Prediktivni modeli bazirani na metodi logističke regresije

Štimac, Karlo

SVEUČILIŠTE U RIJECI

TEHNIČKI FAKULTET

Prijediplomski sveučilišni studij elektrotehnike

Završni rad

PREDIKTIVNI MODELI BAZIRANI NA METODI LOGISTIČKE REGRESIJE

Rijeka, rujan 2023.

Karlo Štimac
0069086012

SVEUČILIŠTE U RIJECI

TEHNIČKI FAKULTET

Prijediplomski sveučilišni studij elektrotehnike

Završni rad

PREDIKTIVNI MODELI BAZIRANI NA METODI LOGISTIČKE REGRESIJE

Mentor: izv. prof. dr. sc. Ivan Dražić

Komentor: doc. dr. sc. Angela Bašić-Šiško

Rijeka, rujan 2023.

Karlo Štimac
0069086012

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
POVJERENSTVO ZA ZAVRŠNE ISPITE

Rijeka, 13. ožujka 2023.

Zavod: **Zavod za matematiku, fiziku i strane jezike**
Predmet: **Inženjerska matematika ET**
Grana: **1.01.07 primijenjena matematika i matematičko modeliranje**

ZADATAK ZA ZAVRŠNI RAD

Pristupnik: **Karlo Štimac (0069086012)**
Studij: Sveučilišni prijediplomski studij elektrotehnike

Zadatak: **Prediktivni modeli bazirani na metodi logističke regresije**

Opis zadatka:

U radu je potrebno objasniti metodu regresije s posebnim osvrtom na logističku regresiju. Potrebno se osvrnuti na parametre kvalitete regresijskog modela, odnosno statističke značajnosti pojedinog prediktora. Metodu logističke regresije potrebno je staviti u kontekst primjene kod klasifikatorskih prediktivnih modela.
U praktičnom dijelu rada logističku regresiju treba primijeniti na izradu klasifikatora koristeći se realnim podacima, analizirati kvalitetu regresije i kvalitetu prediktivnog modela temeljem tipičnih pokazatelja kvalitete dijagnostičkog testa.

Rad mora biti napisan prema Uputama za pisanje diplomskih / završnih radova koje su objavljene na mrežnim stranicama studija.

Karlo Štimac

Zadatak uručen pristupniku: 20. ožujka 2023.

Mentor:

Izv. prof. dr. sc. Ivan Dražić

Doc. dr. sc. Angela Bašić-Šiško (komentor)

Predsjednik povjerenstva za završni ispit:

Prof. dr. sc. Dubravko Franković

IZJAVA

Sukladno članku 7. stavku 1. Pravilnika o završnom radu, završnom ispitu i završetku sveučilišnih prijediplomskih studija Tehničkog fakulteta Sveučilišta u Rijeci od 4. travnja 2023., izjavljujem da sam samostalno izradio završni rad prema zadatku preuzetom dana 20. ožujka 2023.

Rijeka, 11. rujna 2023.

Karlo Štimac

Karlo Štimac

Najprije, želim se zahvaliti svom mentoru, izv. prof. dr. sc. Ivanu Dražiću, na strpljenju, vođenju i podršci koju mi je pružio tijekom izrade ovog završnog rada. Na empatiji, razumijevanju i suosjećanju sa problemima i preprekama pred kojima sam se nalazio. Želim se zahvaliti komentorici, doc. dr. sc. Angeli Bašić-Šiško na ukazanoj pomoći, vaši stručni savjeti i konstruktivna povratna informacija bili su ključni za poboljšanje kvalitete ovog rada. Veliku zahvalu dugujem cijeloj svojoj užoj obitelji, naročito roditeljima i bratu koji su uz mene bili u dobrim i manje dobrim trenutcima i što nisu nikad prestali vjerovati u mene. I ono najbitnije, hvala dragom Bogu jer bez njega ništa od ovoga ne bi bilo izvedivo.
Naposljetku, želim se zahvaliti svima koji su bili uz mene kroz ovaj period života.

Sadržaj

1. Uvod

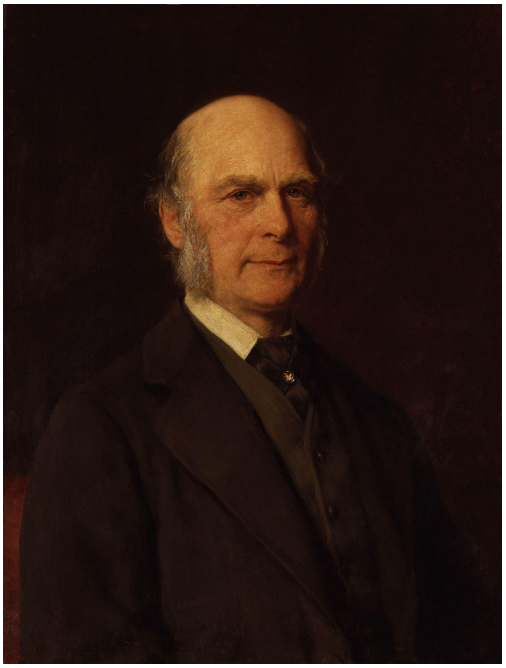
Regresija ili regresijska analiza je statistička metoda, odnosno skup metoda koje se koriste za identifikaciju odnosa između zavisne varijable te jedne ili više nezavisnih varijabli, koje često nazivamo prediktorima. Drugim riječima, smisao regresijske analize je formiranje matematičkog modela kojim se na temelju vrijednosti nezavisnih varijabli mogu predvidjeti vrijednosti zavisne varijable. Koristi se u različitim područjima, uključujući ekonomiju, psihologiju, financije, medicinu, itd. S matematičke strane metoda je kompleksna te zahtijeva pažljivo tumačenje i razmatranje pretpostavki, ali je istovremeno odličan alat za analizu podataka i predviđanja. [18]

Regresijska analiza ima svoje korijene u astronomiji. Još u 18. stoljeću, astronomi su pokušavali pronaći matematičke modele koji bi predvidjeli kretanje nebeskih tijela. Jedan od najpopularnijih primjera je rad njemačkog matematičara i astronoma Johanna Karla Friedricha Gaussa[1] koji je koristio regresiju kako bi predvidio kretanje Ceresa, jednog od najvećih asteroida, te ujedno prvog otkrivenog asteroida koji je lociran između Jupitera i Marsa. Također, Gauss je zaslužan za razvoj metode najmanjih kvadrata, koja je temelj regresijske analize. Metoda najmanjih kvadrata po prvi put objašnjena je u Gaussovim radovima iz 19. stoljeća. Glavna zadaća bila je minimizirati sumu kvadrata razlika između stvarnih i predviđenih vrijednosti kako bi se pronašli optimalni parametri regresijskog modela. [19]

Slika 1.1. Johann Carla Friedrich Gauss. Izvor: [1]

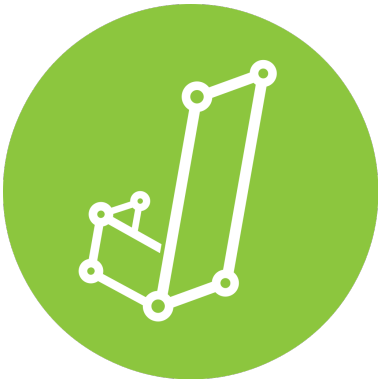
[1]Johann Carl Friedrich Gauss (1777.-1855.) bio je njemački matematičar, astronom, fizičar i statističar. Smatra se jednim od najvećih matematičara u povijesti i pridonosio je mnogim područjima matematike i znanosti.

U 19. stoljeću, britanski znanstvenik Francis Galton[2] donio je značajan doprinos regresijskoj analizi. Koristeći se regresijskom analizom, istraživao je odnose između visine roditelja i visine njihove djece te otkrio fenomen regresije prema srednjoj vrijednosti[3]. Galton je popularizirao regresijsku analizu u znanstvenoj zajednici i otvorio vrata za njezin daljnji razvoj. [19]



Slika 1.2. Sir Francis Galton. Izvor: [2]

Postoji više regresijskih modela poput linearne, polinomne, eksponencijalne, logističke, ... U ovom radu ćemo razmatrati, i najviše se fokusirati na logističku regresiju. Logistička regresija je statistički model koji se koristi za predviđanje vjerojatnosti ili klasifikaciju zavisne varijable koja ima binarnu ili višekategorijalnu prirodu. Logistička regresija povezana je s omjerom izgleda (eng. odds ratio), zbog čega se u radu bavimo i s tim statističkim pojmom.



Slika 1.3. Logo softverskog paketa JASP. Izvor: [8]

[2]Francis Galton (1822.-1911.) bio je engleski znanstvenik, istraživač, polihistor odnosno pionir u mnogim područjima, uključujući statistiku, antropologiju, psihologiju i eugeniku.

[3]Jednostavnim riječima, Galton je uočio da ako dobijemo uzorak s većim odstupanjem od srednje vrijednosti, da ćemo pri sljedećem uzorkovanju vrlo vjerojatno dobiti uzorak s manjim odstupanjem.

Sve eksperimentalne analize u ovom radu, kako za logističke tako i za druge regresijske modele napraviti ćemo u softverskom paketu JASP. JASP je besplatni open-source programski paket kreiran na Sveučilištu u Amsterdamu, a odlikuje se jednostavnošću korištenja. Princip rada u JASP-u sličan je principu rada u poznatom statističkom softverskom paketu SPSS-a[4]. Na slici 1.3 prikazan je logo programskog paketa JASP.

U ovom radu objasniti ćemo više vrsta regresijskih modela kao što su linearna univarijatna i multivarijatna regresija, a poseban naglasak stavit će se na logističku regresiju što je i glavna tema rada. Sve objašnjene regresijske metode potkrijepiti ćemo s više praktičnih primjera, pri čemu će biti i primjera povezanih s elektrotehnikom. Također će biti objašnjeni parametri kvalitete regresije i interpretacija rezultata regresijske analize.

[4]SPSS je statistički softverski paket koji je razvio IBM za upravljanje podacima, naprednu analitiku, multivarijantnu analizu, poslovnu inteligenciju i kriminalističku istragu.

2. Univarijatni modeli

"Univarijatno" je izraz iz statističke i istraživačke metodologije koji označava analizu ili proučavanje samo jedne varijable u istraživanju ili modeliranju. Ova riječ proizlazi iz "uni", što znači "jedan", te "varijatno", što potječe od riječi "varijabla" tj. ona koja ima sposobnost mijenjanja ili variranja. Osnovni oblik regresijske analize su univarijatni modeli koji se orijentiraju na odnos između jedne nezavisne varijable i jedne zavisne varijable, a često se nazivaju i jednostavnim modelima. U ovom poglavlju bavimo se sljedećim univarijatnim modelima:

- jednostavna linearna regresija,
- jednostavna polinomna regresija,
- jednostavna eksponencijalna regresija.

Ovo poglavlje obrađeno je prema izvorima [10], [11], [12] i [13].

