Universitatea Babeș-Bolyai

Facultatea de Științe Economice și Gestiunea Afacerilor

**PROIECT**

**ESTIMAREA COSTULUI SPITALIZĂRII**

Student: Șendroni Mircea

Informatică Economică – ID

Anul 3

**I. INTRODUCERE**

Serviciu esențial pentru comunitate, asistența medicală reprezintă o parte însemnată din bugetele unităților administrative locale, respectiv naționale. Așadar, eficientizarea cheltuielilor cu asistența medicală prezintă un interes major atât dintr-un unghi economic, cât și politic.

Acest deziderat este prezent la nivelul tuturor ordonatorilor de credite, indiferent de nivelul lor. În alte cuvinte, interesul pentru eficientizarea cheltuielilor este actual nu doar la nivel de politică de stat, ci și la nivel descentralizat, la treapta managerilor de spitale.

Aceștia din urmă au responsabilitatea de a întocmi un buget de venituri și cheltuieli care să asigure funcționarea serviciilor de asistență medicală prevăzute pentru unitatea sanitară administrată, dar trebuie să adopte măsuri și pentru eficientizarea acestei activități.

Pornind de la aceste premise, prezentul studiu dorește să ofere un instrument util pentru managerii de spitale în îndeplinirea activităților. Având în vedere că cea mai importantă cheltuială a unei unități sanitare o reprezintă costurile de spitalizare, pe baza datelor furnizate în acest scop, s-a efectuat o analiză predictivă pentru a determina un model care să prezică costurile de spitalizare (predicție numerică). Datele analizate conțin o variabilă dependentă (price) – costul de spitalizare, și patru variabile independente (age, female, los, race) – vârsta, sexul, numărul de zile de spitalizare și rasa pacientului, analiza efectuată în acest proiect dorind să ofere răspunsuri la următoarele întrebări:

*1. Există o legătură între vârsta, sexul, numărul de zile de spitalizare și rasa pacientului și costul de spitalizare ? Dacă răspunsul este pozitiv, cât de puternică este această legătură ?*

*2. Este cel puțin unul dintre predictori relevant pentru a prezice costul de spitalizare ?*

*3. Există posibilitatea estimării costului de spitalizare dacă se cunosc vârsta, sexul, numărul de zile de spitalizare și rasa pacientului ?*

*4. Dacă răspunsul la întrebarea anterioară este pozitiv, cât de precisă este estimarea respectivă ?*

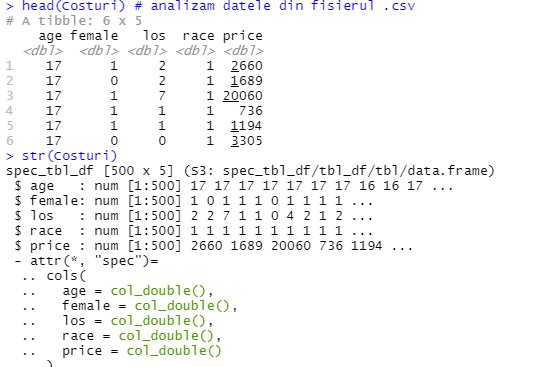
Răspunsul la aceste întrebări va furniza informații prețioase celor însărcinați cu administrarea unităților medicale, care își pot ajusta propunerile de buget astfel încât să reflecte mai bine realitățile din practica medicală. De asemenea, aceste date pot fi utilizate pentru crearea unor programe de prevenire a cauzelor care măresc în mod semnificativ costul de spitalizare, totul în scopul creării unui act medical mai eficient, modern și care să se adreseze mai multor persoane.

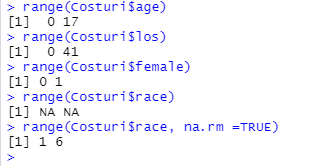
În fine, din informațiile noastre nu am identificat un studiu asemănător ale cărui rezultate să poate fi analizate cu cele stabilite în urma acestui proiect, astfel că datele din acest proiect pot reprezenta un punct de pornire în analiza acestui domeniu.

**II. Setul de date**

Datele analizate în cadrul acestui studiu au fost colectate de pe platforma Kaggle, care conține numeroase seturi de date pentru domenii diferite. În concret, setul nostru a fost descărcat de la adresa [*https://www.kaggle.com/datasets/ravichaubey1506/healthcare-cost*](https://www.kaggle.com/datasets/ravichaubey1506/healthcare-cost)*.*

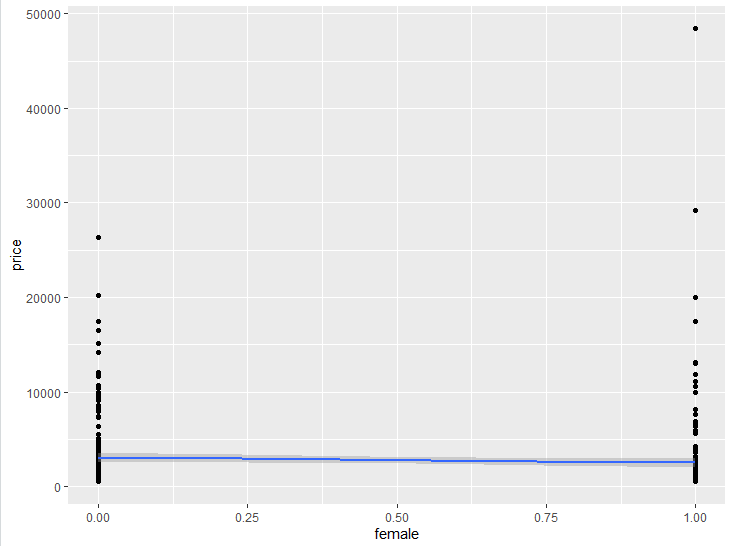
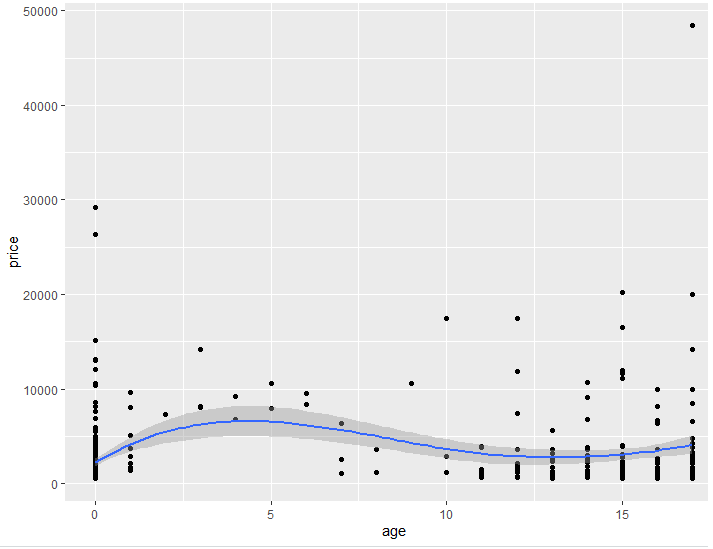
În continuare, se constată că setul de date este compus din cinci coloane care au 500 de înregistrări fiecare. Coloanele sunt reprezentate după cum urmează: price (prețul de spitalizare), age (vârsta pacientului), los (zilele de spitalizare petrecute în unitatea sanitară), female (sexul pacientului) și race (rasa pacientului). Toate coloanele sunt de tipul numerice.

