Peatükk 8

Lihtne lineaarne regressioon

Siin peatükis õpime käejoonte järgi pikka või lühikest elu ennustama
ehk
regressioonmudel, mudeli olulisuse testimine ja tema eeldused, determinatsiooni- ja korrelatsioonikordaja

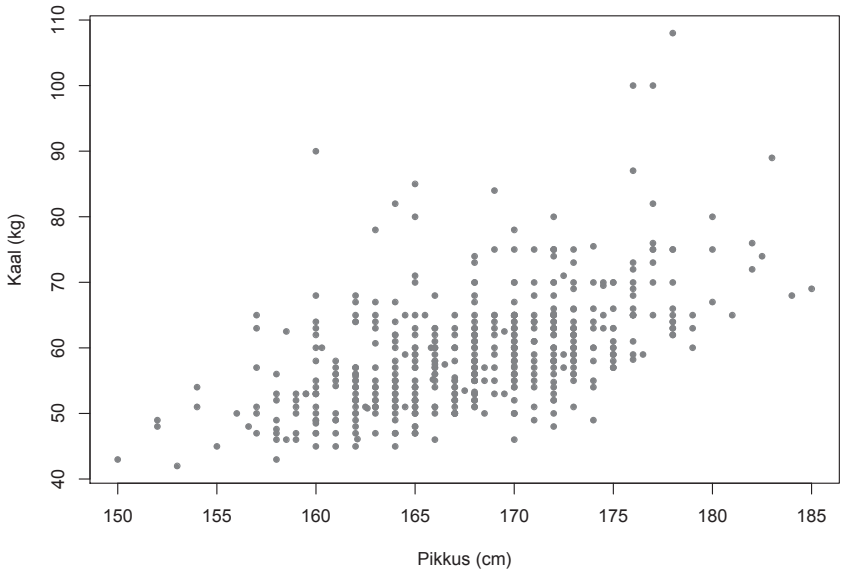
Nähes inimest, kelle nägu kaunistavad rohked kortsud, oskame üht-teist tema kohta arvata ka siis, kui me midagi muud temast ei tea. Võime arvata, et ta on vanem, elukogenud ja küllap ka veidi konservatiivne (arvatavasti eelistab ta ikka veel paberraamatut lugeda ning ehk ta ei ole kuulnudki iPad'ist või Kindlest...). Mõistes ühe tunnuse (kortsude arv näos) seost teise tunnusega (vanus) on meil võimalik "ennustada" meile tundmatu tunnuse väärtuseid (oletada teise inimese vanust). Selline ennustamine võiks olla eriti meelepärane siis, kui ühe tunnuse mõõtmine on suhteliselt lihtne ja odav (piisab ühest pilgust inimese näkku märkamaks ta kortse), samas kui teise tunnuse mõõtmine on raske, kallis või ohtlik (eks proovige küsida daamilt tema vanust!)

Vahel on erinevate tunnuste vahelised seosed tõepoolest intuitiivselt mõistetavad ja piisab ka nende seoste põhjal tehtud ligikaudsetest järeldustest. Mõnikord aga vajame võimalikult täpseid tulemusi (hästi, on elukogenum, aga kui palju elukogenum?), või on tunnustevaheline seos meie jaoks üldse mõistatuseks — kas see üldse eksisteerib, ja kui, siis milline see on? Sellistel juhtudel — kui täpsus on oluline või kui intuitsioon ei aita — võib endale appi paluda regressioonanalüüsi.

8.1 Sissejuhtaus

Kui uurimise all on kaks pidevat tunnust, mille vahelist seost tahame mõista, on enamasti tark esmalt vaadata tunnustevahelist hajuvusgraafikut. Joonisel 8.1 on toodud Tartu Ülikoolis õppivate naistudengite pikkuse ja kaalu vahelist seost kirjeldav hajuvusgraafik (*scatterplot*).

Joonis 8.1: Hajuvusgraafik



Hajuvusgraafikult näeme, et pikemad tudengid kipuvad ka kaalukamad olema. Seost nende kahe tunnuse vahel võiks justnagu kirjeldada sirge abil. Sageli ongi seost kahe tunnuse vahel üsna hästi võimalik kirjeldada sirge abil (muidugi mitte alati, aga esimeseks lähendiks sobib sirge üllatavalt sageli).

