Proiect la disciplina Fundamente de Big Data

Tătar Gabriela-Diana, [gabriela.tatar@stud.ubbcluj.ro](mailto:gabriela.tatar@stud.ubbcluj.ro)

Vlad Irina, [irina.vlad@stud.ubbcluj.ro](mailto:irina.vlad@stud.ubbcluj.ro)

Facultatea de Științe Economice și Gestiunea Afacerilor

Informatică Economică

An: III

2022 - 2023

Analiza costurilor medicale

**Introducere**

Costurile medicale sunt identificate ca fiind cheltuieli implicate în asigurarea de servicii medicale, achiziționarea de medicamente și dispozitive medicale necesare pentru tratamentul pacienților. Costurile respective sunt suportate de către persoanele aflate în cauză, guvern, companiile de asigurări și alți actori care aparțin de sistemul de asigurare.

Aceste costuri includ în mare parte și cheltuielile pentru consultări medicale, analize de laborator, intervenții chirurgicale, medicamente, terapii, dar și alte servicii. Datorită unor anumiți factori, cum ar fi țara de proveniență a omului, condiția medicală a sa, sistemul de sănătate, aceste costuri pot varia semnificativ.

Asigurarea medicală servește la compensarea asiguratului în mod parțial sau integral de costurile medicale, în caz de accident de sănătate, dar și a femeilor gravide. Asigurarea de sănătate de stat se efectuează prin înscrierea unei persoane fizice la un medic de familie. Aceasta persoană trebuie să realizeze venituri din activități unde este obligat prin lege să plătească îndatoririle prin plata contribuției, care acoperă asigurarea obligatorie de sănătate. Altă asigurare de sănătate este cea privată, care poate fi achiziționată de către angajator plătind o taxă lunară integral sau parțial.

Este importantă cumpărarea unei asigurări medicale, deoarece are scopul de a oferi acces unor servicii medicale de calitate, responsabilitate socială menținând un sistem de sănătate sustenabil și protecție bugetară, pentru că tarifele unor servicii pot fi foarte ridicate și greu de suportat din punct de vedere financiar.

Cum influențează acești factori costurile medicale?

Printre factorii pe care îi vom studia se numără vârsta, sexul, regiunea de reședință (din SUA), dacă subiectul este fumător sau nu și indicele de masă corporal. Aceștia sunt factorii primari și cei mai relevanți care influențează costurile medicale.

În Statele Unite ale Americii, asigurarea medicala (medic aid) acoperă majoritatea medicamentelor cu prescripție, un aspect care influențează major costurile medicale finale. Medicamentele prescrise sunt diferite în funcție de factorii enumerați mai sus, având scopuri și cantități unice în funcție de sex, vârstă sau indicele de masă corporal. În ceea ce privește sexul persoanei, este mai probabil pentru femei decât pentru bărbați să achiziționeze medicamente. Aceste diferențe apar datorită unor medicamente cu scop unic în funcție de sex și o utilizare pe termen lung, cum ar fi pastilele contraceptive sau pastilele pentru sănătatea aparatului reproduător, fiind mai predispuse la afecțiuni ale acestuia. De asemenea, vârsta este factorul cu o influență majoră asupra costurilor medicale, având în vedere problemele apărute odată cu înaintarea în vârstă, aceste probleme cauzând o dependență mai mare de medicamente sau spitalizare pentru boli precum cele de inima sau de oase. Obezitatea sau anorexia sunt două cazuri extreme care vin la pachet cu probleme proprii, astfel ca indicele de masa corporal influențează la rândul lui toate aceste sume. Anorexia poate necesita o spitalizare îndelungată pentru revenirea la un indice de masă corporal normal, care rezulta în cheltuieli majore. Obezitatea poate cauza o mulțime de probleme, cum ar fi diabetul sau chiar probleme care necesită intervenții chirurgicale. Fumatul crește la rândul lui riscul bolilor de inima și afecțiunilor aparatului respirator. Regiunea de proveniență poate impune diferite politici asupra asigurărilor medicale, fapt care infleunțează sumele finale.

Toate acestea sunt exemple care rezuma în cateva cuvinte influența acestori factori asupra costurilor medicale și de ce sunt atât de importanți în cercetarea noastră.

Ne propunem să realizăm o predicție numerică pe setul de date ales, pentru a observa mai atent si mai precis cât de mult influențează acești factori costurile medicale. Considerăm că această analiză poate fi una folositoare atât pentru asigurat, pentru a-și putea estima singur valoarea investiției sale, cât și pentru firmele de asigurări medicale, pentru a putea decide mai corect un preț pentru asigurările puse la dispoziție. În continuare, vom încerca sa răspundem la următoarele întrebări de cercetare:

1. Există o corelație între prețul final al asigurării medicale și vârsta, sexul, indicele de masă corporală, numărul de copii, a fi fumător sau nu, ori regiunea de reședință?
2. Dacă într-adevăr există o legătură, cât de semnificativă este aceasta?
3. Care dintre acești factori are influența cea mai mare?
4. Cât de bine putem prezice costurile medicale viitoare pe baza acestor factori?

**Setul de date**

În vederea cercetării noastre, am ales următorul set de date preluate de pe <https://www.kaggle.com/datasets/mirichoi0218/insurance>, set care conține toate informațiile necesare pentru a realiza o analiză completă. Acesta include un număr total de 1338 de indivizi și caracteristicile acestora, precum:

**Variabilele independente**:

Age: vârsta asiguratului exprimată în ani.

Sex: sexul asiguratului, și anume *female* sau *male*.

BMI: indicele de masa corporal, exprimat în cifre.

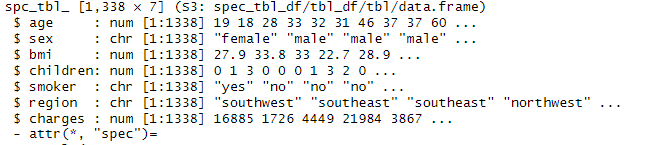
Children: numărul de copii, exprimat în cifre, care are o valoare de maxim 5.

Smoker: dacă asiguratul este fumător sau nu, exprimat într-o valoare binară de *yes* sau *no*.

Region: împărțită în 4 mari categorii, și anume *northwest*, *southwest*, *southeast* si *northeast*.

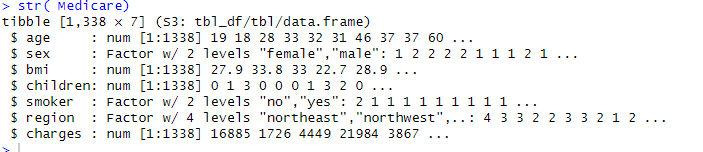
**Variabila dependenta**:

Charges: prețul asigurării medicale.



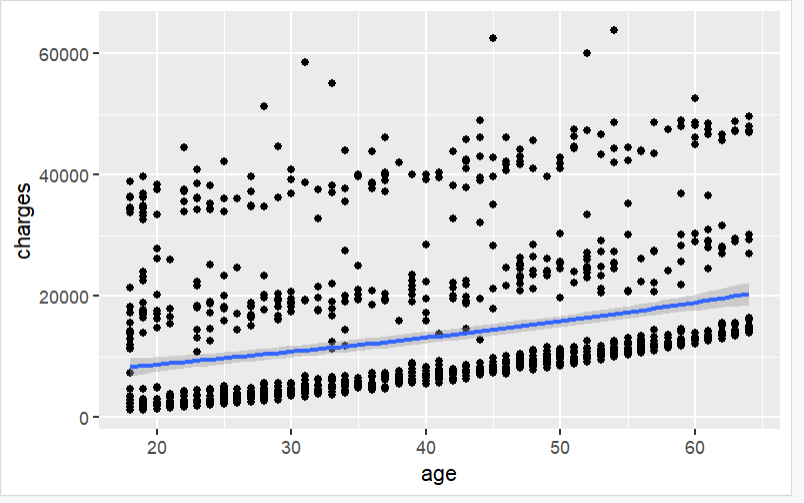
Pentru a putea antrena corect setul nostru de date, folosind predicția numerică, dar și pentru a observa posibilitatea unei relații cu ajutorul graficelor, va trebui să aducem modificări, astfel că variabilele nominale vor fi transformate în variabile de tip factor, cu ajutorul funcției:

Medicare <- mutate\_at(Medicare, vars(sex,smoker,region), as.factor)



Pentru a analiza mai bine datele din setul ales, vom lua fiecare variabila independentă în parte si vom crea mai multe grafice pentru a decide relevanța acestora în studiul nostru.

