Jednostavna linearna regresija i polinomijalna regresija - geometrijski pristup

Stošić, Nikolina

Jednostavna linearna regresija i polinomijalna regresija - geometrijski pristup

Stošić, Nikolina

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Nikolina Stošić

JEDNOSTAVNA LINEARNA REGRESIJA I POLINOMIJALNA REGRESIJA-GEOMETRIJSKI PRISTUP

Diplomski rad

Voditelj rada:
doc.dr.sc. Snježana Lubura Strunjak

Zagreb, ožujak 2021.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem svojoj obitelji koja je bila tu za mene tijekom cijelog obrazovnog puta te mi omogućila studiranje. Hvala prijateljima i kolegama na savjetima, ohrabrenjima te vremenu provedenom zajedno koje je uljepšalo ovaj dio života. Također, veliko hvala mentorici, doc.dr.sc. Snježani Luburi Strunjak, na strpljenju, odvojenom vremenu te pomoći pri pisanju ovog rada.

Sadržaj

Poglavlje 1

Uvod

Statistika je grana matematike koja se bavi prikupljanjem, analizom i obradom podataka te njihovom interpretacijom. Upravo ovi postupci omogućavaju nam izradu predviđanja te donošenje zaključaka o svakodnevnom životu na temelju promatranih podataka pa možemo reći kako nam statistika omogućava donošenje zaključaka o budućnosti na temelju postojećih podataka. Danas, zbog velikog razvoja tehnologije, dostupnost podataka je sve veća te su vještine analize i obrade podataka sve važnije. Donošenje pouzdanih i ispravnih zaključaka na temelju promatranih podataka nije trivijalno te je u tu svrhu podatke važno opisati pomoću odgovarajućeg matematičkog modela.
U ovom diplomskom radu promatrat ćemo model jednostavne linearne regresije i model polinomijalne regresije te ćemo im pristupiti nešto drugačije, koristeći geometrijski pristup. U nastavku ovog poglavlja navesti ćemo motivaciju za korištenje geometrijskog pristupa u statistici, reći nešto o regresijskoj analizi te ukratko opisati ideju kako geometrijski pristup funkcionira. U drugom i trećem poglavlju ćemo primijeniti alate iz geometrije i statistike te ilustrirati kako ovaj pristup funkcionira na modelu jednostavne linearne regresije i modelu polinomiijalne regresije, redom. Kao osnova za pisanje ovog rada poslužit će nam knjiga *Statistical Methods: The Geometric Approach*.

1.1 Motivacija za korištenje geometrije u statistici

Geometrija je jedna od najstarijih grana matematike koja se bavi razmatranjem prostora, odnosno njegovih svojstava koja su povezana s udaljenošću, oblikom, veličinom te relativnim položajem likova. Na prvu pomisao čini se kako spajanje dviju ovako kompleksnih matematičkih grana, geometrije i statistike, može samo dodatno otežati. No, ne mora biti tako. Geometrija nam omogućava sagledavanje problema kroz slike koje predstavljaju snažan i lako shvatljiv sažetak problema pa posljedično, nakon malo analiziranja, dolazimo do rješenja problema. Upravo na taj način geometrija pomaže u razjašnjavanju

osnova statistike. Naime, geometrija pomaže premostiti jaz između problema i rješenja jer vizualizacija omogućuje bolje shvaćanje problema i njegovog rješenja nego korištenje niza algebarskih izraza. U ovom diplomskom radu ćemo se poslužiti vektorima koji će nam omogućiti da proces jednostavne linearne regresije i polinomijalne regresije sagledamo na nešto drugačiji način.

1.2 Regresijska analiza

Pojam regresije uveo je, u devetnaestom stoljeću, engleski znanstvenik Sir Francis Galton. Naime, on je istraživao kako visina djece ovisi o visini njihovih roditelja te zaključke napisao u svom djelu *Regression towards Mediocrity in Hereditary Stature*. Konkretno, zaključio je kako će djeca čiji su roditelji iznadprosječno visoki također biti iznadproječno visoka, no u usporedbi s generacijom, njihova visina neće toliko odstupati od prosjeka. Drugim riječima, opisao je kako visine potomaka visokih predaka imaju tendenciju približavanja prema prosječnoj visini te se riječ regresija od tada zadržala. Danas, u statistici, pojam regresije koristi se kao mjera povezanosti varijable koju želimo opisati i varijabli koje koristimo u svrhu njenog opisivanja. Dakle, regresija je metoda koja proučava ovisnost između varijabli te predstavlja najčešće korištenu statističku metodu. Kako ovisnost među varijablama može biti linearna i nelinearna, u skladu s time imamo i model linerane regresije, odnosno nelinearne regresije. Kada god je to moguće, nelinerne zavisnosti nastoje se linearizirati transformiranjem podataka kako bi se koristio model linearne regresije. Ukoliko nakon provođenja transformacija veza između transformiranih varijabli i dalje nije linearna, koriste se nelinearni modeli.
Kao što smo rekli, model linearne regresije pretpostavlja linearnu vezu među promatranim podacima te ćemo se u nastavku fokusirati upravo na takve modele. Varijable čije ponašanje želimo opisati modelom nazivamo zavisnim varijablama ili varijablama odziva. S druge strane, varijable koju koristimo kako bismo opisali ponašanje drugih varijabli nazivamo nezavisnim, eksplanatornim ili prediktorskim varijablama. U ovisnosti o tome koliko eksplanatornih varijabli koristimo prilikom opisivanja varijable odziva imamo sljedeću podjelu linearne regresije:

- jedna eksplanatorna varijabla
  - Jednostavna linearna regresija
  - Polinomijalna linearna regresija
- više eksplanatornih varijabli
  - Višestruka linearna regresija

Uočimo kako smo ovdje promatrali slučaj kada imamo samo jednu varijablu odziva. Spomenimo kako je moguć i slučaj kada imamo više varijabli odziva te tada govorimo o multivarijantnoj regresiji, no taj slučaj u ovom diplomskom radu nećemo razmatrati.

1.3 Kako geometrijski pristup funkcionira?

Prilikom korištenja geometrijskog pristupa vrijednosti zavisnih, odnosno nezavisnih varijabli zapisujemo u vektorskom obliku. Uključivanje vektora u proces modeliranja otvara nove mogućnosti pristupa predviđanju ponašanja nezavisne varijable na temelju opažanja zavisnih varijabli. Naime, korištenje vektora omogućava nam da iskoristimo alate i zakonitosti iz linearne algebre te koristeći njih na nešto neuobičajeniji način pristupimo procesu modeliranja. Vođeni time, kada se opredijelimo za model koji ćemo koristiti kako bismo modelirali ponašanje varijable odziva, možemo definirati vektorski prostor kojeg ćemo u nastavku nazivati prostor modela te koristiti oznaku M.

Definicija 1.3.1. *Neka je V neprazan skup i $\mathbb{F}$ polje. Uređena trojka $(V, +, \cdot)$ gdje je*

- *$+$ binarna operacija na V takva da je $(V, +)$ Abelova grupa*
- *$\cdot : \mathbb{F} \times V \to V$ preslikavanje sa svojstvima*

  1. $\alpha \cdot (\beta \cdot a) = (\alpha \cdot \beta) \cdot a$,
  2. $(\alpha + \beta) \cdot a = \alpha \cdot a + \beta \cdot a$
  3. $\alpha \cdot (a + b) = \alpha \cdot a + \alpha \cdot b$
  4. $1 \cdot a = a$, *za sve* $\alpha, \beta \in \mathbb{F}, a, b \in V$,

*naziva se **vektorski prostor nad poljem** $\mathbb{F}$. Elementi skupa V nazivaju se **vektori**, a elementi polja $\mathbb{F}$ **skalari**. Tradicionalno, operaciju $+$ zovemo zbrajanje vektora, a preslikavanje $\cdot$ zovemo množenje vektora skalarom.*

Dakle, prostor modela je prostor koji sadrži sve vektore modela, odnosno vektore čije su komponentne očekivane vrijednosti zavisne varijable za danu vrijednost nezavisne varijable. Dodatno, prostor M je i unitaran prostor.

Definicija 1.3.2. *Neka je V vektorski prostor nad poljem $\mathbb{F}$, pri čemu je $\mathbb{F}$ polje $\mathbb{R}$ ili $\mathbb{C}$. Preslikavanje $s : V \times V \to \mathbb{F}$ koje svakom uređenom paru vektora pridružuje skalar $s(a, b) = \langle a|b\rangle \in \mathbb{F}$ naziva se **skalarno množenje** na prostoru V ako su ispunjena sljedeća svojstva:*

1. $\langle a|a\rangle \geq 0$, *za sve* $a \in V$, *pri čemu je* $\langle a|a\rangle = 0$ *akko je* $a = 0_V$
2. $\langle a|b\rangle = \overline{\langle b|a\rangle}$ *za sve* $a, b \in V$
3. $\langle \lambda a|b\rangle = \lambda\langle a|b\rangle$ *za sve* $a, b \in V, \lambda \in \mathbb{F}$
4. $\langle a + b|c\rangle = \langle a|c\rangle + \langle b|c\rangle$ *za sve* $a, b, c \in V$

Skalar $\langle a|b\rangle$ *se zove skalarni produkt ili umnožak vektora a i b. Uređeni par* $(V, \langle\cdot|\cdot\rangle)$ *nazivamo* ***unitarni prostor nad poljem*** $\mathbb{F}$.

U našem slučaju biti će $\mathbb{F} = \mathbb{R}$ pa tada govorimo o realnom vektorskom prostoru te za $x = (x_1, \ldots, x_n)$ i $y = (y_1, \ldots, y_n)$ iz $\mathbb{R}^n$ definiramo

$$\langle x|y\rangle = x_1y_1 + \ldots x_ny_n = \sum_{i=1}^{n} x_iy_i$$

Može se pokazati da su za tako definirano preslikavanje zadovoljena svojstva skalarnog množenja iz gornje definicije pa slijedi da je $\mathbb{R}^n$ realan unitaran prostor. U nastavku ćemo koristiti i pojam norme pa navodimo sljedeću propoziciju.

Propozicija 1.3.3. *Neka je* $(V, \langle\cdot|\cdot\rangle)$ *unitarni prostor. Preslikavanje*

$$a \to \sqrt{\langle a|a\rangle}$$

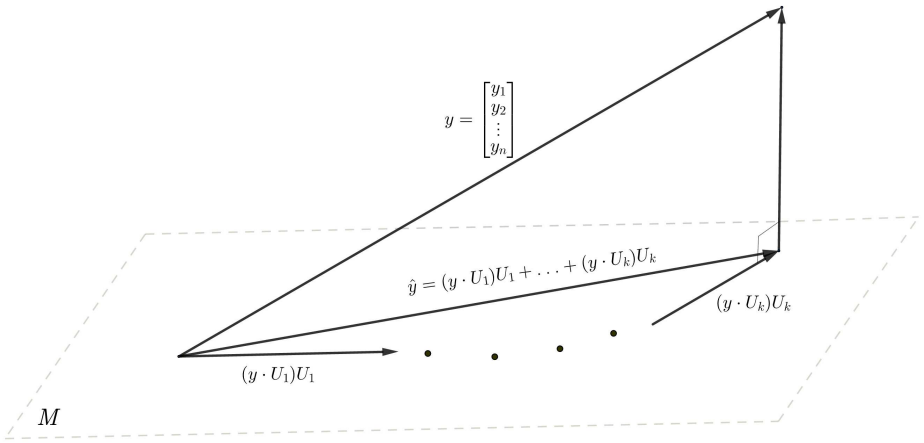
s prostora V u polje $\mathbb{R}$ *je norma na prostoru V.*

Za ovako definirano preslikavanje u nastavku ćemo koristiti oznaku $\|\cdot\|$. Nakon navedenih definicija prirodno se nameće pitanje kako odrediti dimenziju prostora M te njegovu bazu, odnosno skup linearno nezavisnih vektora koji ga razapinju. U tu svrhu pretpostavimo da imamo n opažanja zavisne varijable te da je odgovarajući vektor opažanja $y = [y_1, \ldots, y_n]^T$. Tada možemo zaključiti da je M zasigurno potprostor $n-$ dimennzionalnog prostora. Nadalje, u ovisnosti o promatranom modelu i broju parametara dolazimo i do dimenzije prostora M. Pretpostavimo sada da je M $k-$ dimenzionalan potprostor čija je baza skup $\{U_1, \ldots, U_k\}$ sačinjen od jediničnih i međusobno ortogonalnih vektora. Jasno, navedenu bazu možemo dopuniti do baze čitavog prostora pa slijedi da je skup $\{U_{k+1}, \ldots, U_n\}$ ortonormirana baza prostora kojeg ćemo nazivati prostor greške.
Prilikom provođenja statističke analize od krucijalne važnosti biti će mogućnost projiciranja vektora opažanja y na definirani prostor modela M te ćemo upravo na taj način odrediti vektor modelom prilagođenih vrijednosti $\hat{y}$ pa u tu svrhu navodimo sljedeću propoziciju.

Propozicija 1.3.4. *Neka je L potprostor končanodimenzionalnog unitarnog prostora V i neka je* $\{e_1, \ldots, e_l\}$ *ortonormirana baza tog potprostora L. Ako je x bilo koji vektor prostora V, njegova ortogonalna projekcija* x_L *na potprostor L određena je s*

$$x_L = \sum_{i=1}^{l} \langle x|e_i\rangle e_i$$

Odabir modelom prilagođenog vektora $\hat{y}$ kao ortogonalne projekcije vektora y na prostor M je smislen zato što je u tom slučaju $\hat{y}$ takav vektor iz M za koji vrijedi da je njegova udaljenost od vektora y minimalna. Spomenimo ovdje kako ćemo u nastavku rada za skalarno množenje umjesto oznake $\langle\cdot|\cdot\rangle$ koristiti oznaku $\cdot$ kojom smo na početku definirali množenje vektora skalarom, no naglasimo kako su množenje vektora skalarom i skalarno množenje operacije koje su slične jedino imenom.



Slika 1.1: Prilagodba modela projiciranjem vektora y na prostor M

Dodatno, može se pokazati da je duljina vektora $\hat{y}$ tada jednaka korijenu sume kvadrata projekcija vektora y na vektore $U_1, \ldots, U_k$, što je ekvivalentno sljedećoj jednakosti:

$$\|\hat{y}\|^2 = (y \cdot U_1)^2 + \ldots + (y \cdot U_k)^2$$

Navedenu zakonitnost nazivamo Pitagorin teorem te ćemo ju, između ostalog, koristiti prilikom izračuna testne statistike. Upravo to dovodi nas do pitanja kako ćemo, koristeći ovaj pristup, testirati hipoteze koje će nam biti od interesa. U regresijskoj analizi, kojom ćemo se u nastavku baviti, hipoteze koje ćemo testirati pretpostavljati će činjenicu da je neki parametar iz modela jedank nuli. U tu svrhu biti će potrebno odrediti vektor smjera hipoteze i testnu statistiku. Testna statistika biti će omjer kvadrata duljine projekcije vektora y na

smjer hipoteze i prosjeka kvadrata duljina projekcija na prostor greške. Preostaje objasniti kako ćemo odrediti vektor smjera hipoteze. Odgovarajući vektor smjera određen je činjenicom da se nalazi u bazi prostora M te da očekivana vrijednost projekcije vektora y na taj vektor mora biti jednaka testiranom parametru pomnoženom nekim skalarom.
Prije prelaska na sljedeće poglavlje, dotaknimo se ukratko i statističke strane cijelog procesa. Prije svega, navedimo dvije pretpostavke koje ćemo i u nastavku rada koristiti.

