Peatükk 8

Mudeli headuse kirjeldamine ja mudeli valik

8.1 Determinatsioonikordaja R^2 ja kohandatud determinatsioonikordaja ${R_{adj}}^2$

Üritame kirjeldada, kui palju täpsemaid prognoose me saame tänu oma mudelile, tänu sõltuvate tunnuste kasutamisele. Me võime kasutada selleks mudeli jääkide summat. Vaatame, milline oleks jääkide ruutude summa siis, kui me ühtegi tunnust ei kasuta y-tunnuse prognoosimisel ehk kui me leiame oma prognoosid vaid mudelist $Y_i = \mu + \varepsilon_i$ ehk $\mathbf{y} = \mathbf{1}\mu + \boldsymbol{\varepsilon}$. Sellisel juhul olgu jääkide ruutude summa

$$SSE_1 = \mathbf{y}^T(\mathbf{I} - \mathbf{P}_1)\mathbf{y} = \mathbf{y}^T(\mathbf{I} - \frac{1}{n}\mathbf{J}_{n\times n})\mathbf{y} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

Kasutades Y-tunnuse prognoosimisel teiste tunnuste abi on prognoosivead enamasti veidi väiksemad — tänu lisainformatsioonile — ja prognoosijääkide ruutude summa $SSE = \mathbf{y}^T(\mathbf{I} - \mathbf{P_X})\mathbf{y}$ tuleb väiksem. Täpsemalt öeldes $SSE_1 - SSE$ võrra väiksem. Prognoosijääkide ruutude summa suhtelist vähenemist kutsutaksegi determinatsioonikordajaks:

$$R^2 := \frac{SSE_1 - SSE}{SSE_1} = 1 - \frac{SSE}{SSE_1}.$$

Sageli väljendatakse determinatsioonikordajat ka protsentides,

$$R^2 := \frac{SSE_1 - SSE}{SSE_1} \cdot 100\%.$$

Alternatiivina võime tunda huvi ka jääkide hajuvuse vastu. Mida paremini meie mudel prognoosib, seda väiksem on mudeli jääkide hajuvus (seda konstantsemad nad on — seda lähemal nullile). Vaatame, palju väheneb jääkide hajuvus tänu sõltumatute tunnuste (mudeli) kasutamisele. Nihketa hinnang jääkide hajuvusele on keskmine ruutviga,

$$MSE = SSE/(n - \mathrm{rank}(\mathbf{X})).$$

Kui me sõltuvaid tunnuseid ei kasutaks, oleks jääkide — prognoosivigade — hajuvuse hinnanguks

$$MSE_1 = SSE_1/(n-1).$$

Jääkide hajuvuse suhtelist vähenemist kutsutakse kohandatud determinatsioonikordajaks:

$$R^2_{adj} := \frac{MSE_1 - MSE}{MSE_1} = 1 - \frac{MSE}{MSE_1}.$$

Mõlemat suurust, nii determinatsioonikordajat R^2 kui ka kohandatud determinatsioonikordajat R^2_{adj} saab vaadata kui suuruse $\frac{\sigma_1^2 - \sigma^2}{\sigma_1^2} = 1 - \frac{\sigma^2}{\sigma_1^2}$ hinnangut, kus σ_1^2 on jääkide hajuvus ainult vabaliiget sisaldava mudeli korral (eeldades, et ka x-tunnused on juhuslikud suurused). Lihtsalt determinatsioonikordaja kasutab nihkega suurima tõepära hinnanguid jääkide hajuvusele: $\hat{\sigma}^2 = SSE/n$ ja $\hat{\sigma}_1^2 = SSE_1/n$, kohandatud determinatsioonikordaja aga nihketa hinnanguid $\tilde{\sigma}^2 = SSE/(n - \mathrm{rank}(\mathbf{X}))$ ja $\tilde{\sigma}_1^2 = SSE_1/(n - \mathrm{rank}(\mathbf{1}))$.

Märkus: tunnuste lisamisel mudelisse saab determinatsioonikordaja väärtus vaid kasvada.

