Metode linearne regresije i primjena u marketingu

Ivković, Ana

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO-MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Ana Ivković

METODE LINEARNE REGRESIJE I PRIMJENA U MARKETINGU

Diplomski rad

Voditelj rada:
prof. dr. sc. Siniša Slijepčević

Zagreb, 2022.

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. _____

Hvala svim prijateljima i kolegama koji su mi uljepšali i
olakšali ovaj životni period,
hvala mentoru na razumijevanju, savjetima i podršci,
a najviše hvala mojoj obitelji koja mi je bila najveća podrška tokom studija.

SADRŽAJ:

Uvod

Predmet regresije, ili linearnog modela, središnja je tema statistike. Odnosi se na ono što se može reći o nekoj varijabli koja nas zanima, koju možda nismo u mogućnosti izmjeriti, počevši od informacija o jednoj ili više drugih varijabli, za koje možda nismo zainteresirani, ali koje možemo izmjeriti. Varijablu koja nam je od interesa modeliramo kao njihovu linearnu kombinaciju zajedno s pripadajućom pogreškom. Ispostavilo se da je ovaj jednostavan pristup vrlo fleksibilan te vrlo koristan.

Dvije osnovne vrste regresije su jednostavna linearna regresija i višestruka linearna regresija. Uz njih, postoji još i nelinearna regresija koja je najmanje prisutna u znanstvenoj primjeni.

Nama će od interesa biti linearna regresija.

Nakon analize metoda, posljednji dio rada dat će nam uvid u linearne te multiplikativne modele te odabir funkcionalnih oblika za modeliranje *Marketing Mixa* s ciljem konstruiranja optimalne kombinacije strategija koje bi maksimizirale profit.

1. Linearna regresija

Pri prvom susretu sa statistikom, susrećemo se sa slučajnim varijablama te ih promatramo svaku zasebno.
Međutim, vrlo brzo se susrećemo s više od jedne slučajne varijable odjednom, stoga već sada moramo razmišljati o tome kako su one međusobno povezane.
Za početak, uzmimo najjednostavniji slučaj, dvije varijable te promotrimo najprije dva ekstremna slučaja.
U prvom ekstremu, dvije varijable mogu biti međusobno nezavisne tj. nepovezane. Ono što tada imamo zapravo su dva jednodimenzionalna problema, a ne jedan dvodimenzionalni, te je takve varijable najbolje promatrati zasebno.
S druge strane, dvije varijable mogu biti u biti iste, odnosno jedna nam daje potpunu informaciju o drugoj. Takve varijable su potpuno međusobno zavisne. Ni u tom slučaju se ne radi o dvodimenzionalnom, već jednom jednodimenzionalnom problemu te je takav problem najbolje tako i razmatrati.

Sada kada smo promotrili ekstreme, preostaje nam tipičan i zapravo najvažniji slučaj – dvodimenzionalni podaci $(x_1, y_1), \dots, (x_n, y_n)$ gdje je svaka od varijabli x i y djelomično, ali ne u potpunosti zavisna o drugoj.
Najčešće, naš interes je jedna varijabla, recimo y, te nas zanima što nam x može reći o y. U tom slučaju y nazivamo *varijablom odgovora,* a x *varijablom objašnjenja* ili *prediktorskom varijablom (varijablom predviđanja).* Treći naziv za x je *regresor* što objašnjava zašto cijelu temu nazivamo regresija.

Dakle, prema [1] zaključujemo da je regresija statistička metoda koja pronalazi odnos između zavisne varijable i jedne ili više nezavisnih varijabli. Ako zavisna varijabla ovisi samo o jednoj nezavisnoj varijabli onda se radi o jednostavnoj regresiji. S druge strane, ako ovisi o više nezavisnih varijabli, kažemo da je riječ o višestrukoj regresiji. Kada je odnos zavisne i nezavisne varijable linearan, radi se o *linearnoj regresiji.*

1.1. Metoda najmanjih kvadrata

Primjer 1.1 *Promotrimo kako implementirati metodu najmanjih kvadrata.*

Prvo radimo najjednostavniji slučaj, prilagodbu pravca

$$y = a + bx$$

pomoću najmanjih kvadrata kroz skup podataka $(x_1, y_1), \dots, (x_n, y_n)$.
Sukladno tome, biramo a i b takve da minimiziramo odstupanje. Odstupanje mjerimo tako da promatramo sumu kvadrata pogrešaka ϵ_i, pa imamo

$$SS \coloneqq \sum_{i=1}^{n} {\epsilon_i}^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

Uzimajući da je $\partial SS/\partial a = 0 \; i \; \partial SS/\partial b = 0$ dobivamo:

$$\frac{\partial SS}{\partial a} \coloneqq -2 \sum_{i=1}^{n} \epsilon_i = -2 \sum_{i=1}^{n} (y_i - a - bx_i),$$

$$\frac{\partial SS}{\partial b} \coloneqq -2 \sum_{i=1}^{n} x_i \epsilon_i = -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i).$$

Kako bismo pronašli minimum, obje izjednačimo s nulom:

$$\sum_{i=1}^{n} (y_i - a - bx_i) = 0 \quad i \quad \sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0.$$

Na taj način dobivamo dvije simultane linearne jednadžbe s dvije nepoznanice a i b, koje nazivamo *normalne jednadžbe.* Uvođenjem oznake

$$\bar{x} \coloneqq \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1.1}$$

te dijeljenjem obje strane s n, nakon sređivanja normalne jednadžbe postaju

$$a + b\bar{x} = \bar{y} \quad i \quad a\bar{x} + b\overline{x^2} = \overline{xy}.$$

Kada pomnožimo prvu sa $\bar{x}$ te ju oduzmemo od druge vrijedi

$$b = \frac{\overline{xy} - \bar{x}.\bar{y}}{\overline{x^2} - (\bar{x})^2},$$

iz čega slijedi $a = \bar{y} - b\bar{x}$. Vrijednost $\bar{x}$ iz (1.2) nazivamo *srednja vrijednost uzorka*, ili prosjek, od $x_1, \dots, x_n$. Analogno za $\bar{y}$. *Varijanca uzorka,* u oznaci ${s_x}^2$ *ili* s_{xx}, definirana je kao prosjek $(x_i - \bar{x})^2$. Koristeći linearnost uzorka te $\overline{\bar{x}.\bar{x}} = (\bar{x})^2$ vrijedi iduće

$${s_x}^2 = s_{xx} \coloneqq \overline{(x - \bar{x})^2} = \overline{x^2 - 2x.\bar{x} + \bar{x}^2} = \overline{(x^2)} - 2\bar{x}.\bar{x} + (\bar{x})^2.$$

Slično, *kovarijanca uzorka* x i y definirana je kao prosjek $(x - \bar{x})(y - \bar{y})$, u oznaci s_{xy}.

$$\begin{aligned} s_{xy} &= \overline{(x - \bar{x})(y - \bar{y})} = \overline{xy - x.\bar{y} - \bar{x}.y + \bar{x}.\bar{y}} = \\ &= \overline{(xy)} - \bar{x}.\bar{y} - \bar{x}.\bar{y} + \bar{x}.\bar{y} = \overline{(xy)} - \bar{x}.\bar{y}. \end{aligned}$$

Stoga je nagib b dan *koeficijentom korelacije*, odnosno omjerom kovarijance uzorka te varijance uzorka x

$$b = s_{xy}/s_{xx}.$$

Koristeći alternativnu notaciju pomoću sume kvadrata

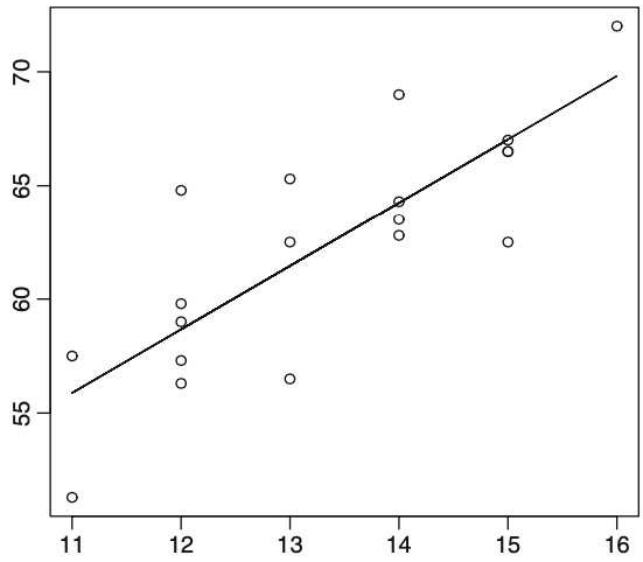
$$S_{xx} \coloneqq \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad S_{xy} \coloneqq \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

$$b = S_{xy}/S_{xx} \quad i \quad a = \bar{y} - b\bar{x}.$$

Pravac najmanjih kvadrata sa ovako dobivenim koeficijentima zove se *regresijski pravac uzorka*

$$y - \bar{y} = b(x - \bar{x}), \qquad b = s_{xy}/s_{xx} = S_{xy}/S_{xx}. \tag{1.2}$$

Primjer 1.2 *Tražimo pravac koji će se najbolje uklopiti u model gdje je y visina, izražena u inčima, na temelju dobi x, izražene u godinama, za sljedeće podatke:*
$x =$ (14, 13, 13, 14, 14, 12, 12, 15, 13, 12, 11, 14, 12, 15, 16, 12, 15, 11, 15),
$y =$ (69, 56.5, 65.3, 62.8, 63.5, 57.3, 59.8, 62.5, 62.5, 59, 51.3, 64.3, 56.3, 66.5, 72, 64.8, 67, 57.5, 66.5).



Slika 1.1 Dijagram rasprostranjenosti podataka iz Primjera 1.2
sa pripadajućim pravcem

Također, možemo izračunati S_{xx} i S_{xy} kao

$$S_{xx} = \sum x_i y_i - n\overline{xy}, \qquad S_{xy} = \sum {x_i}^2 - n\bar{x}^2.$$

Budući da $\sum x_i y_i = 15883,\ \bar{x} = 13.316,\ \bar{y} = 62.337,\ \sum {x_i}^2 = 3409,\ n = 19$, slijedi

$$b = \frac{S_{xy}}{S_{xx}} = \frac{15883 - 19(13.316)(62.337)}{3409 - 19(13.316^2)} = 2.787$$
$$a = 62.337 - 2.787(13.316) = 25.224.$$

Ovaj model sugerira da djeca rastu nešto manje od 3 inča godišnje. Grafički prikaz promatranih podataka te pripadajućeg pravca prikazan na Slici 1.1 izgleda poprilično razumno, podaci prate pravac uz određeno odstupanje.

1.1.1. Koeficijent korelacije uzorka

Definicija 1.3 *Neka je* s_{xy} *kovarijanca uzorka x i y, te* s_x *i* s_y *standardne devijacije uzorka x i y redom. Koeficijent korelacije uzorka* r *definiramo kao*

$$r = r_{xy} := \frac{s_{xy}}{s_x s_y}.$$

Za razliku od ostalih vrijednosti s kojima smo se do sada susreli, r nema dimenziju te se nalazi između -1 i 1, s time da je jednak -1 ili 1 ako i samo ako sve točke $(x_1, y_1), \dots, (x_n, y_n)$ leže na pravcu.
Koristeći $s_{xy} = r_{xy} s_x s_y$ *i* $s_{xx} = {s_x}^2$, regresijski pravac uzorka možemo alternativno zapisati kao

$$y - \bar{y} = b(x - \bar{x}), \qquad b = r_{xy} s_y / s_x. \tag{1.3}$$

Primijetimo da nagib pravca b ima isti predznak kao kovarijanca uzorka i koeficijent korelacije uzorka. Nagib će težiti ka nuli kada y i x nisu u korelaciji – posebno kada su nezavisni, biti će pozitivan (negativan) kada su x i y pozitivno (negativno) korelirani.

1.1.2. Primjena zakona velikih brojeva

Definicija 1.4 *Kažemo da niz slučajnih varijabli* $(X_n; n \in N)$ *konvergira gotovo sigurno (g.s.) prema slučajnoj varijabli X ako je*

$$\mathbb{P}\left\{\lim_{n\to\infty} X_n = X\right\} = 1.$$

To označavamo sa $X_n \xrightarrow{g.s.} X$, $(n \to \infty)$.

Definicija 1.5 *Kažemo da niz slučajnih varijabli* $(X_n; n \in N)$ *konvergira po vjerojatnosti prema slučajnoj varijabli X ako za svaki* $\varepsilon > 0$ *vrijedi*

$$\lim_{n\to\infty} \mathbb{P}\{|X_n - X| > \varepsilon\} = 0.$$

To označavamo sa $X_n \xrightarrow{\mathbb{P}} X$, $(n \to \infty)$.

Napomena 1.6 *Kažemo da je (g.s.) konvergencija jača od konvergencije po vjerojatnosti, odnosno da je konvergencija po vjerojatnosti slabija od (g.s.) konvergencije.*

Postoji nekoliko verzija zakona velikih brojeva.
Slabi zakon velikih brojeva (WLLN) daje konvergenciju po vjerojatnosti, dok jaki zakon velikih brojeva (SLLN) daje konvergenciju s vjerojatnošću jedan (ili gotovo sigurnu, oznaka *g.s.*).

Teorem 1.7 *(Slabi zakon velikih brojeva) Neka je* $x = (x_n; n \in \mathbb{N})$ *niz nezavisnih slučajnih varijabli takvih da je* $\mathbb{E}(x_n) = \mu$ *i* $\mathrm{var}(x_n) = \sigma^2$ *za svaki* $n \in \mathbb{N}$. *Tada*

$$\bar{x} \xrightarrow{\mathbb{P}} \mu, \qquad (n \to \infty).$$

Dokaz: [2, Teorem 8.7]

Teorem 1.8 *(Jaki zakon velikih brojeva) Neka je* $x = (x_n; n \in \mathbb{N})$ *niz nezavisnih jednako distribuiranih slučajnih varijabli takvih da je* $\mathbb{E}(x_n) = \mu$. *Tada*

$$\bar{x} \xrightarrow{g.s.} \mu, \qquad (n \to \infty).$$

Dokaz: [2, Teorem 8.10]

Sve ovo vrijedi na sličan način kada x zamijenimo sa y, x^2, y^2, xy, ukoliko oni imaju očekivanje. Tada je

$$s_x{}^2 = s_{xx} = \overline{x^2} - (\bar{x})^2 \to \mathbb{E}(x^2) - (\mathbb{E}x)^2 = \mathrm{var}(x),$$

varijanca populacije, također u oznaci $\sigma_x{}^2 = \sigma_{xx}$, a

$$s_{xy} = \overline{xy} - \bar{x}.\bar{y} \to \mathbb{E}(xy) - \mathbb{E}x\mathbb{E}y = \mathrm{cov}(x, y),$$

kovarijanca populacije, također u oznaci σ_{xy}.

