**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](#_Toc124973738)

[APLICAȚIA 1: Modele de regresie 4](#_Toc124973739)

[**Capitolul teoretic** 4](#_Toc124973740)

[***1.1. Literature review*** 4](#_Toc124973741)

[***1.2. Metodologia cercetării*** 6](#_Toc124973742)

[**Capitolul aplicativ** 8](#_Toc124973743)

[***2.1. Date utilizate*** 8](#_Toc124973744)

[***2.2.Rezultate empirice ale cercetării*** 11](#_Toc124973745)

[**Concluzii** 24](#_Toc124973746)

[**Bibliografie** 26](#_Toc124973747)

[APLICAȚIA 2: Modele cu date de tip panel 27](#_Toc124973748)

[**Capitolul teoretic** 27](#_Toc124973749)

[***1.2 Metodologia cercetării*** 27](#_Toc124973750)

[**Capitolul aplicativ** 28](#_Toc124973751)

[***2.1 Date utilizate*** 28](#_Toc124973752)

[***2.2 Rezultate empirice ale cercetării*** 30](#_Toc124973753)

[**Concluzii** 36](#_Toc124973754)

[**Bibliografie** 37](#_Toc124973755)

# **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**

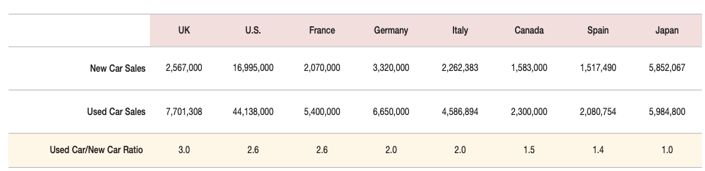
## **Capitolul teoretic**

### ***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:

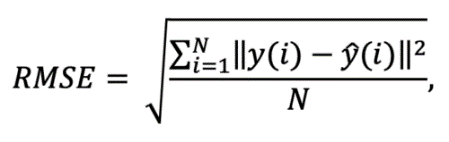


Figura 2. Formula eroare pătrată medie (sursa: Wikipedia)

Testul MAE – eroarea absolută medie

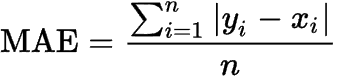


Figura 3. Formula eroare absolută medie (sursa: Wikipedia)

Testul MSE – eroarea medie pătratică

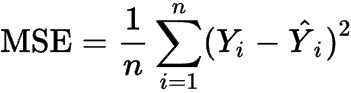


Figura 4. Formula eroare medie pătratică (sursa: Wikipedia)

Testul MAPE - eroarea procentuală medie absolută

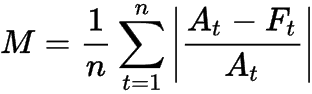


Figura 5. Formula eroare procentuală medie absolută (sursa: Wikipedia)

## **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.

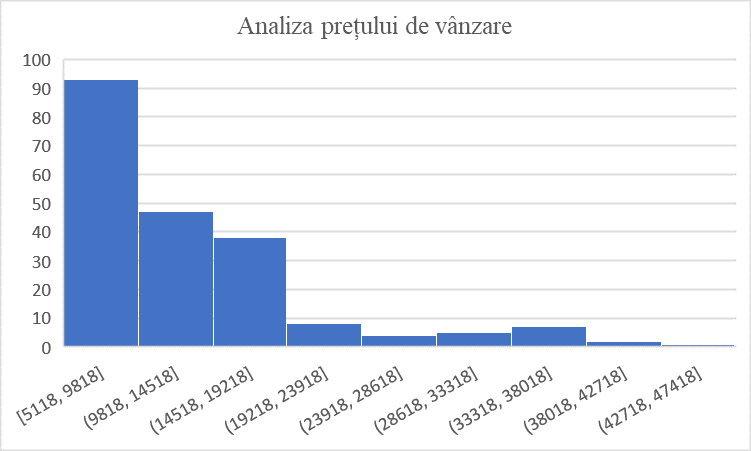


Figura 6. Grafic privind analiza prețului de vânzare (sursa: Excel)

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:

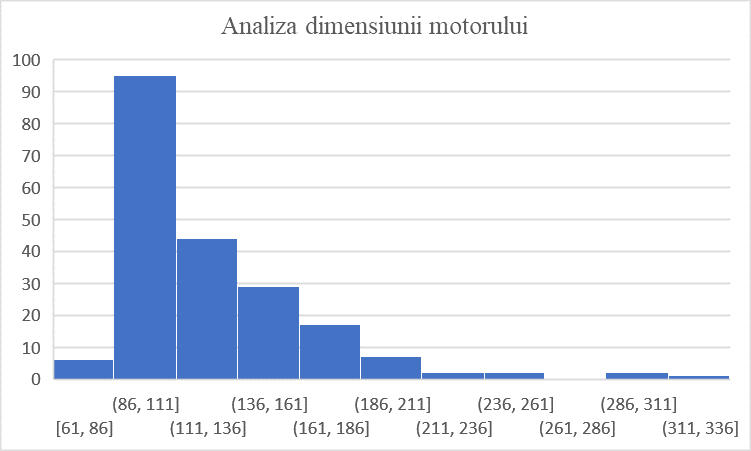


Figura 7. Grafic privind analiza dimensiunii motorului (sursa: Excel)

O altă caracteristică a automobilului se rezumă la performanțele motorului, astfel că, un motor cu o rata de compresie mai mare se dovedește a fi mai economic:

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

Totodată, motoarele pe benzină diferă față de cele pe motorină dat fiind faptul că cele pe benzină au o turație maximă mult mai mare decât cea a celor alimentate cu motorină:

Figura 9. Grafic privind analiza turației maxime a motorului (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:

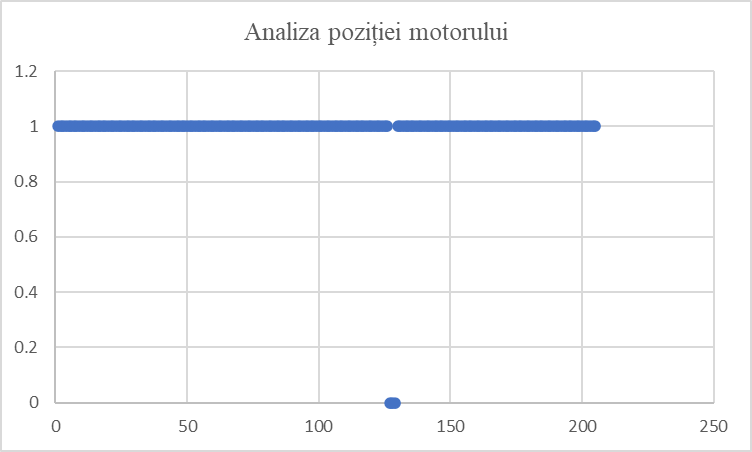


Figura 10. Grafic privind analiza poziției motorului (sursa: Excel)

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 cilindreea motorului, 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).

* Cilindreea 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.

Din punct de vedere al îmbunătățirii modelului de regresie simplă, am luat în considerare raportul de compresie a motorului, fiind un factor destul de important în ceea ce privește cât de economic este automobilul și turația maximă a motorului, rezultând un model de regresie multiplă.

* Raportul de compresie al motorului (compressratio)

Acesta reprezintă raportul dintre volumul camerei de ardere dintr-un piston atunci când se află cel mai jos și volumul camerei de ardere atunci când pistonul se află cel mai sus.

* Turația maximă a motorului (peakrpm)

Se măsoară în rotații pe minut, reprezintă numărul rotațiilor efectuate într-un minut de arborele cotit.

În cele din urmă, am îmbunătățit modelul de regresie multiplă prin adăugarea unei variabile independente dummy, locația motorului, fiind împărțită în două categorii: front = 1, rear = 0 :

* Poziționarea motorului (enginelocation)

Aceasta se poate împărți în mai multe feluri, longitudinal sau transversal în cadrul poziției din față, mijloc sau spate. Majoritatea automobilelor au motorul situat transversal în partea din față din rațiuni de costuri.

### ***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 dimensiunea motorului. Coeficientul de corelație este 0.87, aparținând de intervalul [-1,1].

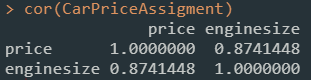


Figura 11. Regresie simplă, indicele de corelație (sursa: RStudio)

Din figura 7 putem deduce faptul că regresia simplă are forma **price = -8005.446 + 167.698 \* enginesize** conform . Această formulă poate fi interpretată astfel: la creșterea cu 1 cm3 a dimensiunii motorului, prețul de achiziție crește cu 167.698$, probabilitatea fiind de 99%. Având în vedere faptul că dimensiunea motorului măsoară valori între [61, 336], cu cât cilindreea motorului este mai mare, cu atât valoarea mașinii va avansa în preț.

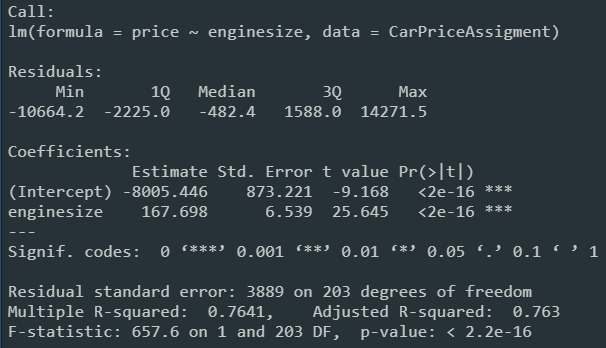


Figura 12. Regresie simplă, output (sursa: RStudio)

Valoarea F-statistic este de 657.6, 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 0.7641  și indică faptul că, cilindreea motorului autovehiculului explică aproximativ 76.41% 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 3889.

