Analiza varijance kao linearni regresijski model

Krajnović, Maja

Master's thesis / Diplomski rad

2020

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **Josip Juraj Strossmayer University of Osijek, Department of Mathematics / Sveučilište Josipa Jurja Strossmayera u Osijeku, Odjel za matematiku**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:126:546017

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-04-25**

Repository / Repozitorij:

Repository of School of Applied Mathematics and Computer Science

Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku
Diplomski studij matematike
Financijska matematika i statistika

Maja Krajnović

Analiza varijance kao linearni regresijski model

Diplomski rad

Osijek, 2020.

Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku
Diplomski studij matematike
Financijska matematika i statistika

Maja Krajnović

Analiza varijance kao linearni regresijski model

Diplomski rad

Mentor: prof. dr. sc. Mirta Benšić

Osijek, 2020.

Sadržaj

Uvod **1**

1 Osnovni pojmovi **3**
1.1 Kvadratna forma 5
1.2 Linearni regresijski model 6
1.3 Procjenitelj metodom najmanjih kvadrata 6
1.4 Normalni regresijski model 8
1.5 F Test 9

2 Anliza varijance **12**
2.1 Jednosmjerna klasifikacija 14
2.1.1 Uvod i motivacija 14
2.1.2 Analiza varijance za jednosmjernu klasifikaciju 16
2.1.3 Primjer 21
2.2 Dvosmjerna klasifikacija 22
2.2.1 Uvod i motivacija 22
2.2.2 Analiza varijance za dvosmjernu klasifikaciju 23
2.2.3 Primjer 29

Literatura **30**

Sažetak **31**

Summary **32**

Životopis **33**

Uvod

U ovom radu bavimo se analizom varijance (engl. analysis of variance - ANOVA) koju je originalno razvio R.A.Fisher[1] 1920-ih godina. Analiza varijance je statistička metoda koja se koristi za testiranje postojanja statistički značajnih razlika između dviju ili više grupa podataka. Prije postojanja ANOVA procedure korišten je t-test, no problem je što on, za razliku od ANOVA procedure, nije primjenjiv na više od dvije grupe podataka.
Analizu varijance možemo provesti na tri načina s obzirom na klasifikaciju podataka pa stoga razlikujemo jednosmjernu, dvosmjernu te višesmjernu klasifikaciju. Kod jednosmjerne klasifikacije pri izgradnji linearnog modela koristimo samo jednu nezavisnu varijablu, u dvosmjernoj klasifikaciji koristimo dvije nezavisne varijable, a u višesmjernoj koristimo tri ili više nezavisnih varijabli. ANOVA metodologija se može objasniti na različite načine. U ovom radu odabran je pristup prezentiranja ANOVA procedura korištenjem teorije linearnog regresijskog modela. Primjena analize varijance je vrlo široka i primjenjiva je u gotovo svim područjima, a neki od primjera su:

- medicina - testiranje učinka više različitih terapija na pacijente sa sličnim dijagnozama,
- poljoprivreda - testiranje učinka različitih pesticida te vremenskih uvjeta na urod poljoprivredne kulture,
- školstvo - utvrđivanje postojanja razlike u rezultatu učenika kod kojih je primijenjen novi program obrazovanja u odnosu na klasični,
- industrija - testiranje kvalitete automobila ovisno o tvornici koja ga proizvodi.

Nakon uvodnog dijela, u prvom poglavlju rada definirat ćemo osnovne pojmove koji će nam biti potrebni za analizu varijance kao što su kvadratna forma, linearan model, procjenitelja metodom najmanjih kvadrata, normalni regresijski model, F test i drugi. U sljedećem, glavnom poglavlju ovog rada dajemo teorijski uvod u sam proces analize varijance definirajući neke od jednakosti potrebnih u nastavku od kojih su najvažniji rastavi ukupne varijabilnosti modela na komponente kako bi se analiza olakšala. Poglavlje smo podijelili na dva dijela. U prvom dijelu provodimo analizu varijance za jednosmjernu klasifikaciju podataka u kojoj pri izgradnji modela koristimo jednu zavisnu, kontinuiranu varijablu te jednu nezavisnu, kategorijalnu varijablu. Ovdje smo definirali sve potrebne sume kvadrata te F-statistike za testiranje hipoteza o jednakosti parametara unutar grupa i među grupama podataka te smo dobivene rezultate obuhvatili u

[1]Ronald Aylmer Fisher (1890.-1962.) britanski statističar i genetičar

obliku ANOVA tablice. Postupak analize varijance smo demonstrirali i na primjeru o korištenju tri različite dijete. U drugom dijelu glavnog poglavlja bavili smo se dvosmjernom klasifikacijom podataka u okviru analize varijance korištenjem linearnog regresijskog modela u kojem smo uz jednu zavisnu, kontinuiranu varijablu koristili dvije nezavisne, kategorijalne varijable pri izgradnji modela. Također smo, kao i u prethodnom dijelu, odredili sve potrebne sume kvadrata te F-statistike kako bismo u okviru modela za dvosmjernu klasifikaciju analizom varijance odgovorili na pitanja o postojanju razlike između podataka klasificiranih u retke, odnosno stupce te nas je zanimalo postojanje interakcijskog efekta između dvije nezavisne varijable modela. Sve dobivene rezultate smo prikazali u obliku ANOVA tablice, a dodatno smo sve demonstrirali na primjeru.

Poglavlje 1

Osnovni pojmovi

Prije samog postupka objašnjenja analize varijance korištenjem linearnog regresijskog modela definirat ćemo neke ključne pojmove koji su nam potrebni u radu. Za početak ćemo definirati bitne distribucije koje će nam biti izuzetno korisne.
Pri analizi distribucije kvadratne forme normalnog slučajnog vektora spomenut ćemo centralnu te necentralnu χ^2 distribuciju. Centralna χ^2 distribucija s n stupnjeva slobode je distribucija sume kvadrata n nezavisnih, standardnih, normalnih slučajnih varijabli. U sljedećoj definiciji definiramo i necentralnu χ^2 distribuciju.

Definicija 1.0.1. *Neka su $Z_1, \ldots, Z_n$ nezavisne, normalne slučajne varijable, $Z_i \sim \mathcal{N}(\mu_i, 1)$. Tada kažemo da slučajna varijabla $W = \sum_{i=1}^{n} Z_i^2$ ima necentralnu χ^2 distribuciju s n stupnjeva slobode i parametrom necentralnosti $\gamma = \frac{1}{2}\sum_{i=1}^{n} \mu_i^2$. Pišemo $W \sim \chi^2(n, \gamma)$.*

Jedna od distribucija zasnovanih na χ^2 distribuciji je Snedecorova[1] F distribucija. U analizi varijance F distribucija je upravo distribucija test statistike koju koristimo.

Definicija 1.0.2. *Neka su $V_1 \sim \chi^2(n)$ i $V_2 \sim \chi^2(k)$ nezavisne slučajne varijable, tada slučajna varijabla*

$$X = \frac{\frac{V_1}{n}}{\frac{V_2}{k}}$$

ima tzv. F distribuciju sa stupnjem slobode brojnika n, a nazivnika k što zapisujemo kao $\mathrm{F}(n, k)$.

Također ćemo koristiti i necentralnu F distribuciju čiju definiciju navodimo u nastavku.

[1] George Waddel Snedecor (1881.-1974.) američki matematičar i statističar

Definicija 1.0.3. *Neka su* $V_1 \sim \chi^2(n, \gamma)$ *i* $V_2 \sim \chi^2(k)$ *nezavisne slučajne varijable, tada slučajna varijabla*

$$X = \frac{\frac{V_1}{n}}{\frac{V_2}{k}}$$

ima necentralnu F *distribuciju sa* n *i* k *stupnjeva slobode i parametrom necentralnosti* γ *što zapisujemo kao* $\mathrm{F}(n, k, \gamma)$.

1.1 Kvadratna forma

Kvadratna forma je općenito homogeni polinom drugog stupnja od n varijabli. Sljedeća definicija definira kvadratnu formu slučajnog vektora koja će nam biti korisna u nastavku rada.

Definicija 1.1.1. *Neka je* $\mathbf{Y}$ *n-dimenzionalni slučajni vektor i* A $n \times n$ *matrica. Kvadratna forma slučajnog vektora* $\mathbf{Y}$ *je slučajna varijabla* $\mathbf{Y}'A\mathbf{Y}$.

Sljedeći teorem daje nam informaciju o distribuciji kvadratne forme normalnog slučajnog vektora.

Teorem 1.1.2. *Neka je* $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathrm{I})$ *slučajan vektor i* M *matrica ortogonalne projekcije. Tada vrijedi da je* $\mathbf{Y}'M\mathbf{Y} \sim \chi^2(rang(M), \frac{1}{2}\boldsymbol{\mu}'M\boldsymbol{\mu})$.

Dokaz. Vidi [7]. □

Nadalje, pretpostavimo da je $\mathbf{Y}$ slučajan vektor koji se sastoji od n nezavisnih, normalnih slučajnih varijabli s očekivanjem $E(\mathbf{Y}) = \boldsymbol{\mu}$ i matricom kovarijanci $\boldsymbol{V}(\mathbf{Y}) = \mathbf{I}$, te da vrijedi

$$\mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{k} \mathbf{Y}'A_i\mathbf{Y},$$

pri čemu matrica A_i ima rang r_i. Možemo primijetiti kako ovaj zapis odgovara sumi k kvadratnih formi. Posebno nas zanimaju distribucije kvadratnih formi $Q_i = \mathbf{Y}'A_i\mathbf{Y}$ te uvjeti njihove nezavisnosti. Prema teoremu 1.1.2 znamo da $\mathbf{Y}'\mathbf{Y}$ ima $\chi^2(n, \boldsymbol{\mu}'\boldsymbol{\mu})$ što znači da očekujemo da sve komponente imaju jednak tip distribucije, a odgovor leži u sljedećem teoremu kojeg zovemo Cochranov teorem.

