Tartu Ülikool

Matemaatika-informaatikateaduskond

Matemaatilise statistika instituut

Ann-Mari Koppel

Determinatsioonikordaja ja prognoosikordaja

Bakalaureusetöö (6 EAP)

Juhendaja: Ene Käärik, PhD

Tartu 2014

Determinatsioonikordaja ja prognoosikordaja

Käesoleva töö eesmärk on anda ülevaade determinatsioonikordajast ja prognoosikordajast. Esimeses peatükis kirjeldatakse determinatsioonikordajat ja parandatud determinatsioonikordajat vabaliikmega ja vabaliikmeta mudeli korral. Lisatud on antud kordajate valemid tarkvarapaketis SAS. Peatüki lõpus on tarkvarapaketi SAS abil läbiviidud näited, mille juurde on lisatud ka mudeli valiku põhimõte. Teises peatükis antakse ülevaade prognoositud jääkidest ja prognoosikordajast. Juurde on lisatud rakendustarkvara SAS valemid ning nende näitajate paremaks mõistmiseks on lisatud näited.

Märksõnad: matemaatiline statistika, andmeanalüüs, statistilised mudelid, mudeli headuse hindamine, tarkvarapakett SAS.

Coefficient of determination and coefficient of prediction

The purpose of this thesis is to give an overview of coefficient of determination and coefficient of prediction. First section describes coefficient of determination and adjusted coefficient of determination in ordinary least-squares regression and in regression through the origin. Including formulas for those statistics in SAS software. At the end of the first section there are examples for choosing the best model. Examples were made by using SAS software. The second section describes *PRESS* statistic and coefficient of prediction. Second section also includes SAS formulas for those statistics and examples.

Keywords: mathematical statistics, data analysis, statistical models, goodness of fit, SAS software.

Sisukord

Sissejuhatus

Lineaarse mudeli headuse iseloomustamiseks on kasutusel üldiselt tuntud determinatsioonikordaja R^2 (*coefficient of determination*), mille arvutamisel on aluseks koguhajuvuse *SST* (*Sum of Squares Total*) jaotamine mudeli poolt kirjeldatud hajuvuseks *SSR* (*Sum of Squares Regression*) ja jääkhajuvuseks *SSE* (*Sum of Squares Error*). Determinatsioonikordaja

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

näitab kui suure osa sõltuva tunnuse *Y* koguhajuvusest (*SST* = *SSR* + *SSE*) mudel kirjeldab.

Kui determinatsioonikordajas asendada mudeli jääkide ruutude summa *SSE* prognoositud jääkide ruutude summaga *PRESS* (*Predicted Residual Sums of Squares*), siis saadakse prognoosikordaja (*coefficient of prediction*)

$$P^2 = 1 - \frac{PRESS}{SST},$$

mille abil hinnatakse mudeli prognoosi täpsust.

Käesoleva bakalaureusetöö eesmärk on anda ülevaade determinatsioonikordajast ja prognoosikordajast, nende omadustest ja kasutamisest statistikapaketis SAS.

Antud töö on jagatud kaheks osaks. Esimeses peatükis antakse ülevaade determinatsioonikordajast R^2 ja parandatud determinatsioonikordajast $\bar{R}^2$. Mõlemat kordajat uuritakse nii vabaliikmega mudeli, kui ka vabaliikmeta mudeli korral. Enamiku teadmiste ja tuletuskäikude puhul on refereeritud erinevaid allikaid, aga parandatud determinatsioonikordaja omaduse $\bar{R}^2 \leq 0$, kui $R^2 \leq \frac{p}{n-1}$, kus *n* on valimi maht ja *p* on mudeli parameetrite arv, tõestuseni jõudis autor iseseisvalt. Teises osas antakse ülevaade prognoosikordajast P^2. Kuna prognoosikordaja arvutamine põhineb *PRESS*-jääkidel, mis on omakorda võimalik saada kasutades mütsi-maatriksi *H* (*hat matrix*) peadiagonaalielemente h_{ii}, siis on eelnevalt antud ülevaade ka mütsi-maatriksist ja tema omadustest. Teises osas on autori poolt iseseisvalt tõestatud väite 2 esimene ja teine omadus ning väites 3 toodud mütsi-maatriksi *H* peadiagonaalielementide esitusviis.

Töö on kirjutatud tekstitöötlusprogrammiga Microsoft Office Word 2010. Näidete läbiviimiseks on kasutatud statistikapaketti SAS versioon 9.2.

Käesoleva töö autor tänab juhendajat Ene Käärikut paranduste, selgituste ning kasulike nõuannete eest.

1. Determinatsioonikordaja

1.1 Seos mitmese korrelatsioonikordajaga

Uuritava tunnuse $Y = (y_1, y_2, \dots, y_n)$ ja argumenttunnuse $X = (x_1,\ x_2, \dots, x_n)$ vahelise seose kirjeldamiseks saab kasutada lihtsat lineaarse regressiooni mudelit

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1.1}$$

kus β_0 on mudeli vabaliige, β_1 on regressioonikordaja ja ε on mudeli juhuslik viga.

Sõltuva tunnuse *Y* ja argumenttunnuse *X* vahelise seose tugevuse uurimiseks saab kasutada lineaarset korrelatsioonikordajat ehk Pearsoni korrelatsioonikordajat. See kordaja on määratud seosega

$$r(Y, X) = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{1.2}$$

kus x_i on argumenttunnuse $X = (x_1,\ x_2, \dots, x_n)$ i-s vaatlus, y_i on sõltuva tunnuse $Y = (y_1, y_2, \dots, y_n)$ i-s väärtus, $i = (1, 2, \dots, n)$ ning $\bar{x}$ ja $\bar{y}$ on vastavalt argumenttunnuse *X* ja sõltuva tunnuse *Y* keskväärtused. Valemi lugejas olev summa, mille liikmeteks on tunnuste keskväärtuste suhtes arvutatud hälvete korrutised, arvestab hajuvusdiagrammi kuju ning nimetajas olevad summad on vajalikud korrelatsioonikordaja normeerimiseks, et väärtused jääksid -1 ja 1 vahele.

Meil on vaja hinnata mudeli (1.1) parameetrid β_0 ja β_1. Parameetrite hindamiseks kasutame vähimruutude meetodit ehk üritame leida sirget, mille puhul vaatluste ja sirge vaheline vertikaalne kaugus oleks minimaalne. Selline sirge annab vaatlustele parima hinnangu ning seda nimetatakse regressioonisirgeks. Vertikaalseid kaugusi sirgest nimetatakse mudeli juhuslikuks veaks.

Mudeli (1.1) juhuslik viga avaldub kujul

$$\varepsilon = Y - \beta_0 - \beta_1 X.$$

Parameetrite β_0 ja β_1 hindamiseks peame minimeerima mudeli juhuslike vigade hajuvust. Selleks kasutame vähimruutude meetodit

$$SSE = \sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

kus $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$ on *i*-nda vaatluse y_i juhuslik viga.

Võttes eelmise võrduse viimasest osast osatuletised parameetrite β_0 ja β_1 järgi ja võrdsustades tuletised nulliga, saame avaldada

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \tag{1.3}$$

ning

$$\sum_{i=1}^{n} y_i x_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2. \tag{1.4}$$

Korrutades esmalt võrdust (1.3) argumenttunnuste summaga $\sum_{i=1}^{n} x_i$ ning avaldades võrdusest (1.4) korrutise $\hat{\beta}_0 \sum_{i=1}^{n} x_i$ ja asendades selle korrutis esimeses võrduses, saame

$$\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = n \sum_{i=1}^{n} x_i\, y_i - n\hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2 + \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i \right)^2,$$

millest saame avaldada parameetri β_1 vähimruutude hinnangu $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{1.5}$$

Parameetri β_0 vähimruutude hinnangu $\hat{\beta}_0$ saame avaldada võrdusest (1.3)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{1.6}$$

Hinnangu $\hat{\beta}_1$ anname enne hinnangut $\hat{\beta}_0$, sest kasutame hinnangus $\hat{\beta}_0$ hinnangut $\hat{\beta}_1$.

Regressioonisirge avaldub kujul

$$\hat{Y} = \hat{\beta}_0 - \hat{\beta}_1 X,$$

kus $\hat{Y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ on sõltuva tunnuse *Y* hinnang. Seosest (1.6) näeme, et regressioonisirge läbib punkti $(\bar{y}, \bar{x})$.

Andmestiku iga vaatluse y_i jaoks saame prognoosi

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad i = 1, 2, \dots, n,$$

kus *i*-s hinnang $\hat{y}_i$ on punkt regressioonisirgel, mis vastab *i*-ndale argumenttunnusele x_i.

Vertikaalne kaugus vaatluse y_i ja hinnangu $\hat{y}_i$ vahel avaldub kujul

$$e_i = y_i - \hat{y}_i, \qquad i = 1, 2, \dots, n.$$

See kaugus on mudeli (1.1) juhusliku vea ε hinnang ja seda kaugust e_i nimetatakse mudeli jäägiks.