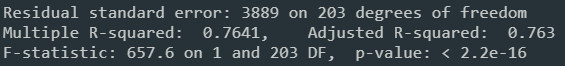


Figura 13. Regresie simplă, indicatori de bonitate (sursa: RStudio)

Modul în care sunt situate observațiile în comparație cu dreapta estimată se reflectă în figura de mai jos, astfel încât majoritatea motoarelor măsoară cilindreea până în pragul de 150, existând și excepții, precum Porsche Cayenne, dezvoltând 203 cm3 și Porsche Panamera cu 194, iar la polul opus se află Buick Century Special cu doar 308 cm3.

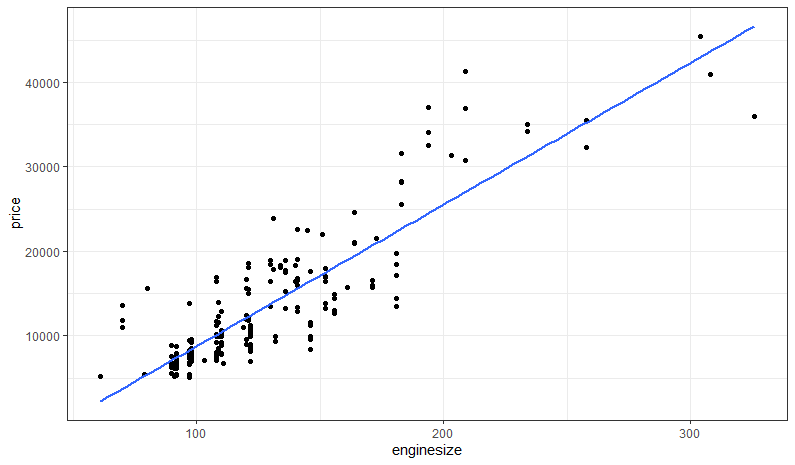


Figura 14. Regresie simplă, observații cu dreapta estimată (sursa: RStudio)

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.

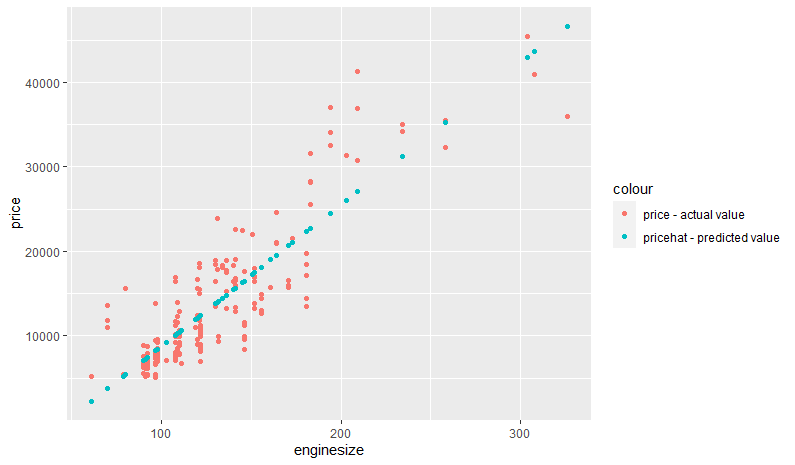


Figura 15. Regresie simplă, valorile estimate pentru variabila dependentă (sursa: RStudio)

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 o dimensiune mică a motorului (până în 150 cm3) 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.

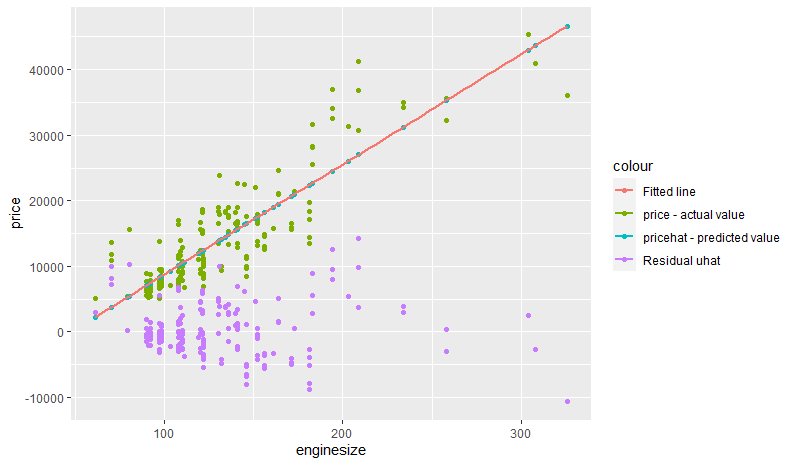


Figura 16. Regresie simplă, ajustarea reziduurilor (sursa: RStudio)

Testând ipotezele pe reziduuri, am început prin heteroschedasticitate, aplicând testul Breusch-Pagan în scopul de a transforma reziduurile din heteroschedastice în homoschedastice, astfel, valoarea lui p-value este de 3.925e-08 rezultând că reziduurile sunt homoschedastice. Rulând testul White, p-value a avut valorea 0, sub 0.1, rezultând faptul că modelul va trebui corectat.

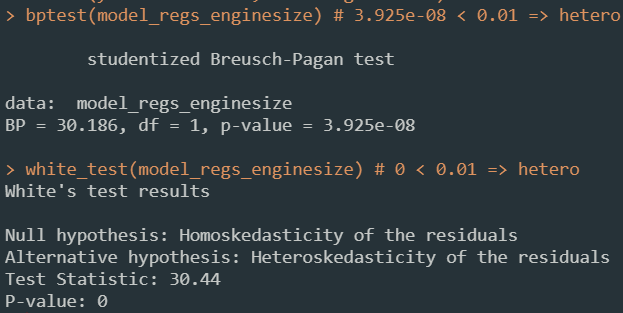


Figura 17. Regresie simplă, testele Breusch-Pagan și White înainte de corectare (sursa:Rstudio)

Corectarea a fost efectuată prin modelul WLS, obținând un p-value generat de către testele pentru homoscedasticitate mai mare decât 0.01:

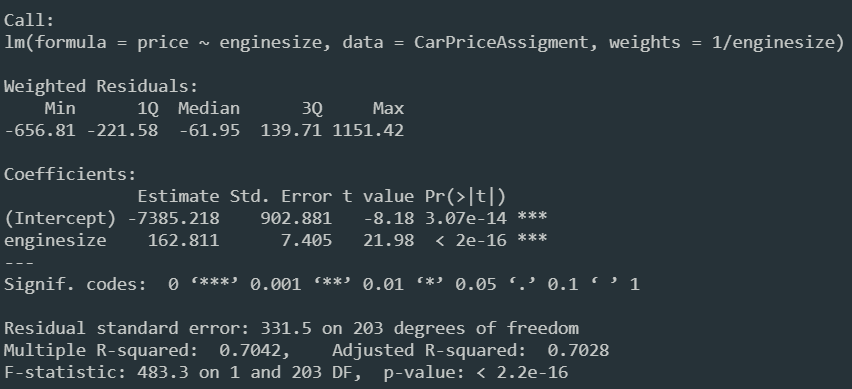


Figura 18. Regresie simplă, modelul WLS (sursa: RStudio)

De această dată, testul Breusch-Pagan a oferit un rezultat pozitiv în ceea ce privește corectitudinea modelului, însă, testul White a semnificat faptul că încă avem heteroscedasticitate:

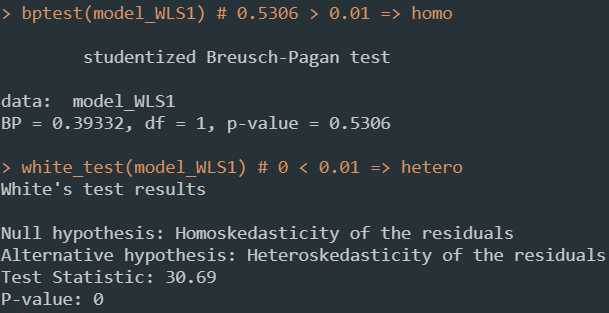


Figura 19. Regresie simplă, testele Breusch-Pagan și White după prima corectare (sursa:Rstudio)

În urma rezultatelor am decis să corectăm modelul prin WLS2, astfel:

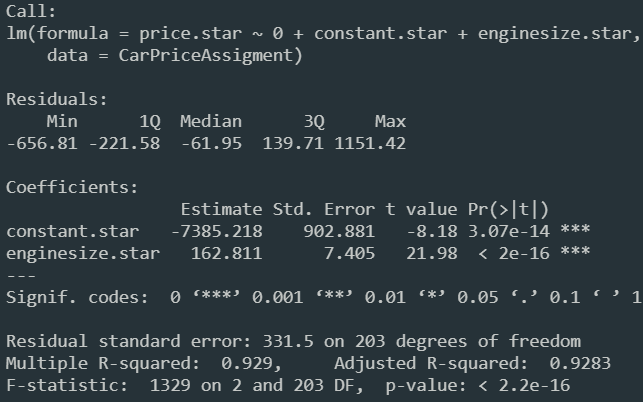


Figura 20. Regresie simplă, modelul WLS2 (sursa: RStudio)

De această dată am putut accepta ipoteza nulă, respingând-o pe cea alternativă, modelul devenind corectat:

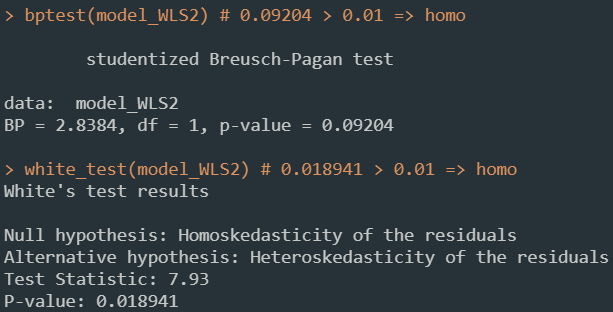


Figura 21. Regresie simplă, testele Breusch-Pagan și White după a doua corectare (sursa:Rstudio)

Continuând cu autocorelarea, am aplicat inspectarea acesteia cu ajutorul graficului ACF. De aici putem observa cum reziduurile nu sunt autocorelate, deoarece există 5 lag-uri care depășesc liniile albastre punctate. Astfel, va trebui să corectăm modelul, însă, vom continua cu testele Durbin-Watson și Breusch-Godfrey.

