Peatükk 7

Mudeli eelduste kontrollimisest

Rääkides mudelite võrdlemisest või ka teiste teemade juures oleme teinud eelduseid — nõudnud, et vaatlused oleksid normaaljaotusega ja konstantse hajuvusega, meie mudel peab olema kindla kujuga jne. Mõningaid tehtud eeldustest saab aga kontrollida. Enamasti nõuab mudeli eelduste kontroll mudeli jääkide põhjalikumat uurimist. Sestap üritame esmalt aru saada, milliste omadustega on mudeli jäägid.

7.1 Jääkidest

Praktikas ei saa me vaadelda mitte tegelikke prognoosivigu $\boldsymbol{\varepsilon}$, vaid mudeli (hinnatud) jääke, $\boldsymbol{e} := \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ ehk $\boldsymbol{e} = (\mathbf{I} - \mathbf{P_X})\mathbf{y}$. Järgnevalt toome ära mõned tähelepanekud vektori $\boldsymbol{e} = (e_1; \ldots; e_n)^T$ kohta.

Esiteks, kui $\mathbf{1} = (1, 1, \ldots, 1)^T \subset \mathcal{C}(\mathbf{X})$, siis $\mathbf{1}^T\mathbf{P_X} = \mathbf{1}^T$ ja järelikult $\mathbf{1}^T\boldsymbol{e} = \mathbf{1}^T(\mathbf{I} - \mathbf{P_X})\mathbf{y} = \mathbf{0}^T \cdot \mathbf{y} = 0$ ehk jääkide summa (ja keskmine) on alati null (ükskõik, kas kehtib võrdus $\mathrm{E}\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ või mitte...). Seega ei saa hinnatud jääkide keskmise $\overline{e_i}$ abil kontrollida, kas prognoosivigade keskväärtus ikka on 0, $\mathrm{E}\boldsymbol{\varepsilon} = \mathbf{0}$.

Teiseks, isegi kui $\mathrm{D}\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$, on $\mathrm{D}\boldsymbol{e} = \mathrm{D}(\mathbf{I} - \mathbf{P_X})\mathbf{y} = (\mathbf{I} - \mathbf{P_X})\mathrm{D}\mathbf{y}(\mathbf{I} - \mathbf{P_X}) = \sigma^2(\mathbf{I} - \mathbf{P_X})$, ehk vaadeldud jäägid pole sõltumatud ja ei pruugi olla ka sama dispersiooniga (sõltuvalt maatriksi $\mathbf{I} - \mathbf{P_X}$ kujust) isegi siis, kui mudeli tegelikud jäägid on sõltumatud ja konstantse dispersiooniga. See, et jäägid ei pruugi olla sama dispersiooniga (isegi kui tegelikud jäägid $\boldsymbol{\varepsilon}$ seda on), tekitab mitmetes kontrollprotseduurides mõnevõrra segadust. Sestap kasutatakse sageli nn. standardiseeritud jääke: hinnatakse jääkide e_i disper-

sioon, eeldades prognoosivigade $\boldsymbol{\varepsilon}$ konstantsent dispersiooni (leitakse maatriksi $\hat{\sigma}^2(\mathbf{I}-\mathbf{P_X}) = MSE(\mathbf{I}-\mathbf{P_X})$ diagonaali elemendid). Seejärel leitakse suurused

$$r_i = e_i/\sqrt{\hat{D}(e_i)} = e_i/\sqrt{MSE(1-h_{ii})},$$

kus h_{ii}-ga on tähistatud maatriksi P_X (tuntud ka mütsimaatriksi H — *hat-matrix* — nime all) diagonaali i. element. Standardiseeritud jääke tuleks kasutada jääkide konstantse hajuvuse ja normaaljaotuse eelduste kontrollimisel. Märkus: vanemas kirjanduses on vahel (harva) kasutatud standardiseeritud jääkide mõistet tähistamaks suurust $e/\sqrt{MSE}$. Nimetatud suurustega pole suurt midagi peale hakata.