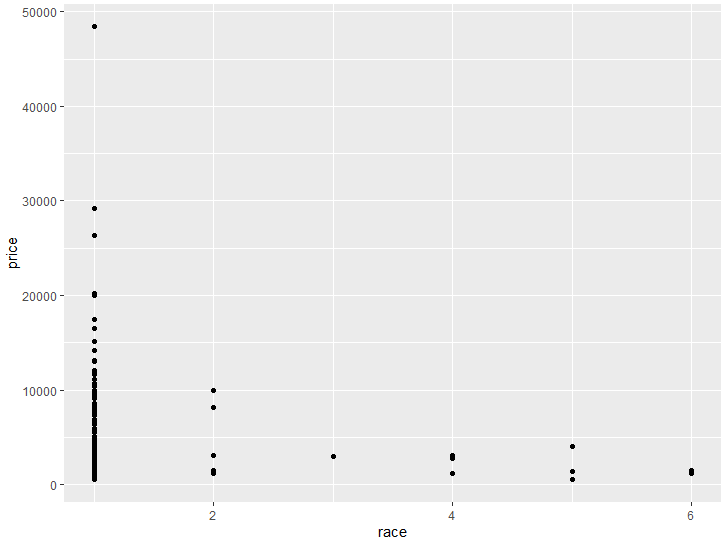
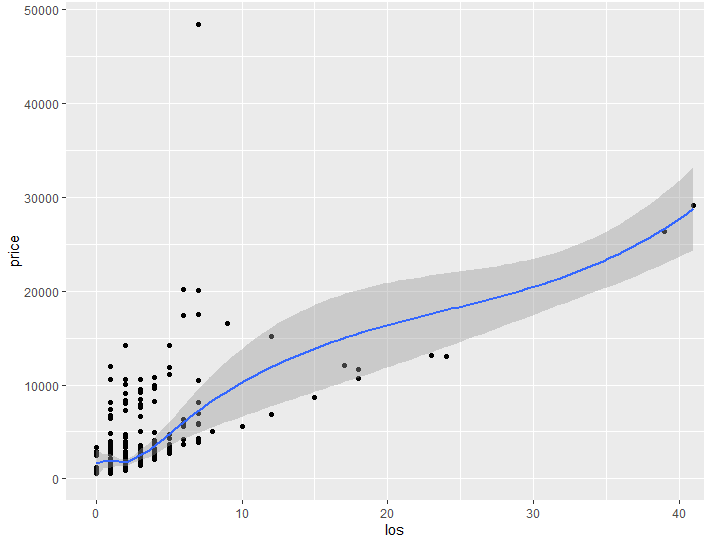




După cum se poate observa, variabila *age* conține valori între 0 și 17, ceea ce înseamnă că aceste date privesc costul de spitalizare aferent unor pacienți minori. În continuare, reținem că variabila *female* este binară, unde 1 arată că pacientul este o persoană de sex feminin, că variabila *los* (zilele de spitalizare) este cuprinsă între 0 și 41, iar variabila *race* are valori între 1 și 6, unde 1 reprezintă rasa europoidă, 2 rasa negroidă, 3 rasa asiatică, 4 rasa amerindiană, 5 rasa oceanică și 6 – alte rase.

Deoarece dorim să stabilim modul în care costul de spitalizare (variabila dependentă) este influențat de către variabilele independente age (vârsta pacientului), los (zilele de spitalizare petrecute în unitatea sanitară), female (sexul pacientului) și race (rasa pacientului), am efectuat o analiză vizuală preliminară a fiecărei variabile independente în raport cu costul de spitalizare (price).





La o analiză vizuală, între variabila price și variabilele age, respectiv los poate există o legătură sub forma unei regresii. În același timp, variabilele female și race nu prezintă indicii în acest sens.

**III. Rezultate și discuții**

Setul de date este compus din variabile numerice și, în esență, se dorește să se stabilească dacă variabilele independente influențează variabila dependetă *price*, la rândul ei numerică. Așadar, în cadrul acestui proiect se va realiza o predicție numerică.

Pornind de la aceste constatări, prezentul studiu va analiza setul de date prin prisma a două metode: regresia liniară și arborele de decizie. Metodele au fost alese pentru că nu sunt complexe și astfel pot fi înțelese de destinatarii acestui studiu, știut fiind faptul că persoanele care trebuie să ia decizii nu vor lua în considerare rezultate unor metode mai complexe, ale căror mecanisme nu le-au înțeles.

**III.1 Regresia liniară**

Prima metodă pe care o vom folosi pentru a analiza datele de mai sus este regresia. Această metodă presupune că variabila dependentă (price) poate fi exprimată printr-o formulă matematică de forma Y = f(X) + €, unde € reprezintă eroarea ireductibilă, adică variabilele independente care influențează Y, dar care nu sunt cunoscute.

Mai departe, Y = f(X) + €, este estimat prin crearea unui Y^= f^(X), care este creat pe baza unui eșantion din datele analizate. Scopul acestui procedeu este de realiza un Y^ cât mai precis, adică prin reducerea erorii reductibile (rezultată din estimarea lui f prin f^), la valori cât mai apropiate de eroare ireducbilă (€).

Pe cale de consecință, este necesar să identificăm o relație f între Y (price) și X=(X1, X2,..., Xp) (age, female, los, race), astfel încât Y = f(X) + €. Mai departe, vom presupune că f are o formă liniară, ceea ce înseamnă că Y = β0 + β1 X1 + ... + βp Xp. În continuare, algoritmul metodei va pondera fiecare parametru prin obținerea unor estimări pentru valorile acestora, prin testarea datelor pe un set de antrenament.

Metoda mai presupune analiza fiecărei variabile independente în corelație cu variabila depedentă, precum și analiza interacțiunii tuturor variabilelor dependente sau o parte dintre acestea în raport cu variabila dependentă.

**III.1.1. Regresia simplă**

**III.1.1.1. Variabila age**

Prima variabilă analizată este *age*. Astfel, am evidențiat variabila dependentă *price* prin prisma variabilei independente *age*, reținând că price ≈ β0 + β1 x age, unde β0 reprezintă intercept, iar β1 panta (slope). Folosind R-Studio am calculat valorile corespunzătoare interceptului, penții, erorii standard, t-statistic și p-value, rezultatele fiind cele de mai jos:

Text, letter

Description automatically generated

Conform acestor date, reținem că:

* variabila dependentă *price* crește cu 2399,66 u.m., în contextul în care variabila independentă *age* lipsește;
* în medie, estimarea acestui parametru diferă de valoarea reală cu 213,90 u.m.;
* numărul de deviații standard cu care paramentrul se îndepărtează de valoarea reală este de 11.218;
* la fiecare creștere cu 1 u.m. variabilei *age*, costul de spitalizare crește cu 73,68 u.m;
* în medie, estimarea acestui parametru diferă de valoarea reală cu 24,85 u.m.;
* numărul de deviații standard cu care paramentrul se îndepărtează de valoarea reală de 2.964;
* probabilitatea ca asocierea dintre variabile să se datoreze șansei este de 0,00318, o astfel de situație excluzând această ipoteză.

Prin urmare, dacă ținem cont doar de variabila *age*: price (costul de spitalizare) = 2399,66 + 73,68 x age.

Deoarece nu se cunoaște f real, pe baza datelor de mai sus, am stabilit **intervalele de încredere** pentru parametrii βi, cu o încredere de 95%, utilizând formula βi [βˆ ∈ i − 2SE(βˆi ), βˆi + 2SE(βˆi )], unde SE este eroare standard.

Astfel, am obținut următoarele intervale:

* pentru β0 : [1979.39805 , 2819.9241], respectiv costul de spitalizare se va situa în acest interval, dacă nu ținem cont de vârsta pacientului;
* pentru β1 : [24.84531 , 122.5109], respectiv o creștere cu o unitate a vârstei pacientului va determina o creștere a costului de spitalizare cu o valoare din acest interval.

Următorul pas constă în **testarea ipotezelor**, respectiv:

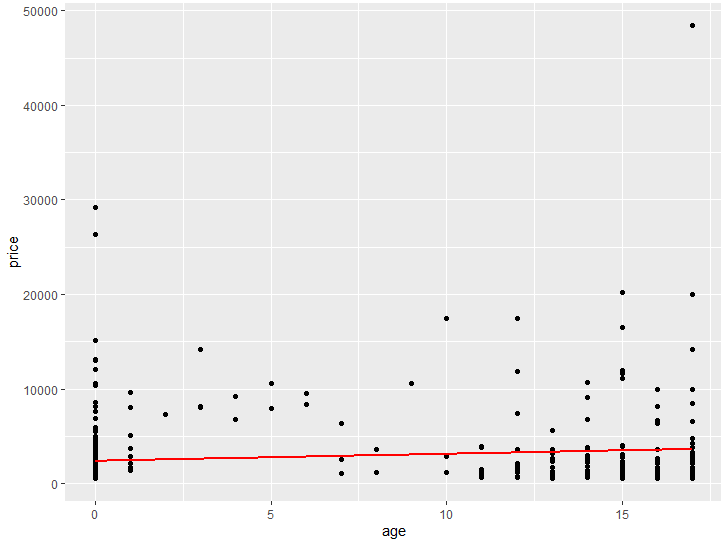
H0 – Ipoteza nulă, care presupune că nu există nicio relație între X (age) și Y (price) sau că aceasta se datorează șansei, hazardului. Testarea ipotezei este simplă și presupune că β1 = 0; În caz contrar, ipoteze este respinsă.