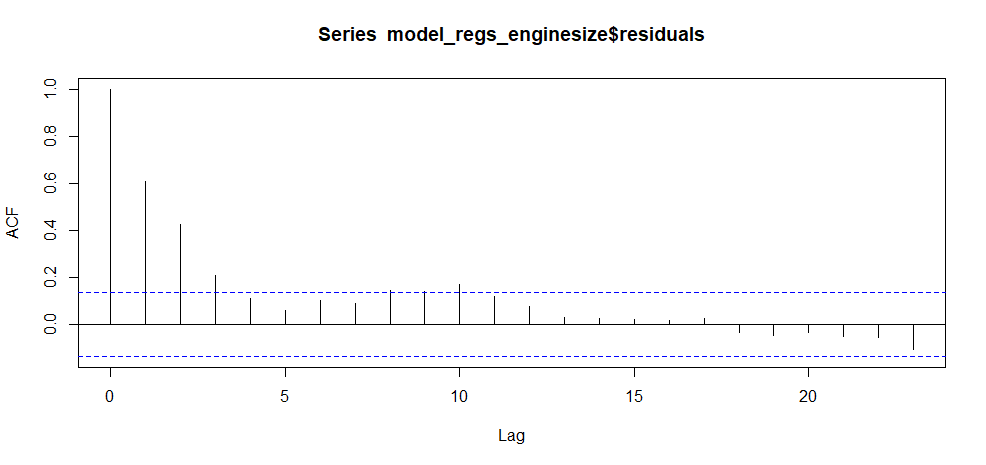


Figura 21. Regresie simplă, ACF înainte de corectare (sursa:RStudio)

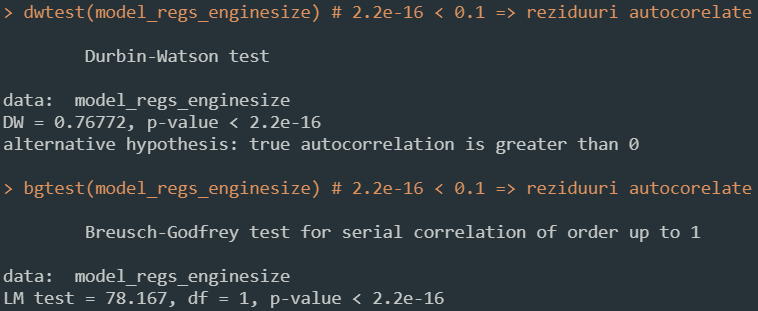


Figura 22. Regresie simplă, testele Durbin-Watson și Breusch-Godfrey înainte de corectare (sursa: RStudio)

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 generat un p-value de 0.6534, care, fiind mai mare decât 0.1, rezultă faptul că reziduurile sunt nonoautocorelate. Din testul Breusch-Godfrey, p-value a reieșit cu valoarea de 0.3648, depășind pragul de 0.1.

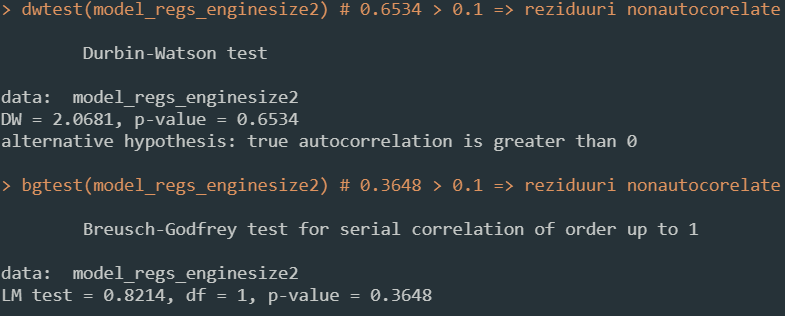


Figura 23. Regresie simplă, testele Durbin-Watson și Breusch-Godfrey după corectare (sursa: RStudio)

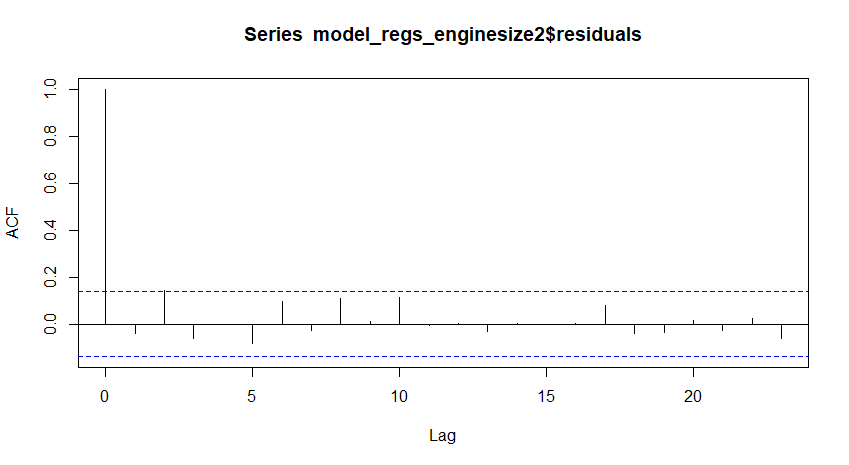


Figura 24. Regresie simplă, ACF după corectarea autocorelării (sursa:RStudio)

În continuare vom testa normalitatea prin testul Jarque-Bera:

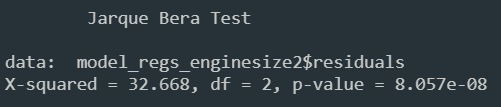


Figura 25. Regresie simplă, testul Jarque-Bera înainte de corectare (sursa:RStudio)

Dat fiind faptul că p-value este mai mic decât 0.05, vom respinge ipoteza nulă și o vom accepta pe .

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

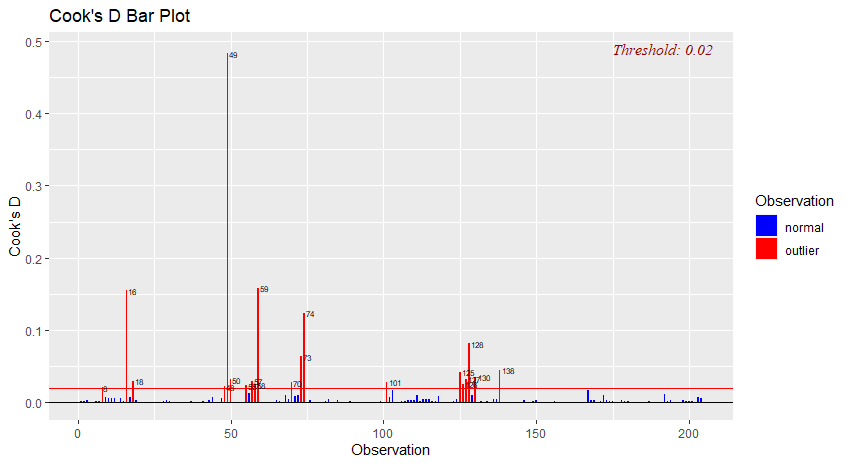


Figura 26. Regresie simplă. distanțele lui Cook înainte de corectare (sursa:RStudio)

În urma corectării am obținut un p-value oferit de către testul Jarque-Bera de 0.07432, mai mare decât 0.05, astfel, modelul este corectat, chiar dacă mai prezintă valori de tip outlier prin intermediul graficului specific:

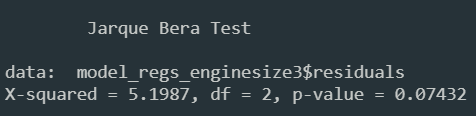


Figura 27. Regresie simplă, testul Jarque-Bera după corectare (sursa:RStudio)

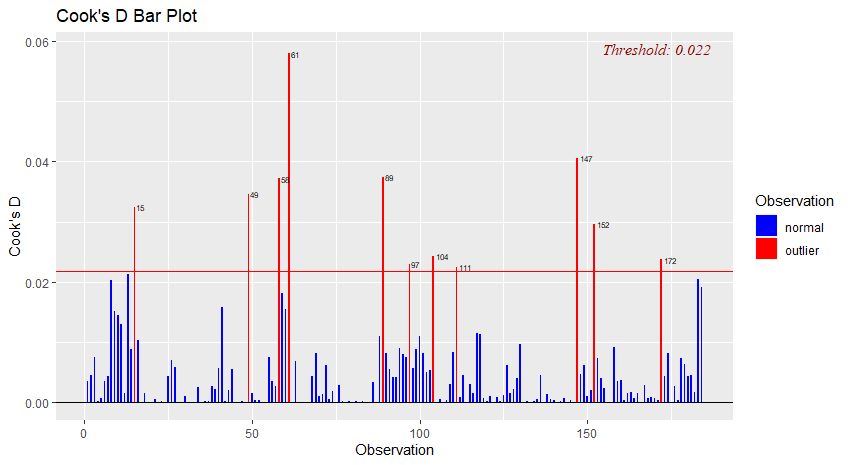


Figura 28. Regresie simplă, distanțele lui Cook după corectare (sursa:Rstudio)

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 0.2590731 (predicție bună), pentru MAE de 0.2010948 (predicție bună), pentru MSE 0.0671189 (predicție slabă), iar pentru MAPE 0.02148163 (predicție bună).