- $Y_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$
- Y_i i Y_j su nezavisne slučajne varijable $\forall i, j$

Ponovno kao i ranije, neka je naš vektor opažanja $y = [y_1, \ldots, y_n]^T$ te neka je $U = [u_1, \ldots, u_n]^T$ jedinični vektor u čijem smjeru želimo projicirati vektor y. Svaku vrijednost $y_i, i, \ldots, n$ doživljavamo kao jednu realizaciju slučajne varijable Y_i. Kako u geometrijskom pristupu promatramo ortogonalne projekcije, duljinu ortogonalne projekcije

$$y \cdot U = u_1 y_1 + \ldots + u_n y_n$$

doživljavamo kao jednu realizaciju slučajne varijable

$$Y \cdot U = u_1 Y_1 + \ldots + u_n Y_n$$

Napomenimo kako bi preciznije bilo reći da je $\mid y \cdot U \mid$ duljina ortogonalne projekcije jer gornji izraz može biti negativan. Koristeći da je $U_1, \ldots, U_n$ ortonormirana baza $n-$ dimenzionalnog prostora, možemo reći kako originalni set nezavisnih varijabli Y_i transformiramo u novi set normalnih, nezavisnih slučajnih varijabli $Y \cdot U_i$ te je samo upitno s kojim parametrima distribucije. U tu svrhu, koristeći svojstva očekivanja i varijance te nezavisnost slučajnih varijabli računamo:

$$\mathbb{E}[Y \cdot U] = u_1 \mathbb{E}[Y_1] + \ldots + u_n \mathbb{E}[Y_n]$$

$$\begin{aligned} \mathrm{Var}[Y \cdot U] &= \mathrm{Var}(u_1 Y_1 + \ldots + u_n Y_n) \\ &= u_1^2 \mathrm{Var}(Y_1) + \ldots + u_n^2 \mathrm{Var}(Y_n) \\ &= u_1^2 \sigma^2 + \ldots + u_n^2 \sigma^2 \\ &= \sigma^2 (u_1^2 + \ldots + u_n^2) \\ &= \sigma^2 \end{aligned}$$

Zaključujemo kako će varijanca "novih" varijabli uvijek biti σ^2 dok će očekivanje varirati. Uzimanjem u obzir pretpostavke da je skup $\{U_1, \ldots, U_k\}$ ortonormirana baza za prostor modela slijedi da slučajne varijable $Y \cdot U_i, i = 1, \ldots, n$ možemo podijeliti u sljedeće dvije kategorije.

- $Y \cdot U_1, \ldots, Y \cdot U_k \Rightarrow$ povezane s prostorom modela
- $Y \cdot U_{k+1}, \ldots, Y \cdot U_n \Rightarrow$ povezane s prostorom greške

Varijable povezane s prostorom modela će većinom imati očekivanje različito od nule, dok će varijable povezane s prostorom greške imati očekivanje nula. Varijable povezane s prostorom greške biti će korištene u procjeni varijance σ^2 čija je procjena od velike važnosti u provođenju statističke analize. Nepristrani procjenitelj nekog parametra, u ovom slučaju varijance, je slučajna varijabla čija je očekivana vrijednost jednaka upravo tom parametru. Na temelju toga možemo zaključiti kako postoji puno izbora za nepristranog procjenitelja varijance. Naime, po definiciji varijance za varijable povezane s prostorom greške slijedi

$$\begin{aligned} \mathrm{Var}[Y \cdot U_i] &= \mathbb{E}[(Y \cdot U_i - \mathbb{E}[Y \cdot U_i])^2] \\ &= \mathbb{E}[(Y \cdot U_i)^2] \end{aligned}$$

gdje drugi redak slijedi zbog $\mathbb{E}[Y \cdot U_i]) = 0, i = k+1, \ldots, n$. Iz navedenoga slijedi $\mathbb{E}[(Y \cdot U_i)^2] = \sigma^2$ pa iz toga zaključujemo da je $(Y \cdot U_i)^2$ nepristrani procjenitelj varijance. Ovu činjenicu koristimo kako bismo došli do najbolje procjene, odnosno do nepristranog procjenitelja varijance s najmanjom varijancom koji glasi

$$s^2 = \frac{(Y \cdot U_{k+1})^2 + \ldots + (Y \cdot U_n)^2}{n-k}$$

Ovakva procjena varijance poslužit će nam i pri konstrukciji traženih pouzdanih intervala što ćemo kroz primjere u narednim poglavljima pokazati.

Poglavlje 2

Geometrijski pristup jednostavnoj linearnoj regresiji

2.1 Jednostavna linearna regresija

Jednostavna linearna regresija statistička je metoda koja se koristi kada želimo modelirati odnos između dvije varijable, odnosno donjeti zaključke o ponašanju jedne, na temelju ponašanja druge varijable. Pridjev jednostavna u nazivu ove široko primjenjivane statističke metode dolazi upravo od činjenice kako koristimo samo jednu varijablu kojom želimo opisati drugu. Konkretno, jednostavna linearna regresija daje nam informaciju koliko varijacije varijable odziva Y je opisano eksplanatornom varijablom X. Linearna veza između varijable odziva Y i eksplanatorne varijable X znači da vrijedi

$$Y = \beta_0 + \beta_1 \cdot X$$

Kako u praksi nećemo naići na podatke koje je moguće savršeno opisati pravcem, odnosno među podacima će neizostavno postojati šum, javljat će se greška prilikom aproksimacije te stoga model jednostavne linearne regresije glasi

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i, \forall i$$

gdje je

- y_i vrijednost varijable odziva za i-ti podatak
- x_i vrijednost eksplanatorne varijable za i-ti podatak
- ϵ_i vrijednost greške (koju ne možemo opaziti) za i-ti podatak
- β_0 nepoznati parametar, parametar presjeka
- β_1 nepoznati parametar, parametar nagiba

Pretpostavke modela

Ono što u modelu zasad pretpostavljamo jest kako je linearna veza između odzivne i eksplanatorne varijable razumna. Kako bismo mogli donjeti statističke zaključke moramo uvesti pretpostavke na grešku koja predstavlja odstupanja odzivne varijable od prilagođenog pravca. Dakle, za grešku ϵ pretpostavljamo da je slučajna varijabla sa sljedećim svojstvima:

- $\mathbb{E}[\epsilon] = 0$
- $\mathbf{Var}[\epsilon] = \sigma^2$
- ϵ_i su nezavisne slučajne varijable, $Cov(\epsilon_i, \epsilon_j) = 0, \quad \forall i, j$
- Greška ϵ je normalno distribuirana, $\epsilon \sim N(0,\sigma^2)$

Sada iz ovih pretpostavki slijedi

$$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 \cdot x$$

Dodatno pretpostavljamo da je za Y normalno distriubirana s konstatntom varijancom σ^2 za poznate vrijednosti. Dakle, za dano x vrijedi

$$Y \sim N(\beta_0 + \beta_1 \cdot x, \sigma^2)$$

pri čemu parametre β_0 i β_1 procjenjujemo iz podataka metododom najmanjih kvadrata.

2.2 Ilustracija geometrijskog pristupa

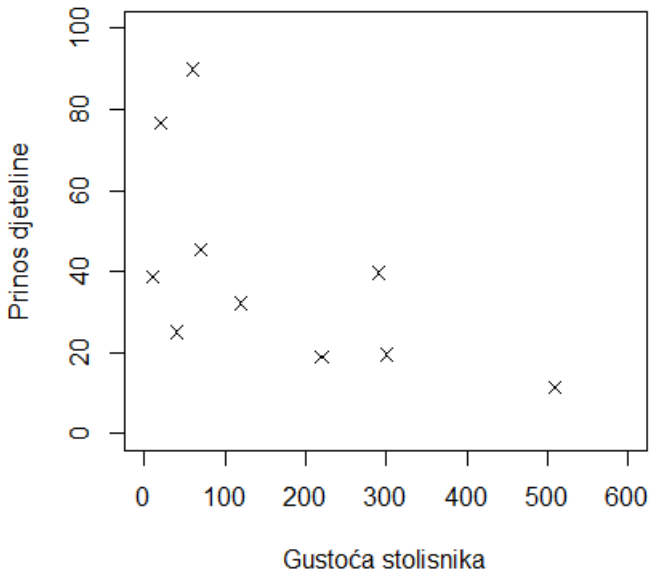
Krenimo sada s jednostavnim primjerom koji će nam dati zorniji uvid u metodu jednostavne linearne regresije koristeći geometrijski pristup.

Podaci

Za početak, promatramo dvije biljke, poljoprivredni korov stolisnik, *Achillea millefolium* i bijelu djetelinu, *Trifolium repens*. Zanima nas na koji način rasprostranjenost stolisnika na nekom usjevu djeteline utječe na urod djeteline. Intuitivno, ono što očekujemo jest kako će veća gustoća korova utjecati na manji prinos zasađene bijele djeteline. Prvo, potrebno je prikupiti podatke. U ovom slučaju, podaci su prikupljeni tako da su na većem usjevu djeteline na slučajan način odabrana tri različita dijela površine 0.1 kvadratnih metara te potom na svakom od njih ubrana i izvagana količina sjemena djeteline te izbrojen broj stabljiki stolisnika. Od tako prikupljenih nekoliko uzoraka biramo jedan s jednakom vjerojatnošću. Sada navodimo podatke s kojima ćemo raditi u nastavku, poredane od onih s najmanjom prema onima s najvećom gustoćom korova.

Broj stabljika cvijeta stolisnika po m^2	Prinos sjemena djeteline u g/m^2
10	38.8
20	76.7
40	25.1
60	89.9
70	45.3
120	32.2
220	19.0
290	39.7
300	19.5
510	11.4

Kako bismo stekli uvid u podatke prvo ćemo ih grafički prikazati na dvodimenzionalnom grafu.



Slika 2.1: Graf rasipanja prinosa bijele djeteline u odnosu na gustoću stolisnika

Podatke o broju stabljiki cvijeta stolisnika koje koristimo za predviđanje prinosa sje-

mena djeteline ćemo centralizirati, odnosno oduzeti im srednju vrijednost kako bismo postigli da očekivanje prediktorske varijable bude nula. Uočimo, centraliziranje neće dovesti do promjene vrijednosti parametra nagiba, ali će olakšati interpretaciju parametra presjeka. Naime, iz izraza

$$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 \cdot x$$

slijedi da je parametar presjeka očekivana vrijednost varijable odziva kada je vrijednost prediktora jednaka nula. Ukoliko prediktorska varijabla ne poprima vrijednost nula, primjerice nosi podatke o visini ili težini, interpretacija nema smisla. Centraliziranjem postižemo

$$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 \cdot (x - \bar{x})$$

gdje je $\bar{x}$ srednja vrijednost, pa parametar presjeka postaje očekivana vrijednost varijable odziva kada je prediktor jednak svojoj srednjoj vrijednosti.

Model jednostavne linearne regresije

Kako bismo krenuli s primjenom modela jednostavne linearne regresije za početak ćemo pretpostaviti da su sve pretpostavke modela zadovoljene. Kako ćemo koristiti geometrijski pristup, naše podatke zapisujemo u vektorskom obliku, odnosno

$$y = \begin{bmatrix} 38.8 \\ 76.7 \\ \vdots \\ 11.4 \end{bmatrix}, \; x - \bar{x} = \begin{bmatrix} 10 - 164 \\ 20 - 164 \\ \vdots \\ 510 - 164 \end{bmatrix}$$

gdje je

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_{10}}{10} = 164$$

Slijedi da je vektor očekivanih vrijednosti $\mathbb{E}[Y|X = x]$ jednak

$$\beta_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} 10 - 164 \\ 20 - 164 \\ \vdots \\ 510 - 164 \end{bmatrix}$$

Dakle, linearna ljuska prostora M razapetog našim modelom je

$$\left\{ \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \begin{bmatrix} -154 \\ -144 \\ \vdots \\ 346 \end{bmatrix} \right\}$$

Uočimo, centraliziranje prediktora osiguralo nam je ortogonalnost navedenih vektora. Prije nego projiciramo vektor y na potprostor M u svrhu pronalska njemu najbližeg prilagođenog vektora preostaje još svesti ih na jedinične. Provođenjem navedenog dobivamo da je naš prostor M razapet vektorima

$$U_1 = \frac{1}{\sqrt{10}} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \; \mathrm{U}_2 = \frac{1}{\sqrt{238640}} \begin{bmatrix} -154 \\ -144 \\ \vdots \\ 346 \end{bmatrix}$$

gdje je $\sqrt{238640} = \| (x - \bar{x})\| = \sqrt{\sum_{i=1}^{10} (x_i - \bar{x})^2}$.

Određivanje smjera hipoteze od interesa

Prisjetimo se, ono što želimo testirati koristeći model jest utječe li gustoća korova na urod djeteline. Upravo parametar nagiba otkriva nam kakva je, pozitivna ili negativna, te postoji li uopće veza među promatranim podacima pa želimo testirati je li on jednak ili različit od nule. Kako bismo proveli testiranje, pritom koristeći geometrijski pristup, potrebno je odrediti smjer nulte hipoteze, odnosno odrediti jedinični vektor U koji se može zapisati kao linearna kombinacija vektora U_1 i U_2 te za koji vrijedi $\mathbb{E}[\mathrm{y} \cdot U] = \mathrm{k} \cdot \beta_1$, pri čemu je k konstanta. U tu svrhu računamo

$$y \cdot U_2 = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{10} \end{bmatrix} \frac{1}{\| (x - \bar{x})\|} \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_{10} - \bar{x} \end{bmatrix}$$

$$= \frac{1}{\| (x - \bar{x})\|} \left[y_1(x_1 - \bar{x}) + y_2(x_2 - \bar{x}) + \ldots + y_{10}(x_{10} - \bar{x}) \right]$$

Uzimanjem očekivanih vrijednosti dobivamo

$$\frac{1}{\| (x - \bar{x})\|} \sum_{i=1}^{10} \left[\beta_0 + \beta_1 (x_i - \bar{x}) \right] (x_i - \bar{x})$$

pa slijedi

$$\mathbb{E}[Y \cdot U_2] = \frac{1}{\| (x - \bar{x})\|}[\beta_0 \underbrace{(x_1 - \bar{x} + x_2 - \bar{x} + \ldots + x_{10} - \bar{x})}_{=0} + \beta_1 \sum_{i=1}^{10} (x_i - \bar{x})^2] \quad = \beta_1 \| (x - \bar{x})\|$$

Iz dobivenog izraza zaključujemo da vektor U_2 zadovoljava ranije navedene pretpostavke uz $k = \| (x - \bar{x})\|$ te je on vektor smjera nulte hipoteze.

Određivanje modelom prilagođene vrijednosti

Nakon što smo odredili prostor M koji sadrži sve mogućnosti za prilagođeni vektor $\hat{y}$ te odredili smjer hipoteze, prelazimo na ključne dijelove u analizi podataka. Prvo, prilagođeni vektor dobiven modelom je vektor

$$\hat{y} = (y \cdot U_1)U_1 + (y \cdot U_2)U_2$$

što slijedi iz svojstava ortogonalne projekcije vektora. U promatranom primjeru uvrštavanjem dolazimo do

$$\hat{y} = (y \cdot U_1)\frac{1}{\sqrt{10}}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} + (y \cdot U_2)\frac{1}{\| (x - \bar{x})\|}\begin{bmatrix}-154\\-144\\\vdots\\364\end{bmatrix}$$

odnosno

$$\hat{y} = \hat{\beta}_0\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} + \hat{\beta}_1\begin{bmatrix}-154\\-144\\\vdots\\364\end{bmatrix}$$

Sada jasno slijedi

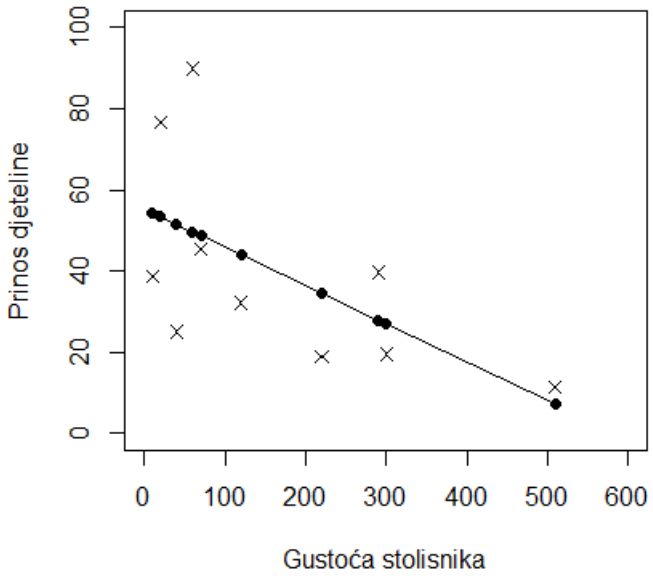
$$\hat{\beta}_0 = \frac{y \cdot U_1}{\sqrt{n}} = \frac{\frac{1}{\sqrt{10}}\sum_{i=1}^{10} y_i}{\sqrt{10}} = \bar{y} = \frac{397.6}{10} = 39.76$$

$$\hat{\beta}_1 = \frac{y \cdot U_2}{\| (x - \bar{x})\|} = \frac{\sum_{i=1}^{10} y_i(x_i - \bar{x})}{\| (x - \bar{x})\| \, \| (x - \bar{x})\|} = \frac{\sum_{i=1}^{10} y_i(x_i - \bar{x})}{\sum_{i=1}^{10} (x_i - \bar{x})^2} = \frac{-22494.4}{238640} = -0.0943$$

gdje su $\hat{\beta}_0$ i $\hat{\beta}_1$ procjene parametara β_0 i β_1 metodom najmanjih kvadrata. Uvrštavanjem dobivamo da je prilagođeni vektor dobiven modelom

$$\hat{y} = \begin{bmatrix}54.3\\53.3\\\vdots\\7.1\end{bmatrix}$$

Iskoristimo sada prednost ovog modela te podatke prikažimo na dvodimenzionalnom grafu zajedno s dobivenim pravcem $\hat{y} = 39.76 - 0.0943 \cdot (x - 164)$.