Igaüks võiks tõmmata seost iseloomustava sirge läbi hajuvusgraafikul kujutatud punktipilve. Üldises vestluses, sõprade seltsis, võime tõepoolest rahulduda sellise käega veetud sirgega. Paraku tõmbab üks inimene veidi ühtemoodi sirge kui teine, isegi läbi sellesama punktipilve. Tekivad küsimused — milline sirge on parem, milline joonis viiks täpsemate prognoosideni?

Parem on muidugi see sirge, mis viib võimaldab uusi, tulevasi väärtuseid, võimalikult täpselt prognoosida. Paraku see, milline sirgetest osutus paremaks, selgub alles tagantjärgi, siis kui prognoositavad sündused on juba

toimunud. Enamasti tahaksime teha valiku sirgete vahel enne seda hetke.

Üks võimalus valida kõikvõimalike seost iseloomustavate sirgete vahel on järgmine. Vaatame, milline sirge prognoosib kõige paremini olemasolevaid andmeid. Peaksime muidugi täpsustama, mis mõttes täpsemalt (kuidas mõõdame täpsust). Üheks parimaks täpsuse mõõdupuuks on osutunud vähimruutude nõue — eelistame sirget, mille puhul prognoosivigade ruutude summa oleks minimaalne (alternatiivne sõnastus: keskmine ruutviga oleks minimaalne). Täpsemalt, kui i. vaatlus on y_i, tema prognoos regressioonisirge abil (kui meid prognoosimisel abistava tunnuse väärtus on x_i) on $\hat{y}_i = \hat{c}_0 + \hat{c}_1 x_i$, siis valime $\hat{c}_0$ ja $\hat{c}_1$ väärtused selliselt, et prognoosivigade

$$\begin{aligned} e_i &:= y_i - \hat{y}_i \\ &= y_i - \hat{c}_0 + \hat{c}_1 x_i \end{aligned}$$

ruutude summa

$$\sum_{i=1}^{n} e_i^2$$

oleks minimaalne. Joonisel 8.1 kujutaud pikkuse ja kaalu andmetega sobib kõige paremini järgmine regressioonsirge:

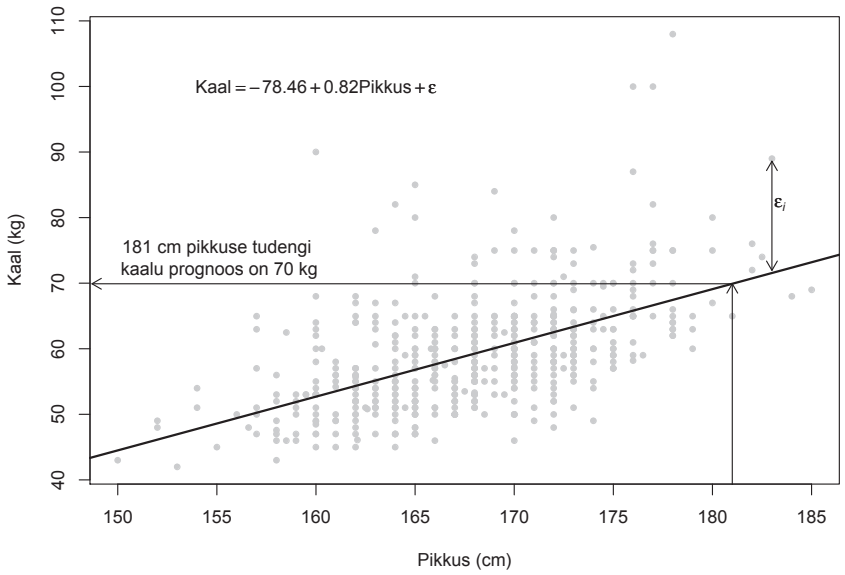
$$\hat{kaal} = -78{,}46 + 0{,}82 \cdot pikkus,$$

ehk sirge, kus vabaliige $c_0 = -78{,}46$ ja sirge tõus $c_1 = 0{,}82$, vaata ka joonist 8.10.

8.2 Regressioonanalüüsi mudel

Kui kahe pideva tunnuse vahel eksisteerib statistiline seos, siis peaks ühe tunnuse väärtuste muutudes muutuma ka teise tunnuse väärtuste jaotus. Sageli (kuid mitte alati) seisneb muutus selles, et ühe tunnuse väärtuste kasvades kipuvad teise tunnuse väärtused olema ka suuremad (või vastupidi väiksemad). Näiteks pikemad inimesed kipuvad ka rohkem kaaluma; suuremates metsades kipub kasvama rohkem puid; mida rohkem inimene joob karastusjooke, seda lühem kipub olema tema eluiga. Kuigi suuremas metsas kipub olema rohkem puid kui väiksemas metsas, ei tähenda see veel seda, et ei võiks esineda mõnda suurt ja hõredat metsa, kus kasvab puid vähem kui mõnes tillukeses tihedalt väikseid noori puid täis metsas.