De asemenea, în cadrul prognonei am ales 3 valori pentru dimensiunea motorului: 70, 200 și 300 pentru a putea afla valori ale prețului ce diferă în cadrul cilindreei. Astfel, cu cât avem o cilindree mai mare, cu atât prețul va crește:



Figura 29. Regresie simplă, prognoză (sursa: RStudio)

Din figura de mai jos, luând în calcul variabilele semnificative în proporție de 99%, enginesize, compressratio și peakrpm rezultă faptul că la creșterea cu un cm3 al motorolui prețul va scădea cu -2.832e+04$, 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 = -2.832e+04 + 1.761e+02 \* enginesize + 2.535e+02 \* enginesize + 2.535e+02 \* compressratio + 3.253e+00 \* peakrpm**

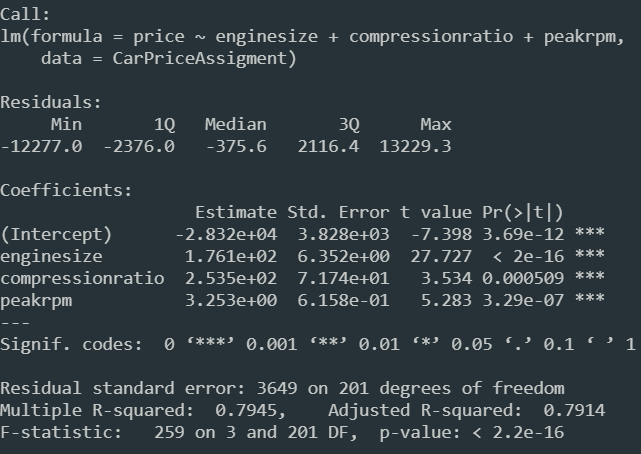


Figura 30. Regresie multiplă cu trei regresori, output (sursa: RStudio)

Interpretarea Indicatorilor de bonitate se rezumă la R-squared (coeficientul de determinație) care este egal cu 0.7945  și indică faptul că, atât dimensiunea motorului, cât și raportul de compresie alături turația maximă a motorului explică aproximativ 79.45% din prețul de vânzare. Pentru modelul nostru, eroarea standard este 3649.

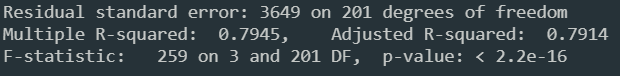


Figura 31. Regresie multiplă, indicator de bonitate (sursa: RStudio)

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 un model de regresie multiplă să verifice următorii pași:

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



Figura 32. Regresie multiplă, număr observații > număr independente (sursa: RStudio)

- variabilitatea trebuie să fie pozitivă

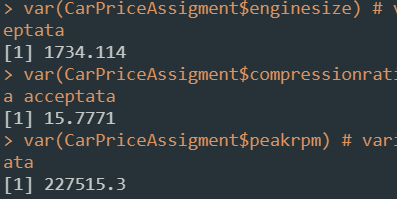


Figura 33. Regresie multiplă, variabilitate (sursa: RStudio)

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



Figura 34. 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

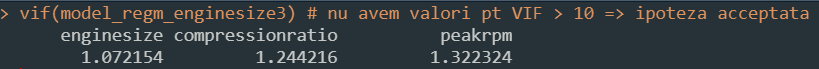


Figura 35. Regresie multiplă, multicoliniaritate (sursa: RStudio)

- reziduurile nu sunt corelate cu variabilele independente

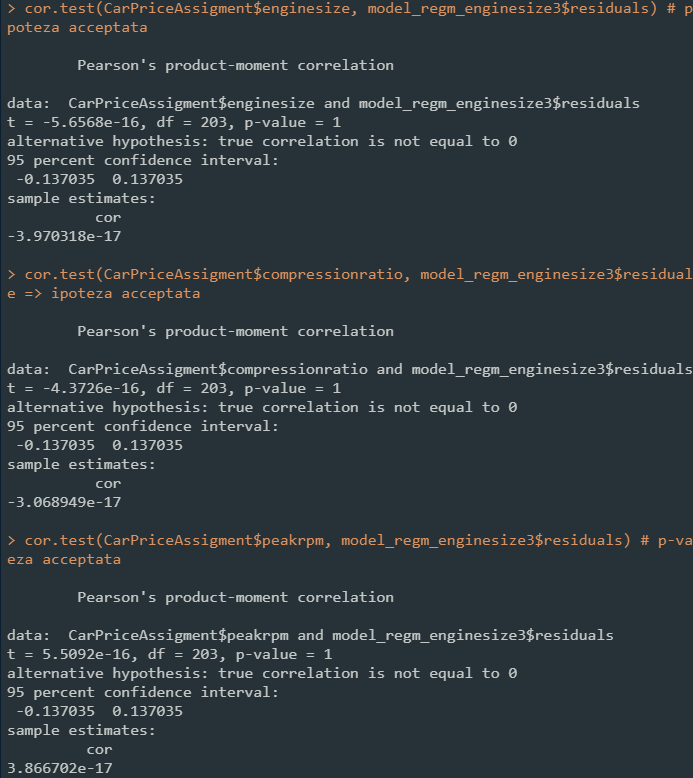


Figura 36. Regresie multiplă, corelația dintre reziduuri și var independente (sursa: RStudio)

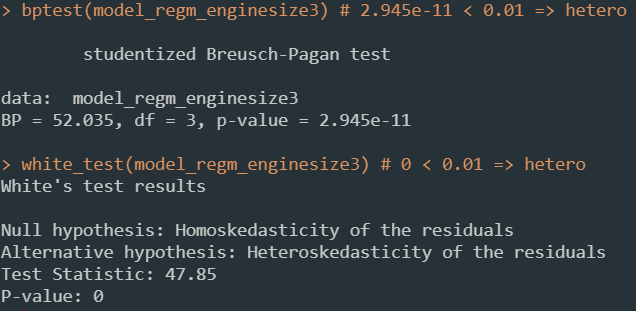


Figura 37. Regresie multiplă, testele Breusch-Pagan și White înainte de corectare (sursa:Rstudio)

P-value a luat valoarea de sub 0.1 rezultând faptul că acceptăm ipoteza alternativă () și respingem ipoteza nulă (.  Prin utilizarea modelului WLS în scopul corectării modelului, nu am reușit să corectăm heteroscedasticitatea, astfel, am utilizat forma de tip log-log în cadrul modelului WLS obținând următoarele rezultate:

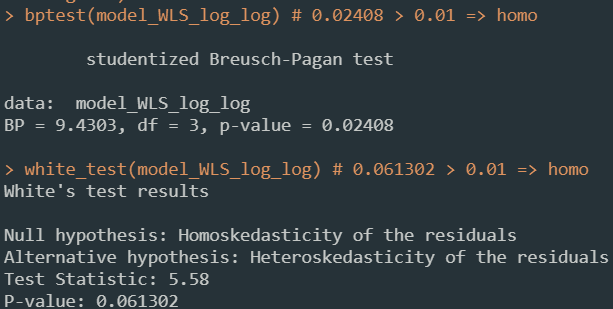


Figura 38. Regresie multiplă, testele Breusch-Pagan și White după corectare (sursa:Rstudio)

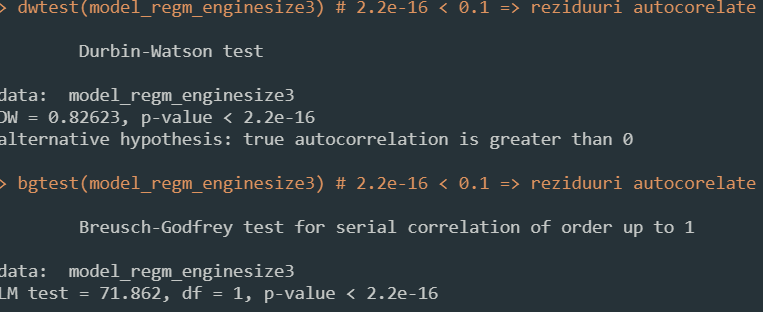


Figura 39. Regresie multiplă, testele Durbin-Watson și Breusch-Godfrey înainte de corectare (sursa: RStudio)

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.

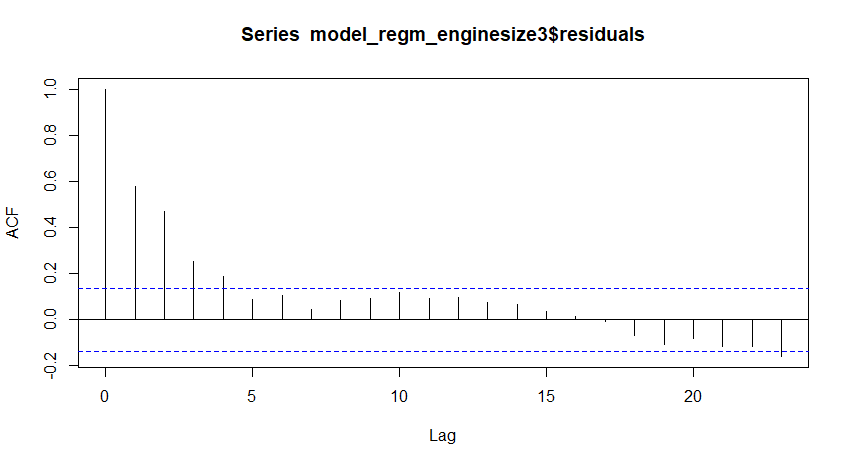


Figura 40. Regresie multiplă, ACF înainte de corectare (sursa: RStudio)

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.