Dakle, kako se veličina uzorka n povećava, regresijski pravac uzorka

$$y - \bar{y} = b(x - \bar{x}), \qquad b = s_{xy}/s_{xx}$$

teži pravcu

$$y - \mathbb{E}y = \beta(x - \mathbb{E}x), \qquad \beta = \sigma_{xy}/\sigma_{xx}. \tag{1.4}$$

Ovaj pravac nazivamo *regresijski pravac populacije.*
Postoji verzija koja uključuje korelaciju, pa je *koeficijent korelacije populacije*

$$\rho = \rho_{xy} := \frac{\sigma_{xy}}{\sigma_x \sigma_y}:$$

$$y - \mathbb{E}y = \beta(x - \mathbb{E}x), \qquad \beta = \rho_{xy}\sigma_y/\sigma_x. \tag{1.5}$$

1.2. Povijest regresije

Modernu eru na području regresije otvorio je Sir Francis Galton (1822. – 1911.). Zanimao se za inteligenciju te kako se ona nasljeđuje. Ali inteligencija, iako vitalno važna, nedostižan je pojam – ljudska inteligencija je beskonačno promjenjiva, i iako postoje numerička mjerenja (kvocijent inteligencije ili IQ), oni mogu poslužiti samo kao zamjena za samu inteligenciju. Galton je imao strast za mjerenjem te je odlučio proučavati nešto što se lako može izmjeriti; izabrao je ljudsku visinu. Promatrao je visinu roditelja (varijabla x) te njihovih potomaka (varijabla y). Kada je prikazao podatke u tabelarnoj formi, primijetio je da dobiva obris elipse, to jest, čini se da kvadrati u *(x, y)* ravnini koji sadrže jednake brojeve leže otprilike na elipsama. Objašnjenje za to leži u bivarijantnoj normalnoj distribuciji koju ćemo analizirati u nastavku. Ono što je najrelevantnije jest Galtonova interpretacija regresijskog pravca uzorka i populacije ((1.3) i (1.5)). Pošto su $\sigma_x \, i \, \sigma_y$ mjere varijabilnosti u roditelja i potomaka, a nemamo razloga misliti da se mijenja, Galton pojednostavljuje (1.5) na:

$$y - \mathbb{E}y = \rho_{xy}(x - \mathbb{E}x) \tag{1.6}$$

Otuda Galtonovo tumačenje: za svaki inč visine iznad (ili ispod) prosjeka, roditelji svojoj djeci prenose prosječno ρ inča, gdje je ρ populacijski koeficijent korelacije između visine roditelja i visine potomaka. Sljedeća generacija uvest će dodatan faktor ρ, tako da će roditelji prenijeti ponovo, u prosjeku, ρ^2 inča svojim unucima. Ovo postaje ρ^3 za praunuke, i tako dalje. Stoga za svaki inč iznad (ili ispod) prosjeka, roditelji prenose svojim potomcima nakon n generacija prosječno ρ^n inča visine. Sada znamo da je $0 < \rho < 1$ (strogo pozitivan jer se prenose geni za visinu, a visina roditelja i potomaka su pozitivno korelirani; te različit od 1 jer bi to značilo da je visina roditelja potpuno zavisna o visini potomaka, što nije slučaj). Dakle,

$$\rho^n \to 0 \qquad (n \to \infty)$$

to jest, učinak svakog inča visine iznad ili ispod prosjeka prigušuje se sa sljedećim generacijama i nestaje u limesu. Galton je to sažeo kao regresiju prema očekivanju.

1.3. Bivarijantna normalna distribucija

Prisjetimo se još jednog ključnog pojma iz statistike.

Definicija 1.9 *Normalna distribucija, u oznaci* $N(\mu, \sigma^2)$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

gdje je očekivanje $\mathbb{E}X = \mu$ *i varijanca* $\mathrm{var}X = \sigma^2$.

Drugi ključan pojam je *linearna regresija metodom najmanjih kvadrata* kao što smo opisali u prethodnom odjeljku. Pišemo

$$\phi(x) \coloneqq \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}x^2\right\}$$

za gustoću jedinične normalne varijable.
Iduće što nas zanima je dvodimenzionalan skup podataka.

Baš kao što nam je u jednoj dimenziji trebalo dva parametra, tako će nam u dvije dimenzije trebati pet parametara. Promotrimo bivarijantnu gustoću:

$$f(x,y) \coloneqq c\,\exp\left\{-\frac{1}{2}Q(x,y)\right\},$$

gdje je c konstanta, Q pozitivno definitna kvadratna forma od x i y. Specijalno,

$$c = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}},$$

$$Q = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right].$$

Ovdje su $\sigma_i > 0, \mu_i$ su realni, $-1 < \rho < 1$.
Svojstva:

1) *f(x,y)* je dvodimenzionalna funkcija gustoće sa marginalnim jednodimenzionalnim funkcijama gustoće $f_1(x), f_2(y)$ pa možemo pisati

$$f(x,y) = f_{X,Y}(x,y), \qquad f_1(x) = f_X(x), \qquad f_2(y) = f_Y(y).$$

2) $X,\ Y$ su normalne slučajne varijable: $X \sim N(\mu_1, {\sigma_1}^2),\ Y \sim N(\mu_2, {\sigma_2}^2)$.

$$f_1 = f_X, \qquad f_2 = f_Y.$$

3) $\mathbb{E}X = \mu_1,\ \mathbb{E}Y = \mu_2,\ \mathrm{var}\,X = {\sigma_1}^2,\ \mathrm{var}\,Y = {\sigma_2}^2$

4) Uvjetna distribucija od y uz dano $X = x$ je

$$N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), \qquad {\sigma_2}^2(1-\rho^2)\right).$$

5) Uvjetno očekivanje $\mathbb{E}(Y|X = x)$ je linearno po x i vrijedi:

$$\mathbb{E}(Y|X = x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1).$$

6) Uvjetna varijanca od Y uz dano $X = x$ *je*

$$\text{var}(Y|X = x) = {\sigma_2}^2(1 - \rho^2).$$

7) Kovarijanca je definirana kao

$$\begin{aligned}\text{cov}(X, Y) &:= \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}[(X - \mu_1)(Y - \mu_2)],\\ &= \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y),\end{aligned}$$

a koeficijent korelacije ρ ili $\rho(X, Y)$ definiran je kao

$$\rho = \rho(X, Y) := \frac{\text{cov}(X, Y)}{\sqrt{\text{var}X}\sqrt{\text{var}Y}} = \frac{\mathbb{E}[(X - \mu_1)(Y - \mu_2)]}{\sigma_1\sigma_2}$$

8) Ako je $(X, Y)^T$ bivarijantna normalna, koeficijent korelacije od $X,\ Y$ je ρ.

9) Funkcija generiranja momenta dana je sa

$$M_{X,Y}(t_1, t_2) = M(t_1, t_2) =$$

$$= \exp\left(\mu_1 t_1 + \mu_2 t_2 + \frac{1}{2}[{\sigma_1}^2 {t_1}^2 + 2\rho\sigma_1\sigma_2 t_1 t_2 + {\sigma_2}^2 {t_2}^2\right] \tag{1.7}$$

10) X i Y su nezavisne ako i samo ako $\rho = 0$.

1.4. Maksimalna vjerodostojnost i najmanji kvadrati

Iz svojstva 4) bivarijantne normalne distribucije, slijedi da je uvjetna distribucija od y uz dano $X = x$

$$N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \qquad {\sigma_2}^2(1 - \rho^2)\right).$$

Stoga se y rastavlja na dvije komponente, linearni trend od x – sustavni dio, i normalnu pogrešku sa očekivanjem 0 i varijancom koja je konstanta – slučajni dio. Mijenjajući zapis, ovo možemo zapisati kao

$$y = a + bx + \epsilon, \qquad \epsilon \sim N(0, \sigma^2).$$

Ukoliko razmatramo n vrijednosti prediktorske varijable x, slično zapisujemo

$$y_i = a + bx_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2).$$

Da bismo završili specifikaciju modela, moramo specificirati zavisnost ili koreliranost pogrešaka $\epsilon_1, \dots, \epsilon_n$. To možemo učiniti na razne načine, ali ovdje ćemo promatrati restringirani najjednostavniji ali i najvažniji slučaj, a to je gdje su ϵ_i nezavisne jednako distribuirane slučajne varijable:

$$y_i = a + bx_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2). \tag{1.8}$$

Ovo je osnovni model za jednostavnu linearnu regresiju.
Pomoću [4], uvodimo iduću definiciju.

Definicija 1.10 *Neka je* $y_i = a + bx_i + \epsilon_i$, *i* $\epsilon\{1, \dots, n\}$ *osnovni model za jednostavnu linearnu regresiju. Neka je svaka* y_i *normalno distribuirana te neka su* ϵ_i *nezavisne jednako distribuirane slučajne varijable* $\epsilon_i \sim N(0, \sigma^2)$. *Vjerodostojnost L definiramo kao funkciju gustoće od* $y_1, \dots, y_n$

$$\begin{aligned} L(a, b, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}n}} \prod\nolimits_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}(y_i - a - bx_i)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}n}} \exp\left\{-\frac{1}{2\sigma^2}\sum\nolimits_{i=1}^{n}(y_i - a - bx_i)^2\right\}. \end{aligned}$$

Fisher je predložio da kao procjene parametara odaberemo vrijednosti koje maksimiziraju vjerodostojnost. Ovo je *metoda maksimalne vjerodostojnosti* koja rezultira procjeniteljem maksimalne vjerodostojnosti, oznaka MLE.
Obzirom da je logaritmiranje strogo rastuća funkcija, maksimiziranje funkcije L jednako je maksimiziranju funkcije $\ell(a, b, \sigma) \coloneqq \log L(a, b, \sigma)$. Budući da je

$$\ell(a, b, \sigma^2) \coloneqq \log L(a, b, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum\nolimits_{i=1}^{n}(y_i - a - bx_i)^2,$$

maksimiziranje obzirom na a i b, isto je kao i minimiziranje zbroja kvadrata $SS \coloneqq \sum_{i=1}^{n}(y_i - a - bx_i)^2$, baš kao u metodi najmanjih kvadrata.

Teorem 1.11 *Za normalan model (1.8), metoda najmanjih kvadrata i metoda maksimalne vjerodostojnosti su ekvivalentni načini procjene parametara a i b.*

Dokaz: [3]

Ostaje nam procjena parametra σ^2, odnosno varijance. Korištenjem maksimalne vjerodostojnosti

$$\frac{\partial \ell}{\partial \sigma}(a, b, \sigma^2) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = 0,$$

ili

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2.$$

U maksimumu, a i b imaju svoje maksimalne vrijednosti $\hat{a}\, i\, \hat{b}$, a tada je maksimizirajuća vrijednost $\hat{\sigma}$ dana sa

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{a} - \hat{b}x_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Primijetimo da suma kvadrata SS iznad uključuje nepoznate parametre a i b. Budući da su oni nepoznati, iz podataka se ne može numerički izračunati ova suma kvadrata. U trećem poglavlju susrest ćemo se s višedimenzionalnim analogonima svega ovoga, te ćemo ih obraditi matičnom algebrom.

1.5. Rastav varijance

Prisjetimo se regresijskog pravca uzorka

$$y = \bar{y} + b(x - \bar{x}), \qquad b = s_{xy}/s_{xx} = S_{xy}/S_{xx}.$$

Pitamo se koliko varijacija od y objašnjeno znanjem o x to jest regresijom.
Podaci su y_i. Procijenjene vrijednosti su $\hat{y}_i$. Zapišimo

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}),$$

kvadrirajmo obje strane te zbrojimo. S lijeve strane imamo

$$SS := \sum_{i=1}^{n} (y_i - \bar{y})^2$$

što je *ukupna suma kvadrata* ili skraćeno *suma kvadrata.*
S desne strane, dobivamo tri člana:

$$SSR := \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

kojeg nazivamo *suma kvadrata za regresiju,*

$$SSE := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

kojeg nazivamo *suma kvadrata za pogreške* (pošto ova suma kvadrata mjeri grešku između vrijednosti na regresijskom pravcu i podataka),
te mješoviti član

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = n\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = n.\overline{(y - \hat{y})(y - \bar{y})}.$$

Teorem 1.12 *Vrijedi*

$$SS = SSR + SSE$$

Dokaz: Po (1.2)

$$\hat{y}_i - \bar{y} = b(x_i - \bar{x}), \qquad uz\ b = S_{xy}/S_{xx} = S_{xy}/{S_x}^2,$$
$$y_i - \hat{y} = (y_i - \bar{y}) - b(x_i - \bar{x}).$$

Dakle, u gornjoj jednakosti imamo n puta

$$\frac{1}{n}\sum_{i=1}^{n} b(x_i - \bar{x})[(y_i - \bar{y}) - b(x_i - \bar{x})] = bS_{xy} - b^2{S_x}^2 = b\left(S_{xy} - b{S_x}^2\right) = 0,$$

pošto je $b = S_{xy}/{S_x}^2$. Iz navedenog slijedi tvrdnja.

□

U pojmovima koeficijenta korelacije uzorka r^2, ovo daje iduću tvrdnju.

Teorem 1.13 *Vrijedi*

$$r^2 = SSR/SS, \qquad 1 - r^2 = SSE/SS.$$

Dokaz: Dovoljno je dokazati iduće.

$$\frac{SSR}{SS} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\sum b^2(x_i - \bar{x})^2}{\sum(y_i - \bar{y})^2} = \frac{b^2 {S_x}^2}{{S_y}^2} = \frac{{S_{xy}}^2}{{S_x}^4} \cdot \frac{{S_x}^4}{{S_y}^2} = \frac{{S_{xy}}^2}{{S_x}^2 {S_y}^2} = r^2,$$

uz $b = S_{xy}/{S_x}^2$.

□

Interpretiramo $r^2 = SSR/SS$ kao udio varijabilnosti od y koji se objašnjava poznavanjem x, to jest regresijom, dok je $1 - r^2$ „neobjašnjeni dio“, to jest pogreška. Prisjetimo se da po Zakonu velikih brojeva limes od r^2 teži ka ρ^2.

.

2. Analiza varijance (ANOVA)

Dok linearna regresija iz prvog poglavlja seže u devetnaesto stoljeće, analiza varijance datira iz dvadesetog stoljeća, u Fisherovom primijenjenom radu motiviranom poljoprivrednim problemima o kojem ćemo reći nešto više kasnije. Ovo poglavlje započinjemo nekim potrebnim pojmovima; posebnim distribucijama statistike potrebnim za teoriju malih uzoraka: χ^2 distribucija $\chi^2(n)$, Fisherova F-distribucija *F(m, n)*, te nezavisnosti očekivanja i varijanci normalnog uzorka. Analiza varijance će nam pomoći da u trećem poglavlju generaliziramo linearnu regresiju iz prvog poglavlja na višestruku regresiju.

2.1. χ^2 distribucija

Za početak ćemo definirati χ^2 *distribuciju* $\chi^2(n)$ s *n* stupnjeva slobode, a kasnije ćemo navesti neka njena svojstva.