Ha – Această ipoteză presupune că există o relație între X (age) și Y (price). Incidența ipotezei presupune că β1 =/= 0.

Având în vedere valoarea foarte mică a lui p-value (0,00318), probabilitatea ca legătură dintre variabile să fie datorată întâmplării este improbabilă, motiv pentru ca ipoteza H0 va fi respinsă. Prin urmare, putem afirma cu certitudine că există o relație între *age* și *price,* care nu este cauzată de hazard.

În continuare, pentru a determina potrivirea modelului, s-a creat **un set de date de test**, luând în considerare 100 de valori ale variabilei *age*, în raport cu care se vor calcula valorile variabilei *price*. Aceste valori au fost aplicate la modelul creat, de unde a rezultat o variabilă nouă,

grid\_age ce cuprinde predicțiile. Rezultatele au fost cuprinse într-un nou grafic, utilizând o dreaptă, după cum se poate observa în figura de mai jos.

****

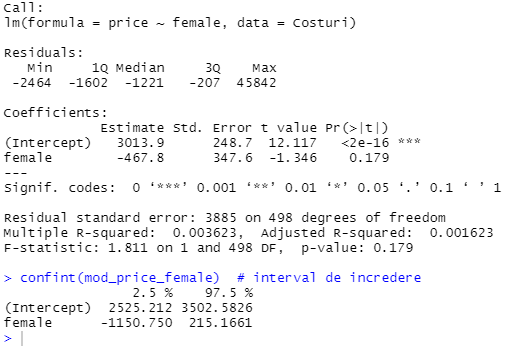
**Acuratețea modelului** se stabilește prin doi indicatori. Primul este eroarea standard reziduală (RSE), ce reprezintă mărimea medie cu care Y va devia de la linia de regresie. Pentru a exprima un model corespunzător valoarea lui RSE trebuie să fie cât mai mică. Al doilea, respectiv R2, măsoară proporția din variabilitatea lui Y care poate fi explicată pe baza lui X. Pentru a descrie un model folositor, valoarea lui R2 trebuie să fie cât mai aproape de 1.

Pornind de la aceste considerente teoretice, constatăm că variabila *age* explică variația varibilei *price* în procent de 1,7% (0,01734), restul variației de 98,5% neputând fi explicată pe baza acestui model.

De asemenea, reținem că eroare standard reziduală (RSE) este de 3858. Valoarea ridicată a acestui indicator, coroborat cu valoarea lui R2, conduce la concluzia că **modelul de mai sus nu este potrivit** pentru a descrie relația dintre cele două variabile.

**III.1.1.2 Variabila female**

A doua variabilă analizată este *female*. Astfel, am evidențiat variabila dependentă *price* prin prisma variabilei independente *female*, reținând că price ≈ β0 + β1 x female, unde β0 reprezintă intercept, iar β1 panta (slope). Folosind R-Studio am calculat valorile corespunzătoare interceptului, penții, erorii standard, t-statistic și p-value, rezultatele fiind cele de mai jos:



Conform acestor date, reținem că:

* variabila dependentă *price* crește cu 3013,9 u.m., în contextul în care variabila independentă *female* lipsește;
* în medie, estimarea acestui parametru diferă de valoarea reală cu 248,7 u.m.;
* numărul de deviații standard cu care paramentrul se îndepărtează de valoarea reală este de 12.117;
* la o creștere cu 1 u.m. a variabilei *female*, costul de spitalizare scade cu -467,8 u.m – relație negativă[[1]](#footnote-1);
* în medie, estimarea acestui parametru diferă de valoarea reală cu 347,6 u.m.;
* numărul de deviații standard cu care paramentrul se îndepărtează de valoarea reală de -1.346;
* probabilitatea ca asocierea dintre variabile să se datoreze șansei este de 0,179;

Prin urmare, dacă ținem cont doar de variabila *female*: price (costul de spitalizare) = 3013,19 – 467,8 x female.

Deoarece nu se cunoaște f real, pe baza datelor de mai sus, am stabilit **intervalele de încredere** pentru parametrii βi, cu o încredere de 95%, utilizând formula βi [βˆ ∈ i − 2SE(βˆi ), βˆi + 2SE(βˆi )], unde SE este eroare standard.

Astfel, am obținut următoarele intervale:

* pentru β0 : [2525.212 , 3502.5826], respectiv costul de spitalizare se va situa în acest interval, dacă nu ținem cont de vârsta pacientului;
* pentru β1 : [-1150.750 , 215.661], respectiv o creștere cu o unitate a vârstei pacientului va determina o creștere a costului de spitalizare cu o valoare din acest interval.

Următorul pas constă în **testarea ipotezelor**, respectiv:

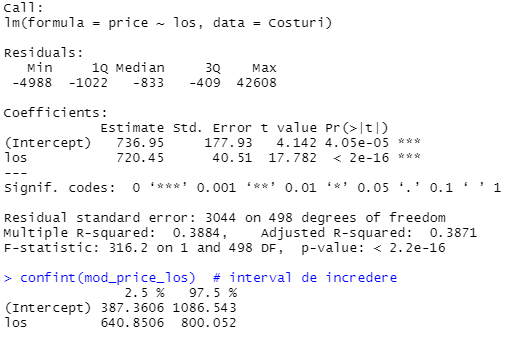
H0 – Ipoteza nulă, care presupune că nu există nicio relație între X (age) și Y (price) sau că aceasta se datorează șansei, hazardului. Testarea ipotezei este simplă și presupune că β1 = 0; În caz contrar, ipoteze este respinsă.

Ha – Această ipoteză presupune că există o relație între X (age) și Y (price). Incidența ipotezei presupune că β1 =/= 0.

Având în vedere valoarea lui p-value (0,179), probabilitatea ca legătura dintre variabile să fie datorată întâmplării este în proporție de 18%. Un astfel de procent este semnificativ, deoarece, în general, se consideră că indicatorul p-value este relevant dacă valoarea acestuia este mai puțin de 0,05. În consecință, **variabila trebuie scoasă din model, fiind irelevantă** **în modelul regresiei simple.**

**III.1.1.3. Variabila los (zile de spitalizare)**

Următoarea variabilă analizată este *los*. Astfel, am evidențiat variabila dependentă *price* prin prisma variabilei independente *los*, reținând că price ≈ β0 + β1 x los, unde β0 reprezintă intercept, iar β1 panta (slope). Folosind R-Studio am calculat valorile corespunzătoare interceptului, penții, erorii standard, t-statistic și p-value, rezultatele fiind cele de mai jos:

****

Conform acestor date, reținem că:

* variabila dependentă *price* crește cu 736,95 u.m., în contextul în care variabila independentă *los* lipsește;
* în medie, estimarea acestui parametru diferă de valoarea reală cu 177,93 u.m.;
* numărul de deviații standard cu care paramentrul se îndepărtează de valoarea reală este de 4.142;
* la fiecare creștere cu 1 u.m. variabilei *los*, costul de spitalizare crește cu 720,45 u.m;
* în medie, estimarea acestui parametru diferă de valoarea reală cu 40,51 u.m.;
* numărul de deviații standard cu care paramentrul se îndepărtează de valoarea reală de 17.782;
* probabilitatea ca asocierea dintre variabile să se datoreze șansei este mai mică de 2e-16 (0.0000000000000002), o astfel de situație excluzând această ipoteză.

Prin urmare, dacă ținem cont doar de variabila *age*: price (costul de spitalizare) = 736,95 + 720,45 x los.

