**ACADEMIA DE STUDII ECONOMICE BUCUREŞTI**

**FACULTATEA DE CIBERNETICĂ, STATISTICĂ ŞI INFORMATICĂ ECONOMICĂ**

***Factorii contribuabili în formarea prețului automobilului***

**Profesor coordonator**

Conf. Univ. Dr. Adriana Ana Maria DAVIDESCU

**Studenţi**

                                                                         Crețu Costin-Răzvan

                                                                            Dan Alex-Silviu

Diță Alexandru

**Bucureşti**

**2022**

**Cuprins**

[Introducere 3](#_Toc122976719)

[APLICAȚIA 1: Modele de regresie 4](#_Toc122976720)

[Capitolul teoretic 4](#_Toc122976721)

[*1.1. Literature review* 4](#_Toc122976722)

[*1.2. Metodologia cercetării* 6](#_Toc122976723)

[Capitolul aplicativ 8](#_Toc122976724)

[*2.1. Date utilizate* 8](#_Toc122976725)

[*2.2.Rezultate empirice ale cercetării* 12](#_Toc122976726)

[Concluzii 27](#_Toc122976727)

[Bibliografie 29](#_Toc122976728)

# **Introducere**

Oamenii sunt diferiți, fiecare simte emoția în felul lui, pentru fiecare o pasiune se simte diferit, poate apărea la o vârstă fragedă și să dureze toată viața sau se poate ivi și flacăra ei să se stingă într-un timp foarte scurt. Lucrurile nu sunt diferite nici în privința mașinilor, astfel atenția fiecăruia este distribuită diferit atunci când privesc o mașină, unele persoane sunt fascinate de putere, altele de aspect, ‘vechime’, numărul de kilometri, marca sau chiar prețul acestora. Pentru unii oameni mașinile sunt simple obiecte care îi pot ajuta să se deplaseze dintr-un loc în altul în timp ce pentru alții mașinile au devenit un simbol al pasiunii lor pentru succes, putere și libertate.

În cadrul acestui proiect, punem accentul pe factorii decizionali care influențează prețul unui autoturism, factori bazați în principal pe prețul de vânzare, persoana/instituția care vinde autoturismul, tipul de motor, cilindreea motorului și puterea autoturismului. Această analiză va fi realizată pe un set de date care conține 205 de autoturisme de diferite mărci, preț, consumul în oraș, caii putere și alte variabile.

În prima aplicație a proiectului, va fi reprezentat un model de regresie simplă care se va realiza pe baza a două variabile, prețul de vânzare fiind o variabilă dependentă și puterea autoturismului, o variabilă independentă. De asemenea, se estimează parametrii implicați în modelele econometrice, se examinează proprietățile estimatorilor rezultați și se generalizează rezultatele analitice la modelele mai complexe.

În plus, ilustrăm și un model de regresie multiplă în care vom folosi variabilele din cadrul primei aplicații și la care vom adăuga dimensiunea motorului și consumul mediu urban ca variabile independente. De asemenea, pe lângă toate aceste variabile se va adăuga și o variabilă dummy, o variabilă ce este populată cu valoarea 0 sau 1, și care indică prezența altor factori de influență în modelul analizat, putând să modifice semnificativ output-ul analizei. ‘Fuel type’ este variabila dummy adăugată și reprezintă tipul de combustibil pe care îl folosește autoturismul, 1 pentru benzină iar 0 pentru diesel.

# **APLICAȚIA 1: Modele de regresie**

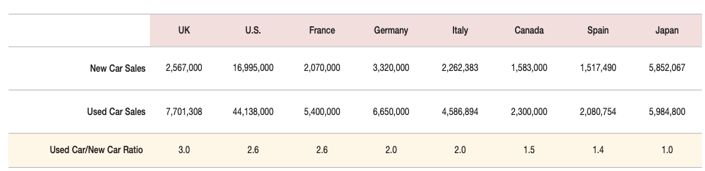
## **Capitolul teoretic**

### ***1.1. Literature review***

Piața de mașini se află într-o creștere continuă, și cu toate acestea producătorii de mașini se confruntă cu mai multe provocări pe piața de autoturisme noi și second-hand. Un exemplu în acest sens este tendința către mașinile mai ecologice, un factor care scade marjele de vânzări. Potrivit datelor Autovista (FleetEurope.com), cererea de mașini a scăzut și pe piețele europene cu preponderență în Germania, înregistrând un procent de peste 14% și cu peste 12% în Franța în primele șase luni ale anului 2022, având repercusiuni și în urma contextului pandemic. Cu toate acestea, prețurile mașinilor nu au început să scadă, iar în următoarea perioadă datele au arătat creșteri lunare ale prețurilor de la 0,1% în Spania la 1,3% în Austria. România este una dintre cele mai ieftine piețe, dar prețurile mașinilor noi în Europa sunt cu aproximativ 15% mai mari decât media globală.[[1]](#footnote-1)

Un alt articol științific relevant pentru tema aleasă a fost publicat pe Econ Stor și arată cum autovehiculele, bunuri de consum moderne, înglobează o multitudine de caracteristici care relevă valoarea lor pe piața. Evaluăm aceste bunuri pe baza unor diverse atribute, astfel în contextul dezvoltării accelerate a tehnologiei, internetul a oferit o modalitate mai ușoară pentru achiziționarea autoturismelor, clienții având acces la o multitudine de informații legate de istoricul unui autoturism evitând în acest fel majoritatea riscurilor de achiziție a unei autoturism indiferent dacă este nou sau second-hand. [[2]](#footnote-2)

Conform articolului “Anatomy and Physiology of the Used Car Business” puterea pieței auto variază în funcție de țară, de zona geografică și de percepția cetățenilor. Așa cum se poate observa și în tabelul de mai jos, în toate țările prezentate, rata vânzărilor de mașini second-hand este semnificativ mai mare decât cea a vânzărilor de mașini noi. O excepție de la această regulă o reprezintă Japonia, țara în care deținerea de autovehicule second-hand generează costuri suplimentare pentru deținător și astfel este încurajată achiziționarea de autoturisme noi. [[3]](#footnote-3)

Figura 1. Vânzări pe piața de vehicule uzate vs vehicule noi [3] (sursa: Google Academics)

Alți factori care influențează rata vânzărilor de autoturisme sunt puterea motorului, dimensiunea motorului și prețul autoturismului. Automat, o persoană care dorește să își achiziționeze un autoturism va lua în calcul atât prețul autoturismului cât și puterea acestuia. Cu cât puterea motorului este mai mică și dimensiunea direct proporțională cu puterea, cu atât prețul de achiziție este mai mic. [[4]](#footnote-4)

De asemenea, piața de automobile începe să devină o sursă profitabilă de investiții, fiind luați în calcul factori precum fluxul comercial, cererea de energie și poluarea aerului. De asemenea, prin încercarea de a aduce inovații noi pe piață, eșecul nu a ezitat să-și facă simțită prezența, astfel guvernul a intervenit pe piața auto prin adăugarea diferitor taxe și impozite pentru a corecta divergențele financiare apărute în urma reglementărilor menite să sporească siguranța autovehiculelor, reducând în același timp poluarea aerului și consumul de energie. [[5]](#footnote-5) În plus, în contextul pandemic și ulterior acestuia, au apărut și persistat probleme legate de aprovizionarea cu diverse materii prime pentru construcția de autoturisme. Astfel cumpărătorii au întâmpinat dificultăți în achiziționarea de autoturisme noi și s-au îndreptat către autoturismele second-hand care uneori au fost vândute la un preț mai mare decât cele noi.

### ***1.2. Metodologia cercetării***

Metodele folosite în ceea ce privește metodologia cercetării sunt regresia simplă și cea multiplă în cadrul aplicației 1, urmărind factorii ce alcătuiesc prețul de achiziție al automobilului.

Forma modelului de regresie simplă este:

.

Unde, variabila dependentă este prețul de vânzare, iar cea independentă este puterea motorului.

Formula regresiei multiple este:

Unde, variabila dependentă este prețul de vânzare, iar cele independente sunt puterea motorului, dimensiunea motorului și consumul în mediul urban.

Intensitatea legăturii dintre variabile va fi măsurată prin indicele de corelație.

Validitatea modelului va fi testată folosind testul Fisher ale cărui ipoteze sunt:

* din  rezultă că modelul nu este valid din punct de vedere statistic
* din rezultă că modelul este valid din punct de vedere statistic

Validitatea parametrilorva fi testată folosind testul T cu ipotezele următoare:

* : = 0; = 0 > parametrii nu sunt semnificativi din punct de vedere statistic
* : ≠ 0; ≠ 0 > parametrii sunt semnificativi din punct de vedere statistic

Ipoteza de homoscedasticitate ce constă în următoarele:

* : erorile aleatoare sunt homoscedastice
* : erorile aleatoare sunt heteroscedastice

Testul Breusch-Pagan având ipotezele următoare:

* : reziduurile sunt distribuite cu varianță egală
* : reziduurile nu sunt distribuite cu o variație egală

Testul White cu ipotezele respective:

* : variațiile pentru erori sunt egale
* : varianțele nu sunt egale

Testul Durbin-Watson de ordin 1 cu ipotezele aferente:

* : corelația de ordinul 1 nu există.
* : corelația de ordinul întâi există.

