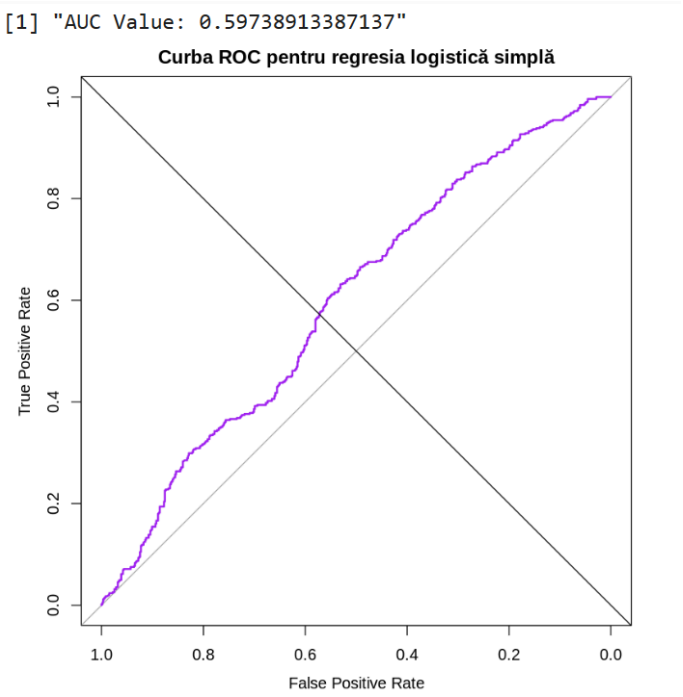
- **Devianța nulă** și **devianța reziduală** sunt utilizate pentru a evalua cât de potrivit este modelul. O devianță mai mică indică o potrivire mai bună a modelului la datele observate.

- (Criteriul de Informare Akaike): **AIC** este o măsură a calității relative a modelului, luând în considerare și complexitatea acestuia. Pentru comparație între mai multe modele, se alege modelul cu AIC cel mai mic. În acest caz, AIC este 1485.

Pentru a vizualiza performanța modelului de regresie logistică simplă, utilizăm o curbă ROC. Curba ROC este o metodă de evaluare a calității unui model de clasificare binară și ne permite să evaluăm cât de bine poate distinge între clasele pozitive și negative. AUC este o altă măsură a performanței modelului, cu o valoare AUC mai mare indicând o performanță mai bună a modelului.

Figura următoare reprezintă curba ROC, ce sugerează că modelul de regresie logistică performează bine în diferențierea între cele două clase, însă este mult de pentru îmbunătățiri .

Valoarea noastră AUC este de aproximativ 0,597 ceea ce semnalează că există loc pentru îmbunătățire.



3.2. Regresie logistică multiplă în R

Pentru a modela regresia logistică multiplă, folosim datele din coloanele prezentate mai jos. Folosim regresia logistică multiplă pentru a analiza cum variază probabilitatea de a atinge sau depăși un anumit scor de educație în funcție de aceste caracteristici socio-economice și de performanță educațională.

- **education_score** reprezintă coloana cu scorul indice calculat folosind datele elevilor care frecventează școala obligatorie în respectivul oraș
- **population_2011** reprezintă coloana cu măsura populației rezidente obișnuite în orașe
- **level4qual_residents35_64_2011** reprezintă coloana cu proporția rezidenților orașului cu vârsta cuprinsă între 35-64 de ani cu o calificare de Nivel 4 sau mai mare
- **key_stage_2_attainment_school_year_2007_to_2008** reprezintă coloana cu proporția elevilor care au atins nivelul 4 sau mai mult (nivelul așteptat) în etapa cheie 2 la engleză și matematică în anul școlar 2007-2008
- **key_stage_4_attainment_school_year_2012_to_2013** reprezintă coloana cu proporția elevilor care au obținut 5 GCSE sau mai mult, inclusiv engleză și matematică, cu notele A*-C în anul școlar 2012-2013
- **activity_at_age_19_full_time_higher_education** reprezintă coloana cu proporția persoanelor cu studii superioare cu normă întreagă la vârsta de 19 ani
- **activity_at_age_19_sustained_further_education** reprezintă coloana cu proporția persoanelor ce continuă să susțină o formă de educație până la 19 ani
- **activity_at_age_19_appprenticeships** reprezintă coloana cu proporția din totalul de elevi din etapa cheie 4 din anul 2012/2013 din oraș care urmează un program de ucenicie la vârsta de 19 ani