Teorem 1.1.3. *Cochranov*[2] *teorem*
Neka je $\mathbf{Y} = Y_1, Y_2, \ldots, Y_n$ *pri čemu su* $Y_i \sim \mathcal{N}(0, 1)$, $i = 1, \ldots, n$ *te pretpostavimo da vrijedi*

$$\sum_{i=1}^{n} Y_i^2 = Q_1 + Q_2 + \cdots + Q_k,$$

gdje su $Q_1, Q_2, \ldots, Q_k$ *pozitivne, semi-definitne kvadratne forme slučajnih varijabli* $Y_1, Y_2, \ldots, Y_n$ *pri čemu su* $Q_j = \mathbf{Y}'A_j\mathbf{Y}$ *za* $j = 1, \ldots, k$ *te neka je* $rang A_j = r_j$. *Ukoliko vrijedi*

$$r_1 + r_2 + \cdots + r_k = n$$

tada su kvadatne forme $Q_1, Q_2, \ldots, Q_k$ *nezavisne i vrijedi* $Q_j \sim \chi^2(r_j)$.

Dokaz. Vidi [10]. □

Napomenimo da u nastavku rada, zbog jednostavnosti, u oznakama nećemo praviti razliku između slučajne varijable i njene realizacije, podrazumijevajući da izraz u kojem se koristi određuje značenje.

[2] William Gemmell Cochran (1909.-1980.) američki statističar

1.2 Linearni regresijski model

Kako se u ovom radu bavimo analizom varijance pomoću linearnog regresijskog modela potrebno je definirati i sam linearni regresijski model te grešku modela.
Neka su dani podaci $(\mathrm{y}_i, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, vektor koeficijenata $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_k)'$ te vektor grešaka $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$. Linearni regresijski model pretpostavlja da vrijedi

$$\mathrm{y}_i = \theta_0 + \theta_1 x_{i1} + \cdots + \theta_k x_{ik} + \epsilon_i = \mathbf{x}_i'\boldsymbol{\theta} + \epsilon_i, \ i = 1, \ldots, n,$$

što u matričnoj notaciji možemo zapisati kao

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{1.1}$$

gdje je $\mathbf{y}$ vektor opažaja, a $\boldsymbol{X}$ fiksna matrica dizajna sljedećeg oblika

$$\mathbf{y} = \begin{bmatrix} \mathrm{y}_1 \\ \mathrm{y}_2 \\ \vdots \\ \mathrm{y}_n \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix}.$$

Grešku modela $\boldsymbol{\epsilon}$ $(n \times 1)$ možemo zapisati kao $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\theta}$ te pretpostavljamo da za njeno očekivanje i matricu kovarijanci vrijedi

$$E[\boldsymbol{\epsilon}|\boldsymbol{X}] = 0, \tag{1.2}$$

$$\boldsymbol{V}(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\mathbf{I}. \tag{1.3}$$

1.3 Procjenitelj metodom najmanjih kvadrata

Neka je dan linearan model oblika (1.1) pri čemu su podaci $\{(\mathrm{y}_1, \mathbf{x}_1), \ldots, (\mathrm{y}_n, \mathbf{x}_n)\}$ jednako distribuirani. Procjena parametra $\boldsymbol{\theta}$ metodom najmanjih kvadrata (eng. least squares - LS) podrazumijeva minimizaciju srednje kvadratne greške $S(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$ po $\boldsymbol{\theta}$, odnosno mora vrijediti

$$\frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0.$$

Diferenciranjem dobivamo sljedeću jednadžbu

$$2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = 0.$$

Rješavanjem prethodne jednadžbe po $\boldsymbol{\theta}$ dobivamo procjenitelja metodom najmanjih kvadrata kojeg označavamo s $\hat{\boldsymbol{\theta}}$ te on iznosi

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{1.4}$$

Zanima nas je li dobiveni procjenitelj nepristran. Kako bismo to pokazali prvo ćemo procjenitelj $\hat{\boldsymbol{\theta}}$ zapisati koristeći (1.1). Slijedi da je

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}) = \boldsymbol{\theta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}. \tag{1.5}$$

Kako je $\boldsymbol{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ iz prethodnog zapisa zaključujemo da je $\boldsymbol{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$, što znači da je $\hat{\boldsymbol{\theta}}$ nepristran procjenitelj za parametar $\boldsymbol{\theta}$.
Zanima nas i matrica kovarijanci procjenitelja $\hat{\boldsymbol{\theta}}$,

$$\begin{aligned}
\boldsymbol{V}(\hat{\boldsymbol{\theta}}) &= \boldsymbol{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'] \\
&= \boldsymbol{E}[((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon})'] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\boldsymbol{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned} \tag{1.6}$$

Kao nusprodukte procjene definiramo:

- teorijske vrijednosti $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\theta}}$,
- reziduale $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}}$.

Kako su greške nemjerljive veličine reziduali će nam biti od velike koristi pri analizi.
Sada ćemo navesti i dva bitna svojstva reziduala koja ćemo u nastavku rada koristi pri zaključivanju o ortogonalnosti.

$$\begin{aligned}
\sum_{i=1}^{n} \mathbf{x}_i\hat{\epsilon}_i &= \sum_{i=1}^{n} \mathbf{x}_i(\mathrm{y}_i - \mathbf{x}_i'\hat{\boldsymbol{\theta}}) \\
&= \sum_{i=1}^{n} \mathbf{x}_i\mathrm{y}_i - \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\hat{\boldsymbol{\theta}} \\
&= \sum_{i=1}^{n} \mathbf{x}_i\mathrm{y}_i - \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{n} \mathbf{x}_i\mathrm{y}_i\right) \\
&= \sum_{i=1}^{n} \mathbf{x}_i\mathrm{y}_i - \sum_{i=1}^{n} \mathbf{x}_i\mathrm{y}_i = 0,
\end{aligned} \tag{1.7}$$

odnosno zaključujemo da je uzoračka korelacija između regresora i reziduala jednaka nula. Također, ukoliko matrica dizajna $\mathbf{X}$ sadrži konstantu vrijedi

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i = 0, \tag{1.8}$$

odnosno, zaključujemo da je uzoračka aritmetička sredina reziduala jednaka nula.

1.4 Normalni regresijski model

Normalni regresijski model je linearni regresijski model s nezavisnom, normalno distribuiranom greškom, odnosno $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. Funkcija vjerodostojnosti normalnog regresijskog modela je

$$L(\boldsymbol{\theta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(\mathrm{y_i} - \mathbf{x}_i'\boldsymbol{\theta})^2},$$

a prema tome je logaritam vjerodostojnosti jednak

$$l(\boldsymbol{\theta}, \sigma^2) = \log L(\boldsymbol{\theta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathrm{y}_i - \mathbf{x}_i'\boldsymbol{\theta})^2. \tag{1.9}$$

Kako bismo odredili procjenitelje za parametar $\boldsymbol{\theta}$ i varijancu greške σ^2, logaritam vjerodostojnosti ćemo maksimizirati po $\boldsymbol{\theta}$ i σ^2, odnosno mora vrijediti

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i(\mathrm{y}_i - \mathbf{x}_i'\boldsymbol{\theta}) = 0,$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{n}(\mathrm{y}_i - \mathbf{x}_i'\boldsymbol{\theta})^2 = 0.$$

Rješavanjem ovih jednadžbi dobivamo procjenitelje

$$\hat{\boldsymbol{\theta}} = \left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i\mathrm{y}_i\right),$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i,$$

pa $\hat{\boldsymbol{\theta}}$ odgovara procjenitelju metodom najmanjih kvadrata (1.4). Uvrstimo li ove vrijednosti u (1.9) dobivamo vrijednost maksimuma logaritama vjerodostojnosti

$$l(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{n}{2}. \tag{1.10}$$

1.5 F Test

Kako u nastavku rada želimo testirati hipoteze na skupu regresijskih koeficijenata to u normalnom regresijskom modelu možemo učiniti pomoću F testa koji se izvodi iz testa kvocijenta vjerodostojnosti (engl. likelihood ratio test). Za početak ćemo regresore rastaviti na sljedeći način $\mathbf{x}_i = (\mathbf{x}'_{1i}, \mathbf{x}'_{2i})$, a slično i vektor koeficijenata $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2)$. Uz ove oznake regresijski model možemo zapisati kao

$$\mathrm{y}_i = \mathbf{x}'_{1i}\boldsymbol{\theta}_1 + \mathbf{x}'_{2i}\boldsymbol{\theta}_2 + \epsilon_i,$$

uz pretpostavku da je $k = dim(\mathbf{x}_i)$, $r = dim(\mathbf{x}_{1i})$ te $q = dim(\mathbf{x}_{2i})$ pri čemu vrijedi $k = r + q$. Želimo testirati sljedeće hipoteze

$$H_0 : \boldsymbol{\theta}_2 = 0,$$
$$H_1 : \boldsymbol{\theta}_2 \neq 0$$

te zaključujemo da u uvjetima istinitosti hipoteze H_0 regresore $\mathbf{x}'_{2i}$ možemo izostaviti iz regresije čime dobivamo sljedeći model

$$\mathrm{y}_i = \mathbf{x}'_{1i}\boldsymbol{\theta}_1 + \epsilon_i. \tag{1.11}$$

Prilikom kreiranja testa metodom kvocijenta vjerodostojnosti pravilo je odbaciti H_0 u korist H_1 za velike vrijednosti kvocijenta vjerodostojnosti koji je kvocijent maksimizirane funkcije vjerodostojnosti u uvjetima H_1, odnosno za puni model i maksimizirane funkcije vjerodostojnosti u uvjetima H_0, odnosno za restringirani model. Želimo konstruirati ovakvu statistiku u normalnom regresijskom modelu koristeći se maksimumom logaritama vjerodostojnosti (1.10) koja odgovara punom modelu. Potrebno je još odrediti maksimumom logaritama vjerodostojnosti za model (1.11). Procjenitelje za restringirani model računamo na sličan način kao i za puni model pri čemu će procjenitelj metodom maksimalne vjerodostojnosti odgovarati procjenitelju metodom najmanjih kvadrata za y_i i $\mathbf{x}_{1i}$. Dobiveni procjenitelji su:

$$\tilde{\boldsymbol{\theta}}_1 = (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{y},$$