Definicija 2.1 χ^2 *distribuciju s n stupnjeva slobode definiramo kao*

$$X_1^2 + \cdots + X_n^2,$$

gdje su $X_i \sim N(0,1)$ nezavisne jednako distribuirale slučajne varijable.

Prisjetimo se da je (1.7) funkcija generiranja momenta, te iz definicije *Gamma funkcije*

$$\Gamma(t) \coloneqq \int_0^\infty e^{-x} x^{t-1}\, dx \qquad (t > 0).$$

Parcijalnom integracijom lako se provjeri da

$$\Gamma(n+1) = n! \qquad (n = 0, 1, 2, \ldots),$$

što će nam biti potrebno u nastavku.

Teorem 2.2 *Za* χ^2 *distribuciju* $\chi^2(n)$ *s n stupnjeva slobode vrijedi:*

i) *očekivanje je jednako broju stupnjeva slobode n, a varijanca je jednaka dvostrukom n*

ii) *funkcija generiranja momenta je*

$$M(t) = 1/(1-2t)^{\frac{1}{2}n} \quad za \quad t < \frac{1}{2},$$

iii) *funkcija gustoće dana je sa*

$$f(x) = \frac{1}{2^{\frac{1}{2}n}\Gamma\left(\frac{1}{2}n\right)}.x^{\frac{1}{2}n-1}exp\left(-\frac{1}{2}x\right) \qquad (x > 0).$$

Dokaz: Vidi [5, Teorem 2.1].

Propozicija 2.3 (*Svojstvo zbrajanja*) *Ako su* X_1 *i* X_2 *nezavisne slučajne varijable iz, redom,* $\chi^2(n_1)$ *i* $\chi^2(n_2)$*, tada je* $X_1 + X_2$ *iz* $\chi^2(n_1 + n_2)$*.*

Dokaz: Neka su

$$X_1 = {U_1}^2 + \cdots + {U_{n1}}^2, \qquad X_2 = {U_{n1+1}}^2 + \cdots + {U_{n1+n2}}^2$$

takve da su U_i nezavisne jednako distribuirane slučajne varijable iz jedinične normalne razdiobe. Slijedi, $X_1 + X_2 = {U_1}^2 + \cdots + {U_{n1}}^2 + {U_{n1+1}}^2 + \cdots + {U_{n1+n2}}^2$, pa je $X_1 + X_2$ *iz* $\chi^2(n_1 + n_2)$.

□

Propozicija 2.4 (*Svojstvo oduzimanja) Ako je* $X = X_1 + X_2$*, te su* X_1 *i* X_2 *nezavisne,* $X \sim \chi^2(n_1 + n_2)$ *i* $X_1 \sim \chi^2(n_1)$*, tada je* $X_2 \sim \chi^2(n_2)$*.*

Dokaz: X je nezavisna suma od X_1 i X_2, stoga je funkcija generiranja momenta od *X* jednaka produktu funkcija generiranja momenta od X_1 i X_2. Te su funkcije za *X* i X_1 redom $(1-2t)^{-\frac{1}{2}(n1+n2)}, (1-2t)^{-\frac{1}{2}n1}$. Dijeljenjem slijedi da je funkcija generiranja momenta za X_2 upravo $(1-2t)^{-\frac{1}{2}n2}$. Dakle, $X_2 \sim \chi^2(n_2)$.

□

2.2. Fisherova F-distribucija

Definicija 2.5 *Ako su U i V dvije nezavisne slučajne varijable takve da* $U \sim \chi^2(m)$, $V \sim \chi^2(n)$, *tada slučajna varijabla*

$$F := \frac{U/m}{V/n}$$

dolazi iz Fisherove ili F-distribucije s (m, n) stupnjeva slobode i pišemo $F \sim F(m, n)$.

Također je poznata kao Fisherova distribucija omjera varijance.
Prije nego zapišemo njenu funkciju gustoće, definirajmo *Beta funkciju*

$$B(\alpha, \beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx.$$

Prema Eulerovom integralu za Beta funkciju,

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Tada se može pokazati da je gustoća od *F(m, n)* dana sa

$$f(x) = \frac{m^{\frac{1}{2}m} n^{\frac{1}{2}n}}{B\left(\frac{1}{2}m, \frac{1}{2}n\right)} \cdot \frac{x^{\frac{1}{2}(m-2)}}{(mx+n)^{\frac{1}{2}(m+n)}} \qquad (m, n > 0, \qquad x > 0).$$

Dvije su važne značajke ove gustoće. Prva je da se oko nule ponaša kao potencija $x^{\frac{1}{2}(m-2)}$, a kako ide prema beskonačnosti kao potencija $x^{-\frac{1}{2}n}$, funkcija je glatka te unimodalna (ima jedan ekstrem). Druga je ta da se, kao i ostale distribucije u statistici, vrijednosti po postocima prikazani u tablici. Korištenje tablice F-distribucije uključuje značajku da imamo dva stupnja slobode (za razliku od χ^2 ili Studentove t-distribucije gdje imamo jedan stupanj slobode) te je vrlo bitno uzeti ih ispravnim redoslijedom.

2.3. Ortogonalnost

Prisjetimo se da je kvadratna, regularna (n x n) matrica A *ortogonalna* ako vrijedi:

$$A^{-1} = A^T.$$

Sada želimo pokazati da ortogonalna transformacija čuva svojstvo nezavisnosti varijable iz $N(0, \sigma^2)$.

Teorem 2.6 *(Teorem o ortogonalnosti) Ako je* $X = (X_1, \dots, X_n)^T$ *n-vektor čije su komponente nezavisne, normalno distribuirane slučajne varijable s očekivanjem 0 i varijancom* σ^2*, te uvedemo supstituciju*

$$Y \coloneqq AX$$

gdje je matrica A ortogonalna, tada su komponente Y_i *od Y ponovo nezavisne, normalno distribuirane slučajne varijable s očekivanjem 0 i varijancom* σ^2*.*

Dokaz: Prema [5], koristimo formulu za Jakobijana. Naša supstitucija dana je sa $Y_i = \sum_{j=1}^{n} a_{ij} X_j$, pa vidimo da je $\partial Y_i / \partial X_i = a_{ij}$, pa je Jakobijan $\partial Y / \partial X = |A|$.
Pošto je A ortogonalna,

$$AA^T = AA^{-1} = I.$$

Gledajući determinantu vidimo da je determinanta od A i od A^T jednaka 1. Budući da ortogonalnost čuva svojstvo duljine,

$$\sum\nolimits_1^n Y_i^2 = \sum\nolimits_1^n X_i^2.$$

Gustoća od $(X_1, \dots, X_n)$ je, radi nezavisnosti, produkt marginalnih gustoća te je ona

$$f(x_1, \dots, x_n) = \prod\nolimits_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} x_i^2\right\} = \frac{1}{(2\pi)^{\frac{1}{2}n}} \exp\left\{-\frac{1}{2} \sum\nolimits_1^n x_i^2\right\}.$$

Iz ovoga i formule za Jakobijana, dobivamo formulu za gustoću od $(Y_1, \dots, Y_n)$

$$f(y_1, \dots, y_n) = \frac{1}{(2\pi)^{\frac{1}{2}n}} \exp\left\{-\frac{1}{2}\sum\nolimits_1^n y_i^2\right\} = \prod\nolimits_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}y_i^2\right\}.$$

Ali ovo je upravo gustoća n nezavisnih jediničnih normalnih varijabli, stoga smo dobili da su upravo $(Y_1, \dots, Y_n)$ nezavisne, normalno distribuirane slučajne varijable što smo i željeli pokazati.

□

2.4. Očekivanje i varijanca normalnog uzorka

Za $X_1, \dots, X_n$ nezavisne jednako distribuirane slučajne varijable s očekivanjem μ i varijancom σ^2 pišemo

$$\overline{X} := \frac{1}{n}\sum\nolimits_1^n X_i$$

za *očekivanje uzorka* i

$$S^2 := \frac{1}{n}\sum\nolimits_1^n \left(X_i - \overline{X}\right)^2$$

za *varijancu uzorka.*

Teorem 2.7 *Ako su* $X_1, \dots, X_n$ *nezavisne jednako distribuirane slučajne varijable iz* $N(\mu, {\sigma_1}^2)$ *tada:*

i) *očekivanje uzorka* $\overline{X}$ *i varijanca uzorka* S^2 *su nezavisne,*
ii) $\overline{X} \sim N(\mu, {\sigma_1}^2/n)$,
iii) $nS^2/\sigma^2 \sim \chi^2(n-1)$.

Dokaz: Vidi [5, Teorem 2.4].

Lema 2.8 *(Fisherova lema) Neka su* $X_1, \dots, X_n$ *nezavisne jednako distribuirane slučajne varijable iz* $N(0, {\sigma_1}^2)$. *Neka je*

$$Y_i = \sum_{j=1}^{n} c_{ij} X_j \qquad (i = 1, \dots, p, \qquad p < n),$$

gdje su redak vektora $(c_{i1}, \dots, c_{in})$ *ortogonalni za i=1, … , p. Ako je*

$$S^2 = \sum_1^n X_i^2 - \sum_1^p Y_i^2,$$

onda

i) S^2 *je nezavisan obzirom na* $Y_1, \dots, Y_p$,
ii) $S^2 \sim \chi^2(n-p)$.

Dokaz: Proširimo p x n matricu (c_{ij}) do n x n ortogonalne matrice $C = (c_{ij})$ Gram-Schmidt ortogonalizacijom. Zatim stavimo

$$Y \coloneqq CX,$$

te smo tako definirali $Y_1, \dots, Y_p$ i $Y_{p+1}, \dots, Y_n$. Budući da je C ortogonalna, $Y_1, \dots, Y_n$ su nezavisne jednako distribuirane slučajne varijable iz $N(0, {\sigma_1}^2)$ i vrijedi $\sum_1^n Y_i^2 = \sum_1^n X_i^2$.
Dakle,

$$S^2 = \left(\sum_1^n - \sum_1^p\right) Y_i^2 = \sum_{p+1}^n Y_i^2$$

je nezavisna obzirom na $Y_1, \dots, Y_p$ i $S^2 \sim \chi^2(n-p)$.

□

2.5. Analiza varijance

Iz [5], znamo da analiza varijance seže sve do ranog Fisherovog rada o matematičkoj genetici iz 1918. Fisherova motivacija bila je usporediti prinose

nekoliko sorti usjeva ili jednog usjeva tretiranog različitim gnojivima. Krenuo je u usporedbu srednjih vrijednosti analizirajući varijabilnost.

Kako bismo usporedili dva normalna uzorka, koristimo *Studentov t-test.*

Označimo srednju vrijednost, to jest očekivanje, i-te sorte sa μ_i za $(j = 1,\dots,n_i,\ i = 1,\dots,r)$. Za svaki i kreiramo n_i nezavisnih očitanja X_{ij}. One su nezavisne te pretpostavimo da su normalno distribuirane s nepoznatom varijancom σ^2:

$$X_{ij} \sim N(\mu_i, \sigma^2) \qquad (j = 1,\dots,n_i, \qquad i = 1,\dots,r).$$

Za ukupnu veličinu uzorka pišemo:

$$n := \sum\nolimits_1^r n_i$$

Uz dvije oznake, i i j, koristimo i punu točku koja indicira da je na toj poziciji sufiks koji je prosjek. Za srednju vrijednost i-te skupine pišemo

$$X_{i\bullet}, \quad \text{ili} \quad \overline{X_i}, := \frac{1}{n_i}\sum\nolimits_{j=1}^{n_i} X_{ij} \qquad (i = 1,\dots,r),$$

za ukupnu srednju vrijednost pišemo

$$X_{\bullet\bullet}, \quad \text{ili} \quad \overline{X}, := \frac{1}{n}\sum\nolimits_{i=1}^{r}\sum\nolimits_{j=i}^{n_i} X_{ij} = \frac{1}{n}\sum\nolimits_{i=1}^{r} n_i X_{i\bullet},$$

a za varijancu i-tog uzorka pišemo

$$S_i^2 := \frac{1}{n_i}\sum\nolimits_{j=1}^{n_i}\left(X_{ij} - X_{i\bullet}\right)^2.$$

Definiramo *ukupnu sumu kvadrata*

$$SS := \sum\nolimits_{i=1}^{r}\sum\nolimits_{j=1}^{n_i}\left(X_{ij} - X_{\bullet\bullet}\right)^2 = \sum\nolimits_i\sum\nolimits_j\left[\left(X_{ij} - X_{i\bullet}\right) + \left(X_{i\bullet} - X_{\bullet\bullet}\right)\right]^2.$$

Teorem 2.9 *Pod gornjim uvjetima i nultom hipotezom* H_0 *da nema razlike u očekivanjima od tretmana, imamo rastav varijance*

$$SS = SSE + SST,$$

gdje je $SS/\sigma^2 \sim x^2(n-1)$, $SSE/\sigma^2 \sim x^2(n-r)$, *i* $SSE/\sigma^2 \sim x^2(r-1)$.

Dokaz: Pošto je $\sum_j (X_{ij} - X_{i\bullet}) = 0$, ako proširimo gornji kvadrat, mješoviti članovi nestaju, te imamo

$$SS = \sum_i \sum_j (X_{ij} - X_{i\bullet})^2 + \sum_i \sum_j (X_{ij} - X_{i\bullet})(X_{i\bullet} - X_{\bullet\bullet}) + \sum_i \sum_j (X_{i\bullet} - X_{\bullet\bullet})^2$$

$$= \sum_i \sum_j (X_{ij} - X_{i\bullet})^2 + \sum_i \sum_j (X_{i\bullet} - X_{\bullet\bullet})^2$$

$$= \sum_i n_i S_i^2 + \sum_i n_i (X_{i\bullet} - X_{\bullet\bullet})^2.$$

Prvi član s desne je mjera je za varijabilnost unutar skupina, a drugi mjera za varijabilnost među skupinama. Prvog zovemo *suma kvadrata pogreške, SSE,* ili *suma kvadrata reziduala,* a drugog *suma kvadrata tretmana* (ili *suma kvadrata među grupama).* Dakle,

$$SS = SSE + SST,$$

gdje je

$$SSE := \sum_i n_i S_i^2, \qquad SST := \sum_i n_i (X_{i\bullet} - X_{\bullet\bullet})^2.$$

Neka je H_0 nulta hipoteza bez učinka tretiranja:

$$H_0: \quad \mu_i = \mu \quad (i = 1, \ldots, r).$$

Ako je H_0 istinita, imamo samo jedan veliki uzorak veličine n, koji dolazi iz distribucije $N(\mu, \sigma^2)$ i to

$$SS/\sigma^2 = \frac{1}{\sigma^2}\sum_i \sum_j (X_{i\bullet} - X_{\bullet\bullet})^2 \sim X^2(n-1).$$

Posebno,

$$\mathbb{E}[SS/(n-1)] = \sigma^2.$$

Bilo da je H_0 istinita ili ne,

$$n_i S_i^2/\sigma^2 = \frac{1}{\sigma^2}\sum_j \left(X_{ij} - X_{i\bullet}\right)^2 \sim x^2(n_i - 1).$$

Dakle, iz svojstva zbrajanja χ^2 distribucije

$$SSE/\sigma^2 = \sum_i n_i\, S_i^2/\sigma^2 = \frac{1}{\sigma^2}\sum_i \sum_j \left(X_{ij} - X_{i\bullet}\right)^2 \sim x^2(n_i - 1),$$

budući da je $n = \sum_i n_i$, slijedi $\sum_{i=1}^{r}(n_i - 1) = n - r$.
Posebno,

$$\mathbb{E}[SSE/(n-r)] = \sigma^2.$$

Nadalje, $SST := \sum_i n_i\,(X_{i\bullet} - X_{\bullet\bullet})^2$, gdje je $X_{\bullet\bullet} = \frac{1}{n}\sum_i n_i\, X_{i\bullet}$, $\quad SSE := \sum_i n_i\, S_i^2$.