- **activity_at_age_19_employment_with_earnings_above_0** reprezintă coloana cu proporția din totalul de elevi din etapa cheie 4 din orașul respectiv pentru anul 2012/2013 care sunt angajați pe termen lung la vârsta de 19 ani
- **activity_at_age_19_employment_with_earnings_above_10_000** reprezintă coloana cu proporția din totalul de elevi din etapa cheie 4 din orașul respectiv pentru anul 2012/2013 care sunt angajați pe termen lung și câștigă £10,000 sau mai mult la vârsta de 19 ani
- **activity_at_age_19_out_of_work** reprezintă coloana cu proporția din totalul de elevi din etapa cheie 4 din orașul respectiv pentru anul 2012/2013 care solicită beneficii de șomaj la vârsta de 19 ani

Primul model de regresie logistică se folosește 4 caracteristici, iar cele ce influențează cel mai mult modelul sunt cele de performanță educațională.

```
# setul de date
data <- read_csv("english_education.csv")
selected_data <- data %>%
  mutate(education_binary = ifelse(education_score >= 0, 1, 0))
# regresia logistică multiplă
logit_model_m <- glm(education_binary ~ population_2011
+ level4qual_residents35_64_2011
+ key_stage_2_attainment_school_year_2007_to_2008
+ key_stage_4_attainment_school_year_2012_to_2013,
                     family = binomial, data = selected_data)
# rezumatul modelului
summary(logit_model_m)
```

```
Coefficients:
                                              Estimate Std. Error z value
(Intercept)                                 -5.028e+01  4.272e+00 -11.771
population_2011                              9.599e-07  3.009e-06   0.319
level4qual_residents35_64_2011Low           -3.563e+00  8.505e-01  -4.189
level4qual_residents35_64_2011Medium        -1.285e+00  8.339e-01  -1.541
key_stage_2_attainment_school_year_2007_to_2008  3.865e-01  3.719e-02  10.394
key_stage_4_attainment_school_year_2012_to_2013  3.823e-01  3.515e-02  10.875
                                            Pr(>|z|)
(Intercept)                                  < 2e-16 ***
population_2011                                0.750
level4qual_residents35_64_2011Low            2.8e-05 ***
level4qual_residents35_64_2011Medium           0.123
key_stage_2_attainment_school_year_2007_to_2008  < 2e-16 ***
key_stage_4_attainment_school_year_2012_to_2013  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 1517.55  on 1099  degrees of freedom
Residual deviance:  354.08  on 1094  degrees of freedom
  (4 observations deleted due to missingness)
AIC: 366.08

Number of Fisher Scoring iterations: 8
```

- Pentru variabilele **key_stage_2_attainment_school_year_2007_to_2008** și **key_stage_4_attainment_school_year_2012_to_2013**, coeficienții sunt 3.865e-01 și 3.823e-01, indicând o asociere semnificativă și pozitivă cu probabilitatea de a avea un scor educațional mai mare sau egal cu 0.
- **Devianța nulă** și **devianța reziduală** sunt 1517.55, respectiv 354.08, ceea ce indică o potrivire bună a modelului la datele observate (mai bună decât la regresia simplă).