Uurime hajuvust sõltuva tunnuse *Y* ja tema hinnangute vahel. Mida lähemal on vaatlused regressioonisirgele, seda suuremas lineaarses sõltuvuses on uuritav tunnus *Y* ja argument-tunnus *X*. Selle sõltuvuse tugevuse mõõtmiseks saab kasutada sõltuva tunnuse *Y* ja tema prognoosi $\hat{Y}$ vahelist korrelatsioonikordajat

$$r(Y, \hat{Y}) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{1.7}$$

kus $\hat{y}_i$ on uuritava tunnuse *Y* *i*-nda vaatluse prognoos. Sellist korrelatsioonikordajat nimetatakse mitmeseks korrelatsioonikordajaks.

Ühe argumenttunnusega mudeli sõltuva tunnuse *Y* ning argumenttunnuse *X* hajuvus-diagramm on identne uuritava tunnuse *Y* ja tema hinnangute $\hat{Y}$ hajuvusdiagrammiga, kui $r(X, Y) > 0$. Seega on nende tunnuste korrelatsioonikordajad seotud järgmiselt:

$$r(Y, \hat{Y}) = |r(X, Y)|.$$

Kuigi lihtsas regressioonanalüüsis pole uuritava tunnuse ning tema hinnangutevaheline hajuvus oluline annab see aimduse mudeli sobivuse kohta (Chatterjee ja Hadi, 2006). Lineaarses regressioonanalüüsis on korrelatsioonikordaja $r(Y, \hat{Y})$ seotud ühe teise mudeli sobivuse näitajaga – determinatsioonikordajaga.

1.2 Determinatsioonikordaja

Argumenttunnuse mõju kindlaks tegemine põhineb sõltuva tunnuse *Y* hajuvuse uurimisele valimis. Vaatluste varieerumist üldkeskmise ümber iseloomustab hälvete ruutude summa *SST*:

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

kus $(y_i - \bar{y})$ on iga vaatluse hajuvus keskväärtuse suhtes. Kui kõikidel vaatlustel y_i on sama väärtus, siis $SST = 0$, vastasel juhul on $SST > 0$.

Regressioonimudeli poolt kirjeldatud hajuvust kirjeldab hälvete ruutude summa *SSR*:

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

kus $(\hat{y}_i - \bar{y})$ on iga hinnangu hajuvus keskväärtuse suhtes.

Jääkide hajuvust kirjeldab hälvete ruutude summa *SSE*:

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

kus $(y_i - \hat{y}_i)$ on vaatluse hinnangu hajuvus vaatluse suhtes ehk jääk e_i.

Paneme tähele, et

$$\begin{aligned} SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \end{aligned} \tag{1.8}$$

Saadud summa keskmine liige on null

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

$$= 2\sum_{i=1}^{n}\big(\hat{y}_i(y_i - \hat{y}_i) - \bar{y}(y_i - \hat{y}_i)\big) = 2\sum_{i=1}^{n}\hat{y}_i e_i - 2\bar{y}\sum_{i=1}^{n} e_i = 0,$$

sest

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}\big(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\big) = \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1\sum_{i=1}^{n} x_i$$

$$= \sum_{i=1}^{n} y_i - n\bar{y} + n\hat{\beta}_1\bar{x} - \hat{\beta}_1\sum_{i=1}^{n} x_i = 0$$

ja

$$\sum_{i=1}^{n}\hat{y}_i e_i = \sum_{i=1}^{n}\big(\hat{\beta}_0 + \hat{\beta}_1 x_i\big)\,e_i = \hat{\beta}_0\sum_{i=1}^{n} e_i + \hat{\beta}_1\sum_{i=1}^{n} x_i e_i$$

$$= \sum_{i=1}^{n} x_i\big(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\big) = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i y_i = 0.$$

Seega saame hajuvustevahelise seose

$$SST = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SSR + SSE. \qquad (1.9)$$

Seosest (1.9) saame, et mudeli koguhajuvus *SST* on suurem või võrdne mudeli jääkide ruutude summaga *SSE*

$$SST \geq SSE$$

ja suurem või võrdne mudeli poolt kirjeldatud hajuvusega *SSR*

$$SST \geq SSR.$$

Vaatlus y_i võrdub tema hinnangu $\hat{y}_i$ ja vastava jäägi e_i summaga ehk

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

Lahutades selle võrduse mõlemast poolest uuritava tunnuse keskväärtuse $\bar{y}$ saame:

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

ehk vaatluse hajuvus keskväärtuse suhtes võrdub hinnangu hajuvusega keskväärtusest ja jäägi summaga. Seega seose (1.9) järgi saab uuritava tunnuse *Y* koguhajuvuse jagada kaheks liidetavaks - *SSR*, mis mõõdab uuritava tunnuse *Y* prognoosi ja argumenttunnuse *X* vastavust ning *SSE*, mis mõõdab prognoosiviga (Chatterjee ja Hadi, 2006).

Suhet

$$R^2 = \frac{SSR}{SST} \tag{1.10}$$

nimetatakse determinatsioonikordajaks.

Determinatsioonikordajat võib tõlgendada kui osa tunnuse *Y* koguhajuvusest, mis on kirjeldatud argumenttunnuse *X* poolt. Arvestades, et $SST = SSR + SSE$, saame

$$R^2 = 1 - \frac{SSE}{SST}. \tag{1.11}$$

Kuna kehtib seos

$$SST \geq SSE \geq 0,$$

siis determinatsioonikordaja R^2 korral kehtib seos

$$0 \leq R^2 \leq 1.$$

Kui determinatsioonikordaja väärtus jääb 1 lähedale, siis argumenttunnus *X* kirjeldab suure osa uuritava tunnuse *Y* hajuvusest.

Statistikut R^2 nimetatakse determinatsioonikordajaks, kuna see annab aimduse, kuidas argumenttunnus *X* määrab (ingl *determines*) sõltuva tunnuse *Y*.

Väide 1. Determinatsioonikordaja R^2 korral kehtib seos

$$[r(Y,X)]^2 = \left[r\left(Y,\hat{Y}\right)\right]^2 = R^2$$

ehk determinatsioonikordaja R^2 on võrdne uuritava tunnuse Y ja argumenttunnuse X vahelise korrelatsioonikordaja ruuduga või uuritava tunnuse Y ja tema hinnangute $\hat{Y}$ vahelise korrelatsioonikordaja ruuduga (Chatterjee ja Hadi, 2006).

Tõestus. Väite 1 tõestamisel kasutame Donald Wittmani tööd (Wittman, 2005). Võtame seosega (1.2) antud lineaarse korrelatsioonikordaja ruutu ning pärast seda korrutame lugejat ja nimetajat $\sum_{i=1}^{n}(x_i - \bar{x})^2$

$$[r(Y,X)]^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}$$
$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Arvestades seosega (1.5) antud vähimruutude hinnangut $\hat{\beta}_1$, saame

$$[r(Y,X)]^2 = \hat{\beta}_1 \hat{\beta}_1 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\hat{\beta}_1{}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}\hat{\beta}_1{}^2 (x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$
$$= \frac{\sum_{i=1}^{n}(\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Kuna i-nda vaatluse prognoos $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ ja arvestades seost (1.6), saame determinatsioonikordaja avaldada kujul

$$[r(Y,X)]^2 = \frac{\sum_{i=1}^{n}(\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST} = R^2.$$

Veel on vaja näidata, et kehtib ka $\left[r\left(Y,\hat{Y}\right)\right]^2 = R^2$. Selleks avaldame seosega (1.7) antud mitmese korrelatsioonikordaja

$$r\left(Y,\hat{Y}\right) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
$$= \frac{\sum_{i=1}^{n}[(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2]}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
$$= \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

mille ruutu võtmisel saame

$$\left[r\left(Y,\hat{Y}\right)\right]^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST} = R^2.$$

Järelikult kehtib seos $[r(Y,X)]^2 = \left[r\left(Y,\hat{Y}\right)\right]^2 = R^2$.

■

1.3 Vabaliikmeta mudeli determinatsioonikordaja

Eelnevalt uurisime vabaliikmega β_0 mudelit

$$Y = \beta_0 + \beta_1 X + \varepsilon\ ,$$

kuid vahel on vaja hinnata mudelit, milles vabaliige β_0 puudub

$$Y = \beta_1 X + \varepsilon. \tag{1.12}$$

Vabaliikmeta mudeleid kasutatakse harva ja nende kasutamine peab olema põhjendatud. Sellise mudeli kasutamine võib tuleneda mingist kindlast teoreetilisest kaalutlusest. Üksnes teadmisest, et sõltuv tunnus võrdub nulliga, kui argumenttunnus võrdub nulliga, ei piisa.