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\tilde{\epsilon}_i,$$

pri čemu su reziduali

$$\tilde{\epsilon}_i = \mathrm{y}_i - \mathbf{x}'_{1i}\tilde{\boldsymbol{\theta}}_1.$$

Uvrštavanjem dobivenih procjenitelja u jednadžbu (1.9) dobivamo vrijednost maksimuma logaritama vjerodostojnosti za restringirani model

$$l(\tilde{\boldsymbol{\theta}}_1, \tilde{\sigma}^2) = -\frac{n}{2}\log(2\pi\tilde{\sigma}^2) - \frac{n}{2}.$$

Sada možemo definirati test statistiku za testiranje H_0 nasuprot H_1 kao

$$\begin{aligned} LR &= 2[l(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) - l(\tilde{\boldsymbol{\theta}}_1, \tilde{\sigma}^2)] \\ &= 2\left[\left(-\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{n}{2}\right) - \left(-\frac{n}{2}\log(2\pi\tilde{\sigma}^2) - \frac{n}{2}\right)\right] \\ &= n\log\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right). \end{aligned}$$

Test kvocijenata vjerodostojnosti (engl. Likelihood ratio test (LR)) odbacuje hipotezu H_0 za velike vrijednosti test statistike LR, a ekvivalent tome je F statistika za testiranje H_0 nasuprot H_1:

$$F = \frac{\dfrac{\tilde{\sigma}^2 - \hat{\sigma}^2}{q}}{\dfrac{\hat{\sigma}^2}{n-k}}, \tag{1.12}$$

pri čemu F test također odbacuje hipotezu H_0 za velike vrijednosti test statistike F. Ekvivalencija test statistike LR i F statistike dolazi od ekvivalencije sljedećih tvrdnji "Odbaci H_0 za $LR \geq c_1$" i "Odbaci H_0 za $F \geq c_2$" za $c_2 = (\exp(c_1/n) - 1)(n-k)/q$. Distribuciju ove F statistike odredit ćemo koristeći matricu projekcije $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ i matricu ortogonalne projekcije $\mathbf{M} = \mathbf{I}_n - \mathbf{P}$ koje imaju sljedeća svojstva:

1. $\mathbf{P}' = \mathbf{P}$, $\mathbf{M}' = \mathbf{M}$, $\mathbf{P}^2 = \mathbf{P}$, $\mathbf{M}^2 = \mathbf{M}$,
2. $\mathbf{PX} = \mathbf{X}$, $\mathbf{MX} = 0$,
3. $\mathrm{tr}\mathbf{P} = k$, $\mathrm{tr}\mathbf{M} = n - k$.

Matrica P u metodi najmanjih kvadrata kreira teorijske vrijednosti dok matrica M kreira reziduale, odnosno vrijedi

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y} = \mathbf{P}\mathbf{y},$$
$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - P)\mathbf{y} = \mathbf{M}\mathbf{y}.$$

Koristeći svojstvo 2. i jednadžbu modela (1.1) reziduale možemo zapisati kao $\hat{\boldsymbol{\epsilon}} = \mathbf{M}\boldsymbol{\epsilon}$.
Ranije smo odredili procjenitelje varijance greške $\hat{\sigma}^2$ i $\tilde{\sigma}^2$ koje u matričnoj notaciji možemo zapisati kao

$$\hat{\sigma}^2 = \frac{1}{n}\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}},$$
$$\tilde{\sigma}^2 = \frac{1}{n}\tilde{\boldsymbol{\epsilon}}'\tilde{\boldsymbol{\epsilon}}.$$

Koristeći izraz $\hat{\boldsymbol{\epsilon}} = \mathbf{M}\mathbf{y} = \mathbf{M}\boldsymbol{\epsilon}$ i svojstva matrice $\mathbf{M}$ za procjenitelj $\hat{\sigma}^2$ vrijedi

$$\hat{\sigma}^2 = \frac{1}{n}\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = \frac{1}{n}\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}.$$

U uvjetima istinitosti hipoteze H_0 vrijedi $n\tilde{\sigma}^2 = \boldsymbol{e}'\mathbf{M}_1\boldsymbol{e}$ pri čemu je $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{P}_1$, gdje je $\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$. Primjećujemo da je $\mathbf{M}_1 - \mathbf{M} = \mathbf{P} - \mathbf{P}_1$ idempotentna s rangom q te je $(\mathbf{M}_1 - \mathbf{M})\mathbf{M} = 0$ pa slijedi da je $\boldsymbol{e}'(\mathbf{M}_1 - \mathbf{M})\boldsymbol{e} \sim \chi^2(q)$ i nezavisna je od $\boldsymbol{e}'\mathbf{M}\boldsymbol{e}$. Test statistiku (1.12) sada možemo zapisati u matričnom obliku

$$F = \frac{\boldsymbol{e}'(\mathbf{M}_1 - \mathbf{M})\boldsymbol{e}/q}{\boldsymbol{e}'\mathbf{M}\boldsymbol{e}/(n-k)} \sim \frac{\chi^2(q)/q}{\chi^2(n-k)/(n-k)}$$

i zaključujemo da ona ima F distribuciju sa stupnjem slobode brojnika q, a nazivnika $n-k$, odnosno $F \sim \mathrm{F}(q, n-k)$.
Označimo li sa S_E sumu kvadrata grešaka tako da je

$$S_E(\boldsymbol{\theta}) = \sum_{i=1}^{n} (\mathrm{y}_i - \mathbf{x}_i'\boldsymbol{\theta})^2,$$

tada test statistiku (1.12) možemo zapisati i u sljedećem obliku

$$F = \frac{S_E(\tilde{\boldsymbol{\theta}}) - S_E(\hat{\boldsymbol{\theta}})}{qs^2},$$

gdje je $s^2 = \dfrac{1}{n-k}\displaystyle\sum_{i=1}^{n}(\mathrm{y}_i - \mathbf{x}_i'\hat{\boldsymbol{\theta}})^2$.

Poglavlje 2

Anliza varijance

Cilj analize varijance je testirati hipoteze koje se odnose na vrijednosti odabranih podskupina parametara kako bismo mogli donijeti zaključke o postajanju razlika među njima, a ideja je da sumu kvadrata $\mathbf{y}'\mathbf{y}$ prikažemo kao sumu nenegativnih komponenti od kojih svaka odgovara podskupini parametara linearnog modela čime ćemo dodatno pojednostaviti analizu.
Za početak ćemo linearni model (1.1) zapisati pomoću matrice projekcije $\mathbf{P}$ i matrice ortogonalne projekcije $\mathbf{M}$ koje smo definirali u prethodnom poglavlju. Vrijedi:

$$\mathbf{y} = \mathbf{P}\mathbf{y} + \mathbf{M}\mathbf{y} = \hat{\mathbf{y}} + \hat{\boldsymbol{\epsilon}}. \tag{2.1}$$

Primjenom svojstava matrica $\mathbf{P}$ i $\mathbf{M}$ zaključujemo da je dekompozicija (2.1) ortogonalna, odnosno vrijedi

$$\hat{\mathbf{y}}'\hat{\boldsymbol{\epsilon}} = (\mathbf{P}\mathbf{y})'(\mathbf{M}\mathbf{y}) = \mathbf{y}'\mathbf{P}\mathbf{M}\mathbf{y} = 0.$$

Sada sumu kvadrata $\mathbf{y}'\mathbf{y}$, koristeći prethodno definiran model (2.1) i svojstvo ortogonalnosti, možemo zapisati na sljedeći način

$$\mathbf{y}'\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + 2\hat{\mathbf{y}}'\hat{\boldsymbol{\epsilon}} + \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}, \tag{2.2}$$

odnosno vrijedi

$$\sum_{i=1}^{n} \mathrm{y}_i^2 = \sum_{i=1}^{n} \hat{\mathrm{y}}_i^2 + \sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

Oduzmemo li $\overline{\mathbf{y}}$, odnosno prosjek podataka, s obje strane jednakosti (2.1) dobivamo:

$$\mathbf{y} - \mathbf{1}_n\overline{\mathbf{y}} = \hat{\mathbf{y}} - \mathbf{1}_n\overline{\mathbf{y}} + \hat{\boldsymbol{\epsilon}}$$

Pokazat ćemo da je i ova dekompozicija ortogonalna u slučaju kada matrica dizajna sadrži konstantu. Koristeći svojstvo reziduala (1.8) dobivamo:

$$(\hat{\mathbf{y}} - \mathbf{1}_n\overline{\mathbf{y}})'\hat{\boldsymbol{\epsilon}} = \hat{\mathbf{y}}'\hat{\boldsymbol{\epsilon}} - \overline{\mathbf{y}}\mathbf{1}_n'\hat{\boldsymbol{\epsilon}} = 0.$$

Kako je dekompozicija ortogonalna dobivamo sumu kvadrata

$$(\mathbf{y} - \mathbf{1}_n\overline{\mathbf{y}})'(\mathbf{y} - \mathbf{1}_n\overline{\mathbf{y}}) = (\hat{\mathbf{y}} - \mathbf{1}_n\overline{\mathbf{y}})'(\hat{\mathbf{y}} - \mathbf{1}_n\overline{\mathbf{y}}) + \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}$$

koju možemo zapisati i u sljedećem obliku:

$$\sum_{i=1}^{n}(\mathrm{y}_i - \overline{\mathbf{y}})^2 = \sum_{i=1}^{n}(\hat{\mathrm{y}}_i - \overline{\mathbf{y}})^2 + \sum_{i=1}^{n}\hat{\epsilon}_i^2 \tag{2.3}$$