2.1. Jednostavna linearna regresija

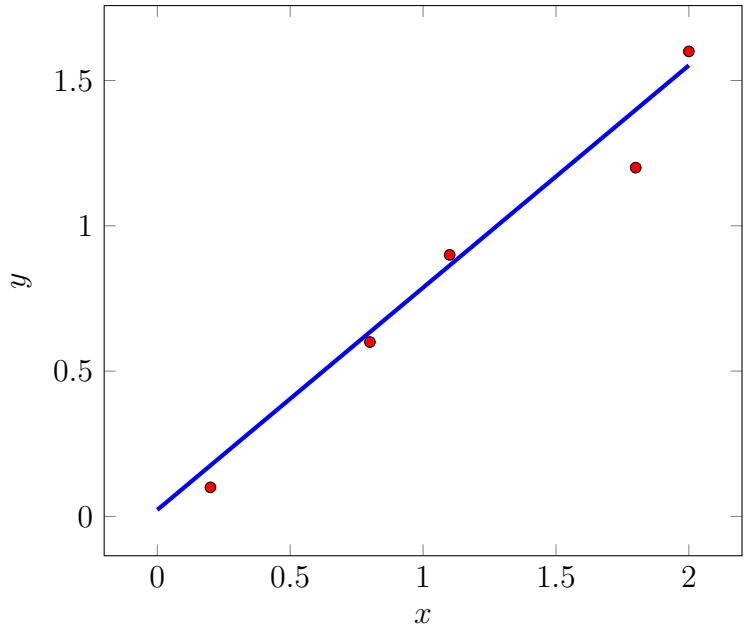
Jednostavna linearna regresija je statistička metoda koja se koristi za modeliranje linearne veze između jedne nezavisne varijable x i jedne zavisne varijable y. Spomenuta linearna veza između varijabli x i y može se interpretirati kao pravac u Descartesovu koordinatnom sustavu koji tada nazivamo pravcem regresije.

Općenito se jednostavna linearna regresija zapisuje kao jednadžba oblika:

$$y = \beta_0 + \beta_1 \cdot x + \varepsilon, \tag{2.1}$$

gdje je:

- y - zavisna varijabla koju predviđamo,
- x - nezavisna varijabla koju koristimo za predviđanje,
- β_0 - presjek s y-osi, tj. predviđena vrijednost varijable y kada je $x = 0$,
- β_1 - koeficijent regresije,
- ε - pogreška procjene ili varijacija.



Slika 2.1. Grafički prikaz linearne regresije. Izvor: izrada autora

Koeficijent β_1 računamo pomoću izraza:

$$\beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}, \tag{2.2}$$

gdje su x_i i y_i opservirane vrijednosti zavisne i nezavisne varijable, a n broj opservacija.

Izraz za koeficijent β_0 glasi:

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{y} - \beta_1 \bar{x}, \tag{2.3}$$

gdje su $\bar{y}$ i $\bar{x}$ aritmetičke sredine opserviranih vrijednosi zavisne, odnosno nezavisne varijable, dok je pogreška procjene ili varijacija ε_i pojedine varijable dana s:

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i). \tag{2.4}$$

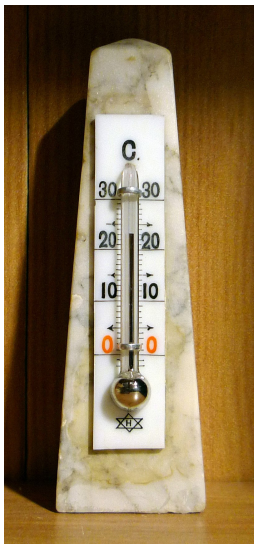
Linearna regresija usko je povezana sa metodom najmanjih kvadrata. Metoda najmanjih kvadrata je postupak ili metoda koja se u suštini bavi minimizacijom sume kvadrata pogrešaka. U ovom slučaju koeficijente β_1 i β_0 dobivamo tom metodom, pri čemu se minimizira funkcija

$$D(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{2.5}$$

Jedan od primjera moguće implementacije jednostavne linearne regresije je kod analize odnosa između temperature i širenja žive[1] u termometru koji se može modelirati linearnom funkci-

[1]Živa je jedini metal koji je pri sobnoj temperaturi u tekućem stanju. Njeno tekuće agregatno stanje čini ju korisnom u raznim industrijskim primjenama, uključujući termometre, barometre i električne uređaje. Izrazito je otrovna pri isparavanju te je stoga sve manje u primjeni.

jom. Drugim riječima, kako temperatura raste, živa se linearno širi. Ovaj linearni odnos je toliko relevantan da možemo upotrebljavati živin termometar u svrhu mjerenja temperature, a metodu linearne regresije koristiti za testiranje ispravnosti termometra. Naime, ako je termometar ispravan iz izmjerenih podataka trebali bi dobiti teorijski utvrđene koeficijente regresije. Živin termometar možemo vidjeti na sljedećoj slici.



Slika 2.2. Živin stakleni termometar za mjerenje sobne temperature. Izvor: [4]

U kontekstu elektrotehnike, jednostavna linearna regresija može se koristiti kod mjerenja otpora, pri čemu je baza Ohmov zakon za strujni krug i strujno-naponska karakteristika.

Ohmov zakon opisuje vezu između električne struje, napona i otpora. Prema Ohmovom zakonu struja kroz strujni krug je proporcionalna naponu, a obrnuto proporcionalna otporu u krugu. Ohmov zakon može se zapisati sljedećim izrazom:

$$R = \frac{U}{I}, \tag{2.6}$$

pri čemu su oznake objašnjene u sljedećoj tablici.

Tablica 2.1. Fizikalne veličine, oznake i mjerne jedinice povezane s Ohmovim zakonom.

Fizikalna veličina	Oznake fizikalne veličine	Mjerna jedinica	Oznaka mjerne jedinice
Električna struja	I	Amper	A
Električni napon	U	Volt	V
Električni otpor	R	Ohm	Ω

Vrijednost otpora R možemo interpretirati kao nagib pravca u koordinatnom sustavu gdje se vrijednost I nalazi na okomitoj osi (y - osi), a vrijednost U se nalazi na vodoravnoj osi (x - osi). Naponsko-strujna $(U - I)$ karakteristika otpora električnog strujnog kruga je grafički prikaz ovisnost jakosti struje I o promjeni napona U na otporu. U ovisnosti o tipu $U - I$ karakteristike razlikujemo:

- Linearni otpor ($U - I$ karakteristika je pravac čiji nagib je konstantan i jednak upravo R),
- Nelinearni otpor ($U - I$ karakteristika nije pravac).

U sljedećem primjeru pokazati ćemo kako funkcionira mjerenje otpora metodom linearne regresije.

Primjer 2.1. *Cilj nam je što preciznije izmjeriti iznos otpora, pri čemu provodimo mjerenja jakosti struje pri različitim naponima. Nakon toga izračunavamo koeficijente regresije kako bismo dobili optimalno prilagođeni pravac kojim je određena $(U - I)$ karakteristika, odnosno samu vrijednost otpora. Izmjerene vrijednosti dane su u sljedećoj tablici.*

Tablica 2.2. Izmjerene vrijednosti jakosti struje i napona. Izvor: Mjerenje autora

Redni broj mjerenja	1	2	3	4	5	6
$y_i(U[V])$	0	0.83	1.34	1.70	1.90	2.10
$x_i(I[A])$	0	0.40	0.66	0.83	0.93	1.13

Da bi izračunali koeficijente pravca regresije β_0 i β_1 koristimo izraze (2.3) *i* (2.2) *te formiramo sljedeću pomoćnu tablicu:*

Tablica 2.3. Pomoćna tablica za izračunavanje koeficijenata regresije.

i	$y_i(U[V])$	$x_i(I[A])$	x_i^2	$x_i y_i$
1	0	0	0	0
2	0.83	0.40	0.16	0.332
3	1.34	0.66	0.4356	0.8844
4	1.70	0.83	0.6889	1.411
5	1.90	0.93	0.8649	1.767
6	2.10	1.13	1.2769	2.373
$\sum$	7.87	3.95	3.4263	6.7674

Korištenjem ove tablica i spomenutih izraza, koeficijente regresije računamo na sljedeći način:

$$\beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{6 \cdot 6.7674 - 31.0865}{6 \cdot 3.4263 - 15.6025} = 1.9208. \tag{2.7}$$

Kako bismo izračunali koeficijent β_0, potrebno je izračunati $\bar{x}$ i $\bar{y}$, pa vrijedi:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6} \cdot 3.95 = 0.6583, \tag{2.8}$$

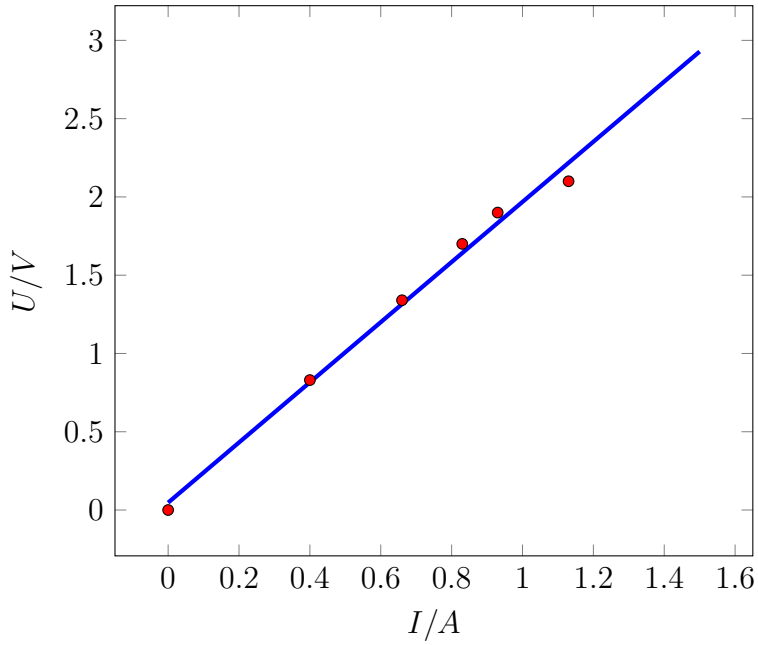
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{6} \cdot 7.87 = 1.3117. \tag{2.9}$$

Sada je:

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 1.3117 - 1.9208 \cdot 0.6583 = 0.0472. \tag{2.10}$$

Prema Ohmovom zakonu je $U = R \cdot I$ pa koeficijent β_1 odgovara traženoj vrijednosti otpora.

Na sljedećem grafikonu prikazana je $(U - I)$ karakteristika za dane podatke te pripadni pravac regresije.



Slika 2.3. Graf raspršenja za primjer 2.1. Izvor: izrada autora

2.2. Jednostavna eksponencijalna regresija

Statistička metoda jednostavne eksponencijalne regresije koristi se za modeliranje eksponencijalne veze između jedne nezavisne varijable x i jedne zavisne varijable y. Drugim riječima, ovu ćemo metodu koristiti kada na dijagramu raspršenja uočimo eksponencijalni rast ili pad.

Jednadžba za model jednostavne eksponencijalne regresije glasi:

$$y = ab^x, \tag{2.11}$$

gdje je:

- y - zavisna varijabla koju predviđamo,
- x - nezavisna varijabla,
- a - početna vrijednost zavisne varijable,
- b - stopa promjene eksponencijalnog rasta ili pada.

Cilj jednostavne eksponencijalne regresije je procijeniti vrijednosti koeficijenata a i b iz postojećih podataka kako bi se stvorio model koji odgovara podacima. Eksponencijalnu regresiju koristimo i kada imamo sporiji rast, ali naglo ubrzanje u nekom trenutku.