R'is saab hinnatud mudeli *mudel* standardiseeritud jääke tellida käsuga *rstandard*(*mudel*), SAS'is GLM-protseduuris saab standardiseeritud jääke salvestada kasutades OUTPUT-käsu võtmesõna STUDENT.

Iseloomustamaks võimaliku muutuse suurust, vaatame järgmist mudelit:

$$y_i = c_0 + c_1 x_{1i} + c_2 x_{2i} + c_3 I_{linn=A} + c_4 I_{linn=B} + \varepsilon_i.$$

Mudeli hindamiseks kasutatud andmed on toodud tabelis 7.1.

Tabel 7.1: Tavalised ja standardiseeritud jäägid.

y	x_1	x_2	$linn$	e_i	$e_i/\sqrt{MSE}$	r_i
28,3	5,0	0,1	A	0,053	0,061	0,752
34,6	1,0	4,0	A	-0,053	-0,061	-0,752
10,0	2,0	0,2	B	-0,082	-0,094	-0,119
9,7	1,5	0,3	B	0,723	0,825	0,890
5,7	1,2	0,1	B	-1,334	-1,522	-1,765
7,0	1,1	0,1	B	0,285	0,326	0,399
9,2	1,8	0,3	B	-0,735	-0,839	-0,995
8,9	1,3	0,5	B	-0,421	-0,481	-0,520
7,3	1,0	0,2	B	0,413	0,471	0,576
10,3	1,4	0,4	B	1,151	1,313	1,417

Kahtleme, kas linnas A ja linnas B on ikka uuritava tunnuse hajuvus samasugune. Võrdleme jääkide (e) ja standardiseeritud jääkide (r) hajuvust — standardhälvet — linnas A ja linnas B:

$\sqrt{\hat{D}(e\|linn = A)}$	$\sqrt{\hat{D}(e\|linn = B)}$	suhe
0.075	0.811	0.093

$\sqrt{\hat{D}(r\|linn = A)}$	$\sqrt{\hat{D}(r\|linn = B)}$	suhe
1.06	1.05	1.02

Märkame, et jääkide põhjal tehtud otsus hajuvuse konstantsuse (*homoscedasity*) kohta ja standardiseeritud jääkide põhjal tehtud otsus võivad olla märkimisväärselt erinevad.

Märkus:
Kui soovime mudeli jääke kasutada sisestusvigade vms leidmiseks, siis võib juhtuda, et tänu mõnele vigaselt sisestatud vaatlusele hinnatakse jääkide hajuvus (MSE) ekslikult liiga suureks. Kui aga MSE on ekslikult liiga suur, siis tulevad ka standardiseeritud jäägid pisikesed. Sestap tekib inimestel vahel soov vaadata selliseid jääke, kus jääkide dispersiooni hindamisel pole kasutatud seda vaatlust, mille jääki parajasti leitakse. Sellisel juhul kasutatakse nn Studentiseeritud jääke *Studentized residuals*'i:

$$r_i{}^* = e_i/\sqrt{MSE_{(-i)}(1 - h_{ii})},$$

kus $MSE_{(-i)}$ on MSE (jääkide dispersiooni hinnang), mille saaksime, kui hindaksime jääkide dispersiooni ilma i. vaatlust kasutamata. Statistikapaketis R saab selliseid jääke leida funktsiooni `rstudent` abil, SAS'i GLM-protseduuris saab neid jääke salvestada kasutades OUTPUT-käsu võtmesõna RSTUDENT. Rusikareegel — kui standardiseeritud või Studentiseeritud jääk pole vahemikus $[-2 \ldots 2]$, siis on tegemist veidrikuga ja vajalikuks võib osutuda täiendav uurimine (äkki on andmete sisestamisel tehtud viga?).

