Sveučilište Josipa Jurja Strossmayera u Osijeku
Odjel za matematiku
Diplomski studij - Financijska matematika i statistika

Regresijski model prosječnog rezultata škole na standardiziranom testu

Seminarski rad

Student: Lara Juzbašić

Kolegij: Multivarijatna analiza

Osijek, 2021.

Sadržaj

1 Uvod

Zbog nejednake kvalitete edukacije u školama, standardizirani testovi korisni su kako bi školama dali informaciju o uspjehu vlastitih učenika u odnosu na uspjeh učenika drugih škola. U ovom radu korišteni su podaci prikupljeni iz 420 osnovnih škola u Kaliforniji, koje su sudjelovale u provođenju standardiziranog testa Stanford 9. Test je proveden među učenicima petih razreda, a uključuje pitanja s višestrukim izborom za testiranje engleskog jezika, matematike, čitanja i znanosti. U ovom radu usmjerit ćemo pažnju na prosječan rezultat postignut iz čitanja, te izgraditi linearan regresijski model koji ga dobro opisuje. Opis varijabli korištenih u izgradnji modela nalazi se u Tablici 1.

Ime varijable	**Opis varijable**
students (Učenici)	Broj učenika u školi
teachers (Učitelji)	Broj učitelja u školi
lunch (Školski obrok)	Postotak učenika koji ispunjavaju uvjete za školski obrok po sniženoj cijeni
computer (Računala)	Broj računala po učionici
expenditure (Izdaci)	Izdaci po učeniku
income (Prihodi)	Prosječni prihodi u okrugu u kojemu se se škola nalazi
english (Engleski)	Postotak učenika koji uče engleski, tj. kojima engleski nije materinji jezik
read (Čitanje)	Prosječan rezultat učenika postignut iz čitanja

Tablica 1: Opis varijabli

Umjesto varijable *Učitelji*, promatrat ćemo varijablu dobivenu kao omjer između broja učenika i broja učitelja, koju ćemo nazvati *Omjer*.

2 Analiza podataka

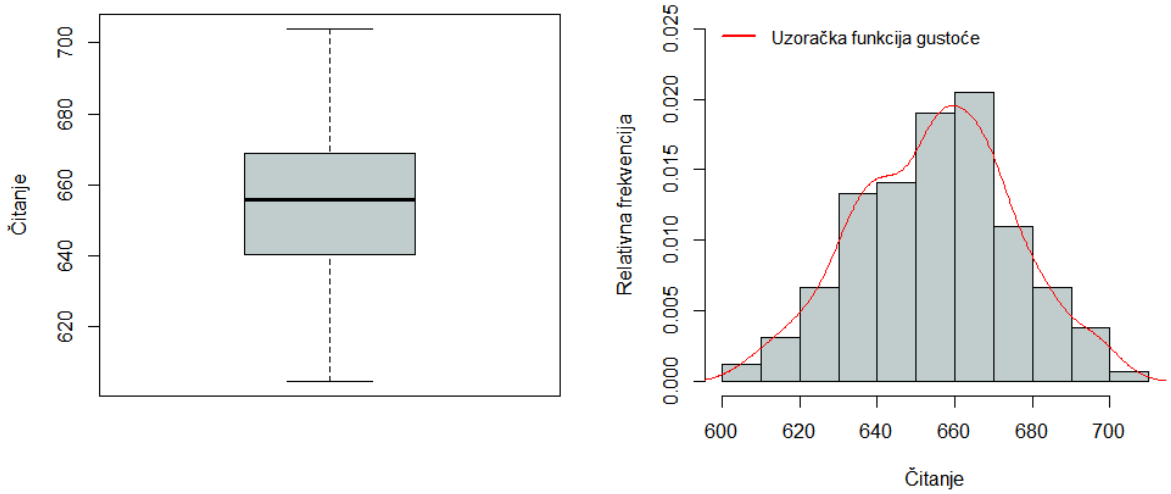
Provest ćemo deskriptivnu statistiku ovisne varijable i neovisnih varijabli za koje će se pokazati kako su najviše korelirane s ovisnom varijablom, a to su Školski obrok, Prihodi i Engleski, te ćemo analizirati njihov odnos s ovisnom varijablom.

2.1 Analiza varijable *Čitanje*

Čitanje je varijabla numeričkog tipa koju želimo modelirati pomoću preostalih varijabli u bazi. Predstavlja prosječan rezultat učenika postignut u pojedinoj školi iz čitanja, na standardiziranom testu Stanford 9. U Tablici 2 nalazi se deskriptivna statistika ove varijable te na Slici 1 pripadni grafički prikazi.