Sada je S_i^2 nezavisan od $X_{i\bullet}$, budući da su to upravo varijanca i očekivanje i-tog uzorka, a čiju smo nezavisnost dokazali u Tm 2.7. Također, S_i^2 je nezavisan od $X_{j\bullet}$ za $j \neq i$, pošto su formirani iz nezavisnih uzoraka. U konačnici, S_i^2 je nezavisan o svim $X_{j\bullet}$, kako u njihovom prosjeku $X_{\bullet\bullet}$, tako i o SST (funkciji od $X_{j\bullet}$ i $X_{\bullet\bullet}$).
Dakle, SSE je također nezavisan od SST.
Sada možemo koristiti svojstvo oduzimanja iz Propozicije 2.4. Imamo nezavisnu sumu

$$SS/\sigma^2 = SSE/\sigma^2 + SST/\sigma^2.$$

Lijeva strana je $\chi^2(n-1)$, dok je na desnoj strani prvi član $\chi^2(n-r)$. Dakle, drugi član na desnoj strani mora biti $\chi^2(r-1)$. Iz navedenog slijedi tvrdnja.

□

Kada imamo sumu kvadrata, χ^2 distribuiranu, te ju podijelimo njenim stupnjem slobode, omjer koji dobijemo nazivamo *srednja vrijednost sume kvadrata* te ga označavamo sa *MS*

$$MS := SS/\mathrm{df}\,(SS) = SS/(n-1).$$

Srednje vrijednosti sume kvadrata tretmana te pogreške su, redom,

$$MST := SST/\mathrm{df}\,(SST) = SST/(r-1),$$
$$MSE := SSE/\mathrm{df}\,(SSE) = SSE/(n-r).$$

Bilo da je H_0 istina ili ne, vrijedi

$$\mathbb{E}[MSE] = \mathbb{E}[SSE]/(n-r) = \sigma^2,$$
$$\mathbb{E}[MS] = \mathbb{E}[SS]/(n-1) = \sigma^2,$$
$$\mathbb{E}[MST]/(r-1) = \sigma^2.$$

Formiramo *F-statistiku:* $F := MST/MSE$.
Ako vrijedi nulta hipoteza, *F* ima distribuciju *F(r – 1, n – r).* Fisher je shvatio da pomoću ove distribucije i statistike možemo testirati istinitost nulte hipoteze.
Bilo da je H_0 istina ili ne,

$$\begin{aligned} SST &= \sum_i n_i (X_{i\bullet} - X_{\bullet\bullet})^2 \\ &= \sum_i n_i X_{i\bullet}^2 - 2X \sum_i n_i X_{i\bullet} + X_{\bullet\bullet}^2 \sum_i n_i \\ &= \sum_i n_i X_{i\bullet}^2 - n\, X_{\bullet\bullet}^2, \end{aligned}$$

jer $\sum_i n_i X_{i\bullet} = n\, X_{\bullet\bullet}$ *i* $\sum_i n_i = n$.
Tako je:

$$\begin{aligned} \mathbb{E}[SST] &= \sum_i n_i \mathbb{E}[X_{i\bullet}^2] - n\mathbb{E}[X_{\bullet\bullet}^2] \\ &= \sum_i n_i [\mathrm{var}(X_{i\bullet}) + (\mathbb{E}X_{i\bullet})^2] - n[\mathrm{var}(X_{\bullet\bullet}) + (\mathbb{E}X_{\bullet\bullet})^2]. \end{aligned}$$

Ali $\mathrm{var}(X_{i\bullet}) = \sigma^2/n_i$, pa je

$$\begin{aligned} \mathrm{var}(X_{\bullet\bullet}) &= \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{r} n_i X_{i\bullet}\right) \\ &= \frac{1}{n^2}\sum_1^r n_i^2 \mathrm{var}(X_{i\bullet}) \end{aligned}$$

$$= \frac{1}{n^2}\sum_1^r n_i^2\,\sigma^2/n_i = \sigma^2/n$$

jer je $(\sum_i n_i = n)$. Pišemo

$$\overline{\mu} := \frac{1}{n}\sum_i n_i\mu_i = \mathbb{E}X_{\bullet\bullet} = \mathbb{E}\left[\frac{1}{n}\sum_i n_i X_{i\bullet}\right],$$

$$\begin{aligned}\mathbb{E}(SST) &= \sum_1^r n_i\left[\frac{\sigma^2}{n_i} + \mu_i^2\right] - \left[\frac{\sigma^2}{n} + \overline{\mu}^2\right] \\ &= (r-1)\sigma^2 + \sum_i n_i\mu_i^2 - n\overline{\mu}^2 \\ &= (r-1)\sigma^2 + \sum_i n_i(\mu_i - \overline{\mu})^2.\end{aligned}$$

Iz ovog slijedi nejednakost

$$\mathbb{E}(SST) \geq (r-1)\sigma^2,$$

a jednakost vrijedi ako i samo ako $\mu_i = \overline{\mu},\ (i = 1,\ldots,r)$, to jest ako je nulta hipoteza istinita.

Teorem 2.10 *Kada je nulta hipoteza* H_0 *istinita (odnosno,* $\mu_1, \ldots, \mu_r$ *su jednaki), F-statistika* $F := \frac{MST}{MSE} = (SST/(r-1))/(SSE/(n-r))$ *ima F-distribuciju* $F(r-1, n-r)$.

Kada je nulta hipoteza H_0 *neistinita, F se povećava. Dakle, velike vrijednosti od F su dokaz neistinitosti nulte hipoteze, a mi testiramo* H_0 *koristeći jednostrani F-test, odbacujući na razini značajnosti* α *ako je F prevelik, to jest, s kritičnim područjem*

$$F > F_{tab} = F_\alpha(r-1, n-r).$$

Jednadžbe modela za jednosmjernu ANOVU.

$$X_{ij} = \mu_i + \epsilon_{ij} \qquad (i = 1,\ldots,r, \qquad j = 1,\ldots,r),\ \ \epsilon_{ij} \sim N(0,\sigma^2).$$

Izračuni.

U svakom izračunu koji uključuje varijance potrebno je poništavanje, što nam je važno. To proizlazi iz formule za računanje varijance

$$\sigma^2 := \mathbb{E}[(X - \mathbb{E}X)^2] = \mathbb{E}[X^2] - (\mathbb{E}X)^2$$

te iz

$$S^2 := \overline{\left(X - \overline{X}\right)^2} = \overline{X^2} - \overline{X}^2.$$

Pišemo T, T_i za ukupni zbroj te zbroj po grupama i definiramo ih kao

$$T := \sum_i \sum_j X_{ij}, \quad T_i := \sum_j X_{ij},$$

a uz $X_{\bullet\bullet} = T/n, \;\; nX_{\bullet\bullet}^2 = T^2/n$:

$$SS = \sum_i \sum_j X_{ij}^2 - T^2/n,$$

$$SST = \sum_i T_i^2/n_i - T^2/n,$$

$$SSE = SS - SST = \sum_i \sum_j X_{ij}^2 - \sum_i T_i^2/n_i.$$

Ove formule pomažu u smanjenju pogreške zaokruživanja i najlakše ih je koristiti ako ručno provodimo analizu varijance. Uobičajeno je prikazati *output* analize varijance pomoću ANOVA tablice, kao što je prikazano u Tablici 2.1.

Izvor	*df*	*SS*	S^2	*F*
Tretmani	$r-1$	SST	$MST = SST/(r-1)$	MST/MSE
Rezidual/Greška	$n-r$	SSE	$MSE = SSE/(n-r)$	
Ukupno	$n-1$	SS		

Tablica 2.1 Jednosmjerna ANOVA tablica

Primjer 2.11 *Ovaj primjer nam pokazuje kako ručno izračunati tablicu analize varijance.* Podaci u Tablici 2.2. dolaze iz poljoprivrednog pokusa.

Gnojivo	Prinos
A	14.5, 12.0, 9.0, 6.5
B	13.5, 10.0, 9.0, 8.5
C	11.5, 11.0, 14.0, 10.0
D	13.0, 13.0, 13.5, 7.5
E	15.0, 12.0, 8.0, 7.0
F	12.5, 13.5, 14.0, 8.0

Tablica 2.2 Podaci za Primjer 2.10

Želimo ispitati različite prinose obzirom na različita gnojiva. Napomenimo da imamo 6 tretmana, tako da imamo 6 – 1 = 5 stupnjeva slobode za tretmane. Ukupan broj stupnjeva slobode je broj opažanja – 1, dakle 23. To ostavlja 18 stupnjeva slobode za sumu kvadrata unutar tretmana. Ukupnu sumu kvadrata računamo rutinski $\sum(y_{ij} - \bar{y}^2) = \sum y_{ij}^2 - n\bar{y}^2$, što se najučinkovitije računa kao $\sum y_{ij}^2 - \left(\frac{1}{n}\right)\left(\sum y_{ij}\right)^2$. Ovo daje $SS = 3119.25 - \left(\frac{1}{24}\right)(266.5)^2 = 159.990$. Najlakši sljedeći korak je izračunati *SST*, što onda znači da možemo dobiti *SSE* oduzimanjem. Formula za *SST* je relativno jednostavna: $\sum_i T_i/n_i - T^2/n$, gdje T_i označava broj opažanja koja odgovaraju i-tom tretmanu, a $T = \sum_{ij} y_{ij}$. Iz ovog slijedi $SST = \left(\frac{1}{4}\right)(42^2 + 41^2 + 46.5^2 + 47^2 + 42^2 + 48^2) - \frac{1}{24}(266.5)^2 = 11.802$. Potpuna ANOVA tablica prikazana je u Tablici 2.3.

Izvor	*df*	*SS*	S^2	*F*
Gnojiva	5	11.802	2.360	0.287
Rezidual/Greška	18	148.188	8.233	
Ukupno	23	159.990		

Tablica 2.3 Jednosmjerna ANOVA tablica za Primjer 2.11

Na ovaj način smo dobili da je p-vrijednost (0.914) neznačajna u odnosu na $F_{3,16}(0.95) = 3.239$. Zaključujemo da nema razlike između različitih vrsta gnojiva.

3. Višestruka regresija

3.1. Normalne jednadžbe

U prvom poglavlju vidjeli smo osnovni model linearne regresije (1.8). Međutim, moramo uzeti u obzir i slučaj s dva ili više regresora. Općenito, naš cilj je rukovati bilo kojim brojem regresora te ćemo koristiti jezik vektora i matrica.
Za slučajan vektor $\mathbf{X}$, pisat ćemo $\mathbb{E}\mathbf{X}$ za njegovo očekivanje, te var($\mathbf{X}$) za kovarijantu matrica. Koristit ćemo p regresora, u oznaci $x_1, \dots, x_p$, te će svaki od njih imati pripadajući parametar $\beta_1, \dots, \beta_p$. Pretpostavimo sada da je u modelu (1.8) a kratica za $a.1$, gdje je 1 regresor koji odgovara konstantnom članu. Tako za uzorak veličine 1 imamo model

$$y = \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon, \qquad \epsilon_i \sim N(0, \sigma^2).$$

U općem slučaju uzorka veličine n, potrebna su nam dva sufiksa, te tako dobivamo jednadžbu modela

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \qquad (i = 1, \dots, n),$$

gdje su ϵ_i nezavisne jednako distribuirane slučajne varijable.
S desne strane prepoznajemo zapis oblika produkta matrica. Vrijednosti y_i formiramo u vektor stupac $\mathbf{y}$, vrijednosti ϵ_i u vektor stupac ϵ, β_i u vektor stupac β, a x_{ij} u matricu X ($\mathbf{y}$ i ϵ su n x 1, β je p x 1, a X je n x p). Tada naš sustav jednadžbi postaje jedna matrična jednadžba, jednadžba modela

$$\mathbf{y} = X\beta + \epsilon. \tag{3.1}$$

Kao i prije, $\mathbf{y}$ je vektor odgovora, ϵ je vektor pogreške, a β vektor parametra.

Prisjetimo se da je n veličina uzorka (što veći to bolji), dok je p broj parametara (što manji, a da je dovoljan). Očekujemo da je n puno veći od p, to jest $n \gg p$.
Konkretno, matrica X, koju nazivamo *matrica dizajna*, nije invertibilna budući da nije čak niti kvadratna.

U daljnjoj notaciji koristit ćemo oznaku A za matricu X, stoga je jednadžba modela

$$\mathbf{y} = A\beta + \epsilon. \tag{3.2}$$

Iz (3.2) slijedi

$$y_i = \sum_{j=1}^{p} a_{ij}\beta_j + \epsilon_i\,, \qquad \epsilon_i \sim N(0, \sigma^2),$$

vjerodostojnost je

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}n}} \prod_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - \sum_{j=1}^{p} a_{ij}\beta_j\right)^2\right\}$$
$$= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}n}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} a_{ij}\beta_j\right)^2\right\}.$$

Kada ju logaritmiramo, imamo:

$$\ell(\beta, \sigma^2) \coloneqq \log L(\beta, \sigma^2) = \mathrm{c} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} a_{ij}\beta_j\right)^2\right]$$

gdje je c konstanta. Kao i prije, koristimo Fisherovu metodu za maksimalnu vjerodostojnost te maksimiziramo obzirom na β_r: $\partial\ell/\partial\beta_r = 0$ iz čega slijedi

$$\sum_{i=1}^{n} a_{ir}\left(y_i - \sum_{j=1}^{p} a_{ij}\beta_j\right) = 0 \qquad (r = 1, \ldots, p),$$

ili

$$\sum_{j=1}^{p}\left(\sum_{i=1}^{n} a_{ir}a_{ij}\right)\beta_j = \sum_{i=1}^{n} a_{ir}y_i.$$

Označimo sa $C = (c_{ij})$ matricu p x p $C \coloneqq A^T A$, koja je simetrična to jest $C^T = C$. Tada

$$c_{ij} = \sum_{k=1}^{n} (A^T)_{ik}A_{kj} = \sum_{k=1}^{n} a_{ki}a_{kj}.$$

Dakle,

$$\sum_{j=1}^{p} c_{rj}\beta_j = \sum_{i=1}^{n} a_{ir}y_i = \sum_{i=1}^{n} (A^T)_{ri}y_i.$$

U matričnoj notaciji

$$(C\beta)_r = (A^T y)_r \qquad (r = 1, \dots, p),$$

ili kombinirano

$$C\beta = A^T y, \qquad C \coloneqq A^T A. \tag{3.3}$$

Ove jednadžbe nazivamo normalne jednadžbe te su one analogoni normalnih jednadžbi za slučajeve jednog regresora iz prvog poglavlja.