7.2 Mõjukus

Mitte kõik vaatlused pole sündinud võrdsena. Mõnel vaatlusel võib olla suurem roll tulemuste (hinnangud, prognoosid) leidmisel kui mõnel teisel. Enamasti on vaatluste mõju tingitud sellest, et nad on erilised. Üks vaatlus ehk objekt võib olla eriline mitmel moel. Tal võib olla ebaharilik Y-tunnuse väärtus või võib ebatüüpiliseks osutuda just tema rida mudeli maatriksis X (tema sõltumatute tunnuste väärtused on veidrad). Esmalt vaatame neid vaatluseid, kellel sõltumatute tunnuste väärtused on veidrad. Neid kutsutakse mõjukateks vaatlusteks. Terminit mõjukus tuntakse inglise keeles kui *leverage*.

Täpsem definitsioon oleks järgmine: i. vaatluse mõjukuseks (mõjukus on inglise keeles *leverage*) kutsume maatriksi $\mathbf{P_X}$ diagonaali i. elementi h_{ii}. Märkus: R'is saab mudeli hindamiseks kasutatud vaatluste mõjukust arvutada käsuga *hatvalues(mudel)*.

Üks võimalus mõelda mõjukuse peale oleks järgmine. Leiame mudeli maatriksi ridade keskmise (keskmise rea), $\overline{\mathbf{x}}$. Arutleme, kui kaugel on i. vaatlusele vastav mudeli maatriksi rida $\mathbf{x}_i$ keskmisest reast $\overline{\mathbf{x}}$. Siin ei sobi kasutada lihtlabast kaugust, nagu näiteks Eukleidilist kaugust (kas mõistad, miks eukleidiline kaugus ei sobi? Mõtle näiteks, mis saaks siis, kui muudaksime ühe seletava tunnuse mõõtühikut — hakkaksime pikkust mõõtma sentimeetrite asemel meetrites?). Sobivam kauguse mõõt on Mahalanobise kaugus:

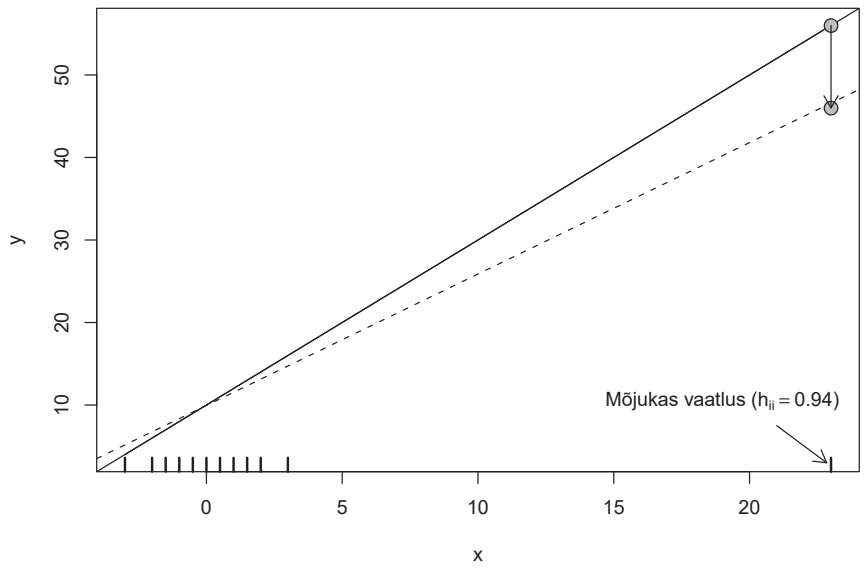
$$D_i{}^2 = (\mathbf{x}_i - \overline{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}),$$

kus $\mathbf{S}$ on sõltumatute tunnuste (mudeli maatriksi $\mathbf{X}$ veergude) kovariatsioonimaatriksi hinnang (tõsi, käsitleme mudeli maatriksit ju fikseerituna ja konstandil puudub rangelt võttes dispersioon... aga siiski võime kasutada sama valemit, mille abil tavaliselt hindame juhuslike suuruste kovariatsioonimaatriksit...).