Testul Breusch-Godfrey de ordin superior cu ipotezele următoare:

* : variațiile erorii sunt egale.
* : variațiile erorii nu sunt egale.

Testul Shapiro-Wilk ce constă în următoarele ipoteze:

* : dacă p value > 0.5, distribuția eșantionului nu este semnificativ diferită de o distribuție normală
* : dacă p value < 0.5, testul este semnificativ, iar distribuția este semnificativ diferită de o distribuție normală

Testul Jarque-Bera având următoarele ipoteze:

* : erorile aleatoare urmează o distribuție normală
* : erorile aleatoare nu urmează o distribuție normală

Testul RMSE - eroarea pătrată medie a cărui formulă este următoarea:

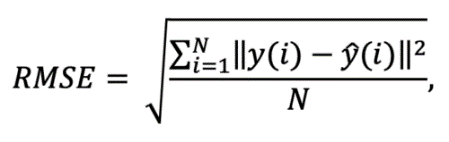


Figura 2. Formula eroare pătrată medie (sursa: Wikipedia)

Testul MAE – eroarea absolută medie

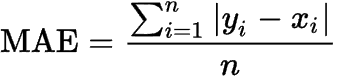


Figura 3. Formula eroare absolută medie (sursa: Wikipedia)

Testul MSE – eroarea medie pătratică

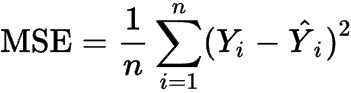


Figura 4. Formula eroare medie pătratică (sursa: Wikipedia)

Testul MAPE - eroarea procentuală medie absolută

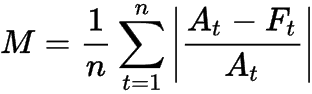


Figura 5. Formula eroare procentuală medie absolută (sursa: Wikipedia)

## **Capitolul aplicativ**

### ***2.1. Date utilizate***

Din punct de vedere al sursei de colectare a datelor, am extras de pe plaftorma Kaggle următorul set de date cu variabilele aferente. Acest set de date pune în ecuație problema unei companii de automobile chineză Geely Auto ce dorește să se extindă pe piața din America prin crearea unei fabrici în zona respectivă. Înainte de a se extinde, compania dorește să înțeleagă ce factori sunt decizionali în privința prețului automobilelor pe piața americană, deoarece aceștia pot diferi față de cei de pe piața chineză.

Din punct de vedere al transformării aduse datelor, tipul de combustibil (fueltype) a fost convertit într-o variabilă de tip TRUE și FALSE, respectiv 1 și 0, astfel: gas = 1 ; diesel = 0. Alimentarea cu benzină a fost considerată ca fiind TRUE, deoarece sunt mai numeroase decât cele pe motorină. Totodată, variabila enginelocation a fost convertită într-o variabilă de tip dummy: front = 1, rear = 0, deoarece automobilele cu motorul situat în partea frontală sunt cele mai răspândite.

Pentru a putea analiza cu acuratețe aceste date, am început cu prețul, unde se poate concluziona faptul că, cu cât crește mai mult, cu atât oferta pentru automobile devine mai slabă.

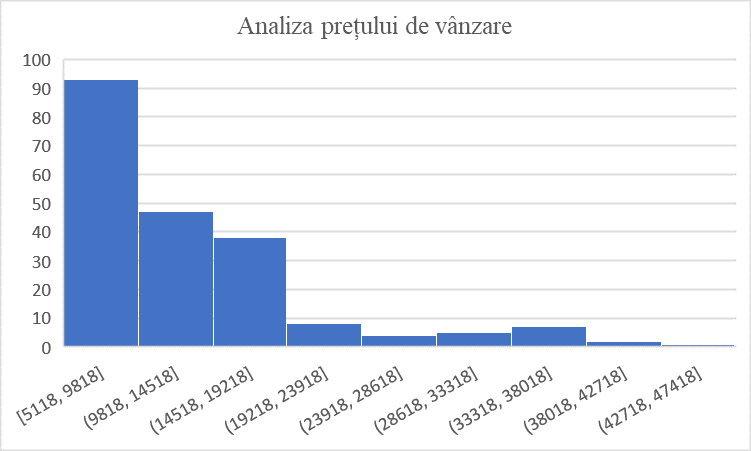


Figura 6. Grafic privind analiza prețului de vânzare (sursa: Excel)

Continuând cu puterea dezvoltată de către motorizare, putem observa faptul că automobilele cu spectrul de putere situat între valorile de 94 și 117 cai putere sunt cele mai numeroase:

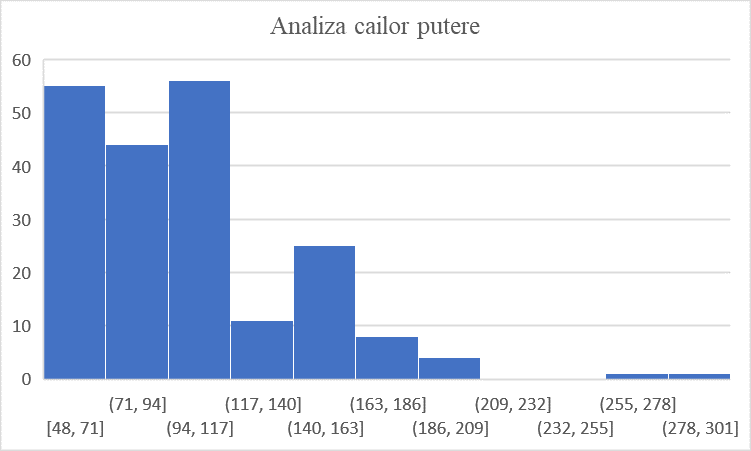


Figura 7. Grafic privind analiza cailor putere (sursa: Excel)

Cilindreea motorului se dovedește a fi unul din factorii importanți din motive economice, atât din punct de vedere al consumului, cât și a fabricației, cu cât aceasta este mai mare, cu atât motorul devine mai gurmand din punct de vedere al consumului, cât și a complexității:

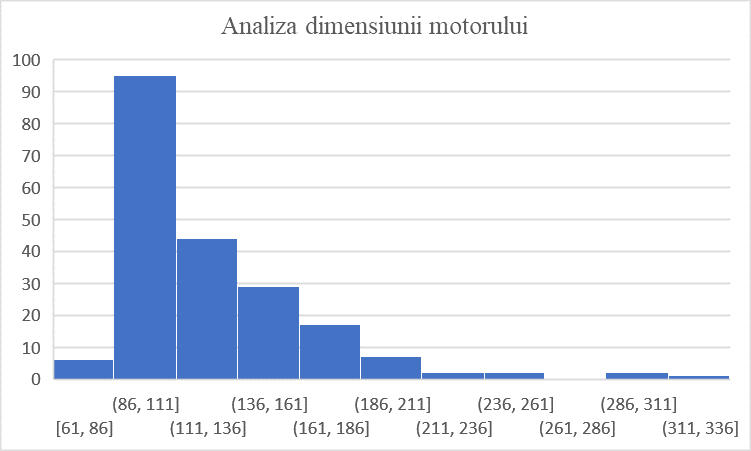


Figura 8. Grafic privind analiza dimensiunii motorului (sursa: Excel)

Din punct de vedere al carburantului cu care se alimentează motorul, autovehiculele alimentate cu benzină domină piața pe care activează în acest set de date:

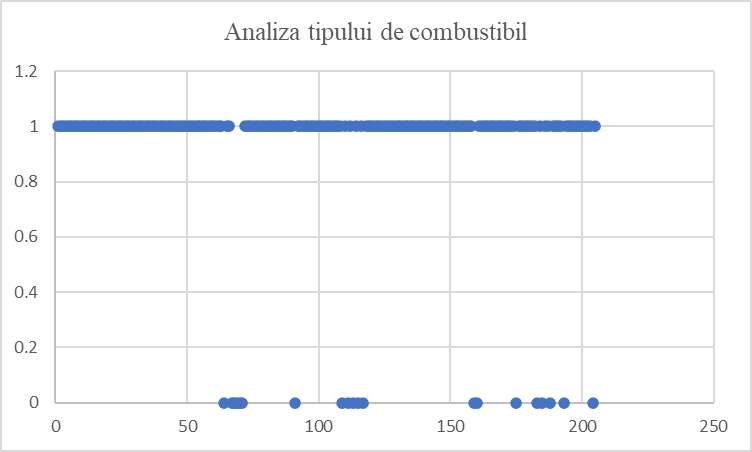


Figura 9. Grafic privind analiza tipului de combustibil (sursa: Excel)

