Peatükk 1

Sissejuhatus

1.1 Lineaarse mudeli definitsioon

Lineaarne mudel on mudeli parameetrite lineaarne funktsioon. Tüüpiline lineaarne mudel näeb välja järgmine:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

kus $\mathbf{y}$ on $n \times 1$ vektor, mis sisaldab n vaadeldavat (või mõõdetavat) väärtust, $\mathbf{X}$ on $n \times p$ maatriks, mille elementideks olevaid konstante me teame. Maatriksit $\mathbf{X}$ tuntakse sageli ka mudeli- või disainimaatriksi nime all; $\boldsymbol{\beta}$ on $p \times 1$ hindamist vajavate (tundmatute) parameetrite vektor; $\boldsymbol{\varepsilon}$ on $n \times 1$ juhuslike vigade vektor (sageli räägitakse temast ka kui mudeli jääkide või prognoosivigade vektorist). Nii $\mathbf{y}$ kui $\boldsymbol{\varepsilon}$ on juhuslikud vektorid. Eeldades, et tehtavad mõõtmistulemused pole süstemaatiliselt valed ehk nihkega ($\mathrm{E}\boldsymbol{\varepsilon} = 0$ või, kui $\mathbf{X}$ on juhuslik, $\mathrm{E}(\boldsymbol{\varepsilon}|\mathbf{X}) = 0$) saame, et

$$\mathrm{E}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}. \tag{1.2}$$

Lisaks eeldame, et mudeli jäägid — juhuslikud vead $\boldsymbol{\varepsilon}$ — pole korreleeritud,

$$(\mathrm{D}(\mathbf{y}|\mathbf{X}) =) \quad \mathrm{D}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}. \tag{1.3}$$

Mõningate tulemuste tarvis läheb tarvis veel eelduseid juhuslike suuruste jaotuste kohta. Sellisel juhul eeldame, et $\boldsymbol{\varepsilon}$ on n-mõõtmelise normaaljaotusega juhuslik vektor, $\boldsymbol{\varepsilon} \sim N(0; \sigma^2\mathbf{I})$ ja sõltumatu mudeli maatriksist ($\boldsymbol{\varepsilon} \perp \mathbf{X}$). Selliste eelduste kehtides peab aga ka $\mathbf{y}$ tinglik jaotus (tingimusel, et $\mathbf{X}$ on antud) olema normaaljaotus:

$$\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}; \sigma^2\mathbf{I}). \tag{1.4}$$

Tihti vaadeldakse praktikas ka situatsioone, kus $\mathbf{X}$ on uurija poolt fikseeritavad/määratavad katsetingimused, seega pole maatriksi $\mathbf{X}$ elemendid juhuslikud. Sellisel juhul taandub jaotuse eeldus 1.4 nõudeks, et uuritav tunnus $\mathbf{y}$ oleks normaaljaotusega:

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}; \sigma^2\mathbf{I}). \tag{1.5}$$

Näide 1.1 *Lihtne lineaarne regressioon.*
Vaata näiteks mudelit

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, 6,$$

kus $(x_1, x_2, \ldots, x_6) = (1, 2, 3, 4, 5, 6)$ *ja mudeli vead* ε_i *on sõltumatud juhuslikud suurused jaotusega* $\varepsilon_i \sim N(0; \sigma^2)$. *Maatrikskujul kirjapandult näeks antud regressioonimudel välja järgmine:*

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix}$$

$$\mathbf{y} \quad = \quad \mathbf{X} \quad \boldsymbol{\beta} \quad + \quad \boldsymbol{\varepsilon}.$$

Näide 1.2 *Dispersioonanalüüs.*
Ühe faktoriga dispersioonanalüüsi mudeli

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

kus $i = 1 \ldots 3,\ j = 1 \ldots n_i,\ (n_1, n_2, n_3) = (3{,}1{,}2)$, *saab maatrikskujul kirja panna järgmiselt:*

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix}$$

$$\mathbf{y} \quad = \quad \mathbf{X} \quad \boldsymbol{\beta} \quad + \quad \boldsymbol{\varepsilon}.$$

1.2 Lineaarse mudeli eeldustest

Lineaarse mudeli eeldustel ja sellel, kuidas tehtud eelduste paikapidavust kontrollida (ning mida teha, kui eeldused paika ei pea) peatume tõsisemalt edaspidi. Siiski tahaks juba siin tähelepanu juhtida paarile asjaolule, mis ehk aitaksid paremini mõista lineaarse mudeli definitsiooni ennast.