Analoogiliselt punktis 1.1 vabaliikmega mudeli parameetrite vähimruutude hinnangu leidmisele leiame vabaliikmeta mudeli (1.12) regressioonikordaja β_1 vähimruutude hinnangu. Selleks vaatame *i*-ndat vaatlust

$$y_i = \beta_1 x_i + \varepsilon_i\ ,$$

millest saame avaldada juhusliku vea ε_i

$$\varepsilon_i = y_i - \beta_1 x_i. \tag{1.13}$$

Hälvete (1.13) ruutude summa avaldub kujul

$$\sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n}(y_i - \beta_1 x_i)^2 = \sum_{i=1}^{n} {y_i}^2 - 2\beta_1 \sum_{i=1}^{n} x_i y_i + {\beta_1}^2 \sum_{i=1}^{n} {x_i}^2,$$

millest võttes tuletise parameetri β_1 järgi ning võrdsustades selle nulliga saame

$$2\sum_{i=1}^{n} x_i y_i - 2\hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2 = 0.$$

Seega vabaliikmeta mudeli (1.12) regressioonikordaja β_1 vähimruutude hinnang on

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} {x_i}^2}.$$

Prognoos *i*-ndale vaatlusele avaldub kujul

$$\hat{y}_i = \hat{\beta}_1 x_i, \qquad i = 1, 2, \dots, n$$

ning *i*-nda prognoosi jääk

$$e_i = y_i - \hat{y}_i, \qquad i = 1, 2, \dots, n. \tag{1.14}$$

Uurime seose (1.9) kehtivust vabaliikmeta mudeli korral. Selleks kontrollime seosega (1.8) antud summa keskmist liiget

$$\begin{aligned} 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= 2\sum_{i=1}^{n} \hat{y}_i e_i - 2\bar{y}\sum_{i=1}^{n} e_i = 2\sum_{i=1}^{n} \hat{\beta}_1 x_i e_i - 2\bar{y}\sum_{i=1}^{n} e_i \\ &= 2\hat{\beta}_1 \sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_1 x_i\right) - 2\bar{y}\sum_{i=1}^{n} e_i \\ &= 2\hat{\beta}_1 \sum_{i=1}^{n} x_i y_i - 2{\hat{\beta}_1}^2 \sum_{i=1}^{n} {x_i}^2 - 2\bar{y}\sum_{i=1}^{n} e_i = -2\bar{y}\sum_{i=1}^{n} e_i. \end{aligned}$$

Erinevalt vabaliikmega mudelist ei pruugi vabaliikmeta mudeli prognooside jääkide summa võrduda nulliga (Chatterjee ja Hadi, 2006). Seega me ei saa vabaliikmeta mudeli korral kasutada seost (1.9) ning seostega (1.10) ja (1.11) antud determinatsioonikordajaid. Mistõttu tuleb vabaliikmeta mudeli jaoks leida eraldi determinatsioonikordaja.

Vabaliikmeta mudel ei anna tavaliselt paremat hinnangut kui vabaliikmega mudel, sest regressioonisirge, mis läbib koordinaatide alguspunkti ei ole üldiselt kõige sobivam andmestiku kirjeldamiseks. Kasutades vabaliikmeta mudelit andmestiku puhul, kus ei esine punkti (0; 0), on regressioonisirge ikka sunnitud seda punkti läbima, mistõttu vaatluste varieerumine regressioonisirge ümber on suurem ning mudeli täpsus väheneb. Kui

regressioonisirge on sunnitud läbima punkti (0; 0), kuigi sellist punkti andmestikus pole, võib see põhjustada situatsiooni, kus regressioonisirge ei läbi punkti $(\bar{x}, \bar{y})$ (Eisenhauer, 2003). Mistõttu sobiva determinatsioonikordaja saamiseks võrdsustame seoses (1.9) sõltuva tunnuse Y keskväärtuse $\bar{y}$ nulliga ning saame

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

ehk arvestades seost (1.14)

$$\sum_{i=1}^{n} {y_i}^2 = \sum_{i=1}^{n} {\hat{y}_i}^2 + \sum_{i=1}^{n} {e_i}^2. \tag{1.15}$$

Ehk vabaliikmeta mudeli puhul jääb jääkide hajuvuse definitsioon $SSE_0 = \sum_{i=1}^{n} {e_i}^2$ samaks, aga koguhajuvus $SST_0 = \sum_{i=1}^{n} {y_i}^2$ ja mudeli poolt kirjeldatud hajuvus $SSR_0 = \sum_{i=1}^{n} {\hat{y}_i}^2$ on muutunud.

Kontrollime viimase seose kehtivust

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

Saadud summa keskmine liige on null, sest

$$2\sum_{i=1}^{n} \hat{y}_i e_i = 2\sum_{i=1}^{n} \hat{\beta}_1 x_i e_i = 2\hat{\beta}_1 \sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_1 x_i\right) = 2\hat{\beta}_1 \sum_{i=1}^{n} x_i y_i - 2{\hat{\beta}_1}^2 \sum_{i=1}^{n} {x_i}^2 = 0.$$

Seega seos (1.15) kehtib.

Suhet

$${R_{(0)}}^2 = \frac{\sum_{i=1}^{n} {\hat{y}_i}^2}{\sum_{i=1}^{n} {y_i}^2} \tag{1.16}$$

nimetatakse vabaliikmeta mudeli determinatsioonikordajaks.

Arvestades seost (1.15), saame

$$R_{(0)}{}^2 = \frac{\sum_{i=1}^{n} \hat{y}_i{}^2}{\sum_{i=1}^{n} y_i{}^2} = 1 - \frac{\sum_{i=1}^{n} e_i{}^2}{\sum_{i=1}^{n} y_i{}^2}.$$

Vabaliikmega mudeli korral on determinatsioonikordaja suhe prognoosi $\hat{y}$ ja vaatluse y hajuvustest sõltuva tunnuse keskväärtuse $\bar{y}$ ümber. Vabaliikmeta mudeli korral kirjeldavad determinatsioonikordaja lugeja ja nimetaja hajuvust nullpunkti ümber. Seetõttu pole seosega (1.10) antud vabaliikmega mudeli determinatsioonikordaja ning seosega (1.16) antud vabaliikmeta mudeli determinatsioonikordajad võrreldavad. Kuna vabaliikmeta mudelis on keskväärtus võetud võrdseks nulliga, siis võib vabaliikmeta mudeli determinatsioonikordaja väärtus olla suurem kui vabaliikmega determinatsioonikordaja väärtus, kuigi mudel ei pruugi teiste näitajate poolest parem olla (Myers, 1990).

Tarkvarapakett SAS kasutab determinatsioonikordaja arvutamiseks valemit

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

Vabaliikmeta mudeli korral arvutab SAS hälvete ruutude summa $SST_0 = \sum_{i=1}^{n} y_i{}^2$ ja vabaliikmega mudeli korral $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

1.4 Determinatsioonikordaja suurus

Vabaliikmega lihtsa lineaarse regressioonimudeli korral on determinatsioonikordajaks sõltuva tunnuse ja argumenttunnuse vahelise lineaarse korrelatsioonikordaja ruut, mis näitab, kui suurt osa uuritava tunnuse hajuvusest mudel kirjeldab. Kui mudelis on vabaliige ja seletavaid tunnuseid on rohkem kui üks, siis on determinatsioonikordajaks mitmese korrelatsioonikordaja ruut. Seega mõlemal juhul peavad determinatsioonikordaja R^2 väärtused jääma nulli ja ühe vahele.

Hea mudeli korral on vaatluste ja prognooside väärtused lähedased. Sel juhul on jääkide hajuvus $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ väike, mistõttu determinatsioonikordaja R^2 väärtus on lähedane ühele. Seega, mida lähemal on determinatsioonikordaja ühele, seda paremini kirjeldab mudel andmestikku. Kui kordaja väärtus on võrdne ühega, siis kirjeldab mudel andmestikku täielikult ehk kõik jäägid on võrdsed nulliga. Nullilähedane determinatsiooni-

kordaja väärtus viitab sellele, et regressioonisirge ei sobi andmestikuga ehk mudel ei kirjelda andmestikku hästi.

Kui sõltuva tunnuse *Y* ja argumenttunnuse *X* vahel puudub igasugune lineaarne suhe, siis annab mudel halva hinnangu. Sel juhul loetakse uuritava tunnuse parimaks hinnanguks tema keskväärtust $\bar{y}$, sest kasutades keskväärtust $\bar{y}$ saame väikseima hälvete ruutude summa. Seda hinnangut kasutatakse ainult juhul kui sõltuval tunnusel puudub igasugune seos argumenttunnusega. Seega saame seose puudumise korral determinatsioonikordaja R^2 väärtuseks nulli (Chatterjee ja Hadi, 2006).

Regressioonimudeli parameetrite hindamisel järgitakse vähimruutude printsiipi ehk mudeli parameetrite väärtused valitakse sellised, et erinevused tegelikult mõõdetud sõltuva tunnuse väärtuste ja mudeli järgi prognoositud väärtuste vahel oleksid minimaalsed. Seega kui mudelisse lisada argumente, siis jääkide ruutude summa *SSE* väheneb või jääb samaks ning seetõttu determinatsioonikordaja R^2 väärtus kasvab monotoonselt või jääb samaks (Myers, 1990). See on ühtlasi üks determinatsioonikordaja kasutamise puuduseid, sest determinatsioonikordaja R^2 väärtust on võimalik kunstlikult tõsta lisades mudelisse ebavajalikke argumente, mille tulemuseks on ülehinnatud mudel. Ülehinnatud mudel kirjeldab rohkem juhuslikku viga, kui uuritavat suhet. Seega determinatsioonikordaja väärtuse kasvamine ei viita sellele, et lisaargumenttunnus on oluline.