Veendume selles. Olgu $\mathbf{X}_0$ lihtsama mudeli mudeli maatriks ja $\mathbf{X}_1$ rikkama mudeli — mudeli, kuhu on lisatud üks või enam tunnust — mudeli maatriks. Siis rikkama mudeli jääkide summa SSE_1 on alati suurem kui vaesema mudeli jääkide summa SSE_0:

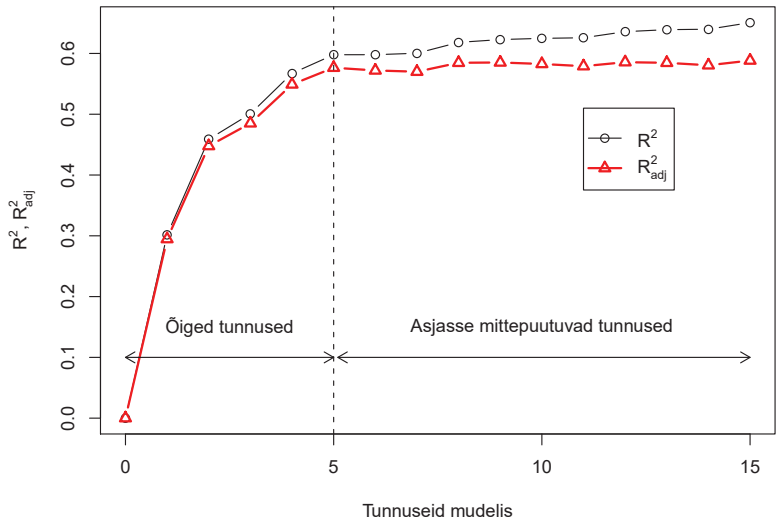
$$\begin{aligned} SSE_0 - SSE_1 &= \mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y} - \mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{y} \\ &= \mathbf{y}^T(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y} \\ &= \mathbf{y}^T(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y} \\ &= \{(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}\}^T(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y} \\ &\geq 0 \end{aligned}$$

sest prognooside erinevuste ruutude summa on mittenegatiivne (analoogne arutelu sellele, mis toimus peatükis 4, F-testi tutvustavas osas). Aga kui

$SSE_0 \geq SSE_1$, sõltumata sellest, kas juurdelisatud tunnused on vajalikud või mitte, siis saabki determinatsioonikordaja tunnuste lisamisel vaid kasvada.

Seevastu kohandatud determinatsioonikordaja väärtus peaks asjasse mittepuutuva tunnuse lisamisel mudelisse jääma ligikaudu samaks (kuid võib kõikuda veidi siia-ja-sinnapoole hinnangu juhuslikkusest tingitult). Determinatsioonikordaja ja kohandatud determinatsioonikordaja käitumist kirjeldab joonis 8.1.

Joonis 8.1: Determinatsioonikordaja ja kohandatud determinatsioonikordaja



Ülesanded. Näita, et kohandatud determinatsioonikordaja ei saa olla suurem kui determinatsioonikordaja.

8.2 Mallows'i C_p kriteerium

Oletame, et teame mudelit, mida võib korrektseks pidada:

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

Paraku võib ülaltoodud mudel sisaldada parameetreid, mida pole tegelikult vaja (osad $\boldsymbol{\beta}$ elemendid võivad olla ka nullid). Selmet et hakata tegelikult

nullilise väärtusega parameetreid otsima võtab Mallows'i C_p kriteerium ette julgema lähenemise. Antud kriteeriumist lähtuvalt peaksime eelistama mudelit mis viib võimalikult täpse prognoosini — ükskõik kas siis prognoosimiseks kasutatav mudel ise on õige või pole ta seda mitte.

Täpsemalt: kui kaalumise all on erinevad mudelid kujul

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

siis Mallows'i arvates peaksime eelisatama mudelit, mille korral keskmine ruutviga on minimaalne:

$$\mathrm{E}\left(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}_i\hat{\boldsymbol{\beta}}_i\right)^T\left(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}_i\hat{\boldsymbol{\beta}}_i\right).$$