Lijeva strana izraza (2.3) označava kvadratno odstupanje podatka od srednje vrijednosti podataka odnosno ukupnu varijabilnost podataka te se stoga izraz (2.3) često naziva i formula za analizu varijance. Mi ćemo ju u nastavku zvati ukupna suma kvadrata i označavati sa S_T.
Nadalje, koristeći definiciju teorijskih vrijednosti i reziduala iz (2.2) dobivamo sljedeći rastav sume kvadrata $\mathbf{y}'\mathbf{y}$:

$$\mathbf{y}'\mathbf{y} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}) + (\mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{X}\hat{\boldsymbol{\theta}}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}) + \hat{\boldsymbol{\theta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\theta}}. \tag{2.4}$$

Kako bismo mogli testirati željene hipoteze o podskupinama parametara potrebno je odrediti distribucije pojedinih komponenti rastava (2.4) te kreirati test statistiku za testiranje hipoteza. Iz (1.5) možemo primijetit da je $\hat{\boldsymbol{\theta}}$ linearna funkcija greške $\boldsymbol{\epsilon}$ te ukoliko je greška $\boldsymbol{\epsilon}$ normalno distribuirana zaključujemo da $\hat{\boldsymbol{\theta}}$ ima multivarijatnu normalnu distribuciju s očekivanjem $\boldsymbol{\theta}$ i matricom kovarijanci $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$ definiranom kao u (1.6). Također, možemo primijetiti da $\hat{\boldsymbol{\theta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\theta}}/\sigma^2$ ima oblik kvadratne forme te primjenom teorema 1.1.2 zaključujemo da ima necentralnu χ^2 distribuciju s k stupnjeva slobode i parametrom necentralnosti $\lambda = \boldsymbol{\theta}'\boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{\theta} = \boldsymbol{\theta}\mathbf{X}'\mathbf{X}\boldsymbol{\theta}/\sigma^2$.
U nastavku rada, cilj nam je testirat hipoteze sljedećeg oblika:

$$\begin{aligned} H_0 &: \boldsymbol{\theta} = \mathbf{0}, \\ H_1 &: \boldsymbol{\theta} \neq \mathbf{0}. \end{aligned} \tag{2.5}$$

Stoga ćemo promotriti specijalni slučaj kada je hipoteza H_0 istinita. U slučaju istinitosti nul hipoteze parametar necentralnosti λ jednak je 0 te prema tome drugi član u (2.4) ima centralnu χ^2 distribuciju s k stupnjeva slobode. Ostaje nam ispitati tip distribucije prve komponente iz (2.4) ukoliko je H_0 istinita. Primjećujemo da $(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\theta}})/\sigma^2$ ima $\chi^2(n-k)$ distribuciju te prema tome $\mathbf{y}'\mathbf{y}/\sigma^2$ ima $\chi^2(n)$ distribuciju. Ukoliko pretpostavka H_0 vrijedi možemo primijeniti Cochranov teorem (Teorem 1.1.3) prema kojem su $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})$ i $\hat{\boldsymbol{\theta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\theta}}$ nezavisne komponente u (2.4). Ukoliko podijelimo navedene komponente s pripadajućim stupnjevima slobode te ih međusobno podijelimo dobivamo F statistiku za ovaj specijalni slučaj koju ćemo koristiti u nastavku:

$$F = \frac{\dfrac{\hat{\boldsymbol{\theta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\theta}}}{k}}{\dfrac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})}{(n-k)}} \tag{2.6}$$

U uvjetima istinitosti hipoteze H_0 F statistika (2.6) ima F distribuciju $\mathrm{F}(k, n-k)$. Uočimo da ovaj F-test odgovara F-testu iz poglavlja 1.5. ukoliko restringirani model pretpostavlja samo postojanje konstantnog člana, dok svi koeficijenti uz prediktor iščezavaju. U nastavku rada analizu varijance ćemo demonstrirati na primjeru jednosmjerne i dvosmjerne klasifikacije koristeći se rezultatima dobivenim u ovom poglavlju.

2.1 Jednosmjerna klasifikacija

2.1.1 Uvod i motivacija

Kako bismo lakše razumjeli problematiku analize varijance u jednosmjernoj klasifikaciji ovo poglavlje započet ćemo s konkretnim primjerom u kojem ćemo proučavati učinke tri različite dijete na gubitak kilograma.

Primjer 2.1.1. Baza podataka koju koristimo u ovom primjeru preuzeta je s internetske stranice `https://www.scribbr.com/`, a sadrži podatke o 72 osobe koje su u ovom eksperimentu isprobale jednu od tri moguće dijete. Radi se 36 žena i 36 muškaraca pri čemu je svaku od dijeta isprobao jednak broj žena i muškaraca, odnosno 12 žena te 12 muškaraca. Varijable korištene u ovom eksperimentu te njihove opise možemo pogledati u tablici (2.1) dok u tablici (2.2) možemo vidjeti prikaz prvih te zadnjih šest redaka baze podataka.

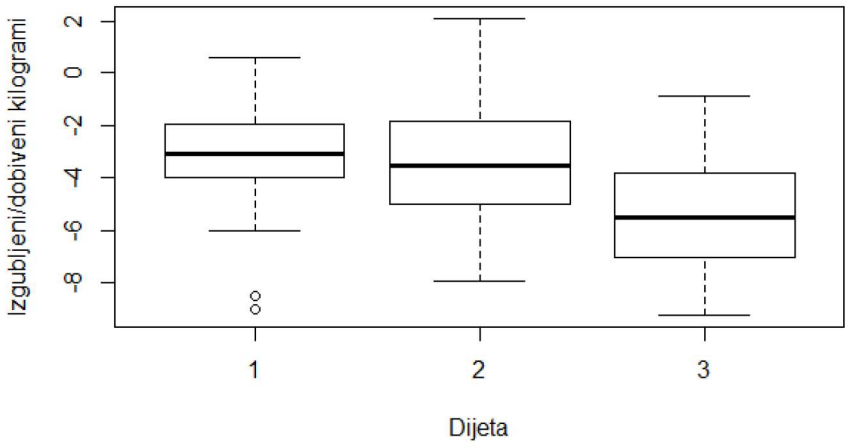
Varijabla	Opis
Osoba	Oznaka za svaku pojedinu osobu u eksperimentu (1-72)
Spol	Spol osobe (0-žensko, 1-muško)
Dob	Dob osobe u navršenim godinama
Visina	Visina osobe u centimetrima
Težina	Težina osobe u kilogramima
Dijeta	Oznaka dijete na kojoj je osoba bila (1, 2 i 3)
Težina_6tjedana	Težina osobe nakon 6 tjedana primjene dijete
Razlika	Varijabla dobivena kao razlika varijable Težina i Težina_6tjedana pri čemu negativan predznak označava da je osoba izgubila kilograme, dok pozitivan predznak znači da je osoba dobila kilograme u eksperimentu

Tablica 2.1: Varijable korištene u eksperimentu

Kao što smo već rekli u uvodu, analiza varijance u jednosmjernoj klasifikaciji podrazumijeva korištenje jedne nezavisne varijable pri izgradnji linearnog modela. U našem primjeru ćemo kao zavisnu varijablu koristiti varijablu Razlika, dok ćemo za nezavisnu varijablu uzeti varijablu Dijeta. S obzirom na ovako postavljen linearan model cilj ovog eksperimenta je utvrditi postoji li statistički značajna razlika u gubitku kilograma ovisno o odabranoj dijeti, odnosno daju li sve dijete iste rezultate ili ne. Iako analizom varijance ne možemo utvrditi koja od dijeta daje najbolje rezultate, pogledamo li sliku 2.1 koja prikazuje kutijaste dijagrame varijable Razlika obzirom na varijablu Dijeta možemo naslutiti kako dijeta 3 u prosijeku daje najbolje rezultate.

Osoba	Spol	Dob	Visina	Težina	Dijeta	Težina_6tjedana	Razlika
1	1	40	167	87	3	77,8	-9,2
2	1	43	162	80	1	71,0	-9,0
3	1	26	179	78	3	69,4	-8,6
4	0	28	176	69	1	60,5	-8,5
5	1	37	198	79	2	71,1	-7,9
6	0	36	160	66	3	58,2	-7,8
67	0	55	170	64	1	63,3	-0,7
68	0	51	174	63	2	62,4	-0,6
69	1	29	169	77	2	77,5	0,5
70	1	39	168	71	1	71,6	0,6
71	1	39	180	80	2	81,4	1,4
72	0	44	174	58	2	60,1	2,1

Tablica 2.2: Prvih i zadnjih šest redaka baze podataka



Slika 2.1: Kutijasti dijagram izgubljenih/dobivenih kilograma s obzirom na dijete 1, 2 i 3

Također možemo primijetit kako dijeta 1 i 2 daju podjednake rezultate te samim pogledom na podatke i kutijaste dijagrame ne možemo potvrditi postajanje razlike između ovih dijeta. Stoga ćemo testiranje postojanja razlika između ove tri dijete provesti kroz postupak analize varijance u jednosmjernoj klasifikaciji čiju teorijsku pozadinu dajemo u nastavku.