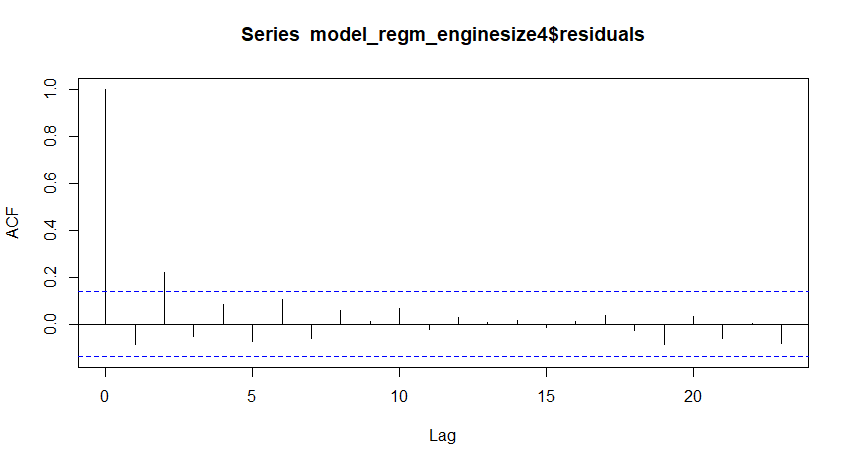


Figura 41. Regresie multiplă, ACF după corectare (sursa: RStudio)

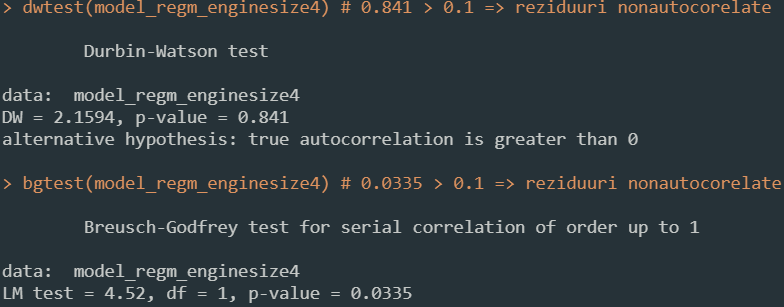


Figura 42. Regresie multiplă, testele Durbin-Watson și Breusch-Godfrey după corectare (sursa: RStudio)

Verificând normalitatea observațiilor prin tesul Jarque-Bera, am acceptat ipoteza alternativă, modelul trebuind a fi corectat:

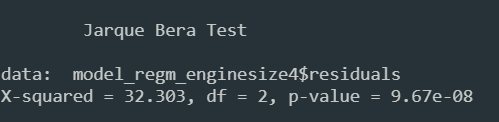


Figura 43. Regresie multiplă, testul Jarque-Bera înainte de corectare (sursa:RStudio)

Normalitatea observațiilor se poate observa în figura de mai jos:

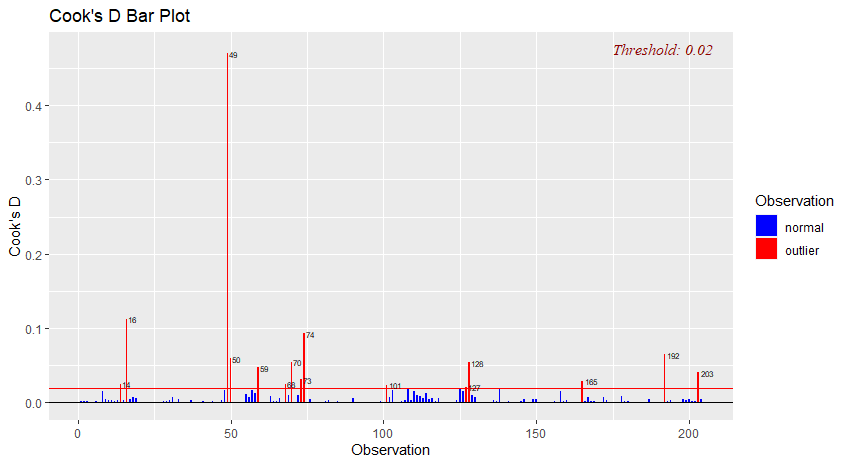


Figura 44. Regresie multiplă, distanța prin graficul Cook înainte de corectare (sursa: RStudio)

După eliminarea valorilor de tip outlier, graficul se prezintă astfel, alături de alături de un p-value generat de către testul Jarque-Bera mai mare decât 0.05:

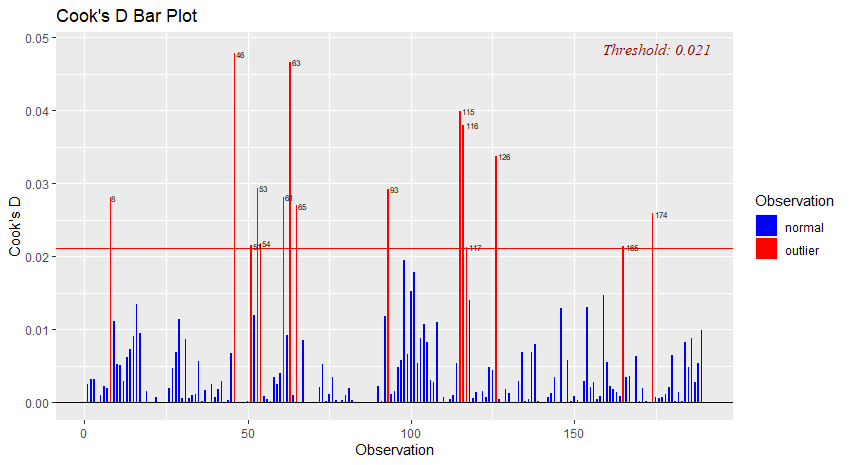


Figura 45. Regresie multiplă, distanța prin graficul Cook după corectare (sursa: RStudio)

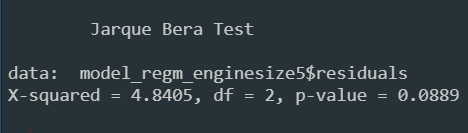


Figura 49. Regresie multiplă, testul Shapiro-Wilk după corectare (sursa:RStudio)

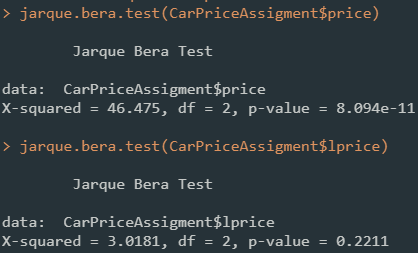


Figura 50. Regresie multiplă, testul Jarque-Bera după corectare (sursa:RStudio)

Formula regresiei multiple prin îmbunătățirea acesteia cu variabila dummy devine: **price = -1.545e+04 + 1.755e+02 \* enginesize + 2.185e+02 \* compressratio + 2.773e+00 \* peakrpm -1.014e+04 \* enginelocation**. Această formulă poate fi interpretată astfel: la creșterea cu un cm3 al motorului, prețul de achiziție al automobilului se va majora cu 1.755e+02 $, dat fiind faptul că am luat în calcul raportul de compresie, turația maximă a motorului și locația acestuia.

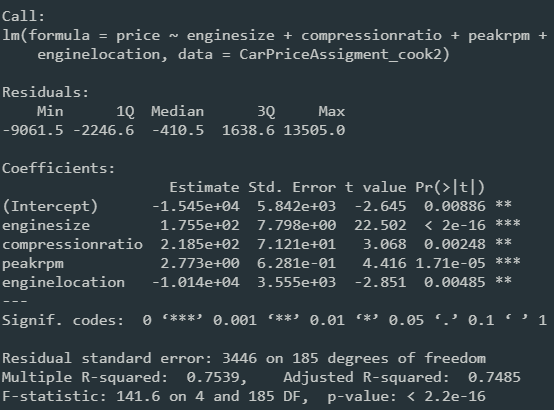


Figura 51. Regresie multiplă cu variabilă dummy, output (sursa: RStudio)

Interpretarea indicatorilor de bonitate se rezumă la R-squared (coeficientul de determinație) este egal cu 0.7539 și indică faptul că, atât dimensiunea, cât și raportul de compresie alături de turația maximă a motorului și poziționarea acestuia explică aproximativ 75.39% din prețul de vânzare. Pentru modelul nostru, eroarea standard este 3446.

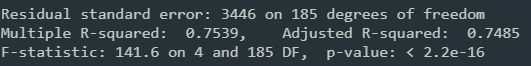


Figura 52. Regresie multiplă cu variabilă dummy, indicatori de bonitate (sursa: RStudio)

Pentru început, vom testa **validitatea modelului**, folosindu-ne de testul fisher. Valoarea F-statistic este de 141.6, 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 185 de observații, rezultând k = 1 și n - k - 1 = 187 - 1 - 1 = 185 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

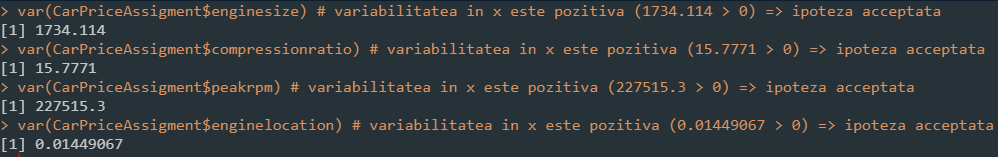


Figura 54. Regresie multiplă, variabilitate (sursa: RStudio)

## **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 cilindreea motorului, astfel indicele de corelație format dintre variabila dependentă și cea independentă a indicat o legătură puternic pozitivă între acestea cu o valoare de 0.87.

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 dimensiunea motorului explică 76.41% din prețul de vânzare. Testând valabilitatea modelului cu un F-statistic de 657.6, modelul este valid cu o probabilitate de 1% fiind garantat cu o probabilitate de 99%, având eroarea standard egală cu 3889.

Testând homoscedasticitatea în scopul de verifica dacă reziduurile sunt homoscedastice, am obținut faptul că sunt heteroschedastice, fapt ce a necesitat corectarea modelului.