În cele din urmă, prezența mașinilor cu motorul situat în față sunt în proporție de peste 95% pe setul acesta de date, cele cu motorul în spate fiind destul de rare:

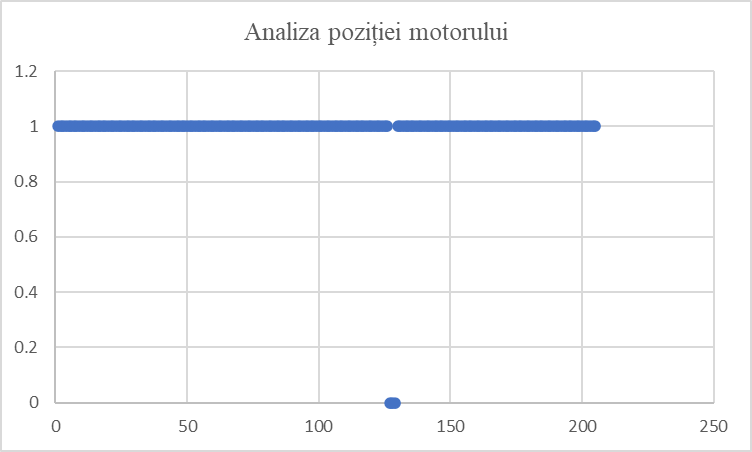


Figura 10. Grafic privind analiza poziției motorului (sursa: Excel)

Pentru realizarea modelelor de tip log-log, log-lin, linear-log am creat variabilele lprice, lhorsepower, lenginesize folosindu-ne de formula LN() în Excel:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| price | lprice | horsepower | lhorsepower | enginesize | lenginesize | citympg | lcitympg |
| 13495 | 9.510075 | 111 | 4.70953 | 130 | 4.867534 | 21 | 3.044522 |
| 16500 | 9.711116 | 111 | 4.70953 | 130 | 4.867534 | 21 | 3.044522 |
| 16500 | 9.711116 | 154 | 5.036953 | 152 | 5.023881 | 21 | 3.044522 |
| 13950 | 9.543235 | 102 | 4.624973 | 109 | 4.691348 | 21 | 3.044522 |
| 17450 | 9.767095 | 115 | 4.744932 | 136 | 4.912655 | 21 | 3.044522 |
| 15250 | 9.632335 | 110 | 4.70048 | 136 | 4.912655 | 21 | 3.044522 |
| 17710 | 9.781885 | 110 | 4.70048 | 136 | 4.912655 | 21 | 3.044522 |

Tabelul 1. Realizarea variabilelor necesare modelelor log-log, log-lin, linear-log (sursa: Excel)

Valoarea mașinii este determinată de preț, iar prețul este de astfel, dependent de celelalte caracteristici ale automobilului, astfel,  variabilele pe care le-am luat în considerare în vederea realizării modelului de regresie simplă sunt prețul de vânzare și nivelul de putere, fiind variabila independentă, deoarece, cu cât o mașină este mai puternică, cu atât devine mai valoroasă.

* Prețul de vânzare (price)

Suma de bani pe care clientul o plătește în vederea achiziționării autovehiculului. Aceasta este stabilită de către dealer, fiind o idee mai mare decât prețul de fabricare al mașinii, setat de către Prețul de vânzare recomandat de producător (PDSM).

* Caii putere (horsepower)

Este unitatea de măsurare a puterii automobilului ce se referă la randamentul motorului, iar, spre deosebire de cuplul motor ce își face simțită prezența doar în momentele de vârf de sarcină, caii putere vizează puterea furnizată în mod continuu.

Din punct de vedere al îmbunătățirii modelului de regresie simplă, am luat în considerare dimensiunea motorului, o altă variabilă independentă, fiind un factor destul de important în fabricarea autovehiculului din punct de vedere al economiei, rezultând un model de regresie multiplă.

* Dimensiunea motorului (enginesize)

Suma volumului tuturor cilindrilor motorului. Spre exemplu, dacă un motor are patru cilindri cu o cilindree de 500 cm³, atunci volumul motorului va fi de aproximativ 2.0 litri

* Consum în mediu urban (citympg)

Cantitatea necesară pentru ca mașina să poată parcurge o anumită distanță în mediul urban măsurată în mile pe galon în cazul de față. MPG este una dintre măsurile principale ale eficienței automobilului din punct de vedere economic.

În cele din urmă, am îmbunătățit modelul de regresie multiplă prin adăugarea unei variabile independente dummy, tipul de combustibil fiind împărțit în două categorii: 1 -> gas (benzină) ; 0 -> diesel (motorină):

* Tipul de combustibil (fueltype):

Sistemul de alimentare al motorului se poate împărți în mai multe categorii, însă, pe acest set de date, s-a rezumat la combustibilii lichizi, cum ar fi motorina și benzina.

### ***2.2.Rezultate empirice ale cercetării***

Pentru formarea modelului de regresie simplă, o primă semnificație asupra legăturii dintre cele două variabile este reprezentată de indicele de corelație, ce, în cazul de față, arată o legătură puternic pozitivă dintre preț și caii putere. Coeficientul de corelație este 0.808, aparținând de intervalul [-1,1].

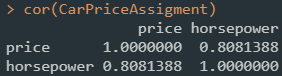


Figura 11. Regresie simplă, indicele de corelație (sursa: RStudio)

Din figura 7 putem deduce faptul că regresia simplă are forma **price = -3721.761 + 163.263 \* horsepower** conform . Această formulă poate fi interpretată astfel: la creșterea cu 1 cal putere a motorului, prețul de achiziție crește cu 163.263$, probabilitatea fiind de 99%. Având în vedere faptul că puterea mașinii măsoară valori între [48, 301], cu cât puterea mașinii este mai mare, cu atât valoarea mașinii va avansa în preț.

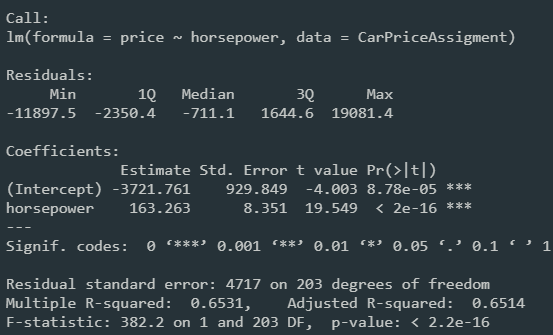


Figura 12. Regresie simplă, output (sursa: RStudio)

Valoarea F-statistic este de 382.2, de unde rezultă că se respinge ipoteza și se acceptă , modelul fiind valid pentru un nivel de semnificație de 1%, rezultatul fiind garantat cu o probabilitate de 99% și 205 de observații, rezultând k = 1 și n - k - 1 = 205 - 1 - 1 = 203 grade de libertate.

Urmând cu validitatea parametrilor astfel, respingem , acceptând , deoarece ≠ 0, rezultatele sunt garantate pe o probabilitate de 99%, deoarece prezintă valori mai mici de pragul de 1%.

Interpretarea Indicatorilor de bonitate se rezumă la R-squared (coeficientul de determinație) este egal cu 0.6514  și indică faptul că, caii putere ai motorului autovehiculului explică aproximativ 65.14% din prețul de vânzare. Adjusted R-squared este un coeficient de determinație corectat cu grade de libertate și are aceeași seminficație ca și R-squared. Eroarea standard a regresiei arată cu cât se abat în medie, valorile observate de la valorile teoretice aflate pe dreapta de regresie. Pentru modelul nostru, eroarea standard este 4717.

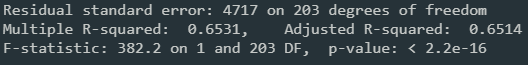


Figura 13. Regresie simplă, indicatori de bonitate (sursa: RStudio)

Modul în care sunt situate observațiile în comparație cu dreapta estimată se reflectă în figura de mai jos, astfel încât majoritatea mașinilor măsoară puterea până în pragul de 100 de cai, existând și excepții, precum Porsche Cayenne, dezvoltând 288 CP și Porsche Panamera cu 207 CP, iar la polul opus se află Chevrolet Impala cu doar 48 CP.

**Chart, scatter chart

Description automatically generated**

Figura 14. Regresie simplă, observații cu dreapta estimată (sursa: RStudio)

Continuând cu predicția după regresie, am început prin a analiza variabila pricehat fiind egală cu extragerea valorilor ajustate ale modelului obținând următorul grafic: valorile actuale și cele prezise ale variabilei dependente.