Procjena parametara kod jednostavne eksponencijalne regresije radi se prelaskom na linearni model na sljedeći način. Logaritmiranjem izraza (2.11) dobivamo:

$$\ln(y) = \ln(a \cdot b^x). \tag{2.12}$$

Nadalje, primjenom svojstava logaritama dobivamo:

$$\ln(y) = \ln(a) + \ln(b^x), \tag{2.13}$$

odnosno

$$\ln(y) = \ln(a) + x \cdot \ln(b). \tag{2.14}$$

Ovime smo izrazili $\ln(y)$ kao linearnu funkciju od x, s nagibom $\ln(b)$ i sjecištem $\ln(a)$. Drugim riječima, provodimo linearnu regresiju na parovima opservacija oblika $(x_i, \ln(y_i))$.

Još je potrebno iz regresijskog pravca oblika

$$\ln(y) = c + x \cdot m, \tag{2.15}$$

odrediti početne koeficijente a i b. Vrijedi:

$$\ln(a) = c, \qquad \ln(b) = m, \tag{2.16}$$

odakle antilogaritmiranjem dobivamo:

$$a = \exp(c), \qquad b = \exp(m). \tag{2.17}$$

Jedan od primjera primjene jednostavne eksponencijalne regresije je u području demografije, točnije kod proučavanja promjena u stanovništvu ili populaciji kroz određeni period. Koristeći postojeće podatke, ova metoda omogućuje predviđanje budućeg rasta ili pada populacije te pruža važne smjernice za planiranje i političko djelovanje, što ćemo ilustrirati u sljedećem primjeru.

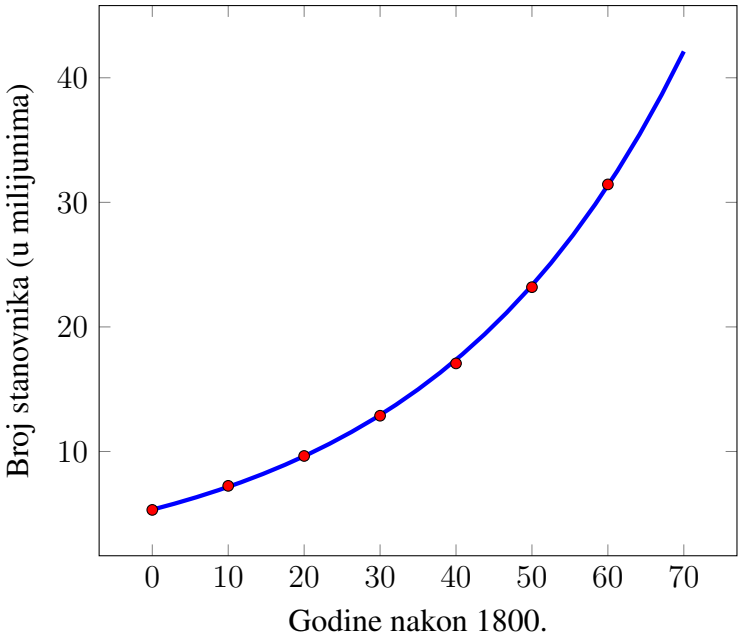
Primjer 2.2. *U tablici je prikazan broj stanovnika SAD-a mjeren u milijunima za različite godine. Pretpostavimo da varijabla x_i predstavlja broj godina od 1800. pa nadalje.*

Tablica 2.4. Broj stanovnika u SAD-u. Izvor: [5]

Godina $[x_i]$	0	10	20	30	40	50	60
Populacija $[y_i]$	5.31	7.24	9.64	12.87	17.07	23.19	31.44

Temeljem ovih podataka dobiva se eksponencijalni regresijski model opisan je sljedećom jednadžbom.

$$y = 5.33 \cdot 1.03^x. \tag{2.18}$$



Slika 2.4. Dijagram raspršenja i eksponencijalni regresijski model za broj stanovnika u SAD-u. Izvor: izrada autora

Na slici 2.4 prikazan je dijagram raspršenja zajedno s dobivenim eksponencijalnim modelom. Sa grafikona je vidljivo da populacija iz desetljeća u desetljeće blago eksponencijalno raste, a to je i dokazano dobivenim eksponencijalnim modelom.

Koristeći se opisanom jednadžbom možemo prediktirati populaciju za 2021. godinu kada je proveden popis stanovništva te usporediti sa stvarnim podacima. Dakle, uvrštavanjem $x = 221$, što odgovara 2021. godini dobivamo rezultat 3662.30 što znači da bi po ovom modelu broj stanovnika u SAD-u trebao iznositi oko 3.662 milijardi stanovnika u 2021. godini. Usporedimo li to sa stvarnim podacima koji govore da je u 2021. godini zabilježeno 331.9 milijuna stanovnika što je višestruko manje od dobivenog rezultata.

Iako se na prvu može činiti da je model neupotrebljiv, problem je u pretjeranoj ekstrapolaciji[2]*. Naime, regresijski model je dobar za predikciju vrijednosti koje su bliske opserviranima, a razlika veća od 100 godina nikako se ne može smatrati malom. Problem bi se djelomično mogao riješiti da se umjesto univarijatnog pristupa promatra multivarijatni, tj. da se u obzir uzima više čimbenika koji imaju utjecaj na veličinu populacije kroz godine, a to će biti objašnjeno u idućim poglavljima.*

2.3. Jednostavna polinomna regresija

Jednostavna polinomna regresija je statistička metoda koju koristimo za modeliranje veza između nezavisne varijable x i zavisne varijable y pomoću polinomne funkcije. Ova metoda omogućuje modeliranje krivulje koja se prilagođava podacima i omogućuje fleksibilnost u odnosu između

[2]Kod predikcije temeljem regresijskih modela razlikujemo problem interpolacije i problem ekstrapolacije. Interpolacija pretpostavlja predviđanje za neku vrijednost unutar granica vrijednosti nezavisne varijable, dok ekstrapolacija znači predviđanje temeljem vrijednosti koje nisu nužno bliske zadanim vrijednostima nezavisne varijable.

varijabli. Drugim riječima, pogodna je opisivanje nelinearnih modela kod kojih nema eksponencijalnog rasta. U modelu polinomne regresije pretpostavljamo da je odnos između zavisne varijable i jedne nezavisne varijable opisan polinomom nekog proizvoljnog stupnja.

Jednadžba za jednostavnu polinomnu regresiju glasi:

$$y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \cdots + \beta_n \cdot x^n + \varepsilon, \tag{2.19}$$

gdje je:

- y - zavisna varijabla koju predviđamo,
- x - nezavisna varijabla,
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ - koeficijenti regresije kojima je opisan utjecaj nezavisne varijable na zavisnu varijablu.

U sljedećem primjeru pokazati ćemo jedan primjer korištenja polinomne regresije u elektrotehnici.

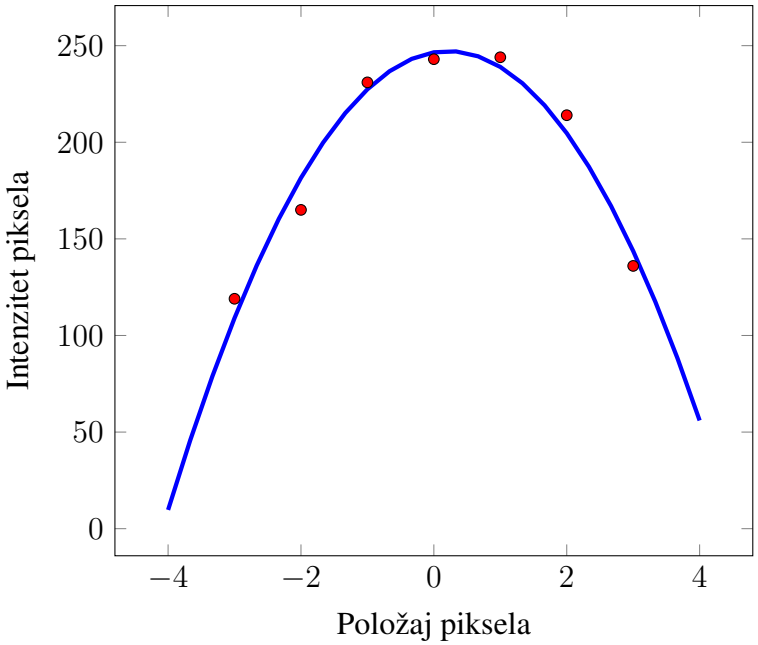
Primjer 2.3. *U ovom primjeru bavimo se vezom intenziteta svijetlosti kao zavisne varijable te položaja piksela kao nezavisne varijable. Pripadni podaci dobiveni opseriviranjem navedeni su u sljedećoj tablici.*

Tablica 2.5. Veza položaja piksela i intenziteta svijetlosti. Izvor: [20]

Položaj piksela, x	Intenzitet piksela, y
-3	119
-2	165
-1	231
0	243
1	244
2	214
3	136

Na slici 2.5 možemo vidjeti da veza između ove dvije varijable nije linearna, odnosno kao regresijska krivulja može se očekivati parabola. To znači da ćemo za regresijski model koristiti polinom drugog stupnja koji u ovom slučaju glasi:

$$y = 246.57 + 5.7857x - 13.357x^2. \tag{2.20}$$



Slika 2.5. Dijagram raspršenja i polinomni regresijski model za položaj piskela u ovisnosti o intenzitetu. Izvor: izrada autora

2.4. Analiza kvalitete univarijatnog regresijskog modela

Iz prethodnih primjera je jasno da regresijski model može biti manje ili više točan, tj. da predikcije koje dobijemo mogu biti bliske opserviranim vrijednostima, ali od njih i značajno odstupati. Time se nameće potreba za uvođenjem metrika kojima se mjeri kvaliteta regresijskog modela. Kvaliteta regresije najčešće se mjeri koeficijentom determinacije koji označavamo s R^2 te različitim statističkim testovima, pri čemu se najčešće koriste F-test i t-test.

2.4.1. Koeficijent determinacije

Koeficijent determinacije izračunava se pomoću izraza

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{2.21}$$

gdje je:

- y_i - opservirana vrijednost zavisne varijable y za opservaciju i,
- $\hat{y}_i$ - prediktirana vrijednost zavisne varijable y za opservaciju i,
- $\bar{y}$ - aritmetička sredina opserviranih vrijednosti zavisne varijable y.

Vrijednost R^2 kreće se između 0 i 1, gdje vrijednost 1 označava savršen model, a vrijednost 0 odsutnost veze između zavisne i nezavisne varijable. U praksi se vrijednosti koeficijenta veće od 0.7 kod univarijatnih modela smatraju prihvatljivima. Koeficijent determinacije može se interpretirati i kao postotak objašnjene varijabilnosti. Primjerice, ako je $R^2 = 0.8$ to znači da je nezavisna varijabla objasnila 80% varijabilnosti zavisne varijable.

Pokažimo sada kako se izračunava koeficijent determinacije na podacima iz primjera 2.1.