Minimum	**Donji kvartil**	**Medijan**	**Prosjek**	**Gornji kvartil**	**Maksimum**	**St. dev.**
604.5	640.4	655.8	655.0	668.7	704.0	20.11

Tablica 2: Deskriptivna statistika varijable Čitanje



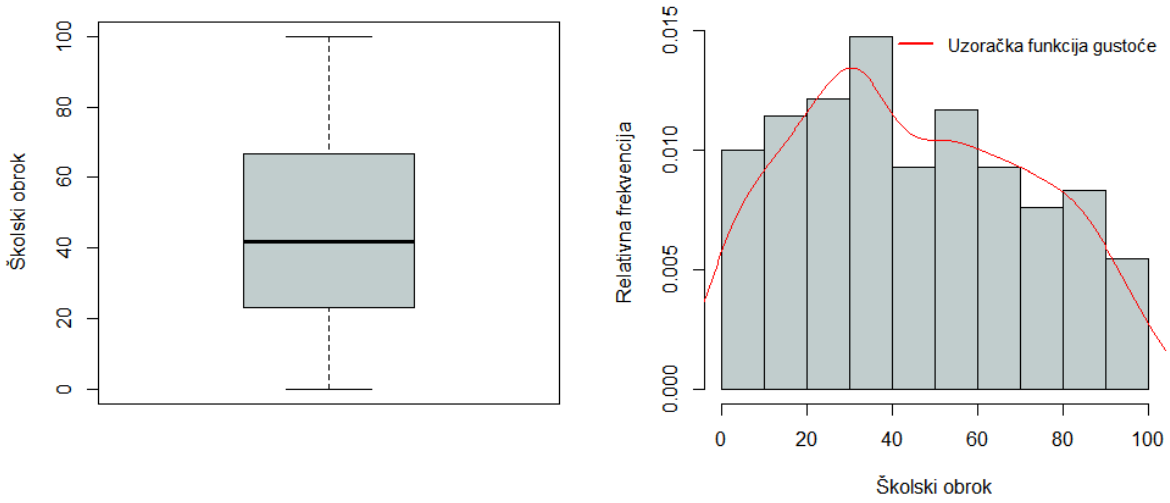
Slika 1: Histogram i kutijasti dijagram varijable Čitanje

2.2 Analiza varijable *Školski obrok*

Školski obrok je numerička varijabla koja sadrži informacije o postotku učenika u školi koji ispunjavaju uvjete za školski obrok po sniženoj cijeni. U Tablici 3 nalazi se deskriptivna statistika ove varijable, a na Slici 2 prikazani su grafički prikazi varijable.

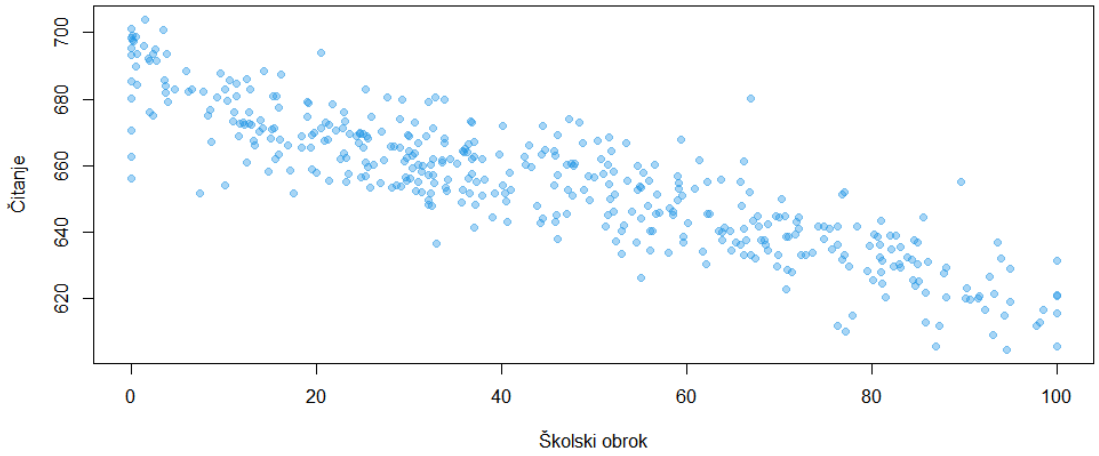
Minimum	**Donji kvartil**	**Medijan**	**Prosjek**	**Gornji kvartil**	**Maksimum**	**St. dev.**
0.00	23.28	41.75	44.71	66.86	100.00	27.12

Tablica 3: Deskriptivna statistika varijable Školski obrok



Slika 2: Histogram i kutijasti dijagram varijable Školski obrok

Na Slici 3 može se vidjeti dijagram raspršenosti varijable *Čitanje* u odnosu na varijablu *Školski obrok*. Iz dijagrama raspršenosti možemo primijetiti kako povećanjem postotka učenika koji imaju pravo na sniženi školski obrok dolazi do smanjenja prosječnog rezultata iz čitanja postignutog na testu. Provođenjem Pearsonovog korelacijskog testa dobivamo izrazito malu p-vrijednost ($p < 2.2e-16$), stoga na razini značajnosti 0.05 možemo tvrditi kako postoji korelacija između navedenih varijabli. Nadalje, provođenjem Kendallovog korelacijskog testa, također dobivamo izrazito malu p-vrijednost ($p < 2.2e-16$), te na razini značajnosti 0.05 možemo tvrditi kako je veza između varijabli monotona. Po vrijednosti koeficijenta korelacije (-0.8788077) veza je padajuća.