1.5 Parandatud determinatsioonikordaja

Kui mudeliga haaratud objektide arv on ligikaudselt võrdne argumenttunnuste arvuga, siis osutub sageli, et determinatsioonikordaja hindab regressiooniseost üle. See tuleneb sellest, et determinatsioonikordaja R^2 on nihkega hinnanguks vastavale üldkogumi determinatsioonikordajale, kusjuures nihe on seda suurem, mida väiksem on valimi maht n ja suurem on parameetrite arv p (Gayawan ja Ipinyomi, 2009). Nihke parandamise tulemusena saadakse parandatud determinatsioonikordaja $\bar{R}^2$, mis arvestab nii valimi mahtu n kui ka mudelis esinevate parameetrite arvu p. Parandatud determinatsioonikordajal $\bar{R}^2$ leidub mitmeid erinevaid kujusid. Erinevate kujude eristamine on keeruline, sest mõnel valemil on mitu erinevat nime ja mõni nimi on kasutusel mitme erineva valemi jaoks. Punktis 1.5

vaatleme põhjalikumalt laialdaselt levinud Ezekiel'i (tuntud ka Wherry ja McNemar'i nime all) parandatud determinatsioonikordaja valemit.

Anname seosega (1.11) antud determinatsioonikordajale R^2 kuju

$$R^2 = 1 - \frac{VAR_{err}}{VAR_{tot}},$$

kus $VAR_{err} = SSE/n$ ja $VAR_{tot} = SST/n$ on vastavalt valimi prognoosi jääkide hajuvus ja sõltuva tunnuse hajuvus, mida võib ühtlasi tõlgendada, kui nihkega hinnanguid üldkogumi jääkide ja sõltuva tunnuse hajuvusele. Nihketa hinnangud üldkogumi jääkide ja sõltuva tunnuse hajuvustele on vastavalt $VAR_{err} = SSE/(n-p-1)$ ja $VAR_{tot} = SST/(n-1)$.

Parandatud determinatsioonikordajaks $\bar{R}^2$ nimetatakse seost

$$\bar{R}^2 = 1 - \frac{(n-1)SSE}{(n-p-1)SST}, \tag{1.17}$$

kus n on valimi maht ja p parameetrite arv.

Arvestades seost (1.11) saame parandatud determinatsioonikordajale (1.17) anda kuju

$$\begin{aligned}\bar{R}^2 &= 1 - \frac{n-1}{n-p-1}(1-R^2) = 1 - \frac{n-p+p-1}{n-p-1}(1-R^2)\\ &= 1 - \left(1 + \frac{p}{n-p-1}\right)(1-R^2) = 1 - (1-R^2) - \frac{p}{n-p-1}(1-R^2)\\ &= R^2 - \frac{p}{n-p-1}(1-R^2).\end{aligned}$$

Järelikult saab parandatud determinatsioonikordaja anda seosega

$$\bar{R}^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) = R^2 - \frac{p}{n-p-1}(1-R^2). \tag{1.18}$$

Parandatud determinatsioonikordaja leidmisel kasutatakse vabadusastmete arvu, seega on tegemist pigem ruutkeskmiste kui ruutude summa suhtega.

Ruutkeskmine viga *MSE* (*Mean Squared Error*) avaldub kujul

$$MSE = \frac{SSE}{n-p-1},$$

mis annab hinnangu juhuslike vigade hajuvusele. Võttes ruutkeskmisest veast ruutjuure, saame mudeli standardhälbe ehk mudeli täpsuse $\sqrt{MSE}$.

Sõltuva tunnuse koguhajuvuse ruutkeskmine *MST* (*Mean Squared Total*) avaldub kujul

$$MST = \frac{SST}{n-1}.$$

Seega saame seosega (1.17) antud parandatud determinatsioonikordaja avaldada kujul

$$\bar{R}^2 = 1 - \frac{MSE}{MST}.$$

Parandatud determinatsioonikordajat $\bar{R}^2$ saab kasutada erineva argumenttunnuste arvuga mudelite võrdlemiseks. Erinevalt tavalisest determinatsioonikordajast ei saa parandatud determinatsioonikordajat interpreteerida kui argumenttunnuse poolt kirjeldatud sõltuva tunnuse varieeruvuse osa. Parandatud determinatsioonikordajat $\bar{R}^2$ võib interpreteerida kui kordajat, mille abil kontrollitakse mudeli alternatiivse kuju sobivust.

Parandatud determinatsioonikordaja $\bar{R}^2$ väärtus on determinatsioonikordaja R^2 väärtusest alati väiksem või võrdne sellega. Kuna determinatsioonikordaja R^2 korral kehtis omadus $0 \leq R^2 \leq 1$ ehk $0 \leq 1 - R^2$ ja arvestades seost (1.18), siis $\bar{R}^2 \leq R^2 \leq 1$ ning $\bar{R}^2 = R^2$, kui $R^2 = 1$ või parameetrite arv $p = 0$ (Dufour, 2011).

Heaks mudeliks loetakse mudelit, mille korral kehtib seos $R^2 \approx \bar{R}^2$.

Parandatud determinatsioonikordaja $\bar{R}^2$ on väiksem või võrdne nulliga, kui determinatsioonikordaja $R^2 \leq \frac{p}{n-1}$ (Dufour, 2011). Selle näitamiseks avaldame seose (1.18) viimase osa

$$\bar{R}^2 = R^2 - \frac{p}{n-p-1}(1-R^2) = R^2 - \frac{p}{n-p-1} + \frac{pR^2}{n-p-1}$$
$$= \frac{R^2(n-p-1) + pR^2 - p}{n-p-1} = \frac{R^2(n-1)-p}{n-p-1},$$

millest näeme, et $\bar{R}^2 \leq 0$, kui $R^2 \leq \frac{p}{n-1}$.

Teist laialdaselt levinud parandatud determinatsioonikordaja kuju kasutatakse rakendustarkvaras SAS

$$\bar{R}^2 = 1 - \frac{(n-i)(1-R^2)}{n-p},$$

kus n on vaatluste arv, p on mudeli parameetrite arv ning *i* on indikaatortunnus, $i = 1$, kui mudelis on vabaliige ja $i = 0$, kui mudel on vabaliikmeta. See valem on samuti tuntud, kui Wherry parandatud determinatsioonikordaja.

Näide 1. Vabaliikmeta ja vabaliikmega mudeli jaoks on erinevad determinatsioonikordajad. Uurime mõlemat mudelit tarkvarapaketi SAS abil. Kasutame tarkvarapaketis SAS leiduvat andmestikku *class*, mis sisaldab 11-16-aastaste laste pikkuseid ja kaale. Uurime, kuidas avaldub lapse kaal (kg) tema kasvu (cm) kaudu. Vabaliikmega mudeliks saame

$$kaal = -64.877 + 0.696kasv,$$

mille determinatsioonikordaja $R^2 = 0.7705$ ja parandatud determinatsioonikordaja $\bar{R}^2 = 0.7570$, mudeli täpsus $\sqrt{MSE} = 5.092$ kg (Lisa 1). Vabaliikme olulisustõenäosus $p = 0.0004$ ja regressioonikordaja olulisustõenäosus $p < 0.0001$, seega on nii vabaliige kui ka regressioonikordaja statistiliselt olulised.

Eemaldades mudelist vabaliikme saame mudeli kujul

$$kaal = 0.289kasv,$$

mille determinatsioonikordaja $R^2 = 0.9768$ ja parandatud determinatsioonikordaja $\bar{R}^2 = 0.9756$ (Lisa 2). Regressioonikordaja olulisustõenäosus on $p < 0.0001$ ning mudeli täpsus $\sqrt{MSE} = 7.265$ kg.

⬚

Võrreldes vabaliikmega ja vabaliikmeta mudelite determinatsioonikordajaid näeme, et vabaliikmega mudel kirjeldab andmestikku vähem, kui vabaliikmeta mudel, samas on vabaliikmega mudel täpsem. Parandatud determinatsioonikordaja $\bar{R}^2$ on mõlemal juhul

väiksem kui determinatsioonikordaja R^2. Kuna vabaliikmega ja vabaliikmeta mudeli determinatsioonikordajad arvutatakse erinevalt, siis ei saa me nende võrdluse põhjal otsustada, millist mudelit kasutada.

Nagu eelnevalt mainitud peab vabaliikmeta mudeli kasutamine olema põhjendatud. Mõnel juhul on raske otsustada, kumma mudeli kasutamine on õige. Otsuse langetamiseks võib kasutada erinevaid näitajaid. Esiteks võib võrrelda mudelite täpsust $\sqrt{MSE}$. Näites 1 on vabaliikmega mudel täpsem kui vabaliikmeta mudel. Teiseks võib uurida vabaliikmega mudeli vabaliikme olulisust. Kui vabaliige on statistiliselt oluline, siis on soovitatav kasutada vabaliikmega mudelit. Näites 1 antud vabaliikmega mudeli vabaliikme olulisustõenäosus on $p = 0.0004$ ehk vabaliige on statistiliselt oluline. Seega on antud andmestiku abil laste kaalu arvutamisel õigem kasutada vabaliikmega mudelit.

Mitme argumendiga mudeli korral hinnatakse mudeli headust analoogiliselt ühe argumendiga mudelile. Mudelisse tunnuste lisamisel reageerib parandatud determinatsioonikordaja lisatud tunnusele determinatsioonikordajast erinevalt. Kui mudelisse lisada tunnuseid, siis determinatsioonikordaja väärtus suureneb või jääb samaks, aga parandatud determinatsioonikordaja väärtus suureneb ainult siis, kui lisatud tunnused on olulised.