Primjer 2.4. *Izračun koeficijenta determinacije radi se pomoću pomoćne tablice navedene u nastavku.*

Tablica 2.6. Pomoćna tablica za izračun koeficijenta determinacije.

i	y_i	x_i	$\hat{y}_i = \beta_0 + \beta_1 x_i$	$(y_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
1	0	0	0.0472	1.7206	0.00223
2	0.83	0.40	0.8155	0.2320	0.00021
3	1.34	0.66	1.3149	0.00081	0.00063
4	1.70	0.83	1.6415	0.1508	0.00342
5	1.90	0.93	1.8335	0.3461	0.00442
6	2.10	1.13	2.2177	0.6214	0.01385
$\sum$	7.87	3.95	7.8703	3.07171	0.02476

Sada dobivamo:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{0.02476}{3.0717} = 0.9919. \tag{2.22}$$

U ovom slučaju dobili smo koeficijent determinacije koji je gotovo jednak 1, što ukazuje na iznimno kvalitetan regresijski model i opravdava korištenje regresijske metode za mjerenje otpora. Ovaj se rezultat može interpretirati i drugim riječima, odnosno da je čak 99.19% varijabilnosti napona objašnjeno s varijabilnošću struje.

2.4.2. Statistički testovi

Drugi način kojim ispitujemo kvalitetu regresijskog modela su statistički testovi. Njihova svrha je da provjere razlikuju li se regresijski koeficijenti statistički značajno od nule. F-test u obzir uzima sve koeficijente, tj. on odgovara na pitanje razlikuje li se barem jedan koeficijent regresije od nule, dok je t-test specifičan i on se provodi na svakom pojedinom koeficijentu. Dakle, t-testom utvrđujemo razlikuje li se neki pojedini koeficijent od nule. Ručno provođenje statističkih testova zahtjeva značajan matematički aparat koji izlazi iz okvira ovog rada, tako da ovdje objašnjavamo samo interpretaciju rezultata tih testova.

Svaki statistički test tumači se pomoću p-vrijednosti. Ako je p-vrijednost manja od 0.05 to sugerira postojanje statistički značajne razlike. Drugim riječima, da bismo prihvatili neki regresijski model, p-vrijednosti pridružene F-testu i t-testu moraju biti manje od 0.05 budući to znači da su koeficijenti regresije statistički značajno različiti od nule.

Praktičnu primjenu statističkih testova pokazati ćemo u sljedećem poglavlju koje se bavi provođenjem univarijante regresijske analize unutar softverskih paketa.

2.5. Univarijatna regresijska analiza pomoću softverskih paketa

U ovom ćemo poglavlju pokazati kako izgleda regresijska analiza kada se ona provodi pomoću softverskih paketa. Analiza je u svim softverskim paketima slična, svodi se na unos vrijednosti zavisne i nezavisne varijable, a i rezultat obrade je vrlo sličan. Stoga ovdje prikazujemo kako izgleda rezultat obrade u softverskom paketu JASP.

U sljedećem primjeru ponavljamo regresijsku analizu iz primjera 2.1.

Primjer 2.5. *Na slici 2.6 prikazana je regresijska analiza u softerskom paketu JASP s podacima iz primjera 2.1.*

Linear Regression ▾

Model Summary - Yi

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.996	0.992	0.990	0.079

Note. Null model includes Xi

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_0	Regression	3.047	1	3.047	492.169	< .001
	Residual	0.025	4	0.006		
	Total	3.072	5			

Note. Null model includes Xi

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	0.047	0.065		0.721	0.511
	Xi	1.921	0.087	0.996	22.185	< .001

Slika 2.6. Regresijska analiza u programskom paketu JASP. Izvor: izrada autora.

Objasnimo najprije sve elemente ove analize. U prvoj tablici prikazana je kvaliteta regresije kroz koeficijent determinacije. JASP ovdje osim klasičnog koeficijenta determinacije R^2 prikazuje i neke srodne pokazatelje kao što je primjerice RMSE (standardna devijacija rezidula). Za potrebe ovog rada zadržat ćemo se na interpretaciji klasičnog koeficijenta determinacije. Možemo vidjeti da se dobila gotovo ista vrijednost kao i kod ručnog izračuna, a do razlike je došlo samo zbog greške zaokruživanja.

Druga tablica se odnosno na F-test, a kako smo već prije objasnili kod F-testa je dovoljno pogledati zadnji stupac, odnosno p-vrijednost. Ostali elementi tablice samo su pomoćni elementi neophodni za izračun p-vrijednosti i njihovo tumačenje nije relevantno. Ova p-vrijednost je značajno manja od 0.05, što znači da se barem jedan regresijski koeficijent razlikuje od 0, odnosno da regresijski model u ukupnosti ima smisla.

U zadnjoj tablici bitan je prvi stupac označen s "Unstandardized" u kojem su navedeni regresijski koeficijenti. U prvom redu je koeficijent β_0, a u drugom koeficijent β_1. Također vidimo da su vrijednosti bliske onima dobivenim ručno. U ovoj tablici bitan nam je i zadnji stupac u kojem su prikazani rezultati t-testa, odnosno značajnosti svakog pojedinog koeficijenta regresije. Ovdje vidimo da je koeficijent β_1 statistički značajno različit od nule, što je nalaz koji smo željeli.

U sljedećem primjeru baviti ćemo se predikcijom bodova iz kolegija Matematika 2 na osnovu bodova iz kolegija Matematika 1. Analiza je izvršena na realnim podacima slučajno odabranog uzorka studenata Tehničkog fakulteta Sveučilišta u Rijeci. Podatke je ustupio nositelj kolegija.

Primjer 2.6. *Na slici 2.7 prikazana je jednostavna regresijska analiza kojom je analiziran odnos između bodova na kolegiju Matematika 1 kao nezavisne varijable i bodova iz Matematike 2 kao zavisne varijable.*

Kako bi ocijenili kvalitetu regresije i dokazali da varijabla MAT1 ima utjecaj na varijablu MAT2, promatrati ćemo određene parametre sa slike. Koeficijent determinacije R^2 iznosi 0.692 što nam govori da je model reprezentativan. Drugim riječima, bodovi na kolegiju Matematika 1 objašnjavaju čak 69.2% varijabilnosti bodova na kolegiju Matematika 1. Vrijednost p-testa pridružena F-testu je manja od 0.05 što znači da se koeficijenti u modelu statistički razlikuje od 0. Iz rezultata p-testiranja za t-test uočavamo značajan utjecaj nezavisne varijable na zavisnu varijablu. Zaključujemo da će uspjeh pojedinog studenta na kolegiju "Matematika 1" imati značajan utjecaj za ostvarivanje uspjeha na kolegiju "Matematika 2" i taj se odnos može izraziti formulom:

$$MAT2 = 0.647 + 0.827 \cdot MAT1. \tag{2.23}$$

Drugim riječima, za svaki postignuti bod na kolegiju Matematika 1, može se očekivati 0.827 bodova više iz kolegija Matematika 2. Ovaj rezultat može se koristiti kao motivacijski primjer koji studentima dokazuje povezanost ova dva kolegija, odnosno važnost dobre pripreme ispita iz kolegija Matematika 2.

Linear Regression ▼

Model Summary - MAT2

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.832	0.692	0.689	0.592

Note. Null model includes MAT1

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_0	Regression	62.322	1	62.322	177.879	< .001
	Residual	27.678	79	0.350		
	Total	90.000	80			

Note. Null model includes MAT1

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	0.647	0.212		3.055	0.003
	MAT1	0.827	0.062	0.832	13.337	< .001

Slika 2.7. Regresijska analiza u JASP programskom paketu . Izvor: izrada autora.

Primjerice, ako student ima iz "Matematike 1" 40 bodova, tada na kolegiju "Matematika 2" može očekivati 73.08 bodova.

3. Multivarijatni modeli

"Multivarijatno" je izraz iz statističke i istraživačke metodologije koji označava analizu ili proučavanje više varijabli u istraživanju ili modeliranju. Ova riječ proizlazi iz "multi", što znači "više", te "varijatno", što potječe od riječi "varijabla". Multivarijatna regresija obično se koristi za predviđanje ponašanja zavisne varijabli u odnosu na ponašanje više nezavisnih varijabli u svim područjima znanosti, a nezaobilazna je u algoritmima strojnog učenja. Postoji više multivarijatnih regresijskih modela, a najčešći je multivarijatni linearni model koji ćemo ovdje detaljnije objasniti kako bi razumjeli razliku između univarijatnog i multivarijatnog pristupa. Ovo poglavlje obrađeno je prema izvorima: [14], [15].

3.1. Multivarijatna linearna regresija

Multivarijatna linearna regresija je statistička metoda koja se koristi za modeliranje linearne veze između više nezavisnih varijabli (prediktora) i jedne zavisne varijable. Ova metoda omogućuje analizu istovremenih utjecaja nezavisnih varijabli na zavisnu varijablu, uzimajući u obzir i njihove međusobne veze.

Matematički zapis modela multivarijatne linearne regresije dan je s:

$$y = \beta_0 + \beta_1 x + \beta_2 x + ... + \beta_n x_n + \varepsilon_i, \tag{3.1}$$

gdje je:

- y - zavisna varijabla,
- x_1, x_2,...,x_n - nezavisne varijable (prediktori),
- β_0, β_1, β_2,..., β_n - regresijski koeficijenti,
- ε - pogreška procjene.

Za procjenu regresijskih koeficijenata iz (3.1.) nužno je imati niz opservacija zavisne varijable i pripadnih opservacija nezavisnih varijabli:

$$y_1, x_{11}, x_{12}, \ldots, x_{1n}, \tag{3.2}$$

$$y_2, x_{21}, x_{22}, \ldots, x_{2n}, \tag{3.3}$$

$$\ldots \tag{3.4}$$

$$y_k, x_{k1}, x_{k2}, \ldots, x_{kn}, \tag{3.5}$$

$$\tag{3.6}$$

gdje je n broj varijabli, a k broj dostupnih opservacija. Na taj način procjenu koeficijenata radimo rješavanjem linearnog sustava koji možemo matrično zapisati na sljedeći način:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_k \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1n} \\ 1 & x_{21} & x_{22} & ... & x_{2n} \\ . & & & & \\ . & & & & \\ 1 & x_{n1} & x_{n2} & ... & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_k \end{bmatrix}, \tag{3.7}$$

što skraćeno možemo zapisati na sljedeći način:

$$Y = X\beta + \varepsilon, \tag{3.8}$$

gdje je:

- Y - vektor opserviranih vrijednosti zavisne varijable,
- X - matrica opserviranih vrijednosti nezavisnih varijabli,
- β - vektor regresijskih koeficijenata,
- ε - vektor pogrešaka.

Vektor β možemo izraziti kao:

$$\beta = (X'X)^{-1}X'Y, \tag{3.9}$$

gdje je X' transponirana matrica matrice X. Do ovog smo izraza došli množenjem izraza (3.8) matricom X'. Time smo dobili kvadratni sustav koji se rješava množenjem s inverznom matricom.

Grafički prikaz multivarijatne linearne regresije postaje složeniji jer se ne može prikazati na jednostavnom dvodimenzionalnom grafu. Drugim riječima, za grafički prikaz ovog modela trebamo toliko dimenzija koliko imamo varijabli, što je već za tri nezavisne varijable problem. Stoga se u praksi multivarijatni modeli uglavnom ne prikazuju grafički.

3.2. Multilinearna regresijska analiza pomoću softverskih paketa

U ovom poglavlju, slično kao u Poglavlju 2.5. baviti ćemo se multilinearnom analizom u softverskom programu JASP, odnosno unositi ćemo jednu zavisnu varijablu i više nezavisnih varijabli i prikazati kako nezavisne varijable utječu na zavisnu varijablu kroz pojedine primjere u nastavku.