3.2. Rješavanje normalnih jednadžbi

Naš idući zadatak je riješiti normalne jednadžbe. Prije nego to napravimo, prema [6] ćemo ustanoviti da postoji jedinstveno rješenje, to jest da je matrica C iz (3.3) regularna. Podsjetimo se da je *rang matrice* broj neovisnih redaka ili stupaca. Znamo da je matrica A reda n x p, te da je $n \gg p$. Ukoliko je rang matrice jednak p (to jest najveći obzirom na zadanu matricu), tada kažemo da matrica ima puni rang [7]. Također, iz [9], znamo da je kvadratna matrica C *nenegativno definitna* ako vrijedi $x^T Cx \geq 0$ za sve vektore x, te *pozitivno definitna* ako $x^T Cx > 0 \qquad \forall x \neq 0$.
Pozitivno definitna matrica je regularna, invertibilna, dok nenegativno definitna matrica ne treba biti.

Lema 3.1 *Ako matrica A (n* x *p, n > p) ima puni rang p, tada je* $C \coloneqq A^T A$ *pozitivno definitna.*

Dokaz: A ima puni rang, stoga ne postoji vektor x takav da $Ax = 0$ osim nultog vektora. Dakle,

$$(Ax)^T Ax = x^T A^T Ax = x^T Cx = 0$$

samo ako je $x = 0$, u suprotnom strogo pozitivno. Iz ovog slijedi da je C pozitivno definitna što smo i htjeli pokazati.

□

Teorem 3.2 *Za matricu punog ranga A, normalne jednadžbe imaju jedinstveno rješenje*

$$\hat{\beta} = C^{-1}A^T y = (A^T A)^{-1}A^T y. \tag{3.4}$$

Dokaz: U slučaju punog ranga, C je pozitivno definitna po Lemi 3.2, pa je invertibilna, tako da rješavanjem normalnih jednadžbi dobivamo gornje rješenje. □

Kao i u prvom poglavlju, funkcionalni oblik za normalnu vjerodostojnost znači da maksimiziranje vjerodostojnosti minimizira sumu kvadrata

$$SS \coloneqq (y - A\beta)^T(y - A\beta) = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} a_{ij}\beta_j\right)^2.$$

Slično kao i prije, imamo idući teorem.

Teorem 3.3 *Rješenja (3.4) normalnih jednadžbi (3.3) su ujedno i procjenitelji maksimalne vjerodostojnosti te procjenitelji najmanjih kvadrata parametara* β.

Preostaje nam još procijeniti parametar σ.
Kada funkciju *SS* deriviramo po σ, dobit ćemo funkciju log-vjerodostojnosti koja za maksimalan $\hat{\beta}$ daje

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} a_{ij}\beta_j\right)^2 = 0.$$

Sređivanjem izraza, slijedi da je

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} a_{ij}\hat{\beta}_j\right)^2.$$

Ova suma kvadrata je minimalna vrijednost ukupne sume kvadrata *SS* kako parametar β varira, a minimum se postiže procjenom najmanjeg kvadrata za $\beta = \hat{\beta}$.
Ta minimizirana suma kvadrata naziva se *suma kvadrata pogreške, SSE,*

$$SSE = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} a_{ij}\hat{\beta}_j \right)^2 = \left(y - A\hat{\beta}\right)^T \left(y - A\hat{\beta}\right),$$

a nepristrani procjenitelj varijance pogreške je $\hat{\sigma}^2 = SSE/(n-p)$.

$\hat{y} := A\hat{\beta}$ nazivamo *procijenjene vrijednosti*, a $e := y - \hat{y}$, razliku između stvarnih i procijenjenih vrijednosti, vektor *reziduala.*
Ako je $e = (e_1, \dots, e_n)$, gdje su e_i reziduali, a suma kvadrata pogreške

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

je suma kvadrata reziduala.

Numeričko rješenje normalnih jednadžbi ((3.3), (3.4)) pojednostavljeno je ukoliko je matrica A (koja je n x p, te punog ranga p) dana svojom *QR dekompozicijom*

$$A = QR$$

gdje je Q reda n x p i ortonormirana matrica, to jest vrijedi $Q^TQ = I$;
a R reda n x p i gornjetrokutasta regularna matrica.
Normalne jednadžbe $A^TA\hat{\beta} = A^Ty$ postaju

$$R^TQ^TQR\hat{\beta} = R^TQ^Ty,$$

ili

$$R^TR\hat{\beta} = R^TQ^Ty \quad jer\ je \quad Q^TQ = I$$

ili

$$R\hat{\beta} = Q^Ty \quad jer\ su\ R\ i\ R^T\ regularne.$$

Ovaj sustav linearnih jednadžbi sadrži gornje trokutastu matricu R, pa se može riješiti povratnom supstitucijom, počevši od donje jednadžbe te uvrštavanjem u gornje.
Zapišimo A kao niz njenih stupaca $A = (a_1, \dots, a_p)$ gdje su vektori a_i linearno nezavisni jer A ima puni rang. Zapišimo $q_1 := a_1/\|a_1\|$, a za $j = 2, \dots, p,$

$$q_j = w_j / \|w_j\|, \qquad w_j \coloneqq a_j - \sum_{k=1}^{j-1} (a_k^T q_k)\, q_k.$$

Tada su q_j ortonormirani vektori koji se protežu kroz prostor stupaca od A. Svaki q_j je linearna kombinacija od $a_1, \ldots, a_j$, a također je i svaki a_j linearna kombinacija od $q_1, \ldots, q_j$.
Dakle, postoje skalari r_{kj} takvi da

$$a_j = \sum_{k=1}^{j} r_{kj} q_k \qquad (j = 1, \ldots, p).$$

Stavimo $r_{kj} = 0$ za $k > j$. Zatim, sastavljanjem p stupaca a_j u matricu A, ova jednadžba postaje

$$A = QR$$

što smo i željeli.

3.3. Svojstva procjenitelja najmanjih kvadrata

U jednadžbama modela (3.2) pretpostavili smo pogrešku. Ali, možemo gledati i općenitije, te pretpostaviti samo

$$\mathbb{E}y = A\beta, \qquad \operatorname{var}(y) = \sigma^2 I. \tag{3.5}$$

Obzirom da nije zadovoljena pretpostavka o distribuciji, nemamo funkciju vjerodostojnosti stoga ne možemo koristiti metodu maksimalne vjerodostojnosti. Preostaje nam ograničiti se na metodu najmanjih kvadrata.

Linearnost.
Procjenitelj najmanjih kvadrata $\hat{\beta} = C^{-1}A^T y$ linearan je.

Nepristranost.

$$\mathbb{E}\hat{\beta} = C^{-1}A^T \mathbb{E}y = C^{-1}A^T A\beta = C^{-1}C\beta = \beta$$

Dakle, $\hat{\beta}$ je nepristrani procjenitelj od β.
Matrica kovarijance.

$$\begin{aligned}&\mathrm{var}(\hat{\beta}) = \mathrm{var}(C^{-1}A^T y) = C^{-1}A^T\big(\mathrm{var}(y)\big)(C^{-1}A^T)^T\\&= C^{-1}A^T.\sigma^2 I.AC^{-1} \qquad (C = C^T)\\&= \sigma^2.C^{-1}A^T.AC^{-1}\\&= \sigma^2.C^{-1} \qquad (C = A^T A).\end{aligned}$$

Želimo zadržati male varijance naših procjenitelja p parametara β_i (dijagonalni elemente matrice kovarijance iznad), slično i za kovarijance (izvandijagonalni elementi). Što su odstupanja manja, to su naše procjene preciznije i imamo više informacija.

Definicija 3.4 *Matricu $C \coloneqq A^T A$, gdje je A matrica dizajna, nazivamo matrica informacije.*

Nepristrani linearni procjenitelji.
Neka je sada $\hat{\beta} \coloneqq By$ (B je matrica p x n) bilo koji nepristrani linearni procjenitelj od β. Tada

$$\mathbb{E}\tilde{\beta} = B\mathbb{E}y = BA\beta = \beta$$

pa je $\tilde{\beta}$ nepristrani procjenitelj od β ako i samo ako $BA = I$.
Treba imati na umu da

$$\mathrm{var}(\tilde{\beta}) = B\mathrm{var}(y)B^T = B.\sigma^2 I.B^T = \sigma^2 BB^T.$$

Teorem 3.5 *(Gauss-Markovljev teorem) Među svim nepristranim linearnim procjeniteljima $\tilde{\beta} = By$ od β, procjenitelj dobiven metodom najmanjih kvadrata $\hat{\beta} = C^{-1}A^T y$ ima najmanju varijancu u svakoj komponenti, to jest $\hat{\beta}$ je najbolji linearni nepristrani procjenitelj.*

Dokaz: Matrica kovarijance proizvoljne nepristrane linearne procjene $\tilde{\beta} = By$ i procjenitelja najmanjih kvadrata $\hat{\beta}$ dana je sa

$$\mathrm{var}(\tilde{\beta}) = \sigma^2 BB^T \;\; i \;\; \mathrm{var}(\hat{\beta}) = \sigma^2 C^{-1}.$$

Njihova razlika (za koju želimo pokazati da je nenegativna) je

$$\mathrm{var}(\tilde{\beta}) - \mathrm{var}(\hat{\beta}) = \sigma^2[BB^T - C^{-1}].$$

Sada koristeći simetriju od C, C^{-1}, te $BA = I$ (tako da $A^T B^T = I$) slijedi

$$(B - C^{-1}A^T)(B - C^{-1}A^T)^T = (B - C^{-1}A^T)(B^T - AC^{-1}).$$

Nadalje,

$$\begin{aligned}(B - C^{-1}A^T)(B^T - AC^{-1}) &= BB^T - BAC^{-1} - C^{-1}A^T B^T + C^{-1}A^T AC^{-1} \\ &= BB^T - C^{-1} - C^{-1} + C^{-1} \qquad (C = A^T A) \\ &= BB^T - C^{-1}\end{aligned}$$

Kombinirajući,

$$\mathrm{var}(\tilde{\beta}) - \mathrm{var}(\hat{\beta}) = \sigma^2(B - C^{-1}A^T)(B - C^{-1}A^T)^T.$$

Sada za matricu $M = (m_{ij})$ neka vrijedi,

$$\begin{aligned}(MM^T)_{ii} &= \sum\nolimits_k m_{ik}(M^T)_{ki} = \\ &= \sum\nolimits_k m_{ik}^2.\end{aligned}$$

Sada je i-ti dijagonalni element jednak $\mathrm{var}(\tilde{\beta}_i) = \mathrm{var}(\hat{\beta}_i) + \sigma^2$, to jest suma kvadrata i-tog retka matrice $B - C^{-1}A^T$.

Dakle, $\mathrm{var}(\tilde{\beta}_i) \geq \mathrm{var}(\hat{\beta}_i)$, i $\mathrm{var}(\tilde{\beta}_i) = \mathrm{var}(\hat{\beta}_i)$ ako i samo ako $B - C^{-1}A^T$ ima i-ti nul-redak.
Dakle, neki $\tilde{\beta}_i$ ima veću varijancu od $\hat{\beta}_i$ osim ako su svi redovi $B - C^{-1}A^T$ jednaki nula – to jest osim ako je $\tilde{\beta} = By = C^{-1}A^T y = \hat{\beta}$ procjenitelj najmanjih kvadrata što smo i željeli pokazati.

□

Procjenjivost.

Linearnu kombinaciju $c^T\beta = \sum_{i=1}^{p} c_i\beta_i$, $sa\ c = \left(c_1, \ldots, c_p\right)^T$ p-vektorom, nazivamo *procjenjivom* ako ima nepristranog linearnog procjenitelja $b^T y = \sum_{i=1}^{n} b_i y_i$, $sa\ b = (b_1, \ldots, b_n)^T$ n-vektorom. Tada vrijedi

$$\mathbb{E}(b^T y) = b^T \mathbb{E}(y) = b^T A\beta = c^T\beta.$$

Ova jednakost može vrijediti i za neki nepoznati β ako i samo ako

$$c^T = b^T A$$

to jest ako je c linearna kombinacija n-redaka matrice dizajna A. U slučaju punog ranga matrice koji ovdje promatramo, redovi matrice A obuhvaćaju prostor pune dimenzije p, stoga su sve linearne kombinacije procjenjive. U slučaju da rang nije pun, to jest $k < p$, procjenjive funkcije obuhvaćaju prostor dimenzije k, a neprocjenjive linearne kombinacije postoje.

3.4. Rastav varijance

Definiramo *sumu kvadrata regresije, SSR,* kao

$$SSR := \left(\hat{\beta} - \beta\right)^T C\left(\hat{\beta} - \beta\right).$$

Budući da se ovdje radi o kvadratnoj formi s matricom C koja je pozitivno definitivna, imamo da je $SSR \geq 0$ i $SSR > 0$ osim ako je $\hat{\beta} = \beta$, to jest, osim ako je procjenitelj najmanjih kvadrata jednak stvarnoj vrijednosti, što se u praksi nikada neće dogoditi.

Teorem 3.6 *(Rastav varijance)*

$$SS = SSR + SSE \tag{3.6}$$

Dokaz: Zapišimo

$$y - A\beta = \left(y - A\hat{\beta}\right) + A\left(\hat{\beta} - \beta\right).$$

Vektor s obje strane pomnožimo s njegovim transponiranim vektorom.
S lijeve strane dobivamo

$$SS = (y - A\beta)^T (y - A\beta),$$

ukupnu sumu kvadrata. S desne, dobivamo tri člana.
Prvi je suma kvadrata pogreške

$$SSE = \left(y - A\hat{\beta}\right)^T \left(y - A\hat{\beta}\right),$$

drugi je suma kvadrata regresije

$$\left(A\left(\hat{\beta} - \beta\right)\right)^T A\left(\hat{\beta} - \beta\right) = \left(\hat{\beta} - \beta\right)^T A^T A\left(\hat{\beta} - \beta\right) = \left(\hat{\beta} - \beta\right)^T C\left(\hat{\beta} - \beta\right) = SSR.$$

Treći , to jest mješoviti član je $\left(y - A\hat{\beta}\right)^T A\left(\hat{\beta} - \beta\right)$ i njegov transponirani vektor, koji je jednak kao i on jer su oboje skalari.

Međutim, zbog (3.3)

$$A^T\left(y - A\hat{\beta}\right) = A^T y - A^T A\hat{\beta} = A^T y - C\hat{b} = 0.$$

Transponiranjem,

$$\left(y - A\hat{\beta}\right)^T A = 0.$$

Dakle, mješoviti član nestaje, što nam rezultira upravo sa (3.6), što smo i željeli pokazati.