Joonis 8.2: Prognoosivigade ruutude summat minimiseeriv sirge



Regressioonanalüüsi mudel kirjeldabki, kuidas muutub Y-tunnuse keskväärtus siis, kui vaatleme erineva X-tunnuse väärtustega objekte. Lihtsa lineaarse regressioonanalüüsi mudel ütleb, et Y-tunnuse keskväärtus muutub lineaarselt X-tunnuse väärtuste muutumisel:

$$\mathrm{E}Y = c_0 + c_1 x. \tag{8.1}$$

Kui räägime naistudengite pikkusest ja kaalust, siis võiks sobiv mudel välja näha järgmine:

$$\mathrm{E}kaal = -78{,}46 + 0{,}82 \cdot pikkus. \tag{8.2}$$

Kuidas tõlgendada antud mudelit? Kui vaataksime vaid 150cm pikkuseid tudengeid, siis selliste tudengite keskmine kaal on antud mudeli järgi

$$-78{,}46 + 0{,}82 \cdot 150 = 44{,}54kg.$$

See ei tähenda, et iga 150cm pikkune tudeng peaks kaaluma täpselt 44,54 kg. Antud mudel lihtsalt väidab, et 150cm pikkuste tudengite keskväärtus on 44,54kg. Samuti väidab ta midagi ka 151cm pikkuste tudengite keskmise

kaalu kohta. Sentimeetri võrra pikemate tudengite keskmine kaal peaks antud mudeli järgi olema $-78{,}46 + 0{,}82 \cdot 151 = 44{,}36kg$ ehk $0{,}82kg$ rohkem.

Toodud näidet võime üldistada andmaks regressioonsirge tõusule (c_1) üldisemat interpretatsiooni. Kui vaatleme ühe sentimeetri võrra pikemaid tudengeid, siis kasvab tudengite keskmine kaal sirge tõusu ehk $0{,}82kg$ võrra. Üldjuhul võiksime öelda järgmist: kui võrdleksime kahte gruppi uuritavaid, kus ühes grupis on x-tunnuse väärtus ühe ühiku võrra suurem kui teises uuritavate grupis, siis oleksid y-tunnuse keskväärtused gruppide vahel c_1 võrra erinevad.

Sageli arvatakse ekslikult, et kui me suurendaksime x-tunnuse väärtust 1 ühiku võrra, siis suureneb y-tunnuse keskväärtus c_1 ühikut. See võib mõningatel juhtudel ka nii olla, aga vaatlusandmete puhul ei pea see enamasti paika: kui võtaksime puntra tudengeid ja venitaksime neid piinapingil 1 cm võrra pikemaks, siis antud piinaprotsedur ei tee neid õnnetuid tudengeid veel 0,82kg võrra raskemateks!

Ka vabaliiget c_0 on vahel võimalik interpreteerida. Kui vaatleme vaid neid uurimisobjekte uuritavas populatsioonis, kelle puhul x-tunnuse väärtus on 0, siis uuritava populatsiooni taoliselt valitud alamhulga y-tunnuse keskväärtust näitabki vabaliige. Kuna aga 0cm pikkust naistudengit ei eksisteeri, siis pole ka antud juhul vabaliikmel sisulist tähendust. Samuti ei tohiks antud sirget kasutada näiteks vastsündinud 50cm pikkuse lapse kaalu leidmiseks: antud regressioonsirge leidmiseks kasutatud tudengite valim võib olla küll esindav kõigi 2. kursuse tudengite jaoks, kuid pole kindlasti esindavaks valimiks vastsündinute jaoks. Näiteks tudengineiude keskmine kaal on 59kg, vastsündinute keskmine kaal on aga kindlasti midagi muud. Arvamine, et tudengite jaoks leitud regressioonsirge võiks sobida vastsündinute kirjeldamiseks, on sama absurdne, kui arvata, et tudengite keskmine kaal kirjeldab vastsündinute keskmist kaalu. Regressioonmudelite kasutamisel tuleb hoolega jälgida, et uuritav, kelle y-tunnuse väärtust soovime prognoosida, kuuluks ikka samasse uuritavasse populatsiooni mida kirjeldab kasutatav regressioonmudel.