Slika 2.2: Regresijski pravac s prilagođenim vrijednostima (•) te podacima (×)

Sljedeće što ćemo učiniti jest odrediti vektor reziduala, odnosno vektor greške. Kako je vektor greške jednak razlici između stvarnih podataka i prilagođenih vrijednosti imamo

$$e = y - \hat{y}$$

$$e = y - (y \cdot U_1)U_1 - (y \cdot U_2)U_2$$

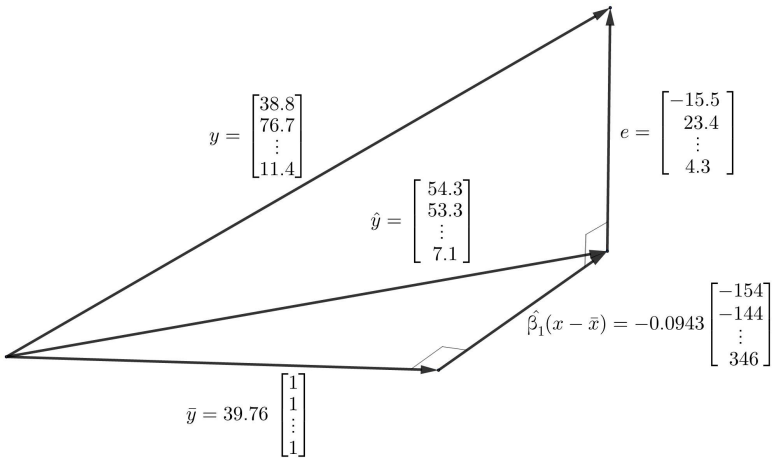
$$e = \begin{bmatrix} 38.8 \\ 76.7 \\ \vdots \\ 11.4 \end{bmatrix} - 39.76 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + 0.0943 \begin{bmatrix} -154 \\ -144 \\ \vdots \\ 346 \end{bmatrix} = \begin{bmatrix} -15.5 \\ 23.4 \\ \vdots \\ 4.3 \end{bmatrix}$$

Konačno, prilagođen model zapisan u cjelosti glasi

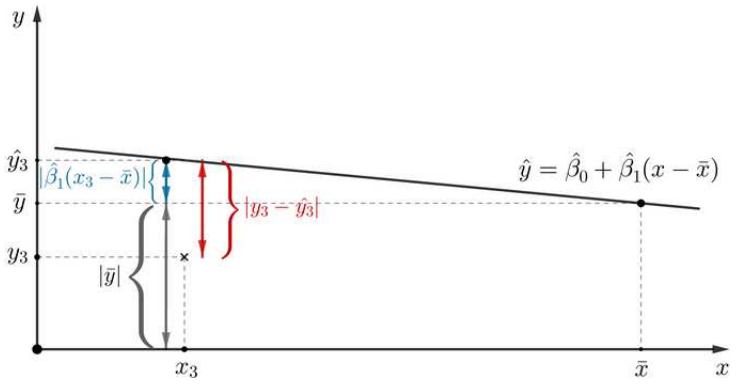
$$y = \hat{\beta}_0 + \hat{\beta}_1(x - \bar{x}) + e$$

$$\begin{bmatrix} 38.8 \\ 76.7 \\ \vdots \\ 11.4 \end{bmatrix} = 39.76 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} - 0.0943 \begin{bmatrix} -154 \\ -144 \\ \vdots \\ 364 \end{bmatrix} + \begin{bmatrix} -15.5 \\ 23.4 \\ \vdots \\ 4.3 \end{bmatrix}$$

Ortogonalna dekompozicija

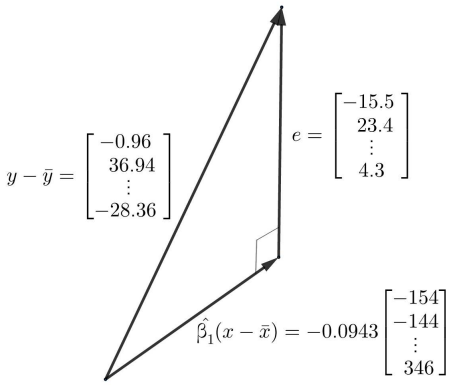


Slika 2.3: Ortogonalna dekompozicija vektora y



Slika 2.4: Ortogonalna dekompozicija prikazana u (x, y) ravnini za jedno opažanje $(x_3, y_3) = (40, 25.1)$

Češće se promatra nešto jednostavnija ortogonalna dekompozicija, dekompozicija vektora $y - \hat{y}$, koja direktno slijedi pa ćemo i nju grafički prikazati.



Slika 2.5: Pojednostavljena ortogonalna dekompozicija vektora $y - \bar{y}$

Iz Slike 2.5 jasno je da primjenom Pitagorinog teorema dobivamo

$$\hat{\beta}_1^{\,2} \,\|\, (x - \bar{x})\|^2 + \|y - \hat{y}\|^2 = \|\, y - \bar{y}\|^2$$

što je, kako smo pokazali, ekvivalentno

$$(y \cdot U_2)^2 = \|\, y - \bar{y}\|^2 - \|y - \hat{y}\|^2$$

odnosno vrijedi

$$(y \cdot U_2)^2 = \sum_{i=1}^{10}(y_i - \bar{y})^2 - \sum_{i=1}^{10}(y_i - \hat{y})^2 \tag{2.1}$$

S obzirom kako je prostor našeg modela dvodimenzionalan potprostor prostora dimenzije deset slijedi da je prostor greške dimenzije osam. Vektore koji čine bazu prostora greške radi jednostavnosti nećemo navoditi, no njihova egzistencija je neupitna. Kako od svake baze možemo doći do jedinične i ortogonalne, u nastavku će nam $U_3, \ldots, \mathrm{U}_{10}$ predstavljati upravo takve vektore koji razapinju prostor greške. U skladu s time imamo da zapis vektora greške u toj ortonormiranoj bazi glasi

$$\| e\|^2 = \sum_{i=1}^{10}(y_i - \hat{y})^2 = (y \cdot U_3)^2 + \ldots + (y \cdot U_{10})^2 \tag{2.2}$$

Vektore $U_3, \ldots, \mathrm{U}_{10}$ koristimo za procjenu varijance

$$s^2 = \frac{\| e\|^2}{8} = 463.458$$

Testiranje hipoteze

Kako bismo donjeli zaključke o podacima potrebno je provesti odgovarajući statistički test. Mi želimo provjeriti je li paramater β_1 jednak ili različit od nule, odnosno zaključiti je li model jednostavne linearne regresije značajan. Ukoliko je $\beta_1 = 0$ model nije značajan što znači da linearna veza nije dobar način za opisati vezu između podataka. Da bismo to provjerili koristit ćemo test prihvatljivosti koji uspoređuje naš model s modelom u kojem preostaje samo parametar presjeka te nema prediktora. Dakle, testiramo sljedeće:

- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

Testna statistika je

$$F = \frac{\sum_{i=1}^{10}(y_i - \bar{y})^2 - \sum_{i=1}^{10}(y_i - \hat{y})^2}{(n-2) - (n-1)} \div \frac{\sum_{i=1}^{10}(y_i - \hat{y})^2}{n-2} \overset{H_0}{\sim} F(1, n-2)$$

gdje su $n-2$ i $n-1$ stupnjevi slobode potpunog i reduciranog modela, redom. Intuitivno, od n podataka imamo n stupnjeva slobode, no u ovisnosti o tome koliko parametara procjenjujemo toliko stupnjeva slobode gubimo. Konkretno, u modelu jednostavne linearne regresije procjenjujemo dva parametra pa imamo $n-2$ stupnjeva slobode. Koristeći 2.1 i 2.2 te $n = 10$ slijedi

$$F = \frac{(y \cdot U_2)^2}{[(y \cdot U_3)^2 + \ldots + (y \cdot U_{10})^2]/8}$$

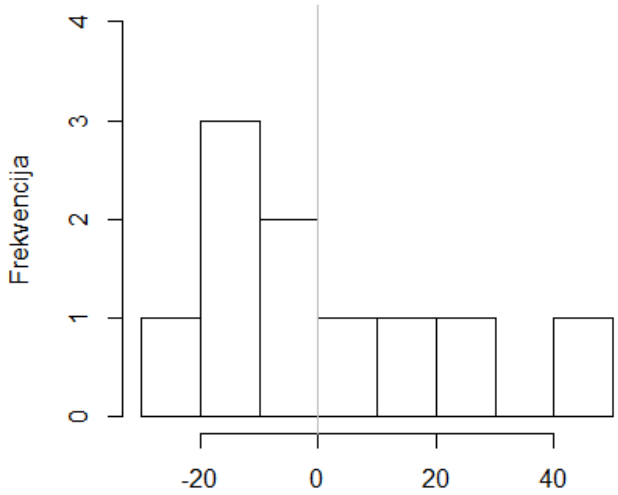
Time smo dobili da je testna statistika koju mi koristimo u geometrijskom pristupu zaista ekvivalentna klasičnom izrazu. Uvrštavanjem odgovarajućih vrijednosti imamo

$$F = \frac{(y \cdot U_2)^2}{\| e\|^2 /8} = \frac{2120.3}{463.5} = 4.57$$

Kako je 0.95 - kvantil $F_{1,8}$ distribucije jednak 5.32, ne možemo odbaciti nultu hipotezu na razini značajnosti 5%. Vrijednost $F_{1,8}$ kvantila iščitavamo iz Tablice A.1 u prilogu i to na način da se spomenuti kvantil nalazi na presjeku prvog retka i osmog stupca te ćemo odgovarajuće kvantile i u nastavku uzimati iz spomenute tablice.

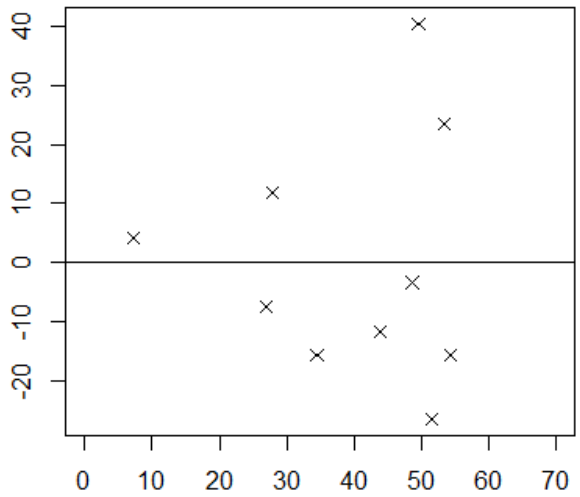
Provjera pretpostavki modela

Pretpostavke koje smo ranije naveli sada moramo provjeriti. Kako bismo ispitali pretpostavku linearnosti podatke prikažemo na dvodimenzionalnom grafu te pogledamo je li razumno koristiti regresijski pravac za opis veze među njima. Sa Slike 2.2 možemo vidjeti kako su naši podaci slučajno raspoređeni te ne primjećujemo nikakav uzorak po kojem se ponašaju. S obzirom da ne uočavamo ništa neobično, zaključujemo kako je ova pretpostavka ispunjena. Napomenimo ovdje da ukoliko je pretpostavka linearnosti neispunjena u daljnjem računu to neće predstavljati problem, no zaključci koje donesemo neće imati smisla.
Zatim, ukoliko su ranije navedene pretpostavke na greške ϵ istinite, u tom bi se slučaju dobivena procijenjena greška e trebala ponašati slično. Dakle, za provjeru normalnosti prikazat ćemo histogram ranije dobivenih reziduala.



Slika 2.6: Histogram opaženih reziduala

Kao ni kod linearnosti, ni ovdje ne uočavamo ništa neobično, odnosno ne uočavamo narušavanje normalnosti, pa uzimamo i ovu pretpostavku kao razumnu. Nezavisnost ne možemo provjeriti jer nemamo dovoljno podataka, no čak i ukoliko ona nije ispunjena, to neće imati značajan utjecaj. Preostaje provjeriti homoskedastičnost grešaka te ćemo u tu svrhu prikazati reziduale u odnosu na procjenjene vrijednosti.



Slika 2.7: Graf reziduala u odnosu na procjenjene vrijednosti $\hat{y}$

Sa Slike 2.7 se čini kako se varijanca reziduala povećava kako raste $\hat{y}$ i to narušava pretpostavku konstantne varijance grešaka pa ta pretpostavka u ovom slučaju nije razumna. Postoji nekoliko načina kako možemo pristupiti rješavanju ovog problema. Jedan od načina, kojeg ćemo mi sada iskoristiti, jest transformacija modela.

Transformacija modela

U početnom modelu originalnim podacima smo prilagođavali pravac, a u transformiranom modelu ćemo umjesto y korisiti logaritimirane podatke $\log y$ te njima prilagođavati regresijski pravac. Analogno kao ranije procijenimo parametre.

$$\hat{\beta_0} = \frac{\log y \cdot U_1}{\sqrt{10}} = 3.504$$

$$\hat{\beta_1} = \frac{\log y \cdot U_2}{\| (x - \bar{x})\|} = -0.0029$$

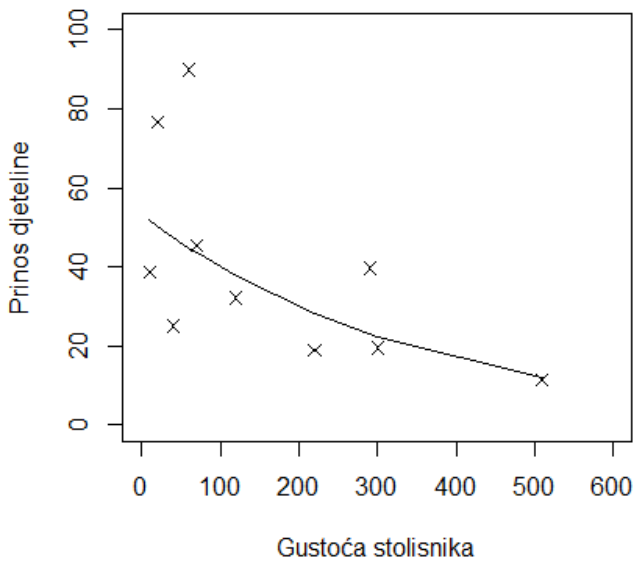
Sada imamo

$$\hat{\log y} = 3.504 - 0.0029 \cdot (x - 164)$$

što je ekvivalentno s

$$\hat{y} = e^{3.504-0.0029(x-164)}$$

odnosno, prilagodba regresijskog pravca podacima $(x, \log y)$ ekvivalenta je prilagodbi eksponencijalne krivulje podacima (x, y).