Deoarece nu se cunoaște f real, pe baza datelor de mai sus, am stabilit **intervalele de încredere** pentru parametrii βi, cu o încredere de 95%, utilizând formula βi [βˆ ∈ i − 2SE(βˆi ), βˆi + 2SE(βˆi )], unde SE este eroare standard.

Astfel, am obținut următoarele intervale:

* pentru β0 : [387.3606 , 1086.543], respectiv costul de spitalizare se va situa în acest interval, dacă nu ținem cont de zilele de spitalizare ale pacientului;
* pentru β1 : [640.8506 , 800.052], respectiv o creștere cu o unitate a zilelor de spitalizare va determina o creștere a costului de spitalizare cu o valoare din acest interval.

Următorul pas constă în **testarea ipotezelor**, respectiv:

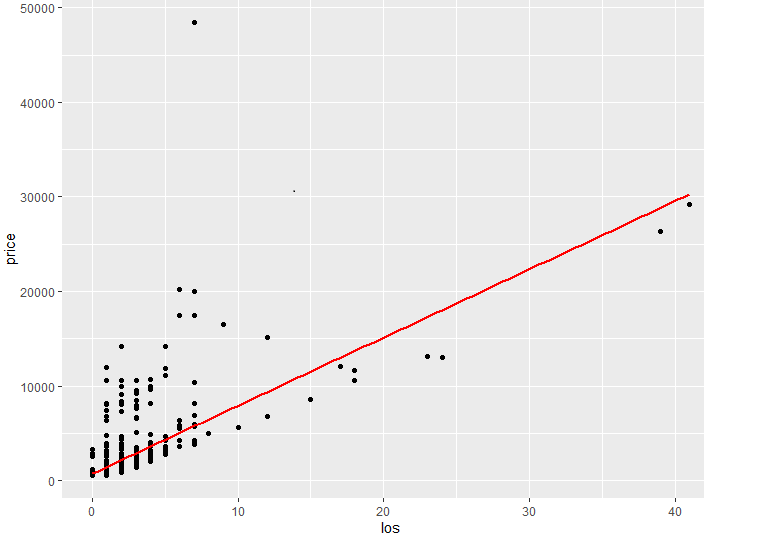
H0 – Ipoteza nulă, care presupune că nu există nicio relație între X (age) și Y (price) sau că aceasta se datorează șansei, hazardului. Testarea ipotezei este simplă și presupune că β1 = 0; În caz contrar, ipoteze este respinsă.

Ha – Această ipoteză presupune că există o relație între X (age) și Y (price). Incidența ipotezei presupune că β1 =/= 0.

Având în vedere valoarea foarte mică a lui p-value, respectiv de 2e-16 (0.0000000000000002), probabilitatea ca legătură dintre variabile să fie datorată întâmplării este improbabilă, motiv pentru ca ipoteza H0 va fi respinsă. Prin urmare, putem afirma cu certitudine că există o relație între *los* și *price,* care nu este cauzată de hazard.

În continuare, pentru a determina potrivirea modelului, s-a creat **un set de date de test**, luând în considerare 100 de valori ale variabilei *los*, în raport cu care se vor calcula valorile variabilei *price*. Aceste valori au fost aplicate la modelul creat, de unde a rezultat o variabilă nouă,

grid\_los ce cuprinde predicțiile. Rezultatele au fost cuprinse într-un nou grafic, utilizând o dreaptă, după cum se poate observa în figura de mai jos.

****

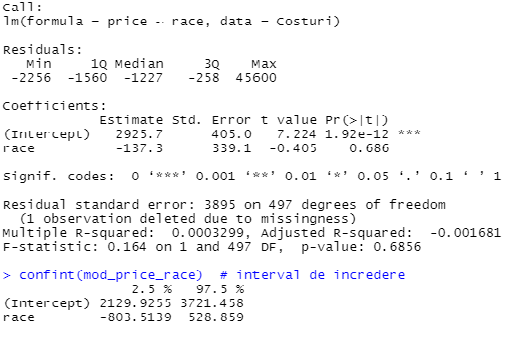
**Acuratețea modelului** se stabilește prin doi indicatori. Primul este eroarea standard reziduală (RSE), ce reprezintă mărimea medie cu care Y va devia de la linia de regresie. Pentru a exprima un model corespunzător valoarea lui RSE trebuie să fie cât mai mică. Al doilea, respectiv R2, măsoară proporția din variabilitatea lui Y care poate fi explicată pe baza lui X. Pentru a descrie un model folositor, valoarea lui R2 trebuie să fie cât mai aproape de 1.

Pornind de la aceste considerente teoretice, constatăm că variabila *los* explică variația variabilei *price* în procent de 39% (0,3884), restul variației de 61% neputând fi explicată pe baza acestui model.

De asemenea, reținem că eroare standard reziduală (RSE) este de 3044. Valoarea ridicată a acestui indicator, coroborat cu valoarea lui R2, conduce la concluzia că **modelul de mai sus nu este potrivit** pentru a descrie relația dintre cele două variabile.

**III.1.1.2 Variabila race**

Ultima variabilă analizată este *race*. Astfel, am evidențiat variabila dependentă *price* prin prisma variabilei independente *race*, reținând că price ≈ β0 + β1 x race, unde β0 reprezintă intercept, iar β1 panta (slope). Folosind R-Studio am calculat valorile corespunzătoare interceptului, penții, erorii standard, t-statistic și p-value, rezultatele fiind cele de mai jos:

****

Conform acestor date, reținem că:

* variabila dependentă *price* crește cu 2925,7 u.m., în contextul în care variabila independentă *race* lipsește;
* în medie, estimarea acestui parametru diferă de valoarea reală cu 405 u.m.;
* numărul de deviații standard cu care paramentrul se îndepărtează de valoarea reală este de 7.224;
* la o creșterea cu 1 u.m. a variabilei *race*, costul de spitalizare scade cu -137,3 u.m – relație negativă[[2]](#footnote-2);
* în medie, estimarea acestui parametru diferă de valoarea reală cu 339,1 u.m.;
* numărul de deviații standard cu care paramentrul se îndepărtează de valoarea reală de -0.405;
* probabilitatea ca asocierea dintre variabile să se datoreze șansei este de 0,686;

Prin urmare, dacă ținem cont doar de variabila *race*: price (costul de spitalizare) = 2925,7 – 137,3 x race.

Deoarece nu se cunoaște f real, pe baza datelor de mai sus, am stabilit **intervalele de încredere** pentru parametrii βi, cu o încredere de 95%, utilizând formula βi [βˆ ∈ i − 2SE(βˆi ), βˆi + 2SE(βˆi )], unde SE este eroare standar.

Astfel, am obținut următoarele intervale:

* pentru β0 : [2129.9255 , 3721.458], respectiv costul de spitalizare se va situa în acest interval, dacă nu ținem cont de vârsta pacientului;
* pentru β1 : [-803.5139 , 528.859], respectiv o creștere cu o unitate a vârstei pacientului va determina o creștere a costului de spitalizare cu o valoare din acest interval.

Următorul pas constă în **testarea ipotezelor**, respectiv:

H0 – Ipoteza nulă, care presupune că nu există nicio relație între X (age) și Y (price) sau că aceasta se datorează șansei, hazardului. Testarea ipotezei este simplă și presupune că β1 = 0; În caz contrar, ipoteze este respinsă.

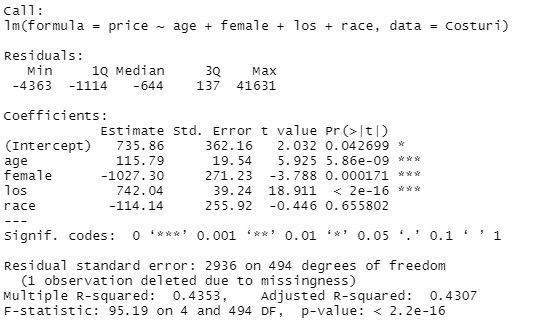
Ha – Această ipoteză presupune că există o relație între X (age) și Y (price). Incidența ipotezei presupune că β1 =/= 0.