Slika 3: Dijagram raspršenosti varijable Čitanje u odnosu na varijablu Školski obrok

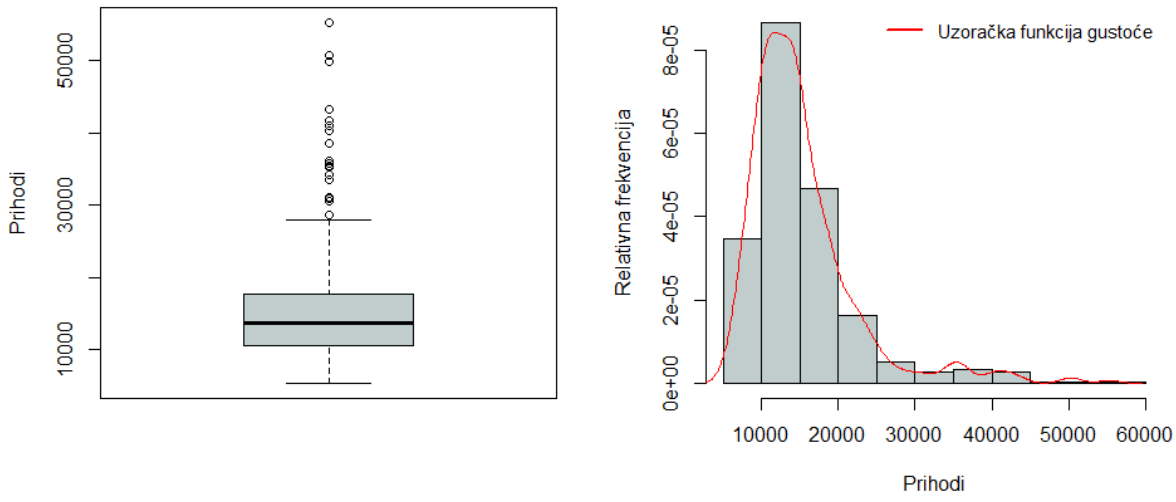
2.3 Analiza varijable *Prihodi*

Varijabla *Prihodi* je numeričkog tipa te predstavlja prosječne prihode okruga u kojem se nalazi pojedina škola. Deskriptivna statistika ove varijable nalazi se u Tablici 4, a pripadni grafički prikazi mogu se vidjeti na Slici 4.

Minimum	**Donji kvartil**	**Medijan**	**Prosjek**	**Gornji kvartil**	**Maksimum**	**St. dev.**
5335	10639	13728	15317	17629	55328	7225.89

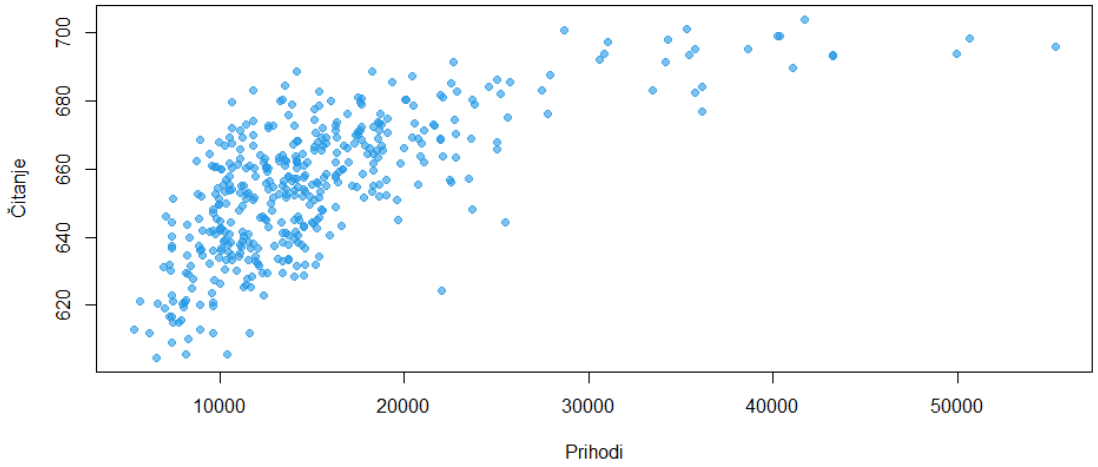
Tablica 4: Deskriptivna statistika varijable Prihodi

Kako bismo provjerili koreliranost varijabli *Prihodi* i *Čitanje*, provodimo Pearsonov korelacijski test. Dobivamo izrazito malu p-vrijednost ($p < 2.2e-16$), te na razini značajnosti 0.05 možemo tvrditi kako postoji korelacija između navedenih varijabli. Zatim, provođenjem Kendallovog korelacijskog testa također dobivamo izrazito malu p-vrijednost ($p < 2.2e-16$), te na razini značajnosti 0.05 možemo tvrditi kako je veza između varijabli monotona. Po vrijednosti koeficijenta korelacije (0.6978189), korelacija je pozitivna.



Slika 4: Histogram i kutijasti dijagram varijable Prihodi

Dijagram raspršenosti na Slici 5 upućuje na nelinearnu vezu između varijabli *Prihodi* i *Čitanje*.