2.1.2 Analiza varijance za jednosmjernu klasifikaciju

Za početak pretpostavimo da je uzorak nezavisnih opažanja podijeljen u k grupa sa n_i, $i = 1, 2, \ldots, k$, opažaja u i-toj grupi pri čemu je $\sum_{i=1}^{k} n_i = n$. Ovu klasifikaciju možemo prikazati i u obliku sljedeće tablice:

$$\begin{array}{cccc} \mathrm{y}_{11} & \mathrm{y}_{12} & \cdots & \mathrm{y}_{1n_1} \\ \mathrm{y}_{21} & \mathrm{y}_{22} & \cdots & \mathrm{y}_{2n_2} \\ \vdots & \vdots & & \vdots \\ \mathrm{y}_{k1} & \mathrm{y}_{k2} & \cdots & \mathrm{y}_{kn_k} \end{array}$$

Tablica 2.3: Tablica klasifikacije

Ako se grupe razlikuju samo u njihovim očekivanjima te označimo li s y_{iq} q-ti opažaj u i-toj grupi možemo pisati

$$\mathrm{y}_{iq} = \theta_i + \epsilon_{iq}, \;\; i = 1, 2, \ldots, k; \;\; q = 1, 2, \ldots, n_i,$$

pri čemu su greške ϵ_{iq} nezavisne s očekivanjem 0 i varijancom σ^2 što zadovoljava formu linearnog modela (1.1), gdje je

$$\mathbf{y} = \begin{bmatrix} \mathrm{y}_{11} \\ \mathrm{y}_{12} \\ \vdots \\ \mathrm{y}_{1n_1} \\ \mathrm{y}_{21} \\ \vdots \\ \mathrm{y}_{2n_2} \\ \vdots \\ \mathrm{y}_{k1} \\ \vdots \\ \mathrm{y}_{kn_k} \end{bmatrix}, \;\; \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{bmatrix}, \;\; \mathbf{X} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \ldots & 0 \\ & & \ddots & \\ 0 & 0 & \ldots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\\vdots\\1\end{matrix}}\right\} n_1 \\ \left.\vphantom{\begin{matrix}1\\\vdots\\1\end{matrix}}\right\} n_2 \\ \vdots \\ \left.\vphantom{\begin{matrix}1\\\vdots\\1\end{matrix}}\right\} n_k \end{matrix}, \;\; \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \vdots \\ \epsilon_{k1} \\ \vdots \\ \epsilon_{kn_k} \end{bmatrix}.$$

Kako je analiza varijance modela zasnovana na ideji da ukupnu sumu kvadrata $\mathbf{y}'\mathbf{y}$ rastavimo na način kao u (2.4) potrebno je u okviru modela u jednosmjernoj klasifikaciji izračunati njene komponente. Za početak računamo komponente procjenitelja parametra $\boldsymbol{\theta}$ metodom najmanjih kvadrata, odnosno $\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. S obzirom da matrica dizajna $\mathbf{X}$ sadrži konstante slijedi da i matrica $\mathbf{X}'\mathbf{X}$ također sadrži konstante, odnosno ona je jednaka

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n_1 & & & 0 \\ & n_2 & & \\ & & \ddots & \\ 0 & & & n_k \end{bmatrix}.$$

Stoga koristeći svojstvo reziduala (1.8) zaključujemo da je analiza ortogonalna bez obzira na vrijednosti n_i, odnosno one ne moraju biti jednake. Nadalje računamo sljedeću komponentu procjenitelja $\hat{\boldsymbol{\theta}}$:

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum_{q=1}^{n_i} \mathrm{y}_{1q} \\ \sum_{q=1}^{n_i} \mathrm{y}_{2q} \\ \vdots \\ \sum_{q=1}^{n_i} \mathrm{y}_{kq} \end{bmatrix}.$$

Koristeći dobivene rezultate procjenitelja metodom najmanjih kvadrata za $\boldsymbol{\theta}$ možemo zapisati na sljedeći način:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} n_1 & & & 0 \\ & n_2 & & \\ & & \ddots & \\ 0 & & & n_k \end{bmatrix}^{-1} \begin{bmatrix} \sum_{q=1}^{n_i} \mathrm{y}_{1q} \\ \sum_{q=1}^{n_i} \mathrm{y}_{2q} \\ \vdots \\ \sum_{q=1}^{n_i} \mathrm{y}_{kq} \end{bmatrix} = \begin{bmatrix} \mathrm{y}_{1.} \\ \mathrm{y}_{2.} \\ \vdots \\ \mathrm{y}_{k.} \end{bmatrix},$$

pri čemu smo uveli novu oznaku $\mathrm{y}_{i.} = \sum_{q=1}^{n_i} \mathrm{y}_{iq}/n_i$.

Koristeći oznake koje smo uveli preostaje nam odrediti i drugu komponentu rastava ukupne sume kvadrata (2.4), odnosno $\mathbf{X}\hat{\boldsymbol{\theta}}$. Vrijedi:

$$\mathbf{X}\hat{\boldsymbol{\theta}} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mathrm{y}_{1.} \\ \mathrm{y}_{2.} \\ \vdots \\ \mathrm{y}_{k.} \end{bmatrix} = \begin{bmatrix} \mathrm{y}_{1.} \\ \vdots \\ \mathrm{y}_{1.} \\ \mathrm{y}_{2.} \\ \vdots \\ \mathrm{y}_{2.} \\ \vdots \\ \mathrm{y}_{k.} \\ \vdots \\ \mathrm{y}_{k.} \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} n_1 \\ \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} n_2 \\ \vdots \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} n_k \end{matrix}$$

te je prema tome suma kvadrata $\hat{\boldsymbol{\theta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\theta}}$ koju u nastavku označavamo sa S_1 dana sljedećim izrazom

$$S_1 = \hat{\boldsymbol{\theta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\theta}} = \sum_{i=1}^{k} n_i \mathrm{y}_{i.}^2.$$

Sumu kvadrata $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})$ nazivamo rezidualna suma kvadrata i označavamo ju sa S_R, a prema (2.4) ona iznosi:

$$\begin{aligned}
S_R &= \mathbf{y}'\mathbf{y} - S_1 \\
&= \sum_{i=1}^{k}\sum_{q=1}^{n_i} \mathrm{y}_{iq}^2 - \sum_{i=1}^{k} n_i \mathrm{y}_{i.}^2 \pm \sum_{i=1}^{k} n_i \mathrm{y}_{i.}^2 \\
&= \sum_{i=1}^{k}\sum_{q=1}^{n_i} \mathrm{y}_{iq}^2 - 2\sum_{i=1}^{k} n_i \mathrm{y}_{i.}^2 + \sum_{i=1}^{k} n_i \mathrm{y}_{i.}^2 \\
&= \sum_{i=1}^{k}\sum_{q=1}^{n_i} \mathrm{y}_{iq}^2 - 2\sum_{i=1}^{k}\sum_{q=1}^{n_i} \mathrm{y}_{iq}\mathrm{y}_{i.} + \sum_{i=1}^{k}\sum_{q=1}^{n_i} \mathrm{y}_{i.}^2 \\
&= \sum_{i=1}^{k}\sum_{q=1}^{n_i} (\mathrm{y}_{iq} - \mathrm{y}_{i.})^2
\end{aligned}$$

Suma kvadrata S_R predstavlja odstupanje opažaja od aritmetičke sredine podataka, odnosno varijabilnost unutar grupe (engl. within-group variation), a još se naziva i rezidualnom ili neobjašnjenom varijabilnošću.
Nadalje, kako je cilj analize varijance testirati postojanje razlika među grupama želimo testirati hipoteze s k ograničenja

$$H_0^1 : \boldsymbol{\theta} = 0,$$
$$H_1^1 : \boldsymbol{\theta} \neq 0$$

pri čemu koristimo F-statistiku (2.6). S obzirom na prethodno definirane sume kvadrata S_1 i S_R ona iznosi

$$F = \Big(\frac{n-k}{k}\Big)\frac{S_1}{S_R},$$

te ima necentralnu F distribuciju $\mathrm{F}(k, n-k, \lambda_1)$ s parametrom necentralnosti

$$\lambda_1 = \frac{\sum_{i=1}^{k} n_i \theta_i^2}{\sigma^2}$$

koja se reducira na centralnu $\mathrm{F}(k, n-k)$ u uvjetima istinitosti hipoteze H_0^1.

Kako nas često u primjeni više zanima jesu li svi parametri θ_i jednaki bez da određujemo njihovu pravu vrijednost, umjesto H_0^1 možemo testirati hipoteze s $(k-1)$ ograničenja:

$$\begin{aligned} &H_0^2 : \theta_1 - \theta_k = \theta_2 - \theta_k = \cdots = \theta_{k-1} - \theta_k = 0, \\ &H_1^2 : \theta_1 - \theta_k \neq \theta_2 - \theta_k \neq \ldots \neq \theta_{k-1} - \theta_k \neq 0, \text{ za barem jedan k.} \end{aligned} \tag{2.7}$$

U uvjetima istinitosti H_0^2 su svih n opažaja jednako distribuirani te uvodimo oznaku

$$\theta_* = \sum_{i=1}^{k} n_i\theta_i/n = \begin{bmatrix} n_1/n \\ n_2/n \\ \vdots \\ n_k/n \end{bmatrix}' \boldsymbol{\theta},$$

odakle slijedi da je procjenitelj metodom najmanjih kvadrata za θ_* u ovom slučaju

$$\hat{\theta}_* = \mathrm{y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{k}\sum_{q=1}^{n_i} \mathrm{y}_{iq} = \frac{1}{n}\sum_{i=1}^{k} n_i \mathrm{y}_{i\cdot}.$$

Neka je sada $\mathbf{1}$ vektor jedinica dimenzije $(n \times 1)$. Linearni model (1.1) možemo zapisati koristeći navedeni vektor jedinica te θ_* na sljedeći način:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} + \mathbf{1}\theta_* - \mathbf{1}\theta_* = \mathbf{1}\theta_* + \left[\mathbf{X} - \mathbf{1}\begin{bmatrix} n_1/n \\ n_2/n \\ \vdots \\ n_k/n \end{bmatrix}'\right]\boldsymbol{\theta} + \boldsymbol{\epsilon}.$$