Annab näidata, et vaatluse mõjukus ja tema Mahalanobise kaugus „keskmisest $\mathbf{X}$-ist“ on teineteisest saadavad lihtsa lineaarteisenduse abil — mida suurem on vaatluse i mõjukus, seda suurem on temale vastava mudeli maatriksi $\mathbf{X}$ rea $\mathbf{x}_i$ kaugus keskmisest reast $\overline{\mathbf{x}}$ (tõestusel tuleb jälgida, et enamasti pole $\mathbf{S}$ pööratav — ja seega peame end Mahalanobise kauguse mõttes täpsemalt väljendama õppima...):

$$h_{ii} = \frac{1}{n} + \frac{D_i{}^2}{n-1}.$$

Suure mõjukusega ($h_{ii} \approx 1$) vaatlustele vastavate jääkide dispersioon $\sigma^2(1 - h_{ii})$ on väike, nullilähedane. Kuna jääkide keskväärtus on null, siis väikesest dispersioonist järeldub, et vastavad jäägid peavad olema väikesed, nullilähedased. Põhjuseks on asjaolu, et suure mõjukusega punktid sikutavad lineaarse mudeli parameetrite hinnangud endale sobivaks — näiteks tirib mõjukas vaatlus regressioonjoone selliseks, et regressioonjoon ikka mõjuka punkti vahetus läheduses püsiks, vaata ka alljärgnevat joonist.



Kui suur h_{ii} väärtus on suur? Leiame esmalt kõigi mõjukuste summa. Kõigi mõjukuste summa on aga idempotentse maatriksi $\mathbf{P_X}$ kõigi omaväärtuste summa. Viimane on aga idempotentse maatriksi puhul võrdne maatriksi astakuga:

$$\begin{aligned} \sum_{i=1}^{n} h_{ii} &= \operatorname{rank}(\mathbf{P_X}) \\ &= \operatorname{rank}(\mathbf{X}) =: p. \end{aligned}$$

Kuna vaatluseid on n, saame keskmiseks mõjukuseks p/n ($\approx$ parameetrite arv jagatud vaatluste arvuga).

Milliseid mõjukuse väärtuseid võime näha? Kuna $\mathbf{P_X}$ on idempotentne, $\mathbf{P_X} = \mathbf{P_X}\mathbf{P_X}$, ja sümmeetriline, siis

$$h_{ii} = \sum_{j=1}^{n} h_{ij}h_{ji} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{j\neq i}^{n} h_{ij}^2$$

ja järelikult (esitatav ruutude summana) peavad mõjukused olema mittene-

gatiivsed, $h_{ii} \geq 0$. Kuna

$$\begin{aligned} h_{ii} - h_{ii}^2 &= \sum_{j \neq i}^{n} h_{ij}^2 \\ h_{ii}(1 - h_{ii}) &= \sum_{j \neq i}^{n} h_{ij}^2 \end{aligned}$$

ja $h_{ii} \geq 0$, siis järelikult ka $1 - h_{ii} \geq 0$ ehk $h_{ii} \leq 1$. Seega vaatluse mõjukus on vahemikus 0..1 ja keskmine mõjukus on p/n. Kui mingi vaatluse mõjukus on tugevalt suurem hallist keskmisest, siis järelikult on tegemist VIP-iga ja vaatlus väärib erikohtlemist. Rusikareegeleid on palju, aga mainime siin ära mõned: kui $h_{ii} > \frac{2p}{n}$ või kui $h_{ii} \in [0{,}2..0{,}5]$, siis on tegemist mõjuka vaatlusega, kui $h_{ii} > 0{,}5$, siis on tegemist juba äärmiselt mõjuka vaatlusega.

Mõjukad vaatlused pole iseenesest halvad vaatlused — pigem vastupidi. Suur mõjukus ju näitab, et me vajame seda vaatlust hädasti parameetrite hindamisel. Pigem võib mõjukust silmas pidada uuringuplaani koostades — hea uuringuplaani/katseplaani puhul ei jää uuringu tulemused sõltuma ühest- kahest mõõtmisest! Muuseas, mõjukust saab välja arvutada juba enne y-tunnuse väärtuste mõõtmist, seega võime vaatluste mõjukust uurida juba katse planeerimise ajal.