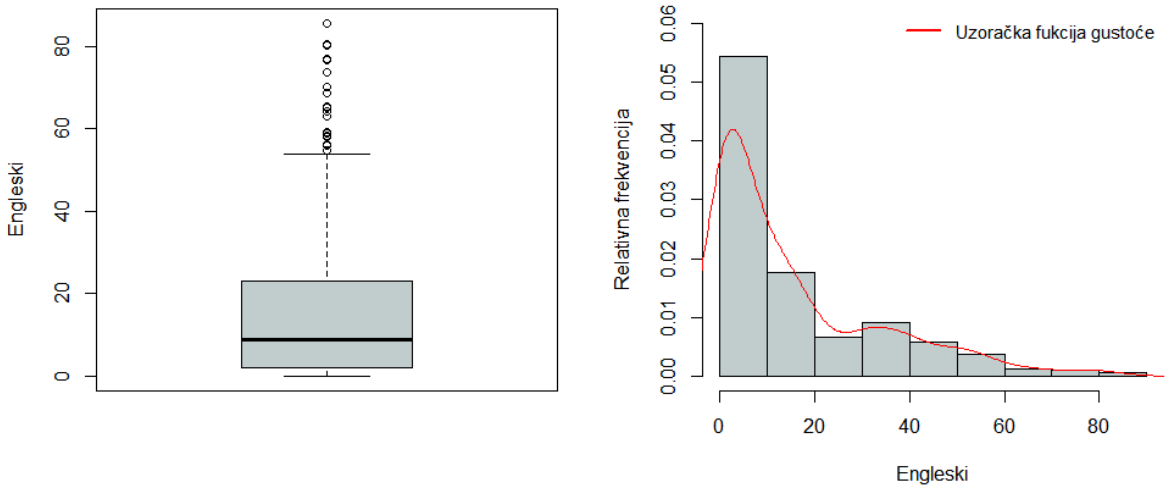
Slika 5: Dijagram raspršenosti varijable Čitanje u odnosu na varijablu Prihodi

2.4 Analiza varijable *Engleski*

Varijabla *Engleski* numerička je varijabla koja sadrži informacije o postotku učenika u školi koji uče engleski jezik, tj. kojima engleski nije materinji jezik. U Tablici 5 nalazi se deskriptivna statistika ove varijable te na Slici 6 njeni grafički prikazi.

Minimum	**Donji kvartil**	**Medijan**	**Prosjek**	**Gornji kvartil**	**Maksimum**	**St. dev.**
0.000	1.941	8.778	15.768	22.970	85.540	18.29

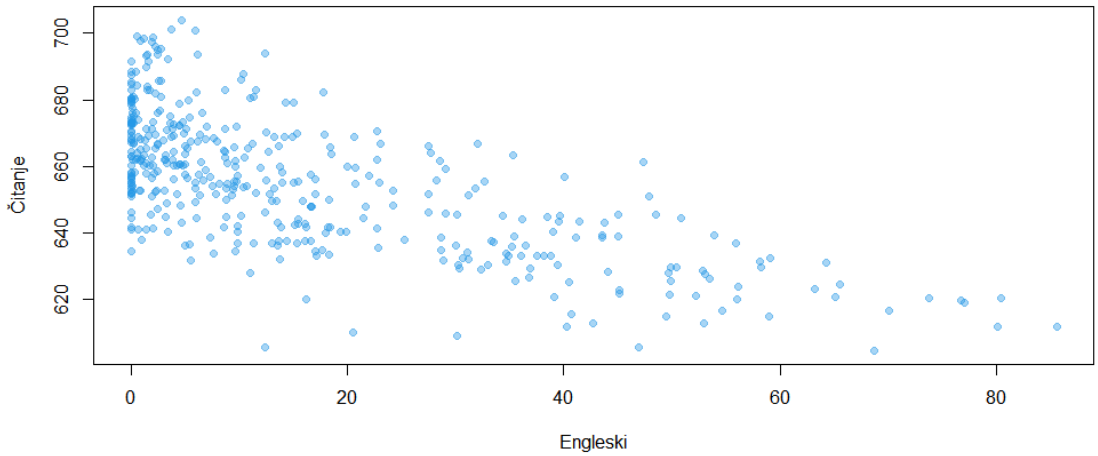
Tablica 5: Deskriptivna statistika varijable Engleski



Slika 6: Histogram i kutijasti dijagram varijable Engleski

Iz dijagrama raspršenosti na Slici 7 vidljivo kako povećanjem postotka učenika koji uče

engleski jezik dolazi do smanjenja prosječnog rezultata škole na testu iz čitanja. Provjerimo jesu li navedene varijable korelirane. U tu svrhu provodimo Pearsonov korelacijski test i dobivamo p-vrijednost koja je izrazito mala ($p < 2.2e-16$), te na razini značajnosti 0.05 možemo tvrditi kako postoji korelacija između navedenih varijabli. Kako bismo provjerili monotonost veze ovih varijabli provodimo Kendallov korelacijski test, također dobivamo izrazito malu p-vrijednost ($p < 2.2e-16$), te na razini značajnosti 0.05 možemo tvrditi kako je veza između varijabli monotona. Po vrijednosti koeficijenta korelacije (-0.6902859) veza je padajuća.