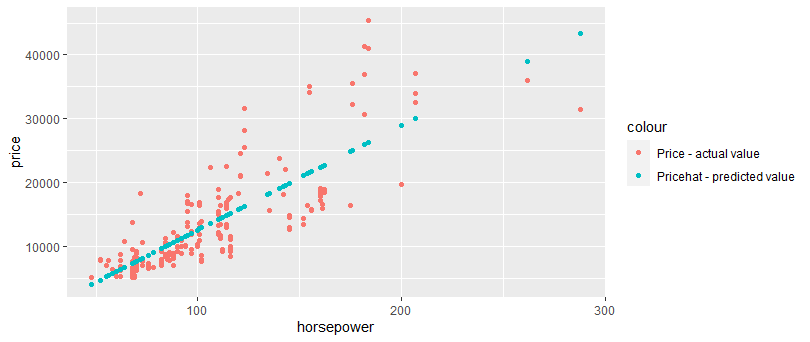


Figura 15. Regresie simplă, valorile estimate pentru variabila dependentă (sursa: RStudio)

Am analizat în continuare prezența reziduurilor comparate cu valorile prețului de achiziție, urmând să ajustăm modelul. Din figura de mai jos reiese faptul că oamenii preferă mașini cu un număr mic spre mediu de cai (până în 100 cai putere) datorită siguranței și pentru impozitarea cât mai mică. Pe de altă parte, există și clienți pasionați de cultura auto, care doresc mașini puternice, valoarea impozitării nefiind un factor deciziv în achiziția automobilului.

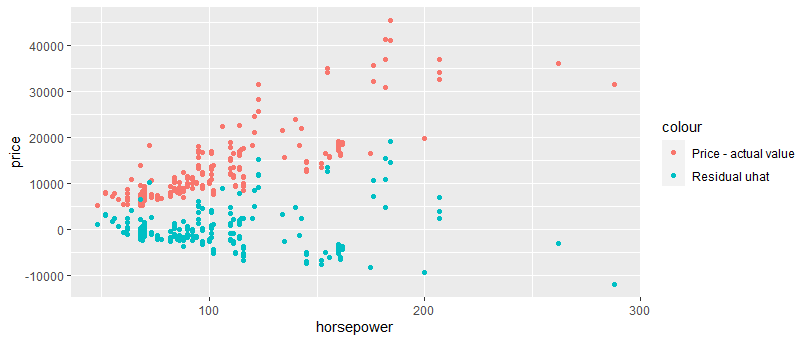


Figura 16. Regresie simplă, ajustarea reziduurilor (sursa: RStudio)

Testând ipotezele pe reziduuri, am început prin heteroschedasticitate, fiind nevoiți în a schimba forma funcțională a variabilelor în cea de tip log-log din cauza faptului că avem  4717 erori standard reziduale, modelul explicând 65.14%. După logaritmarea variabilelor am obținut un surplus în ceea ce privește explicarea modelului de R pătrat ajustat de aproximativ 5%, iar erorile standard reziduale au scăzut până la 0.2719.

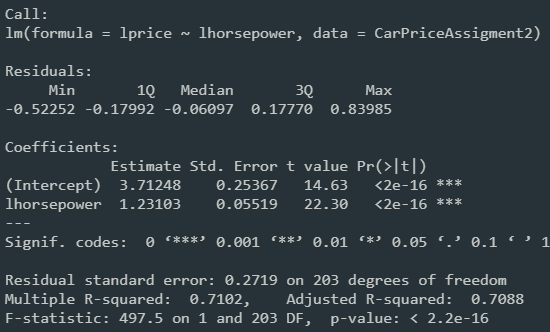


Figura 17. Regresie simplă, output forma funcțională de tip log-log (sursa: RStudio)

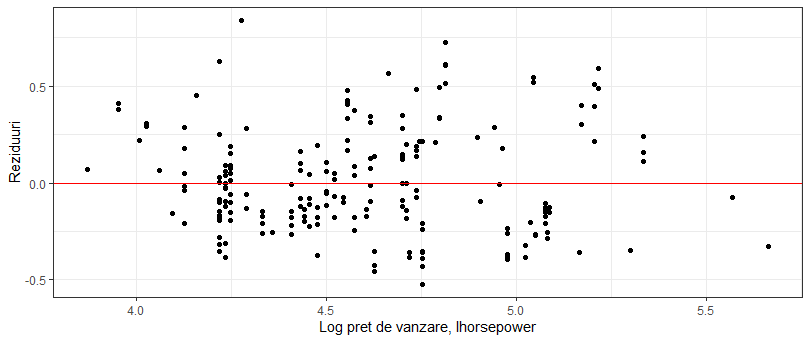


Figura 18. Regresie simplă, reziduuri față de variabila lhorsepower (sursa: RStudio)

Aplicând testul Breusch-Pagan în scopul de a transforma reziduurile din heteroschedastice în homoschedastice, astfel, valoarea lui p-value este de 1.497e-13 rezultând că reziduurile sunt homoschedastice. Rulând testul White, p-value a avut valorea 0, sub 0.1, rezultând faptul că modelul va trebui corectat.

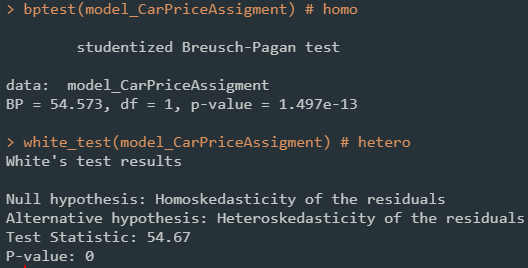


Figura 19. Regresie simplă, testele Breusch-Pagan și White înainte de corectare (sursa:Rstudio)

După ce am corectat modelul folosind forma log-linear, am obținut următoarele rezultate ce specifică faptul că modelul reziduurile încă sunt heteroschedastice.

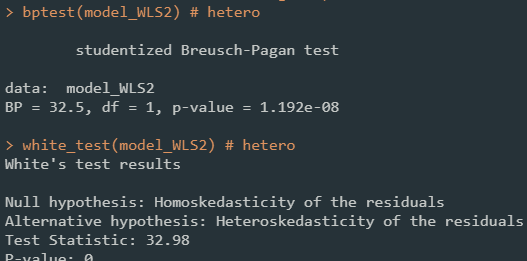


Figura 20. Regresie simplă, testele Breusch-Pagan și White după corectare (sursa:Rstudio)

Continuând cu autocorelarea, am aplicat inspectarea acesteia cu ajutorul graficului ACF. De aici putem observa cum reziduurile nu sunt autocorelate, deoarece există 6 lag-uri care depășesc liniile albastre punctate. Astfel, va trebui să corectăm modelul, însă, vom continua cu testele Durbin-Watson și Breusch-Godfrey.

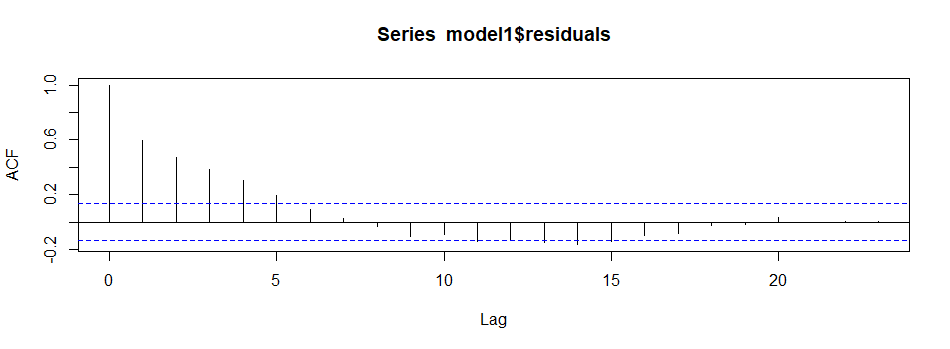


Figura 21. Regresie simplă, ACF înainte de corectare (sursa:RStudio)

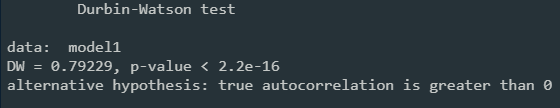


Figura 22. Regresie simplă, testul Durbin-Watson (sursa: RStudio)

Aplicând testul Durbin-Watson, testul statistic obținut ne sugerează faptul că nu există autocorelare între variabile, deoarece valoarea se încadrează în intervalul [0,2]. Acceptăm ipoteza nulă și o respingem pe cea alternativă.

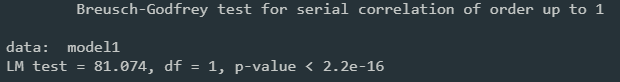


Figura 23. Regresie simplă, testul Breusch-Godfrey (sursa: RStudio)

Reieșind o valoarea a lui p-value foarte apropiată de 0 din testul Breusch-Godfrey rezultă că vom accepta ipoteza nulă și o vom respinge pe cea alternativă.

Corectarea autocorelării a fost efectuată prin adaugarea variabilei independente lag1 în modelul original de regresie simplă, astfel, retestând prin modelul ACF, putem observa cum autocorelarea a dispărut, în condițiile în care doar un singur lag a depășit pragul, însă, nu este semnificativ. Testul Durbin-Watson a schimbat p-value în 0.7819, care, fiind mai mare decât 0.1, rezultă faptul că reziduurile sunt nonoautocorelate. Din testul Breusch-Godfrey, p-value a reieșit cu valoarea de 0.1398, depășind pragul de 0.1.