Slika 2.8: Procijenjena eksponencijalna krivulja veze prinosa djeteline i gustoće korova

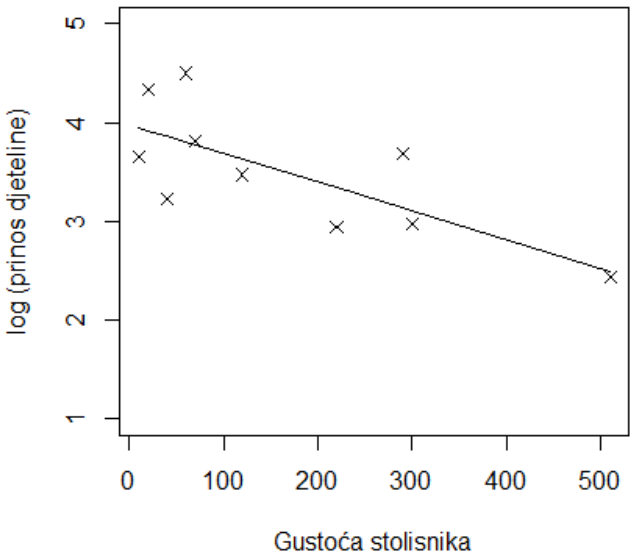
Ponovno provodimo test:

- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

Testna statistika je sada

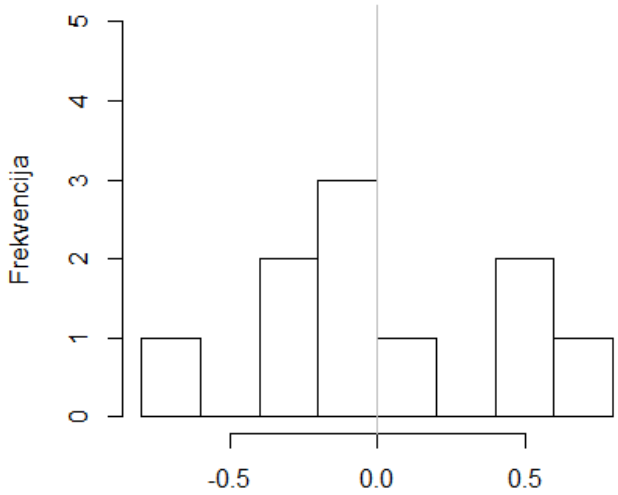
$$F = \frac{(\log y \cdot U_2)^2}{[(\log y \cdot U_3)^2 + \ldots + (\log y \cdot U_{10})^2]/8} = 9.71$$

pa možemo odbaciti nultu hipotezu na razini značajnosti 5%. Sada je još samo potrebno ponovno provjeriti pretpostavke. Sa Slike 2.9 vidimo da je linearnost zadovoljena.



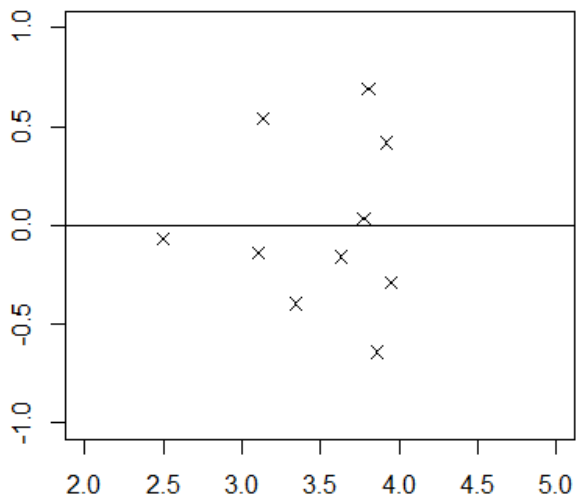
Slika 2.9: Prikaz logarimiranih podataka o prinosu djeteline u ovisnosti o gustoći korova

Zatim, pogledom na graf reziduala u odnosu na procjenjene vrijednosti, kao i ranije, ne uočavamo ništa problematično.



Slika 2.10: Histogram reziduala

Posljednje provjeravamo ranije narušenu pretpostavku homoskedastičnosti. Sa Slike 2.11 je jasno da to sada nije slučaj te je i ova pretpostavka sada zadovoljena.



Slika 2.11: Graf reziduala u odnosu na procjenjene vrijednosti $\hat{logy}$

Napomenimo ovdje kako ćemo u nastavku poglavlja raditi s ovim modelom te ćemo umjesto $\log y$ i $\hat{logy}$, radi jednostavnosti, koristiti oznake y i $\hat{y}$.

Pouzdani intervali

Pouzdani interval za parametar nagiba β_1

Cilj nam je odrediti $(1-\alpha)\%$ pouzdani interval za parametar nagiba β_1. Kako smo ranije pokazali

$$\mathbb{E}[Y \cdot U_2] = \beta_1 \| (x - \bar{x})\|$$

$$\text{Var}[Y \cdot U_2] = \sigma^2$$

Također, vrijedi da $y \cdot U_2 = \hat{\beta}_1 \| (x - \bar{x})\|$ dolazi iz normalne distribucije s očekivanjem $\beta_1 \| x - \bar{x}\|$ i varijancom σ^2 pa imamo

$$(\hat{\beta}_1 - \beta_1) \| x - \bar{x}\| \sim N(0, \sigma^2)$$

S obzirom da je varijanca nepoznata, slijedi

$$\frac{(\hat{\beta_1} - \beta_1) \parallel x - \bar{x} \parallel}{\sqrt{s^2}} \sim t_{n-2}$$

odnosno promatrani izraz ima t-distribuciju s $n-2$ stupnjeva slobode što nas dovodi do

$$t_{n-2}\left(\frac{\alpha}{2}\right) \leq \frac{(\hat{\beta_1} - \beta_1) \parallel x - \bar{x} \parallel}{s} \leq t_{n-2}\left(1 - \frac{\alpha}{2}\right)$$

Nadalje, vrijedi i

$$\hat{\beta_1} \sim N\left(\beta_1, \frac{\sigma^2}{\parallel x - \bar{x} \parallel^{\,2}}\right)$$

pa standardnu grešku nepristranog procjenitelja za β_1 možemo procijeniti s

$$\frac{s}{\parallel x - \bar{x} \parallel}$$

te ona u našem konkretnom primjeru iznosi -0.0009. Iz svega navedenog slijedi da je $(1-\alpha)\%$ pouzdani interval za parametar nagiba β_1 oblika

$$\hat{\beta_1} - 0.0009 \cdot t_{n-2}\left(\frac{\alpha}{2}\right) \leq \beta_1 \leq \hat{\beta_1} + 0.0009 \cdot t_{n-2}\left(1 - \frac{\alpha}{2}\right)$$

Na razini značajnosti $\alpha = 5\%$ iz našeg primjera uvrštavanjem dobivamo

$$-0.0029 - 0.0009 \cdot t_8(0.975) \leq \beta_1 \leq -0.0029 + 0.0009 \cdot t_8(0.025)$$

iz čega slijedi da je 95% pouzdani interval za β_1

$$[-0.0051, -0.0008]$$

Konačno, ukoliko promotrimo dobiveni pouzdani interval možemo vidjeti kako se nula ne nalazi u njemu te je ta činjenica u skladu s ranije dobivenim rezultatom testiranja. Kako bismo mogli donjeti značajnije zaključke o odnosu promatranih podataka trebali bismo ovakav postupak ponoviti puno puta na različitim podacima prikupljenim u različitim trenucima te na različitim mjestima. Također, spomenimo kako nam procjenjena standardna greška ne otkriva informaciju kako bi se parametar nagiba ponašao u tim mjerenjima već se u njoj samo krije podatak za naš konkretan slučaj.

Pouzdani interval za očekivanu vrijednost $\mathbb{E}[Y|x = x_0]$

Za neku poznatu vrijednost od x, primjerice x_0, modelom procjenjena očekivana vrijednost $\mathbb{E}[\widehat{Y|x = x_0}]$ jednaka je $\hat{\beta}_0 + \hat{\beta}_1 \cdot (x - x_0)$. Njena varijanca je

$$\begin{aligned}\text{Var}[\hat{\beta}_0 + \hat{\beta}_1 \cdot (x - x_0)] &= \text{Var}(\hat{\beta}_0) + (x_0 - x)^2 \text{Var}(\hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + (x_0 - x)^2 \frac{\sigma^2}{\| x - \bar{x}\|^2} \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - x)^2}{\| x - \bar{x}\|^2} \right]\end{aligned}$$

gdje drugi redak slijedi iz nezavisnosti nepristranih procjenitelja $\hat{\beta}_0$ i $\hat{\beta}_1$ za parametre β_0 i β_1, redom. Uočimo, varijanca će biti najmanja za vrijednosti najbliže srednjoj vrijednosti $\bar{x}$. Iz gornjeg izraza direktno slijedi da je procjenjena standardna greška za $\hat{y}$

$$\sqrt{s^2 \left[\frac{1}{n} + \frac{(x_0 - x)^2}{\| x - \bar{x}\|^2} \right]}$$

95% pouzdani intervala za procjenjenu vrijednost u $x = x_0$ je

$$\hat{\beta}_0 + \hat{\beta}_1 (x_0 - \bar{x}) \pm t_{n-2}(0.975) \sqrt{s^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\| x - \bar{x}\|^2} \right]}$$

Uvrštavanjem promatranih podataka slijedi

$$3.504 - 0.0029 (x_0 - 164) \pm 2.306 \sqrt{0.2068 \left[\frac{1}{10} + \frac{(x_0 - 164)^2}{238640} \right]}$$

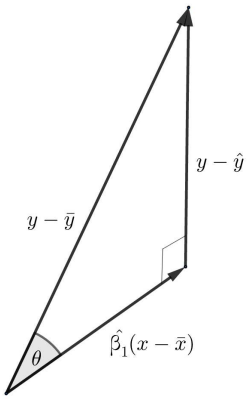
Naglasimo na kraju kako ne možemo tvrditi da $(1 - \alpha)100\%$ pouzdani intervali sadrže nepoznati parametar s vjerojatnošću da $(1 - \alpha)$ već jedino možemo reći da kada bismo za puno uzoraka izračunali pouzdani interval tada bi $(1 - \alpha)100\%$ pouzdanih intervala sadržavalo taj nepoznati parametar.

Koeficijent korelacije

U statistici, kako bismo odredili koliko je snažna veza među varijablama koristimo koeficijent korelacije. Postoji nekoliko različitih koeficijenata korelacije no najčešće korišten je Pearsonov koeficijent determinacije, oznaka r. Pearsonov koeficijent mjeri smjer te jačinu linearne veze među varijablama te ćemo u nastavku njega koristiti. Intuitivno, u jednostavnoj linearnoj regresiji on je zapravo pokazatelj koliko su promatrani podaci blizu regresijskog pravca, odnosno kolika je greška e. Koeficijent korelacije može poprimiti vrijednosti u rasponu od -1 do 1 pri čemu:

- -1 ukazuje na snažnu negativnu povezanost što znači da s povećanjem jedne varijable dolazi do smanjenja druge varijable
- 0 predstavlja slučaj kad linearna veze među podacima ne postoji
- 1 ukazuje na savršenu pozitivnu povezanost, odnosno povećanje jedne varijable vodi povećanju druge varijable

Napomenimo da se u prvom i trećem slučaju podaci nalaze upravo na regresijskom pravcu. U svrhu daljnjeg računanja koeficijenta r, prisjetimo se pojednostavljene ortogonalne dekompozicije vektora $y - \bar{y}$.



Slika 2.12: Ortogonalna dekompozicija vektora y-$\bar{y}$

Sa Slike 2.12 je jasno da ukoliko je duljina vektora greške manja da će kut θ biti manji. Dakle, manji kut θ sugerira veći koeficijent korelacije koji može poprimiti vrijednosti između -1 i 1. Ovi zaključci navode nas na uspostavljanje veze između koeficijenta korelacije među podacima te kosinusa kuta θ. Naime, koeficijent r upravo je definiran kao kosinus kuta između vektora $(y - \bar{y})$ i $(x - \bar{x})$ pri čemu $\hat{\beta_1}$ izostavljamo jer on naravno ne utječe na iznos kuta. Iz izraza za skalarni produkt imamo

$$(x - \bar{x}) \cdot (y - \bar{y}) = \| x - \bar{x}\| \cdot \| y - \bar{y}\| \cdot cos\,(\theta)$$

iz čega slijedi

$$cos\,(\theta) = \frac{(x - \bar{x})}{\| x - \bar{x}\|} \cdot \frac{(y - \bar{y})}{\| y - \bar{y}\|}$$

U našem primjeru koeficijent korelacije između logaritmiranih podataka o prinosu djeteline i gustoće korova je

$$cos\,(\theta) = \frac{-692.3}{\sqrt{238640} \cdot \sqrt{3.6627}} = -0.740$$

Dobivena vrijednost ukazuje na negativnu, no ne pretjerano snažnu vezu među podacima te odgovara kosinusu kuta $\theta = 138°$. Možemo reći kako podaci o gustoći korova pomažu u procjeni prinosa djeteline, no prinos djeteline nije u potpunosti određen gustoćom korova. Koeficijent korelacije r je ustvari procjenitelj za stvarni koeficijent korelacije ρ koji je jednak

$$\rho = \frac{Cov(X, Y)}{\sigma_x \cdot \sigma_y}$$

Koeficijent determinacije

Nakon što su sve pretpostavke modela ispunjene uobičajeno je pitanje koliko je model uopće dobar. Odgovor na to pitanje krije se u koeficijentu determinacije. Naime, upravo on pokazuje koliko je varijabilnosti varijable odziva opisano modelom te ga računamo koristeći izraz

$$r^2 = \frac{\| \hat{y} - \bar{y} \|^2}{\| y - \bar{y} \|^2}$$

Sa Slike 2.12 jasno je da vrijedi

$$\cos^2(\theta) = \frac{\left\| \hat{\beta}_1 \cdot (x - \hat{x}) \right\|^2}{\| (y - \hat{y}) \|^2} = \frac{\| y - \bar{y} \|^2 - \| y - \hat{y} \|^2}{\| y - \bar{y} \|^2}$$

gdje je brojnik izraza jednak razlici ukupne varijabilnosti i varijabilnosti koja nije opisana modelom. Dakle, r^2 je omjer varijabilnosti od y opisane modelom i ukupne varijabilnosti te je jednak kvadratu ranije računatog koeficijenta determinacije. U našem primjeru izračunom dobivamo

$$r^2 = (-0.740)^2 = 0.548$$

Iz svega navedenog jasno je da r^2 može poprimiti vrijednosti između 0 i 1, pri čemu veća vrijednost znači da je prilagodba modela bolja.

2.3 Rekapitulacija modela

Kako bismo napravili kratku rekapitulaciju navedenoga, ponovimo ukratko cijeli postupak na drugim, simuliranim podacima. Podatke ćemo simulirati tako da za poznate vrijednosti x generiramo y takve da vrijedi

$$Y \sim N(60 + 15 \cdot x, 400)$$

Za svaku vrijednost od x, simulirat ćemo 25 vrijednosti y te na slučajan način za svaku vrijednost od x odabrati jednu vrijednost y gdje će nam x predstavljati količinu proizvedenih proizvoda u tvornici, a y ukupan prihod od proizvodnje. Za simuliranje podataka koristimo statistički paket R te dobivene podatke navodimo u tablici.