Figura 1

Din *Figura 1* putem observa că proprietatea **age** oferă o variație semnificativă costurilor medicale, dar și o creștere ușoara a costurilor asigurării cu vârsta. Totodată, se poate observa că punctele sunt dispersate destul de mult în jurul liniei regresiei, fapt care indică și mai mult relevanța acestui factor în suma finală a asigurării.

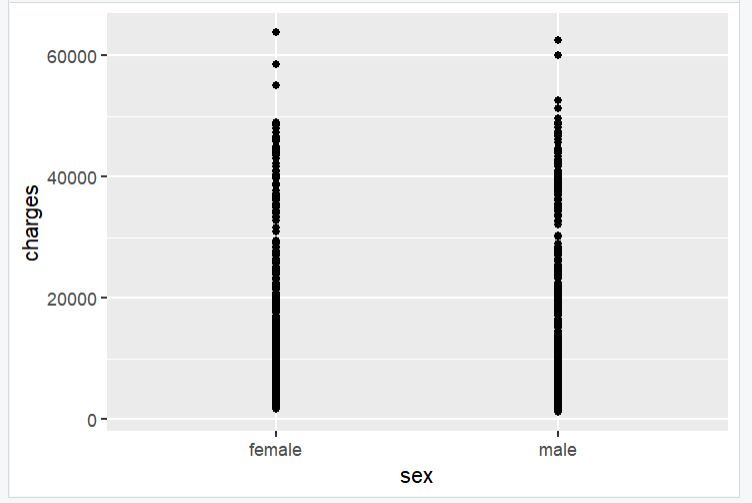
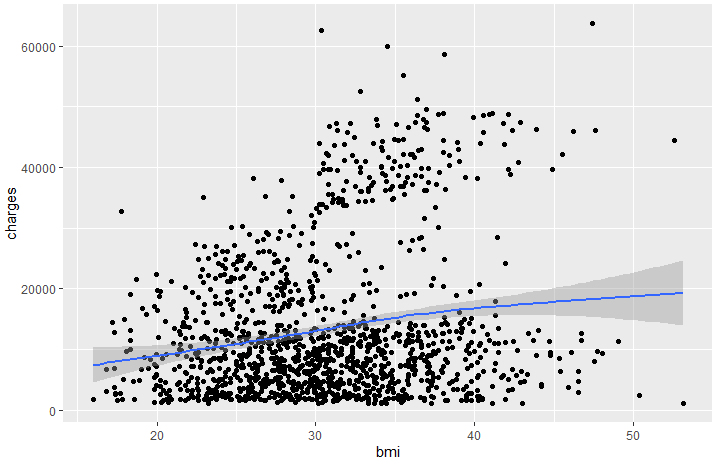
Din *Figura 2* se poate observa o suprapunere destul de mare asupra celor două valori, *male* și *female*, ceea ce indică faptul că variabila **sex** nu ar fi foarte relevantă într-o regresie liniară simplă, de exemplu, dar ar putea să reprezinte o influență asupra altor variabile independente (cum ar fi într-o regresie liniară multiplă). Cu toate acestea, putem observa costurile care ajung la sume mai mari în cazul sexului feminin.

Figura 2



În *Figura 3* este reprezentată variabila **BMI** si corelația acesteia cu variabila dependentă. Și această variabilă este una semnificativă, observând acest aspect din dispersia majoră a punctelor în jurul liniei regresiei, precum și din creșterea ușoară a liniei regresiei și a costurilor medicale odată cu un BMI mai crescut.

Figura 3

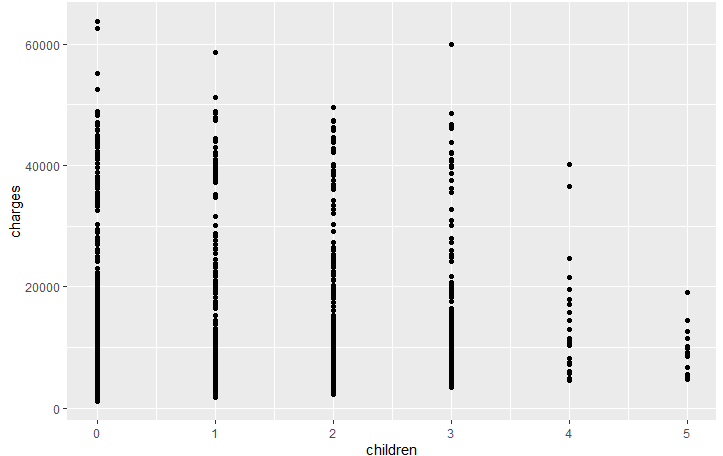
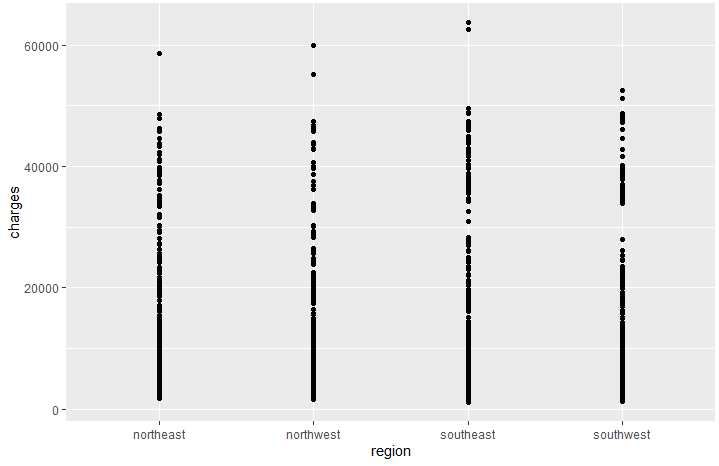
În *Figura 4* am analizat importanța variabilei **children**. Din păcate, nu putem stabili o tendință clara, observând ca în cazul asiguraților cu un număr crescut de copii, costurile medicale sunt mai scăzute, contrar așteptărilor. De asemenea, există foarte multe suprapuneri, indicând faptul că nu există o influență atât de mare din partea acestei variabile asupra costurilor medicale.

Figura 4



În *Figura 5* observăm că regiunile au un nivel de suprapunere mare, ceea ce face ca variabila **region** să nu aibă o influență atât de mare asupra costurilor medicale, rezultând ca acestea sunt în mare parte constante indiferent de regiunea din care face parte asiguratul.

Figura 5

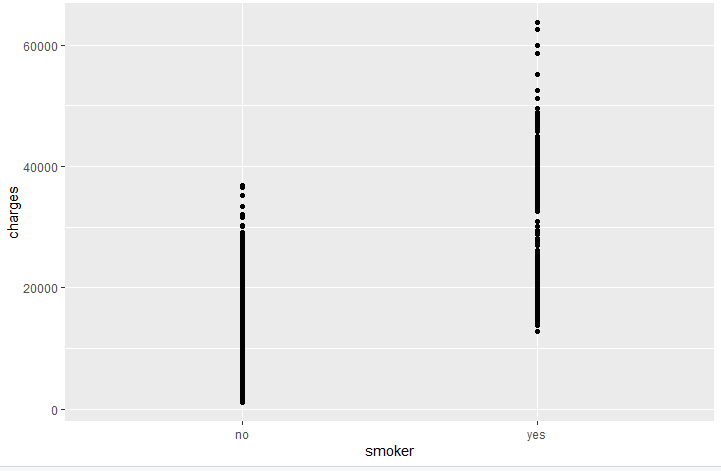


Figura 6

În *Figura 6* am analizat influența factorului **smoker**. Observăm clar ca fumătorii prezintă costuri medicale mai ridicate, dar că există și o arie de suprapunere. Costurile fumătorilor au un punct de start mai ridicat decât cel al nefumătorilor.

**Interpretarea datelor**

Regresia liniară simplă

Primul pas în analiza noastră, este să antrenăm datele. Pentru partea de antrenament a setului nostru de date, am ales sa folosim mai multe metode de învățare. Având in vedere graficele create anterior, am observat ca se poate stabili o relație liniara între variabilele independente și variabila dependentă, astfel că am ales ca prima metoda de învățare regresia liniară (sau media celor mai mici pătrate).