U sljedećem primjeru prikazati ćemo koliki utjecaj imaju broj cilindara te snaga motora na potrošnju pojedinog automobila.

Primjer 3.1. *Na slici 3.1 prikazana je multivarijatna regresijska analiza kojom je analiziran odnos između snage, broja cilindara i potrošnje automobila. Izvor: [21].*

Linear Regression ▼

Model Summary - Potrosnja, l/100 km ▼

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.836	0.698	0.675	2.658

Note. Null model includes Broj cilindara, Snaga, kW

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_0	Regression	424.982	2	212.491	30.070	< .001
	Residual	183.733	26	7.067		
	Total	608.715	28			

Note. Null model includes Broj cilindara, Snaga, kW

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	1.914	2.071		0.924	0.364
	Broj cilindara	0.762	0.510	0.282	1.494	0.147
	Snaga, kW	0.068	0.022	0.589	3.119	0.004

Slika 3.1. Regresijska analiza u programskom paketu JASP. Izvor: izrada autora.

Koeficijent determinacije R^2 iznosi 0.698 što nam govori da je model reprezentativan. Možemo reći da broj cilindara i snaga objašnjavaju 69.8% varijabilnosti potrošnje automobila izražene u litrama na 100 km. Također, p-vrijednost pridružena F-testu je manja od 0.05 što znači da se barem jedan regresijski koeficijent u modelu statistički razlikuje od 0, odnosno da regresijski model ima smisla.

p-vrijednost pridružena t-testu za varijablu "Broj cilindara" je veća 0.05 što znači da ovaj koeficijent nema statistički značajan utjecaj na zavisnu varijablu. Za varijablu "Snaga, kW" p-vrijednost pridružena t-testu je manja od 0.05, što znači ovaj koeficijent ima značajniji utjecaj na zavisnu varijablu. Kako je regresijski koeficijent pozitivan, možemo očekivati da što je veća snaga motora, da će i potrošnja automobila biti sve veća. Konkretno, svaki dodatni kW snage povećava potrošnju za 0.068 litara na 100 km.

Analizirani model može se prikazati formulom:

$$y = 1.914 + 0.762 \cdot x_1 + 0.068 \cdot x_2, \tag{3.10}$$

pri čemu je:

- x_1 *- broj cilindara,*
- x_2 *- snaga motora.*

Primjerice, ako se motor automobila sastoji od 3 cilindra te njegova snaga iznosi 60 kW, potrošnja automobila će prema ovoj formuli iznositi 8.82 litre na 100 km.

U sljedećem primjeru nadovezati ćemo se na prethodni primjer, te ćemo provesti analizu dodavanjem nekolicine nezavisnih varijabli i vidjeti njihov kompletan utjecaj na zavisnu varijablu.

Primjer 3.2. *Na slici 3.2 prikazana je regresijska analiza u kojoj smo analizirali utjecaj mase, snage, zapremine motora i broja cilindara na potrošnju automobila. Izvor: [21].*

Linear Regression ▼

Model Summary - Potrosnja, l/100 km ▼

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.939	0.881	0.862	1.735

Note. Null model includes Broj cilindara, Snaga, kW, Masa, kg, Zapremina, kubicni centimetri

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_0	Regression	536.458	4	134.114	44.545	< .001
	Residual	72.258	24	3.011		
	Total	608.715	28			

Note. Null model includes Broj cilindara, Snaga, kW, Masa, kg, Zapremina, kubicni centimetri

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	−5.787	1.879		−3.080	0.005
	Broj cilindara	1.772	0.605	0.656	2.928	0.007
	Snaga, kW	0.083	0.026	0.712	3.140	0.004
	Masa, kg	0.007	0.002	0.551	2.741	0.011
	Zapremina, kubicni centimetri	−0.002	5.858×10^{-4}	−1.031	−4.024	< .001

Slika 3.2. Regresijska analiza u programskom paketu JASP. Izvor: izrada autora.

Koeficijent determinacije R^2 iznosi 0.881 što nam govori da je model reprezentativan. Možemo reći da nam nezavisne varijable objašnjavaju 88.1% varijabilnosti potrošnje automobila. Vrijednost p-testa pridruženog F-testu je manja od 0.05 što kao i u prethodnom primjeru govori

da model ima statističkog smisla, odnosno da se barem jedan regresijski koeficijent razlikuje od nule.

p-vrijednost pridružena t-testu je za sve varijable manja od 0.05 što nam govori da sve varijable imaju značajan utjecaj na zavisnu varijablu. Stupac "standardized" nam pomaže kod uspoređivanja utjecaja među koeficijentima, pa vidimo da nam najveći utjecaj na potrošnju ima zapremina motora, dok najmanji utjecaj ima masa automobila.

Zaključujemo da u ovom slučaju snaga, broj cilindara, masa te zapremina motora imaju značajan utjecaj na to koliki će biti iznos potrošnje automobila. Konkretno, svaki dodatni cilindar povećao bi potrošnju za 1.772 litre na 100 km, svaki dodatni kW snage za 0.083 litre, a svaki dodatni kilogram mase za 0.007 litara. Svaki dodatni kubični centimetar zapremnine motora smanjio bi potrošnju za 0.002 litre.

Ovaj model možemo prikazati formulom:

$$y = -5.787 + 1.772 \cdot x_1 + 0.083 \cdot x_2 + 0.007 \cdot x_3 - 0.002 \cdot x_4, \tag{3.11}$$

pri čemu je :

- x_1 *- broj cilindara,*
- x_2 *- snaga motora,*
- x_3 *- masa automobila,*
- x_4 *- zapremina motora.*

Pokažimo još primjenu dobivene formule. Kao i u prethodnom primjeru, pretpostaviti ćemo da je broj cilindara motora 4, snaga 89 kW, masa 1692 kg, te da je zapremnina motora 1998 cm^3. Potrošnja automobila prema ovom izrazu bi u tom slučaju iznosila 16.54 l/100 km. Ove konkretne karakteristike opisuju model automobila "Mazda CX3" kod koje je proizvođač specificirao prosječnu potrošnju od 6.2 l/100 km, što sa značajno razlikuje u odnosu na našu predikciju. Na temelju toga zaključujemo da ovaj model nije dobar, a to možemo objasniti činjenicom da je predikcija rađena za novije vozilo koje ne odgovara podacima na temelju kojih je formiran model.

U sljedećem primjeru ćemo prediktirati kako prisutnost na nastavi, te bodovi iz prvog, drugog i trećeg kolokvija utječu na ukupan broj bodova na kontrolnoj zadaći iz jednog matematičkog kolegija. Ova analiza je bazirana na stvarnim podacima Tehničkog fakulteta Sveučilišta u Rijeci na slučajno odabranom uzorku studenata. Podatke je ustupio nositelj kolegija.

Primjer 3.3. *Na slici 3.3 prikazana je multivarijatna regresijska analiza u kojoj smo analizirali utjecaj prisutnosti i bodova sa sva 3 kolokvija na ukupan broj bodova ostvaren na kolegiju.*

Koeficijent determinacije iznosi 0.940 što nam govori da je model reprezentativan i da nezavisne varijable u ovom primjeru objašnjavaju čak 94% varijabilnosti ukupnih bodova. p-vrijednost

Linear Regression ▼

Model Summary - Kontrolne zadaće ukupno (60)

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.969	0.940	0.935	3.149

Note. Null model includes Prisutnost (100), 1. KOLOKVIJ (15), 2. KOLOKVIJ (15), 3. KOLOKVIJ (15)

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_0	Regression	8337.466	4	2084.367	210.172	< .001
	Residual	535.542	54	9.917		
	Total	8873.008	58			

Note. Null model includes Prisutnost (100), 1. KOLOKVIJ (15), 2. KOLOKVIJ (15), 3. KOLOKVIJ (15)

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	−0.489	1.951		−0.251	0.803
	Prisutnost (100)	0.056	0.031	0.083	1.814	0.075
	1. KOLOKVIJ (15)	1.104	0.150	0.348	7.366	< .001
	2. KOLOKVIJ (15)	1.368	0.186	0.371	7.343	< .001
	3. KOLOKVIJ (15)	1.030	0.172	0.321	5.976	< .001

Slika 3.3. Regresijska analiza u programskom paketu JASP. Izvor: izrada autora.

pridružena F-testu je manja od 0.05, što znači da model ima statističkog smisla. Iz p-vrijednosti pridružene t-testu uočavamo da sva 3 kolokvija imaju statistički značajan utjecaj na ukupne bodove dok prisutnost nema značajan utjecaj.

Ovaj model možemo prikazati formulom:

$$y = -0.489 + 0.056 \cdot x_1 + 1.104 \cdot x_2 + 1.368 \cdot x_3 + 1.030 \cdot x_4, \tag{3.12}$$

pri čemu je :

- x_1 *- prisutnost u postotku,*
- x_2 *- bodovi s 1. kolokvija,*
- x_3 *- bodovi s 2. kolokvija,*
- x_4 *- bodovi s 3. kolokvija.*

Ako pretpostavimo da je neki student na prvom kolokviju ostvario 5 bodova, na drugom 10, te na trećem 7 uz 80 % prisustva na nastavi, student će imati ukupno 30.4 boda.

Promatrajući koeficijente, zaključujemo da nam je drugi kolokvij najvažniji jer je koeficijent uz njega najveći. Ovaj nalaz može se i logički objasniti. Prosječni je student na prvom kolokviju relativno opušten zbog činjenice da mu preostaju još dva kolokvija na kojima može skupiti bodove. Ako ne ostvari zadovoljavajući broj bodova na prvom kolokviju, student se ipak odlučuje maksimalno potruditi na drugom kolokviju kako bi osigurao dovoljno bodova za prolaz. Što se tiče trećeg kolokvija, student već zna na čemu je zbog čega vjerojatno taj kolokvij ima najmanji utjecaj na ukupan broj bodova iz kolegija.

4. Omjer izgleda (eng. odds ratio)

U ovom se poglavlju bavimo vjerojatnošću i izgledima, što će nam biti neophodno za analizu logističke regresije. Naime, cilj logističke regresije je procijeniti koliko neka nezavisna varijabla utječe na vjerojatnost realizacije nekog događaja. Međutim, taj je utjecaj iskazan na specifičan način kroz omjer izgleda (eng. odds ratio) koji ćemo u ovom poglavlju objasniti. Poglavlje je obrađeno prema izvorima [16] i [17].

Vjerojatnost i izgledi mjere su šansi realizacije događaja. Dok je vjerojatnost direktna mjera šanse realizacije, izgledi prikazuju odnos između šansi realizacije događaja i šansi realizacije njemu suprotnog događaja. Izgledima se ponekad daje drugačije značenje i interpretacija u odnosu na vjerojatnost. Na primjer, izgledi se često koriste kao mjera "vrijednosti" u sportskom klađenju, dok se vjerojatnost koristi za razumijevanje apsolutnih šansi događaja.

Definirajmo sada pojam izgleda (engl. odds).

Definicija 4.1. *Neka je zadan događaj A čija je vjerojatnost $p(A) = P$. Izgledi događaja A računaju se formulom*

$$Odds = \frac{P}{1-P}, \tag{4.1}$$

gdje brojnik predstavlja vjerojatnost realizacije događaj A, a nazivnik vjerojatnost da se događaj A neće realizirati.