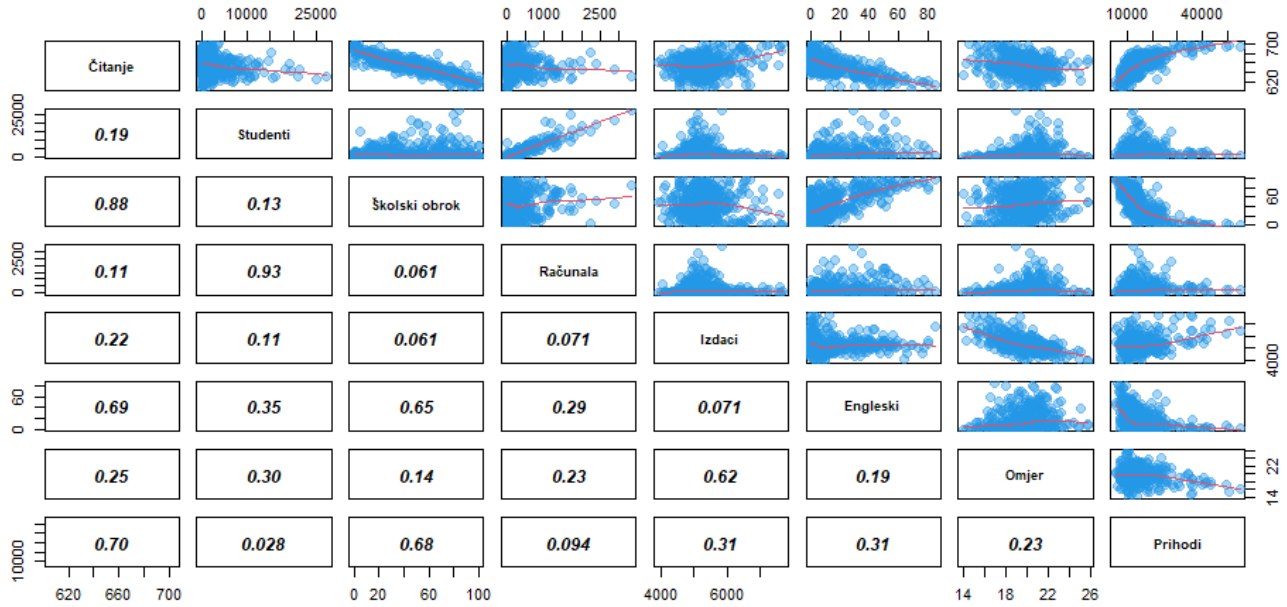
Slika 7: Dijagram raspršenosti varijable Čitanje u odnosu na varijablu Engleski

3 Postupak izgradnje modela

Kao što je najavljeno u uvodu, zadatak je pronaći regresijski model koji dovoljno dobro opisuje varijablu *Čitanje*. Osim što dobro opisuje ovisnu varijablu, takav model treba zadovoljavati teorijske pretpostavke regresijskog modela. Osim samog odabira modela i njegove interpretacije, poglavlje sadrži i provjeru postojanja stršećih vrijednosti u uzorku te utjecajnih mjerenja.

3.1 Selekcija modela

Postupak selekcije modela započinjemo time što provjeravamo matricu korelacija te matrični dijagram raspršenosti između svih varijabli, prikazane na Slici 8.



Slika 8: Dijagrami raspršenosti i koeficijenti korelacije među varijablama

Kao što je već spomenuto, veza između ovisne varijable *Čitanje* i varijable *Prihodi* ne izgleda linearno. Dijagram raspršenosti ukazuje na to da je veza logaritamska, stoga će u daljnjoj analizi umjesto varijable *Prihodi* biti promatran njezin prirodni logaritam. Uključimo li sve varijable kao prediktore u model, dobijemo sljedeću tablicu.

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	5.689e+02	1.898e+01	29.983	$< 2e-16$
students	-4.516e-04	2.934e-04	-1.539	0.124513
lunch	-4.252e-01	2.990e-02	-14.221	$< 2e-16$
computer	2.609e-03	2.519e-03	1.036	0.300891
expenditure	3.309e-03	8.512e-04	3.888	0.000118
english	-2.383e-01	3.170e-02	-7.515	3.57e-13
omjer	-1.511e-01	2.866e-01	-0.527	0.598369
lincome	9.902e+00	1.796e+00	5.512	6.27e-08

Tablica 6: Puni model

Vidimo kako nisu svi prediktori statistički značajni pa provedimo selekcijske procedure koje će nam pomoći u odabiru varijabli. Prva procedura koju provodimo je leaps koja bira podskup varijabli čijim bi uključivanjem dobili model koji ima najveći prilagođeni R^2. Ovom procedurom sugeriran je model koji se dobije isključenjem prediktora *Omjer* iz punog modela. No, u tom modelu javlja se problem multikolinearnosti, čiji je pokazatelj faktor inflacije varijance (VIF) dan u Tablici 7.