Folosind Rstudio, am creat funcții de regresie liniară, atât simplă cât și multiplă, pentru a stabili cu certitudine cât de puternice sunt relațiile dintre variabilele independente și variabila dependentă.

Primul pas a constat în determinarea și interpretarea coeficienților regresiei liniare, și anume intercept (care determină valoarea medie a variabilei dependente când variabila independentă este nulă) și panta (care determină valoarea medie a variabilei dependente când variabila independentă crește cu o unitate). Pentru fiecare coeficient al fiecărui model de regresie, vom stabili intervale de încredere (CI) cu o încredere de 95%, determinând astfel dacă acești parametrii resping sau nu ipoteza nulă. Calculăm intervalele de încredere pentru a putea prezice unde se va situa variabila dependentă în funcție de starea variabilei independente. De exemplu, pentru o regresie simplă, vom avea două intervale de încredere, unul pentru fiecare coeficient al regresiei. Formula pe care o vom folosi pentru a afla intervalele de încredere este:

βi ∈ [βˆ i − 2SE(βˆ i ), βˆ i + 2SE(βˆ i )]

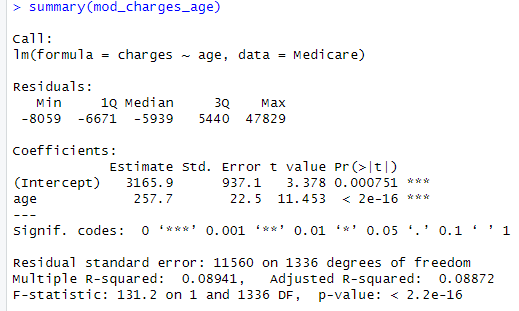
Pentru a ajunge la o concluzie corectă în ceea ce privește parametrii regresiei, vom lua în considerare mai mulți factori, printre aceștia numarându-se t-value, p-value si stde. Error.

Având in vedere aceste aspecte, vom aplica metoda regresiei liniare pentru fiecare variabila independenta, începând prin a stabili care dintre ipoteze este cea acceptată.

**Variabila independentă AGE**

Primul predictor pe care îl vom analiza în relația cu variabila independentă este age. Calculăm coeficienții regresiei, B0 si B1, având în considerare că charges ≈ β0 + β1 × age. Așadar, în Rstudio, introducem următoarea funcție:

mod\_charges\_age <- lm(data = Medicare, charges ~ age)

**Rezultatul în urma executării liniei de cod anterioare:

Intercept: valoarea medie a variabilei dependente când variabila independentă este nulă este 3165.9 u.m.

age: valoarea medie a variabilei dependente crește cu 257.7 u.m. atunci când variabila age crește cu o unitate.

În continuare vom testa ipoteza nulă și cea alternativă:

H0 (ipoteza nulă): nu se identifică nicio relație între X si Y. Testăm acest lucru prin a verifica dacă B1 = 0.

H1 (ipoteza alternativă): exista o relație între X si Y. Testăm acest lucru prin a verifica daca B1 este diferit de 0.

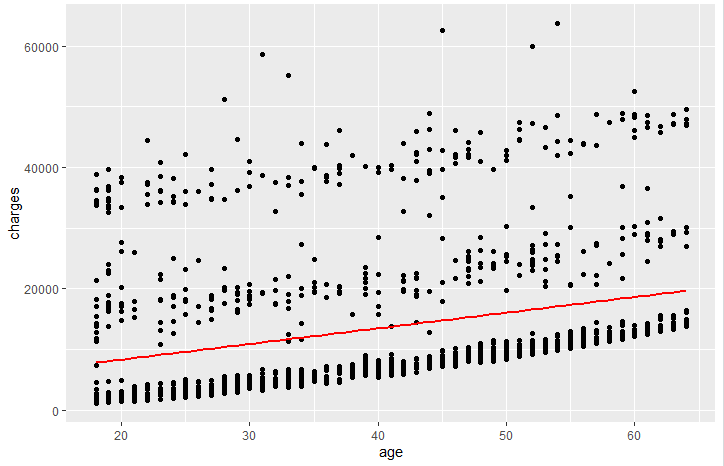
În urma acestor rezultate, putem observa că t-statistics pentru coeficientul de regresie are o valoare cu mult peste valoarea critică, și anume 11.453. De asemenea p-value are o valoare foarte apropiata de 0, astfel ca p-value < 2e-16. Putem concluziona că coeficientul este semnificativ diferit de 0 (argument întărit de valoarea lui t-statistics) și că această valoare nu este datorată șansei (argument întărit de valoarea lui p-value).

La nivel de model, putem măsura calitatea acestuia prin intermediul valorilor lui F-Statistics si p-value, dar și R-Squared si RSE. Dat fiind faptul ca F-statistic este o valoare mult mai mare decât 1, mai exact 131.2, duce la o cât mai bună ajustare a modelului pentru datele analizate. În plus, p-value reprezintă o valoare foarte mică (<2.2e-16), ceea ce sugerează că F-statistic este foarte puțin probabil să demonstreze să fie adevărată ipoteza nulă. Totuși, R-Squared are o valoare foarte apropiata de 0 (8.94%). În concluzie, deși rezultatele la nivel de model susțin ipoteza alternativă si demonstrează că există o relație între variabilele independente și cea dependentă, această relație nu este una foarte puternică.

**Urmează să identificăm intervalele de încredere a acestor coeficienți pe baza formulei aferente, ajutându-ne de valorea erorii standard. In R, am introdus comanda confint(mod\_charges\_age) pentru a identifica mai ușor capetele intervalelor.

Pentru β0, CI = [1327.4403, 5004.3297], așadar costul final al asigurării se va încadra în acest interval dacă nu luăm in calcul vârsta persoanei asigurate.

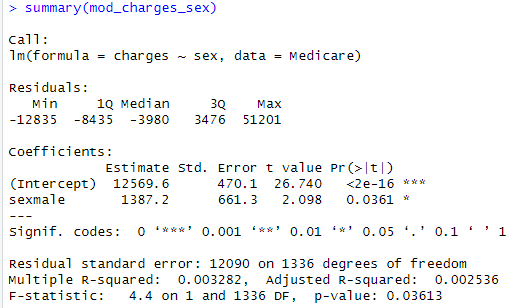
Pentru β1, CI = [213.5788 301.8665], așadar creșterea cu o unitate a vârstei persoanei asigurate va rezulta într-o creștere a asigurării situată între 213.5788 u.m și 301.8665 u.m.

**Având în vedere modelul antrenat anterior, putem construi pe baza acestuia o grilă nouă cu 100 de valori ale variabilei age, pentru a prezice valorile viitoare ale variabilei dependente charges în funcție de variabila dependentă age. Aceasta ia valori intre 0 si 65, iar variabila charges ia valori intre 0 si 60000. Observăm și în acest grafic relația liniară dintre cele două, cu o ușoară creștere a costurilor medicale atunci când asiguratul înaintează în vârstă. De asemenea, există o dispersie mare când costurile trec de suma de 1000 u.m.

**Variabila independentă SEX**

Cel de-al doilea predictor asupra căruia vom realiza analiza este sex. Calculăm coeficienții regresiei, B0 si B1, având în considerare că charges ≈ β0 + β1 × sex. Așadar, în Rstudio, introducem următoarea funcție:

mod\_charges\_sex <- lm(data = Medicare, charges ~ sex)

 Rezultatul în urma executării liniei de cod anterioare:

Intercept: valoarea medie a variabilei dependente când variabila independentă este nulă este 12569.6u.m.

sexmale: valoarea medie a variabilei dependente crește cu 1387.2 u.m. atunci când asiguratul este bărbat.

În continuare vom testa ipoteza nulă și cea alternativă:

H0 (ipoteza nulă): nu se identifică nicio relație între X si Y. Testăm acest lucru prin a verifica dacă B1 = 0.

H1 (ipoteza alternativă): exista o relație între X si Y. Testăm acest lucru prin a verifica daca B1 este diferit de 0.