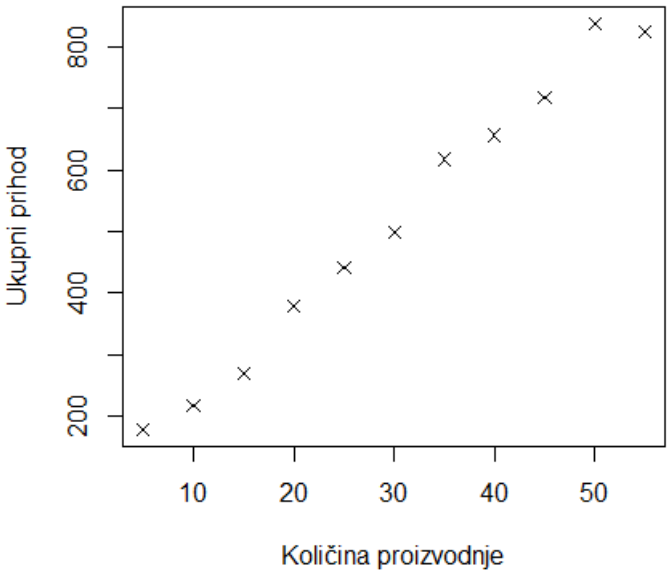
Ukupni troškovi proizvodnje y za količinu proizvodnje x										
$x = 5$	$x = 10$	$x = 15$	$x = 20$	$x = 25$	$x = 30$	$x = 35$	$x = 40$	$x = 45$	$x = 55$	$x = 60$
128	195	274	344	430	538	592	642	727	810	826
154	214	276	333	474	515	568	655	749	804	766
170	211	297	351	450	544	587	673	717	798	810
149	199	275	395	441	537	583	662	732	798	816
144	200	269	363	393	493	570	692	722	823	793
129	175	266	388	430	530	606	640	751	799	788
139	222	295	349	440	521	590	665	734	811	847
158	221	261	322	473	499	583	666	738	837	783
175	208	253	370	419	527	573	656	746	828	824
159	217	288	348	454	532	586	662	744	800	815
179	193	270	336	422	479	606	656	707	848	831
117	232	236	382	444	505	578	685	742	793	824
117	239	306	333	426	538	543	661	727	816	791
77	200	298	363	430	528	620	647	726	872	820
160	231	312	344	401	492	571	667	761	777	797
149	218	298	371	442	482	617	631	742	789	835
118	189	284	359	428	516	621	670	714	787	797
146	218	275	382	414	516	600	653	729	775	786
144	205	284	360	436	517	563	631	749	804	772
164	181	286	324	446	503	582	649	749	793	814
117	206	290	357	381	512	549	659	736	817	803
135	190	296	384	436	535	575	653	708	822	824
111	232	272	380	430	510	604	655	726	832	790
138	200	286	337	446	508	592	672	770	832	846
134	197	279	343	430	499	603	679	717	824	824

Tablica 2.1: Uzorci 25 normalno distribuiranih vrijednosti s očekivanjem $\mathbb{E}[Y] = 60+15\cdot x$ i varijancom $\sigma^2 = 400$ za svaku vrijednost x

Sada u svakom stupcu Tablice 2.1 izaberemo po jedan podatak na slučajan način, odnosno tako da za svaku vrijednost u stupcu postoji vjerojatnost $\frac{1}{25}$ da baš nju izaberemo. Na taj

način dolazimo do sljedećih podataka s kojima ćemo raditi u nastavku koje odmah grafički prikazujemo na Slici 2.13..

Količina proizvodnje	Ukupni prihod
5	179
10	218
15	269
20	380
25	441
30	449
35	617
40	656
45	717
50	837
55	824



Slika 2.13: Graf rasipanja ukupnog prihoda u odnosu na količinu proizvodnje

Kako je model jednostavne linearne regresije u ortogonalnom zapisu

$$y = \beta_0 + \beta_1 \cdot (x - \bar{x}) + \epsilon$$

uvrštavanjem opaženih podataka slijedi

$$y = \beta_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} 5 - 30 \\ 10 - 30 \\ \vdots \\ 55 - 30 \end{bmatrix}$$

Prostor M u kojem će se modelom prilagođena vrijednost nalaziti razapet je ortogonalnim i jedinčnim vektorima

$$U_1 = \frac{1}{\sqrt{11}} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \; \mathrm{U}_2 = \frac{1}{\sqrt{2750}} \begin{bmatrix} -25 \\ -20 \\ \vdots \\ 25 \end{bmatrix}$$

Procijenimo parametre β_0 i β_1

$$\hat{\beta_0} = \frac{y \cdot U_1}{\sqrt{n}} = \frac{1699.619}{\sqrt{11}} = 512.4545$$

$$\hat{\beta_1} = \frac{y \cdot U_2}{\| (x - \bar{x})\|} = \frac{741.1265}{52.44044} = 14.1327$$

Napravimo kratki osvrt na zasad dobivene procjene. Sjetimo se, za podatke koje smo simulirali vrijedi $\mathbb{E}[Y] = 60{+}15x$. Uzimanjem u obzir činjenice da u promatranom primjeru vrijedi $\bar{x} = 30$, iz jednakosti

$$60 + 15 \cdot x = \beta_0 + \beta_1 \cdot (x - \bar{x})$$

slijedi

$$\mathbb{E}[Y] = 510 + 15 \cdot (x - \bar{x})$$

Dakle, možemo zaključiti kako će nakon puno ponavljanja na različitim, slučajno izabranim podacima iz Tablice 2.1, prosjek svih procjena parametara β_0 i β_1 težiti u 510 i 15, redom. Uvrštavanjem procjenjenih vrijednosti dobivamo sljedeći vektor prilagođenih vrijednosti

$$\hat{y} = 512.4545 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + 14.1327 \begin{bmatrix} -25 \\ -20 \\ \vdots \\ 25 \end{bmatrix} = \begin{bmatrix} 159.1364 \\ 229.8 \\ \vdots \\ 865.7727 \end{bmatrix}$$

Model zapisan u cjelosti glasi

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_1 1 \end{bmatrix} = \hat{\beta}_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \hat{\beta}_1 \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_{11} - \bar{x} \end{bmatrix} + \begin{bmatrix} y_1 - \hat{y} \\ y_2 - \hat{y} \\ \vdots \\ y_{11} - \hat{y} \end{bmatrix}$$

$$\begin{bmatrix} 179 \\ 218 \\ \vdots \\ 284 \end{bmatrix} = 512.4545 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + 14.1327 \begin{bmatrix} -25 \\ -20 \\ \vdots \\ 25 \end{bmatrix} + \begin{bmatrix} 19.8636 \\ -11.8 \\ \vdots \\ -41.7727 \end{bmatrix}$$

Kada nam je poznat vektor greške možemo procijeniti varijancu

$$s^2 = \frac{\| (y - \hat{y})\|^2}{n-2} = \frac{6492.282}{9} = 721.3646$$

Kao i kod procjene parametara, nakon puno ponavljanja, prosjek procjena svih varijanci će težiti u 400, iako individualne procjene imaju nešto veća odstupanja.
Računamo testnu statistku kako bismo testrali je li β_1 jednak ili različit od nule te dobivamo

$$F = \frac{(y \cdot U_2)^2}{[(y \cdot U_3)^2 + \ldots + (y \cdot U_{11})^2]/9} = \frac{549268.4}{721.3646} = 761.4297$$

Kako 0.95 - kvantil $F_{1,9}$ distribucije iznosi 5.12, možemo odbaciti nultu hipotezu na razini značajnosti 5%

Navedimo sada pouzdane intervale za β_1 i $\mathbb{E}[Y|x = x_0]$.

- 95% pouzdani interval za β_1

$$14.1327 - 1.4767 \cdot t_9(0.975) \leq \beta_1 \leq 14.1327 + 1.4767 \cdot t_9(0.975)$$

$$12.9517 \leq \beta_1 \leq 15.3138$$

- 95% pouzdani interval za $\mathbb{E}[Y|x = x_0]$

$$512.4545 + 14.1327\,(x_0 - 30) \pm t_9(0.975) \sqrt{721.3646 \left[\frac{1}{9} + \frac{(x_0 - 30)^2}{\| x - 30\|^2}\right]}$$

Koeficijent korelacije jednak je 0.9941, a koeficijent determinacije 0.9883.
Svi dobiveni rezultati u skladu su sa činjenicom da su podaci simulirani upravo tako da među njima postoji linearna zavisnost.

Poglavlje 3

Geometrijski pristup polinomijalnoj regresiji

3.1 Polinomijalna regresija

Jednostavna linearna regresija, kojom smo se bavili u prošlom poglavlju, temelj je za razumijevanje ostalih, kompliciranijih modela. Međutim, pretpostavljanje linearne veze između zavisne i nezavisne varijable, što odgovara pronalasku pravca koji najbolje opisuje podatke, dosta je strogo te vrlo često nije dovoljno jer ukoliko veza između varijabli i postoji ona ne mora biti linearna. Kao i kod jednostavne linearne regresije, u polinomijalnoj regresiji također promatramo jednu nezavisnu i jednu zavisnu varijablu. No, za razliku od jednostavne linearne regresije, vezu između varijabli ne opisujemo isključivo pravcem, već u obzir dolaze i krivulje višeg reda. Dakle, polinomijalna linearna regresija je regresijska analiza u kojoj je veza između nezavisne varijable Y i zavisne varijable X opisana polinomom n-tog stupnja pa stoga model glasi

$$y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot x_i^2 + \cdots + \beta_n \cdot x_i^n + \epsilon_i, \forall i$$

gdje je

- y_i vrijednost varijable odziva za i-ti podatak
- x_i vrijednost eksplanatorne varijable za i-ti podatak
- ϵ_i vrijednost greške (koju ne možemo opaziti) za i-ti podatak
- $\beta_0, \beta_1, \ldots, \beta_n$ nepoznati parametri
- $n \in \mathbb{N}_0$, stupanj polinomijalne regresije

Pitanje koje se može javiti jest zašto se javlja riječ linearna u nazivu ovog regresijskog modela s obzirom da se u njemu javljaju potencije. Međutim, odgovor na to pitanje krije se u tome što kada kažemo linearna ustvari mislimo na linearnost u parametarima te činjenicu da nezavisnu varijablu prikazujemo kao linearnu kombinaciju potencija zavisne varijable. Sljedeće što se možemo zapitati jest kako odrediti optimalan stupanj polinoma n. Naime, potrebno je odabrati n takav da u modelu ne nedostaje parametara, odnosno da su podaci adekvatno opisani, a opet s druge strane, da ne dođe do korištenja previše parametara, što dovodi do toga da model jako dobro opisuje određeni skup podataka pa možda neće odgovarati drugom setu podataka ili neće pouzdano predvijeti buduća opažanja. Za takav n, koji nam govori koliko potencija zavisne varijable ćemo uključiti, imamo ograničenje da ne može biti veći od broja opažanja s kojima radimo. Osim toga, ne postoji drugo ograničenje te različiti statističari zagovaraju različite metode njegova određivanja.

Pretpostavke modela

U modelu polinomijalne regresije reda n zasad pretpostavljamo kako je veza između odzivne i eksplanatorne varijable opisana polinomom stupnja n, odnosno podatke opisujemo krivuljom n-tog reda. Primjerice, u slučaju $n = 2$ dobivamo kvadratnu jednadžbu te vezu među podacima opisujemo parabolom. Kao i u prvom poglavlju, uvodimo sljedeće pretpostavke na grešku ϵ koja predstavlja razliku procijenjene vrijednosti odzivne varijable od njene očekivane vrijednosti:

- $\mathbb{E}[\epsilon] = 0$
- $\mathbf{Var}[\epsilon] = \sigma^2$
- ϵ_i su nezavisne slučajne varijable, $Cov(\epsilon_i, \epsilon_j) = 0, \quad \forall i, j$
- ϵ je normalno distribuirana, $\epsilon \sim N(0,\sigma^2)$

Iz navedenoga slijedi

$$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \cdots + \beta_n \cdot x^n$$

Dodatno pretpostavljamo da za poznate vrijednosti x vrijedi

$$Y \sim N(\beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \cdots + \beta_n \cdot x^n, \sigma^2)$$

3.2 Ilustracija geometrijskog pristupa

Model polinomijalne linearne regresije alternativno možemo zapisati u obliku

$$y_i = \beta_0 + \beta_1 \cdot (x_i - \bar{x}) + \beta_2 \cdot (x_i - \bar{x})^2 + \cdots + \beta_n \cdot (x_i - \bar{x})^n + \epsilon_i, \forall i$$

odnosno u vektorskom zapisu

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \beta_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix} + \beta_2 \begin{bmatrix} (x_1 - \bar{x})^2 \\ (x_2 - \bar{x})^2 \\ \vdots \\ (x_n - \bar{x})^2 \end{bmatrix} + \ldots + \beta_n \begin{bmatrix} (x_1 - \bar{x})^n \\ (x_2 - \bar{x})^n \\ \vdots \\ (x_n - \bar{x})^n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Prilagođavanjem modela podacima opredjeljujemo se za $n \in \mathbb{N}$ takav da dobiveni polinom najbolje opisuje promatrani set podataka pa tako imamo sljedeće mogućnosti za vektor očekivanih vrijednosti $\mathbb{E}[Y|X = x]$ u ovisnosti o odabranom stupnju polinoma n:

- $n = 0 \Rightarrow \beta_0$
- $n = 1 \Rightarrow \beta_0 + \beta_1 \cdot (x - \bar{x})$
- $n = 2 \Rightarrow \beta_0 + \beta_1 \cdot (x - \bar{x}) + \beta_2 \cdot (x - \bar{x})^2$
- $n = 3 \Rightarrow \beta_0 + \beta_1 \cdot (x - \bar{x}) + \beta_2 \cdot (x - \bar{x})^2 + \beta_3 \cdot (x - \bar{x})^3$

  $\vdots$

- $n \in \mathbb{N} \Rightarrow \beta_0 + \beta_1 \cdot (x - \bar{x}) + \beta_2 \cdot (x - \bar{x})^2 + \beta_3 \cdot (x - \bar{x})^3 + \cdots + \beta_n \cdot (x - \bar{x})^n$

Procijenjeni parametri $\hat{\beta}_i, i = 1, \ldots, n$, mijenjaju se s odabirom različitog n. Uočimo, ovako zapisan model polinomijalne regresije u vektorskom obliku i dalje nije ortogonalan, odnosno vektori koji se nalaze uz parametre nisu međusobno ortogonalni. Prilikom korištenja geometrijskog pristupa zahtijevamo da model bude ortogonalan, što postižemo ortogonalizacijom spomenutih vektora te u konačnici dolazimo do ortogonalnog zapisa modela polinomijalne regresije koji glasi

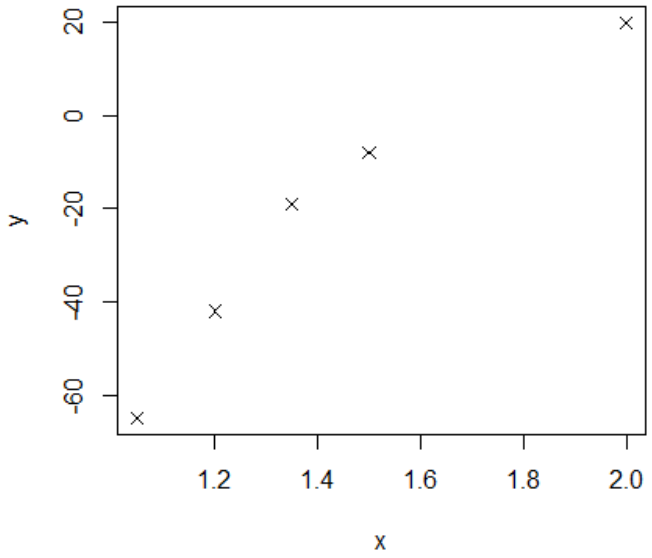
$$y = \beta_0 \cdot p_0(x) + \beta_1 \cdot p_1(x) + \beta_2 \cdot p_2(x) + \cdots + \beta_n \cdot p_n(x) + \epsilon$$

gdje $p_i(x), i = 1, \ldots, n$ predstavljaju polinome i-tog stupnja. Napomenimo ovdje kako je model polinomijalne regresije hijerarhijski, odnosno ukoliko se u modelu pojavljuje n-ta polinomijalna komponenta, tada se moraju pojavljivati i sve polinomijalne komponentne do tog stupnja.
Nakon kratkog upoznavanja s modelom, vrijeme je da kroz primjere pokažemo kako konkretno geometrijski pristup funkcionira u slučaju polinomijalne linearne regresije.