Primjećujemo da vrijednost $\hat{\theta}_*$ nije uključena u hipotezu H_0^2 te sumu kvadrata $(\mathbf{1}\hat{\theta}_*)'(\mathbf{1}\hat{\theta}_*) = n\mathrm{y}_{\cdot\cdot}^2$, moramo oduzeti od sume kvadrata S_1 kako bismo dobili sumu kvadrata s obzirom na ostalih $(k-1)$ ograničenja koju označavamo s S_2:

$$S_2 = S_1 - n\mathrm{y}_{\cdot\cdot}^2 \equiv \sum_{i=1}^{k} n_i(\mathrm{y}_{i\cdot} - \mathrm{y}_{\cdot\cdot})^2$$

te ona predstavlja sumu kvadrata među grupama, odnosno varijabilnost među grupama (engl. between-group variation).
Test statistika za testiranje hipoteza (2.7) s obzirom na definirane test statistike ima sljedeći oblik

$$F = \left(\frac{n-k}{k-1}\right)\frac{S_2}{S_R},$$

s ne centralnom F distribucijom $\mathrm{F}(k-1, n-k, \lambda_2)$, gdje je parametar necentralnosti jednak

$$\lambda_2 = \frac{\sum_{i=1}^{k} n_i(\theta_i - \theta_*)^2}{\sigma^2},$$

dok u uvjetima istinitosti hipoteze H_0^2 test statistika F ima centralnu distribucija $\mathrm{F}(k-1, n-k)$. Radi jednostavnijeg izračuna sume kvadrata S_2 i S_R možemo zapisati i na sljedeći način

$$S_2 = \sum_{i=1}^{k} \frac{\left(\sum_{q=1}^{n_i} \mathrm{y}_{iq}\right)^2}{n_i} - \frac{\left(\sum_{i=1}^{k}\sum_{q=1}^{n_i} \mathrm{y}_{iq}\right)^2}{n},$$

$$S_R = \sum_{i=1}^{k}\sum_{q=1}^{n_i} \mathrm{y}_{iq}^2 - \sum_{i=}^{k} \frac{\left(\sum_{q=1}^{n_i} \mathrm{y}_{iq}\right)^2}{n_i}.$$

Ove rezultate vrlo pregledno zapisujemo i u obliku ANOVA tablice pri čemu srednje kvadratno odstupanje označavamo s MS i računamo kao

$$MS = \frac{\text{Suma kvadrata}}{\text{Stupanj slobode}}.$$

Izvor varijabilnosti	Stupanj slobode	Suma kvadrata	Srednje kvadratno odstupanje
Među grupama	$k-1$	S_2	MS_2
Unutar grupe	$n-k$	S_R	MS_R
Ukupno	$n-1$	S_T	MS_T

Tablica 2.4: ANOVA tablica za jednosmjernu klasifikaciju

2.1.3 Primjer

Nakon što smo definirali sve potrebno za analizu varijance u jednosmjernoj klasifikaciji vraćamo se na primjer 2.1.1 s početka poglavlja. Kao što smo na početku naveli, cilj je testirati ovise li izgubljeni/dobiveni kilogrami o odabranoj dijeti. Za početak, s obzirom da imamo tri različite vrste dijeta u ovom eksperimentu, možemo zaključiti da je broj grupa $k = 3$, a s obzirom da imamo 72 osobe koje su sudjelovale u eksperimentu broj podataka je $n = 72$. Kako imamo velik broj podataka za računanje potrebnih suma kvadrata te srednje kvadratnih grešaka korištena je funkcija *aov*() u programskom jeziku R. Dobiveni rezultati pri analizi varijance prikazani su u tablici 2.5.

Izvor varijabilnosti	Stupanj slobode	Suma kvadrata	Srednje kvadratno odstupanje
Među grupama	2	65.2	32.61
Unutar grupe	69	379.1	5.49
Ukupno	71	444.3	38.1

Tablica 2.5: ANOVA tablica za jednosmjernu klasifikaciju za Primjer 2.1.1

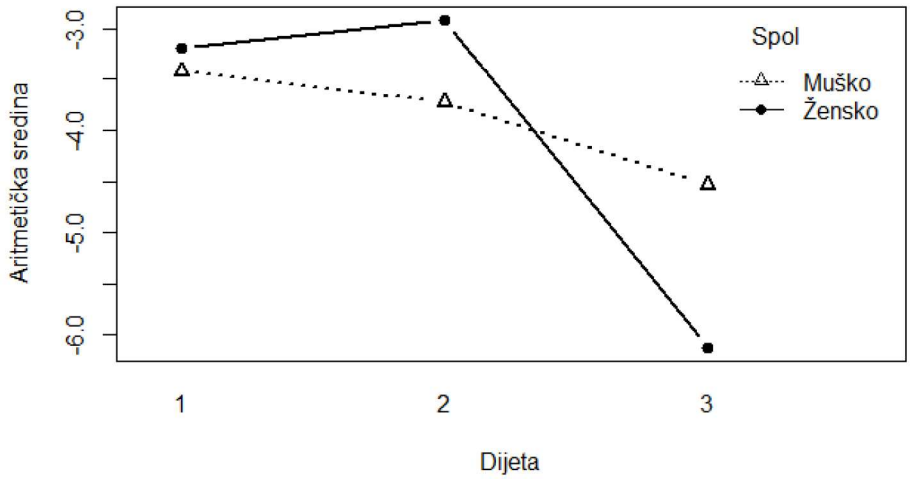
Korištenjem iste funkcije dobili smo i vrijednost F-statistike $F = 5.935$ te P-vrijednost $P = 0.004180$. Također je uz nivo značajnosti $\alpha = 0.05$ kritična vrijednost F testa jednaka $F_{\alpha,2,69} = 3.1296$ što je manje od dobivene vrijednosti F-statistike. S obzirom na sve dobivene vrijednosti zaključujemo da odbacujemo nul hipotezu H_0 o jednakosti parametara i prihvaćamo alternativnu hipotezu odnosno možemo zaključiti kako postoji razlika između tri dijete u gubitku kilograma. Dakle, zaključujemo da primjena različitih dijeta daje i različite rezultate. Bitno je naglasiti da samom analizom varijance ne možemo potvrditi koja od dijeta daje najbolje rezultate te da je za takve zaključke potrebno provesti post-hoc testove odnosno procedure višestrukog uspoređivanja.

2.2 Dvosmjerna klasifikacija

2.2.1 Uvod i motivacija

Dvosmjerna klasifikacija za analizu varijance je jedna vrsta proširenja jednosmjerne klasifikacije koja proučava utjecaj dvije različite nezavisne kategorijalne varijable na jednu kontinuiranu zavisnu varijablu. Ovdje će nas zanimati i interakcije između dvije nezavisne varijable.
Slično kao i u prethodnom poglavlju analizu varijance za dvosmjernu klasifikaciju započinjemo s primjerom kako bismo jasnije shvatili problematiku, a koristimo se istom bazom podataka kao u primjeru 2.1.1.

Primjer 2.2.1. *Za razliku od jednosmjerne klasifikacije gdje smo pri izgradnji linearnog modela uz zavisnu varijablu Razlika koristili jednu nezavisnu varijablu (Dijeta) kod dvosmjerne klasifikacije koristimo dvije nezavisne varijable koje će u našem primjeru biti Dijeta i Spol. Cilj je testirati postoji li razlika u gubitku kilograma s obzirom na spol i odabranu vrstu dijete. Za početak pogledajmo sliku 2.2 koja grafički prikazuje interakciju između varijabli Dijeta i Spol.*



Slika 2.2

Na slici 2.2 oznakom kružića prikazane su aritmetičke sredine varijable Razlika za žene za svaku od tri dijete dok su oznakom trokutića prikazane aritmetičke sredine varijable Razlika za muškarce za svaku od tri dijete. Spajanjem ovih oznaka linijama dobili smo interakcijski graf. Ukoliko su linije interakcijskog grafa paralelne one ukazuju na nepostojanje interakcije. Kako linije na našem grafu nisu paralelne nego se sijeku možemo zaključiti da postoji interakcija između spola i odabrane dijete pri gubitku kilograma što ćemo u nastavku pokušati potvrditi kroz postupak analize varijance u dvosmjernoj klasifikaciji.

2.2.2 Analiza varijance za dvosmjernu klasifikaciju

Pretpostavimo da sada, za razliku od prethodnog poglavlja gdje smo uzorak klasificirali u k grupa, slučajan uzorak od n opažaja klasificiramo na način prikazan u tablici 2.6 te frekvencijama prikazanim u tablici 2.7.

	B_1	B_2	$\dots$	B_c
A_1	$y_{111}, y_{112}, \dots, y_{11n_{11}}$	$y_{121}, y_{122}, \dots, y_{12n_{12}}$	$\dots$	$y_{1c1}, y_{1c2}, \dots, y_{1cn_{1c}}$
A_2	$y_{211}, y_{212}, \dots, y_{21n_{21}}$	$y_{221}, y_{222}, \dots, y_{22n_{22}}$	$\dots$	$y_{2c1}, y_{2c2}, \dots, y_{2cn_{2c}}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
A_r	$y_{r11}, y_{r12}, \dots, y_{r1n_{r1}}$	$y_{r21}, y_{r22}, \dots, y_{r2n_{r2}}$	$\dots$	$y_{rc1}, y_{rc2}, \dots, y_{rcn_{rc}}$

Tablica 2.6: Tablica dvosmjerne klasifikacije uzorka

n_{11}	n_{12}	$\dots$	n_{1c}	$n_{1.}$
n_{21}	n_{22}	$\dots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
n_{r1}	n_{r2}	$\dots$	n_{rc}	$n_{r.}$
$n_{.1}$	$n_{.2}$	$\dots$	$n_{.c}$	n

Tablica 2.7: Tablica frekvencija

Napomenimo da točka koja zamjenjuje indeks u n označava sumaciju po tom indeksu, odnosno vrijedi

$$\sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{c} n_{.j} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = n.$$

U ovom radu promatrat ćemo slučaj proporcionalnih frekvencija odnosno pretpostavljamo da vrijedi da je

$$n_{ij} = \frac{n_{i.}n_{.j}}{n}.$$

Možemo primijetiti da Tablica 2.7 ima oblik tablice kontingencije, ali je ovdje važno da je vrijednost y poznata za svaki opažaj dok su u tablici kontingencije poznate samo frekvencije odabranih kategorija.
Cilj analize varijance u dvosmjernoj klasifikaciji je odgovoriti na sljedeća tri bitna pitanja koja si postavljamo tijekom analize:

1. Postoje li razlike u klasifikacijama po redcima $(A_1, A_2, \dots, A_r)$?
2. Postoje li razlike u klasifikacijama po stupcima $(B_1, B_2, \dots, B_c)$?
3. Postoji li interakcijski efekt između stupaca i redaka?