Kui aga on juhtunud nii, et meie poolt analüüsitavas andmestikus on sees mõni mõjukas vaatlus, peame tegema kõik mis võimalik selleks, et selle vaatlusega midagi korrast ära poleks — sest sellest vaatlusest võib sõltuda väga palju... .

7.3 Cook'i kaugus

Cook'i kaugust on võimalik defineerida päris mitmel moel, kusjuures kõigile neile võimalustele võib anda ka erineva kuid ilusa interpretatsiooni.

Esimene võimalus Cook'i kaugust konstrueerida.
Meie mudeli prognoos vaatlusvektorile on $\hat{\mathbf{y}} = \mathbf{P_X y}$. Kui viskaksime andmestikust välja i. vaatluse, saaksime veidi teistsuguse prognoosi neile samadele n vaatlusele, $\hat{\mathbf{y}}_{[-i]}$. Leiame prognooside erinevuste ruutude summa, $(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{[-i]})^T(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{[-i]})$. Edasi arvutame prognoosi muutuse parameetri kohta, mõõdetuna standardhälbe ühikutes. Saadud suurust kutsutaksegi (i. vaatluse) Cook'i kauguseks:

$$C_i = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{[-i]})^T(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{[-i]})}{p \cdot MSE}.$$

Teine võimalus Cook'i kaugust defineerida on vaadata parameetrite hinnangute muutust i. vaatluse eemaldamisel andmestikust. Kui kõiki vaatluseid kasutame, saame parameetervektori hinnanguks $\hat{\beta}$. Ilma i. vaatluseta saaksime hinnanguks $\hat{\beta}_{[i]}$. Soovime nüüd kirjeldada, kui erinevad need kaks parameetervektori hinnangut on. Erinevuste ruutude summa või nende vektorite eukleidiline kaugus aga praegu eriti hästi ei sobi — sest erinevad X-tunnused on mõõdetud erinevates mõõtühikutes ja kaugus jääks seega sõltuma sõltumatute tunnuste mõõtmiseks kasutatud mõõtühikutest. Sestap kasutame Mahalanobise kaugust (täpsemalt: Mahalanobise kauguse ruutu), $(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[-i]})^T \mathrm{D}(\hat{\boldsymbol{\beta}})^{-1} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[-i]})$. Kuna $\mathrm{D}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1}$, siis $\mathrm{D}(\hat{\boldsymbol{\beta}})^{-1}$ hinnanguks võiks sobida $(\mathbf{X}^T\mathbf{X})/MSE$. Kui me nüüd vaatame kahe parameetervektori hinnangu vahelist kaugust per parameeter, saamegi Cook'i kauguse:

$$C_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[-i]})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[-i]})}{p \cdot MSE}.$$

Muuseas, Cook'i kaugus on leitav mudeli jääki ja mõjukust kasutades:

$$C_i = \frac{{e_i}^2}{p \cdot MSE} \left(\frac{h_{ii}}{(1 - h_{ii})^2} \right).$$

Seega, kui mõjukus on väike, $h_{ii} \approx 0$, siis on raske saada suurt Cook'i kaugust — sellest vaatlusest ei sõltu midagi, hinnangud ei muutu ükskõik kas me jätame ta mudelisse või mitte...

Samuti, kui mudeli jääk $e_i = 0$, siis olgu vastav vaatlus kuitahes mõjukas — kui tema arvamus ühtib ülejäänute arvamusega, siis tema eemaldamine mudelist ei muuda samuti midagi...

Kui suur Cook'i kaugus on suur? Oletame,et tahame kontrollida hüpoteesi $H_0 : \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{[-i]}$. Milline näeks välja sellisele hüpoteesile vastava F-testi teststatistik? Sageli vaadeldakse sellist F-testi olulisuse nivool 0,5 (ühe vaatluse pärast ülejäänud vaatluste põhjal arvutatud parameetervektori hinnangu välistamine on nagunii ennekuulmatu, sestap olgem olulisuse nivooga veidi leebemad kui tavaliselt...). Aga F-jaotuse $F_{p;n-p}$ 0,5-kvantiil on ligikaudu 1, $f_{0,5;p;n-p} \approx 1$, seega $C_i > 1$ loetakse vägagi meelerahu häirivaks olukorraks.