Pokažimo sada na nekoliko primjera kako se šanse realizacije iskazuju pomoću izgleda.

U prvom primjeru odrediti ćemo izglede za pad šestice na igraćoj kocki. Vjerojatnost tog događaja dana je s:

$$p(\text{pad na 6}) = \frac{1}{6}, \tag{4.2}$$

dok je pripadni izled jednak

$$Odds = \frac{\frac{1}{6}}{\frac{5}{6}} = \frac{1}{5} = 0.2. \tag{4.3}$$

Sada ćemo analizirati pad parnog broja na igraćoj kocki. Kao i u prethodnom primjeru dobivamo:

$$p(\text{pad na paran broj}) = \frac{3}{6} = \frac{1}{2}, \qquad Odds = \frac{\frac{1}{2}}{\frac{1}{2}} = 1, \tag{4.4}$$

U trećem primjeru određujemo izglede da će na igraćoj kocki pasti na broj manji od. Slijedi:

$$p(\text{pad na broj manji od 6}) = \frac{5}{6}, \qquad Odds = \frac{\frac{5}{6}}{\frac{1}{6}} = 5. \tag{4.5}$$

Možemo uočiti da su izgledi kod događaja čija je vjerojatnost manja od 50% manji od 1, dok su za događaje čija je vjerojatnost veća od 50% izgledi veći od 1.

Uz pojam izgleda često se uvodi i omjer izgleda (OR) koji je pogodan za iskazivanje razlike u vjerojatnosti realizacije pri različitim uvjetima. Uobičajeno se omjer izgleda koristi kod identificiranja čimbenika rizika procjenom odnosa između izloženosti čimbeniku rizika i medicinskog ishoda. Na primjer, postoji li povezanost između izloženosti kemikaliji i bolesti. Pokažimo ovo na jednom primjeru.

Primjer 4.1. *Promotrimo grupu pušača (izloženih) i nepušača (ne izloženih) te istražimo koliko je pušenje rizično za oboljevanje od raka pluća. U promatranoj grupi ima 100 pušača od kojih 20 ima rak pluća, te 100 nepušača od kojih rak pluća ima njih dvoje.*

Prema ovim podacima vjerojatnost da će pušač oboliti od raka pluća jednaka je 20%, a pripadni izgled će biti jednak

$$odds = \frac{0.2}{1-0.2} = \frac{1}{4} = 0.25. \tag{4.6}$$

Vjerojatnost da će nepušač oboliti od raka pluća jednaka je 2%, a pripadni izgled

$$odds = \frac{0.02}{1-0.02} = \frac{1}{49} = 0.02041. \tag{4.7}$$

Omjer izgleda za obolijevanje od raka pluća kod pušača u odnosu na nepušače sada je jednak

$$OR = \frac{0.25}{0.02041} = 12.25, \tag{4.8}$$

što znači da su izgledi za oboljenje od raka pluća kod pušača oko 12 puta veći u odnosu na nepušače.

Iz ovog primjera možemo zaključiti sljedeće:

- $OR = 1$ - ukazuje da ne postoji povezanost između izloženosti i realizacije događaja,
- $OR > 1$ - ukazuje da izloženost povećava šansu realizacije događaja,
- $OR < 1$ - ukazuje da izloženost smanjuje šansu realizacije događaja.

Pokažimo izračun i interpretaciju omjera izgleda na još jednom primjeru.

Primjer 4.2. *Promatra se skupina od 100 ljudi sa probavnim tegobama koji su primali antibiotsku terapiju. Promatrati ćemo koji utjecaj će imati konzumiranje ili ne konzumiranje probiotika na smanjenje probavnih tegoba. Pritom su dobiveni podaci prikazani u sljedećoj tablici*[1].

Prema podacima iz tablice, vjerojatnost da osoba koja je konzumirala probiotik uslijed probavnih tegoba izazvanih antibiotskom terapijom smanji probavne tegobe je 73%, a pripadni izgled će biti:

$$odds = \frac{0.73}{1-0.73} = \frac{19}{24} = 2.7037. \tag{4.9}$$

[1]Podaci iz tablice nisu rezultat realnog istraživanja, već imaju samo ilustrativan karakter

Tablica 4.1. Podaci o utjecaju probiotika na smanjenje probavnih tegoba pri liječenju antibioticima.

	Smanjili tegobe	Nisu smanjili tegobe
Konzumirali probiotik	38	14
Nisu konzumirali probiotik	21	27

Vjerojatnost da će osoba koja nije konzumirala probiotik smanjiti probavne tegobe biti 44%, a pripadni izgled je:

$$odds = \frac{0.44}{1-0.44} = \frac{21}{79} = 0.7857. \tag{4.10}$$

Omjer izgleda za smanjenje tegoba kod osoba koje su konzumirale probiotik u odnosu na osobe koje nisu jednak je:

$$OR = \frac{2.7037}{0.7857} = 3.44. \tag{4.11}$$

Što znači da izgledi za smanjenje tegoba kod osoba koji su konzumirali probiotik je 3.44 puta veća nego u osoba koje nisu konzumirale.

Promotrimo sada ovaj problem iz drugog kuta. Vjerojatnost da osoba koja je konzumirala probiotik uslijed probavnih tegoba izazvanih antibiotskom terapijom ne smanji probavne tegobe je 27%, a pripadni izgled je:

$$odds = \frac{0.27}{1-0.27} = \frac{27}{73} = 0.3699. \tag{4.12}$$

Vjerojatnost da osoba koja nije konzumirala probiotik ne smanji probavne tegobe bit će 56%,a pripadni izgled je:

$$odds = \frac{0.56}{1-0.56} = \frac{27}{73} = 1.2727. \tag{4.13}$$

Omjer izgleda za ne smanjivanje tegoba kod osoba koje su konzumirale probiotik u odnosu na osobe koje nisu jednak je:

$$OR = \frac{0.3699}{1.2727} = 0.29, \tag{4.14}$$

što znači da se ne uzimanjem probiotika izgled smanjivanja tegoba smanjuje 0.29 puta. Primijetimo također da je

$$0.29 \approx \frac{1}{3.44}. \tag{4.15}$$

4.1. Logistička funkcija i logaritam izgleda

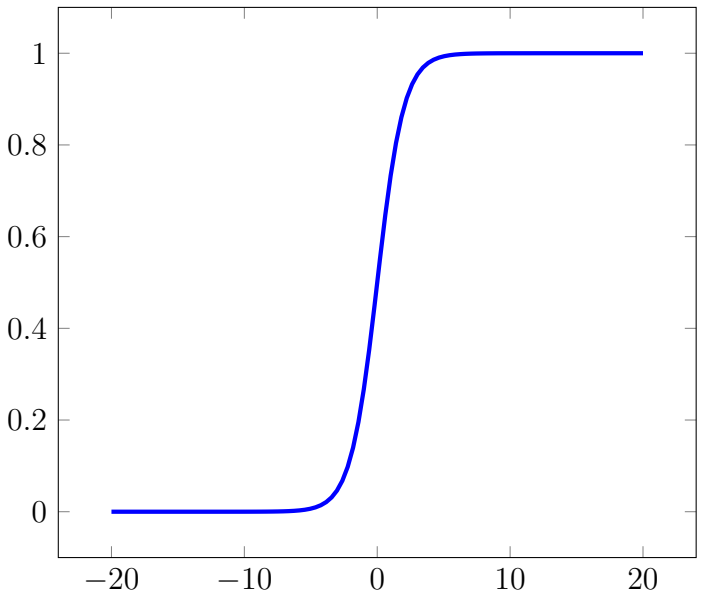
Sigmoidna funkcija često zvana i logistička funkcija ima važnu ulogu u logističkoj regresiji i drugim statističkim analizama, a usko je povezana s izgledima. Njena ključna svrha je transformirati početne vrijednosti tako da budu smještene u rasponu od 0 do 1. Poglavlje je obrađeno prema izvoru : [22].

Definicija 4.2. *Funkciju*

$$f(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}, \tag{4.16}$$

pri čemu su β_0 i β_1 slobodni realni parametri nazivamo logističkom funkcijom.

Karakteristični sigmoidni oblik logističke funkcije prikazan je na sljedećoj slici.



Slika 4.1. Grafički prikaz logističke funkcije za $\beta_0 = 0$ i $\beta_1 = 1$. Izvor: izrada autora

Kako bi odredili inverz logističke funkcije, u njenom definicijskom obliku (4.16) zamijeniti ćemo zavisnu i nezavisnu varijablu, pa imamo:

$$x = \frac{1}{1 + e^{-(\beta_0 + \beta_1 y)}}. \tag{4.17}$$

Slijedi

$$e^{-(\beta_0 + \beta_1 y)} = \frac{1}{x} - 1, \tag{4.18}$$

odnosno

$$e^{-(\beta_0 + \beta_1 y)} = \frac{1 - x}{x}. \tag{4.19}$$

Logaritmiranjem ovog izraza dobivamo

$$-\beta_0 - \beta_1 y = \ln\left(\frac{1 - x}{x}\right), \tag{4.20}$$

odnosno

$$y = \frac{1}{\beta_1} \ln\left(\frac{x}{1 - x}\right) - \frac{\beta_0}{\beta_1}. \tag{4.21}$$

Primijetimo da se u inverzu logističke funkcije pojavljuje izraz

$$\ln\left(\frac{x}{1 - x}\right), \tag{4.22}$$

što ukoliko x označava vjerojatnost predstavlja logaritam izgleda, a ponekad se naziva i logit. Uočimo također da ako je $\beta_0 = 0$, $\beta_1 = 1$ inverz logističke funkcije nije ništa drugo nego logaritam izgleda.

Kako je u inverznoj logističkoj funkciji logaritam izgleda linearno povezan s nezavisnim varijablama, to motivira mogućnost njegovog korištenja kod povezivanja vrijednosti varijabli i vjerojatnosti realizacije nekog događaja, što će i biti zadatak logističke regresije.

5. Logistička regresija

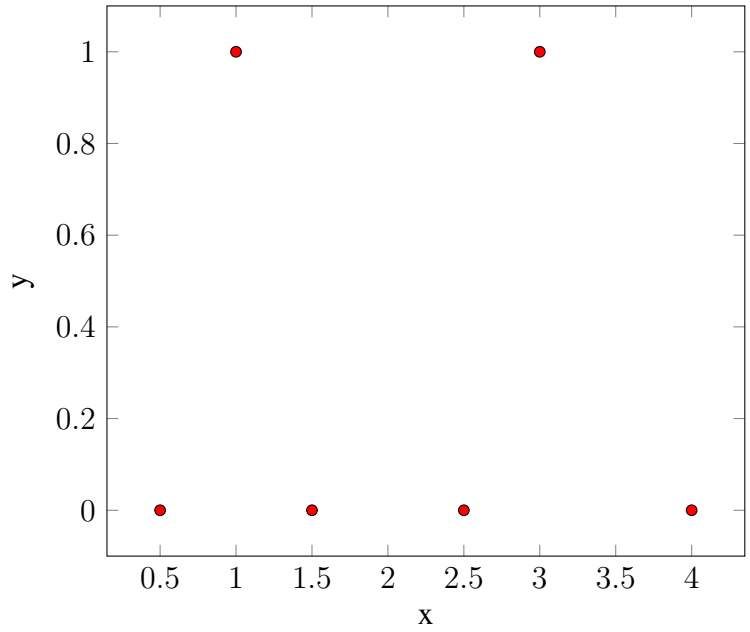
Kod problema logističke regresije analiziramo ulazne podatke u vidu informacije o nezavisnoj varijabli i realizaciji događaja koji nas zanima, dakle u obliku prikazanom u sljedećoj tablici.