□

Korolar 3.7 *Imamo da je*

$$SSE = \min_{\beta} SS,$$

to jest, minimum se postiže za procjenitelja najmanjih kvadrata $\hat{\beta} = C^{-1}A^T y$.

Dokaz: $SSR \geq 0, a\ SSR = 0$ ako i samo ako $\beta = \hat{\beta}$.

□

Sada ćemo vidjeti kako izvođenje regresije s p regresora predstavlja ortogonalnu projekciju na odgovarajući p-dimenzionalni potprostor n-dimenzionalnog prostora.

Definicija 3.8 *Linearnu transformaciju* $P: V \to V$ *nazivamo projekcija na* V_1 *duž* V_2 *ako je* V *direktan zbroj* $V = V_1 \oplus V_2$ *i ako je* $x = (x_1, x_2)^T$ *sa* $Px = x_1$.
Tada je, prema [9] i [10] $V_1 = \mathrm{Im}P = \mathrm{Ker}(I - P)$, $V_2 = \mathrm{Ker}P = \mathrm{Im}(I - P)$.
Podsjetimo se da je kvadratna matrica idempotentna ako vrijedi $M^2 = M$. *M je idempotentna ako i samo ako je projekcija.*

Nadalje, za matrice A i $C = A^TA$, pišemo $P \coloneqq AC^{-1}A^T$.
Primijetimo da je P simetrična, te da po (3.3) vrijedi $Py = AC^{-1}A^Ty = A\hat{\beta}$.

Lema 3.9 *P i I – P su idempotentne, stoga su projekcije.*

Dokaz:

$$P^2 = AC^{-1}A^T.AC^{-1}A^T = AC^{-1}A^T = P:$$
$$P^2 = P$$
$$(I - P)^2 = I - 2P + P^2 = I - 2P + P = I - P.$$

□

Teorem 3.10

$$SSE = y^T(I - P)y = (y - A\beta)^T(I - P)(y - A\beta),$$
$$SSR = (y - A\beta)^TP(y - A\beta).$$

Dokaz: Pošto je $SSE \coloneqq \left(y - A\hat{\beta}\right)^T\left(y - A\hat{\beta}\right)$, a $A\hat{\beta} = Py$, slijedi da je

$$SSE = \left(y - A\hat{\beta}\right)^T\left(y - A\hat{\beta}\right)$$
$$= (y - Py)^T(y - Py) = y^T(I - P)(I - P)y = y^T(I - P)y,$$

pošto je $I – P$ projekcija. Za SSR imamo da je

$$SSR \coloneqq \left(\hat{\beta} - \beta\right)^T C\left(\hat{\beta} - \beta\right) = \left(\hat{\beta} - \beta\right)^T A^TA\left(\hat{\beta} - \beta\right).$$

Znamo da

$$\left(\hat{\beta} - \beta\right) = C^{-1}A^Ty - \beta = C^{-1}A^Ty - C^{-1}A^TA\beta = C^{-1}A^T(y - A\beta)$$

pa je

$$\begin{aligned} SSR &= (y - A\beta)^T\, AC^{-1}.A^TA.C^{-1}A^T(y - A\beta) \\ &= (y - A\beta)^T\, AC^{-1}A^T(y - A\beta) \qquad (A^TA = C) \\ &= (y - A\beta)^T\, P(y - A\beta), \end{aligned}$$

što smo i željeli pokazati.
Sada formula za *SSE* sada slijedi iz ove formule za *SSR* i (3.6) oduzimanjem.

□

Koeficijent determinacije.
Koeficijent determinacije definiran je kao R^2, gdje je R koeficijent korelacije podataka i procijenjenih vrijednosti to jest

$$R \coloneqq \sum (y_i - \overline{y})\left(\hat{y}_i - \overline{\hat{y}}\right) / \sqrt{\sum (y_i - \overline{y})^2 \sum \left(\hat{y}_i - \overline{\hat{y}}\right)^2}.$$

Dakle, $-1 \le R \le 1, 0 \le R^2 \le 1,$ a R^2 je mjera za *dobro poklapanje* procijenjenih vrijednosti sa podacima.

Teorem 3.11 *Vrijedi*

$$R^2 = 1 - \frac{SSE}{\sum (y_i - \overline{y})^2}$$

Dokaz: Vidi [5, Teorem 3.20].

Bitno je uočiti da je $R^2 = 1$ ako i samo ako je *SSE = 0*, to jest, ako su svi reziduali jednaki 0, a procijenjene vrijednosti su točne vrijednosti podataka. Što je R^2 veći, odnosno *SSE* manji, to je procjena regresijskog modela podacima bolja.

Propozicija 3.12 *(Formula za trag)*

$$\mathbb{E}(x^TAx) = \mathrm{tr}\big(A.\,\mathrm{var}(x)\big) + \mathbb{E}x^T.A.\,\mathbb{E}x \tag{3.7}$$

Dokaz:

$$x^T Ax = \sum_{ij} a_{ij} x_i x_j,$$

pa zbog linearnosti očekivanja,

$$\mathbb{E}[x^T Ax] = \sum_{ij} a_{ij} \mathbb{E}\big[x_i x_j\big].$$

Sada $\mathrm{cov}\big(x_i, x_j\big) = \mathbb{E}\big(x_i x_j\big) - (\mathbb{E}x_i)\big(\mathbb{E}x_j\big)$ pa je

$$\begin{aligned}\mathbb{E}[x^T Ax] &= \sum_{ij} a_{ij}\big[\mathrm{cov}\big(x_i x_j\big) + \mathbb{E}x_i . \mathbb{E}x_j\big] \\ &= \sum_{ij} a_{ij}\mathrm{cov}\big(x_i x_j\big) + \sum_{ij} a_{ij} . \mathbb{E}x_i . \mathbb{E}x_j.\end{aligned}$$

Drugi član s desne strane je $\mathbb{E}x^T . A . \mathbb{E}x$. Prvo uočimo da

$$\mathrm{tr}(AB) = \sum_i (AB)_{ii} = \sum_{ij} a_{ij} b_{ji} = \sum_{ij} a_{ij} b_{ij},$$

ako je B simetrična. Ali matrice kovarijance su simetrične, tako da je prvi član s desne strane $\mathrm{tr}\big(A . \mathrm{var}(x)\big)$, što smo i željeli pokazati.

□

Korolar 3.13 *Vrijedi*

$$\mathrm{tr}(P) = p, \qquad \mathrm{tr}(I - P) = n - p, \qquad \mathbb{E}(SSE) = (n - p)\sigma^2.$$

Dakle, $\hat{\sigma}^2 := SSE/(n - p)$ *je nepristrani procjenitelj od* σ^2*.*

Dokaz: Iz Teorema 3.10, *SSE* je kvadratna forma od $y - A\beta$ s matricom $I - P = I - AC^{-1}A^T$. Sada je

$$\mathrm{tr}(I - P) = \mathrm{tr}(I - AC^{-1}A^T) = \mathrm{tr}(I) - \mathrm{tr}(AC^{-1}A^T).$$

Ali $\mathrm{tr}(I) = n$ (jer je I u ovom slučaju jedinična matrica reda n), i $\mathrm{tr}(AB) = \mathrm{tr}(BA)$,

$$\mathrm{tr}(P) = \mathrm{tr}(AC^{-1}A^T) = \mathrm{tr}(C^{-1}A^T A) = \mathrm{tr}(I) = p,$$

jer je ovdje I jedinična matrica reda p. Dakle,

$$\mathrm{tr}(I - P) = \mathrm{tr}(I - AC^{-1}A^T) = n - p.$$

Pošto je $\mathbb{E}y = A\beta$, a $\mathrm{var}(y) = \sigma^2 I$, iz formule za trag (3.7) slijedi

$$\mathbb{E}(SSE) = (n - p)\sigma^2.$$

□

Ova zadnja formula analogna je odgovarajućoj ANOVA formuli $\mathbb{E}(SSE) = (n - r)\sigma^2$.

3.5. χ^2 dekompozicija

Prisjetimo se iz Teorema o ortogonalnosti da ako je $x = x_1, \dots, x_n \sim N(0, I)$ (ako su x_i nezavisne jednako distribuirane slučajne varijable iz *N(0, 1)*, te ako napravimo supstituciju ortogonalnom transformacijom B u $y := Bx$, tada je i $y \sim N(0, I)$.
Sada iz [9] znamo da je λ svojstvena vrijednost matrice A sa svojstvenim vektorom x (koji nije nulvektor) ako vrijedi

$$Ax = \lambda x.$$

Također, prema [8] znamo da ako je A realna simetrična matrica, onda je možemo dijagonalizirati ortogonalnom transformacijom B, te dobiti matricu $D = B^T AB$.
Nadalje, ako je λ svojstvena vrijednost od A, tada

$$|D - \lambda I| = |B^T AB - \lambda I| = |B^T AB - \lambda B^T B| = |B^T||A - \lambda I||B| = 0.$$

Tada je kvadratna forma u normalnim varijablama s matricom A također kvadratna forma u normalnim varijablama s matricom D kao

$$x^T Ax = x^T BDB^T x = y^T Dy, \qquad y \coloneqq B^T x.$$

3.5.1 Idempotentnost, trag i rang

Podsjetimo se da je kvadratna matrica idempotentna ako vrijedi $M^2 = M$.

Propozicija 3.14 *Ako je B idempotentna, tada vrijedi*

i) *njene svojstvene vrijednosti su 0 ili 1,*
ii) *njen trag je jednak njenom rangu.*

Dokaz:

i) Ako je λ svojstvena vrijednost od B, sa svojstvenim vektorom x, tada $Bx = \lambda x,\ x \neq 0$. Nadalje,

$$B^2x = B(Bx) = B(\lambda x) = \lambda(Bx) = \lambda(\lambda x) = \lambda^2 x,$$

pa je λ^2 svojstvena vrijednost od B^2 (vrijedi uvijek, bez obzira na idempotentnost). Dakle,

$$\lambda x = Bx = B^2x = \ldots = \lambda^2 x$$

a obzirom da je $x \neq 0$, $\lambda = \lambda^2, \lambda(\lambda - 1) = 0$, slijedi da je $\lambda = 0\ ili\ 1$.

ii) Znamo da je trag od B suma svojstvenih vrijednosti, što je zbog *i)* jednako broju svojstvenih vrijednosti različitih od nule, što je upravo rang od B.

□

Korolar 3.15 *Vrijedi*

$$r(P) = p, \qquad r(I - P) = n - p.$$

Dokaz: Tvrdnja slijedi direktno iz Korolara 3.13 i Propozicije 3.14.

□

3.5.2 Kvadratne forme normalnih slučajnih varijabli

Od interesa su nam simetrične projekcijske (dakle, idempotentne) matrice *P.* Budući da su njihove svojstvene vrijednosti 0 i 1, možemo ih dijagonalizirati ortogonalnim transformacijama u dijagonalnu matricu nula i jedinica. Stoga ako matrica P ima rang r, kvadratna se forma $x^T Px$ može reducirati na zbroj od r kvadrata standardnih normalnih slučajnih varijabli. Možemo uzeti da jedinice prethode nulama na dijagonali pa imamo

$$x^T Px = y_1^2 + \ldots + y_r^2, \qquad y_i \sim N(0, \sigma^2).$$

gdje su y_i nezavisne, jednako distribuirane.

Dakle, $x^T Px$ je σ^2 puta $\chi^2(r)$ distribuirana slučajna varijabla.

Teorem 3.16 *Ako je P simetrična projekcija ranga r,* $x_i \sim N(0, \sigma^2)$ *nezavisni, tada je kvadratna forma*

$$x^T Px \sim \sigma^2 x^2(r).$$

3.5.3 Suma projekcija

Rastav varijance, koja izražava sumu kvadrata (χ^2 distribuiranu) kao sumu nezavisnih suma kvadrata (također χ^2 distribuirane) koja odgovara dekompoziciji identiteta I kao sumu ortogonalnih projekcija. Gauss-Markovljev teorem nam odgovara za slučaj $I = P + (I - P)$, međutim susreli smo se i sa dekompozicijama koje imaju više od dva sumanda (na primjer $SS = SSB + SST + SSI$ koji ima tri). Promotrimo sada općeniti slučaj.
Pretpostavimo da su $P_1, \ldots, P_k$ simetrične matrice projekcije takve da:

$$I = P_1 + \ldots + P_k.$$

Promotrimo trag obje strane jednakosti. Matrica I reda n ima trag n. Svaki P_i ima za trag svoj rang n_i, pa po Propoziciji 3.14 $n = n_1 + \ldots + n_k$.

Kada kvadriramo, dobivamo

$$I = I^2 = \sum_i P_i^2 + \sum_{i<j} P_i P_j = \sum_i P_i + \sum_{i<j} P_i P_j.$$

Sada gledamo trag,

$$n = \sum n_i + \sum_{i<j} \mathrm{tr}\left(P_i P_j\right) = n + \sum_{i<j} \mathrm{tr}\left(P_i P_j\right):$$
$$\sum_{i<j} \mathrm{tr}\left(P_i P_j\right) = 0.$$

Dakle,

$$\begin{aligned} \mathrm{tr}\left(P_i P_j\right) &= \mathrm{tr}\left(P_i^2 P_j^2\right) && (jer\ su\ P_i P_j\ projekcije) \\ &= \mathrm{tr}\left(\left(P_j P_i\right).\left(P_i P_j\right)\right) && (tr(AB) = tr(BA)) \\ &= \mathrm{tr}\left(\left(P_i P_j\right)^T.\left(P_i P_j\right)\right), \end{aligned}$$

Stoga imamo da je $\mathrm{tr}\left(P_i P_j\right) \geq 0$, jer za matricu M

$$\begin{aligned} \mathrm{tr}(M^T M) &= \sum_i (M^T M)_{ii} \\ &= \sum_i \sum_j (M^T)_{ij} (M)_{ij} \\ &= \sum_i \sum_j m_{ij}^2 \\ &\geq 0 \end{aligned}$$

Dakle, imamo sumu nenegativnih članova koja mora biti nula, pa svaki član mora biti nula. To jest, kvadrat svakog elementa od $P_i P_j$ mora biti nula, pa je svaki element $P_i P_j$ nula, pa je matrica $P_i P_j = 0,\ i \neq j$.