Podaci

Podaci koje ćemo koristiti prikupljeni su na pet velikih skupina starijih ovaca bez zubiju. Svaka od promatranih pet skupina ovaca ponuđena je s različitim površinama pašnjaka, što se iskazuje u kilogramima ponuđene suhe tvari u ispaši po ovci na dan, a prikupljeni podaci nose informaciju o prosječnom povećanju težine ovaca u gramima u svakoj od skupina s obzirom na to. Dobivene podatke navodimo te ćemo ih odmah i grafički prikazati.

Ponuđena količina suhe tvari u ispaši (x)	Povećanje težine (y)
1.05	-65
1.20	-42
1.35	-19
1.50	-8
2	20



Slika 3.1: Graf rasipanja povećanja težine u odnosu na ponuđenu količinu suhe tvari u ispaši

Modeli polinomijalne linearne regresije

Prije nego krenemo s analizom, pretpostavit ćemo da su ranije navedene pretpostavke zadovoljene. S obzirom da promatrani podaci sadrže opažanja za pet različitih vrijednosti nezavisne varijable, možemo zaključiti kako zasigurno podatke nećemo opisivati polinomom stupnja većeg od četiri jer nemamo dovoljno podataka za to. Dakle, sljedeći polinomijalni modeli dolaze u obzir za vektor očekivanih vrijednosti:

- β_0
- $\beta_0 + \beta_1 \cdot (x - \bar{x})$
- $\beta_0 + \beta_1 \cdot (x - \bar{x}) + \beta_2 \cdot (x - \bar{x})^2$
- $\beta_0 + \beta_1 \cdot (x - \bar{x}) + \beta_2 \cdot (x - \bar{x})^2 + \beta_3 \cdot (x - \bar{x})^3$
- $\beta_0 + \beta_1 \cdot (x - \bar{x}) + \beta_2 \cdot (x - \bar{x})^2 + \beta_3 \cdot (x - \bar{x})^3 + \beta_4 \cdot (x - \bar{x})^4$

Uvedimo sljedeće oznake vektora:

$$X_1 = 1, X_2 = (x - \bar{x}), X_3 = (x - \bar{x})^2, X_4 = (x - \bar{x})^3, X_5 = (x - \bar{x})^4$$

Posljednji gore navedeni model nazivamo potpunim modelom te koristeći ovu notaciju vektor očekivanja glasi

$$\beta_0 \cdot X_1 + \beta_1 \cdot X_2 + \beta_2 \cdot X_3 + \beta_3 \cdot X_4 + \beta_4 \cdot X_5$$

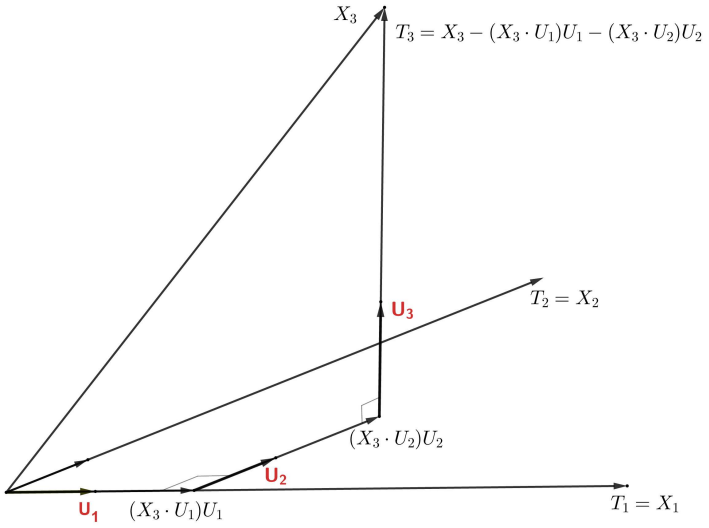
Kako vrijedi $\bar{x} = 1.42$ uvrštavanjem dobivamo

$$\beta_0 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} -0.37 \\ -0.22 \\ -0.07 \\ 0.08 \\ 0.58 \end{bmatrix} + \beta_2 \begin{bmatrix} 0.1369 \\ 0.0484 \\ 0.0049 \\ 0.0064 \\ 0.3364 \end{bmatrix} + \beta_3 \begin{bmatrix} -0.0507 \\ -0.0107 \\ -0.0003 \\ 0.0005 \\ 0.1951 \end{bmatrix} + \beta_4 \begin{bmatrix} 0.01874 \\ 0.00234 \\ 0.00002 \\ 0.00004 \\ 0.11316 \end{bmatrix}$$

Uočimo, ovo je neortogonalni zapis polinomijalnog modela te je baza prostora M razapetog potpunim modelom sačinjena od vektora X_1, X_2, X_3, X_4, X_5. Kako bismo došli do ortogonalnog zapisa modela, bazu prostora M ćemo dovesti do ortonormirane baze koristeći Gram-Schmidtov postupak ortogonalizacije. U narednom postupku ortogonalne vektore označavati ćemo s $T_1, \ldots, T_5$, a njima odgovarajuće jedinične vektore s $U_1, \ldots, U_5$. Iz prethodnog poglavlja znamo da su vektori X_1 i X_2 već ortogonalni pa ih je potrebno samo normirati.

$$T_1 = \begin{bmatrix} 0.1369 \\ 0.0484 \\ 0.0049 \\ 0.0064 \\ 0.3364 \end{bmatrix}, \; T_2 = \begin{bmatrix} -0.37 \\ -0.22 \\ -0.07 \\ 0.08 \\ 0.58 \end{bmatrix} \Rightarrow \mathrm{U}_1 = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \; \mathrm{U}_2 = \frac{1}{\sqrt{0.533}} \begin{bmatrix} -0.37 \\ -0.22 \\ -0.07 \\ 0.08 \\ 0.58 \end{bmatrix}$$

Sada uzimamo treći vektor X_3 te tražimo U_3 takav da prostor razapet vektorima U_1, U_2, U_3 bude jednak prostoru razapetom vektorima X_1, X_2, X_3. Sa Slike 3.2 jasno je da, kako bismo to postigli, od vektora X_3 moramo oduzeti njegovu projekciju na vektore U_1 i U_2 te normiranjem dobivenog vektora dobivamo traženi U_3.



Slika 3.2: Ilustracija pronalska vektora U_3 koristeći Gram - Schmidt ortogonalizaciju

Dakle, imamo

$$\begin{aligned} T_3 &= X_3 - (X_3 \cdot U_1)\, U_1 - (X_3 \cdot U_2)\, U_2 \\ &= \begin{bmatrix} 0.1369 \\ 0.0484 \\ 0.0049 \\ 0.0064 \\ 0.3364 \end{bmatrix} - 0.1066 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} - 0.2514 \begin{bmatrix} -0.37 \\ -0.22 \\ -0.07 \\ 0.08 \\ 0.58 \end{bmatrix} = \begin{bmatrix} 0.1233 \\ -0.0029 \\ -0.0841 \\ -0.1203 \\ 0.084 \end{bmatrix} \end{aligned}$$

pa slijedi

$$U_3 = \begin{bmatrix} 0.5891 \\ -0.0139 \\ -0.4018 \\ -0.5747 \\ 0.4013 \end{bmatrix}$$

Slično, iz

$$T_4 = X_4 - (X_4 \cdot U_1)\, U_1 - (X_4 \cdot U_2)\, U_2 - (X_4 \cdot U_3)\, U_3$$
$$T_5 = X_5 - (X_5 \cdot U_1)\, U_1 - (X_5 \cdot U_2)\, U_2 - (X_5 \cdot U_3)\, U_3 - (X_5 \cdot U_4)\, U_4$$

dobivamo

$$T_4 = \begin{bmatrix} -0.0128 \\ 0.0187 \\ 0.01 \\ -0.0186 \\ 0.0027 \end{bmatrix}, \quad T_5 = \begin{bmatrix} 0.0005 \\ -0.0019 \\ 0.0023 \\ -0.001 \\ 0.00004 \end{bmatrix} \Rightarrow \mathrm{U}_4 = \begin{bmatrix} -0.4108 \\ 0.6013 \\ 0.3212 \\ -0.599 \\ 0.0874 \end{bmatrix}, \quad U_5 = \begin{bmatrix} 0.1655 \\ -0.5894 \\ 0.7254 \\ -0.3144 \\ 0.0129 \end{bmatrix}$$

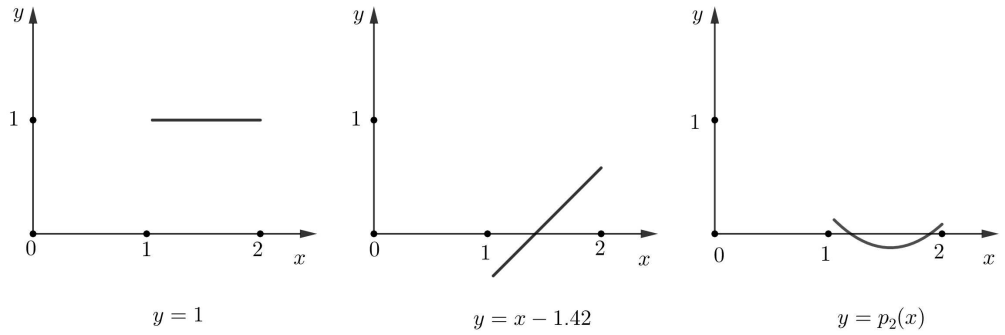
Sada, nakon što smo odredili ortonomormiranu bazu za M sačinjenu od vektora $U_1, \ldots, U_5$ možemo potpuni polinomijalni model zapisati u ortogonalnom obliku. Kako je prostor M dimenzije pet u slučaju potpunog polinomijalnog modela, prostor greške je dimenzije nula te ju zato nećemo navoditi. Dakle, ortogonalni zapis potpunog polinomijalnog modela glasi

$$y = \beta_0 \cdot p_0(x) + \beta_1 \cdot p_1(x) + \beta_2 \cdot p_2(x) + \beta_3 \cdot p_3(x) + \beta_4 \cdot p_4(x)$$

Uvrštavajući dobivene rezultate slijedi

$$y = \beta_0 + \beta_1 \cdot (x - 1.42) + \beta_2 \cdot \left[(x - 1.42)^2 - 0.2514 \cdot (x - 1.42) - 0.1066\right] + \beta_3 \cdot p_3(x) + \beta_4 \cdot p_4(x)$$

Dakle, ranije izračunati vektor T_1 je vektor vrijednosti koje poprima polinom $p_0(x)$ te se koristi za procjenu visine na kojoj se nalazi krivulja za koju pretpostavljamo da opisuje vezu među podacima. Vektor T_2 sadrži vrijednosti linearnog polinoma $p_1(x)$ koje se koriste u procjeni nagiba krivulje, dok su vrijednosti kvadratnog polinoma, koje služe u aproksimaciji zakrivljenosti krivulje, sadržane u vektoru T_3. Kako još nismo proveli analizu i ne znamo koji stupanj polinomijalne regresije će biti optimalan, radi jednostavnosti, polinome $p_3(x)$ i $p_4(x)$ nećemo egzaktno navoditi nego ćemo ih ostaviti u ovakvom obliku te ukoliko će biti potrebno, naknadno ćemo ih odrediti.



Slika 3.3: Konstantna, linearna i kvadratna ortogonalna komponentna u polinomijalnom modelu

Prilagodba potpunog modela polinomijalne regresije

U slučaju potpunog modela, kako smo ranije rekli, prostor M je petodimenzionalan, pa je posljedično, prostor greške dimenzije nula. Zapisan u ortonormiranoj bazi prostora M, model glasi

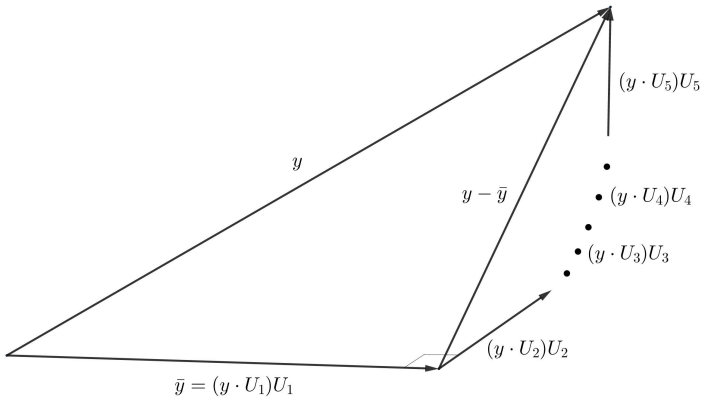
$$y = (y \cdot U_1)U_1 + (y \cdot U_2)U_2 + (y \cdot U_3)U_3 + (y \cdot U_4)U_4 + (y \cdot U_5)U_5$$

Uvrštavanjem odgovarajućih izračunatih vrijednosti slijedi

$$y = -22.8 + 62.4325 \cdot U_2 - 17.4494 \cdot U_3 + 1.8866 \cdot U_4 + 2.9905 \cdot U_5$$

što je ekvivalentno izrazu

$$y - \bar{y} = 62.4325 \cdot U_2 + -17.4494 \cdot U_3 + 1.8866 \cdot U_4 + 2.9905 \cdot U_5$$



Slika 3.4: Ortogonalna dekompozicija vektora y i vektora $y - \bar{y}$

Primjenom Pitagorinog teorema slijedi

$$\| y\|^2 = (y \cdot U_1)^2 + (y \cdot U_2)^2 + (y \cdot U_3)^2 + (y \cdot U_4)^2 + (y \cdot U_5)^2$$

Izračunom dobivamo

$$6814 = 2599.2 + 3897.817 + 304.4809 + 3.5593 + 8.9429$$

Koji stupanj polinomijalne regresije odabrati?

Nakon što smo zapisali ukupnu varijancu koristeći Pitagorin teorem, vidimo kako je ukupna količina varijance u podacima najvećim dijelom opisana s prve tri komponente u modelu. Upravo ta činjenica olakšava nam donošenje odluke koji stupanj polinomijalne regresije odabrati te je u ovom konkretnom primjeru podosta jasno kako će stupanj dva biti optimalan. Međutim, to nije uvijek slučaj te ćemo mi provesti analizu koju bismo u tom slučaju proveli.
Za utvrđivanje optimalnog stupnja polinomijalne regresije potrebno je ustanoviti je li dodavanje određene polinomijalne komponentne statistički značajno. Kako bismo provjerili dodavanje koje kompontetne je, odnosno nije, statistički značajno koristit ćemo testnu statistiku F te ćemo, u tu svrhu, vektor greške procijenjivati na dva načina. Za početak, krećemo od konstantnog polinoma te povećavamo stupanj sve dok za određenu polinomijalnu komponentu ne procijenimo da nije statistički značajna. Dakle, prvo testiramo je li dodavanje linearne komponente statistčki značajno, odnosno testiramo:

- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

Vektor greške procjenjujemo na sljedeća dva načina:

- $(y \cdot U_3)U_3 + (y \cdot U_4)U_4 + (y \cdot U_5)U_5$
- $(y \cdot U_4)U_4 + (y \cdot U_5)U_5$

Dakle, u prvom slučaju grešku procjenjujemo koristeći sve moguće preostale polinomijalne komponente, odnosno izostavljamo konstantnu i linearnu komponentu. U drugoj procjeni greške osim konstantne i linearne, izostavljamo i prvu sljedeću komponentu nakon linearne, kvadratnu. Kako vrijedi

$$y \cdot U_2 = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_5 \end{bmatrix} \frac{1}{\| T_2\|} \begin{bmatrix} \mathrm{p}_1(x_1) \\ \mathrm{p}_1(x_2) \\ \vdots \\ \mathrm{p}_1(x_5) \end{bmatrix}$$

uzimanjem očekivanih vrijednosti imamo

$$\begin{aligned} \mathbb{E}[Y \cdot U_2] &= \frac{1}{\| T_2\|} \sum_{i=1}^{5} \left[\beta_0 \cdot p_0(x_i) + \beta_1 \cdot p_1(x_i) + \beta_2 \cdot p_2(x_i)\right] \cdot p_1(x_i) \\ &= \frac{1}{\| T_2\|} \cdot \beta_1 \sum_{i=1}^{5} p_1(x_i)^2 \\ &= \frac{1}{\| T_2\|} \cdot \beta_1 \| T_2\|^2 \\ &= \beta_1 \| T_2\| \end{aligned}$$

gdje drugi redak slijedi iz ortogonalnosti polinomijalnih komponenata. Iz navedenoga slijedi da je U_2 vektor smjera nulte hipoteze pa sada računamo testnu statistiku F za oba slučaja.