Odgovor na ova pitanja leži u testiranju hipoteza koje su ključne za analizu varijance te ćemo ih postaviti u nastavku. Za početak uvodimo nove oznake kako bismo si olakšali proces analize te definirali model za ovaj tip klasifikacije. Neka y_{ijp} označava p-ti opažaj u i-tom redu i j-tom stupcu tablice 2.6, sukladno tome uvodimo i sljedeće oznake:

$$\begin{aligned}
\mathrm{y}_{ij.} &= \sum_{p=1}^{n_{ij}} \mathrm{y}_{ijp}, & \overline{\mathrm{y}}_{ij.} &= \frac{\mathrm{y}_{ij.}}{n_{ij}}, \\
\mathrm{y}_{i..} &= \sum_{j=1}^{c} \mathrm{y}_{ij.}, & \overline{\mathrm{y}}_{i..} &= \frac{\mathrm{y}_{i..}}{n_{i.}}, \\
\mathrm{y}_{.j.} &= \sum_{i=1}^{r} \mathrm{y}_{ij.}, & \overline{\mathrm{y}}_{.j.} &= \frac{\mathrm{y}_{.j.}}{n_{.j}}, \\
\mathrm{y}_{...} &= \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{p=1}^{n_{ij}} \mathrm{y}_{ijp}, & \overline{\mathrm{y}}_{...} &= \frac{\mathrm{y}_{...}}{n}.
\end{aligned} \tag{2.8}$$

Nadalje, označimo s μ slobodan član, α_i efekt obzirom na i-ti redak, β_j efekt obzirom na j-ti stupac, $(\alpha\beta)_{ij}$ interakcijski efekt između α_i i β_j te ϵ_{ijp} grešku modela povezanu s p-tim opažajem u i-tom retku i j-tom stupcu. Koristeći navedene oznake možemo definirati model za dvosmjernu klasifikaciju s interakcijom koji je dan sljedećim izrazom:

$$\begin{gathered}
\mathrm{y}_{ijp} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijp}, \\
\text{za } i = 1, ..., r; j = 1, ..., c; p = 1, ..., n_{ij}.
\end{gathered} \tag{2.9}$$

Slično kao kod jednosmjerne klasifikacije model (2.9) želimo zapisati u obliku osnovnog linearnog modela (1.1). U svrhu ilustracije pretpostavimo da su parametri $r = 2$ i $c = 2$, matrični zapis modela (2.9) je sljedeći:

$$\mathbf{y} = \begin{bmatrix}
1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
\mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \\ (\alpha\beta)_{21} \\ (\alpha\beta)_{22}
\end{bmatrix} + \begin{bmatrix}
\epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{221} \\ \epsilon_{222}
\end{bmatrix}.$$

Napomenimo još da interakcijski efekt $(\alpha\beta)_{ij}$ predstavlja odstupanje očekivanja u (i, j) ćeliji tablice (2.6), koje označavamo s μ_{ij}, od sume prva tri člana u (2.9), dok za konstante α_i, β_j i $(\alpha\beta)_{ij}$ vrijede sljedeći uvjeti:

$$\sum_{i=1}^{r} \alpha_i = \sum_{j=1}^{c} \beta_j = \sum_{i=1}^{r} (\alpha\beta)_{ij} = \sum_{j=1}^{c} (\alpha\beta)_{ij} = 0.$$

Prvi korak ka analizi varijance je kao i u prethodnom poglavlju rastaviti ukupnu sumu kvadrata, odnosno ukupnu varijabilnost modela (2.9) na ne negativne komponente. Uzmemo li u obzir oznake definirane pod (2.8) odstupanje podatka y_{ijp} od aritmetičke sredine podataka možemo zapisati na sljedeći način

$$y_{ijp} - \overline{y}_{...} = (\overline{y}_{i..} - \overline{y}_{...}) + (\overline{y}_{.j.} - \overline{y}_{...}) + (\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...}) + (y_{ijp} - \overline{y}_{ij.}).$$

Ovakvim rastavljanjem gornjeg izraza te kvadriranjem i sumacijom po odgovarajućim indeksima dobivamo sljedeći izraz za ukupnu sumu kvadrata modela (2.9):

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{p=1}^{n_{ij}}(y_{ijp} - \overline{y}_{...})^2 = \sum_{i=1}^{r} n_{i.}(\overline{y}_{i..} - \overline{y}_{...})^2 + \sum_{j=1}^{c} n_{.j}(\overline{y}_{.j.} - \overline{y}_{...})^2 + \\ + \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{p=1}^{n_{ij}}(y_{ijp} - \overline{y}_{ij.})^2.$$

Na ovaj smo način dobili četiri nove sume kvadrata i time olakšali analizu, odnosno ukupnu varijabilnost modela smo rastavili na sljedeći način

$$S_T = S_A + S_B + S_{AB} + S_E,$$

pri čemu vrijedi:

$$\begin{aligned}
S_T &= \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{p=1}^{n_{ij}}(y_{ijp} - \overline{y}_{...})^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{p=1}^{n_{ij}} y_{ijp}^2 - \frac{y_{...}^2}{n} \\
S_A &= \sum_{i=1}^{r} n_{i.}(\overline{y}_{i..} - \overline{y}_{...})^2 = \sum_{i=1}^{r} \frac{y_{i..}^2}{n_{i.}} - \frac{y_{...}^2}{n} \\
S_B &= \sum_{j=1}^{c} n_{.j}(\overline{y}_{.j.} - \overline{y}_{...})^2 = \sum_{j=1}^{c} \frac{y_{.j.}^2}{n_{.j}} - \frac{y_{...}^2}{n} \\
S_{AB} &= \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{y_{ij.}^2}{n_{ij}} - \sum_{i=1}^{r} \frac{y_{i..}^2}{n_{i.}} - \sum_{j=1}^{c} \frac{y_{.j.}^2}{n_{.j}} + \frac{y_{...}^2}{n} \\
S_E &= \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{p=1}^{n_{ij}}(y_{ijp} - \overline{y}_{ij.})^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{p=1}^{n_{ij}} y_{ijp}^2 - \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{y_{ij.}^2}{n_{ij}}
\end{aligned}$$

Suma kvadrata S_A označava varijabilnost podataka između redaka, suma kvadrata S_B označava varijabilnost podataka između stupaca dok S_{AB} označava interakcijski zbroj kvadrata. Suma kvadrata S_E označava sumu kvadrata pogrešaka odnosno sumu kvadrata odstupanja podataka od odgovarajućih sredina. Sada je potrebno odrediti i stupnjeve slobode za odgovarajuće sume kvadrata. Za sume kvadrata S_A i S_B vrijede sljedeće restrikcije:

$$\sum_{i=1}^{r}(\overline{\mathrm{y}}_{i..} - \overline{\mathrm{y}}_{...}) = \sum_{i=1}^{c}(\overline{\mathrm{y}}_{.j.} - \overline{\mathrm{y}}_{...}) = 0,$$

te su stoga stupnjevi slobode za njih redom $r-1$ i $c-1$. Za sumu kvadrata S_{AB} radi jednostavnosti uvodimo oznaku

$$\theta_{ij} = (\overline{\mathrm{y}}_{ij.} - \overline{\mathrm{y}}_{i..} - \overline{\mathrm{y}}_{.j.} + \overline{\mathrm{y}}_{...})$$

pri čemu vrijede sljedeće restrikcije

$$\begin{aligned} \sum_{i=1}^{r}\theta_{ij} &= 0, \ \forall j, \\ \sum_{j=1}^{c}\theta_{ij} &= 0, \ \forall i. \end{aligned} \tag{2.10}$$

Međutim, postoji samo $r+c-1$ nezavisnih restrikcija na θ_{ij} jer sumacijom restrikcija (2.10) dobivamo

$$\begin{aligned} \sum_{j=1}^{c}\left(\sum_{i=1}^{r}\theta_{ij}\right) &= 0, \\ \sum_{i=1}^{r}\left(\sum_{j=1}^{c}\theta_{ij}\right) &= 0. \end{aligned}$$

Stoga će samo $r-1$ restrikcija od njih r iz (2.10) biti nezavisne i ukupan broj nezavisnih restrikcija na θ_{ij} je $r+c-1$. Kako varijabli θ_{ij} ukupno imamo rc, stupanj slobode za sumu kvadrata S_{AB} je $rc-(r+c-1) = (r-1)(c-1)$. Uz ove rezultate je stupanj slobode za sumu kvadrata S_E jednak $(n-1)-(r-1)-(c-1)-(r-1)(c-1) = n-rc$.
Dobivene rezultate ćemo, kao i u prethodnom poglavlju, zapisati u obliku ANOVA tablice za dvosmjernu klasifikaciju koja je prikazana u tablici 2.8.