Ovo je uvjet da linearni oblici $P_1 x, \ldots, P_k x$ budu nezavisni, pa su i $(P_i x)^T (P_i x) = x^T P_i^T P_i x$ nezavisni, odnosno $x^T P_i x$ jer je P_i simetričan i idempotentan. Kvadratne forme $x^T P_1 x, \ldots, x^T P_k x$ su također nezavisne.
Sada imamo

$$x^T x = x^T P_1 x + \ldots + x^T P_k x.$$

Sumirajmo zaključak:

Teorem 3.17 *(χ^2 dekompozicija) Ako $I = P_1 + \ldots + P_k$, gdje su P_i simetrične matrice projekcije s rangom n_i, tada vrijedi:*

i) *suma ranga je* $n = n_1 + \ldots + n_k$,
ii) *svaka kvadratna forma* $Q_i := x^T P_i x$ *je* $Q_i \sim \sigma^2 x^2(n_i)$,
iii) Q_i *su međusobno nezavisni,*
iv) $P_i P_j = 0 \quad (i \neq j)$ *(svojstvo ortogonalnosti projekcija).*

3.6. Ortogonalna projekcija i Pitagorin poučak

Procjenitelji najmanjih kvadrana procijenjene su vrijednosti $\hat{y} = A\hat{\beta} = A(A^T A)^{-1} A^T y = Py$, gdje je P matrica projekcija (idempotentna, simetrična). Zatim, $e := y - \hat{y} = y - Py = (I - P)y$ je vektor reziduala. Dakle, dobivamo

$$y = A\beta + \epsilon = A\hat{\beta} + e = \hat{y} + e$$

to jest, podaci = točne vrijednosti + pogreška = procijenjene vrijednosti + rezidual.
Sada,

$$\begin{aligned} e^T \hat{y} &= y^T (I - P)^T P y \\ &= y^T (I - P) P y \qquad (P \text{ simetrična}) \\ &= y^T (P - P^2) y \\ &= 0, \end{aligned}$$

jer je P idempotentan. Ovo nam govori da su e i $\hat{y}$ ortogonalni. Također su Gaussovi (multinormalni) kao linearna kombinacija Gaussovih. Za Gaussove vekore, ortogonalnost = nekoreliranost = nezavisnost.
Reziduali e i procijenjene vrijednosti $\hat{y}$ su ortogonalni i nezavisni. Stoga je vektor podataka y hipotenuza pravokutnog trokuta u n-dimenzionalnom prostoru gdje su druge dvije stranice trokuta upravo procijenjene vrijednosti $\hat{y} = (I - P)y$ i rezidual $e = Py$.

Dakle, duljine vektora su povezane Pitagorinim poučkom u n-prostoru:

$$\|y\|^2 = \|\hat{y}\|^2 + \|e\|^2.$$

Konkretno, $\|\hat{y}\|^2 \leq \|y\|^2$:

$$\left\|\hat{P}y\right\|^2 \leq \|y\|^2 \quad za\ sve\ y.$$

Dakle, $\|P\| \leq 1$, to jest P ima normu manju od 1, smanjuje duljinu. Slično za $I - P$, također je projekcija te smanjuje duljinu jer joj je norma manja od 1.
Za realne vektorske prostore, projekcija P je simetrična ako i samo ako smanjuje duljinu ako i samo ako je ortogonalna ili okomita projekcija, tada su slika i jezgra od P ortogonalni ili okomiti potprostori. Pošto je naša $P \coloneqq AC^{-1}A^T,\ C \coloneqq A^TA$ automatski simetrična i idempotentna projekcija, ovo je situacija relevantna za nas.

Teorem 3.18 *Vrijedi*

i) $\hat{y} = Py \backsim N(A\beta, \sigma^2 P)$
ii) $e \coloneqq y - \hat{y} = (I - P)y \sim N\big(0, \sigma^2(I - P)\big)$
iii) *e i* $\hat{y}$ *su nezavisni.*

Dokaz:

i) $\hat{y}$ je linearna transformacija Gaussovog vektora y, pa je također Gaussov. Znamo da je procjenitelj $\hat{b}$ nepristran za β, pa $\hat{y} \coloneqq A\hat{b}$ nepristran za $A\beta$.

$$\begin{aligned} \mathrm{var}(\hat{y}) &= P\mathrm{var}(y)P^T \\ &= \sigma^2 PP^T \qquad (var(y) = \sigma^2 I) \\ &= \sigma^2 P^2 \qquad (P\ simetrična) \\ &= \sigma^2 P \qquad (P\ idempotentna). \end{aligned}$$

ii) Slično, e je Gaussov, $Ee = Ey - E\hat{y} = A\beta - A\beta = 0$.

$$\begin{aligned} \mathrm{var}(e) &= (I - P)\mathrm{var}(y)(I - P)^T \\ &= \sigma^2(I - P)(I - P)^T \qquad (var(y) = \sigma^2 I) \\ &= \sigma^2(I - P)^2 \qquad (I - P\ simetrična) \\ &= \sigma^2(I - P) \qquad (I - P\ idempotentna). \end{aligned}$$

iii) *Vrijedi*

$$\begin{aligned}\operatorname{cov}(\hat{y}, e) &= \mathbb{E}[(\hat{y} - \mathbb{E}\hat{y})^T (e - \mathbb{E}e)] \\ &= \mathbb{E}[(\hat{y} - A\beta)^T e] \qquad (\mathbb{E}\hat{y} = A\beta, \mathbb{E}e = 0) \\ &= \mathbb{E}[(Py - A\beta)^T (I - Py)] \\ &= \mathbb{E}[(y^T P - \beta^T A^T)(y - Py)] \\ &= \mathbb{E}[y^T Py] - \mathbb{E}[y^T P^2 y] - \beta^T A^T \mathbb{E}y + \beta^T A^T A(A^T A)^{-1} A^T \mathbb{E}y \\ &= 0\end{aligned}$$

Dakle, e *i* $\hat{y}$ su nekorelirani, pa su nezavisni jer su Gaussovi.

□

Teorem 3.19 *Vrijedi*

i) $\hat{\beta} \sim N(\beta, \sigma^2 C^{-1})$
ii) $\hat{\beta}$ *i SSE (ili* $\hat{\beta}$ *i* σ^2*) su nezavisni.*

Dokaz: Vidi [5, Teorem 3.31].

Korolar 3.20 *SSR i SSE su nezavisni.*

Dokaz: $SSR := \left(\hat{\beta} - \beta\right)^T C \left(\hat{\beta} - \beta\right)$ je funkcija od $\hat{\beta}$, pa radi Teorema 2.19 *ii)* slijedi tvrdnja.

□

Konačno, Teorem 3.19 u kombinaciji sa Teoremom 3.16 daje metodu za računanje jednodimenzionalnih intervala pouzdanosti za pojedinačne elemente od β. Imamo:

Korolar 3.21 *Neka je* β_i *i-ti element od* β*, i neka je* ${C_{ii}}^{-1}$ *i-ti dijagonalni element od* C^{-1}*. Vrijedi*

$$\frac{\beta_i \hat{\beta}_i}{\hat{\sigma}\sqrt{C_{ii}^{-1}}} \sim t_{n-p}.$$

Dokaz: Vidi [5, Korolar 3.33].

3.7. Primjer

U ovom odjeljku promotrit ćemo primjer. Prva stvar koju je potrebno uočiti je identificirati matricu dizajna A, a zatim pronaći različite matrice, posebno matricu projekcije P, koje su povezane s njom. Ovaj primjer je dovoljno mali pa ćemo ga moći odraditi ručno, a dovoljno je velik da je netrivijalan i da predoči postupak.

Primjer 3.22 *Dva proizvoda, A i B, važu se na vagi prvo odvojeno, a zatim zajedno, kako bi se dobila opažanja* y_1, y_2, y_3.

1. *Pronađimo procjenitelje najmanjih kvadrata od pravih težina* β_A, β_B.
   Imamo

$$y_1 = \beta_A + \epsilon_1,$$
$$y_2 = \beta_B + \epsilon_2,$$
$$y_1 + y_2 = \beta_A + \beta_B + \epsilon_3,$$

   gdje su ϵ_1 nezavisne jednako distribuirane slučajne varijable iz normalne razdiobe s očekivanjem 0 i varijancom σ^2. Dakle,

$$\mathbb{E}y = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \beta_A \\ \beta_B \end{pmatrix}.$$

   Matrica dizajna je:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

   Stoga,

$$C = A^T A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

   Vidimo da je determinanta matrice C jednaka 3, te da je

$$C^{-1} = \frac{1}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix},$$

$$A^T y = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} . \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} y_1 + y_3 \\ y_2 + y_3 \end{pmatrix},$$

$$\hat{\beta} = C^{-1} A^T y = \frac{1}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} y_1 + y_3 \\ y_2 + y_3 \end{pmatrix},$$

$$= \frac{1}{3} \begin{pmatrix} 2y_1 - y_2 + y_3 \\ -y_1 + 2y_2 + y_3 \end{pmatrix}.$$

Prva i druga komponenta ovog vektora su procjenitelji najmanjih kvadrata za $\beta_A \, i \, \beta_B$.

2. *Pronađimo kovarijancu matrice od procjenitelja najmanjih kvadrata.*
Ovo je

$$\text{var}(\hat{\beta}) = \sigma^2 C^{-1} = \frac{\sigma^2}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$$

3. *Pronađimo SSE i procijenjeni* σ^2.

$$P = A . C^{-1} A^T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} . \frac{1}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix},$$

$$I - P = \frac{1}{3} \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix}.$$

Dakle,

$$SSE = y^T (I - P) y = \frac{1}{3} (y_1 \quad y_2 \quad y_3) \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix} . \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

$$= \frac{1}{3} (y_1 + y_2 - y_3)^2.$$

Pošto je $n = 3$, $p = 2$, $n - p = 1$, onda je $\hat{\sigma}^2$ također

$$\hat{\sigma}^2 = \frac{1}{3} (y_1 + y_2 - y_3)^2.$$

4. Primjena u marketingu

Optimalna raspodjela marketinškog budžeta teško je pitanje s kojim se svaka tvrtka suočava. Pojavom novih marketinških tehnika, poput online oglašavanja i oglašavanja putem društvenih mreža, složenost podataka je porasla, što ovaj problem čini još većim izazovom. Statistički alati za modeliranje često se koriste za rješavanje problema raspodjele budžeta. Modeliranje takozvanog *Marketing Mixa* uključuje korištenje niza statističkih metoda koje su prikladne za modeliranje varijable od interesa (u ovom poglavlju, to je prodaja), s ciljem konstruiranja optimalne kombinacije strategija koje bi maksimizirale profit.

4.1. Linearni i multiplikativni modeli

Promotrimo prvo kako najbolje izabrati formu modela. Za početak, iz [13, str. 3] imamo linearni model koji pretpostavlja konstantne povrate te je oblika

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + \epsilon_t, \tag{6.1}$$

gdje prateći notaciju iz [11] imamo da su t vrijednosti varijabli u periodu $t = 1, \ldots, T$ gdje je T broj zapažanja, β_j parametri modela, x_{kt} vrijednost varijable k u periodu t gdje je $k = 1, \ldots, K$, te ϵ_t vrijednost odstupanja to jest pogreška.
Najčešće se prvo isprobava linearni model jer je procjena koeficijenata jednostavna. On pokazuje dobru aproksimaciju nelinearne funkcije, ali samo u ograničenom rasponu. No nedostatak pretpostavke o linearnosti je taj što implicira konstantne povrate što je nerealno u stvarnom svijetu marketinga. Najčešće krivulja odaziva prodaje nema konstantan oblik. Ponovo iz [13, str. 3] vidimo da se jedan takav model naziva *multiplikativni model snage:*

$$y_t = \beta_0 x_{1t}^{\beta_1} \epsilon_t, \qquad x_{1t} \geq 0, \qquad 0 < \beta_1 < 1. \tag{6.2}$$

Taj model može se linearizirati logaritmiranjem obje strane:

$$\ln(y_t) = \ln(\beta_0) + \beta_1 \ln(x_{1t}) + \ln\epsilon_t, \qquad x_{1t} \geq 0, \qquad 0 < \beta_1 < 1. \tag{6.3}$$

Ta jednakost je linearna u parametrima $\beta_0{}^*, \beta_1$ gdje je $\beta_0{}^* = \ln\beta_0$.
Ovaj model još nazivamo i *log-log* model. Slično, iz [13, str. 4], multiplikativna verzija za K marketinških instrumenata je oblika

$$y_t = \beta_0 x_{1t}^{\beta_1} x_{2t}^{\beta_2} \dots x_{Kt}^{\beta_K} \epsilon_t. \tag{6.4}$$

Kada je funkcija odaziva prodaje izložena povećanju povrata, možemo koristiti eksponencijalni model:

$$y_t = \beta_0 e^{\beta_1 x_{1t}} \epsilon_t, \tag{6.5}$$

Logaritmiranjem slijedi

$$\ln(y_t) = \ln(\beta_0) + \beta_1 x_{1t} + \ln(\epsilon_t). \tag{6.6}$$

Ovaj model nazivamo još i *log-linearan* model.
Prema [12], kada imamo log-log ili log-linearan model, potrebna je prilagodba y_t kako bi ostali nepristrani. Ako uzmemo u obzir multiplikativni oblik (6.4) gdje je $\ln\epsilon^t \sim N(0, \sigma^2)$, može se pokazati da:

$$\mathbb{E}[y_t] = \beta_0 x_{1t}^{\beta_1} x_{2t}^{\beta_2} \dots x_{Kt}^{\beta_K} e^{1/2\sigma^2}$$

Procjenitelje računamo na idući način

$$\hat{y}_t = \hat{\beta}_0 x_{1t}^{\hat{\beta}_1} x_{2t}^{\hat{\beta}_2} \dots x_{Kt}^{\hat{\beta}_K} e^{1/2\sigma^2}$$

gdje su $\hat{\beta}_i$ procjene najmanjih kvadrata.