Esindav valim

Lineaarset mudelit kasutatakse sageli prognoosimiseks — kas tulevaste vaatluste prognoosimiseks või mõõtmata jäänud vaatluste prognoosimiseks (seose kirjeldamine populatsioonis). Prognoosimisel tekib aga lisaeeldus, mis nõuab, et ka tulevaste vaatluste jaoks peab kehtima sama mudel, ehk teisisõnu:

$$\mathbf{y}_{uus} = \mathbf{X}_{uus}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{uus},$$

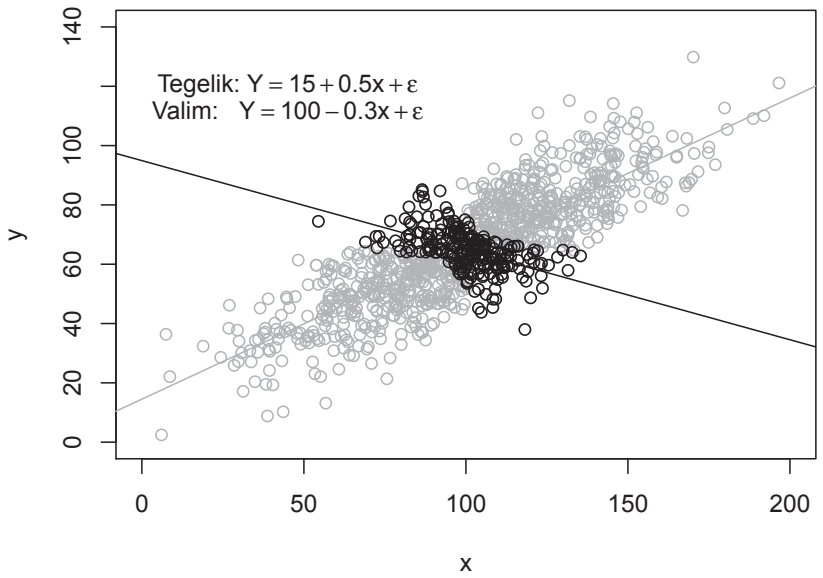
kus tundmatute parameetrite vektor $\boldsymbol{\beta}$ on seesama mis vaatlusandmeid genereerinud mudelis 1.1. Kui vaatluste genereerimisel on kasutatud ühte mehhanismi, hiljem tahetakse teha aga järeldusi teistsuguse olukorra tarvis, siis võib ette tulla pettumusi — hinnatud mudel kirjeldab vaatlusandmeid tekitanud protsessi, mitte aga seda protsessi, mille vastu huvi tunti! Teisisõnu — kui valim pole esindav valim meid huvitava üldkogumi jaoks, siis on ka saadavad prognoosid/tehtavad järeldused valed. Olukorda iseloomustab joonis 1.1.

Kuna võime arvata, et meie valim pole esindav valim? Mõned võimalikud tähelepanekud: küsimustele vastama nõustuvad inimesed pole esindavad kõigi inimeste jaoks (vastama nõustuvad eelkõige need, kel on aega: pensionärid, töötud, lapsed,...; need kellel on mingi probleem, mida küsitlejale kui inimesele kurta – vastajad kipuvad sageli olema haigemad, neil on just katki läinud uurijale huvipakkuv tehnikavidin ning neil on seetõttu tarvidus tehnikavidina valmistajat siunata vms). Probleeme võib tekitada ka andmete kogumise viis – arsti poole pöördunud seenhaiguse all kannatajad ei pruugi olla seenhaiguse all kannatajate jaoks esindav valim – sest arsti poole pöörduvad eelkõige tõsisemate tervisehädadega patsiendid.