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 autocorelarea a dispărut.

Aplicând testul Jarque-Bera am obținut un P-value mai mic de 0.05 rezultând faptul că distribuția este semnificativ diferită de o distribuție normală, fapt ce a necesitat să corectăm modelul 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 raportul de compresie joacă un rol foarte important în construcția motorului automobilului, cât și turația maximă, am îmbunătățit modelul de regresie simplă prin adăugarea acestor două variabile independente în modelul de regresie multiplă. Acest model explică aproximativ 79.45% din prețul de achiziție cu o eroare standard de 3649 si un F-statistic de 259, 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, forma log-log.

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 normalitatății prin Jarque-Bera a constat în valoarea testului statistic ce a sugerat faptul că erorile aleatoare nu urmează o distribuție normală. Am corectat normalitatea prin graficul ce prezintă distanțele lui Cook.

Un alt factor decizional în cadrul stabilirii prețului unui automobil se dovedește a fi poziționarea motorului, astfel am îmbunătățit modelul de regresie multiplă cu această variabilă dummy. De această dată, atât dimensiunea motorului, cât și raportul de compresie alături de turația maximă și poziționarea explică aproximativ 75.39% din prețul de vânzare, valoarea lui F-statistic fiind de 141.6 și eroarea standard de 3446.

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 pozitivă.

## **Bibliografie**

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

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

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

Raportul de compresie al motorului > <https://www.4tuning.ro/ghidul-soferului/ce-inseamna-la-un-motor-raport-de-compresie-si-cu-ce-ne-ajuta-sa-stim-care-e-acela-30147.html>

Turația maximă a motorului > <https://www.autovit.ro/blog/ce-este-turatia-motorului-si-care-este-rolul-senzorului-de-turatie/>

Poziționarea motorului > <https://avtotachki.com/ro/razlichnye-vozmozhnye-polozheniya-dvigatelya/>

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

# **APLICAȚIA 2: Modele cu date de tip panel**

## **Capitolul teoretic**

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

### ***1.2. Metodologia cercetării***

Modelul de regresie liniară simplă pentru datele de tip panel este următoarul, unde prețul bursier este variabila dependentă, iar variabilele independente sunt rata de schimb și venitul pe cap de locuitor:

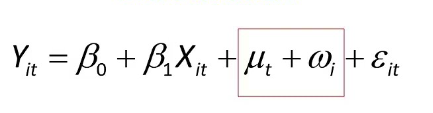


Figura 55. Formulă model de regresie liniară simplă de tip panel (sursa: Wikipedia)



Figura 56. Formulă model clasic de regresie liniară OLS(sursa: Wikipedia)

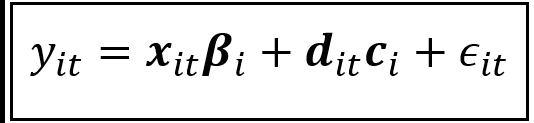


Figura 57. Formulă model de tip FE(sursa: Wikipedia)

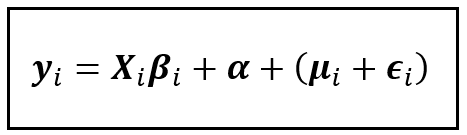


Figura 58. Formulă model de tip RE(sursa: Wikipedia)

Testul FSatistic pentru alegerea celei mai adecvate variante de model prin testarea intre regresie:

: model cu efecte fixe

: model de tip OLS

Testul Hausman il utilizam pentru a decide intre FE si RE:

: model cu efecte random

: model cu efecte fixe

Testul Lagrange Multiplier:

: nu sunt necesare efectele fixe in timp

: sunt necesare efectele fixe in timp

Testul Lagrange Multiplier pentru model de tip pool:

: variațiile in timp sunt 0

: variațiile in timp sunt diferite de 0

Testul Breusch-Pagan LM și Perasan CD având ipotezele următoare:

: reziduurile intre entitati nu sunt corelate

: reziduurile intre entitati sunt corelate

Testul Breusch-Godfrey:

: Nu exista autocorelate

: autocorelarea este prezenta

Testul Breusch-Pagan pentru heteroschedasticitate:

: homoschedasticitate

: heteroschedasticitate

## **Capitolul aplicativ**

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

Din punct de vedere al sursei de corectare a datelor, am extras de pe platforma kaggle următorul set de date cu variabile aferente. Setul de date conține 8 țări și o regiune administrativă specială: China, Franța, Germania, Hong Kong, India, Japonia, Spania, Regatul Unit și Statele Unite ale Americii pentru perioada 1980-2020.

Transformarea datelor a constat în păstrarea acestora doar pentru perioada 2010-2020. alături de eliminarea coloanelor cu numere negative, precum: rata inflației, procentajul PIB, rata de șomaj, balanța comercială și trezoreria.

Pentru a analiza cu acuratețe acest set de date am început cu prețul bursier:

Figura 59. Grafic privind analiza indicelui prețului bursier (sursa: Excel)

Continuând cu rata de schimb:

Figura 60. Grafic privind analiza ratei de schimb (sursa: Excel)

Finalizând cu venitul per cap de locuitor:

Figura 61. Grafic privind analiza venitului pe cap de locuitor (sursa: Excel)

Indicele prețului bursier semnifică medii ponderate în cadrul acțiunilor cotate la o bursă de valori, iar factorii de care depinde sunt următorii indicatori macroeonomici din cadrul setului de date ales:

* Cursul valutar se referă la valoarea unei unități monetare provenită din moneda unei țări exprimată în unități monetare ale altei țări.
* Venitul sau PIB-ul per cap de locuitor este un indicator ce semnifică producția totală a unei țări ce preia produsul intern brut pentru a îl împărți la numărul populației.

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

Pentru început, graficul de mai jos semnifică corelația dintre indicele prețului bursier între spațiu și timp:

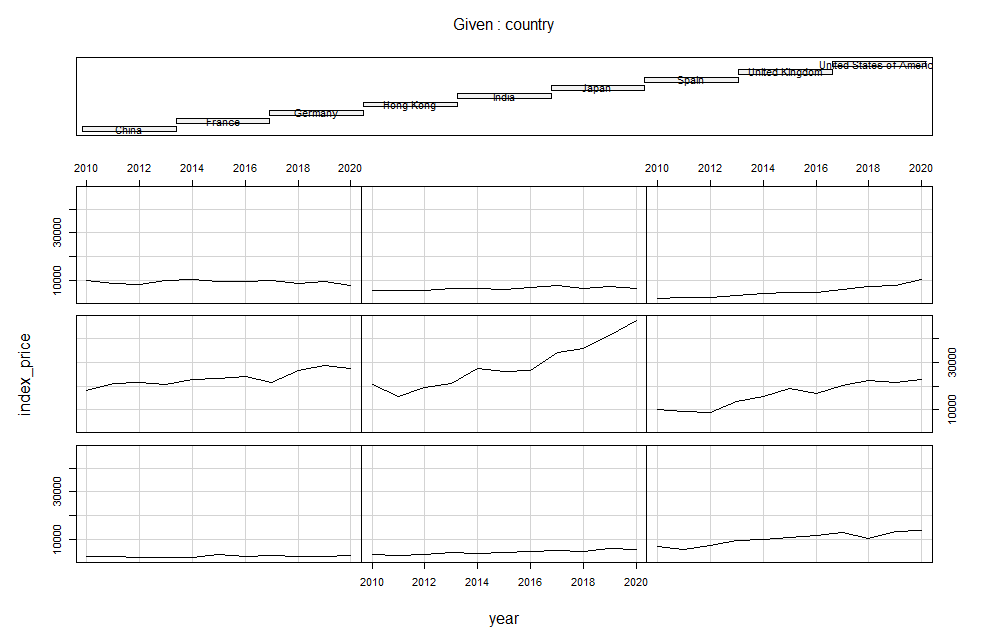


Figura 62. Corelația dintre indicele prețului bursier între spațiu și timp (sursa: RStudio)

Heterogeneitatea se referă la existența diferențelor între unitățile studiate, astfel că, în figura de mai jos avem diferențe destul de mari între țări, semnificând ipoteza, deci avem heterogeneitate transversală.