Având în vedere valoarea lui p-value (0,686), probabilitatea ca legătură dintre variabile să fie datorată întâmplării este în proporție de 69%. Un astfel de procent este semnificativ, deoarece, în general, se consideră că indicatorul p-value este relevant dacă valoarea acestuia este mai puțin de 0,05. În consecință, **variabila trebuie scoasă din model, fiind irelevantă** **în modelul regresiei simple.**

**III.1.2. Regresia multiplă**

Din datele rezultate din analiza regresiei simple obținute a rezultat că există o legătură între o parte dintre variabilele independente și costul de spitalizare (variabila dependentă). În concret, este probabil ca vârsta și zilele de internare să influențeze costul de spitalizare, în timp ce variabilele privind sexul și rasa pacientului să nu aibă vreo influență asupra varibilei dependente.

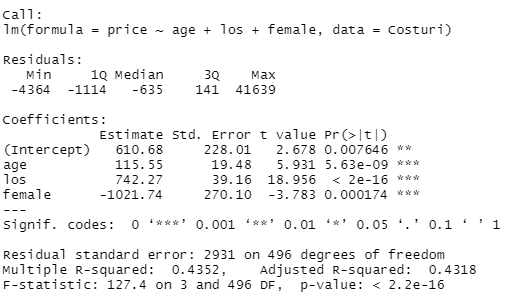
Pentru a determina mai amănunțit modul în care variabilele influențează costul de spitalizare, am considerat util să folosim metoda regresiei liniare cu mai mulți predictori (age, female, los, race). Astfel, am considerat că price ≈ β0 + β1 x age + β2 x female + β3 x los + β4 x race, obținând următoarele valori ale indicatorilor relevanți:



Analizând datele de mai sus reținem următoarele:

* variabila *race* are o probabilitate de 66% ca legătura dintre variabile să fie datorată șansei. În condițiile în care indicatorul p-value trebuie să aibă o valoare mai mică de 0,05 pentru a fi considerat relevant, rețin că variabila *race* trebui eliminată din model;
* variabila *female* este irelevantă în regresia simplă, dar devine relevantă în regresia multiplă prin prisma indicatorului p-value;
* relevante pentru stabilirea însemnătății legăturii cu varabiala depedentă *price* sunt variabilele independente *age, female* și *race.*

Astfel, am obținut următoarele rezultate:



Relația dintre predictori și variabila dependetă se determină cu ajutorul ipotezelor alternative.

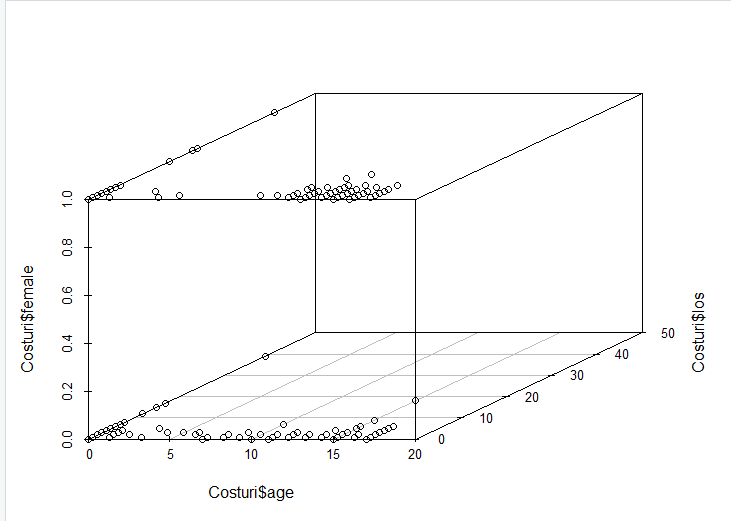
În acest caz, ipoteza nulă este adevărată dacă H0 : β1 = β2 =...= βp = 0, iar ipoteza alternativă Ha este adevărată dacă cel puțin unul dintre βj este diferit de 0. Indicatorul prin care se stabilește care dintre ipoteze este incidentă este F – statistic. Astfel, H0 este acceptată dacă F – statistic se aproprie de valoarea 1, iar Ha este acceptată dacă indicatorul are o valoarea mult mai mare ca 1.

În cazul acestui model se observă că F-statistic este 127,4, cu mult mai mare decât 1. Probabilitatea ca valoarea indicatorului să fie alta este de 0.00000000000000022. Prin urmare, ipoteza nulă trebuie respinsă.

De asemenea, se observă o ușoară îmbunătățire a modelului prin prisma reducerii valorii RSE și creșterea indicatorului F-statistic. Prin urmare, acest model are cea mai mare acuratețe dintre cele prezentate până în acest moment.

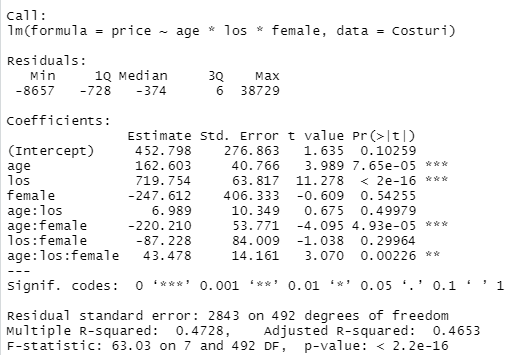
Drept consecință, funcția va fi de forma:

**Yprice = 610,68 + 115,55 x Xage + 742.27 x Xlos – 1021,74 x Xfemale**

****

**III.1.3. Regresia multiplă cu interacțiune între variabile**

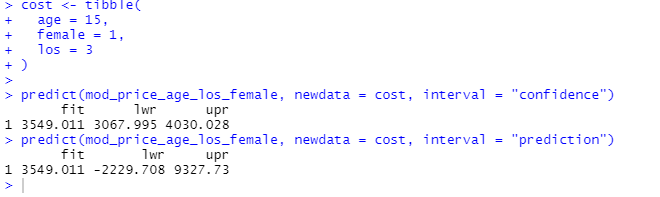
Pentru a determina modul în care costul de spitalizare este influențat de către cele trei variabile și prin luarea în calcul a interacțiunii între variabilele independente, am folosit Rstudio pentru a constata sinergia dintre variabile. În alte cuvinte, am dorit să arătam dacă modificarea valorii unei variabile independente influențează valoarea altei valori independente.

****

În acest model, devin relevante variabilele age, los, age:female și age:los:female. Acest model oferă ar părea că oferă o acuratețe mai bună deoarece RSE este mai redusă, iar coeficientul R2 mai ridicat în raport cu valorile înregistrate în modelul care cuprinde variabilele age, female și los fără interacțiune între ele. Cu toate acestea, rețin că indicele F-statistic a scăzut semnificativ. În raport cu toți acești indicatori, consider că modelul regresiei multiple fără interacțiune de între variabilele independente explică mai bine formarea costului de spitalizare.

**III.1.3. Predicții**

Pe baza modelui de mai sus, am efectuat un test în care am ales aleator următoarele variabile: age = 15, female = 1, los =2.

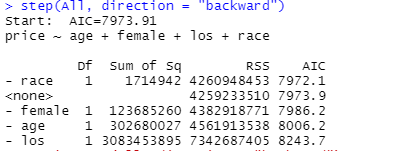
****

Intervalul de încredere este necesar pentru a calcula cât de apropiată este predicția lui Y^ de f(X) estimat. Astfel, se determină incertutidinea costului mediu (încrederea în valori medii). Intervalul de predicție este necesar pentru a calcula cât de aproape este Y^ de Y real (incluzând și eroarea ireductibilă). Astfel, se determină incertitudinea din jurul unei anumite instanțe de test (încrederea în valori punctuale).

Așadar, costul de spitalizare așteptat este de 3549 u.m, valoare care se situează atât în interiorul intervalului de încredere, cât și în cel de predicție.

**III.1.4. Alegerea celui mai bun model**

Astfel cum rezultă din datele expuse mai sus, modelul care conține variabilele age, los și female oferă ceea mai bună acuratețe. Concluzia este întărită și prin apelarea metodei stepwise regression în Rstudio, de unde reiese că prin eliminarea variabilei race, indicatorul de acuratețe se îmbunătățește ușor.