În urma acestor rezultate, putem observa că t-statistics pentru coeficientul de regresie are o valoare peste valoarea critică, dar nu cu foarte mult, și anume 2.098. De asemenea p-value are o valoare apropiată de 0 și sub nivelul critic de 0.05, mai exact 0.0361, însă nu la fel de considerabil apropiata de 0 ca la modelul anterior. Putem concluziona că coeficientul este într-adevăr diferit de 0, (argument întărit de valoarea lui t-statistics), însă această valoare ar putea fi datorata șansei, având in vedere p-value puțin mai crescut.

În ceea ce privește calitatea modelului, F-Statistics trece foarte puțin peste pragul critic, cu o valoare de 4.4 și un p-value de 0.0361, iar R-Squared are de asemenea o valoare foarte mică (0.003282), lucru care ne arată că doar aproximativ 0,32% din variația variabilei dependente depinde de variabila sex. Respingem așadar ipoteza nulă, deoarece există totuși o relație între cele două variabile, dar această relație nu este una semnificativă, existând variabile cu o influență mai mare.

Urmează să identificăm intervalele de încredere a acestor coeficienți pe baza formulei aferente, ajutându-ne de valoarea erorii standard. In R, am introdus comanda confint(mod\_charges\_sex) pentru a identifica mai ușor capetele intervalelor.



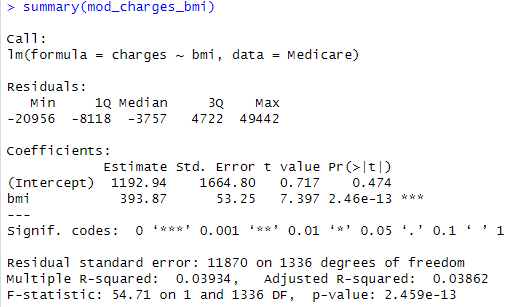
Pentru β0, CI = [11647.41984 13491.738], așadar costul final al asigurării se va încadra în acest interval dacă nu luăm în calcul sexul persoanei asigurate.

Pentru β1, CI = [89.81229 2684.532], dacă asiguratul este bărbat, va determina o valoare medie a costurilor medicale situată între 89.81229 u.m. și 2684.532 u.m.

**Variabila independentă BMI**

Al treilea predictor pe care ne vom axa este bmi. Calculăm coeficienții regresiei, B0 și B1, având în considerare că charges ≈ β0 + β1 × bmi. Așadar, în Rstudio, introducem următoarea funcție:

mod\_charges\_bmi <- lm(data = Medicare, charges ~ bmi)

Rezultatul în urma executării liniei de cod anterioare:

Intercept: valoarea medie a variabilei dependente când variabila independentă este nulă este 1192.94 u.m.

bmi: valoarea medie a variabilei dependente crește cu 393.87 u.m. atunci când variabila bmi crește cu o unitate.

În continuare vom testa ipoteza nulă și cea alternativă:

H0 (ipoteza nulă): nu se identifică nicio relație între X si Y. Testăm acest lucru prin a verifica dacă B1 = 0.

H1 (ipoteza alternativă): exista o relație între X si Y. Testăm acest lucru prin a verifica daca B1 este diferit de 0.

În urma acestor rezultate, putem observa că t-statistics pentru coeficientul de regresie are o valoare cu mult peste valoarea critică, și anume 7.397. De asemenea p-value are o valoare foarte apropiată de 0, și anume 2.46e-13. Putem concluziona că coeficientul este semnificativ diferit de 0 (argument întărit de valoarea lui t-statistics) și că această valoare nu este datorată șansei (argument întărit de valoarea lui p-value).

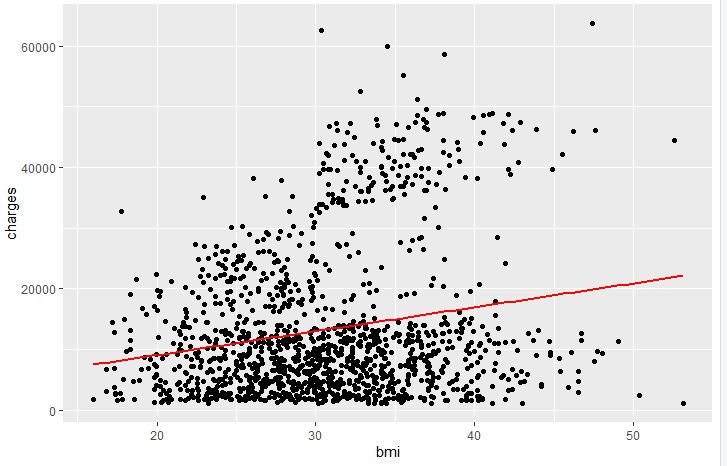
În ceea ce privește modelul, putem măsura calitatea acestuia prin intermediul valorilor lui F-Statistics si p-value, dar și R-Squared si RSE. Dat fiind faptul ca F-statistic este o valoare mult mai mare decât 1, mai exact 54.71, duce la o cât mai bună ajustare a modelului pentru datele analizate. În plus, p-value reprezintă o valoare foarte mică (<2.2e-16), ceea ce sugerează că F-statistic este foarte puțin probabil să demonstreze să fie adevărată ipoteza nulă. Cu toate acestea, R-Squared are o valoare foarte apropiata de 0 (aprox. 3.93%). În concluzie, deși rezultatele la nivel de model susțin ipoteza alternativă și demonstrează că există o relație între variabilele independente și cea dependentă, această relație nu este una foarte puternică.

Urmează să identificăm intervalele de încredere a acestor coeficienți pe baza formulei aferente, ajutându-ne de valoarea erorii standard. În R, am introdus comanda confint(mod\_charges\_bmi) pentru a identifica mai ușor capetele intervalelor.



Pentru β0, CI = [-2072.9743 4458.8487], așadar costul final al asigurării se va încadra în acest interval dacă nu luăm în calcul indicele de masă corporal al persoanei asigurate.

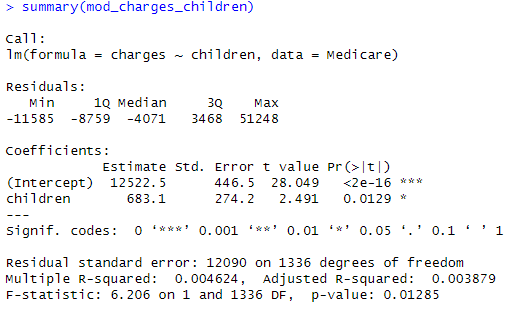
Pentru β1, CI = [289.4089 498.3372], creșterea cu o unitate a indicelui de masă corporal va determina o valoare medie a costurilor medicale situată între 289.4089 u.m. și 498.3372 u.m.

**Pe baza acestui model de antrenament, putem construi o grilă nouă cu 100 de valori ale variabilei bmi, pentru a prezice valorile viitoare ale variabilei dependente charges în funcție de variabila dependentă bmi. Aceasta ia valori între 0 și 55, iar variabila charges ia valori între 0 și 60000. Se poate observa și în acest grafic relația liniară dintre cele două, cu o ușoară creștere a costurilor medicale pentru asigurații cu un indice mărit de masă corporală. De asemenea, există o dispersie considerabilă a costurilor.

**Variabila independentă CHILDREN**

Cel de-al patrulea predictor asupra căruia vom realiza analiza este sex. Calculăm coeficienții regresiei, B0 și B1, având în considerare că charges ≈ β0 + β1 × children. Așadar, în Rstudio, introducem următoarea funcție:

mod\_charges\_children <- lm(data = Medicare, charges ~ children)

Rezultatul în urma executării liniei de cod anterioare:

Intercept: valoarea medie a variabilei dependente când variabila independentă este nulă este 12522.5 u.m.

children: valoarea medie a variabilei dependente crește cu 683.1 u.m. atunci când numărul de copii crește cu unu.

În continuare vom testa ipoteza nulă și cea alternativă:

H0 (ipoteza nulă): nu se identifică nicio relație între X si Y. Testăm acest lucru prin a verifica dacă B1 = 0.

H1 (ipoteza alternativă): exista o relație între X si Y. Testăm acest lucru prin a verifica daca B1 este diferit de 0.

În urma acestor rezultate, putem analiza t-statistics pentru coeficientul de regresie, care are o valoare peste valoarea critică, dar nu semnificativ peste, și anume 2.491. De asemenea p-value are o valoare apropiată de 0 și sub nivelul critic de 0.05, mai exact 0.0129, însă nu atât de apropiată de 0 pe cât ne-am aștepta. Putem concluziona că coeficientul este într-adevăr diferit de 0 și că există o relație (argument întărit de valoarea lui t-statistics), însă această valoare ar putea fi datorată șansei, având în vedere valoarea nu atât de mică pe care o are p-value.