Samuti ei pruugi majanduslanguse ajajärgul kogutud andmetest olla kasu olukorras, kus majandus taas tõusule pöördub ja vastupidi – majandusliku tõusu ajal kogutud andmed pole esindavaks valimiks majanduslanguse ajajärgu jaoks. Konkurendi lisandumisel turule võib olukord turul muutuda ja varasemad andmed osutuda väärtusetuks jne.

Esindusliku valimi saamiseks võib kasutada lihtsat juhuslikku valikut (see pole kaugeltki ainus viis esindusliku valimini jõudmiseks, kuid statistiku vaa-

Joonis 1.1: Halb valim



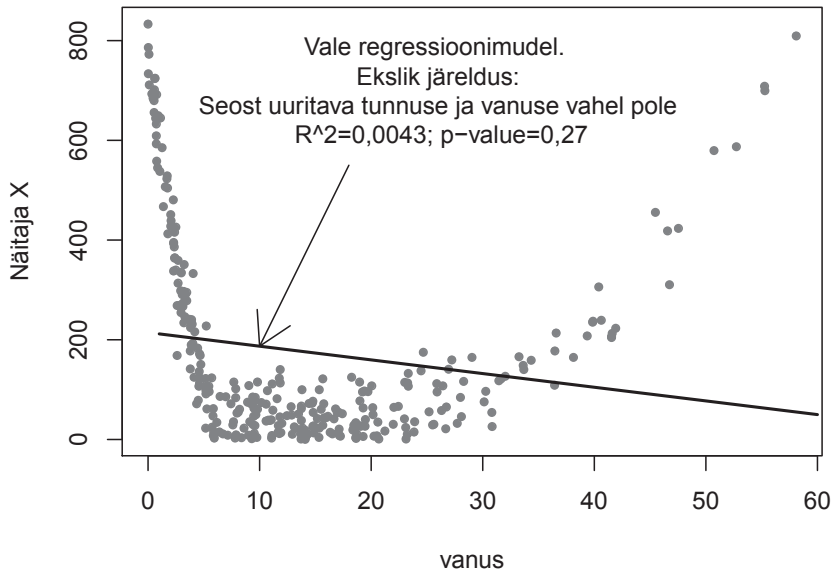
tevinklist ehk kõige lihtsam ja riskivabam meetod). Kuid isegi kui olemas on lihtsa juhusliku valiku abil saadud juhuslik valim, tuleb meeles pidada, et uuritav populatsioon ise võib ajas muutuda!

Kasutatav mudel peab olema õige

Kui seose kuju kirjeldavate tunnuste, mille põhjal moodustati mudeli maatriks $\mathbf{X}$, ja prognoositava tunnuse Y vahel pole selline, nagu mudel väidab, siis võivad tekkida probleemid. Mitte iga kahe pideva tunnuse vahelise seose kirjeldamiseks ei sobi näiteks lihtne lineaarne regressioonimudel: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Sellise mudeli kasutamisel ja pimesi uskudes tema õigsust võime lasta end kergesti eksitada, vaata ka joonist 1.2.

Antud seose iseloomustamiseks, seose olemasolu kontrollimiseks ja prognooside tegemiseks sobivad suurepäraselt lineaarsed mudelid, tuleks valida lihtsalt õige või „õigem“ lineaarne mudel. Eelmisel joonisel kujutatud seose modelleerimiseks sobib näiteks päris hästi järgmine lineaarne mudel $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \beta_5 x_i^5 + \varepsilon_i$, vaata ka joonist 1.3.

Joonis 1.2: Vale mudel



Tasub tähele panna, et mõlematel joonistel kujutatud mudelid on lineaarsed mudelid — st. tegemist on tundmatute parameetrite $\boldsymbol{\beta}$-de lineaarse funktsiooniga (aga mudelis võib sees olla argumenttunnuse x mistahes funktsioon!)