Leidmaks ülaltoodud avaldise väärtust paneme esmalt tähele, et

$$\begin{aligned}\mathrm{E}\left(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}_i\hat{\boldsymbol{\beta}}_i\right) &= \mathrm{E}\left(\mathbf{X}\boldsymbol{\beta} - \mathbf{P}_{\mathbf{X}_i}\mathbf{y}\right)\\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{P}_{\mathbf{X}_i}\mathbf{X}\boldsymbol{\beta}\\ &= (\mathbf{I} - \mathbf{P}_{\mathbf{X}_i})\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

ja

$$\begin{aligned}\mathrm{D}\left(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}_i\hat{\boldsymbol{\beta}}_i\right) &= \mathrm{D}\left(\mathbf{P}_{\mathbf{X}_i}\mathbf{y}\right)\\ &= \mathbf{P}_{\mathbf{X}_i}\sigma^2.\end{aligned}$$

Meenutame, et lemma 3.1 järgi $\mathrm{E}\left(\mathbf{y}^T\mathbf{A}\mathbf{y}\right) = \mathrm{tr}(\mathbf{A}\mathbf{V}) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$, kus $\mathrm{E}\mathbf{y} = \boldsymbol{\mu}$ ja $\mathrm{D}\mathbf{y} = \mathbf{V}$. Seega

$$\begin{aligned}\mathrm{E}\left(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}_i\hat{\boldsymbol{\beta}}_i\right)^T\left(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}_i\hat{\boldsymbol{\beta}}_i\right) &= \mathrm{tr}(\mathbf{P}_{\mathbf{X}_i}\sigma^2) + \boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_i})\mathbf{X}\boldsymbol{\beta}\\ &= \sigma^2 p_i + \boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_i})\mathbf{X}\boldsymbol{\beta},\end{aligned}$$

kus $p_i := \mathrm{rank}(\mathbf{P}_{\mathbf{X}_i})$ on i. mudeli hinnatavate parameetrite arv.

Raske on soovitud keskmist ruutviga ülaltoodud arvutamisvalemi abil leida, sest tundmatute parameetrite $\boldsymbol{\beta}$ ja σ^2 tegelikud väärtused pole enamasti ju teada. Küll aga võime meenutada:

$$\begin{aligned}\mathrm{E}\mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_i})\mathbf{y} &= \sigma^2\mathrm{tr}(\mathbf{I} - \mathbf{P}_{\mathbf{X}_i}) + \boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_i})\mathbf{X}\boldsymbol{\beta}\\ &= \sigma^2(n - p_i) + \boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_i})\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

ja seega annaks statistik kujul

$$\frac{\mathbf{y}^T(\mathbf{I}-\mathbf{P}_\mathbf{X})\mathbf{y}}{n-\text{rank}(\mathbf{X})}(p_i-(n-p_i))+\mathbf{y}^T(\mathbf{I}-\mathbf{P}_{\mathbf{X}_i})\mathbf{y} = MSE(2p_i-n)+\mathbf{y}^T(\mathbf{I}-\mathbf{P}_{\mathbf{X}_i})\mathbf{y}$$

nihketa hinnangu keskmisele ruutveale.

Võiksime seega siis otsustada mudeli kasuks, mille puhul prognooside keskmine ruutvea nihketa hinnang oleks minimaalne. Mallows soovitas kasutada ülatoodud hinnangu skaleeritud varianti:

$$\begin{aligned} C_p &= \frac{MSE(p_i-(n-p_i))+\mathbf{y}^T(\mathbf{I}-\mathbf{P}_{\mathbf{X}_i})\mathbf{y}}{MSE} \\ &= 2p_i - n + \frac{\mathbf{y}^T(\mathbf{I}-\mathbf{P}_{\mathbf{X}_i})\mathbf{y}}{\mathbf{y}^T(\mathbf{I}-\mathbf{P}_\mathbf{X})\mathbf{y}/(n-p)} \end{aligned}$$

ehk tuleks otsustada sellise mudeli kasuks, mille C_p väärtus on väikseim.

Kui kaalumisel olev i. mudel on ka õige, siis

$$\mathbf{y}^T(\mathbf{I}-\mathbf{P}_{\mathbf{X}_i})\mathbf{y} \approx \sigma^2(n-p_i)$$

ja

$$\mathbf{y}^T(\mathbf{I}-\mathbf{P}_\mathbf{X})\mathbf{y}/(n-p) \approx \sigma^2.$$

Seega võiks sellisel juhul oodata, et C_p väärtus oleks ligikaudu võrdne p_i-ga:

$$C_p \approx 2p_i - n + \frac{(n-p_i)\sigma^2}{\sigma^2} = p_i.$$

Täpsemalt:

$$\mathrm{E}C_p = p_i + 2\frac{p-p_i}{n-p-2},$$

mis muidugi on peaaegu võrdne p_i -ga (kui n on suur ja $p-p_i$ on suhteliselt väike). Mallows'i C_p alusel valitud mudel võib seega olla nii õige kui vale, vahel võib väikseima keskmise ruutveani viia tõepoolest ka (veidike) vale mudel.