7.4 Durbin-Watson'i test

Lineaarsete mudelite kursuses oleme eeldanud, et vaatlused on sõltumatud. Paraku alati nii see pole. Eriti tihti võivad tekkida probleemid järjestikuste mõõtmiste korral. Mõõdame näiteks taimede pikkuseid. Kasutame selleks

mingit mõõteriista, näiteks joonlauda. Mõõtmiseid alustades, varahommikul, on veel üsna külm. Päeva edenedes läheb aga aina soojemaks. Mida soojemaks läheb, seda pikemaks aga venib soojuspaisumise tõttu ka meie mittekvaliteetne joonlaud ja seda suuremaid vaatlustulemusi saame. Tulemuseks on sõltuvad vaatlused/vaatlusvead — kui üks mõõtmine näitab ebaloogiliselt pikka taime, siis järgmisena mõõdetud taim on kardetavasti ka „liiga pikk“ — sest ka tema mõõdeti päeva kõige kuumemal ajal. Taolistest mõõteaparatuuri kalibreerimisvigadest, küsitlejate osavuse tõusust tingitud nihetest jne ei pruugi me teadlikud olla. Sestap võib osutuda mõistlikuks kontrollida, ega meie andmetes sedalaadi probleemi ei esine.

Järjestikuste vaatluste korreleeritust kontrollitakse sageli Durbin-Watson'i testi abil. Antud testi teststatistikuks on suurus

$$d = \frac{\sum_{i=1}^{n-1} \left(\hat{\varepsilon}_{i+1} - \hat{\varepsilon}_i\right)^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}$$

Kuidas saadud statistikut interpreteerida?

Kui ühel mõõtmisel tehtud viga „kandub üle“ ka järgmisesse mõõtmisesse, siis võime kirja panna nn. autoregressiivse mudeli lineaarse mudeli jääkidele:

$$\varepsilon_{i+1} = \rho\varepsilon_i + \tau_{i+1},$$

kus $\rho \in [0 \ldots 1]$ iseloomustab, kui suur osa veast kandub edasi järgmisse mõõtmisse. Suurus τ_{i+1} on aga varasematest mõõtmistest sõltumatu, uus, vea komponent: $\tau_{i+1} \perp \tau_i$. Näiteks mõõdetava indiviidi isikupära, konkreetsel mõõtmisel tehtav mõõtmisviga, mis ei sõltu mõõteriista kalibratsiooniveast jne. Sellisel juhul

$$\begin{aligned} \mathrm{E}\left(\varepsilon_{i+1} - \varepsilon_i\right)^2 &= \mathrm{D}\left(\varepsilon_{i+1} - \varepsilon_i\right) \\ &= \mathrm{D}\left(\rho\varepsilon_i + \tau_{i+1} - \varepsilon_i\right) \\ &= \mathrm{D}\left((\rho - 1)\varepsilon_i + \tau_{i+1}\right) \\ &= (\rho - 1)^2\sigma^2 + \mathrm{D}\left(\tau_{i+1}\right). \end{aligned}$$

Kuna aga

$$\begin{aligned} \mathrm{D}(\varepsilon_i) &= \mathrm{D}(\varepsilon_{i+1}) \\ \sigma^2 &= \mathrm{D}(\rho\varepsilon_i + \tau_{i+1}) \\ \sigma^2 &= \rho^2\sigma^2 + \mathrm{D}(\tau_{i+1}) \\ \mathrm{D}(\tau_{i+1}) &= (1 - \rho^2)\sigma^2. \end{aligned}$$

siis saame, et

$$\mathrm{E}\left(\varepsilon_{i+1}-\varepsilon_i\right)^2=(\rho-1)^2\sigma^2+\sigma^2(1-\rho^2).$$