Näide 2. Uurime sportlaste füüsilist võimekust leides mudeli nende hapniku tarbimisele jooksu ajal (Lisa 3). Algselt uurime, kuidas mõjutavad pulss ja jooksja vanus hapniku tarbimist jooksu ajal:

$$hapnik = 121.376 - 0.294pulss - 0.507vanus.$$

Mudeli determinatsioonikordaja $R^2 = 0.3760$ ja parandatud determinatsioonikordaja $\bar{R}^2 = 0.3314$. Vabaliikme olulisustõenäosus on $p < 0.0001$ ning regressioonikordajate olulisustõenäosused on vastavalt $p = 0.0013$ ja $p = 0.0041$, seega on nii vabaliige, kui ka regressioonikordajad statistiliselt olulised. Mudeli täpsus $\sqrt{MSE} = 4.356$ ml/min/kg.

Uurime, kas mudelisse tunnuseid lisades on võimalik mudelit paremaks muuta. Lisame eelmisesse mudelisse jooksule kulunud aja (min). Saame hapniku tarbimise mudeli

$$hapnik = 111.718 - 0.131pulss - 0.256vanus - 2.825aeg.$$

Selle mudeli determinatsioonikordaja $R^2 = 0.8111$ ja parandatud determinatsioonikordaja $\bar{R}^2 = 0.7901$. Võrreldes eelneva mudeliga on mõlemad determinatsioonikordajad suurenenud, seega on hapniku tarbimise arvutamisel jooksu aeg oluline tunnus. Seda fakti kinnitab ka tunnuse „aeg“ olulisustõenäosus $p < 0.0001$. Vabaliikme ja teiste regressioonikordajate olulisustõenäosused on vastavalt $p < 0.0001$, $p = 0.0154$ ja $p = 0.0129$. Mudeli täpsus $\sqrt{MSE} = 2.441$ ml/min/kg. Seega sobib hapniku tarbimise hindamiseks paremini teisena koostatud mudel.

Uurime, kas mudelit on võimalik veel paremaks muuta. Lisame eelmisesse mudelisse sportlaste kaalud (kg). Saame hapniku tarbimise mudeli

$$hapnik = 115.662 - 0.129pulss - 0.276vanus - 2.772aeg - 0.049kaal,$$

mille determinatsioonikordaja $R^2 = 0.8165$ ja parandatud determinatsioonikordaja $\bar{R}^2 = 0.7883$, mudeli täpsus $\sqrt{MSE} = 2.451$ ml/min/kg. Paneme tähele, et võrreldes eelmise mudeliga on determinatsioonikordaja suurenenud, aga parandatud determinatsioonikordaja on vähenenud. Seega on hapniku tarbimise arvutamisel sportlaste kaalud ebaolulised. Seda kinnitab ka tunnuse „kaal“ olulisustõenäosus $p = 0.3898$.

Hapniku tarbimise hindamiseks sobib kõige paremini teine mudel, sest see kirjeldab andmestikku kõige rohkem ja on kõige täpsem.

2. Prognoosikordaja

2.1 Prognoos ja mütsi-maatriks

Mitme argumendiga lineaarse mudeli matemaatiline esitus antakse tavaliselt maatrikskujul. Maatrikskujul avaldub lineaarne regressioonimudel järgmiselt

$$Y = X\beta + \varepsilon,$$

kus Y on n-mõõtmeline funktsioontunnuse vektor, β on p-mõõtmeline tundmatute parameetrite vektor, ε on n-mõõtmeline juhuslike vigade vektor ning X on nxp-mõõtmeline plaanimaatriks.

Juhuslike vigade ruutude summad (*SSE*) avalduvad kujul

$$SSE(\beta) = (Y - X\beta)^T(Y - X\beta).$$

Vähimruutude printsiibi realiseerimiseks tuleb minimeerida vigade ruutude summad. Selleks on vaja leida parameetri β hinnangu $\hat{\beta}$ väärtus, mis minimeeriks

$$SSE(\beta) = (Y - X\beta)^T(Y - X\beta) = (Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta).$$

Kuna $Y^TX\beta = \beta^TX^TY$, siis saame

$$SSE(\beta) = Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta.$$

Võttes parameetri β järgi tuletise ja võrdsustades tuletise nulliga saame avaldada normaalvõrrandisüsteemi

$$(X^TX)\hat{\beta} = X^TY,$$

millel on ühene lahend parajasti siis, kui maatriksi veerud on lineaarselt sõltumatud ehk kui leidub pöördmaatriks $(X^TX)^{-1}$ (Chatterjee ja Hadi, 2006).

Saame, et parameetri β hinnang

$$\hat{\beta} = (X^TX)^{-1}X^TY.$$

Mudeli põhjal arvutatud sõltuva tunnuse väärtust nimetatakse prognoosiks

$$\hat{Y} = X\hat{\beta}.$$

Anname hinnangud mudeli juhuslikele vigadele ehk leiame prognoosijäägid. Arvestades prognoosi $\hat{Y}$ ja hinnangu $\hat{\beta}$ avaldisi, saame mudeli jäägi

$$e = Y - \hat{Y} = Y - X(X^TX)^{-1}X^TY = (I - X(X^TX)^{-1}X^T)Y,$$

kus I on nxn ühikmaatriks ja maatriksit $H = X(X^TX)^{-1}X^T$ nimetatakse nxn mütsi-maatriksiks (*hat matrix*).

Väide 2. Mütsi-maatriksile kehtivad järgmised omadused:

1) maatriks H on idempotentne ehk $HH=H$;
2) maatriks H on sümmeetriline ehk $H^T = H$;
3) maatriks H on positiivselt poolmääratud ehk kehtib seos $w^THw \geq 0$ iga w korral, kus w on nx1 mittenulliline vektor.

Tõestus.

1) näitame kõigepealt, et maatriks H on idempotentne:

$$HH = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T$$
$$= X(X^TX)^{-1}X^T = H.$$

2) näitame, et maatriks H on sümmeetriline:

$$H^T = [X(X^TX)^{-1}X^T]^T = X[(X^TX)^{-1}]^TX^T = X(X^TX)^{-1}X^T = H.$$

3) näitame, et maatriks H on positiivselt poolmääratud.

Kuna maatriks H on sümmeetriline, siis on see positiivselt poolmääratud, kui mõni selle omaväärtus on võrdne nulliga ning ülejäänud omaväärtused on positiivsed. Skalaari λ nimetatakse nxn maatriksi H omaväärtuseks, kui leidub selline nx1 mittenulliline vektor w, mis rahuldab võrdust

$$Hw = \lambda w. \tag{2.1}$$

Korrutame seosega (2.1) antud võrdust maatriksiga H

$$HHw = \lambda Hw$$

$$HHw = \lambda\lambda w.$$

Arvestades, et maatriks H on idempotentne, saame

$$H^2w = Hw = \lambda w$$

ja

$$\lambda^2 w = \lambda w.$$

Saame maatriksi H omaväärtusteks $\lambda_1 = 0$ ja $\lambda_2 = 1$, järelikult on maatriks H positiivselt poolmääratud.

■

2.2 Prognoositud jäägid, *PRESS*-statistik

Determinatsioonikordajat kasutatakse mudeli headuse määramiseks. Samas ei anna determinatsioonikordaja ülevaadet ühe kindla vaatluse potentsiaalsest mõjust prognoosile. Mudeli prognoosimisvõime uurimiseks kasutatakse *PRESS*-statistikut (*Predicted Residual Sums of Squares*).

Prognoositud jäägid leitakse kui vahe tegeliku väärtuse ja ilma *i*-nda vaatluseta prognoositud väärtuse vahel. Olgu ilma *i*-nda vaatluseta prognoositud väärtus $\hat{y}_{(i)}$, mis on arvutatud hinnangu $\hat{Y} = X\hat{\beta}$ abil. Prognoositud ehk *PRESS*-jäägid avalduvad kujul

$$e_{(i)} = y_i - \hat{y}_{(i)},$$

kus *i*-nda vaatluse välja jätmisel saadud prognoos $\hat{y}_{(i)}$ on sõltumatu y_i-st, sest vaatlus y_i ei ole kasutuses regressioonimudeli hindamisel.

Kui prognoositud jääk $e_{(i)}$ on negatiivne, siis mudel ülehindab seost. Kui jääk $e_{(i)}$ on positiivne, siis mudel alahindab seost (Mendez, 2008).

PRESS-jääkide $e_{(i)}$ leidmiseks on olemas ka lihtsam avaldis, mille jaoks ei ole vaja mudelit pärast iga vaatluse eemaldamist uuesti hinnata

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}, \tag{2.2}$$

kus e_i on hinnangu $\hat{Y} = X\hat{\beta}$ prognoosijäägid ja h_{ii} on mütsi-maatriksi peadiagonaali elemendid, $i = 1, 2, \dots, n$. Paneme tähele, et prognoositud ehk *PRESS*-jäägid on kaalutud vähimruutude jäägid, kaaluga $1/(1 - h_{ii})$ (Landram, 2005).