8.2.1 AIC ja Mallows'i C_p

Sobivaima mudeli valikuks kasutatakse sageli ka Akaike informatsioonikriteeriumi ehk AIC väärtust. Olgu hinnatavate parameetrite vektor i. mudelis Θ_i, parameetrite suurima tõepära hinnang $\hat{\Theta}_i$ ja log-tõepärafunktsioon $l(\Theta_i)$. Olgu hinnatavate parameetrite arv k_i. Sellisel juhul on i. mudeli AIC väärtus AIC_i leitav

$$AIC_i = -2 \cdot l(\hat{\Theta}_i) + 2 \cdot k_i.$$

Akaike soovitas valida sobivaimaks mudeliks mudeli, mille korral AIC_i väärtus on võimalikult väike.

Lineaarse mudeli korral tuleb AIC väärtuseks (vaata ka valemit 2.13):

$$AIC_i = n\log(2\pi\hat{\sigma_i}^2) + 2\frac{(\mathbf{y}-\mathbf{X}_i\hat{\boldsymbol{\beta}}_i)^T(\mathbf{y}-\mathbf{X}_i\hat{\boldsymbol{\beta}}_i)}{2\hat{\sigma_i}^2} + 2k_i,$$

kus k_i on hinnatavate parameetrite arv ($k_i = \text{rank}(\mathbf{X}_i) + 1 = p_i + 1$, sest lisaks $\boldsymbol{\beta}_i$-le hinnatakse ka σ^2 väärtust).

Juhul kui teame σ^2 väärtust, siis tuleks AIC-väärtuseks i. mudeli jaoks:

$$AIC_i = n\log(2\pi\sigma^2) + \frac{(\mathbf{y}-\mathbf{X}_i\hat{\boldsymbol{\beta}}_i)^T(\mathbf{y}-\mathbf{X}_i\hat{\boldsymbol{\beta}}_i)}{\sigma^2} + 2p_i,$$

Võrdle viimasena saadud AIC-väärtuse arvutusvalemit Mallows'i C_p arvutusvalemiga (juhul kui σ^2 väärtus on teada):

$$C_p = 2p_i - n + \frac{\mathbf{y}^T(\mathbf{I}-\mathbf{P}_{\mathbf{X}_i})\mathbf{y}}{\sigma^2}$$

Erinevus ülaltoodud valemite vahel seisneb vaid konstandis (-n vs $n\log(2\pi\sigma^2)$), mis jääb samaks kõigi võrreldavate mudelite korral. Suuresti käituvad Mallows'i C_p ja AIC sarnaselt, seega valides väiksema AIC-väärtusega mudeli valime ühtlasi ka keskmise ruutvea mõttes parima mudeli (vähemalt lineaarsete mudelite korral on see nii).

Ülesanded

1. Võrdleme kümmet mudelit. Üheksa võrreldavat mudelit on kujul

   $$y = c_1x_1 + c_2x_2 + c_3x_i + \varepsilon,\ i = 3\ldots 11$$

   ja kümnes mudel on kujul

   $$y = c_1x_1 + c_2x_2 + c_3x_3 + c_3x_4 + \ldots c_{11}x_{11} + \varepsilon.$$

   Oletame, et kõik võrreldavad mudelid on õiged (kuidas see on võimalik?) ja lisaks eeldame, et valimi maht n on suur. Milline võiks olla väljavalitud parima mudeli Mallows'i C_p statistiku väärtus?

2. Soovime kirjeldada seda, kuidas laste pikkus sõltub laste vanusest. Kaalumisel on kolm erinevat katseplaani:

(a) Uurida 4.-6. klassis käivaid lapsi

(b) Uurida 1.-11. klassis käivaid lapsi

(c) Uurida 1. ja 11. klassis käivaid lapsi

Millise katseplaani korral tuleks mudeli determinatsioonikordaja suurim, millise puhul väikseim? Miks?