Din punct de vedere al calității modelului, F-Statistics trece foarte puțin peste pragul critic, cu o valoare de 6.206 și un p-value de 0.01285, iar R-Squared are de asemenea o valoare foarte mică (0.004624), lucru care ne arată că doar aproximativ 0.46% din variația variabilei dependente depinde de variabila children. Respingem așadar ipoteza nulă, deoarece există totuși o relație între cele două variabile, dar această relație nu este una semnificativă, existând variabile cu o influență mai mare.

Urmează să identificăm intervalele de încredere a acestor coeficienți pe baza formulei aferente, ajutându-ne de valoarea erorii standard. În R, am introdus comanda confint(mod\_charges\_children) pentru a identifica mai ușor capetele intervalelor.



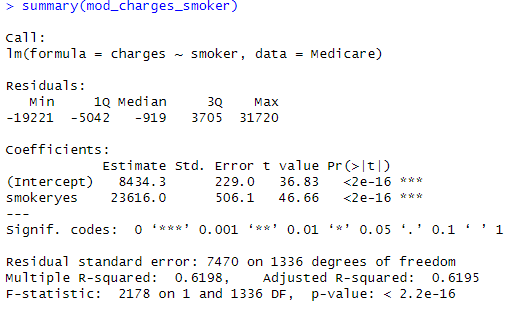
Pentru β0, CI = [11646.6761 13398.315] așadar costul final al asigurării se va încadra în acest interval dacă nu luăm în calcul numărul de copii ai persoanei asigurate.

Pentru β1, CI = [145.1763 1221.002], dacă numărul de copii al asiguratului crește cu o unitate va determina o valoare medie a costurilor medicale situată între 145.1763 um. și 1221.002 um.

**Variabila dependentă SMOKER**

Cel de-al cincilea predictor asupra căruia vom realiza analiza este sex. Calculăm coeficienții regresiei, B0 si B1, având în considerare că charges ≈ β0 + β1 × smoker. Așadar, în Rstudio, introducem următoarea funcție:

mod\_charges\_smoker <- lm(data = Medicare, charges ~ smoker)

Rezultatul în urma executării liniei de cod anterioare:

Intercept: valoarea medie a varibilei dependente când variabila independentă este nulă este 8434.3u.m.

smokeryes: valoarea medie a variabilei dependente crește cu 23616.0 u.m. dacă asiguratul este fumător.

În continuare vom testa ipoteza nulă și cea alternativă:

H0 (ipoteza nulă): nu se identifică nicio relație între X si Y. Testăm acest lucru prin a verifica dacă B1 = 0.

H1 (ipoteza alternativă): exista o relație între X si Y. Testăm acest lucru prin a verifica daca B1 este diferit de 0.

În urma acestor rezultate, putem analiza t-statistics pentru coeficientul de regresie, care are o valoare peste valoarea critică, mai exact 46.66. De asemenea p-value are o valoare foarte apropiată de 0, mai exact <2e-16. Putem concluziona că coeficientul este într-adevăr diferit de 0 și că există o relație și că această valoare nu este una întâmplătoare.

În ceea ce privește modelul, putem măsura calitatea acestuia prin intermediul valorilor lui F-Statistics si p-value, dar și R-Squared și RSE. Dat fiind faptul ca F-statistic este o valoare mult mai mare decât 1, mai exact 2178, duce la o cât mai bună ajustare a modelului pentru datele analizate. În plus, p-value reprezintă o valoare foarte mică (<2.2e-16), ceea ce sugerează că F-statistic este foarte puțin probabil să demonstreze să fie adevărată ipoteza nulă. De asemenea, R-Squared are o valoare foarte apropiata de 1 (0.6198 sau 61.98%), indicând o acuratețe mare. În concluzie, putem infirma ipoteza nulă, deoarece există o relație puternică între aceste variabile.

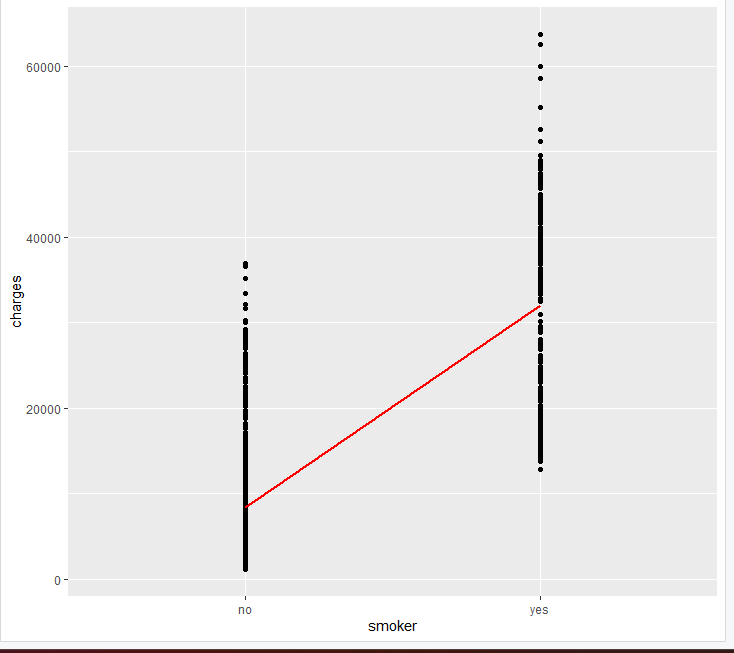
Urmează să identificăm intervalele de încredere a acestor coeficienți pe baza formulei aferente, ajutându-ne de valoarea erorii standard. În R, am introdus comanda confint(mod\_charges\_smoker) pentru a identifica mai ușor capetele intervalelor.



Pentru β0, CI = [7985.002 8883.535], așadar costul final al asigurării se va încadra în acest interval atunci când nu luăm în calcul dacă asiguratul este fumător.

Pentru β1, CI = [22623.175 24608.752], dacă asiguratul este fumător, valoarea medie a costurilor medicale este situată între 22623.175 u.m. și 24608.752 u.m.

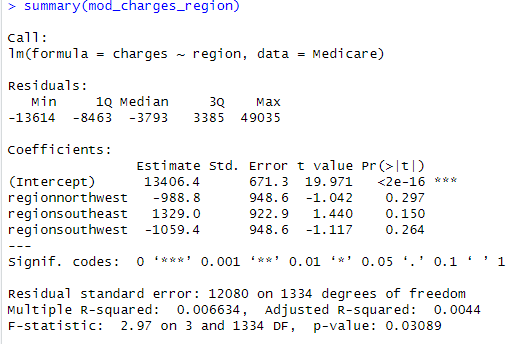
Pe baza acestui model de antrenament, putem construi o grilă nouă cu 100 de valori ale variabilei smoker, pentru a prezice valorile viitoare ale variabilei dependente charges în funcție de starea dacă este fumător sau nu. Variabila smoker ia valori binare „yes” și „no”, iar variabila charges ia valori intre 0 si 60000. Se poate observa și în acest grafic relația liniară dintre cele două, cu o ușoară creștere semnificativă a costurilor medicale pentru asigurații care sunt fumători.

**

**Variabila independentă REGION**

Cel de-al șaselea predictor asupra căruia vom realiza analiza este region. Calculăm coeficienții regresiei, B0 si B1, având în considerare că charges ≈ β0 + β1 × region. Așadar, în Rstudio, introducem următoarea funcție:

mod\_charges\_region <- lm(data = Medicare, charges ~ region)

Rezultatul în urma executării liniei de cod anterioare:

Intercept: valoarea medie a varibilei dependente când variabila independentă este nulă este 13406.4u.m.

region: valoarea medie a variabilei dependente scade cu 988.8 u.m. dacă asiguratul este din regiunea nord-vestică, crește cu 1329.0 u.m. dacă acesta este din regiunea sud-estică și scade cu 1059.4 u.m. dacă este din regiunea sud-estică.

În continuare vom testa ipoteza nulă și cea alternativă:

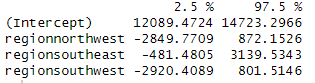
H0 (ipoteza nulă): nu se identifică nicio relație între X si Y. Testăm acest lucru prin a verifica dacă B1 = 0.