- $F = \dfrac{(y \cdot U_2)^2}{[(y \cdot U_3)^2 + (y \cdot U_4)^2 + (y \cdot U_5)^2]/3} \overset{H_0}{\sim} F(1, 5-2) \Rightarrow F = \dfrac{3897.817}{105.661} = 36.8898$
- $F = \dfrac{(y \cdot U_2)^2}{[(y \cdot U_4)^2 + (y \cdot U_5)^2]/2} \overset{H_0}{\sim} F(1, 5-3) \;\Rightarrow\; F = \dfrac{3897.817}{6.2511} = 623.5381$

Kako je 0.95 - kvantil $F_{1,3}$ distribucije jednak 10.13, a 0.95 - kvantil $F_{1,2}$ distribucije iznosi 18.51 zaključujemo kako možemo odbaciti nultu hipotezu koja pretpostavlja kako linearna komponenta nije statistički značajna u modelu. Za linearnu komponentu zaključili bismo da nije statistički značajna kada bismo iz obje testne statistike zaključilli da nultu hipotezu ne možemo odbaciti na promatranom nivou značajnosti. Pitanje koje se intuitivno

javlja jest zašto grešku procjenjujemo na dva načina, odnosno zašto prva procjena nije dovoljna. Naime, prilikom testiranja značajnosti linearne komponente može se dogoditi da su vrijednosti $(y \cdot U_2)^2$ i $(y \cdot U_3)^2$ podjenake pa bi se vrlo lako moglo dogoditi da zbog toga zaključimo kako linearna komponenta nije statistički značajna. Upravo druga procjena greške osigurava nas da se ne dogodi taj propust, pa analogno, prilikom testiranja značajnosti ostalih komponenata radimo dvije procjene greške. S obzirom da je linearna komponenta statistički značajna, nastavljamo s testiranjem značajnosti kvadratne polinomijalne komponente, odnosno testiramo:

- $H_0 : \beta_2 = 0$
- $H_1 : \beta_2 \neq 0$

Slično kao ranije može se pokazati da vrijedi $\mathbb{E}[y \cdot U_3] = \beta_2 \| T_3 \|$ pa je vektor smjera nulte hipoteze u ovom slučaju U_3. Ponovno računamo odgovarajuće testne statistike.

- $F = \dfrac{(y \cdot U_3)^2}{[(y \cdot U_4)^2 + (y \cdot U_5)^2]/2} \overset{H_0}{\sim} F(1,2) \;\Rightarrow\; F = \dfrac{304.4809}{6.2511} = 48.7081$
- $F = \dfrac{(y \cdot U_3)^2}{(y \cdot U_5)^2} \overset{H_0}{\sim} F(1,1) \;\Rightarrow\; F = \dfrac{304.4809}{8.9429} = 34.0471$

Na razini značajnosti 5% odbacujemo nultu hipotezu koja pretpostavlja da kvadratna komponenta nije statistički značajna pa testiramo dalje.

- $H_0 : \beta_3 = 0$
- $H_1 : \beta_3 \neq 0$

Procjena greške, u ovom slučaju samo jedna, je

- $(y \cdot U_5)U_5$

Uz činjenicu da je vektor smjera nulte hipoteze U_3 računamo testnu statistiku.

- $F = \dfrac{(y \cdot U_4)^2}{(y \cdot U_5)^2} \overset{H_0}{\sim} F(1,1) \;\Rightarrow\; F = \dfrac{3.5593}{8.9429} = 0.398$

S obzirom da je 0.95 - kvantil $F_{1,1}$ distribucije jednak 161.45 jasno je da za kubičnu komponentu zaključujemo kako nije statistički značajna. Navedenom analizom došli smo do rezultata da je optimalan stupanj polinomijalne regresija zaista jednak dva te odgovarajući model glasi

$$y = \beta_0 \cdot p_0(x) + \beta_1 \cdot p_1(x) + \beta_2 \cdot p_2(x) + \epsilon$$

Kvadratni model polinomijalne regresije

Nakon što smo odabrali optimalan stupanj polinomijalne regresije, potrebno je procijeniti nepoznate parametre. Za početak, prilagođen vektor dobiven optimalnim modelom glasi

$$\hat{y} = (y \cdot U_1)U_1 + (y \cdot U_2)U_2 + (y \cdot U_3)U_3$$

Uvrštavanjem ranije uvedenih ortogonalnih vektora T_1, T_2 i T_3 slijedi

$$\hat{y} = (y \cdot U_1)\frac{T_1}{\| T_1\|} + (y \cdot U_2)\frac{T_2}{\| T_2\|} + (y \cdot U_3)\frac{T_3}{\| T_3\|}$$

Koristeći činjenicu da su vrijednosti koje polinomi $p_0(x)$, $p_1(x)$ i $p_2(x)$ mogu poprimiti sadržane u vektorima T_1, T_2 i T_3 redom, možemo pisati

$$\hat{y} = \frac{(y \cdot U_1)}{\| T_1\|}p_0(x) + \frac{(y \cdot U_2)}{\| T_2\|}p_1(x) + \frac{(y \cdot U_3)}{\| T_3\|}p_2(x)$$

Dakle, nepoznate parametre β_0, β_1 i β_2, procjenjuemo na sljedeći način:

$$\hat{\beta_0} = \frac{y \cdot U_1}{\| T_1\|} = -22.8$$

$$\hat{\beta_1} = \frac{y \cdot U_2}{\| T_2\|} = 85.516$$

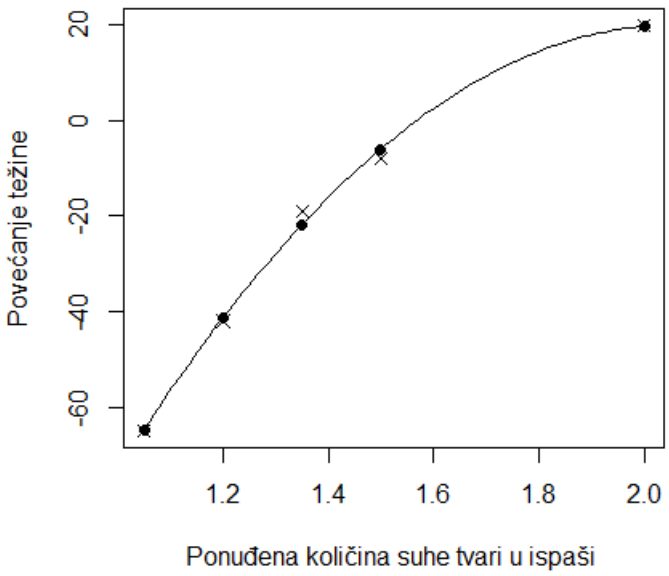
$$\hat{\beta_2} = \frac{y \cdot U_3}{\| T_3\|} = -83.3594$$

Konačno, dolazimo do sljedeće kvadratne aproksimacije veze između promatranih podataka

$$\hat{y} = -22.8 + 85.516 \cdot (x - 1.42) - 83.3594 \cdot \left[(x - 1.42)^2 - 0.2514 \cdot (x - 1.42) - 0.1066\right]$$

Slijedi da je vektor prilagođenih vrijednosti jednak

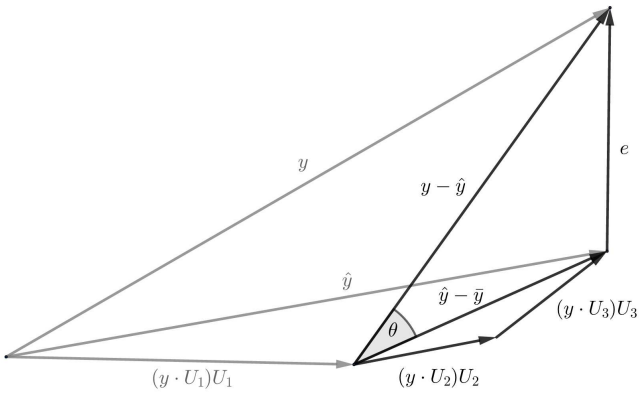
$$\hat{y} = \left[-64.7197, \quad -41.3719, \quad -21.7752, \quad -5.9298, \quad 19.7966\right]^T$$



Slika 3.5: Regresijska krivulja drugog reda s prilagođenim vrijednostima (•) te podacima (×)

Preostaje odrediti vektor greške pa posljedično i procijeniti varijancu.

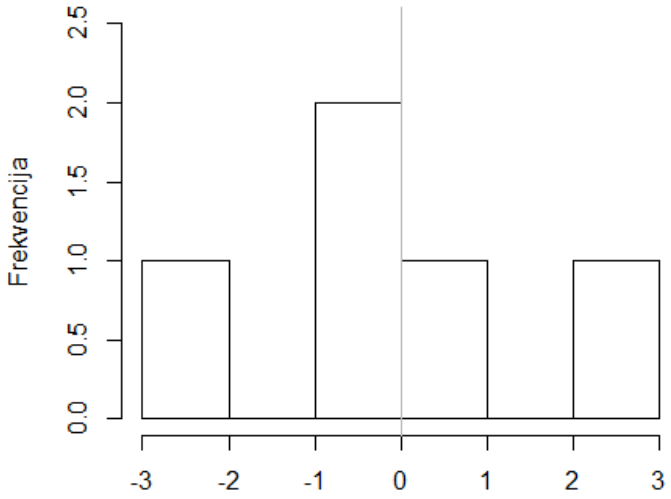
$$e = y - \hat{y} = \begin{bmatrix} -0.2803 \\ 0.6281 \\ 2.7752 \\ -2.0702 \\ 0.2034 \end{bmatrix} \Rightarrow s^2 = \frac{\| e\|^2}{2} = 6.2511$$



Slika 3.6: Ortogonalna dekompozicija vektora y i $y - \bar{y}$ u kvadratnom modelu

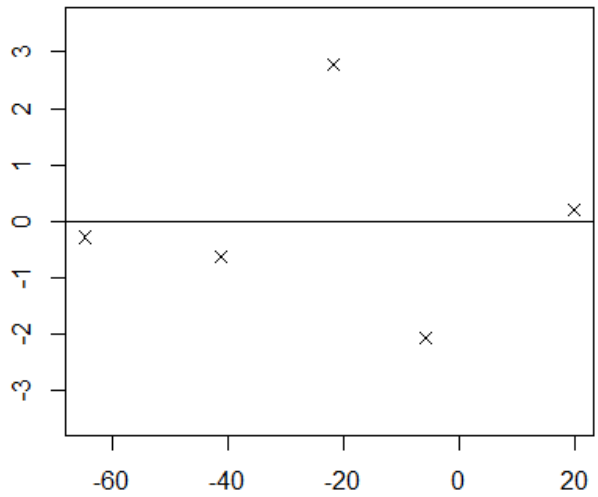
Provjera pretpostavki

Nakon provođenja analize u kojoj smo pretpostavili istinitost pretpostavki, potrebno je provjeriti jesu li ranije navedene pretpostavke ispunjene za slučaj $n = 2$.
Za početak, zadovoljenost pretpostavke nezavisnosti slijedi iz načina prikupljanja podataka u istraživanju. Unatoč malom broju podataka, pretpostavku normalnosti ćemo svejedno provjeriti prikazujući histogram grešaka dobivenih kvadratnim modelom. Iz prikazanog histograma na Slici 3.7 vidimo da je pretpostavka normalnosti razumna.



Slika 3.7: Histogram reziduala

S obzirom na Sliku 3.8 zaključujemo da ni pretpostavka konstantne varijance grešaka, odnosno pretpostavka homoskedastičnosti nije narušena.



Slika 3.8: Graf reziduala u odnosu na procijenjene vrijednosti y

Pouzdani intervali

Nakon što smo se uvjerili u razumnost pretpostavki u modelu, prelazimo na određivanje pouzdanih intervala.

Pouzdani interval za parametre β_0, β_1 i β_2

Prilikom određivanja pouzdanog intervala za navedene parametre koristimo:

$$Y \cdot U_1 = \hat{\beta}_0 \| T_1\| \sim N\left(\beta_0 \cdot \| T_1\|, \sigma^2\right) \Rightarrow \| T_1\| (\hat{\beta}_0 - \beta_0) \sim N\left(0, \sigma^2\right)$$

$$Y \cdot U_2 = \hat{\beta}_1 \| T_2\| \sim N\left(\beta_1 \cdot \| T_2\|, \sigma^2\right) \Rightarrow \| T_2\| (\hat{\beta}_1 - \beta_1) \sim N\left(0, \sigma^2\right)$$

$$Y \cdot U_3 = \hat{\beta}_2 \| T_3\| \sim N\left(\beta_2 \cdot \| T_3\|, \sigma^2\right) \Rightarrow \| T_3\| (\hat{\beta}_2 - \beta_2) \sim N\left(0, \sigma^2\right)$$

Slijedi da su odgovarajući 95% pouzdani intervali:

$$\hat{\beta}_0 \pm t_2(0.975)\frac{s}{\| T_1\|} \Rightarrow -27.611 \leq \beta_0 \leq -17.9891$$

$$\hat{\beta}_1 \pm t_2(0.975)\frac{s}{\| T_2\|} \Rightarrow 70.7809 \leq \beta_1 \leq 100.251$$

$$\hat{\beta}_2 \pm t_2(0.975)\frac{s}{\| T_3\|} \Rightarrow -134.7508 \leq \beta_2 \leq -31.9681$$

Pouzdani interval za očekivanu vrijednost $\mathbb{E}[Y|x = x_0]$

Odredimo sada pouzdani interval za modelom procijenjenu očekivanu vrijednost $\hat{\beta}_0 \cdot p_0(x) + \hat{\beta}_1 \cdot p_1(x) + \hat{\beta}_2 \cdot p_2(x)$ za neku poznatu vrijednost $x = x_0$.
Računamo varijancu procjenitelja u svrhu procjene standardne devijacije

$$\text{Var}[\hat{\beta}_0 p_0(x_0) + \hat{\beta}_1 p_1(x_0) + \hat{\beta}_2 p_2(x_0)] = \text{Var}(\hat{\beta}_0) + p_1(x_0)^2\text{Var}(\hat{\beta}_1) + p_2(x_0)^2\text{Var}(\hat{\beta}_2)$$

Traženi 95% pouzdani interval za $\mathbb{E}[Y|x = x_0]$ je

$$\hat{\beta}_0 + \hat{\beta}_1 p_1(x_0) + \hat{\beta}_2 p_2(x_0) \pm t_{n-2}(0.975)\sqrt{s^2\left[\frac{1}{\| T_1\|^2} + \frac{p_1(x_0)^2}{\| T_2\|^2} + \frac{p_2(x_0)^2}{\| T_3\|^2}\right]}$$

Koeficijent korelacije i koeficijent determinacije

Sa Slike 3.6 slijedi da je u kvadratnom modelu izraz za koeficijent korelacije upravo

$$cos\,(\theta) = \frac{(\hat{y} - \bar{y})}{\| \hat{y} - \bar{y}\|} \cdot \frac{(y - \bar{y})}{\| y - \bar{y}\|}$$

$$cos\,(\theta) = \frac{4202.298}{4208.544} = 0.9985 \Rightarrow \theta = 3^\circ$$

Jasno je kako iz dobivenog slijedi da je prisutna snažna korelacija među podacima.
Kvadriranjem dobivenog koeficijenta korelacije dobivamo da koeficijent determinacije r^2 u kvadratnom modelu iznosi 0.997. Dakle, možemo zaključiti kako je količina varijabilnosti kriterijske varijable opisane modelom vrlo visoka. Međutim, napomenimo ovdje kako s koeficijentom determinacije moramo biti oprezni. Naime, u slučaju višeg odabranog stupnja polinomijalne regresije r^2 će biti samo veći, odnosno r^2 raste kako se broj stupnjeva slobode, preostalih za grešku, smanjuje, a znamo da tada raste rizik pretjerane prilagodbe modela podacima, engl. *overfitting*.