Studenti	Školski obrok	Računala	Izdaci	Engleski	ln(Prihodi)
7.871232	4.039442	7.551210	1.174776	2.061177	3.051504

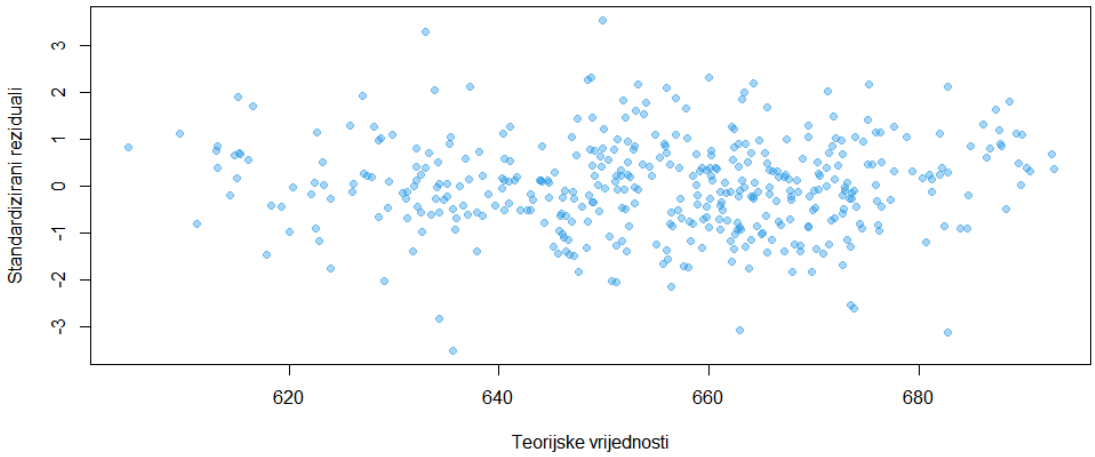
Tablica 7: VIF

Pripadne VIF vrijednosti varijabli *Studenti* i Računala veće su od 5, što ukazuje na jaku korelaciju među prediktorima, te pomoću naredbe linearHypothesis potvrđujemo da smijemo isključiti ove varijable iz modela (p=0.1418). Dakle, u modelu ostaju varijable *Školski obrok*, *Izdaci*, *Engleski* i *ln(Prihodi)*.

Zatim provodimo step funkciju na punom modelu, koja bira podskup varijabli čijim bi uključivanjem dobili model koji ima najniži Akaike informacijski kriterij (AIC). U našem slučaju to su varijable *Školski obrok*, *Izdaci*, *Engleski* i *ln(Prihodi)* i *Studenti*, ali posljednji prediktor se ne pokazuje statistički značajan pa njegovim izbacivanjem dobivamo model jednak konačnom modelu koji smo dobili vodeći se leaps funkcijom. Međutim, provjerom normalnosti standardiziranih reziduala pomoću Shapiro-Wilk testa dobivamo p-vrijednost manju od 0.05 (p=0.003032), te na razini značajnosti odbacujemo hipotezu o normalnoj distribuiranosti reziduala.
Sada želimo pronaći podskup ovih prediktora koji zadovoljava sve pretpostavke regresijskog modela. Iz matrice korelacija vidjeli smo kako među ovim prediktorima, najmanji koeficijent korelacije s ovisnom varijablom ima varijabla *Izdaci*. Odlučujemo se izbaciti iz trenutnog modela varijablu *Izdaci* te dobivamo sljedeći model:
Čitanje = 556.63325−0.39069∗Školski obrok−0.27666∗Engleski+12.57936∗ln(Prihodi)
Prilagođeni R^2 ovog modela iznosi 0.8184, odnosno 81.84% ukupne varijabilnosti u podacima objašnjeno je modelom.

3.2 Pretpostavke modela

Provjerimo prvo homoskedastičnost grešaka modela, pri čemu kao procjenitelje grešaka modela promatramo standardizirane reziduale. Dijagram raspršenosti na Slici 9 ukazuje na homoskedastičnost jer nije vidljiva nikakva pravilnost. Kako bismo bili sigurni provodimo statističke testove, čije su p-vrijednosti prikazane u Tablici 8. Sve p-vrijednosti su veće od 0.05 pa nemamo razloga sumnjati u homoskedastičnost grešaka modela.



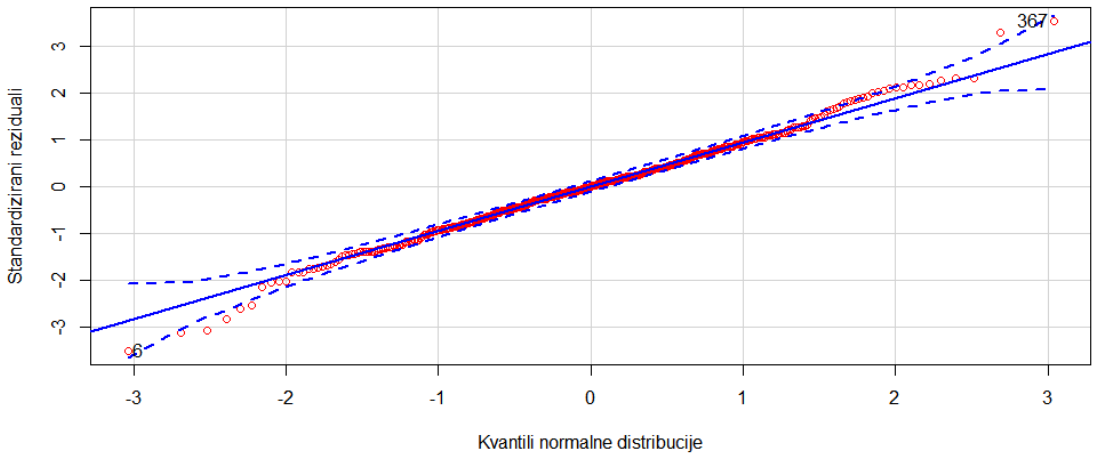
Slika 9: Dijagram raspršenosti teorijskih vrijednosti i standardiziranih reziduala