H1 (ipoteza alternativă): exista o relație între X si Y. Testăm acest lucru prin a verifica daca B1 este diferit de 0.

În urma acestor rezultate, putem analiza t-statistics pentru coeficientul de regresie care are o valoare foarte slabă comparativ cu celelalte modele, mai exact, în aceeași ordine a regiunilor, -1.042, 1.440 și -1.117. De asemenea p-value are valori crescute, mai exact 0.297, 0.150 și 0.264.

În ceea ce privește modelul, putem măsura calitatea acestuia prin intermediul valorilor lui F-Statistics si p-value, dar și R-Squared și RSE. F-Statistics are o valoare foarte mică, 2.97, cu un p-value de 0.03089. De asemenea, R-Squared are o valoare foarte apropiată de 0 (0.006634 sau 0.66%), indicând o variație mică a variabilei dependente în funcție de regiune. În concluzie, ipoteza nulă s-ar putea să se afirme, având în vedere valorile foarte slabe ale modelului, indicând că nu există o relație deloc semnificativă.

Urmează să identificăm intervalele de încredere a acestor coeficienți pe baza formulei aferente, ajutându-ne de valoarea erorii standard. În R, am introdus comanda confint(mod\_charges\_region) pentru a identifica mai ușor capetele intervalelor.



Pentru β0, CI = [12089.4724 14723.2966], așadar costul final al asigurării se va încadra în acest interval atunci când nu luăm în calcul regiunea de proveniență a asiguratului.

Pentru β1, CI = [-2849.7709 872.1526], dacă asiguratul se află în regiunea nord-vestică

Pentru β1, CI = [-481.4805 3139.5343], dacă asiguratul se află în regiunea sud-estică

Pentru β1, CI = [-2920.4089 801.5146], dacă asiguratul se află în regiunea sud-vestică

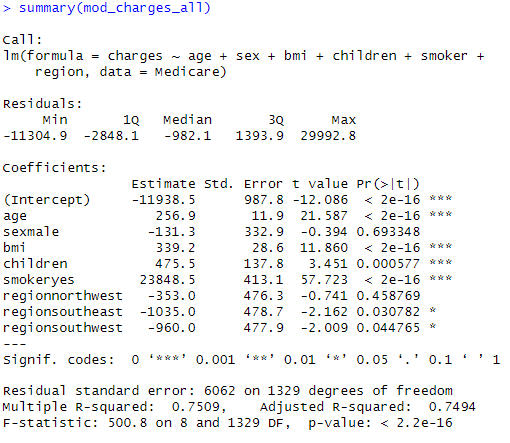
Pe baza modelelor de antrenament anterioare, putem trage concluzia că factorii cu influența cea mai mare sunt statutul de fumător, vârsta și bmi-ul, în această ordine de importanță bazându-ne pe valorile RSE și anume 7470, 11560, 11870. Factori precum numărul de copii ai asiguratului, sexul acestuia sau regiunea de reședință nu au reprezentat o relație atât de puternica cu costurile medicale, având o influență mai mica (children și sex cu un RSE de 12090) sau aproape nesemnificativă (special region cu un RSE de 12080).

Regresia liniară multiplă

Din valorile rezultate până acum, am considerat că acești factori nu au o influență atât de majoră când sunt analizați separați, unii fiind totuși mai relevanți decât alții, așa că am decis să mergem mai departe și să studiem o regresie multiplă aplicată asupra acestora.

În primul rând, am ales să realizăm o regresie multiplă asupra tuturor predictoriilor. Formula matematică pentru modelul analizat este următoarea:

charges ≈ β0 + β1 x age + β2 x sex + β3 x bmi + β4 x children+ β5 x smoker+ β6 x region

Așadar, în Rstudio, introducem următoarea funcție:

mod\_charges\_all <- lm(data = Medicare, charges ~ age + sex + bmi + children + smoker + region)

În urma informațiilor obținute am ajuns la concluzia ca există o legătura mare între variabilele independente age (vârstă), bmi (indicele de masă corporală) și smoker (statutul de fumător/nefumător) și variabila dependentă charges (costul asigurării). Celelalte variabile independente nu depind sau depind foarte puțin de costul asigurării.

Conform rezultatelor, putem sesiza testarea ipotezelor, și anume:

H0 – Ipoteza nulă: Nu s-a determinat nicio relație între variabila dependentă și predictori;

Ha - Ipoteza alternativă: Există cel puțin un termen Bi care va îndeplini condiția Bi ≠ 0;

Dat fiind faptul ca F-statistic este o valoare mult mai mare decât 1, mai exact 500.8, duce la o cât mai bună ajustare a modelului pentru datele analizate. În plus, p-value reprezintă o valoare foarte mică (<2.2e-16), ceea ce sugerează că F-statistic este foarte puțin probabil să demonstreze să fie adevărată ipoteza nulă. În concluzie, rezultatele susțin ipoteza alternativă că există o relație semnificativă între variabilele independente și cea dependentă.

Pe baza rezultatelor obținute mai devreme, am creat un nou model cu predictorii cei mai importanți pentru regresia noastră multiplă, rezultatele se pot observa în *Figura 7*. Aceștia sunt age (vârsta), bmi (indicele de masă corporală), dar și smoker (dacă persoana este fumătoare/nefumătoare). Drept urmare, avem un p-value mic, rezultând că variabilele sunt semnificative static. Cu trecerea unui an de vârstă (age), costul asigurării (charges) va crește cu 259.55 u.m. Mai mult, pentru fiecare unitate de creștere a masei corporale (bmi), costul asigurării (charges) va crește cu 322.62 u.m. În plus, dacă o personă este fumătoare (smokeryes), costul asigurării (charges) va fi cu 23823.68 u.m mai mare decât a unei persoane care nu fumează. Având în vedere ca valoarea lui R-squared este 0.7475, putem spunem că aproximativ 74.75% din variația costului medical poate fi interpretat prin intermediul predictorilor age, bmi și smoker.

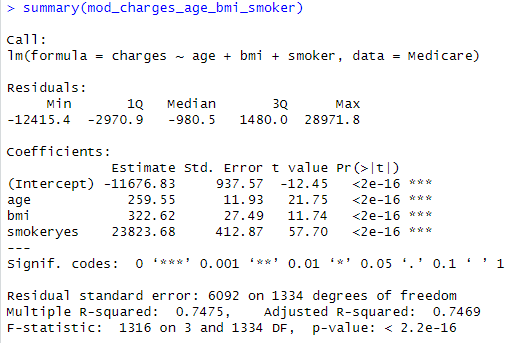
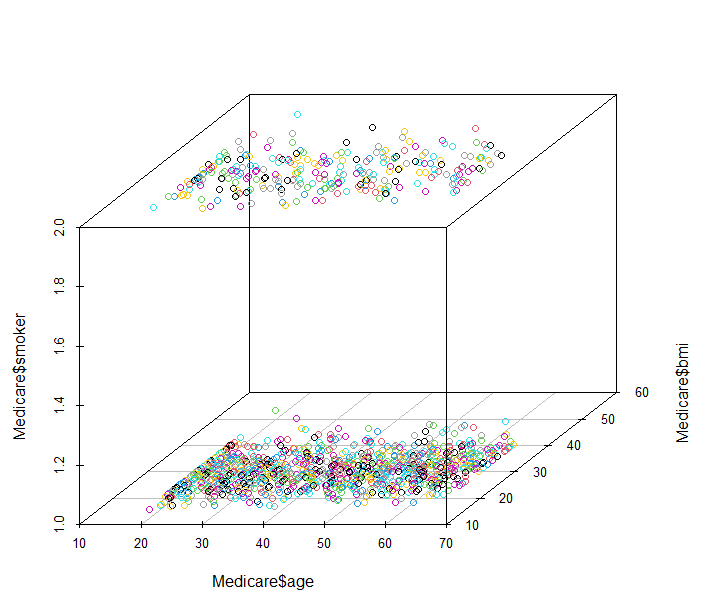
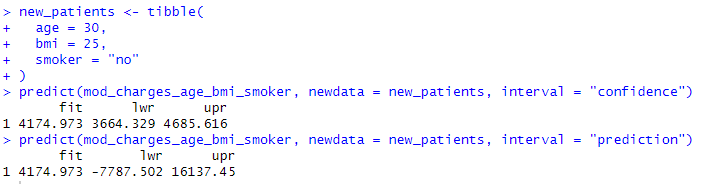