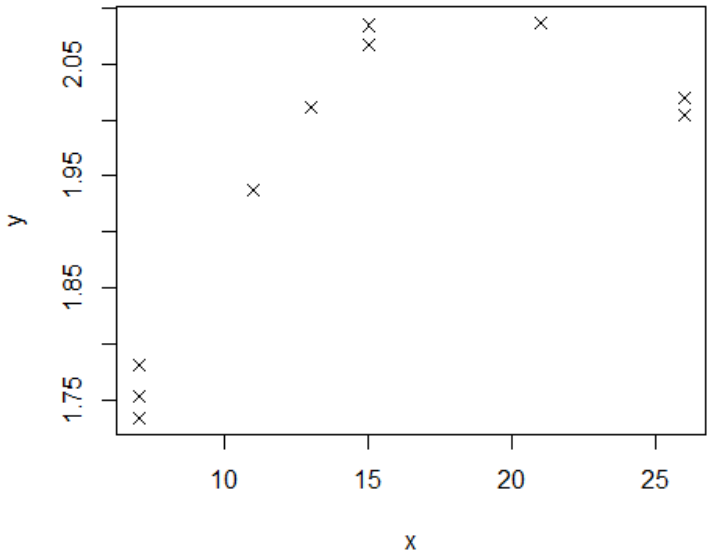
3.3 Rekapitulacija modela

Sada ćemo ukratko ponoviti ranije navedene korake koje koristimo prilikom geometrijskog pristupa polinomijalnoj regresiji te pritom obuhvatiti i nešto drugačiji slučaj, slučaj kada sve vrijednosti, koje nezavisna varijabla poprima, nisu međusobno različite. Podatke s kojima ćemo raditi ćemo simulirati, na način da za svaku poznatu vrijednost x vrijedi

$$Y \sim N(1.3 + 0.08x - 0.002x^2, 0.0004)$$

Postupak biranja simuliranih vrijednosti y je slučajan te analogan kao i u prethodnom poglavlju gdje smo ga detaljnije opisali pa u ovom slučaju dobivene simulirane podatke samo navodimo.

x	7	7	7	11	13	15	15	21	26	26
y	1.78	1.75	1.73	1.94	2.01	2.08	2.07	2.09	2.00	2.02



Slika 3.9: Graf rasipanja za promatrane podatke

Uočimo, unatoč tome što je duljina promatranih podataka 10, maksimalan stupanj polinomijalne regresije za promatrane podatke je 6, zato što imamo samo 6 različitih vrijednosti

zavisne varijable. Upravo zato, za razliku od prethodnog primjera i u potpunom modelu javljat će se greška te on zapisan u ortogonalnom obliku u ovom slučaju glasi

$$\begin{aligned} y = &\beta_0 \cdot p_0(x) + \beta_1 \cdot p_1(x) + \beta_2 \cdot p_2(x) + \beta_3 \cdot p_3(x) + \beta_4 \cdot p_4(x) \\ &+ \beta_5 \cdot p_5(x) + \beta_6 \cdot p_6(x) + \epsilon \end{aligned}$$

Nakon izračuna vektora $T_i, i = 1, \ldots, 6$ koristeći Gram-Schmidtovu ortogonalizaciju vektora $X_i = (x - \bar{x})^i, i = 1, \ldots, 6$ dolazimo do

$$\begin{aligned} y = &\beta_0 + \beta_1 \cdot (x - 14.8) + \beta_2 \cdot \left[(x - 14.8)^2 - 48.96 - 3.1941 \cdot (x - 14.8)\right] \\ &+ \beta_3 \cdot p_3(x) + \beta_4 \cdot p_4(x) + \beta_5 \cdot p_5(x) + \beta_6 \cdot p_6(x) + \epsilon \end{aligned}$$

Vektore $U_1, \ldots, U_6$ koji čine ortonormiranu bazu za M dobijemo normiranjem vektora $T_1, \ldots, T_6$, analogno kao i u prethodnom primjeru te ih također, radi jednostavnosti, nećemo egzaktno navoditi. Jednadžba procijenjenje krivulje u potpunom modelu glasi

$$\hat{y} = (y \cdot U_1)U_1 + (y \cdot U_2)U_2 + (y \cdot U_3)U_3 + (y \cdot U_4)U_4 + (y \cdot U_5)U_5 + (y \cdot U_6)U_6$$

odnosno

$$\hat{y} = 1.947 + 0.3071 \cdot U_2 - 0.288 \cdot U_3 + 0.0071 \cdot U_4 + 0.0189 \cdot U_5 + 0.0071 \cdot U_6$$

Odmah možemo izračunati i vektor greške te procijeniti varijancu pa to i činimo.

$$e = y - \hat{y} = \begin{bmatrix} 0.0267 \\ -0.0033 \\ -0.0233 \\ 0 \\ 0 \\ 0.005 \\ -0.005 \\ 0 \\ -0.01 \\ 0.01 \end{bmatrix} \Rightarrow s^2 = \frac{\| e\|^2}{4} = 0.0004$$

Sljedeće što je potrebno napraviti jest odrediti optimalan stupanj polinomijalne regresije. Za razliku od prethodnog primjera, imamo poznatu grešku te ćemo, u ovom slučaju, nju koristiti prilikom računanja testne statistike. Kao i ranije, krećemo od konstantnog polinoma i dodajemo polinomijalne komponente sve dok za neku ne procijenimo da nije statistčki značajna. Dakle, prvo testiramo:

- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

Testna statistika iznosi

$$F = \frac{(y \cdot U_2)^2}{s^2} = 248.6449$$

Na temelju dobivenog zaključujemo kako možemo odbaciti nultu hipotezu na razini značajnosti 5% pa nastavljamo dalje testirati.

- $H_0 : \beta_2 = 0$
- $H_1 : \beta_2 \neq 0$

Računamo testnu statistiku

$$F = \frac{(y \cdot U_3)^2}{s^2} = 218.7906$$

Analogno slijedi da i u ovom slučaju možemo odbaciti nultu hipotezu. Provjeravamo sada statističku značajnost kubične polinomijalne komponente.

- $H_0 : \beta_3 = 0$
- $H_1 : \beta_3 \neq 0$

$$F = \frac{(y \cdot U_4)^2}{s^2} = 0.1335$$

S obziromm da 0.95 - kvantil $F_{1,6}$ distribucije iznosi 5.99 ne možemo, kao ranije, odbaciti nultu hipotezu na razini značajnosti 5%. Prisjetimo se, u prethodnom primjeru koristili smo dvije testne statistike te tek ukoliko obje pokažu kako ne možemo zaključiti da je komponenta statistički značajna, komponentu ne bismo dodavali u model. Slično činimo i ovdje. Naime, još računamo je li količina varijabilnosti koja neće biti opisana modelom zbog izostavljanja komponenata višeg reda značajna u odnosu na ranije određenu grešku.

$$F = \frac{(y \cdot U_5)^2 + (y \cdot U_6)^2}{s^2} = 1.0054$$

Iz dobivenog zaključujemo kako ne možemo odbaciti nultu hipotezu na razini značajnosti 5% te slijedi da je optimalan stupanj polinomijalne regresije dva. Sada procijenimo parametre analogno kao u prethodnom primjeru te dolazimo do

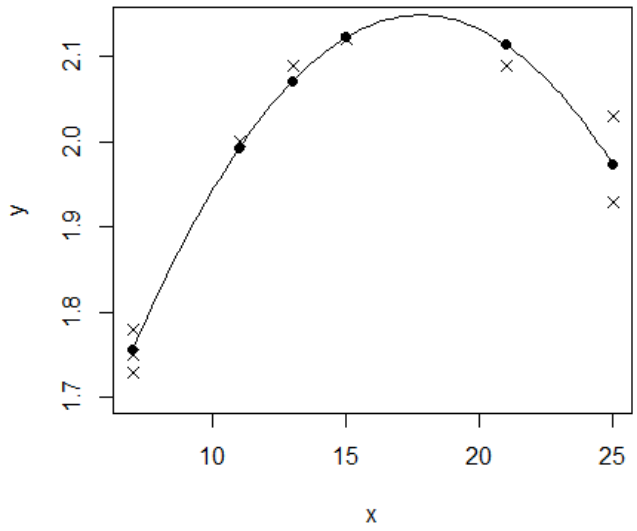
$$\hat{y_2} = 1.947 + 0.0139 \cdot (x - 14.8) - 0.0023 \cdot \left[(x - 14.8)^2 - 3.1941 \cdot (x - 14.8) - 48.96\right]$$

Greška koja se javlja u prilikom korištenja polinomijalne regresije reda dva je

$$e_2 = y - \hat{y_2} = \begin{bmatrix} 0.0254 \\ -0.0046 \\ -0.0246 \\ -0.0077 \\ -0.0063 \\ 0.0137 \\ 0.0037 \\ -0.0145 \\ -0.0083 \\ 0.0117 \end{bmatrix} \Rightarrow s_2^2 = \frac{\| e_2\|^2}{7} = 0.0002$$

Spomenimo kako vrijedi

$$\| e_2\|^2 = \| e\|^2 + \| (y \cdot U_4)U_4 + (y \cdot U_5)U_5 + (y \cdot U_6)U_6\|^2$$



Slika 3.10: Regresijska krivulja drugog reda s prilagođenim vrijednostima (•) te podacima (×)

Naposljetku navedimo još pouzdane intervale za nepoznate parametre te za $\mathbb{E}[Y|x = x_0]$.

- $1.9347 \leq \beta_0 \leq 1.9593$
- $0.0121 \leq \beta_1 \leq 0.0156$
- $-0.0026 \leq \beta_2 \leq -0.002$
- $1.306 \leq \mathbb{E}[Y|x = 1.5] \leq 1.4254$

Koeficijent korelacije iznosi 0.9944, a koeficijent determinacije jednak je 0.9889.

Dodatak A

Prilog

df	1	2	3	4	5	6	7	8	9	10	20
1	161.45	18.51	10.13	7.71	6.61	5.99	5.59	5.32	5.12	4.96	4.35
2	199.50	19.00	9.55	6.94	5.79	5.14	4.74	4.46	4.26	4.10	3.49
3	215.71	19.16	9.28	6.59	5.41	4.76	4.35	4.07	3.86	3.71	3.10
4	224.58	19.25	9.12	6.39	5.19	4.53	4.12	3.84	3.63	3.48	2.87
5	230.16	19.30	9.01	6.26	5.05	4.39	3.97	3.69	3.48	3.33	2.71
6	233.99	19.33	8.94	6.16	4.95	4.28	3.87	3.58	3.37	3.22	2.60
7	236.77	19.35	8.89	6.09	4.88	4.21	3.79	3.50	3.29	3.14	2.51
8	238.88	19.37	8.85	6.04	4.82	4.15	3.73	3.44	3.23	3.07	2.45
9	240.54	19.38	8.81	6.00	4.77	4.10	3.68	3.39	3.18	3.02	2.39
10	241.88	19.40	8.79	5.96	4.74	4.06	3.64	3.35	3.14	2.98	2.35
11	242.98	19.40	8.76	5.94	4.70	4.03	3.60	3.31	3.10	2.94	2.31
12	243.91	19.41	8.74	5.91	4.68	4.00	3.57	3.29	3.07	2.91	2.28
13	244.69	19.42	8.73	5.89	4.66	3.98	3.55	3.26	3.05	2.89	2.25
14	245.36	19.42	8.71	5.87	4.64	3.96	3.53	3.24	3.03	2.86	2.22
15	245.95	19.43	8.70	5.86	4.62	3.94	3.51	3.22	3.01	2.85	2.20
16	246.46	19.43	8.69	5.84	4.60	3.92	3.49	3.20	2.99	2.83	2.18
17	246.91	19.44	8.68	5.83	4.59	3.91	3.48	3.19	2.97	2.81	2.17
18	247.32	19.44	8.67	5.82	4.58	3.9	3.47	3.17	2.96	2.80	2.15
19	247.69	19.44	8.67	5.81	4.57	3.88	3.45	3.16	2.95	2.79	2.14
20	248.01	19.45	8.66	5.80	4.56	3.87	3.44	3.15	2.94	2.77	2.12

Tablica A.1: Tablica kvantila $F-$ razdiobe ($\alpha = 0.05$)

Bibliografija

[1] *The General Linear F-Test*, https://online.stat.psu.edu/stat501/lesson/6/6.2, posjećena 9.1.2020.

[2] *Polynomial Regression Models*, http://home.iitk.ac.in/~shalab/regression/Chapter12-Regression-PolynomialRegression.pdf, posjećena 9.1.2020.

[3] *Regresijska analiza*, https://www.pmf.unizg.hr/_download/repository/PREDAVANJE11.pdf, posjećena 9.1.2020.

[4] *Simple Linear Regression*, https://daviddalpiaz.github.io/appliedstats/simple-linear-regression.html, posjećena 9.1.2020.

[5] https://contentsimplicity.com/machine-learning-simple-linear-regression/, posjećena 9.1.2020.

[6] https://www.youtube.com/watch?v=4otEcA3gjLk&t=3s, posjećena 9.1.2020.

[7] Z. Franušić i J. Šiftar, *Linearna algebra 2*, https://web.math.pmf.unizg.hr/~fran/predavanja-LA2.pdf, posjećena 9.1.2020.

[8] D. Saville i G. R. Wood, *Statistical Methods: The Geometric Approach*, Springer, New York, 1991.

[9] P. Škiljan, *Diplomski rad - Linearna regresija u aktuarstvu*, (2019), https://repozitorij.pmf.unizg.hr/en/islandora/object/pmf%3A8354/datastream/PDF/view.

Sažetak

U ovom radu dan je jedan drugi aspekt provođenja regresijske analize. Konkretno, linearnoj i polinomijalnoj regresiji pristupili smo na nešto slikovitiji način, koristeći geometrijski pristup. Kako bismo to izveli, od krucijalne važnosti bili su nam vektori koji su nam omogućili povezivanje znanja iz područja linearne algebre i područja statistike. Naime, podatke s kojima radimo, kao i sam model zapisivali smo u vektorskom obliku. Prilikom provođenja statističke analize, klasičan izračun modelom prilagođene vrijednosti zamijenili smo izračunom ortogonalne projekcije vektora opažanja na prostor razapet samim modelom. Svojstva ortogonalne projekcije poslužila su nam i prilikom provođenja statističkih testova te računanja testne statistike. Dakle, uobičajenim izrazima koje u regresijskoj analizi inače koristimo dodali smo jednu novu interpretaciju te pokazali da je to zaista ekvivalentno. Upravo ta mogućnost sagledavanja problema te pristupa njegovu rješavanju s više točaka gledišta koje vode istom rješenju jedna je od ljepota matematike koja nam omogućuje povezivanje znanja iz različitih područja te pomaže razjasniti eventualne nedoumice.

Summary

In this thesis, another aspect of conducting regression analysis is given. In particular, we approached linear and polynomial regression in a somewhat more illustrative way, using a geometric approach. To achieve this, the vectors that enablee us to connect knowledge from the field of linear algebra and the field of statistics were of crucial importance to us. Namely, the data we work with, as well as the model itself, were written in vector form. When performing statistical analysis, we replaced the classical calculation of model fitted values by calculating the orthogonal projection of the observation vector on the model space. The properties of orthogonal projection also served us when conducting statistical tests and calculating test statistics. So, we added a new interpretation to the common expressions we usually use in regression analysis and showed that it is really equivalent. It is this possibility of perceiving the problem and approaching its solution from several points of view that leads to the same solution which is one of the beauties of mathematics that enables us to connect knowledge from different fields and helps to clarify possible doubts.

Životopis

Rođena sam 25. kolovoza 1996. godine u Zagrebu. Završila sam osnovnu školu Klinča Sela 2011. godine te potom upisala XI. gimnaziju u Zagrebu koju završavam 2015. godine. Tada nastavljam svoje obrazovanje na Prirodoslovno - matematičkom fakultetu u Zagrebu gdje upisujem Matematiku, smjer nastavnički. Godine 2018. postajem sveučilišna prvostupnica edukacije matematike te zatim, na istom fakultetu, upisujem diplomski studij Financijske i poslovne matematike.