Test	**p-vrijednost**
ncvTest	0.82896
gqtest	0.3676
bptest	0.2196

Tablica 8: Testiranje homoskedastičnosti

Provjerimo sada normalnu distribuiranost standardiziranih reziduala. U tu svrhu promotrimo QQ-graf na Slici 10. Na grafičkom prikazu uočavamo mala odstupanja od normalne distribucije, međutim provedbom Shapiro-Wilk testa dobivamo p-vrijednost 0.06222 pa na razini značajnosti 0.05 nemamo razloga sumnjati u normalnost grešaka modela.

Posljednja pretpostavka koju je potrebno provjeriti je multikolinearnost. U Tablici 9 dan je VIF, a kako su sve vrijednosti manje od 5, zaključujemo da nemamo problema s multikolinearnosti.



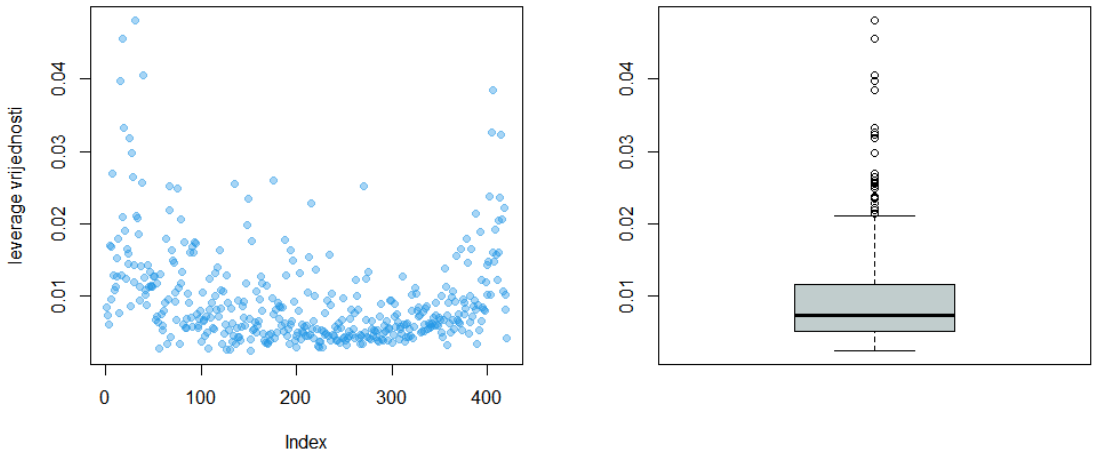
Slika 10: QQ-graf

Školski obrok	Engleski	ln(Prihodi)
3.769789	1.843481	2.538540

Tablica 9: VIF

3.3 Stršeće vrijednosti i utjecajna mjerenja

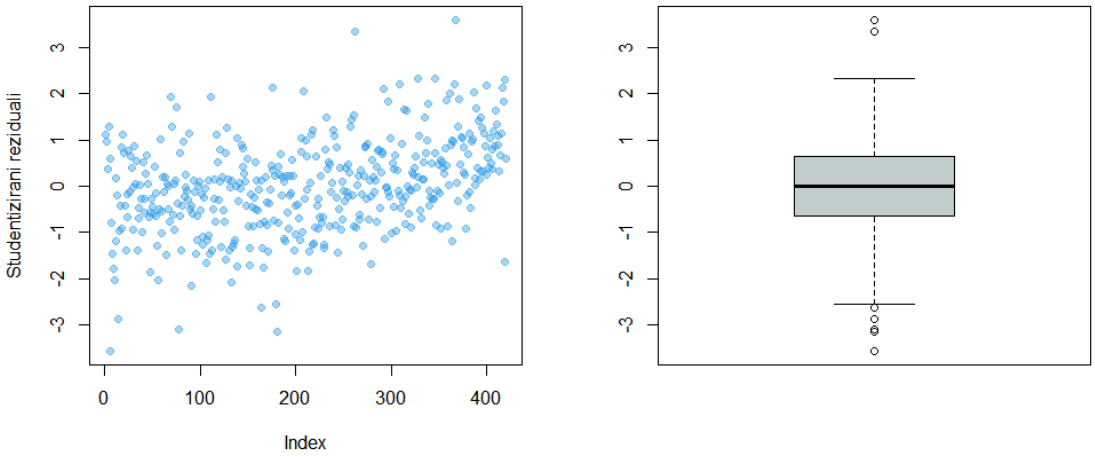
Stršeće vrijednosti su ona mjerenja čija vrijednost značajno odstupa od prosjeka, a utjecajna su mjerenja ako znatno zakreću regresijsku funkciju te utječu na rezultate modela. Na slici 11 nalaze se grafički prikazi leverage score-a.