Figura 8

Figura 7

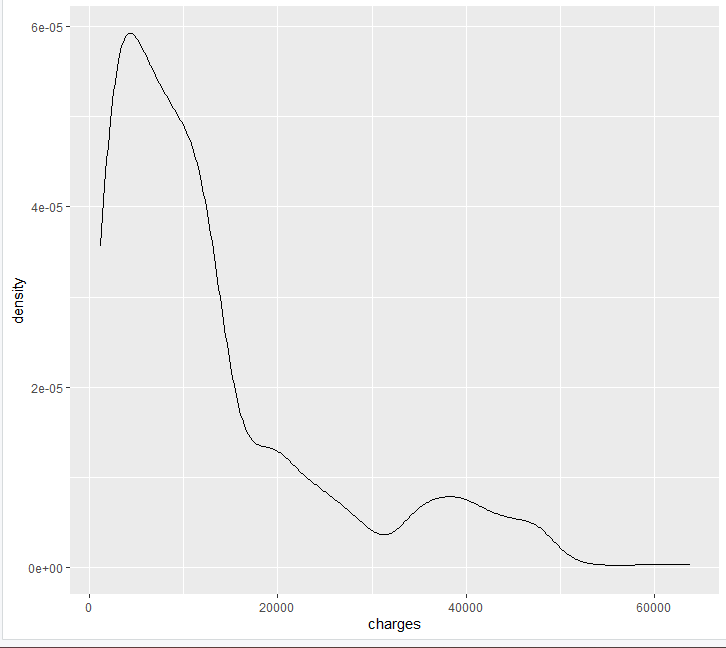
Având în vedere aceste date, am considerat crearea unei diagrame 3D (*Figura 8*) pentru factorii cu influența cea mai mare.

Continuăm prin a face o predicție asupra unui asigurat nou, luând în considerare factorii cu relevanța cea mai mare. Persoana pe baza căreia facem predicția are vârsta de 30 de ani, un indice de masă corporală de 30 si nu este fumătoare. Asigurarea acestei persoane va avea valoarea de 4174.973 u.m., pentru un interval de încredere de [3664.329 4685.616] și un interval de predicție de [-7787.502 16137.45].

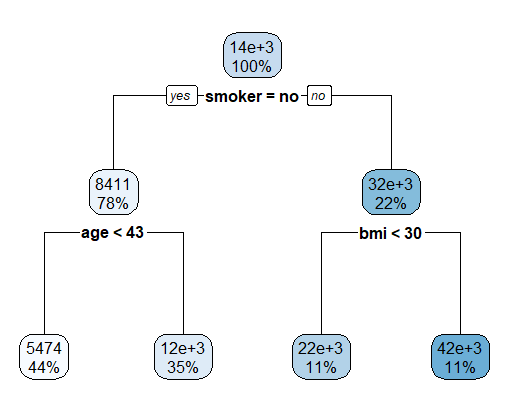
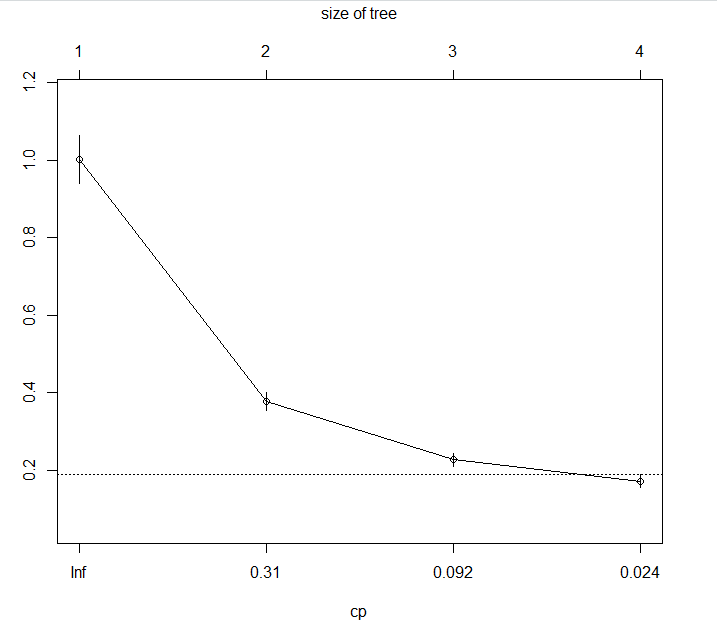


Arbori de decizie și Bagging

Am decis ca analiza noastră să cuprindă mai multe modele de învățare pentru a avea diversitate și perspective multiple. Am ales în continuare să folosim arbori de decizie, deoarece reprezintă o metoda ușor de interpretat, atât de către persoanele care realizează analiza, cât și de persoane mai puțin specializate în domeniu.

**Pentru început, am realizat un grafic pentru a determina densitatea persoanelor în funcție de suma totală a costului asigurării. Din acest grafic, rezulta ca cele mai multe persoane plătesc o suma cuprinsă între aproximativ 1000 si 10000 de unități monetare.

Mai apoi, am împărțit datele noastre cu ajutorul unui split în doua categorii: date de antrenament și date de testare. În această împărțire, s-a respectat proporția de 70% pentru datele de antrenament și 30% pentru datele de testare.

Continuăm prin a crea primul arbore de antrenament din analiza noastră. Pentru a realiza acest lucru ne-am folosit de funcția rpart() din Rstudio, care ne ajută să construim cel mai mic arbore cu cea mai mica eroare (SSE), pe baza parametrului costului de complexitate (cp). Pentru a face acest lucru, algoritmul realizează mai mulți arbori, iar prin cross-validation alege cel mai bun. De asemenea, sunt aleși si parametrii optimi după care se face partiționarea setului inițial în două subseturi, precum și valoarea acestora, pentru a minimiza SSE.

Pentru a vedea care este costul de complexitate ales implicit de către algoritm, vom rula comanda plotcp(m1). Așadar observăm că pentru a minimiza SSE s-a ales un cost de complexitate de 0.02, pentru un arbore cu dimensiunea de 4.

Primul și cel mai important criteriu după care se face partiționarea a fost statutul de fumător al asiguratului. Dacă acesta este fumător, predictorul după care se face următoarea partiționare este indicele de masă corporal. Acest subset este divizat în alte două subseturi, după valoarea de test optimă găsită de algoritm și anume daca indicele se află sub 30 sau nu. Dacă asiguratul nu este fumător, se va lua în considerare vârsta asiguratului, mai exact dacă se află sau nu sub 43 de ani.

Mai departe, vom crea un arbore cu costul de complexitate 0. Astfel, dacă nu se alege un cost de complexitate optim, rezultatul este un arbore de dimensiuni foarte mari, deoarece în lipsa unui cost de complexitate nu se minimizează SSE sau dimensiunea arborelui. (*Figura 9*)