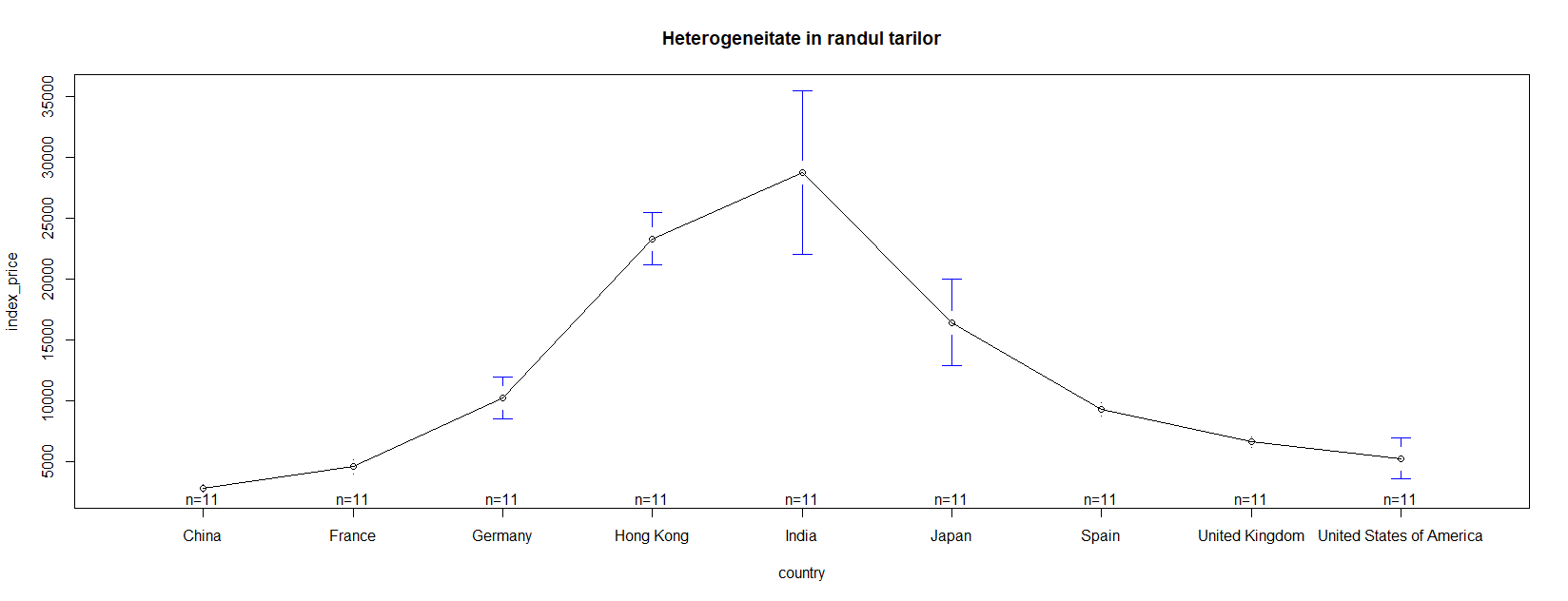


Figura 62. Heterogeneitatea pe țară și indecele prețului bursier (sursa: RStudio)

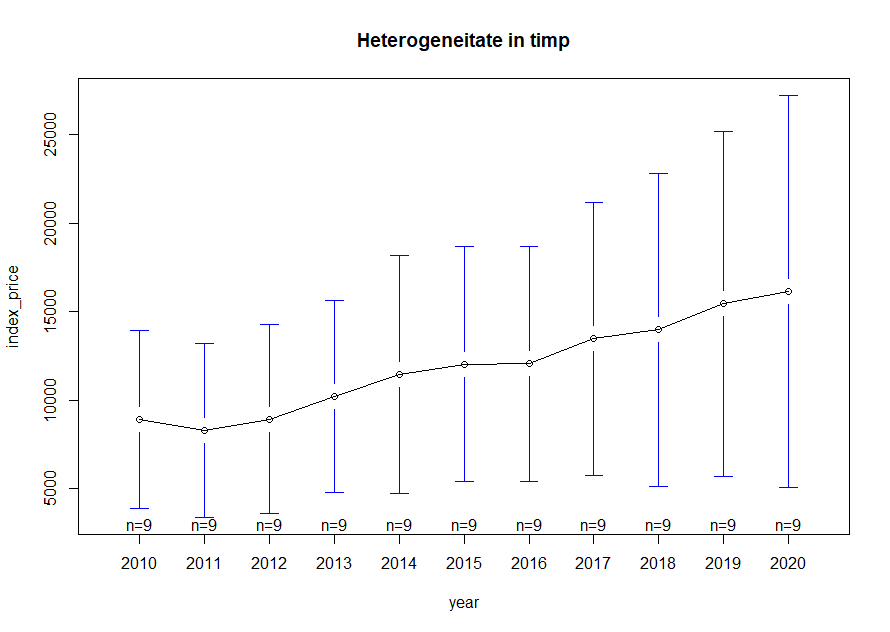


Figura 63. Heterogeneitatea pe an și indicele prețului bursier (sursa: RStudio)

În cazul figurii de mai sus, heterogeneitatea în timp nu afișează diferențe mari între țări, neexistând heterogeneitate în secțiunea transversală.

Modelul OLS nu se pretează a fi unul de actualitate dat fiind faptul că nu ia în calcul heterogeneitatea între spațiu și timp. Acesta explică 65.47%, cu un f-statistic de 34.51 și o eroare standard de 5621. Din acesta rezultă faptul că sunt semnificative în proporție de 99% percapitalincome, unemploymentrate și manufacturingoutput.

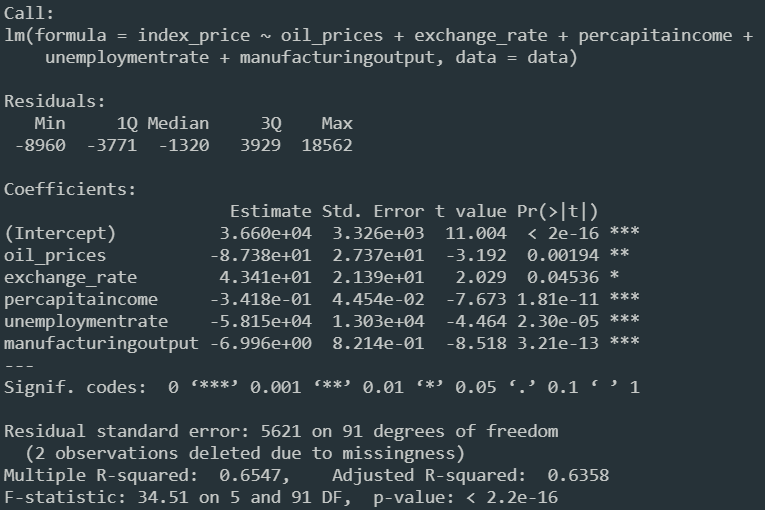


Figura 64. Model OLS, output (sursa: RStudio)

Modelul cu efecte fixe semnifică faptul că variabilele independente semnificative în proporție de 99% sunt exchange\_rate și percapitalincome cu un R-squared de 65.03% și f-statistic de doar 30.86.

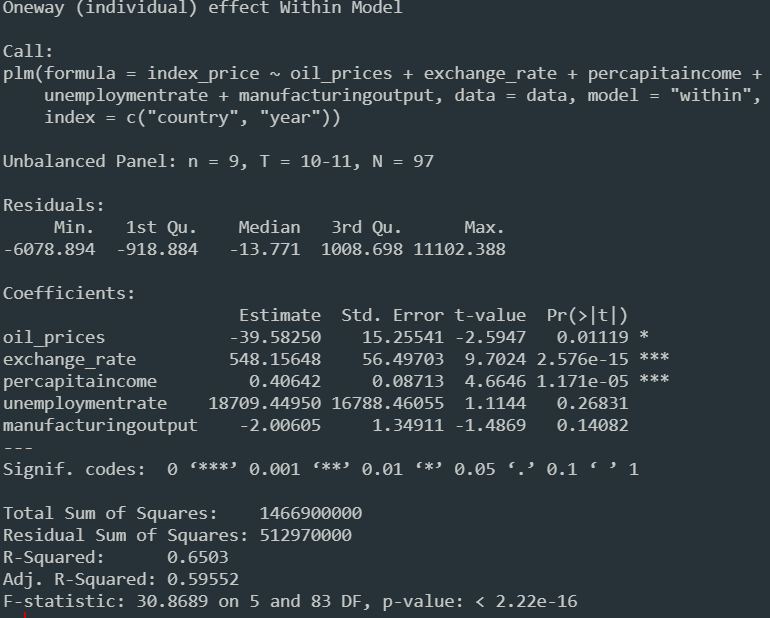


Figura 65. Model FE, output (sursa:RStudio)

Alegând variabilele cele mai semnificative, am creat modelul cu efecte fixe pentru a putea alege metoda cea mai adecvata a variantei de model:

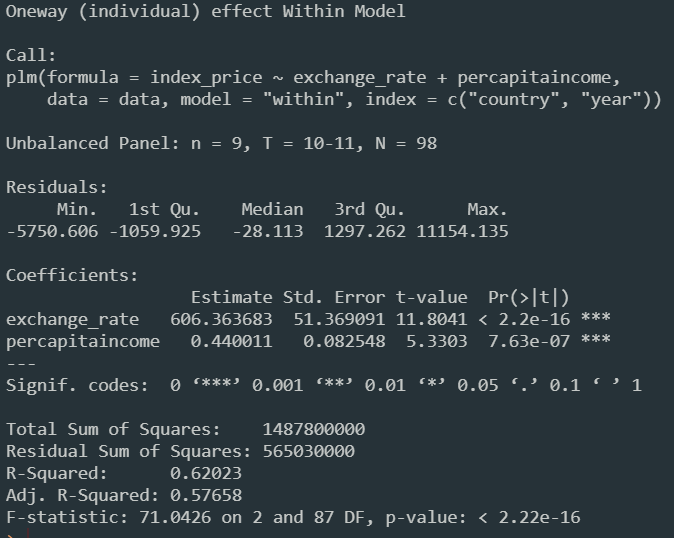


Figura 66. Model FE cu variabilele cele mai semnificative, output (sursa:RStudio)

Dat fiind faptul că testul F a generat un p-value mai mic decât 0.05, am continuat cu modelul de date de tip panel cu efecte fixe.

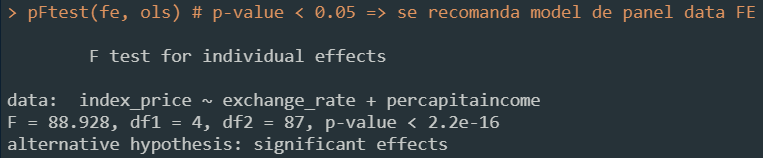


Figura 67. Test Fstatistic(sursa:RStudio)

În cadrul modelului cu efecte aleatorii, am obținut următoarele valori luând în calcul variabilele cele mai semnificative generate de către modelul cu efecte fixe. Modelul explică doar 34.94% cu un f-statistic de 1.61.