****

**III.1.5. Concluzii în raport cu metoda regresiei**

*1. Există o legătură între vârsta, sexul, numărul de zile de spitalizare și rasa pacientului și costul de spitalizare ? Dacă răspunsul este pozitiv, cât de puternică este această legătură ?*

Metoda parametrică a relevat că există o legătură între vârsta, sexul, numărul de zile de spitalizare și rasa pacientului pentru determinarea costului acordării de îngrijiri medicale. Acest aspect a reieșit prin eliminarea ipotezei nule în cadrul modelului care privea toate variabilele (H0 : βage = βfemale = βlos = βrace = 0, ipoteza nulă s-a dovedit a fi neadevărată).

Dintre variabilele analizate, variabilele privind vârsta, sexul și numărul de zile de spitalizare pot explica 44% din costul total de spitalizare (relevanța modelului). Așadar, legătura nu este foarte puternică. Pe cale de consecință, se constată că există alți factori, neluați în considerare, care au un impact semnificativ asupra variabilei dependente.

*2. Este cel puțin unul dintre predictori relevant pentru a prezice costul de spitalizare ?*

Luate individual, variabilavârsta pacientului are un impact redus asupra determinării costului de spitalizare (1,7%), în timp ce variabila los explică 39% din formarea costului. Cu toate acestea, variabilele nu sunt folositoarea pentru a precize costul de spitalizare cu un grad crescut de acuratețe. În ceea ce privește variabilele female și race, luate izolat, acestea nu sunt relevante pentru modul în care se formează costul de spitalizare. Prin urmare, nu există predicatori individuali folositori pentru a prezice costul de spitalizare.

*3. Există posibilitatea estimării costului de spitalizare dacă se cunosc vârsta, sexul, numărul de zile de spitalizare și rasa pacientului ?*

Există posiblitatea efectuării unei predicții.

*4. Dacă răspunsul la întrebarea anterioară este pozitiv, cât de precisă este estimarea respectivă ?*

Dacă dorește determinarea unei valorii medii, atunci vom folosi intervalul de încredere. În schimb, dacă se dorește determinarea unei valori punctuale, atunci vom folosi intervalul de predicție. Estimările nu sunt fiabile, acuratețea modelului fiind limitată în acest caz.

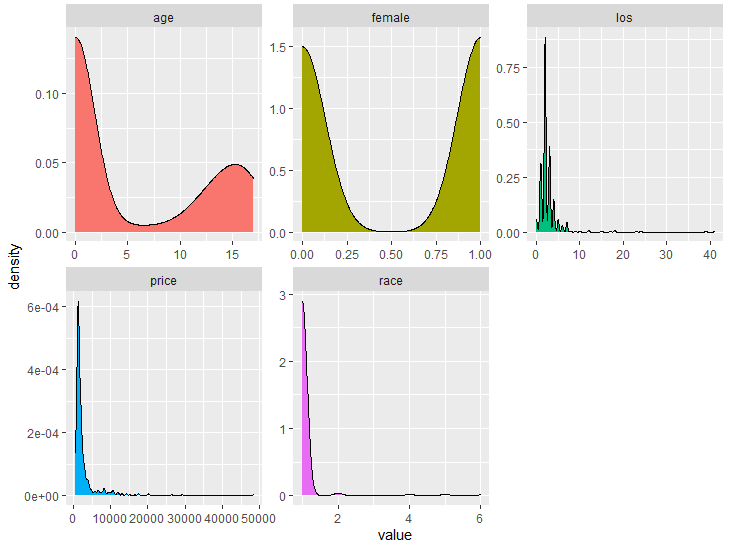
**III.2 Arborele de decizie**

Prin prisma rezultatelor nesatisfăcătoare obținute în urma metodei regresie, am considerat oportună analizarea datelor printr-o nouă metodă, respectiv cea a arborilor de decizie.

Această metodată a fost aleasă pentru că este ușor de explicat și interpretat, fiind cea mai populară metodă chiar dacă performanța nu este cea mai bună. Printre punctele forte ale metodei putem preciza identificarea ușoară a variabilelor importante, rapiditatea algoritmului sau faptul că poate fi aplicată și unor modele non liniare. În schimb, metoda arborilor de regresie singulari are o varianță mare, având drept consecință stabilirea unor predicții nestabile.

**III.2.1 Pașii urmați**

Primul pas a constat în afișarea sub formă de grafice a distrubuției variabilelor numerice, după cum se observă mai jos:

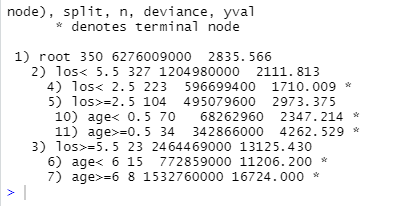


Având în vedere că analiza noastră are drept variabilă dependentă factorul price, reținem că, în cele mai multe situații, constul de spitalizare este situat între 1.000 și 3.000 u.m.

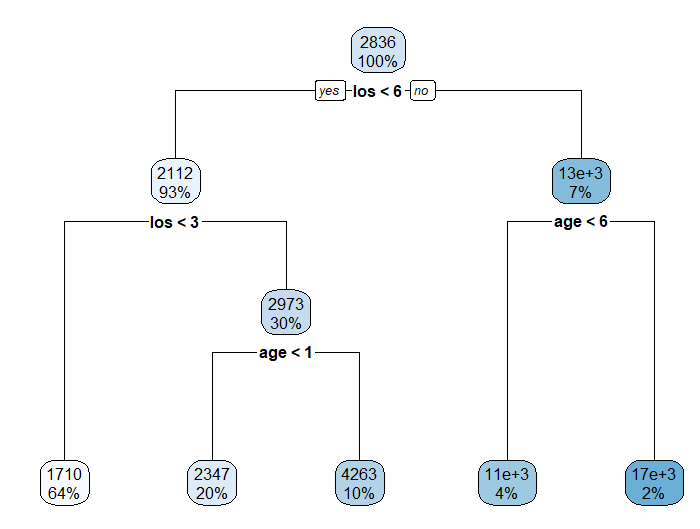
Mai apoi, am procedat la împărțirea setului de date în două subseturi, primul de antrenament, și al doilea pentru testare, respectându-se proporția de 70%, respectiv 30%.

Pe baza datelor din setul de antrenament am creat un arbore în care atributul țintă este variabile *price* în comparație cu restul variabilelor. De asemenea, trebuie precizat că arborele a fost creat prin metoda anova, cu setările implicite. Acest lucru presupune generarea mai multor arbori, care mai apoi sunt comparați prin metoda cross-validation, adică se creează sub seturi pentru valori diferite ale parametrului cost complexity (α), astfel încât să identificăm acel α pentru care relația SSE (sum of squared errors) + α \* |T| (numărul de noduri terminale în arbori) este cea mai mică.

După crearea arborelui am procedat la afișarea lui în mod text:



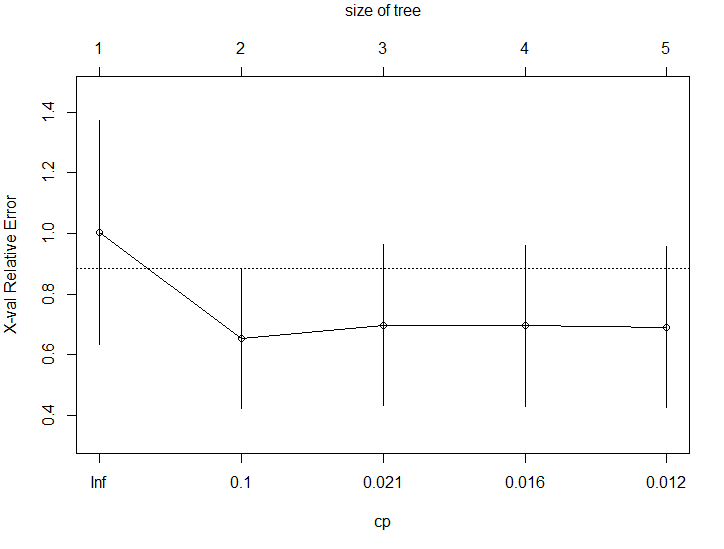
și printr-un grafic:



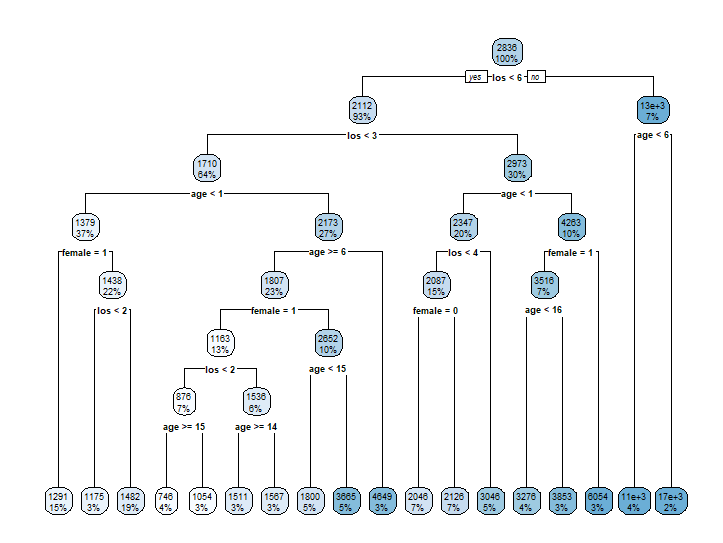
Se observă că s-a creat un arbore în care cea mai importantă variabilă este *los*, adică cea care reprezintă numărul de zile de spitalizare. Mai rețin că valoarea medie a costului de spitalizare este de 2836 u.m., și că arborele a fost împărțit în funcție de numărul efectiv de zile pentru care s-au acordat încrijiri medicale. Astfel, pentru mai puțin de 5 zile de îngrijiri medicale, costul mediu este de 2112, în timp ce pentru 6 zile sau mai mult, costul mediu crește foarte mult, și anume spre o medie de 13125 u.m. Mai reținem, că majoritatea covârșitoarea a cazurilor (93%) numărul de zile de spitalizare a fost mai puțin de 6 zile.

În fine, observăm că în acest arbore mai este relevantă și variabila *age,* reieșind că, pentru pacienții care au beneficiat între 3 și 5 zile de spitalizare și au vârsta mai mică de 1 an de zile, costul de spitalizare este mai ridicat, în medie, decât a pacienților cu vârsta de 1 an de zile sau mai mari și care au beneficiat de aceeași perioadă de tratament. În fine, variabila age mai are relevanță în cazul în care perioada de spitalizare este de 6 zile sau mai mare prin aceea că, în cazul pacienților mai mici de 6 ani, costul pentru tratament este mai ridicat, în medie.

Pentru arborele de mai sus, algoritmul a ales implicit parametrul de complexitate, în urma calculelor expuse grafic mai jos:



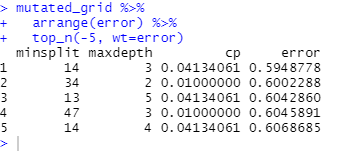
Prin urmare, am creat un arbore în care am setat valoarea parametrului α=0, rezultat fiind un arbore de o dimensiune sensibil diferită.



Analizând acest arbore, observăm că prin împărțirea s-a realizat tot în raport cu variabilele *los* și *price*, dar a fost adăugat în plus și variabila *female*. La nivelul cel mai de jos, frunzele nodului arată valoarea prezisă de arbore. Un astfel de arbore nu este practic, motiv pentru care vom stabili manual anumite criterii în raport cu care să se facă împărțitea. Astfel, cu parametrii minsplit (numărul minim de instanța dintr-un nod necesar pentru ca împărțirea să continue) și maxdept (adâncimea maximă a arborelui, adică numărul dintre rădăcină și frunze) am procedat la crearea altui arbore.

Parametrii au fost selectați în urma rezultatelor arborelui în care parametrul de complexitate α = 0. Astfel, s-a observat că majoritatea nodurilor aveau între 10-30 de instanție, motiv pentru care minsplit a fost selectat pentru o variație între 10 – 30, iar adăncimea arborelui a fost redusă la cel mul 5 pentru micșora dimensiunile acestuia.

În continuare, am creat modelele de arbori printr-un loop de tip for care au fost introduse într-un hyper\_grid. Mai apoi, am stabilit valorile pentru parametrul de complexitate și eroare minimă, după care am afișat cele mai bune cinci modele (adică acelea care au parametrul α și eroarea cele mai mici):



Așadar, modelul cu cele mai bune valori este cel în care minspilt este 14, iar maxdepth este 3. Adică arborele care conține un număr minim de 14 instanțe pentru fiecare nod, și cel mult trei noduri de la rădăcină la frunze. De asemenea, se observă că eroarea este mai mare decât cea înregistrată de arborele în care parametrul complexity a fost 0.

Diagram

Description automatically generated

Asemănător arborelui analizat în precedent, variabila cea mai importană în raport cu care se face împărțirea este numărul de zile de îngrijiri medicale. De asemenea, pentru pacienții în care perioada de spitalizare mai mică de 6 zile valoarea medie a costului de spitalizare este de 2112 u.m., în timp ce pentru ceilalți este de 13125 u.m. Ca diferență, se observă că vârsta pacientului nu mai reprezintă criteriu pentru 93% dintre pacienți, ci doar pentru cei care beneficiază de tratament pentru o perioadă mai îndelungată.

Timeline

Description automatically generated with medium confidence

Pentru a testa acuratețea modelului, am efectuat o predicție pornind de la cele stabilite mai sus.



Astfel, a rezultat RMSE (eroarea pătratică medie) în valoare de 2903,627. În alte cuvinte, această metodă poate precize costul de spitalizare cu o eroare medie de 2903,627 u.m. În raport cu aceste date putem concluziona că predicțiile oferite de acest model nu se bucură de o fiabilitate ridicată.

**III.2.3. Îmbunătățirea modelului**

Deoarece arborii de regresie singulari au o varianța mare (nefiind robuști), performanța lor predictivă poate fi îmbunătățită prin apelarea la proceduri avansate precum bagging.

Arborele de decizie este dependent de setul de antrenament creat, ceea ce presupune că schimbarea unei părți din instanțele setului de date poate genera un arbore total diferit. Pentru a evidenția care dintre arbori este caracterizat de o acuratețe mai bună, folosim procedura avansată bagging.

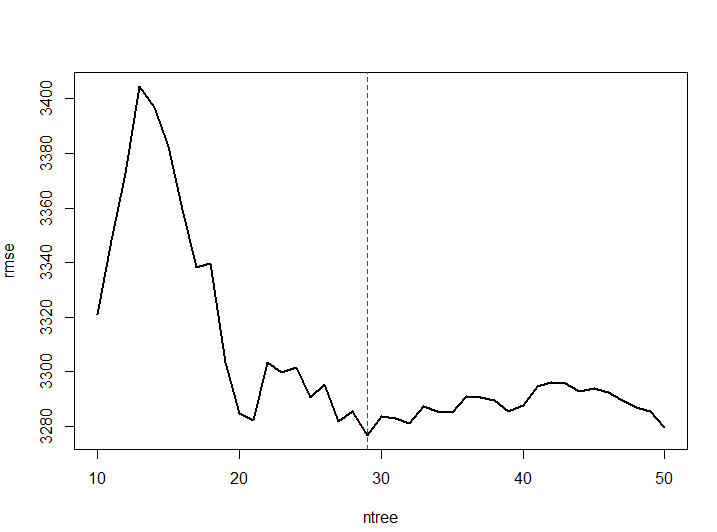
Această procedură extrage submulțimi din setul de antrenament, creeză modele, după care calculează media tuturor predicțiilor modelelor create. Varianța fiecărui model privit individual este redusă prin calcularea mediilor acestora.

Inițial, am creat un set de arbori prin **metoda bagging folosind setările implicite**, după care am testat acuratețea modelului. Rezultate sunt următoarele:



Așadar, s-a constat o reducere a erorii medii pătratice (RMSE) de la 2903,627 la 2406,034. În alte cuvinte, această metodă poate precize costul de spitalizare cu o eroare medie de 2406,034 u.m. În raport cu aceste date putem concluziona că predicțiile oferite de acest model nu se bucură de o fiabilitate ridicată.