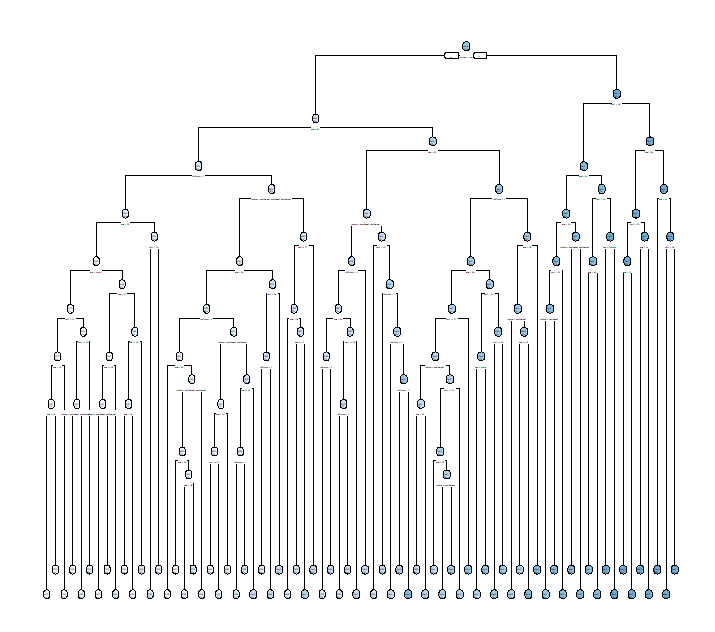
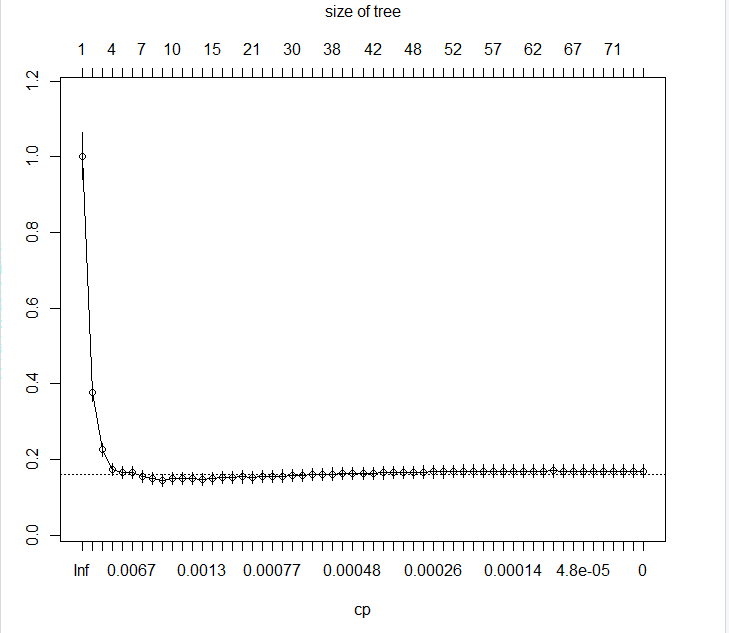


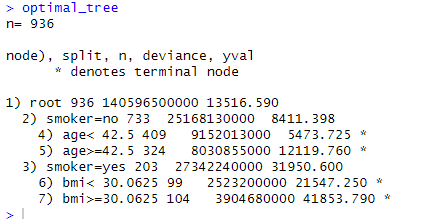
Figura 10

Figura 9

Putem observa si dimensiunea exacta a acestui arbore, fiind de 73 de noduri distincte. Si in acest arbore, criteriul principal după care se face partiționarea este statutul de fumător, urmat de vârstă si indice de masă corporal. (*Figura 10*)

Pentru a gasi arborele optim, o să determinăm valorile necesare pentru minsplit si maxdepth. Minsplit determină numărul minim de instanțe al unui subgrup pentru a continua împățirea acestuia, iar maxdepth determină numarul maxim de noduri interne ale arborelui. În primul rând, creăm un grid care conține toate combinațiile de minsplit si maxdepth dintr-un interval ales pentru fiecare. Într-o lista, stocăm toate modelele bazate pe aceste combinații. S-au creat 128 astfel de modele, dintre care o să alegem primele 5 modele cu cea mai mică eroare. Pentru a construi arborele optim, minsplit si maxdepth vor primi valorile corespondente ale modelului cu cea mai mică eroare.

Așadar, construim un arbore cu un minsplit de 20 și un maxdepth de 8, cu o eroare de 0.1688184. Putem observa diferența dintre eroarea primului arbore din antrenament 0.1699170 si eroarea arborelui optim, fiind o diferență aproape nesemnificativă. Acuratețea modelului nostru este reprezentată de RMSE, având valoarea egală cu 5069.309, deci nu putem afirma ca predicțiile vor fi foarte precise, dat fiind valoarea crescută. Se observă similaritatea crescută între acești doi arbori, iar structura arborelui optim este următoarea.

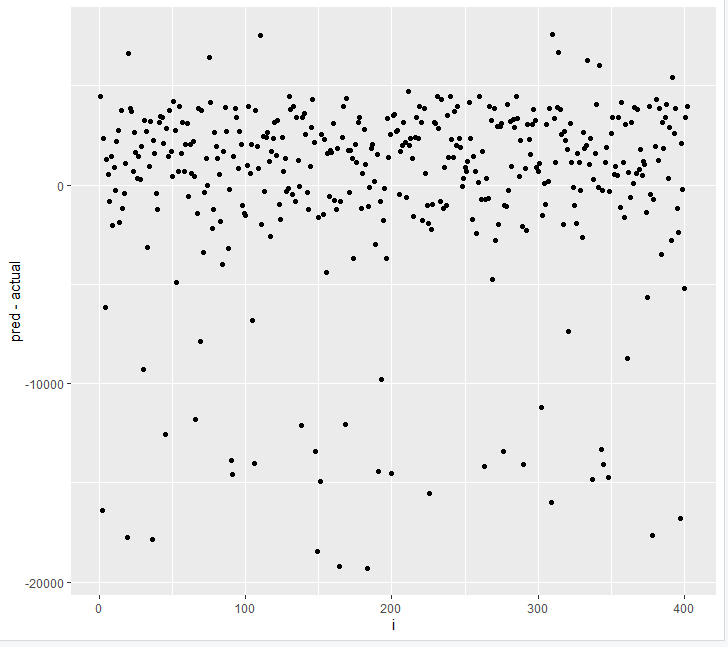


Pentru a face o analiză și mai complexă asupra costurilor medicale, am decis să folosim o a treia metodă de antrenament, și anume bagging. Pentru inceput, vom folosi metoda “bagging” în R pentru a împărți setul de date de antrenament în mai multe eșantioane sau replicări bootstrap, pentru a antrena câte un arbore individual. Aceste seturi de antrenament se folosesc pentru a antrena câte un arbore individual. Am creat un model simplu de bagging, cu un număr implicit de 25 de replicări bootstrap. RMSE rezultat are valoarea de 4828.225, de unde rezultă o discrepanță între valorile prezise de ansamblul de arbori și valorile reale, așadar costurile nu sunt prezise cu precizie.

Pentru a încerca să minimizăm RMSE, vom afla care este numărul optim de replicări bootstrap pentru o eroare cât mai mica. În urma graficului, eroarea minimă pare să fie undeva la 40-43 de replicări.

Având în vedere aceste informații, următorul pas este să realizăm modelele de antrenament propriu-zise pentru fiecare subset, folosind metoda bagging, prin arbori de decizie. Pentru început, specificăm anumite setări asupra modelului de antrenament bagging: va avea 10 pliuri (adică va împărți setul de antrenament în 10 subseturi și va antrena câte un model pentru fiecare subset), iar metoda prin care se face antrenamentul este validarea încrucișată.

Rezultatul antrenamentului are un r-squared de 0.8469122, de unde deducem o acuratețe de aproximativ 85% a modelului de antrenament. RMSE are valoarea de 4732.587, de unde deducem o precizie mai bună față de modelul de antrenament creat anterior.

Am creat un grafic pentru a observa importanța predictorilor, age, bmi și smoker fiind încă o dată cei mai relevanți predictori. Având în vedere datele despre predicții, am calculat RMSE pe setul de test, iar rezultatul a fost 4977.4, indicând încă o dată o precizie nu atât de puternică, așadar nu putem afirma că acești factori reprezintă o influență atât de mare asupra costurilor medicale și nu putem prezice cu exactitate costurile medicale pe baza acestora. Acest aspect se poate observa și în următoarele grafice:



**Concluzii**

În urma acestei analize, considerăm ca am răspuns la toate întrebările adresate la începutul procesului de cercetare. Am aflat că există într-adevăr o corelație între acești factori și suma costurilor și nu putem nega influența acestora asupra cât va avea de plată asiguratul. Totuși, legătura nu este una semnificativă, în sensul că un singur factor luat singur nu are o influență majoră asupra costurilor, dar împreună pot crea diferențe. Totodată, considerăm că există o multitudine de alți factori de influență în afara celor studiați. În ciuda faptului că aceștia nu au reprezentat diferențe majore asupra costurilor medicale, câțiva dintre ei, cum ar fi vârsta, statutul de fumător sau indicele de masă corporal, s-au remarcat mai mult față de ceilalți. Concluzia rămâne totuși aceeași: nu se poate realiza o predicție de acuratețe maximă pe baza acestor factori, dar oferă o idee generală și un punct de plecare pentru o analiză mai complexă. De asemenea, metoda cu acuratețea cea mai mare, pe baza căreia am putut să extragem informații mai clare, a fost regresia, comparativ cu cele două modele de arbori.