Figura următoare reprezintă graficul cu coeficienții estimați și intervalele lor de încredere pentru predictorii din model.

```
# coeficienții modelului
data_coeficient <- tidy(logit_model_m)
ggplot(data_coeficient, aes(x = term, y = estimate, ymin = estimate -
std.error, ymax = estimate + std.error)) + geom_pointrange() +
       coord_flip() + labs(title = "Estimările coeficienților cu
intervale de încredere", y = "Estimare", x = "Predictori") +
       theme_minimal()
```

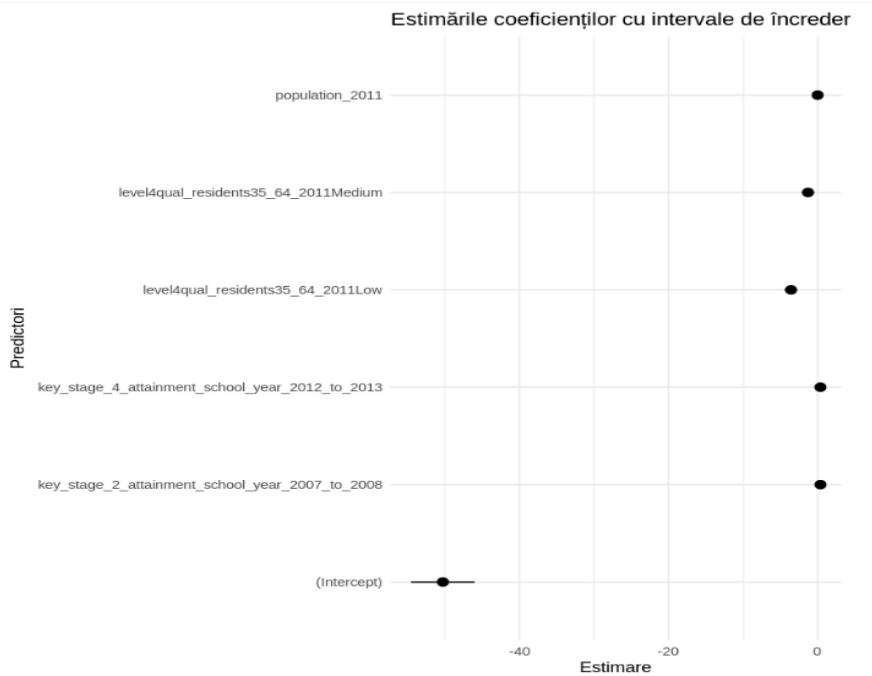
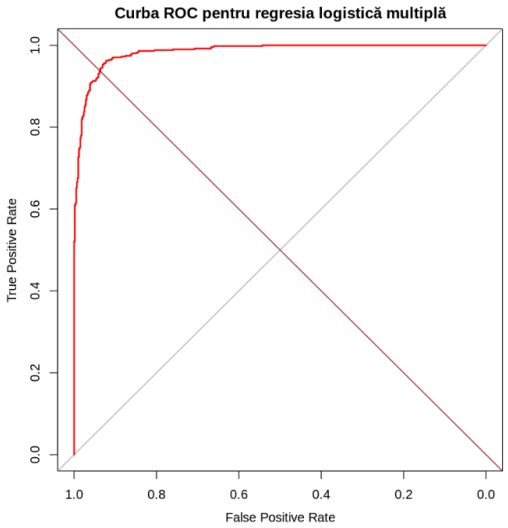


Figura următoare reprezintă curba ROC, ce arată că modelul de regresie logistică multiplă performează bine în diferențierea între cele două clase. Valoarea noastră AUC este de aproximativ 0,9841, ce întărește acest lucru.

```
probabilitati <- predict(logit_model_m, newdata = selected_data, type =
"response")
# curba ROC
rezultat_roc <- roc(response = selected_data$education_binary,
predictor = probabilitati)
plot(rezultat_roc, main = "Curba ROC pentru regresia logistică
multiplă", col = "red")
abline(0, 1, col = "darkred")
```

```
[1] "AUC Value: 0.984181712288876"
```