În continuare, am creat un set de arbori prin metoda **bagging, modificând numărul de bagguri**. Astfel, am stabilit că numărul maxim de bagguri care pot conduce la reducerea erorii este de 29, după cum reiese din graficul de mai jos:

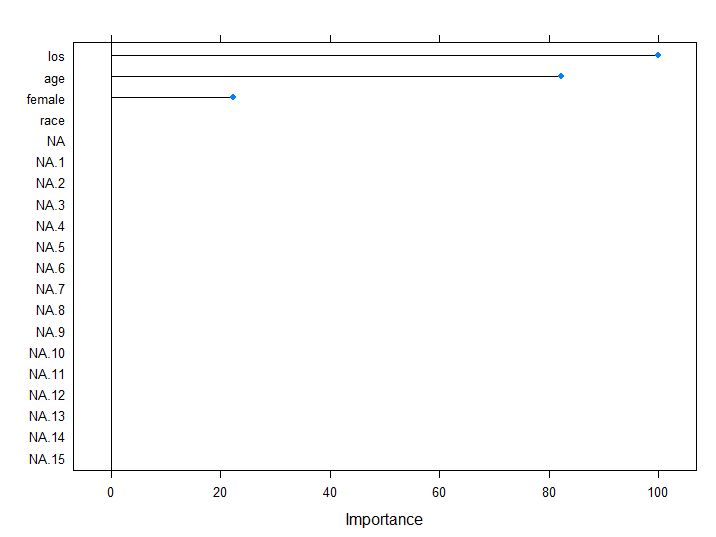


Mai apoi,am testat acuratețea modelului. Rezultate sunt următoarele:

****

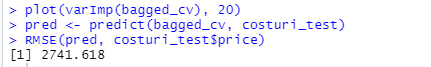
Așadar, s-a constat o creștere a erorii medii pătratice (RMSE) de la 2406,034 la 3811,092. În alte cuvinte, această metodă poate precize costul de spitalizare cu o eroare medie de 3811,092 u.m. În raport cu aceste date putem concluziona că predicțiile oferite de acest model nu se bucură de o fiabilitate ridicată.

În fine, am mai creat un set de arbori folosind procedura **bagging alături de metoda CARET.** Folosind această procedură am stabilit importanța atributelor luând în calcul toți arborii creați prin această metodă. Rezultatele sunt prezentate în figura de mai jos:

****

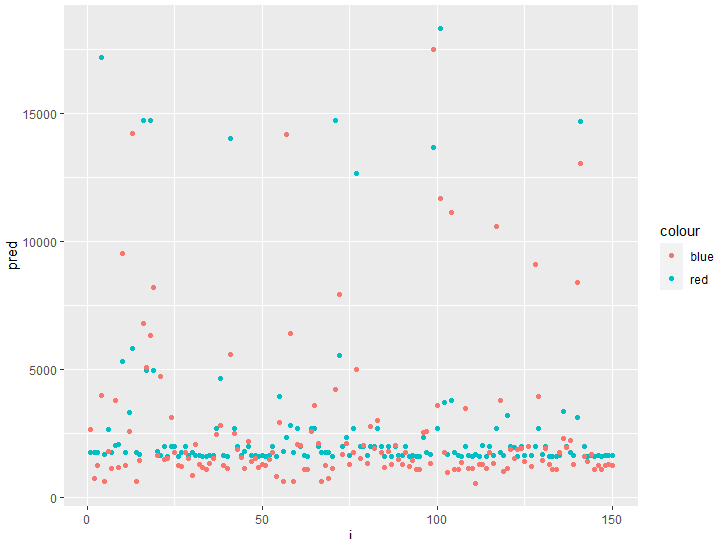
Astfel, a reieșit că cea mai importantă variabilă este *los,* numărul de zile de internare influențând cel mai mult eroarea. Mai apoi, vârsta pacientului are un aport important, în timp ce variabila female are un efect redus. Se observă că variabila race nu afectează în niciun fel acuratețea modelelor.

Mai apoi,am testat acuratețea modelului. Rezultate sunt următoarele:

****

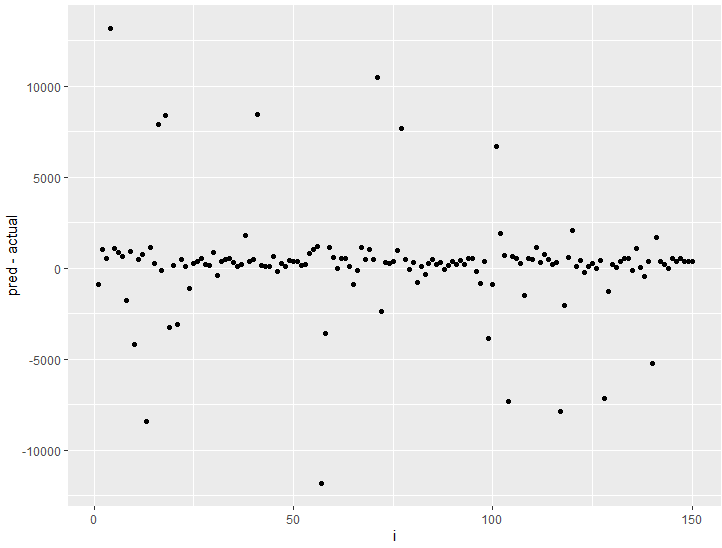
Așadar, eroarea medie pătratică înregistrată de acest model de 2741,618. S-a constat o scădere a erorii medii pătratice (RMSE) față de modelul anterior (3811,092), dar o creștere în raport cu modelul creat prin procedura baggind implicită (2406,034). În alte cuvinte, această metodă poate precize costul de spitalizare cu o eroare medie de 2741,618 u.m. În raport cu aceste date putem concluziona că predicțiile oferite de acest model nu se bucură de o fiabilitate ridicată.

Pentru a ilustra acuratețea acestui model am reprezentat grafic, cu culoare roșie, predicțiile modelului, și cu albastru, valorile reale ale instanțelor.



Se observă că există diferențe între valorile înregistrate de instanțele reale și cele prezise, uneori însemnate.

În fine, am exprimat grafic acuratețea modelului și prin afișarea distribuției predicțiilor în raport cu valoarea 0. Rezultate sunt următoarele:



După cum se observă, majoritatea predicțiilor sunt în jurul valorii de 0, dar există și cazuri când valorea erorii este semnificativă.

**III.2.4. Concluzii în raport cu metoda arborelui de decizie**

*1. Există o legătură între vârsta, sexul, numărul de zile de spitalizare și rasa pacientului și costul de spitalizare ? Dacă răspunsul este pozitiv, cât de puternică este această legătură ?*

Metoda arborelui de deciziie a relevat că există o legătură între vârsta și numărul de zile de spitalizare pentru determinarea costului acordării de îngrijiri medicale. Variabilele privind sexul și rasa pacientului nu au semnificație importantă în stabilirea valorii tratamentului acordat.

*2. Este cel puțin unul dintre predictori relevant pentru a prezice costul de spitalizare ?*

Numărul de zile de îngrijiri medicale poate fi un predicator relevant având în vedere că în toți arborii creați a constituit variabila cea mai importantă. S-a observat că, pentru perioadele de spitalizare mai mari de 5 zile, costul mediu al asistenței medicale crește în mod semnificativ.

*3. Există posibilitatea estimării costului de spitalizare dacă se cunosc vârsta, sexul, numărul de zile de spitalizare și rasa pacientului ?*

Există posiblitatea efectuării unei predicții, dacă se cunosc variabilele.

*4. Dacă răspunsul la întrebarea anterioară este pozitiv, cât de precisă este estimarea respectivă ?*

Având în vedere valoarea ridicată a parametrului RMSE, reținem că predicția nu se bucură de un nivel ridicat de precizie.

1. Relația sau corelația negativă are loc când două variabile ale unei funcții se mișcă în direcții opuse. Pe măsură ce valoarea lui x crește, valoarea lui y scade și viceversa. Existența unei corelații negative indică o relație semnificativă între cele două variabile. [↑](#footnote-ref-1)
2. Relația sau corelația negativă are loc când două variabile ale unei funcții se mișcă în direcții opuse. Pe măsură ce valoarea lui x crește, valoarea lui y scade și viceversa. Existența unei corelații negative indică o relație semnificativă între cele două variabile. [↑](#footnote-ref-2)