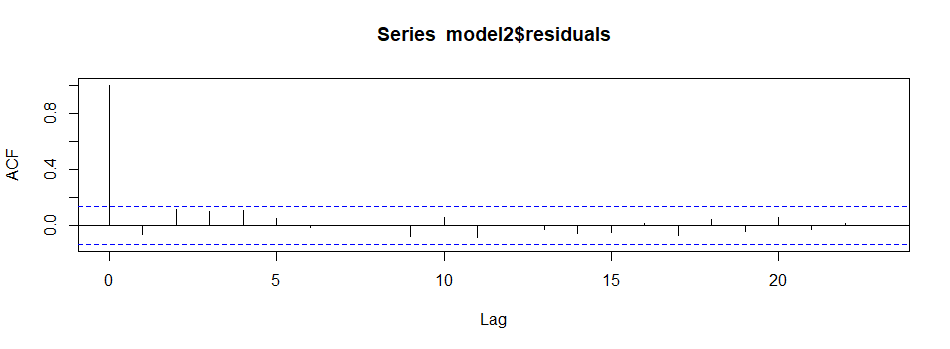


Figura 24. Regresie simplă, ACF după corectarea autocorelării (sursa:RStudio)

În continuare vom testa normalitatea prin testele Shapiro Wilk și Jarque-Bera. Pentru acestea, am logaritmat variabila dependentă pentru a putea observa cât de mult ne poate ajuta schimbarea formei liniare pentru a obține rezultate cât mai bune. Astfel, pentru variabila price și lprice am obținut rezultatele următoare:

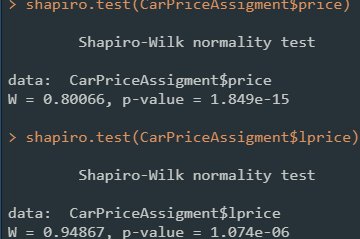


Figura 25. Regresie simplă, testul Shapiro-Wilk înainte de corectare (sursa: RStudio)

Dat fiind faptul că p-value este mai mic decât 0.5, vom respinge ipoteza nulă și o vom accepta pe .

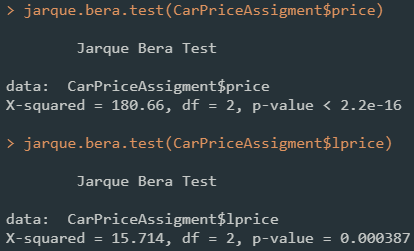


Figura 26. Regresie simplă, testul Jarque-Bera înainte de corectare (sursa: RStudio)

Pentru a corecta modelul din punct de vedere al distribuției de normalitate, am analizat graficul ce prezintă distanțele lui Cook:

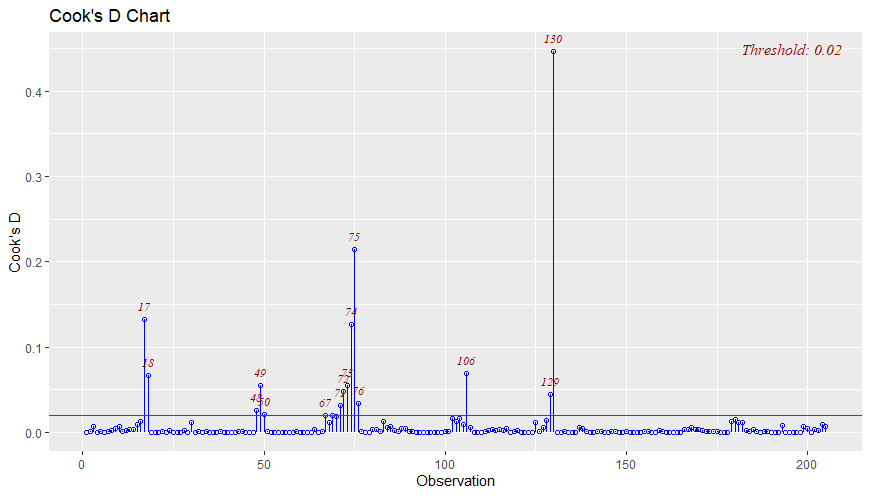


Figura 27. Regresie simplă, distanțele lui Cook (sursa:RStudio)

După corectare, graficul nu mai prezintă valori de tipul outlier:

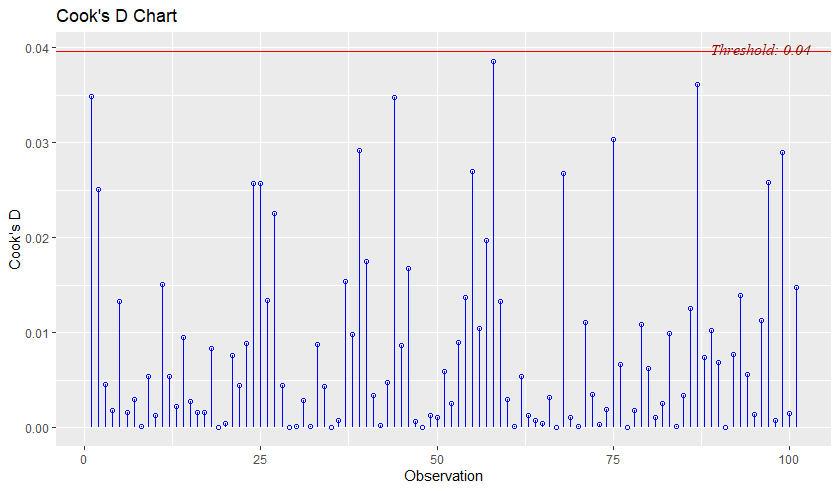


Figura 28. Regresie simplă, distanțele lui Cook după corectare (sursa:Rstudio)

Testele de normalitate au semnificat faptul că logaritmarea variabilei dependente a fost utilă, deoarece p-value a depășit valoarea de 0.1, modelul fiind corectat:

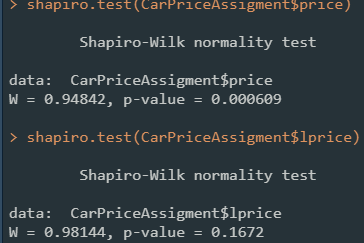


Figura 29. Regresie simplă, testul Shapiro-Wilk după corectare (sursa:Rstudio)

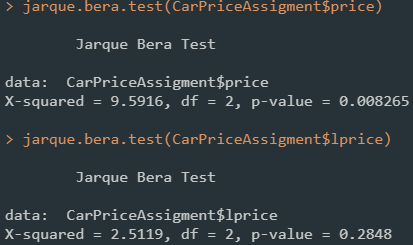


Figura 30. Regresie simplă, testul Jarque-Bera după corectare (sursa:Rstudio)

Pentru final, setul de date se va împărți într-unul de antrenament și unul de test calculând, în primul rând RMSE, apoi MAE, MSE și MAPE testând acuratețea modelului. Folosind forma log-log, output-ul pentru testul RMSE a fost de 0.6197881 (predicție bună), pentru MAE de 0.5372257 (predicție bună), pentru MSE 0.4611295 (predicție slabă), iar pentru MAPE 0.05681061 (predicție bună).

Formula regresiei multiple devine: **price = -8389.73 + 58.85 \* horsepower + 122.19 \* enginesize**. Această formulă poate fi interpretată astfel: la creșterea cu un cal putere prețul de achiziție al automobilului se va majora cu 58.55$, dat fiind faptul că am luat în calcul și dimensiunea motorului.

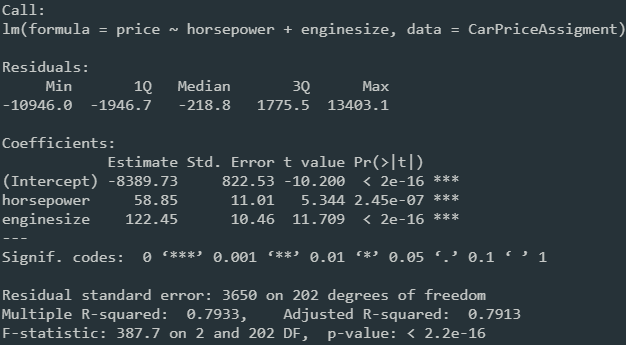


Figura 31. Regresie multiplă cu doi regresori, output (sursa: RStudio)

Dat fiind faptul că prețul de vânzare depinde de mai mulți factori în afară de cei folosiți am mai luat în considerare și consumul automobilului în oraș, deoarece majoritatea populației preferă un mijloc de transport propriu în detrimentul celui comun, motivul plauzibil fiind de a avea parte de cât mai mult comfort, însă eficiența unui autovehicul se măsoară în mare parte de cât de economic este când vine vorba de consum.