Normaaljaotuse eeldus

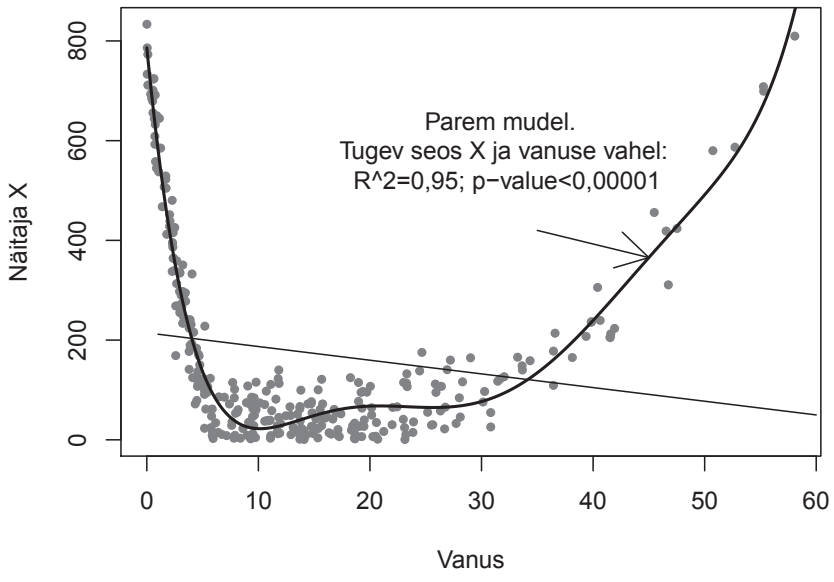
Mõningate tulemuste saamiseks (näiteks prognoosiintervalli leidmiseks) peame eeldama, et uuritava tunnuse tinglikuks jaotuseks on normaaljaotus, $\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}; \mathbf{I}\sigma^2)$. Sageli vaadeldakse ekslikult mainitud eelduse kontrollimiseks uuritava tunnuse marginaaljaotust, näiteks kasutades histogrammi või tõenäosuspaberit (QQ-plot). Miks ei sobi kasutada uuritava tunnuse histogrammi mainitud eelduse paikapidavuse kontrollimiseks?

Jääkide sõltumatuse ja konstantse hajuvuse eeldus

Mudelit kirja pannes eeldasime, et $\mathrm{D}\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$. Aga kas päriselus nimetatud eeldus ikka kehtib? Mis võib minna viltu? Vahel tasub kontrollida, kas mõni

järeldus, mida saame antud eelduse põhjal teha, ikka päriselt ka paika peab. Vahel kontrollitakse, ega sõltuva tunnuse väärtuste (või mudeli prognoosi) kasvades uuritava tunnuse hajuvus ei suurene (kui suureneb, siis pole konstantse hajuvuse eeldus rahuldatud); testitakse järjestikuste mudeli jääkide sõltumatust (mõõtmisaparatuuri kalibreerimisvead võivad näiteks tekitada sellist jääkide vahelist autokorrelatsiooni) jne.

Joonis 1.3: Parem mudel



Peatükk 2

Mudeli parameetrite hindamine

Mudeli parameetrite hindamiseks on mitmeid erinevaid võimalusi. Ühel viisil leitud hinnangutel on ühed head omadused, teisel viisil leitud hinnangutel jälle teised head omadused. Käesolevas peatükis hindame lineaarse mudeli parameetreid kolmel erineval viisil — vähimruutude meetodil; suurima tõepära meetodil ja parima lineaarse nihketa hinnangu abil.

2.1 Vähimruutude hinnang

Vähimruutude meetodi puhul leitakse mudeli parameetrite $\boldsymbol{\beta}$ hinnang $\hat{\boldsymbol{\beta}}$ selliselt, et tekkivate prognoosivigade $\boldsymbol{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ ruutude summa $\sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^T\boldsymbol{e}$ oleks minimaalne; teisisõnu minimeeritakse avaldise

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

väärtus $\boldsymbol{\beta}$ järgi.