Tablica 5.1. Oblik ulaznih podataka kod logističke regresije.

k	x_n	y_n
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
...	...	...
n	x_n	y_n

Vrijednosti zavisne varijable y u tablici poprimaju vrijednosti iz skupa {0,1}, gdje 0 označava da se događaj nije realizirao, a 1 da je.

Iako ovo sugerira na mogućnost korištenja klasične linearne regresije, sa sljedeće je slike jasno da to nema smisla.



Slika 5.1. Dijagram logističke regresije sa nasumičnim koordinatama.

Naime, zavisna varijabla trebala bi biti kontinuirana te se nameće ideja da povežemo nezavisnu varijablu i vjerojatnost realizacije promatranog događaja.

U prošlom smo poglavlju vidjeli da to možemo učiniti pomoću logističke funkcije pa promatramo model oblika:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}. \tag{5.1}$$

Napomenimo da je ovo poglavlje obrađeno prema izvoru: [23]. Kako se u (5.9) ne radi o linearnom modelu, a također ni vjerojatnosti nisu poznate, ne možemo koristiti metodu najmanjih kvadrata. Za procjenu koeficijenata β_0 i β_1 koristiti ćemo metodu najveće vjerodostojnosti (eng. maximum likelihood method).

Svakom uzorku može se pridružiti vrijednost funkcije vjerodostojnosti koja predstavlja vjerojatnost realizacije tog uzorka. Ako se za x_i promatrani događaj realizirao tada je vjerodostojnost od x_i jednaka p_i, a ako se nije realizirao onda je vjerodostojnost $1 - p_i$.

Prema tome, vjerodostojnost našeg uzorka može se opisati funkcijom:

$$f(\beta_0, \beta_1) = \prod_{i,y_i=1}^{n} p(x_i) \cdot \prod_{j,y_j=0}^{n} (1 - p(x_j)), \tag{5.2}$$

a cilj je naći takve vrijednosti β_0 i β_1 za koje je ova funkcija postiže maksimum.

Uvrštavanjem izraza za $p(x)$, logaritmiranjem i dodatnim sređivanjem ovaj se problem može svesti na maksimiziranje funkcije:

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i \ln p(x_i) + (1 - y_i) \ln(1 - p(x_i))\right). \tag{5.3}$$

Jasno je da za rješavanje ovog maksimizacijskog problema ne postoji analitičko rješenje već se do rješenja dolazi numeričkim metodama, stoga je za logističku regresiju nužno korištenje softverskih paketa.

Pozabavimo se sada interpretacijom koeficijenata logističke regresije. Pretpostavit ćemo da se vrijednost nezavisne varijable poveća za 1. Uvrštavanjem u inverznu logističku funkciju dobiva se:

$$x + 1 = \frac{1}{\beta_1} \ln\left(\frac{P_1}{1 - P_1}\right) - \frac{\beta_0}{\beta_1}, \tag{5.4}$$

dok za bazni model vrijedi

$$x = \frac{1}{\beta_1} \ln\left(\frac{P_0}{1 - P_0}\right) - \frac{\beta_0}{\beta_1}. \tag{5.5}$$

Oduzimanjem ovih dviju izraza dobivamo sljedeće:

$$1 = \frac{1}{\beta_1} \left[\ln\left(\frac{P_1}{1 - P_1}\right) - \ln\left(\frac{P_0}{1 - P_0}\right)\right], \tag{5.6}$$

tj.

$$\beta_1 = \ln \frac{\dfrac{P_1}{1 - P_1}}{\dfrac{P_0}{1 - P_0}}. \tag{5.7}$$

Sada je

$$e^{\beta_1} = \frac{\dfrac{P_1}{1 - P_1}}{\dfrac{P_0}{1 - P_0}}, \tag{5.8}$$

ništa drugo nego omjer izgleda koji pokazuje kako se mijenja izgled realizacije ako vrijednost nezavisne varijable povećavamo za 1.

5.1. Univarijatna logistička regresijska analiza pomoću softverskih paketa

U ovom poglavlju baviti ćemo se univarijatnom analizom logističke regresije u softverskom programu JASP. U sljedećem primjeru analizirati ćemo tragediju koja se desila putničkom brodu "Titanic[1]" odnosno koji sve čimbenici vezani uz putnike su utjecali na njihovo preživljavanje. U ovom konkretnom primjeru promatrati ćemo utjecaj dobi na preživljavanje određenog putnika.

Primjer 5.1. *Na slici 5.2 prikazana je univarijatna logistička regresijska analiza u kojoj smo analizirali odnos između nezavisne varijable "Dob" i zavisne varijable "Prezivjeli". Zavisna varijabla je binarnog karaktera, odnosno njene moguće vrijednosti su 0 i 1, pri čemu 0 znači da putnik nije preživio, a 1 da je putnik preživio.*

Logistic Regression ▼

Model Summary - Prezivjeli ▼

Model	Deviance	AIC	BIC	df	X^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2	Cox & Snell R^2
H_0	1186.655	1188.655	1193.447	890						
H_1	1182.207	1186.207	1195.792	889	4.448	0.035	0.004	0.007	0.005	0.005

Coefficients

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	−0.140	0.172	0.869	−0.814	0.662	1	0.416
Dob	−0.011	0.005	0.989	−2.094	4.385	1	0.036

Note. Prezivjeli level '1' coded as class 1.

Slika 5.2. Regresijska analiza u programskom paketu JASP. Izvor: izrada autora.

U prvoj tablici nalaze se metrike koje govore o kvaliteti regresijskog modela. AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria) su pokazatelji koje koristimo ako se između nekoliko modela želimo odlučiti za najbolji, što posebno ima smisla kod multivarijatnih modela. Vrijedi pravilo da je model s nižim AIC-om i BIC-om bolji. U ovom slučaju ne biramo modele pa nam ove dvije metrike nisu potrebne. Istaknimo da pravo razumijevanje te dvije metrike zahtjeva poznavanje teorije informacija, što prelazi okvire ovog rada.

[1]Titanic je bio britanski prekooceanski brod koji je potonuo u sjevernom Atlantskom oceanu u ranim jutarnjim satima 15. travnja 1912. godine, nakon sudara s ledenjakom, za vrijeme svojega prvoga putovanja iz Southamptona u New York.

Iduća tri stupca odnose se na testiranje statističke značajnosti utjecaja nezavisnih varijabli na zavisnu varijablu, što se ovdje radi temeljem Hi-kvadrat testa. Ovo je statističko testiranje ekvivalentno F-testu kod klasične (multi)linearne regresije. Drugim riječima tu nas zanima samo pripadna p-vrijednost koja treba biti manja od 0.05. U ovom slučaju dobila se p-vrijednost od 0.035 što znači da je model statistički značajan.

U zadnja četiri stupca nalaze se četiri pseudo R^2 vrijednosti, čija je uloga identična ulozi klasične R^2 vrijednosti kod (multi)linearne regresije. Treba naglasiti da su ovi parametri kreirani kako bi na neki način imitirali klasični koeficijent determinacije, no njihovo tumačenje nije jednostavno kao u klasičnom slučaju, tj. oni ne govore o postotku objašnjene varijabilnosti i na njihove vrijednosti treba gledati čisto numerički, tj. tražimo da su te vrijednosti čim je moguće veće. Kako nam vrijednosti ovih metrika ne daju nikakvu dodatnu informaciju u odnosu na prethodno objašnjeno testiranje značajnosti u nastavku rada ih nećemo koristiti.

U sljedećoj tablici nalazi se glavni dio ove analize, tj. procjena regresijskih koeficijenata. Kako smo u prethodnim poglavljima objasnili, kod logističke regresije sam koeficijent nije pogodan za tumačenje već se obično tumači omjer izgleda, odnosno vrijednost u stupcu Odds Ratio. Od ostalih parametara u tablici bitan je još Waldov test koji ima istu ulogu kao i t-test te se tumači na isti način. Ovdje, prema tome možemo vidjeti da dob ima statistički značajan utjecaj na preživljavanje.

Sada možemo i komentirati dobivenu vrijednost omjera rizika za dob. Ta je vrijednost manja od 1, što znači da će kod starijih putnika vjerojatnost preživljenja biti manja. Štoviše, ovdje to možemo i dodatno kvantificirati na sljedeći način. Za svaku dodatnu godinu starosti vjerojatnost preživljenja bit će manja za $1/0.989 = 1.011$ puta. Kvalitativno gledano, iako dob ima statistički značajan utjecaj, taj je utjecaj praktički zanemariv.

U sljedećem primjeru promatrati ćemo kako utjecaj previđenog opterećenja utječe cijenu električne energije, odnosno na činjenicu hoće li ona biti pozitivna ili negativna.

Primjer 5.2. *Na slici 5.3 prikazana je univarijatna logistička regresijska analiza u kojoj smo analizirali odnos između nezavisne varijable "predviđeno opterećenje (MW)" i zavisne varijable "day ahead price pozitivna (1) ili negativna (0)". Zavisna varijabla je binarnog karaktera, te ćemo temeljem podataka o predviđenom opterećenju predvidjeti izgled da cijena bude pozitivna. Izvor: [24]*

Iz prve tablice možemo zaključiti temeljem Hi-kvadrat testa da postoji statistički značajna povezanost predviđenog opterećenja i pozitivnosti cijene. U drugoj tablici rezultat Waldova testa pokazuje statističku značajnost utjecaja nezavisne varijable. Iz omjera izgleda možemo zaključiti da se izgled pozitivnosti cijene povećava za 1.001 puta za svaki dodatni MW predviđenog opterećenja.

U sljedećem primjeru promatrati ćemo kakav utjecaj varijabla "Domaće zadaće" ima na zavisnu varijablu "Prolaz" tj. kako domaće zadaće utječu na ostvarenje prolaska na kolegiju. Ova

Logistic Regression

Model Summary - day ahead price pozitivna (1) ili negativna (0)

Model	Deviance	AIC	BIC	df	X^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2	Cox & Snell R^2
H_0	4432.602	4434.602	4443.430	50399						
H_1	4213.278	4217.278	4234.933	50398	219.324	< .001	0.049	0.004	0.004	0.004

Coefficients

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	0.760	0.287	2.138	2.648	7.012	1	0.008
predviđeno opterećenje (MW)	0.001	0.000	1.001	13.614	185.343	1	< .001

Note. day ahead price pozitivna (1) ili negativna (0) level '1' coded as class 1.

Slika 5.3. Regresijska analiza u programskom paketu JASP. Izvor: izrada autora.

analiza je bazirana na stvarnim podacima Tehničkog fakulteta Sveučilišta u Rijeci na slučajno odabranom uzorku studenata iz kolegija Inženjerska matematika ET. Podatke je ustupio nositelj kolegija.