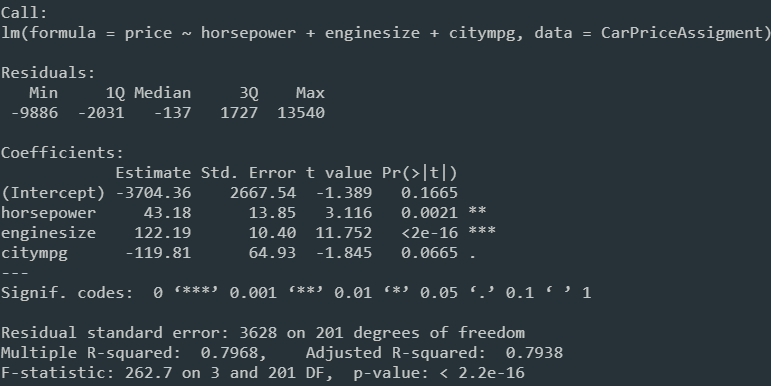


Figura 32. Regresie multiplă cu trei regresori, output (sursa: RStudio)

Din figura de mai sus luând în calcul variabila citympg rezultă faptul că la creșterea cu un cal putere al motorolui prețul se va majora cu doar 43.18$, deci cu cât depindem de mai mulți factori, creșterea variabilei dependente va fi din ce în ce mai mică. Formula modelului de regresie multiplă devine: **price = -3704.36 + 43.18 \* horsepower + 122.19 \* enginesize + (-119.81) \* citympg**.

Interpretarea Indicatorilor de bonitate se rezumă la R-squared (coeficientul de determinație) care este egal cu 0.7938  și indică faptul că, atât caii putere ai motorului, cât și dimensiunea motorului alături de consumul în mediul urban explică aproximativ 79.38% din prețul de vânzare. Pentru modelul nostru, eroarea standard este 3628.

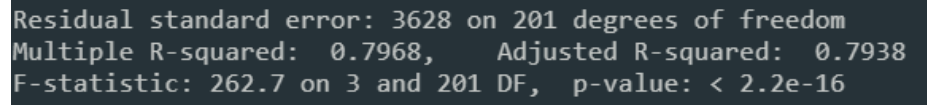


Figura 33. Regresie multiplă, indicator de bonitate (sursa: RStudio)

Pentru început, vom testa validitatea prin testul Fisher, modelul fiind valid pentru un nivel de semnificație de 1%, rezultatul fiind garantat cu o probabilitate de 99% și 203 de observații, rezultând k = 1 și n - k - 1 = 203 - 1 - 1 = 201 grade de libertate.

Validitatea parametrilorconstă în utilizarea testului T, astfel, respingem , acceptând , deoarece ≠ 0, rezultatele sunt garantate pe o probabilitate de 99%, pentru că prezintă valori mai mici de pragul de 1%.

Este nevoie ca acest model de regresie multiplă să testeze următorii pași:

- numărul de observații trebuie să fie mai mare decât numărul de variabile independente

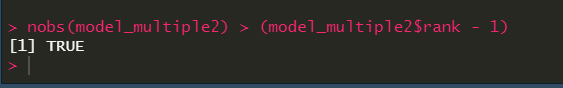


Figura 34. Regresie multiplă, nr observații > nr independente (sursa: RStudio)

- variabilitatea trebuie să fie pozitivă

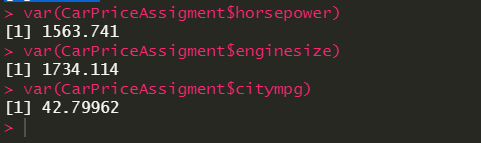


Figura 35. Regresie multiplă, variabilitate (sursa: RStudio)

- media reziduurilor trebuie să fie foarte aproape de 0 sau chiar 0

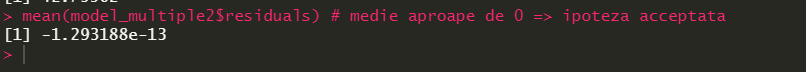
****

Figura 36. Regresie multiplă, media reziduurilor (sursa: RStudio)

- multicoliniaritatea ce poate fi determinată prin funcția vif(*model*)*,* dacă una dintre

variabile generează o valoare mai mare decât 10 atunci va trebui eliminată și înlocuită cu

altă variabilă

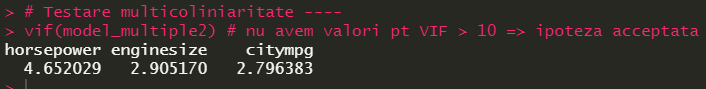


Figura 37. Regresie multiplă, multicoliniaritate (sursa: RStudio)

- reziduurile nu sunt corelate cu variabilele independente

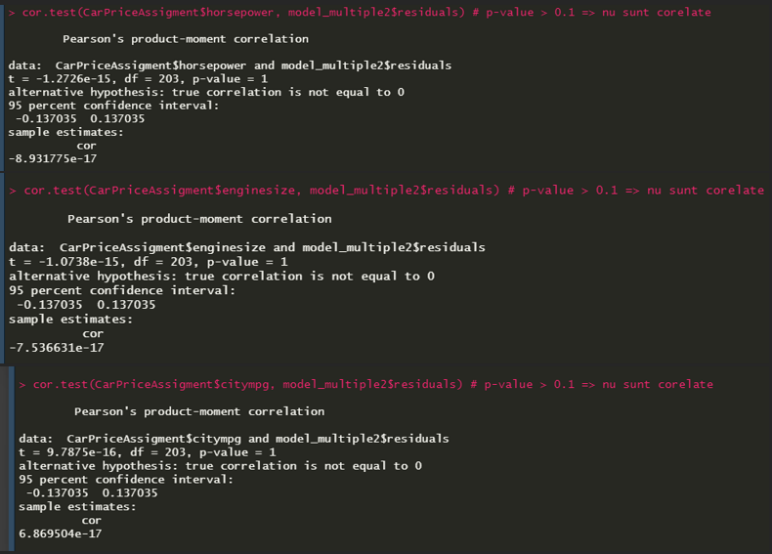


Figura 38. Regresie multiplă, corelația dintre reziduuri și var independente (sursa: RStudio)

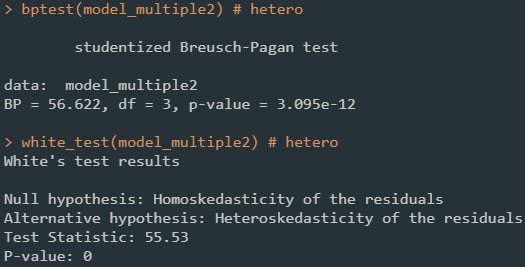


Figura 39. Regresie multiplă, testele Breusch-Pagan și White înainte de corectare (sursa:Rstudio)

P-value a luat valoarea 0.000000000003095 rezultând faptul că acceptăm ipoteza alternativă () și respingem ipoteza nulă (. P-value a luat de data aceasta valoarea 0, ceea ce indică faptul că acceptăm ipoteza alternativă și respingem ipoteza nulă. După corectare am obținut următoarele rezultate:

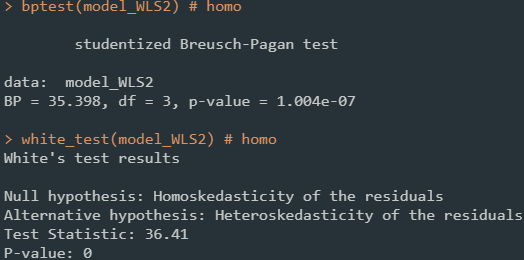


Figura 40. Regresie multiplă, testele Breusch-Pagan și White după corectare (sursa:Rstudio)

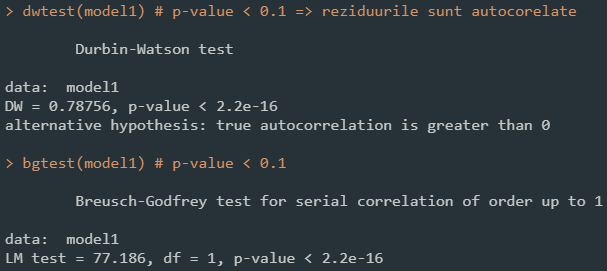


Figura 41. Regresie multiplă, testele Durbin-Watson și Breusch-Godfrey înainte de corectare (sursa: RStudio)

Continuând cu metoda grafică ne vom utiliza de graficul ACF pentru a putea pune accentul pe faptul că variabilele nu sunt autocorelate și va necesita corectarea modelului.