Al doilea model de regresie logistică multiplă se folosește 10 caracteristici, cele ce influențează cel mai mult sunt cele de performanță educațională, dar și cele ce prezintă starea la 19 ani (dacă muncesc sau nu, dacă urmează o formă de educația superioară sau nu).

```
# regresia logistică multiplă cu mai multe variabile
logit_model_m2 <- glm(education_binary ~ population_2011 +
level4qual_residents35_64_2011 +
key_stage_2_attainment_school_year_2007_to_2008 +
key_stage_4_attainment_school_year_2012_to_2013 +
activity_at_age_19_full_time_higher_education +
activity_at_age_19_sustained_further_education +
activity_at_age_19_appprenticeships +
activity_at_age_19_employment_with_earnings_above_0 +
activity_at_age_19_employment_with_earnings_above_10_000 +
activity_at_age_19_out_of_work,
                      family = binomial, data = selected_data)
summary(logit_model_m2) #  rezumatul modelului
```

```
Coefficients:
                                                    Estimate Std. Error
(Intercept)                                       -1.076e+02  2.002e+01
population_2011                                   -6.032e-06  6.915e-06
level4qual_residents35_64_2011Low                 -5.712e+00  1.120e+01
level4qual_residents35_64_2011Medium              -4.375e+00  1.118e+01
key_stage_2_attainment_school_year_2007_to_2008    7.595e-01  1.227e-01
key_stage_4_attainment_school_year_2012_to_2013    6.627e-01  1.134e-01
activity_at_age_19_full_time_higher_education      4.615e-01  9.872e-02
activity_at_age_19_sustained_further_education     1.895e-01  1.201e-01
activity_at_age_19_appprenticeships               -1.389e-01  1.601e-01
activity_at_age_19_employment_with_earnings_above_0  3.897e-02  8.580e-02
activity_at_age_19_employment_with_earnings_above_10_000 -9.802e-02  1.163e-01
activity_at_age_19_out_of_work                    -2.013e-01  1.426e-01
                                                  z value Pr(>|z|)
(Intercept)                                        -5.373 7.73e-08 ***
population_2011                                    -0.872    0.383
level4qual_residents35_64_2011Low                  -0.510    0.610
level4qual_residents35_64_2011Medium               -0.391    0.696
key_stage_2_attainment_school_year_2007_to_2008     6.190 6.00e-10 ***
key_stage_4_attainment_school_year_2012_to_2013     5.843 5.12e-09 ***
activity_at_age_19_full_time_higher_education       4.675 2.94e-06 ***
activity_at_age_19_sustained_further_education      1.578    0.114
activity_at_age_19_appprenticeships                -0.868    0.386
activity_at_age_19_employment_with_earnings_above_0  0.454    0.650
activity_at_age_19_employment_with_earnings_above_10_000  -0.843    0.399
activity_at_age_19_out_of_work                     -1.411    0.158
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 778.22  on 619  degrees of freedom
Residual deviance: 100.09  on 608  degrees of freedom
  (484 observations deleted due to missingness)
AIC: 124.09

Number of Fisher Scoring iterations: 11
```

- **Performanță educațională** (0,7595, respectiv 0,6627): sugerează asocieri pozitive semnificative cu rezultatul educațional
- **Devianța nulă** este 778,22 pe 619 și **devianța reziduală** este 100,09: scăderea semnificativă de la devianța nulă la devianța reziduală indică faptul că modelul cu predictori se potrivește mult mai bine decât modelul gol.
- **AIC** este 124,09 : indică un model bine ajustat

Figura următoare reprezintă graficul cu coeficienții estimați și intervalele lor de încredere pentru predictorii din model.

```
data_coeficient2 <- tidy(logit_model_m2)# coeficienții modelului

ggplot(data_coeficient2, aes(x = term, y = estimate, ymin = estimate - std.error, ymax = estimate + std.error)) +
       geom_pointrange() +
       coord_flip() +
       labs(title = "Estimările coeficienților cu intervale de încredere", y = "Estimare", x = "Predictori") +
       theme_minimal()
```