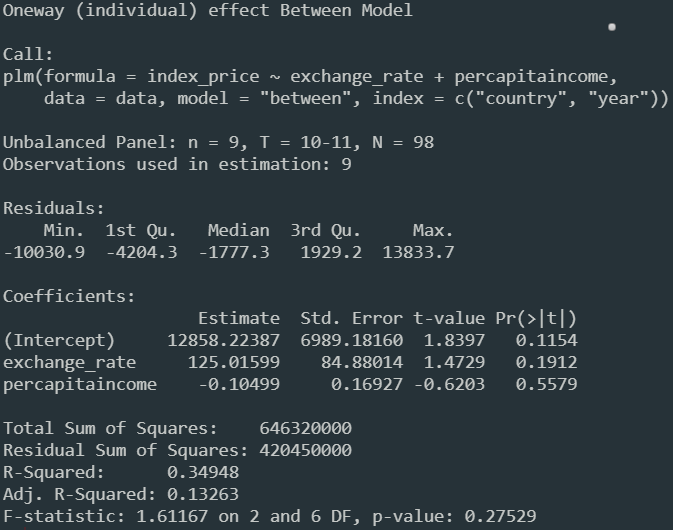


Figura 68. Model RE, output (sursa:RStudio)

Aplicând testul Hausman pentru determinarea deciziei în privința utilizării modelelor FE și RE am obținut valoarea lui p-value mai mică decât 0.05 rezultând faptul că vom continua cu modelul cu efecte fixe.

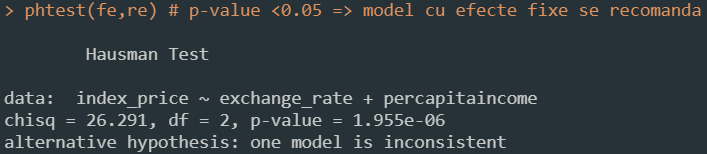


Figura 69. Test Hausman (sursa: RStudio)

Continuând cu testele Fstatistic și Lagrange Multiplier pentru decizia necesității în cadrul efectelor fixe în timp, am acceptat ipoteza nulă în cadrul testului Fstatistic, iar în cadrul celuilalt ipoteza alternativă. Dat fiind faptul că cele două nu prezintă același rezultat, vom alege varianta în care nu vom utiliza efectele fixe în timp.

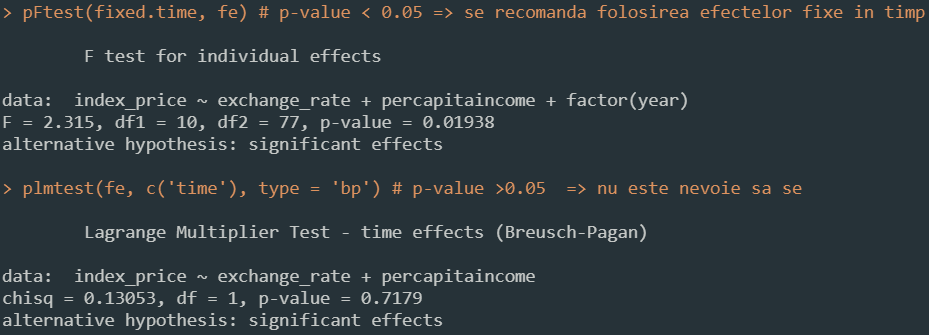


Figura 70. Testarea necesității efector fixe în timp (sursa: RStudio)

Am creat modelul de tip pool cu variabilele cele mai semnificative pentru a determina decizia utilizării a modelului cu efecte random și a celui OLS, prin testarea efectelor aleatorii cu Breusch-Pagan Lagrange Multiplier, ce a rezultat faptul că există diferențe semnificative între țările analizate.

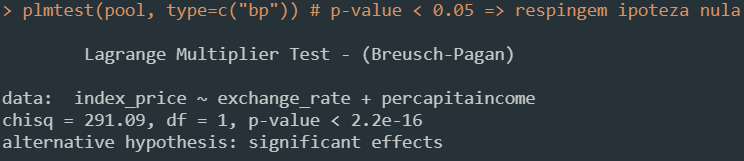


Figura 71. Test Breusch-Pagan Lagrange Multiplier (sursa: RStudio)

Folosind testele Parasan CD și Breusch-Pagan LM pentru testarea dependenței transversale am obținut următoarele rezultate referitoare la valoare lui p, respectiv 0.0002514 în cadrul primului test, rezultând faptul că reziduurile între entități nu sunt corelate, 0.5363 în cadrul celui de al doilea, adică avem independență transversală.

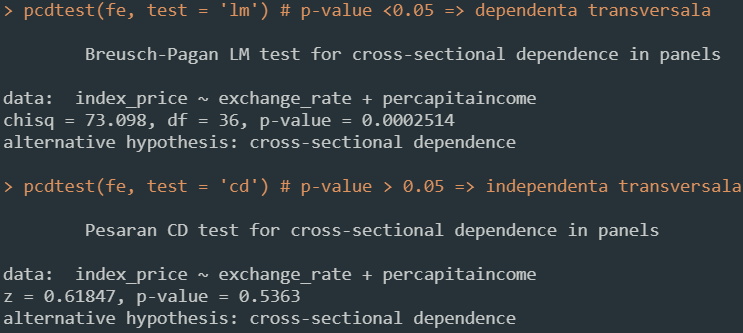


Figura 72. Teste Breusch-Pagan Lagrange Multiplier și Perasan CD (sursa: RStudio)

Continuând cu testarea autocorelării aplicând testul Breusch-Godfrey a rezultat faptul că trebuie să acceptăm ipoteza alternativă și modelul va trebui să fie corectat, însă setul de date este pe 10 ani.

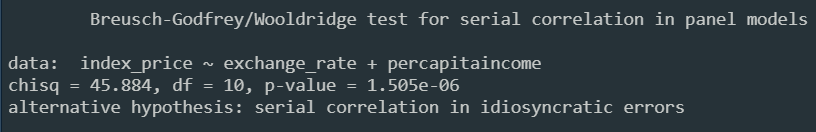


Figura 73. Test Breusch-Godfrey/Wooldridge (sursa: RStudio)

Pentru final am testat heteroschedasticitatea prin intermediul testului Breusch-Pagan rezultând faptul că am acceptat ipoteza alternativă deoarece p-value este mai mic decât 0.05, modelul necesitând corectarea.

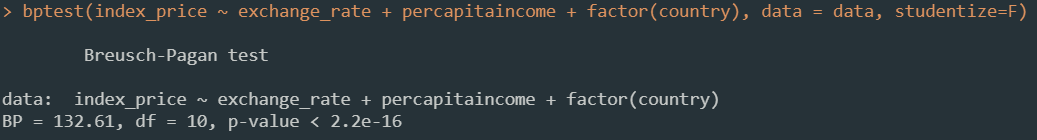


Figura 74. Test Breusch-Pagan (sursa: RStudio)

## **Concluzii**

În urma aplicării modelelor de regresie pe date de tip panel am analizat factorii contribuabili în vederea indicelui prețului bursier în cadrul celor 9 țări din setul de date ales pe perioada anilor 2010-2020.

Explorând heterogeneitatea în secțiunea transversală am obținut faptul că avem diferențe între medii destul de mari, rezultă heterogeneitate, deci avem heterogeneitate transversală. Analizând heterogeneitatea în timp, am obținut că aceasta nu se confirmă prin faptul că nu există diferențe destul de mari între medii raportat la an și indicele prețului bursier.

Creeând modelul OLS luând in calcul toate variabilele independente a rezultat că doar prețul petrolului și rata de schimb sunt nesemnificative, dar acest model nu se pretează pentru a fi unul de actualitate deoarece nu ia în calcul heterogeneitatea între spațiu și timp.

Aplicând modelul FE cu aceleași variabile au reieșit ca variabile semnificative doar rata de schimb și venitul per cap de. Valoarea generată de către testul Fstatistic ne recomandă să folosim un model cu efecte fixe deoarece valoarea acesteia este sub 0.05. Pentru determinarea deciziei utilizării între modelele cu efecte fixe si cele aleatorii am aplicat testul Hausman și a rezultat faptul că a trebuit să utilizăm modelul cu efecte fixe.

În continuare pentru a determina necesitatea efectelor de timp am aplicat testele Fstatistic și de multiplicare Lagrange ce au generat contrariul între acestea, astfel că am mers pe varianta în care nu folosim efecte fixe în timp. Pentru alegerea utilizării dintre modelul cu efecte random și cel OLS, am creat modelul de tip pool, iar prin aplicarea testelor aferente a reieșit faptul că există diferențe semnificative între țările analizate. Nu în cele din urmă am testat dependența transversală prin intermediul testelor Parasan CD și Breusch-Pagan LM din care valoarea lui p a fost diferită, astfel încât în cadrul Parasan CD nu a necesitat corectare. În cadrul ipotezei pentru homoschedasticitate, rezultatele au necesitat corectarea.

## **Bibliografie**

Setul de date > <https://www.kaggle.com/datasets/pratik453609/economic-data-9-countries-19802020>

Indicele prețului bursier > <https://tradeville.ro/tradepedia/indici>

Rata de schimb > <https://intreb.bancatransilvania.ro/ce-este-cursul-valutar/>

Venit pe cap de locuitor > [http://www.stiri-economice.ro/ce-este-pib-ul-pe-cap-de-locuitor.html#:~:text=Conceptul%20de%20PIB%20pe%20cap%20de%20locuitor&text=Acesta%20se%20calculeaz%C4%83%20fie%20prin,%C3%AEn%20%C8%9Bar%C4%83%20pe%20parcursul%20anului.](http://www.stiri-economice.ro/ce-este-pib-ul-pe-cap-de-locuitor.html%23:~:text=Conceptul%20de%20PIB%20pe%20cap%20de%20locuitor&text=Acesta%20se%20calculeaz%C4%83%20fie%20prin,%C3%AEn%20%C8%9Bar%C4%83%20pe%20parcursul%20anului.)

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)