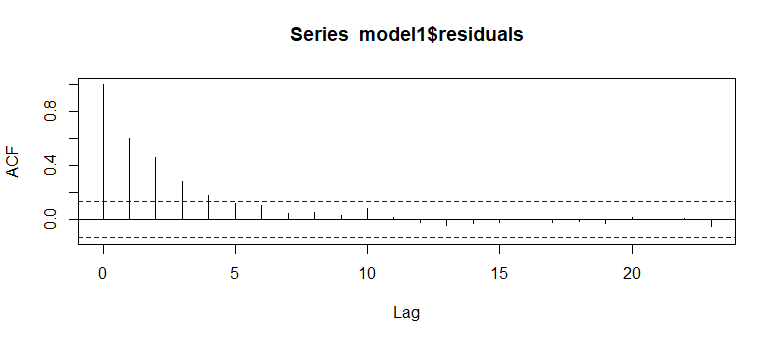


Figura 42. Regresie multiplă, ACF înainte de corectare (sursa: RStudio)

Corectarea autocorelării a fost efectuată prin adaugarea variabilei independente lag1 în modelul original de regresie multiplă, astfel, retestând prin modelul ACF, putem observa cum autocorelarea a dispărut, în condițiile în care doar un singur lag a depășit pragul, însă, nu este semnificativ.

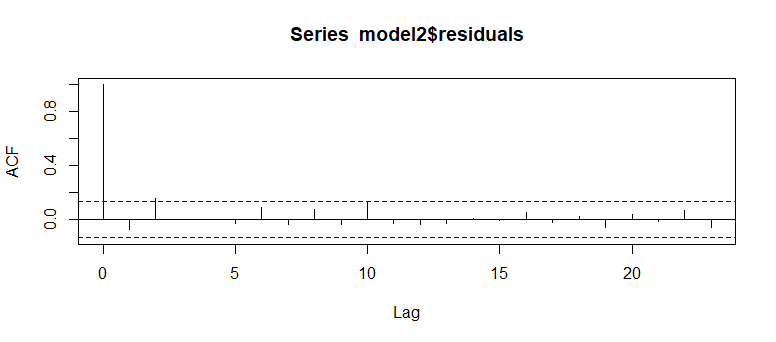


Figura 43. Regresie multiplă, ACF după corectare (sursa: RStudio)

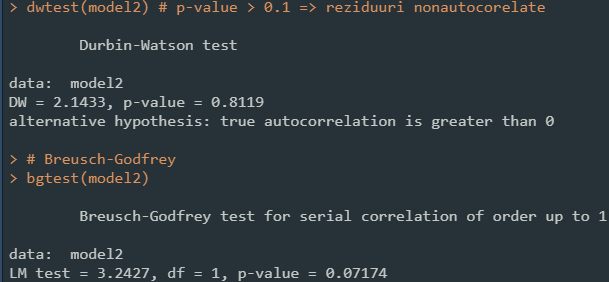


Figura 44. Regresie multiplă, testele Durbin-Watson și Breusch-Godfrey după corectare (sursa: RStudio)

Verificând normalitatea observațiilor prin testele aferente am obținut următoarele rezultate utilizând logaritmarea prețului de achiziție:

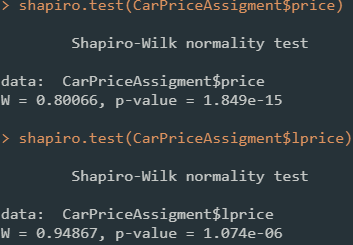


Figura 45. Regresie multiplă, testul Shapiro-Wilk înainte de corectare (sursa:RStudio)

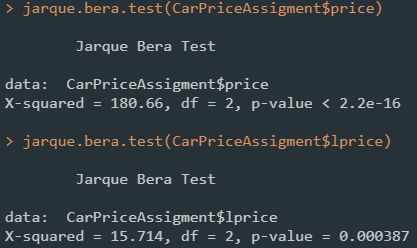


Figura 46. Regresie multiplă, testul Jarque-Bera înainte de corectare (sursa:RStudio)

Normalitatea observațiilor se poate observa în figura de mai jos:

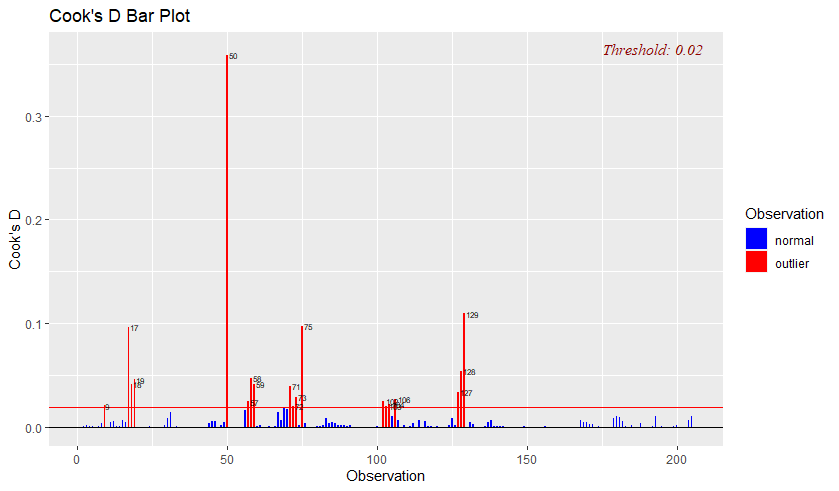


Figura 47. Regresie multiplă, distanța prin graficul Cook înainte de corectare (sursa: RStudio)

După corectare, graficul se prezintă astfel, alături de testele de normalitate ce prezintă valori mai mari decat 0.1:

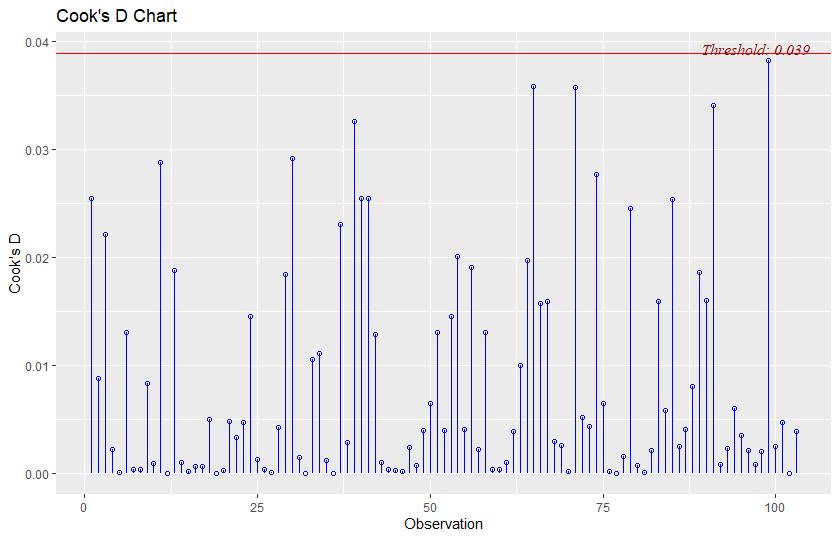


Figura 48. Regresie multiplă, distanța prin graficul Cook după corectare (sursa: RStudio)

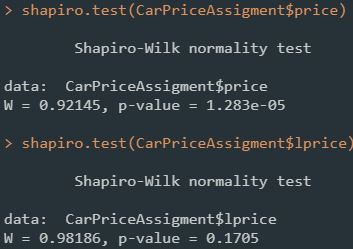


Figura 49. Regresie multiplă, testul Shapiro-Wilk după corectare (sursa:RStudio)

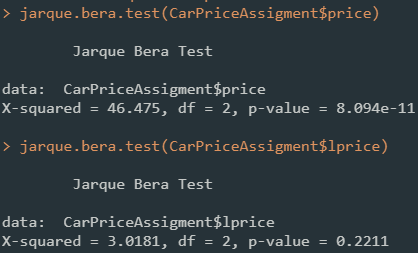


Figura 50. Regresie multiplă, testul Jarque-Bera după corectare (sursa:RStudio)

Pentru final, setul de date se va împărți într-unul de antrenament și unul de test calculând, în primul rând RMSE, apoi MAE, MSE și MAPE testând acuratețea modelului. Folosind forma log-log, output-ul pentru testul RMSE a fost de 0.7064829 (predicție bună) pentru MAE de 0.5114134 (predicție bună), pentru MSE 0.4126816 (predicție slabă), iar pentru MAPE 0.0543246 (predicție bună).

Formula regresiei multiple prin îmbunătățirea acesteia cu variabila dummy devine: **price = -2618.09 + 54.05 \* horsepower + 105.12 \* enginesize + (-185.12) \* citympg + (-4034.38) \* fueltype**. Această formulă poate fi interpretată astfel: la creșterea cu un cal putere prețul de achiziție al automobilului se va majora cu 54.05$, dat fiind faptul că am luat în calcul dimensiunea motorului, consumul mediu urban și tipul de combustibil.