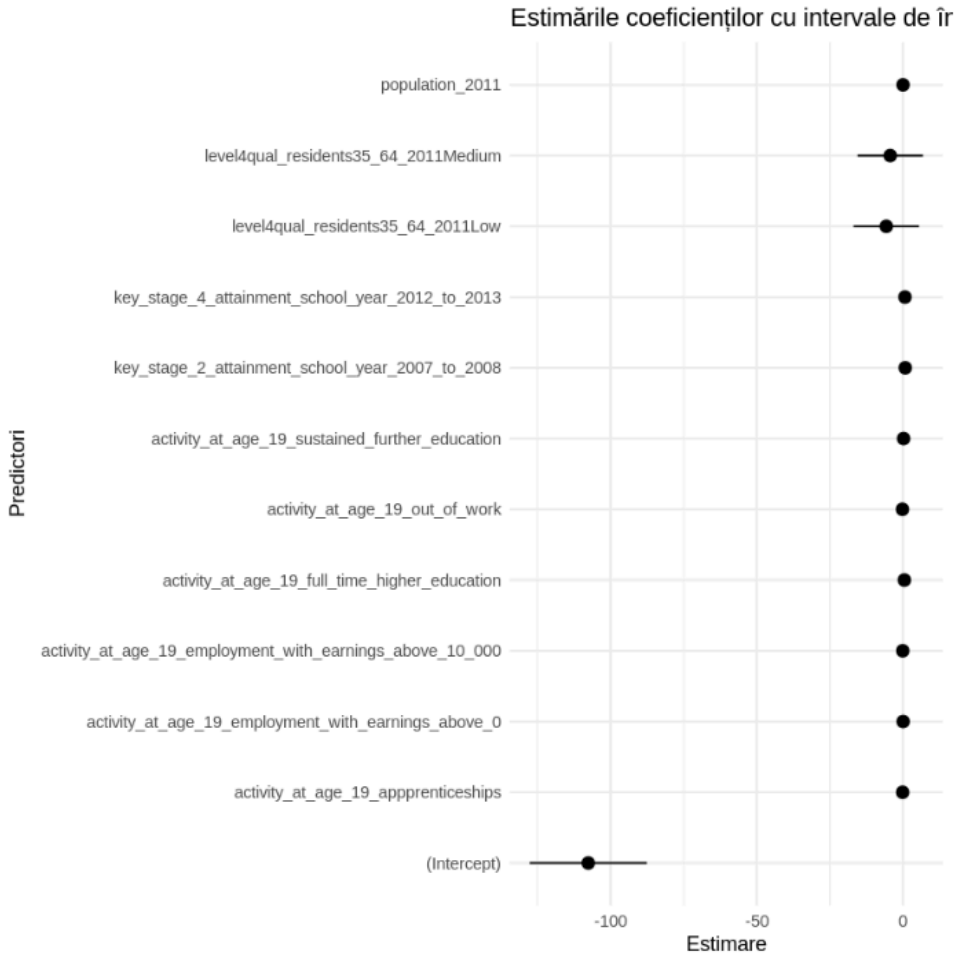
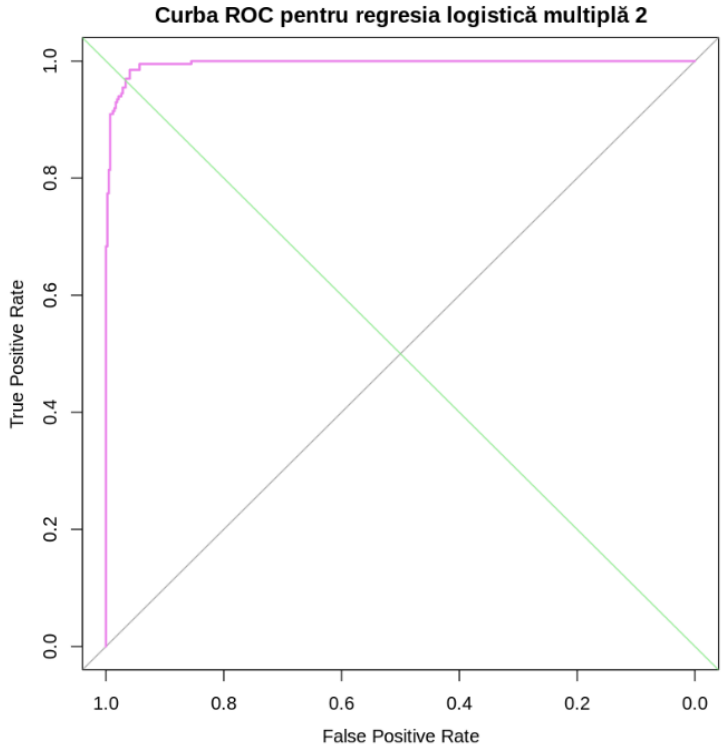