Väide 3. Mütsi-maatriksi *H* peadiagonaali elemendid avalduvad kujul

$$h_{ii} = \sum_{j=1}^{n} {h_{ij}}^2 = {h_{ii}}^2 + \sum_{j \neq i} {h_{ij}}^2, \qquad i = 1, 2, \dots, n.$$

Tõestus. Kuna maatriks *H* on sümmeetriline, siis saame anda selle kujul

$$H = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{12} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{1n} & h_{2n} & \cdots & h_{nn} \end{pmatrix}$$

ja korrutades maatriksit *H* iseendaga saame

$$HH = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{12} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{1n} & h_{2n} & \cdots & h_{nn} \end{pmatrix} \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{12} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{1n} & h_{2n} & \cdots & h_{nn} \end{pmatrix}$$

$$= \begin{pmatrix} {h_{11}}^2 + \cdots + {h_{1n}}^2 & h_{11}h_{12} + \cdots + h_{2n}h_{1n} & \cdots & h_{11}h_{1n} + \cdots + h_{1n}h_{nn} \\ h_{11}h_{12} + \cdots + h_{2n}h_{1n} & {h_{12}}^2 + \cdots + {h_{2n}}^2 & \cdots & h_{12}h_{1n} + \cdots + h_{2n}h_{nn} \\ \vdots & \vdots & \ddots & \vdots \\ h_{11}h_{1n} + \cdots + h_{1n}h_{nn} & h_{12}h_{1n} + \cdots + h_{2n}h_{nn} & \cdots & {h_{1n}}^2 + \cdots + {h_{nn}}^2 \end{pmatrix}.$$

Arvestades maatriksi *H* idempotentsust näeme, et maatriks *H* peadiagonaali elemendid avalduvad kujul

$$h_{ii} = \sum_{j=1}^{n} {h_{ij}}^2 = {h_{ii}}^2 + \sum_{j \neq i} {h_{ij}}^2, \qquad i = 1, 2, \dots, n.$$

■

Väide 4. Maatriksi *H* peadiagonaali elementidele h_{ii} kehtib järgmine omdadus

$$0 \leq h_{ii} \leq 1.$$

Tõestus. Näitame kõigepealt, et maatriksi *H* peadiagonaali elementidele kehtib omadus $0 \leq h_{ii}$. Selle tõestamiseks kasutame maatriksi *H* positiivse poolmääratuse omadust. Olgu *n*-mõõtmeline vektor a_i, mille kõik elemendid peale *i*-nda on nullid, *i*-s element on üks. Korrutame mütsi-maatriksit *H* vasakult vektori a_i transponeeritud kujuga ning paremalt vektoriga a_i, saame

$$a_i{}^T H a_i = h_{ii},$$

mis on nullist suurem või võrdne, sest maatriks *H* on positiivselt poolmääratud.

Nüüd näitame, et kehtib ka teine pool võrratusest ehk $h_{ii} \leq 1$. Olgu lisaks maatriksile *H* nxn ühikmaatriks *I* ning nxn maatriks $M = (I - H)$. Avaldame maatriksi *H* peadiagonaali elemendid kujul

$$h_{ii} = a_i{}^T H a_i = a_i{}^T (I - M) a_i = a_i{}^T a_i - a_i{}^T M a_i = 1 - a_i{}^T M a_i.$$

Maatriks *H* on positiivselt poolmääratud, seega on ka maatriks *M* positiivselt poolmääratud. Kuna maatriks *M* on positiivselt poolmääratud, siis $a_i{}^T M a_i \geq 0$, mistõttu $1 - a_i{}^T M a_i \leq 1$.

Järelikult kehtib mütsi-maatriksi *H* peadiagonaali elementidele omadus $0 \leq h_{ii} \leq 1$ iga $i = 1, 2, \dots, n$ korral.

■

Prognoositud jääkide korral kasutatavat ühe vaatluse väljajätmise protsessi korratakse kõigi vaatluste jaoks. Võttes saadud prognoositud jäägid ruutu ja siis kõiki prognoositud jääkide ruute summeerides saame prognoositud jääkide ruutude summa ehk *PRESS*-statistiku

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{n} e_{(i)}{}^2. \qquad (2.3)$$

PRESS simuleerib prognoosi, jättes välja vaatluse, mida prognoositakse, ning mõõdab, kui hästi mudeli prognoos $\hat{y}_{(i)}$ suudab prognoosida vaatlust y_i. Mida väiksem on *PRESS*-statistiku väärtus, seda paremini regressioonimudel prognoosib.

Tarkvarapakett SAS kasutab prognoositud jääkide arvutamisel valemit

$$e_{(i)} = \frac{e_i}{1 - h_{ii}} \tag{2.4}$$

ja *PRESS*-statistiku arvutamisel valemit (2.3).

2.3 Prognoosikordaja P^2

Regressioonanalüüsis kasutatakse prognoositud jääke, et hinnata, kuidas mudel prognoosib sõltuvat tunnust ilma *i*-ndat vaatlust arvestamata. Prognoositud jäägi $e_{(i)}$ leidmisel vajalik ilma *i*-nda vaatluseta prognoositud väärtus $\hat{y}_{(i)}$ on sõltumatu vaatlusest y_i, kuna $\hat{y}_{(i)}$ arvutamisel pole kasutatud vaatlust y_i. Seega on *PRESS*-statistik

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{n} {e_{(i)}}^2$$

hea hindamaks regressioonimudeli valiidsust ja prognoosimisvõimet. Näeme, et *PRESS*-statistik sarnaneb regressioonanalüüsi jääkide summa ruudule

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

Kui jääkide hajuvus *SSE* kasutab hinnatud väärtusi $\hat{y}_i$, siis *PRESS*-statistik kasutab ilma *i*-nda vaatluseta hinnatud väärtusi $\hat{y}_{(i)}$. Arvestades seost (2.4) näeme, et *PRESS*-statistiku väärtus on alati suurem, kui mudeli jääkide ruutude summa *SSE*, aga nende väärtused peaksid olema ligilähedased. Kui *PRESS*-statistiku väärtus on kordades suurem kui mudeli jääkide ruutude summa, siis pole mudel valiidne ehk mudel ei mõõda seda, mida ta on määratud mõõtma. Kui asendame determinatsioonikordajas

$$R^2 = 1 - \frac{SSE}{SST}$$

mudeli jääkide ruutude summa *PRESS*-statistikuga saame leida prognoosikordaja.

Prognoosikordajaks P^2 nimetatakse suurust

$$P^2 = 1 - \frac{PRESS}{SST},$$

kus $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ ja *PRESS* on antud seosega (2.3).

Prognoosikordaja sarnaneb oma kujult seosega (1.11) antud determinatsioonikordajale, kuid erinevalt determinatsioonikordajast ei mõõda prognoosikordaja mudeli sobivust vaid mudeli prognoosivõimet.

Prognoosikordaja P^2 väärtust tarkvarapakett SAS ei arvuta, kuid selle leidmine pole väga keerukas, sest mudeli moodustamisel väljastab SAS vajalikud andmed (*PRESS*-statistiku ja mudeli koguhajuvuse *SST*) prognoosikordaja arvutamiseks.

2.4 Prognoosikordaja suurus

Prognoositud jäägid $e_{(i)}$, mis on antud seosega (2.2) on saadud kasutades vähimruutude jääke e_i ja mütsi-maatriksi H peadiagonaalielemente h_{ii}. Mütsi-maatriks on ühtlasi projektsioonimaatriks, mistõttu mudelisse argumenttunnuste lisamisel tavaliselt peadiagonaali elementide väärtused suurenevad (Landram, 2005).

Kui argumenttunnuse lisamisel prognoosikordaja P^2 väärtus väheneb, siis on see märk sellest, et mudelisse on lisatud ebavajalik tunnus ja mudel hakkab ülehindama. Seega lisades argumenttunnuseid võib see põhjustada prognoosi täpsuse kahanemise.

Prognoosikordaja P^2 leidmisel pole mudeli tuletamisel kasutatud kõiki vaatlusi, mistõttu sõltumatud, ühe vaatluse välja jätmisel saadud hinnangud $\hat{y}_{(i)}$ ei ole nii täpsed vaatluse y_i prognoosimisel, kui kõikide vaatluste kaasamisel saadud hinnangud. Seega on tavaliselt prognoosikordaja väärtus väiksem kui determinatsioonikordaja R^2 ja parandatud determinatsioonikordaja $\bar{R}^2$ väärtused. Lisaks on *PRESS*-jäägid $e_{(i)} = e_i/(1 - h_{ii})$ kaalutud vähimruutude jäägid, mistõttu $e_{(i)} > e_i$ ning sellest omakorda saame, et $PRESS > SSE$ ja $P^2 < R^2$. Seega prognoosikordaja väärtus ei saa ületada kõiki vaatlusi hõlmavate determinatsioonikordajate R^2 ja $\bar{R}^2$ väärtusi.

Halva mudeli korral võib prognoosikordaja P^2 omandada negatiivse väärtuse, sel juhul on *PRESS*-statistiku väärtus suurem mudeli koguhajuvus *SST*.