Izvor varijabilnosti	Stupanj slobode	Suma kvadrata	Srednje kvadratno odstupanje
Između redaka	$r-1$	S_A	MS_A
Između stupaca	$c-1$	S_B	MS_B
Interakcije	$(r-1)(c-1)$	S_{AB}	MS_{AB}
Greška	$n-rc$	S_E	MS_E
Ukupno	$n-1$	S_T	

Tablica 2.8: ANOVA tablica za dvosmjernu klasifikaciju

Kako bismo odgovorili na pitanja postavljena na početku poglavlja postavljamo odgovarajuće hipoteze te za njih kreiramo test statistike. Želimo li testirati postoji li interakcijski efekt između stupaca i redaka postavljamo sljedeće hipoteze

$$H_0^{AB} : (\alpha\beta)_{ij} = 0 \text{ za sve } i, j,$$
$$H_1^{AB} : (\alpha\beta)_{ij} \neq 0 \text{ za barem jedan } i, j.$$

Test statistika koju koristimo za testiranje ovih hipoteza je

$$F_{AB} = \frac{MS_{AB}}{MS_E},$$

Potrebno je odrediti distribucije za MS_{AB} i MS_E kako bismo mogli odrediti distribuciju test statistike. Prema definiciji centralne i ne centralne χ^2 distribucije lako je zaključiti da vrijedi

$$\frac{MS_E}{\sigma_e^2} \sim \frac{\chi^2(n-rc)}{n-rc}$$

$$\frac{MS_{AB}}{\sigma_e^2} \sim \frac{\chi^2((r-1)(c-1), \lambda_{AB})}{(r-1)(c-1)},$$

gdje je

$$\lambda_{AB} = \frac{1}{2\sigma_e^2} \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} (\alpha\beta)_{ij}^2.$$

U uvjetima istinitosti hipoteze H_0^{AB}, test statistika F_{AB} ima F distribuciju $\mathrm{F}((r-1)(c-1), n-rc)$, pri čemu hipotezu H_0^{AB}odbacujemo za velike vrijednosti statistike F_{AB}.
Nadalje, postojanje razlike u klasifikacijama obzirom na retke testiramo postavljanjem hipoteza

$$H_0^A : \alpha_i = 0 \text{ za sve } i,$$
$$H_1^A : \alpha_i \neq 0 \text{ za barem jedan } i,$$

te koristimo test statistiku

$$F_A = \frac{MS_A}{MS_E}$$

pri čemu je

$$\frac{MS_A}{\sigma_e^2} \sim \frac{\chi^2(r-1, \lambda_A)}{r-1},$$

dok je parametar necentralnosti λ_A jednak

$$\lambda_A = \frac{1}{2\sigma_e^2} \sum_{i=1}^{r} n_{i\cdot} \alpha_i^2.$$

Test statistika F_A u uvjetima istinitosti hipoteze H_0^A ima F distribuciju $\mathrm{F}(r-1, n-rc)$. Slično prethodnom, postojanje razlike u klasifikacijama obzirom na stupce testiramo hipotezama

$$H_0^B : \beta_j = 0 \text{ za sve } j,$$
$$H_1^B : \beta_j \neq 0 \text{ za barem jedan } j,$$

pri čemu je test statistika

$$F_B = \frac{MS_B}{MS_E}.$$

Na sličan način određujemo i distribuciju za MS_B te parametar necentralnosti λ_B

$$\frac{MS_B}{\sigma_e^2} \sim \frac{\chi^2(c-1, \lambda_B)}{c-1},$$

$$\lambda_B = \frac{1}{2\sigma_e^2} \sum_{j=1}^{c} n_{\cdot j} \beta_j^2.$$

U uvjetima istinitosti hipoteze H_0^B test statistika F_B također ima F distribuciju $\mathrm{F}(c-1, n-rc)$.

2.2.3 Primjer

Vratimo se sada na primjer 2.2.1. Ranije smo zaključili da je broj podataka $n = 72$, a s obzirom da promatramo tri različite dijete vrijedi $r = 3$. Nadalje, osobe smo klasificirali i po spolu te je stoga $c = 2$. Korištenjem programskog jezika R dobivamo podatke koje prikazujemo u sljedećoj tablici.

Izvor varijabilnosti	Stupanj slobode	Suma kvadrata	Srednje kvadratno odstupanje	F-statistika	P-vrijednost
Dijeta	2	65.2	32.61	5.984	0.00409
Spol	1	0.6	0.61	0.111	0.74003
Interakcija	2	18.8	9.42	1.729	0.18539
Greška	66	359.6	5.45		
Ukupno	71	444.2	48.09		

Tablica 2.9: ANOVA tablica za dvosmjernu klasifikaciju

Hipoteze	Test statistika	Distribucija u uvjetima H_0
$H_0^{AB} : (\alpha\beta)_{ij} = 0$ za sve i, j $H_1^{AB} : (\alpha\beta)_{ij} \neq 0$ za barem jedan i, j	$F_{AB} = \frac{MS_{AB}}{MS_E}$	$\mathrm{F}((r-1)(c-1), n-rc)$
$H_0^{A} : (\alpha)_{i} = 0$ za sve i $H_1^{A} : (\alpha)_{i} \neq 0$ za barem jedan i	$F_{A} = \frac{MS_{A}}{MS_E}$	$\mathrm{F}(r-1, n-rc)$
$H_0^{B} : (\beta)_{j} = 0$ za sve j $H_1^{B} : (\beta)_{j} \neq 0$ za barem jedan j	$F_{B} = \frac{MS_{B}}{MS_E}$	$\mathrm{F}(c-1, n-rc)$

Za početak želimo testirati postojanje interakcije između varijabli Dijeta i Spol. Za nivo značajnosti uzimamo $\alpha = 0.05$. Kako je p-vrijednost interakcije veća od nivoa značajnosti α te je kritična vrijednost F testa $F_{\alpha,2,66} = 3.135918$ veća od vrijednosti F-statistike za interakciju zaključujemo da ne odbacujemo nul-hipotezu odnosno nemamo razloga sumnjati u postojanje statistički značajne interakcije između efekta dijete i spola na gubitak kilograma. Slično kao i kod jednosmjerne klasifikacije iz tablice 2.9 te kritične vrijednosti F testa $F_{\alpha,2,66} = 3.135918$, koja je manja od vrijednosti F-statistike zaključujemo da odbacujemo nul hipotezu o nepostojanju efekta, odnosno možemo zaključiti da efekt dijete postoji što znači da različite dijete daju i različite rezultate u gubitku kilograma. Za kraj ostaje nam testirati postojanje efekta varijable Spol na gubitak kilograma. Iz tablice 2.9 vidimo da je p-vrijednost za varijablu Spol veća od nivoa značajnosti α, te je kritična vrijednost F testa $F_{\alpha,1,66} = 3.986269$ što je veće od vrijednosti F-statistike za varijablu Spol. Sve navedene vrijednosti upućuju nas da nemamo razloga sumnjati u istinitost nul-hipoteze, odnosno u nepostojanje efekta. Tako zaključujemo da nemamo razloga sumnjati da postoji razlika između gubitka kilograma kod žena i muškaraca.

Literatura

[1] M. Benšić, N. Šuvak, Uvod u vjerojatnost i statistiku, Sveučilište J. J. Strossmayera, Odjel za matematiku, Osijek, 2014.

[2] R. Christensen, Analysis Of Variance, Design and Regression, University of New Mexico, Albuquerque, USA, CRC Press, 2016.

[3] B. E. Hansen, Econometrics, University of Wisconsin
`https://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf`

[4] B. E. Hansen, Statistical Theory For Economists, University of Wisconsin
`https://www.ssc.wisc.edu/~bhansen/probability/Probability.pdf`

[5] M. G. Kendall, A. F. Stuart, The Advanced Theory Of Statistics, Volume 1: Distribution Theory, Charles Griffin & Company Limited, London, 1945.

[6] M. G. Kendall, A. F. Stuart, The Advanced Theory Of Statistics, Volume 2: Inference and Relationship, Charles Griffin & Company Limited, London, 1961.

[7] M. G. Kendall, A. F. Stuart, The Advanced Theory Of Statistics, Volume 3: Design and Analysis, and Time-Series, Charles Griffin & Company Limited, London, 1966.

[8] H. Sahai, M. I. Ageel, The Analysis of Variance, Fixed, Random and Mixed Models, Springer Science+Buisiness Media, 2000.

[9] H. Sahai, M. M. Ojeda, Analysis of Variance for Random Models, Volume 1: Balanced Data, Springer Science+Buisiness Media, New York, 2004.

[10] H. Scheffe, The Analysis Of Variance, John Wiley & Sons, Inc., New York, 1959.

Sažetak

U ovom diplomskom radu upoznali smo se s postupkom analize varijance korištenjem linearnog regresijskog modela. U uvodnom dijelu definirali smo osnovne pojmove potrebne u nastavku rada, a nakon toga dajemo teorijski uvod u analizu varijance. Definirajući i rastavljajući ukupnu varijabilnost modela na komponente dobivamo jednostavnu formulu za analizu varijance koju ćemo koristiti u nastavku. Postupak analize varijance korištenjem linearnog regresijskog modela objasnili smo u okviru jednosmjerne i dvosmjerne klasifikacije podataka. Sve dobivene sume kvadrata i F-statistike prikazali smo u obliku ANOVA tablice. Također smo na primjeru prikazali postupak izrade ANOVA tablice te interpretaciju njenih koeficijenata.

Ključne riječi: analiza varijance, F test, greška modela, linearni model, procjenitelj metodom najmanjih kvadrata , reziduali, suma kvadrata, test-statistika, jednosmjerna klasifikacija, dvosmjerna klasifikacija, kvadratna forma, srednje kvadratno odstupanje

Summary

In this thesis, we introduced the procedure of analysis of variance using a linear regression model. In the first chapter we defined the basic concepts that we will need for the analysis of variance. In the next, main chapter of this thesis, we give a theoretical introduction to the process of variance analysis itself. By defining and breaking down the total variability of the model into components, we obtain a simple formula for analysis of variance. The procedure of analysis of variance using a linear regression model was explained within the one-way and two-way data classification. All obtained sums of squares and F-statistics are presented in the form of ANOVA table. We also presented the process of making the ANOVA table and the interpretation of its coefficients as an example.

Key words: analysis of variance, linear model, least squares estimator, sum of squares, error, test-statistics, F test, residuals, one-way classification, two-way classification, quadratic form, mean of squares

Životopis

Rođena sam 12.8.1994. godine u Osijeku. Nakon završene Osnovne škole Milka Cepelića u Vuki upisujem I. gimnaziju u Osijeku. Gimnaziju završavam 2013. godine te iste godine upisujem preddiplomski studij matematike na Odjelu za matematiku u Osijeku. Preddiplomski studij završavam 2017. godine sa završnim radom na temu Transformacije diskretnih i neprekidnih slučajnih varijabli te time stječem akademski stupanj prvostupnice matematike. Iste godine upisujem diplomski studij na Odjelu za matematiku u Osijeku, smjer Financijska matematika i statistika. U rujnu 2019. godine odrađujem stručnu praksu u tvrtki Osijek - Koteks d.d.