4.2. Marketing Mix

Podaci dolaze od tvrtke Nepa koja je jedan od najvećih trgovaca elektroničkom opremom u Švedskoj. Skup podataka sadrži podatke o tjednoj prodaji i marketinškim aktivnostima specifične za model, kao i podatke o okruženju, za dvije godine.
Za procjenu parametara za jednadžbe, koristimo iduće podatke:

$y_t = prodaja\ u\ tjednu\ t$
$\mathrm{TV}_t = troškovi\ oglašavanja\ za\ televiziju\ u\ tjednu\ t$
$\mathrm{DR}_t = troškovi\ oglašavanja\ za\ oglase\ koji\ stižu\ poštom\ u\ tjednu\ t$
$\mathrm{DR.POSTEN}_t = ulaganja\ u\ oglase\ koji\ dolaze\ poštom$
$\mathrm{OUTDOOR}_t = ulaganja\ u\ oglase\ postavljene\ vani, npr.\ na\ stanici\ u\ tjednu\ t$
$\mathrm{RADIO}_t = radio\ troškovi\ oglašavanja\ u\ tjednu\ t$
$\mathrm{PRINT}_t = print\ troškovi\ oglašavanja\ u\ tjednu\ t$
$\mathrm{SOCIALMEDIA}_t = troškovi\ oglašavanja\ na\ društveni\ mrežama\ u\ tjednu\ t$
$\mathrm{Rain}_t = količina\ kiše\ u\ tjednu\ t$
$\mathrm{sal}_t = varijabla\ koja\ pokazuje\ je\ li\ u\ pitanju\ tjedan\ primanja\ plaće$
$\mathrm{HOLIDAY}_t = varijabla\ koja\ pokazuje\ dal\ je\ u\ tjednu\ t\ bio\ praznik$

Sa tako definiranim varijablama, linearni model iz (6.1) postaje:

$$y_t = \beta_0 + \beta_1 \mathrm{TV}_t + \beta_2 \mathrm{DR}_t + \beta_3 \mathrm{DR.POSTEN}_t + \beta_4 \mathrm{OUTDOOR}_t + \\ + \beta_5 \mathrm{RADIO}_t + \beta_6 \mathrm{PRINT}_t + \beta_7 \mathrm{SOCIALMEDIA}_t + \beta_8 \mathrm{Rain}_t + \\ + \beta_9 \mathrm{sal}_t + \beta_{10} \mathrm{HOLIDAY}_t + {\epsilon_t}^{(1)}.$$

Eksponencijalni model iz (6.5) sada je oblika

$$y_t = \beta_0 e^{\beta_0 + \beta_1 \mathrm{TV}_t + \beta_2 \mathrm{DR}_t + \beta_3 \mathrm{DR.POSTEN}_t + \beta_4 \mathrm{OUTDOOR}_t + \beta_5 \mathrm{RADIO}_t} \\ e^{\beta_6 \mathrm{PRINT}_t + \beta_7 \mathrm{SOCIALMEDIA}_t + \beta_8 \mathrm{Rain}_t + \beta_9 \mathrm{sal}_t + \beta_{10} \mathrm{HOLIDAY}_t} {\epsilon_t}^{(2)}$$

koji logaritmiranjem obje strane postaje

$$\ln(y_t) = \ln(\beta_0) + \beta_1 \mathrm{TV}_t + \beta_2 \mathrm{DR}_t + \beta_3 \mathrm{DR.POSTEN}_t + \beta_4 \mathrm{OUTDOOR}_t + \\ + \beta_5 \mathrm{RADIO}_t + \beta_6 \mathrm{PRINT}_t + \beta_7 \mathrm{SOCIALMEDIA}_t + \beta_8 \mathrm{Rain}_t + \\ + \beta_9 \mathrm{sal}_t + \beta_{10} \mathrm{HOLIDAY}_t + \ln({\epsilon_t}^{(2)}).$$

Multiplikativni model (6.4) za ovaj slučaj je

$$y_t = \gamma_0 \mathrm{TV}_t^{\gamma_1} \mathrm{DR}_t^{\gamma_2} \mathrm{DR.POSTEN}_t^{\gamma_3} \mathrm{OUTDOOR}_t^{\gamma_4} \mathrm{RADIO}_t^{\gamma_5} \\ \mathrm{PRINT}_t^{\gamma_6} \mathrm{SOCIALMEDIA}_t^{\gamma_7} {\gamma_8}^{\mathrm{Rain}_t} {\gamma_9}^{\mathrm{sal}_t} {\gamma_{10}}^{\mathrm{HOLIDAY}_t} {\epsilon_t}^{(3)}.$$

te nakon transformacije

$$\ln(y_t) = \ln(\gamma_0) + \gamma_1 \ln(\mathrm{TV}_t) + \gamma_2 \ln(\mathrm{DR}_t) + \gamma_3 \ln(\mathrm{DR.POSTEN}_t) + \\ + \gamma_4 \ln(\mathrm{OUTDOOR}_t) + \gamma_5 \ln(\mathrm{RADIO}_t) + \gamma_6 \ln(\mathrm{PRINT}_t) + \gamma_7 \ln(\mathrm{SOCIALMEDIA}_t) + \\ + {\gamma_8}^* \mathrm{Rain}_t + {\gamma_9}^* \mathrm{sal}_t + {\gamma_{10}}^* \mathrm{HOLIDAY}_t + \ln\left({\epsilon_t}^{(3)}\right),$$

$$\gamma_8^* = \ln(\gamma_8), \qquad \gamma_9^* = \ln(\gamma_9), \qquad \gamma_{10}^* = \ln(\gamma_{10}).$$

Sljedeće tablice ilustriraju rezultate procjene metodom najmanjih kvadrata za linearan, log-linearan, te log-log model redom.

```
##
## Call:
## lm(formula = "SALES_TOT ~ TV+DR+DR.POSTEN+OUTDOOR+RADIO+PRINT+SOCIALMEDIA+Rain..
    mm.+sal+HOLIDAY",
##     data = regdata)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -23290744   -7598418   -1641095    7793738   82440104
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.226e+07  3.988e+06  10.598  < 2e-16 ***
## TV           2.239e+01  2.567e+00   8.721 9.59e-14 ***
## DR           1.453e+01  8.034e+00   1.809  0.07369 .
## DR.POSTEN    1.742e+01  5.417e+00   3.216  0.00178 **
## OUTDOOR      3.756e+01  1.035e+01   3.631  0.00046 ***
## RADIO       -6.633e+01  2.957e+01  -2.243  0.02724 *
## PRINT        6.304e+00  5.876e+00   1.073  0.28610
## SOCIALMEDIA  1.832e+02  2.400e+01   7.636 1.84e-11 ***
## Rain..mm.    2.796e+05  1.461e+05   1.913  0.05877 .
## sal          6.885e+06  3.734e+06   1.844  0.06834 .
## HOLIDAY      2.906e+06  6.383e+06   0.455  0.64999
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14380000 on 94 degrees of freedom
## Multiple R-squared:  0.8203, Adjusted R-squared:  0.8012
## F-statistic: 42.92 on 10 and 94 DF,  p-value: < 2.2e-16
```

Tablica 4.1 Rezultat procjene linearnog modela [13, Tablica 2]

Obzirom da su korišteni podaci vremenski niz, za linearnu formu očekivana je relativno visoka vrijednost za koeficijent determinacije R^2 koji je mjera za *dobro poklapanje* procijenjenih vrijednosti sa podacima. Vrijednosti F-statistike u svim slučajevima pokazuje da su sva tri modela (osobita prva dva) velike značajnosti. Broj značajnih parametara blago varira u svakom modelu. Značajni parametri koji su zajednički svim modelima su TV, DR.POSTEN, OUTDOOR, RADIO, SOCIALMEDIA. Za njih je p-vrijednost manja od 0.05. U log-linearnom modelu također su značajni parametri za DR, Rain i sal, dok su u log-log modelu log(PRINT) te Rain.

```
##
## Call:
## lm(formula = "log(SALES_TOT) ~ TV+DR+DR.POSTEN+OUTDOOR+RADIO+PRINT+SOCIALMEDIA+
    Rain..mm.+sal+HOLIDAY",
##     data = regdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20887 -0.07726 -0.00648  0.06248  0.32894
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.787e+01  3.063e-02 583.581  < 2e-16 ***
## TV           1.661e-07  1.972e-08   8.425 4.07e-13 ***
## DR           1.273e-07  6.171e-08   2.062 0.041929 *
## DR.POSTEN    1.423e-07  4.161e-08   3.420 0.000929 ***
## OUTDOOR      3.334e-07  7.946e-08   4.196 6.15e-05 ***
## RADIO       -6.494e-07  2.271e-07  -2.860 0.005224 **
## PRINT        7.119e-08  4.513e-08   1.577 0.118097
## SOCIALMEDIA  1.547e-06  1.843e-07   8.393 4.76e-13 ***
## Rain..mm.    2.251e-03  1.122e-03   2.006 0.047757 *
## sal          5.934e-02  2.868e-02   2.069 0.041258 *
## HOLIDAY     -4.279e-02  4.903e-02  -0.873 0.385054
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1105 on 94 degrees of freedom
## Multiple R-squared:  0.8332, Adjusted R-squared:  0.8154
## F-statistic: 46.95 on 10 and 94 DF,  p-value: < 2.2e-16
```

Tablica 4.2 Rezultat procjene log-linearnog modela [13, Tablica 3]

Napomenimo da su brojevi u Tablici 4.3 procjene za parametre u lineariziranoj verziji log-log modela. Da bismo pronašli procjene za nezavisne varijable koje su bile zabilježene, bilo je potrebno upotrijebiti antilogaritamsku transformaciju. Umjesto uzimanja eksponencijalne procjene iz Tablice 4.3 da bismo dobili odgovarajuće procjene za parametre u log-log modelu, mora se primijeniti iduća korekcija:

$$\hat{\gamma} = \exp(\hat{\gamma}^*)\exp\left(-\frac{1}{2}\sigma^2_{\hat{\gamma}^*}\right).$$

Procjene $\hat{\gamma}_8, \hat{\gamma}_9, \hat{\gamma}_{10}$ iz jednadžbe multiplikativnog modela postaju:

$$\hat{\gamma}_8 = e^{0.004394} \cdot e^{-\frac{1}{2}0.001918^2} = 1.0044$$
$$\hat{\gamma}_9 = e^{0.049464} \cdot e^{-\frac{1}{2}0.0471644^2} = 1.0495$$
$$\hat{\gamma}_{10} = e^{-0.050739} \cdot e^{-\frac{1}{2}0.0864388^2} = 0.9470$$

```
##
## Call:
## lm(formula = "log(SALES_TOT) ~ log(TV)+log(DR)+log(DR.POSTEN)+log(OUTDOOR)+log(
    RADIO)+log(PRINT)+log(SOCIALMEDIA)+Rain..mm.+sal+HOLIDAY",
##     data = regdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36417 -0.10845  0.01072  0.06971  0.79316
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      14.953425   0.501883  29.795  < 2e-16 ***
## log(TV)           0.012518   0.003359   3.727 0.000331 ***
## log(DR)           0.005367   0.007390   0.726 0.469484
## log(DR.POSTEN)    0.013922   0.006767   2.057 0.042410 *
## log(OUTDOOR)      0.052534   0.014639   3.589 0.000530 ***
## log(RADIO)       -0.012448   0.004007  -3.107 0.002500 **
## log(PRINT)        0.181852   0.036741   4.950 3.26e-06 ***
## log(SOCIALMEDIA)  0.012168   0.003768   3.229 0.001709 **
## Rain..mm.         0.004394   0.001918   2.292 0.024164 *
## sal               0.049464   0.047164   1.049 0.296974
## HOLIDAY          -0.050739   0.086438  -0.587 0.558615
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1858 on 94 degrees of freedom
## Multiple R-squared:  0.5279, Adjusted R-squared:  0.4777
## F-statistic: 10.51 on 10 and 94 DF,  p-value: 1.001e-11
```

Tablica 4.3 Rezultat procjene log-log modela [13, Tablica 4]

Kako bi se testirala pretpostavka za očekivanje reziduala različitog od nule za sva tri modela, prema [13, str. 26-27] ispitani su dijagrami reziduala u odnosu na prediktorsku varijablu. Iz [13, str. 27], donesen je zaključak da je najprikladniji oblik log-linearni.

Bibliografija

[1] Y. Dodge, *The Concise Encyclopedia of Statistics,* Springer, 2008.

[2] Z. Vondraček, N. Sandrić, *Vjerojatnost,* https://web.math.pmf.unizg.hr/nastava/vjer/files/vjer_predavanja.pdf

[3] N. M. Nathan, *Equivalence of MLE and OLS in linear regression,* https://snaveenmathew.medium.com/equivalence-of-mle-and-ols-in-linear-regression-d3e44e47df3c

[4] M. Huzak, *Vjerojatnost i matematička statistika,* http://aktuari.math.pmf.unizg.hr/docs/vms.pdf

[5] N. H. Bingham, J. M. Fry, *Regression,* Springer, 2010.

[6] T. S. Blyth, E. F. Robertson, *Basic linear algebra,* Springer, 2002.

[7] V. Hari, *Linearna algebra,* https://web.math.pmf.unizg.hr/~hari/LA.pdf

[8] Z. Drmač, *Numerička matematika*, https://web.math.pmf.unizg.hr/~drmac/na001.pdf

[9] T. S. Blyth, E. F. Robertson, *Further linear algebra,* Springer, 2002.

[10] P.R. Halmos, *Finite-dimensional vector spaces,* Undergraduate Texts in Mathematics, Springer, 1979.

[11] *Modeling Markets: Analyzing Marketing Phenomena and Improving Marketing Decision Making,* International Series in Quantitive Marketing, 2015.

[12] *Market Response Models: Econometric and Time Series Analysis,* Volume 12, 2001.

[13] E. Mhitarean-Cuvsinov, *Marketing Mix Modelling from multiple regression perspective,* https://www.math.kth.se/matstat/seminarier/reports/M-exjobb17/170524b.pdf

Sažetak

U ovom diplomskom radu proučili smo linearnu regresiju s naglaskom na jednostavnu i višestruku regresiju. Za početak, rekli smo nešto o povijesti regresije, bivarijantnoj normalnoj distribuciji te njenim svojstvima.
Proučili smo metodu maksimalne vjerodostojnosti te metodu najmanjih kvadrata, koja je i najstarija metoda, te je radi svoje jednostavnosti najčešće korištena za procjenu parametara. Od interesa nam je bila i analiza varijance te smo kroz primjer vidjeli i njenu primjenu.
Kraj ovog diplomskog rada ilustrira primjenu suvremenih pristupa na skupu podataka tvrtke Nepa. Cilj posljednjeg poglavlja bio je konstruirati strategiju kako napraviti model prikladan za visoku razinu složenosti podataka.
Marketing Mix model bavi se svim elementima problema koji se proučava. Jedan od tih elemenata je izbor odgovarajuće funkcionalne forme. U ovom slučaju, dijagram reziduala sugerira da je log-linearan model najprikladniji.

Summary

In this thesis, we are introduced to the linear regression with an emphasis on simple and multiple regression. To begin with, we said something about the origins of regression, the bivariate normal distribution and its facts.
We have seen maximum likelihood method and the method of least squares, which is also the oldest method and, due to its simplicity, is most often used for parameter estimation.
The analysis of variance was also of interest to us, and we saw its purpose of using through an example.
The end of this thesis illustrates an application of modern approaches of statistical learning on a set of data provided by Nepa. The goal was to construct a model building strategy suitable for a high level of complexity of the data. A Marketing Mix model must address all elements of the problem being studied. One of such elements is the choice of the appropriate functional form. At the end, the plot of the residuals against each predictor variable suggest that the log-linear specification is appropriate in this case.

Životopis

Rođena sam 25. veljače 1995. u Zagrebu. Nakon završetka Osnovne škole „Matija Gubec“ u Zagrebu, upisujem XI. Opću gimnaziju. Tokom osnovnoškolskog obrazovanja sudjelovala sam na natjecanjima iz kemije, biologije i matematike, a tokom srednjoškolskog iz hrvatskog jezika. Kao dodatnu aktivnost, sudjelovala sam u plesnoj te odbojkaškoj skupini te natjecanjima iz Prve pomoći. Kroz čitavo dotadašnje obrazovanje bila sam član dramske skupine te sudjelovala na natjecanjima LiDraNo. Završetkom srednje škole, 2013. godine upisujem preddiplomski sveučilišni studij Matematika na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu. Titulu sveučilišne prvostupnice matematike stječem 2019. kada upisujem diplomski sveučilišni studij Financijska i poslovna matematika na istom fakultetu kojeg završavam ovim Radom.