Näide 3. Näite läbiviimiseks kasutame sama andmestikku, mis näite 1 puhul. Laste kaalu mudeli

$$kaal = -64.877 + 0.696kasv,$$

jääkide ruutude summa $SSE \approx 440.823$ ja prognoositud jääkide ruutude summa $PRESS \approx 545.524$ (Lisa 1). Kuna *PRESS*-statistiku väärtus ei erine kordades jääkide hajuvusest *SSE* väärtusest, siis võib öelda, et mudel on valiidne ehk mudel mõõdab seda, mida ta on määratud mõõtma. Mudeli koguhajuvus $SST \approx 1920.855$, seega mudeli prognoosikordaja $P^2 \approx 0.7160$, mis on väiksem näites 1 arvutatud determinatsioonikordajast ja parandatud determinatsioonikordajast (vastavalt $R^2 = 0.7705$ ja $\bar{R}^2 = 0.7570$).

⬚

Nagu eelnevalt mainitud, kui argumenttunnuse lisamisel prognoosikordaja P^2 väärtus väheneb, siis on see märk sellest, et mudelisse on lisatud ebavajalik tunnus ja mudel hakkab ülehindama. Näites 4 uurime prognoosikordaja käitumist mudelisse tunnuste lisamisel.

Näide 4. Uurime näites 2 antud mudelite prognoosikordajaid. Esimese mudeli

$$hapnik = 121.376 - 0.294pulss - 0.507vanus$$

jääkide ruutude hajuvus $SSE \approx 531.266$ ja prognoositud jääkide ruutude summa $PRESS \approx 658.199$ (Lisa 3). Kuna *PRESS*-statistiku väärtus ei erine kordades jääkide hajuvusest *SSE* väärtusest, siis võib öelda, et mudel on valiidne. Mudeli koguhajuvus $SST \approx 851.382$, seega mudeli prognoosikordaja $P^2 \approx 0.2269$, mis on väiksem näites 2 arvutatud determinatsioonikordajast ja parandatud determinatsioonikordajast (vastavalt $R^2 = 0.3760$ ja $\bar{R}^2 = 0.3314$**)**.

Näite 2 teise mudeli

$$hapnik = 111.718 - 0.131pulss - 0.256vanus - 2.825aeg,$$

prognoositud jääkide ruutude summa $PRESS \approx 205.125$, mudel on valiidne. Mudeli prognoosikordaja $P^2 \approx 0.7590$, mis on eelmise mudeli prognoosikordajast suurem, see viitab asjaolule, et lisatud tunnus on oluline. Mudeli prognoosikordaja on väiksem näites 2 arvutatud determinatsioonikordajast ja parandatud determinatsioonikordajast (vastavalt $R^2 = 0.8111$ ja $\bar{R}^2 = 0.7901$).

Kolmanda mudeli

$$hapnik = 115.662 - 0.129pulss - 0.276vanus - 2.772aeg - 0.049kaal$$

prognoositud jääkide ruutude summa $PRESS \approx 212.272$. Mudeli prognoosikordaja $P^2 \approx 0.7506$. Selle mudeli prognoosikordaja on eelmise mudeli omast väiksem, mis viitab, et lisatud tunnus on ebaoluline.

Kokkuvõte

Käesolevas bakalaureusetöös vaatlesime determinatsioonikordajat ja prognoosikordajat, nende omadusi ja rakendamist statistikapaketis SAS.

Töö esimeses peatükis kirjeldasime determinatsioonikordajat R^2 ja parandatud determinatsioonikordajat $\bar{R}^2$. Determinatsioonikordajat tõlgendatakse kui osa uuritava tunnuse *Y* koguhajuvusest, mis on kirjeldatud mudeli poolt.

Vabaliikmega ja vabaliikmeta mudeli jaoks eristatakse erinevaid determinatsioonikordajaid. Vabaliikmega mudeli jaoks on determinatsioonikordaja R^2 antud kujul

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST},$$

kus *SST* on vaatluste varieerumine üldkeskmise ümber, *SSR* on regressioonimudeli poolt kirjeldatud hajuvus ja *SSE* on vea poolt kirjeldatud hajuvus. Vabaliikmeta mudeli jaoks on determinatsioonikordaja antud kujul

$$R_{(0)}{}^2 = \frac{\sum_{i=1}^{n} \hat{y}_i{}^2}{\sum_{i=1}^{n} y_i{}^2} = 1 - \frac{\sum_{i=1}^{n} e_i{}^2}{\sum_{i=1}^{n} y_i{}^2},$$

kus x_i on argumenttunnuse $X = (x_1, x_2, \dots, x_n)$ i-s vaatlus, y_i on sõltuva tunnuse $Y = (y_1, y_2, \dots, y_n)$ i-s väärtus, $\hat{y}_i$ i-nda vaatluse prognoos ning e_i on i-nda prognoosi jääk, $i = (1, 2, \dots, n)$.

Determinatsioonikordaja R^2 on nihkega hinnang vastavale üldkogumi determinatsioonikordajale. Nihke parandamise tulemusena saadakse parandatud determinatsioonikordaja $\bar{R}^2$, mis võtab arvesse nii valimi mahu n, kui ka mudelis esinevate parameetrite arvu p. Parandatud determinatsioonikordaja on antud kujul

$$\bar{R}^2 = 1 - \frac{(n-1)SSE}{(n-p-1)SST}.$$

Parandatud determinatsioonikordajat saab kasutada erineva argumenttunnuste arvuga mudelite võrdlemiseks ning selle väärtus on determinatsioonikordaja väärtusest alati väiksem või võrdne sellega.

Töö teises peatükis kirjeldasime prognoosikordajat P^2. Determinatsioonikordaja arvutamisel kasutatud jääkide ruutude summa *SSE* asendamisel ühe vaatluse välja jätmisel saadud prognoositud jääkide $e_{(i)}$ ruutude summaga ehk *PRESS*-statistikuga, saame prognoosikordaja

$$P^2 = 1 - \frac{PRESS}{SST}.$$

Prognoosikordaja kasutab kaalutud vähimruutude jääke, et minimeerida *PRESS*-jääkide ruutude summat. Prognoosikordaja sarnaneb oma kujult determinatsioonikordajale, kuid erinevalt determinatsioonikordajast ei mõõda prognoosikordaja mudeli sobivust vaid mudeli prognoosivõimet.

Kasutatud kirjandus

1) Chatterjee, S. ja Hadi, Ali S., 2006. Regression analysis by example. – 4th ed. New Jersey: John Wiley & Sons, lk 40-62, lk 82-90.
2) Dufour, J-M., 2011. Coefficient of determination. McGill University. Allikas: http://www2.cirano.qc.ca/~dufourj/Web_Site/ResE/Dufour_1983_R2_W.pdf [24.04.2014]
3) Eisenhauer, J.G., 2003. Regression through the origin, *Teaching Statistics*, vol 25(3), lk 76-80.
4) Gayawan, E. ja Ipinyomi, R.A., 2009. A Comparison of Akaike, Schwarz and R Square for Model Selection Using Some Fertility Models, *Australian Journal of Basic and Applied Sciences*, vol 3(4), lk 3524-3530.
5) Landram, F. G., Abdullat, A. ja Shah, V., 2005. The coefficient of prediction for model specification, *Southwestern Economic Review*, vol 32(1), lk 149-156.
6) Mendez Mediavilla, F.A., Landram, F. ja Shah, V., 2008. A Comparison of the Coefficient of Predictive Power, Coefficient of Determination and *AIC* for Linear Regression, *Journal of Applied Business and Economics*, vol 8(4). Allikas: http://www.na-businesspress.com/JABE/MendezWeb.pdf , [24.04.2014].
7) Myers, R.H., 1990. *Classical and modern regression with applications*. – 2nd ed. Belmont: Duxbury Press.
8) Wittman, D., 2005. A refresher in Statistics and econometrics, lk 52-53. Allikas: http://people.ucsc.edu/~wittman/classes/econ-113/c.05.pdf , [04.05.2014]

Lisad

Lisa 1. Tarkvarapaketi SAS kood ja väljavõte, vabaliikmega mudel

Kood:

```
Proc reg data=andmed.class;
model kaal=kasv/p; /* P väljastab: tegelikud väärtused, prognoosid, jäägid, SSE ja PRESS*/
run;
```

Väljavõte:

The REG Procedure
Model: MODEL1
Dependent Variable: Kaal 0.4536 * Weight

Number of Observations Read	19
Number of Observations Used	19

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1480.03230	1480.03230	57.08	<.0001
Error	17	**440.82319**	25.93078		
Corrected Total	18	**1920.85549**			

Root MSE	**5.09223**	**R-Square**	**0.7705**
Dependent Mean	45.37194	**Adj R-Sq**	**0.7570**
Coeff Var	11.22330		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	**-64.87701**	14.63975	-4.43	**0.0004**
Kasv	2.54 * Height	1	**0.69630**	0.09217	7.55	**<.0001**

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: Kaal 0.4536 * Weight

Output Statistics

Obs	Dependent Variable	Predicted Value	Residual
1	51.0300	57.1564	-6.1264
2	38.1024	35.0489	3.0535
3	44.4528	50.6126	-6.1598
4	46.4940	46.1911	0.3029
5	46.4940	47.4291	-0.9351
6	37.6488	36.4638	1.1850
7	38.3292	40.8853	-2.5561
8	51.0300	45.6605	5.3695
9	38.1024	45.6605	-7.5581
10	45.1332	39.4704	5.6628
11	22.9068	25.8522	-2.9454
12	40.8240	48.8440	-8.0200
13	34.9272	34.6952	0.2320
14	50.8032	52.7349	-1.9317
15	68.0400	62.4622	5.5778
16	58.0608	49.7283	8.3325
17	60.3288	53.6192	6.7096
18	38.5560	36.8175	1.7385
19	50.8032	52.7349	-1.9317