Primjer 5.3. *Na slici 5.4 prikazana je univarijatna logistička regresijska analiza u kojoj smo analizirali odnos između nezavisne varijable "Domaće zadaće" i zavisne varijable "Prolaz". Zavisna varijabla je binarnog karaktera, te ćemo temeljem podataka o domaćim zadaćama predvidjeti šansu za prolaz na matematičkom kolegiju.*

Logistic Regression ▼

Model Summary - Prolaz ▼

Model	Deviance	AIC	BIC	df	X^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2	Cox & Snell R^2
H_0	50.397	52.397	54.475	58						
H_1	39.835	43.835	47.990	57	10.562	0.001	0.210	0.285	0.197	0.164

Coefficients

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	−0.344	0.717	0.709	−0.480	0.230	1	0.631
Domaće zadaće ukupno (10-14)	0.475	0.178	1.608	2.673	7.146	1	0.008

Note. Prolaz level '1' coded as class 1.

Slika 5.4. Regresijska analiza u programskom paketu JASP. Izvor: izrada autora.

U prvoj tablici opažamo da postoji statistički značajna povezanost domaćih zadaća sa prola-

skom kolegija. U drugoj tablici Waldov test ukazuje na statističku značajnost utjecaja nezavisne varijable na način da se izgled prolaska na kolegiju povećava za 1.608 puta za svaki ostvareni bod na domaćim zadaćama.

5.2. Multivarijatna logistička regresija

Multivarijatna ili višestruka logistička regresija je statistička metoda koju koristimo kako bi modelirali vezu između više nezavisnih varijabli i jedne binarne zavisne varijable. Ova metoda se koristi kada želimo predvidjeti vjerojatnost da se dogodi određeni binarni ishod na temelju kombinacije nezavisnih varijabli.

Izražava se kao:

$$p(x_1, x_2, ..., x_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 ... + \beta_n x_n)}}. \tag{5.9}$$

gdje su :

- x_1, x_2, x_n - nezavisne varijable,
- $\beta_0, \beta_1, \beta_2, \beta_n$ - koeficijenti regresije,
- p - vjerojatnost analiziranog događaja.

5.3. Multivarijatna logistička regresijska analiza pomoću softverskih paketa

U ovom poglavlju baviti ćemo se multivarijatnom analizom u softverskom paketu JASP.

U sljedećem primjeru nadovezati ćemo se na primjer 5.2 tako što ćemo dodati još jednu nezavisnu varijablu kako bi prediktirali hoće li cijena električne energije biti pozitivna ili negativna, a ta je varijabla "sat u godini".

Primjer 5.4. *Na slici 5.5 prikazana je multivarijatna logistička regresijska analiza u kojoj smo analizirali odnos između dvije nezavisne varijable "predviđeno opterećenje(MW)", "sat u godini" i zavisne varijable "day ahead price pozitivna(1) ili negativna(0)". Zavisna varijabla je binarnog karaktera te ćemo temeljem podataka o satu u godini i predviđenom opterećenju predvidjeti izgled da cijena bude pozitivna. Izvor: [24]*

U prvoj tablici vidimo temeljen Hi-kvadrat testa da postoji statistički značajna povezanost između nezavisnih varijabli i pozitivnosti cijene. U drugoj tablici iz Waldova testa vidimo statističku značajnost utjecaja nezavisne varijable "predviđeno opterećenje" na način da se izgled pozitivnosti cijene povećava za 1.001 puta za svaki dodatni MW predviđenog opterećenja, dok za nezavisnu varijablu "sat u godini" ne vidimo statistički značajan utjecaj na zavisnu varijablu.

Logistic Regression

Model Summary - day ahead price pozitivna (1) ili negativna (0)

Model	Deviance	AIC	BIC	df	X^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2	Cox & Snell R^2
H_0	4432.602	4434.602	4443.430	50399						
H_1	4213.182	4219.182	4245.665	50397	219.420	< .001	0.050	0.004	0.004	0.004

Coefficients

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	0.803	0.319	2.233	2.516	6.330	1	0.012
predviđeno opterećenje (MW)	0.001	0.000	1.001	13.504	182.347	1	< .001
sat u godini	−0.000	0.000	1.000	−0.310	0.096	1	0.757

Note. day ahead price pozitivna (1) ili negativna (0) level '1' coded as class 1.

Slika 5.5. Regresijska analiza u programskom paketu JASP. Izvor: izrada autora.

U sljedećem primjeru nadovezati ćemo se na primjer 5.3 tako što ćemo dodati još jednu nezavisnu varijablu kako bi prediktirali uspješnost prolaska, a dodatna nezavisna varijabla će nam biti "prisutnost".

Primjer 5.5. *Na slici 5.6 prikazana je multivarijatna logistička regresijska analiza u kojoj smo analizirali odnos između dvije nezavisne varijable "Domaće zadaće" i "Prisutnost" i zavisne varijable "Prolaz". Zavisna varijabla je binarnog karaktera te ćemo temeljem podataka o prisutnosti na nastavi i domaćim zadaćama predvidjeti šansu za prolazak na matematičkom kolegiju.*

Logistic Regression

Model Summary - Prolaz

Model	Deviance	AIC	BIC	df	X^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2	Cox & Snell R^2
H_0	50.397	52.397	54.475	58						
H_1	32.671	38.671	44.903	56	17.726	< .001	0.352	0.452	0.347	0.260

Coefficients

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	−4.817	2.371	0.008	−2.031	4.126	1	0.042
Domaće zadaće ukupno (10-14)	0.195	0.192	1.215	1.016	1.031	1	0.310
Prisutnost (100)	0.074	0.033	1.076	2.247	5.049	1	0.025

Note. Prolaz level '1' coded as class 1.

Slika 5.6. Regresijska analiza u programskom paketu JASP. Izvor: izrada autora.

U prvoj tablici vidimo temeljem Hi-kvadrat testa da promatrani model ima statističku značajnost. U drugoj tablici iz Waldova testa opažamo statističku značajnost utjecaja nezavisne varijable

"Prisutnost" na način da se izgled prolaska na kolegiju poveća za 1.076 puta za svaki ostvareni bod iz prisustva. Također kod nezavisne varijable "Domaće zadaće" ne primjećujemo statistički značajan utjecaj na zavisnu varijablu.

6. Zaključak

U ovome radu bavili smo se regresijskom analizom i njenom primjenom u inženjerstvu. U prvom dijelu obradili smo klasični linearni i multilinearni model gdje smo na primjerima iz struke pokazali korisnost regresijskog modela. Objasnili smo bitne parametre kvalitete regresije te ih komentirali uz provedene analize, a kod provođenja analiza koristili smo se softverskim paketom JASP.

U drugom dijelu rada obradili smo problem logističke regresije. Objasnili smo što je omjer izgleda i koja je njegova povezanost s logističkom regresijom. Logističkoj regresiji prišli smo i kroz kontekst primjene, koristeći se s više primjera iz inženjerske struke, elektrotehnike i obrazovanja.

Zaključujemo da logistička regresija predstavlja snažan statistički alat, često primijenjen za rješavanje izazova klasifikacije i prognoze u raznolikim domenama. Ovaj statistički pristup omogućava konstrukciju modela vjerojatnosti i temelji se na empirijskim podacima kako bi donijeli relevantne zaključke i odluke.

Jedan od pozitivnih aspekata logističke regresije leži u njezinoj relativnoj transparentnosti u tumačenju rezultata. Koeficijenti u modelu pružaju uvid u utjecaj svakog pojedinog prediktora na vjerojatnost pripadanja određenoj kategoriji.

Iako donosi niz koristi, logistička regresija neizbježno nosi sa sobom i svoja ograničenja. Bitno je ispoštovati pretpostavke o linearnosti i neovisnosti prediktora, jer u protivnom model može pružiti netočne rezultate.

Literatura

[1] https://en.wikipedia.org/wiki/Carl_Friedrich_Gauss, s Interneta, 10.6.2023.

[2] https://en.wikipedia.org/wiki/Francis_Galton, s Interneta, 15.6.2023.

[3] https://www.ssla.co.uk/wp-content/uploads/2020/08/regression-in-AI.png, s Interneta, 25.6.2023.

[4] https://en.wikipedia.org/wiki/Mercury-in-glass_thermometer, s Interneta, 5.7.2023.

[5] https://www.math.uh.edu/ irina/MATH1311/Notes1311/1311S44.pdf, s Interneta, 10.7.2023.

[6] Szczepanek, A., https://www.omnicalculator.com/statistics/coefficient-of-determination, s Interneta, 19.7.2023.

[7] https://www.statology.org/wp-content/uploads/2021/07/polycurve1.png, s Interneta, 24.7.2023.

[8] https://en.wikipedia.org/wiki/JASP, s Interneta, 24.7.2023.

[9] https://christophm.github.io/interpretable-ml-book/logistic.html, s Interneta, 29.7.2023.

[10] https://www.hackerearth.com/practice/machine-learning/linear-regression/univariate-linear-regression/tutorial/, s Interneta, 1.8.2023.

[11] https://www.voxco.com/blog/exponential-regression/, s Interneta, 4.8.2023.

[12] https://en.wikipedia.org/wiki/Polynomial_regression, s Interneta, 8.8.2023.

[13] Pant, A., https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb, s Interneta, 8.8.2023.

[14] https://www.datacamp.com/tutorial/multiple-linear-regression-r-tutorial, s Interneta, 10.8.2023.

[15] https://en.wikipedia.org/wiki/Polynomial_regression, s Interneta, 14.8.2023.

[16] https://psychscenehub.com/psychpedia/odds-ratio-2/, s Interneta, 14.8.2023.

[17] https://en.wikipedia.org/wiki/Odds_ratio, s Interneta, 17.8.2023.

[18] Beers, B., https://www.investopedia.com/terms/r/regression.asp, s Interneta, 18.8.2023.

[19] https://priceonomics.com/the-discovery-of-statistical-regression/, s Interneta, 20.8.2023.

[20] Kaw, A., https://resources.saylor.org/wwwresources/archived /site/wp-content/uploads/2011/11/ME205-6.3-TEXT2EXAMPLE.pdf, s Interneta, 20.8.2023

[21] https://regressit.com/data.html, s Interneta, 22.8.2023.

[22] https://mathresearch.utsa.edu/wiki/index.php?title=The_Logistic_Equation, s Interneta, 3.9.2023.

[23] Swaminathan, S., https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc, s Interneta, 3.9.2023.

[24] https://data.open-power-system-data.org/time_series/2018-06-30, s Interneta, 5.9.2023.

Sažetak i ključne riječi

Ovaj se rad bavi regresijskim modelima s naglaskom na model logističke regresije, kako u univarijatnom, tako i u multivarijatnom smislu. Regresijski modeli primjenjivani su na primjere iz inženjerske struke, obrazovanja i elektrotehnike. Definirani su parametri kvalitete regresije koji su analizirani na svim promatranim primjerima. Kod provođenja analiza korišten je softverski paket JASP. U kontekstu logističke regresije, dodatno je objašnjen omjer izgleda, logistička funkcija i logaritam izgleda.

Ključne riječi: regresijski modeli, logistička regresija, omjer izgleda, logaritam izgleda, logistička funkcija

Summary and key words

This paper deals with regression models with emphasis on the logistic regression model, both in the univariate and multivariate sense. The regression models were applied to examples from engineering, education, and electrical engineering. Regression quality parameters were defined and analyzed for all observed examples. The software package JASP was used for the analysis. In the context of logistic regression, the odds ratio, the logistic function and the logarithm of the odds are explained in more detail.

Keywords: regression models, logistic regression, odds ratio, logistic function, log-odds