Figura următoare reprezintă curba ROC, ce arată că modelul de regresie logistică multiplă performează mult mai bine decât celelalte în diferențierea între cele două clase, fapt evidențiat și de valoarea AUC de aproximativ 0,9955.

```
probabilitati <- predict(logit_model_m2, newdata = selected_data, type
= "response")
# curba ROC
rezultat_roc <- roc(response = selected_data$education_binary,
predictor = probabilitati)
plot(rezultat_roc, main = "Curba ROC pentru regresia logistică multiplă
2", xlab = "False Positive Rate", ylab = "True Positive Rate", col =
"violet")
abline(0, 1, col = "lightgreen")
```

```
[1] "AUC Value: 0.995583618806622"
```



Pentru acest model am decis să realizez și o matrice de confuzie. Această matrice arată numărul de predicții corecte și incorecte făcute de model. Aceasta are acuratețea de 96% (proporția totală de predicții care au fost corecte).

```
predicted_clas <- ifelse(probabilitati > 0.5, 1, 0)
conf_matrix <- confusionMatrix(as.factor(predicted_clas),
as.factor(selected_data$education_binary))
print(conf_matrix)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 410  12
         1  11 187

               Accuracy : 0.9629
                 95% CI : (0.9449, 0.9763)
    No Information Rate : 0.679
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9148

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9739
            Specificity : 0.9397
         Pos Pred Value : 0.9716
         Neg Pred Value : 0.9444
             Prevalence : 0.6790
         Detection Rate : 0.6613
   Detection Prevalence : 0.6806
      Balanced Accuracy : 0.9568

       'Positive' Class : 0
```

În concluzie, în urma acestor experimente, am observat gradual că performanța modelelor este mai bună în cazul regresiilor logistice multiple. În setul nostru de date, ”education_score” este identificată drept variabilă răspuns, și este o variabilă continuă. Astfel, pe baza setului de date ales și a ipotezei de la care am pornit, am obținut că elevii din orașele mai mici depășesc pragul educațional ales și se descurcă mai bine din punct de vedere academic decât cei din orașele mai mari.

Aceste modele oferă informații utile despre asocierea dintre variabilele independente și probabilitatea unui eveniment binar (în acest caz, obținerea unui scor educaționale peste un anumit prag). Analiza modelelor poate ajuta la înțelegerea și identificarea factorilor care influențează obținerea acestor scoruri educaționale și poate servi ca bază pentru luarea deciziilor și intervențiilor în domeniul educației.

4. Bibliografie:

tidytuesday/data/2024/2024-01-23 at master · rfordatascience/tidytuesday (github.com)

youngpeoplesattainmentintownsreferencetable1.xlsx - Foi de calcul Google

1 Introduction | Advanced Statistical Computing (bookdown.org)

Why do children and young people in smaller towns do better academically than those in larger towns? - Office for National Statistics (ons.gov.uk)

Regresie logistică - UC Business Analytics Ghid de programare R | Market tay

https://stats.libretexts.org/Bookshelves/Applied_Statistics/Biological_Statistics_(McDonald)/05%3A_Tests_for_Multiple_Measurement_Variables/5.06%3A_Simple_Logistic_Regression

http://math.ucv.ro/~gorunescu/courses/en/curs/7.pdf