Funktsiooni $S(\boldsymbol{\beta})$ miinimumi leidmiseks leiame esmalt tuletise vektori $\boldsymbol{\beta}$ järgi:

$$\begin{aligned} \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial(\mathbf{y}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X} - 2\mathbf{y}^T\mathbf{X}, \end{aligned}$$

võrdsustame leitud tuletise nulliga ja teisendame veidi saadud võrrandisüs-

teemi:

$$\begin{aligned} \left.\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} &= 0 \\ 2\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X} - 2\mathbf{y}^T\mathbf{X} &= 0 \\ \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}^T\mathbf{y}. \end{aligned} \tag{2.1}$$

Kui maatriks $(\mathbf{X}^T\mathbf{X})$ on pööratav, siis saame võrrandisüsteemi (2.1) lahendiks

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Paraku pole maatriks $\mathbf{X}^T\mathbf{X}$ sageli pööratav — isegi näites 1.2 toodud lihtsa dispersioonanalüüsi mudeli korral pöördmaatriksit $(\mathbf{X}^T\mathbf{X})^{-1}$ ei eksisteeri.

Mida siis teha? Kui otsime lahendeid võrrandisüsteemile $\mathbf{Ax} = \mathbf{b}$, siis juhul, kui lahend leidub, avaldub ta kujul $\mathbf{x} = \mathbf{A}^-\mathbf{b}$, kus $\mathbf{A}^-$ tähistab maatriksi $\mathbf{A}$ üldistatud pöördmaatriksit ($\mathbf{A}^-$ on maatriksi $\mathbf{A}$ üldistatud pöördmaatriks, kui $\mathbf{A} = \mathbf{AA}^-\mathbf{A}$). Tasub tähele panna, et selline lahend pole üldjuhul üheselt määratud — kuna üldistatud pöördmaatriks ei pruugi olla üheselt määratud. Seega, kui võrrandisüsteem (2.1) on lahenduv, avaldub lahend kujul

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}. \tag{2.2}$$

Leitud vähimruutude hinnang $\hat{\boldsymbol{\beta}}$ on üheselt määratud vaid juhul, kui maatriks $\mathbf{X}^T\mathbf{X}$ on pööratav. Kui aga $\mathbf{X}^T\mathbf{X}$ on kõdunud, siis eksisteerib rohkem kui üks lahend. Täpsemalt, iga $\hat{\boldsymbol{\beta}}$, mis avaldub kujul

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y} + (\mathbf{I} - (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X})\mathbf{u},$$

kus $\mathbf{u}$ on suvaline n-mõõtmeline vektor, osutub samuti võrrandisüsteemi 2.1 lahendiks.

Kuidas jääb võrrandisüsteemi lahenduvuse küsimusega? Selgub, et vaadeldav võrrandisüsteem on alati lahenduv. Paraku läheb lahenduvuse korrektseks näitamiseks vaja mõningaid tehnilisi oskuseid ja abitulemusi. Sestap lükkame tõestuse korrektse vormistamise ja vähimruutude hinnangu omadustega tutvumise seniks edasi, kuni oleme tõestanud kõik vajalikud abitulemused. Kirjutame siinkohal vaid välja vähimruutude meetodil saadud hinnangud $\mathbf{y}$-le (või $\mathbf{y}$ keskväärtusele) ja anname ka valemi prognoosivigade

arvutamiseks:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}\end{aligned}$$

$$\begin{aligned}\boldsymbol{e} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T)\mathbf{y}.\end{aligned}$$

Kuna maatriksit $\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$ läheb antud kursuse raames sageli vaja, siis võtame tema jaoks kasutusele uue tähise:

$$\mathbf{P}_\mathbf{X} := \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T. \tag{2.3}$$

Kasutades uudset tähistus võime kirjutada: $\hat{\mathbf{y}} = \mathbf{P}_\mathbf{X}\mathbf{y}$ ja $\boldsymbol{e} = (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{y}$.

Ülesanded

Vaata vabaliikmeta dispersioonanalüüsi mudelit

$$y_{ij} = \mu_i + \varepsilon_{ij},\ i = 1 \ldots k, j = 1 \ldots n_i.$$

Tõesta, et vektori $\boldsymbol{\beta}^T = (\mu_1, \ldots, \mu_k)$ hinnang (2.2) on kirja pandav kujul

$$\hat{\beta} = (\overline{y}_{1.}, \ldots \overline{y}_{k.}),$$

kus $\overline{y}_{i.}$ tähistab i. grupist pärit vaatluste aritmeetilist keskmist.