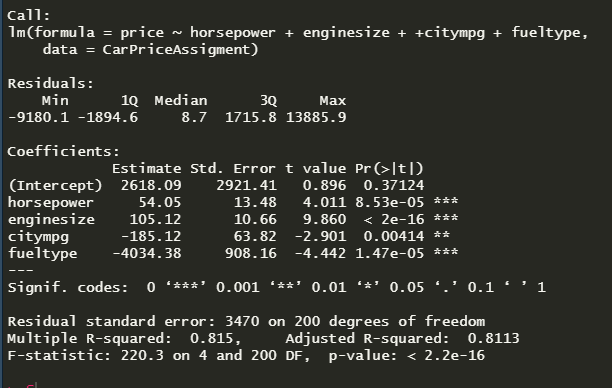


Figura 51. Regresie multiplă cu variabilă dummy, output (sursa: RStudio)

Interpretarea indicatorilor de bonitate se rezumă la R-squared (coeficientul de determinație) este egal cu 0.8113 și indică faptul că, atât caii putere ai motorului, cât și dimensiunea motorului alături de consumul în mediul urban și tipul de combustibil explică aproximativ 81.13% din prețul de vânzare. Pentru modelul nostru, eroarea standard este 3470.



Figura 52. Regresie multiplă cu variabilă dummy, indicatori de bonitate (sursa: RStudio)

Pentru început, vom testa **validitatea modelului**, folosindu-ne de testul fisher. Valoarea F-statistic este de 220.3, de unde rezultă că se respinge ipoteza și se accepta , modelul fiind valid pentru un nivel de semnificație de 1%, rezultatul fiind garantat cu o probabilitate de 99% și 202 de observații, rezultând k = 1 și n - k - 1 = 202 - 1 - 1 = 200 grade de libertate.

**Validitatea parametrilor** constă în utilizarea testului T, astfel, respingem , acceptăm , deoarece ≠ 0, rezultatele sunt garantate pe o probabilitate de 99%, pentru că prezintă valori mai mici de pragul de 1%.

Este nevoie ca acest model de regresie multiplă să testeze următorii pași:

- numărul de observații trebuie să fie mai mare decât numărul de variabile independente



Figura 53. Regresie multiplă, nr observații > nr independente (sursa: RStudio)

- variabilitatea trebuie să fie pozitivă

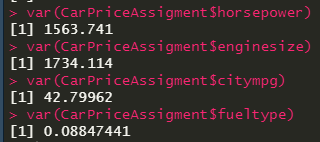


Figura 54. Regresie multiplă, variabilitate (sursa: RStudio)

# **Concluzii**

În urma aplicării modelelor de regresie ale aplicației 1 am concluzionat faptul că factorii contribuabili în vederea prețului de achiziție al automobilului sunt formați in primul rând de puterea pe care o dezvoltă motorul, astfel indicele de corelație format dintre variabila dependentă și cea independentă a indicat o legătură puternic pozitivă între acestea cu o valoare de 0.8.

Creând modelul de regresie simplă ipoteza indicelui de corelație s-a dovedit a fi relevantă prin faptul că variabila independentă a avut o semnificație de 99% generând coeficientul de determinație de 65% indicând faptul că, caii putere explică 65% din prețul de vânzare. Testând valabilitatea modelului cu un F-statistic de 382.2, modelul este valid cu o probabilitate de 1% fiind garantat cu o probabilitate de 99%, având eroarea standard egală cu 4717.

Testând heteoschedasticitatea în scopul de a transforma reziduurile în homoscedastice, chiar și prin corectare, testele corespondente heteroscedasticității au semnificat faptul că reziduurile au rămas homoschedastice.

În urma aplicării autocorelării, prin inspectarea acesteia cu ajutorul graficului ACF, a rezultat faptul că reziduurile nu sunt autocorelate. Astfel a trebuit să corectăm modelul având un singur lag care depășește pragul de autocorelare. De asemenea, am aplicat testul Durbin-Watson obținând P-value mai mare de 0.1 și testul Breusch-Godfrey obținând P-value mai mic de 0.1 rezultând faptul că reziduurile sunt non-autocorelate.

Aplicând testele Shapiro-Wilk utilizând variabila dependentă am obținut un P-value mai mic de 0.5 rezultând faptul că distribuția este semnificativ diferită de o distribuție normală și Jarque-Bera obținând faptul că erorile aleatoare nu urmează o distribuție normală. Am corectat normalitatea prin graficul ce semnifică distanțele lui Cook.

Împărțind setul de date în unul de antrenare și altul de testare am implementat regresia pe cel de antrenare și am testat acuratețea pe setul de testare obținând indicatorul abaterii pătratice medii mai mic decât 1 rezultând o predicție bună, indicatorul erorii pătratice medie rezultând astfel o predicție slabă, aplicând eroarea medie absolută a rezultat o predicție bună, iar eroarea procentuală medie absolută fiind mai mică decât 1 a rezultat o predicție bună.

Prin faptul că cilindreea motorului joacă un rol foarte important în construcția automobilului, cât și consumul în mediul urban, am îmbunătățit modelul de regresie simplă prin adăugarea acestor două variabile independente în modelul de regresie multiplă. Acest model explică aproximativ 79% din prețul de achiziție cu o eroare standard de 3628 si un F-statistic de 262.7, rezultatul fiind garantat cu o probabilitate de 99%.

Autenticitatea modelului a rezultat din numărul de observații care este mai mare decât numărul de variabile independente, din variabilitatea fiecărui argument independent ca fiind pozitiv, media reziduurilor foarte aproape de 0, multicoliniaritatea prin faptul că niciunul dintre factorii determinanți nu a generat o valoare mai mare decât 10, reziduurile nu sunt corelate cu variabilele independente.

În plus autenticitatea modelului a rezultat și din testarea ipotezelor pe reziduuri începând cu testul Breusch-Pagan din care a rezultat că reziduurile nu sunt distribuite cu o variație egală si testul White din care a rezultat ca variațiile pentru erori sunt egale. Am corectat modelul folosind WLS, însă, rezultatele au concluzionat faptul că reziduurile au ramas heteroschedastice.

Și în cazul modelului de regresie multiplă am testat autocorelarea prin testele Durbin-Watson și Breusch-Godfrey rezultând că nu există autocorelare între variabile și că variațiile erorii sunt egale. Folosind metoda grafică a rezultat faptul că reziduurile nu sunt autocorelate, astfel a trebuit să corectăm modelul.

Testând normalitatea prin Shapiro-Wilk valoarea testului statistic indică faptul că distribuția este semnificativ diferită de o distribuție normală, iar prin Jarque-Bera valoarea testului statistic a sugerat faptul că erorile aleatoare nu urmează o distribuție normală. Am corectat normalitatea prin graficul ce prezintă distanțele lui Cook pentru a nu mai avea valori de tipul outlier.

Acuratețea modelului a rezultat din valoarea testului RMSE ce a generat o predicție bună, la fel si pentru MEA si MAPE având o predicție slabă în cadrul lui MSE.

Un alt factor decizional în cadrul stabilirii prețului unui automobil se dovedește a fi sursa de alimentare a motorului, astfel am îmbunătățit modelul de regresie multiplă cu această variabilă dummy. De această dată, atât caii putere ai motorului, cât și dimensiunea motorului alături de consumul în mediul urban și tipul de combustibil explică aproximativ 81.13% din prețul de vânzare, valoarea lui F-statistic fiind de 220.3 și eroarea standard de 3470.

Asemănător cu modelul de regresie multiplă, acesta a îndeplinit condițiile necesare pentru un model valid precum numărul de observații mai mare decât numărul de variabile independente și variabilitatea să fie pozitivă.

# **Bibliografie**

Setul de date > <https://www.kaggle.com/datasets/hellbuoy/car-price-prediction>

PDSM > <https://ro.kamiltaylan.blog/manufacturers-suggested-retail-price-msrp/>

Puterea motorului > <https://www.autovit.ro/blog/ce-inseamna-cal-putere-si-cum-se-calculeaza/>

Dimensiunea motorului > <https://avtotachki.com/ro/chto-znachit-obem-dvigatelya/>

Consumul în mile pe galon ><https://www.carwow.co.uk/guides/running/what-is-mpg-0255#gref>

Tipuri de combustibil > <https://sdfauto.ro/instalatia-de-alimentare-a-motoarelor-cu-ardere-interna/>

1. Resale Price Prediction in the Used Car Market - Stefan Lessmann, Institute of Information Systems, [↑](#footnote-ref-1)
2. Conference paper - Price Discontinuities in an online used Car Market, Englmaier, Florian; Schmöller, Arno; Stowasser, Till [↑](#footnote-ref-2)
3. Capgemini - The Anatomy and Physiology of the Used Car Business [↑](#footnote-ref-3)
4. Transportation Research Part D: Transport and Environment - Bert Van Wee, Henri C. Moll, Jessica Dirks [↑](#footnote-ref-4)
5. New Car Sales and Used Car Stocks: A Model of the Automobile Market - James Berkovec [↑](#footnote-ref-5)