Slika 11: Dijagram raspršenosti i kutijasti dijagram leverage score-a

Stršećim vrijednostima smatramo one čija je vrijednost leverage score-a veća od $\frac{2k}{n}$. Pronađeno je 33 takva podatka za koje ne možemo očekivati dobre rezultate primjenom

ovog modela. Radi lociranja velikih grešaka predikcije potrebno je provjeriti stršeće vrijednosti u rezidualima. Na Slici 12 nalaze se grafički prikazi studentiziranih reziduala.

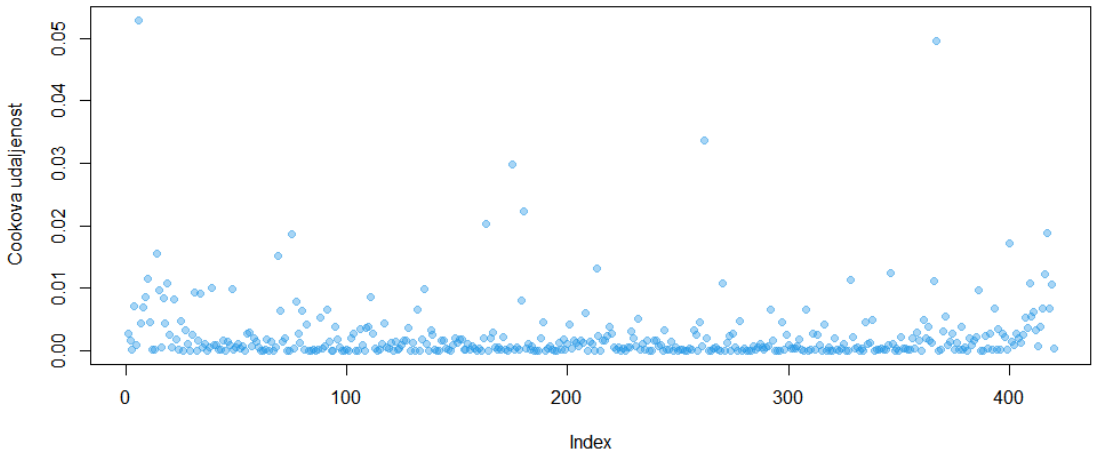


Slika 12: Dijagram raspršenosti i kutijasti dijagram studentiziranih reziduala

Iz kutijastog dijagrama vidljivo je nekoliko stršećih vrijednosti, a najveća iznosi 3.600412 i pripada podatku 367. Utjecajna mjerenja detektiraju se pomoću Cookove udaljenosti koja mjeri kako se cijela regresijska funkcija mijenja ukoliko izbacimo i-tu opservaciju. Promatrajući Cookovu udaljenost, kao najutjecajnija pokazala se opservacija 6. Kako ne sumnjamo u vjerodostojnost podataka, ovu opservaciju nećemo izbaciti.

Učenici	Školski obrok	Računala	Izdaci	Prihodi	Engleski	Čitanje	Omjer
137	86.9565	25	5580.147	10415	12.40876	605.7	21.40625

Tablica 10: Tablica 8: Najutjecajnije mjerenje



Slika 13: Grafički prikaz Cookove udaljenosti

3.4 Interpretacija modela

Naposljetku, interpretirajmo konačni model koji je oblika:
Čitanje = 556.63325−0.39069∗Školski obrok−0.27666∗Engleski+12.57936∗ln(Prihodi)

- Slobodan član iznosi 556.63325, a on predstavlja procijenjenu vrijednost ovisne varijable ukoliko su vrijednosti svih neovisnih varijabli jednake nuli (s tim da je pouzdani interval [522.9652815, 590.3012143]).

- Jedinično povećanje postotka učenika koji imaju pravo na školski obrok po sniženoj cijeni prosječno se reflektira smanjenjem prosječnog rezultata postignutog na testu za 0.39069, uz ostale nepromijenjene varijable (s tim da je pouzdani interval [-0.4496038, -0.3317790]).

- Jedinično povećanje postotka učenika koji uče engleski jezik prosječno se reflektira smanjenjem prosječnog rezultata postignutog na testu za 0.27666, uz ostale nepromijenjene varijable (s tim da je pouzdani interval [-0.3377654, -0.2155504]).

- Jedinično povećanje prirodnog logaritma prosječnih prihoda okruga u kojemu se škola nalazi, prosječno se reflektira povećanjem prosječnog rezultata postignutog na testu za 12.57936, uz ostale nepromijenjene varijable (s tim da je pouzdani interval [9.2375274, 15.9211961]).

4 Zaključak

Analizom podataka prikupljenih iz 420 osnovnih škola, izgrađen je model koji dobro opisuje prosječan rezultat na standardiziranom testu Stanford 9 iz dijela čitanje. Pri odabiru modela korištene su selekcijske procedure, te su zadovoljene sve pretpostavke linearnog regresijskog modela. Prilagođeni R^2 modela iznosi 0.8184 te ga zbog toga smatramo prediktivnim.