Statistiku d arvutusvalemis nimetajas asuvad liidetavad on kujul $\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = n\overline{\hat{\varepsilon}_i^2} \approx n\mathrm{E}\left(\varepsilon_i\right)^2 = n\sigma^2$. Lugejaga saame läbi viia samasuguse lähenduse, $\sum_{i=1}^{n-1}\left(\hat{\varepsilon}_{i+1}-\hat{\varepsilon}_i\right)^2 \approx (n-1)\cdot\mathrm{E}\left(\varepsilon_{i+1}-\varepsilon_i\right)^2$ ja järelikult, ligikaudu,

$$\begin{aligned} d &\approx \frac{(n-1)\left((\rho-1)^2\sigma^2+\sigma^2(1-\rho^2)\right)}{n\sigma^2} \\ &\approx (\rho-1)^2+(1-\rho^2) \\ &\approx 2-2\rho. \end{aligned}$$

Järelikult, kui Durbin-Watsoni statistik on ligikaudu 2, võiks mudeli eeldustega kõik korras olla. Kui aga statistiku väärtus on märkimisväärselt kahest väiksem, on põhjust muretsemiseks.

Muidugi tuleb meeles pidada, et Durbin-Watsoni testi on mõtet kasutada vaid siis, kui andmestik on eelnevalt järjestatud mingis mõtekas järjekorras — näiteks andmete kogumise/mõõtmise järjekorras vms.

7.5 Mudeli piisavuse test (Lack of fit test)

Vahel kerkib esile küsimus selle kohta, kas me oleme seost tunnuste vahel kirjeldanud piisavalt hästi. Kas mudel, mille abil kirjeldame y-tunnuse sõltuvust pidevast tunnusest x sobib? Kas mudel, kus modelleerime keskmist palka maakonna, inimese soo ja vanuseklassi abil on piisavalt hea?

Sellistele küsimustele vastamiseks tehakse sageli nn. mudeli piisavuse test. Mudeli piisavuse testi puhul võrreldakse kahte mudelit. Üks neist mudelitest on kõige täielikum, kõige rikkam mõeldav mudel. Teine aga meie mudel, see mudel, mida oleme otsustanud kasutada. Lineaarse regressioonmudeli korral võime näiteks võrrelda lineaarse regressioonimudelit sellise mudeliga, kus iga x-tunnuse väärtuse jaoks hinnatakse uus y-tunnuse keskmine. Selline mudelite võrdlemine on muidugi võimalik vaid siis, kui andmestikus vähemalt mõned x-tunnuse väärtused esinevad mitu korda. Kolme faktortunnust sisaldava mudeli puhul võiksime näiteks peamõjudega mudelit võrrelda mudeliga, kus on sees ka kõik mõeldavad koosmõjud (sugu*maakond, sugu*vanuseklass, vanuseklass*maakond, sugu*vanuseklass*maakond). Kui mudelite võrdlemiseks kasutatav F-test näitab, et lihtsam mudel on sama

hea kui kõige keerulisem mõeldav mudel, võime enda poolt valitud mudeliga rahule jääda. Kui aga keerukam mudel osutub tõestatavalt paremaks, siis peaksime oma mudeliga edasi töötama. Enamasti pole mõtekas siis kohe kõige keerukamat mõeldavat mudelit kasutama hakata — sageli piisab vaid oma mudeli väikesest parandamisest (näiteks lisame ruutliikme ka mudelisse või...).

Ülesanded

Hinnati kaks mudelit kasutades tudengite andmeid — prooviti prognoosida tudengite pikkuseid tudengi sugu ja kaalu kasutades (mudel m1) ja teise mudeli abil prognoositi tudengite kaalu kasutades tudengi sugu ja pikkust (mudel m2):

```
print(load(url("http://www.ms.ut.ee/mart/linmud2021/kysitlus.RData")))
m1= lm(pikkus~factor(sugu)+kaal, data=tudengid)
m2= lm(kaal~factor(sugu)+pikkus, data=tudengid)
```

Tee mõlema mudeli jaoks mudeli piisavuse test (ignoreeri ülejäänuid, mudelites mitteolevaid tunnuseid)! Milliste otsusteni jõuad?