Sum of Residuals	0
Sum of Squared Residuals	440.82319
Predicted Residual SS (PRESS)	**545.52354**

Lisa 2. Tarkvarapaketi SAS kood ja väljavõte, vabaliikmeta mudel

Kood:

```
Proc reg data=andmed.class;
model kaal=kasv/noint;
run;
```

Väljavõte:

The REG Procedure
Model: MODEL1
Dependent Variable: Kaal 0.4536 * Weight

Number of Observations Read	19
Number of Observations Used	19

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	40084	40084	759.44	<.0001
Error	18	950.07147	52.78175		
Uncorrected Total	19	41034			

Root MSE	**7.26510**	**R-Square**	**0.9768**
Dependent Mean	45.37194	**Adj R-Sq**	**0.9756**
Coeff Var	16.01233		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Kasv	2.54 * Height	1	**0.28916**	0.01049	27.56	**<.0001**

Lisa 3. Tarkvarapaketi SAS kood ja väljavõte, mudelisse tunnuste lisamine

Esimene mudel

Kood:

```
proc reg data=andmed.fitness;
model hapnik=pulss vanus/P;
run;
```

Väljavõte

The REG Procedure
Model: MODEL1
Dependent Variable: hapnik hapniku tarbimine

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	320.11557	160.05779	8.44	0.0014
Error	28	**531.26597**	18.97378		
Corrected Total	30	**851.38154**			

Root MSE	**4.35589**	**R-Square**	**0.3760**
Dependent Mean	47.37581	**Adj R-Sq**	**0.3314**
Coeff Var	9.19434		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	**121.37601**	18.13570	6.69	**<.0001**
Pulss	pulsisagedus jooksu ajal	1	**-0.29382**	0.08242	-3.56	**0.0013**
vanus	vanus aastates	1	**-0.50665**	0.16214	-3.12	**0.0041**

The REG Procedure
Model: MODEL1
Dependent Variable: hapnik hapniku tarbimine

Output Statistics

Obs	Dependent Variable	Predicted Value	Residual
1	39.4070	41.3730	-1.9660
2	46.0800	48.1817	-2.1017
3	45.4410	46.8444	-1.4034
4	54.6250	53.1464	1.4786
5	45.1180	45.0005	0.1175
6	39.2030	44.6559	-5.4529
7	45.7900	40.8871	4.9029
8	50.5450	49.0122	1.5328
9	48.6730	41.9004	6.7726
10	47.9200	47.1081	0.8119
11	47.4670	45.0815	2.3855
12	50.5410	49.7223	0.8187
13	37.3880	43.9270	-6.5390
14	44.7540	46.8652	-2.1112
15	47.2730	49.9653	-2.6923
16	51.8550	45.2435	6.6115
17	49.1560	43.6633	5.4927
18	40.8360	46.1758	-5.3398
19	46.6720	47.9387	-1.2667
20	46.7740	49.4586	-2.6846
21	50.3880	47.1891	3.1989
22	44.6090	46.7842	-2.1752
23	45.3130	46.7541	-1.4411
24	54.2970	53.2481	1.0489
25	59.5710	51.3233	8.2477
26	49.8740	49.8241	0.0499
27	44.8110	45.8519	-1.0409
28	45.6810	49.3984	-3.7174
29	49.0910	51.9919	-2.9009
30	39.4420	47.9594	-8.5174
31	60.0550	52.1746	7.8804

Sum of Residuals	0
Sum of Squared Residuals	531.26597
Predicted Residual SS (PRESS)	**658.19879**

Teine mudel

Kood:

```
proc reg data=andmed.fitness;
model hapnik=pulss vanus aeg/P;
run;
```

Väljavõte:

The REG Procedure
Model: MODEL1
Dependent Variable: hapnik hapniku tarbimine

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	690.55086	230.18362	38.64	<.0001
Error	27	**160.83069**	5.95669		
Corrected Total	30	**851.38154**			

Root MSE	**2.44063**	**R-Square**	**0.8111**
Dependent Mean	47.37581	**Adj R-Sq**	**0.7901**
Coeff Var	5.15165		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	**111.71806**	10.23509	10.92	**<.0001**
Pulss	pulsisagedus jooksu ajal	1	**-0.13091**	0.05059	-2.59	**0.0154**
vanus	vanus aastates	1	**-0.25640**	0.09623	-2.66	**0.0129**
aeg	aeg 1.5 miili läbimiseks (min)	1	**-2.82538**	0.35828	-7.89	**<.0001**

The SAS System
The REG Procedure
Model: MODEL1
Dependent Variable: hapnik hapniku tarbimine

Output Statistics

Obs	Dependent Variable	Predicted Value	Residual
1	39.4070	38.6407	0.7663
2	46.0800	45.8913	0.1887
3	45.4410	49.7079	-4.2669
4	54.6250	54.5831	0.0419
5	45.1180	44.8203	0.2977
6	39.2030	39.4890	-0.2860
7	45.7900	44.7110	1.0790
8	50.5450	49.6729	0.8721
9	48.6730	48.2470	0.4260
10	47.9200	44.6646	3.2554
11	47.4670	46.4644	1.0026
12	50.5410	49.8228	0.7182
13	37.3880	36.1911	1.1969
14	44.7540	45.7220	-0.9680
15	47.2730	48.5111	-1.2381
16	51.8550	46.9556	4.8994
17	49.1560	50.3038	-1.1478
18	40.8360	45.7112	-4.8752
19	46.6720	49.1808	-2.5088
20	46.7740	49.2436	-2.4696
21	50.3880	48.6821	1.7059
22	44.6090	45.0102	-0.4012
23	45.3130	48.7925	-3.4795
24	54.2970	55.5753	-1.2783
25	59.5710	56.1351	3.4359
26	49.8740	52.6232	-2.7492
27	44.8110	43.7683	1.0427
28	45.6810	44.6589	1.0221
29	49.0910	48.8304	0.2606
30	39.4420	40.7025	-1.2605
31	60.0550	55.3374	4.7176

Sum of Residuals	0
Sum of Squared Residuals	160.83069
Predicted Residual SS (PRESS)	**205.12466**

Kolmas mudel

Kood

```
proc reg data=andmed.fitness;
model hapnik=pulss vanus aeg kaal/P;
run;
```

Väljavõte

The REG Procedure
Model: MODEL1
Dependent Variable: hapnik hapniku tarbimine

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	695.14669	173.78667	28.92	<.0001
Error	26	**156.23485**	6.00903		
Corrected Total	30	**851.38154**			

Root MSE	**2.45133**	**R-Square**	**0.8165**
Dependent Mean	47.37581	**Adj R-Sq**	**0.7883**
Coeff Var	5.17423		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	**115.66230**	11.22578	10.30	**<.0001**
Pulss	pulsisagedus jooksu ajal	1	**-0.12932**	0.05084	-2.54	**0.0173**
vanus	vanus aastates	1	**-0.27642**	0.09932	-2.78	**0.0099**
aeg	aeg 1.5 miili läbimiseks (min)	1	**-2.77237**	0.36492	-7.60	**<.0001**
kaal	kaal kg	1	**-0.04932**	0.05640	-0.87	**0.3898**

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: hapnik hapniku tarbimine

Output Statistics

Obs	Dependent Variable	Predicted Value	Residual
1	39.4070	38.7703	0.6367
2	46.0800	45.6786	0.4014
3	45.4410	49.6172	-4.1762
4	54.6250	54.7351	-0.1101
5	45.1180	45.2865	-0.1685
6	39.2030	38.7818	0.4212
7	45.7900	44.8485	0.9415
8	50.5450	50.3229	0.2221
9	48.6730	48.2390	0.4340
10	47.9200	45.5064	2.4136
11	47.4670	46.1107	1.3563
12	50.5410	50.0874	0.4536
13	37.3880	35.9493	1.4387
14	44.7540	46.3563	-1.6023
15	47.2730	48.4292	-1.1562
16	51.8550	46.5297	5.3253
17	49.1560	50.0110	-0.8550
18	40.8360	46.0468	-5.2108
19	46.6720	49.0481	-2.3761
20	46.7740	48.5075	-1.7335
21	50.3880	48.8271	1.5609
22	44.6090	44.5455	0.0635
23	45.3130	49.0603	-3.7473
24	54.2970	55.1105	-0.8135
25	59.5710	56.5734	2.9976
26	49.8740	52.1868	-2.3128
27	44.8110	43.8470	0.9640
28	45.6810	44.9672	0.7138
29	49.0910	48.7411	0.3499
30	39.4420	40.7191	-1.2771
31	60.0550	55.2098	4.8452

Sum of Residuals	0
Sum of Squared Residuals	156.23485
Predicted Residual SS (PRESS)	**212.27170**

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Ann-Mari Koppel,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Determinatsioonikordaja ja prognoosikordaja“,

mille juhendaja on Ene Käärik,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.
3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 02.05.2014