On väheusutav, et ühe konkreetse tudengi kaal oleks täpselt võrdne tudengite keskmise kaaluga. Kui tahame kirja panna mudelit juhuslikult valitud tudengi kaalu jaoks, siis peame arvestama, et tema kaal erineb rohkem või vähem keskväärtusest. Isiku eripära, tema kaalu erinevust keskväärtusest (hälvet) tähistatakse sageli sümboliga ε:

$$kaal = -78{,}46 + 0{,}82 \cdot pikkus + \varepsilon. \tag{8.3}$$

Kui varem kirja pandud regressioonmudel (8.2) kirjeldas, kuidas kaalu kesk-

väärtus muutub erineva pikkusega tudengitel, siis regressioonmudel (8.3) kirjeldab, kuidas (juhuslikult valitud) tudengineiu kaal sõltub pikkusest.

Valimis on mõõdetud paljude uuritavate y-tunnuse väärtused. Kui soovime rääkida näiteks i. tudengi kaalust, siis võime kasutada ka alaindekseid:

$$kaal_i = -78{,}46 + 0{,}82 \cdot pikkus_i + \varepsilon_i.$$

Nii on i. tudengil oma isiklik kaal ($kaal_i$), pikkus ($pikkus_i$) ja hälve ehk erinevus keskmisest (ε_i).

Keskmine erinevus keskmisest peab muidugi olema null ($\mathrm{E}\varepsilon = 0$), sest muidu poleks keskmine enam keskmine. Vahel aga tekib vajadus täpsustada, kui suured ja millise jaotusega võivad olla hälbed ehk üksikindiviidide erinevused keskmisest. Sageli eeldatakse (ja tihti ka on) hälbed normaaljaotusega, $\varepsilon \sim N(0; {\sigma_\varepsilon}^2)$. Sellisel juhul on ka y-tunnuse jaotuseks normaaljaotus. Tudengite kaalu-pikkuse näite korral saame näiteks hälvete normaaljaotust eeldades kirja panna, millise jaotusega on mingi kindla pikkusega tudengite kaalud:

$$kaal \sim N(-78{,}46 + 0{,}82 \cdot pikkus; {\sigma_\varepsilon}^2),$$

või üldisema juhu tarvis kirjapandult:

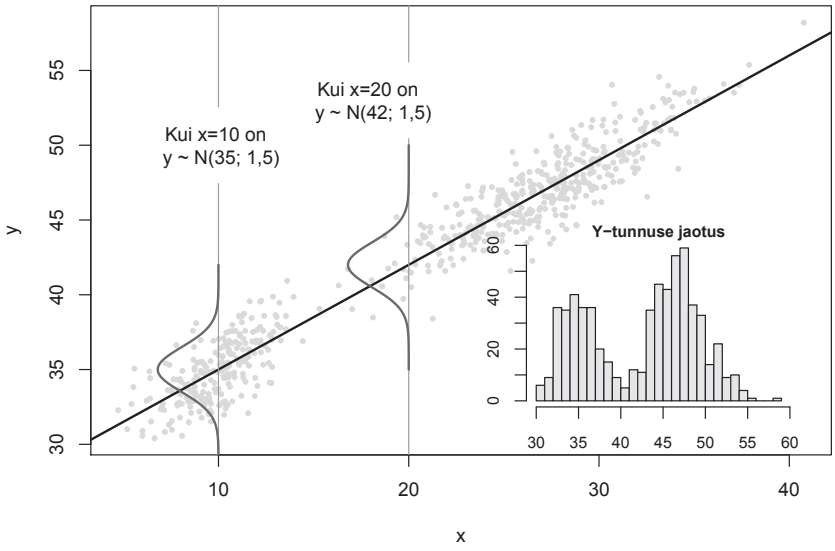
$$y \sim N(c_0 + c_1 x; {\sigma_\varepsilon}^2) \tag{8.4}$$

Taolises olukorras kiputakse rääkima, et uuritav tunnus on normaaljaotusega (näiteks kaal on normaaljaotusega). Paraku tekib siin oht segi ajada kahte täiesti erinevat asja. Regressioonanalüüsi juures peetakse silmas, et y-tunnuse väärtused mingi fikseeritud x-tunnuse väärtuse korral on normaaljaotusega. Näiteks 160cm pikkuste tudengite kaalud on normaaljaotusega. Samuti eeldatakse, et 170 cm pikkuste tudengite jaotus on normaaljaotusega. Aga need kaks normaaljaotust on erinevad, erinevate keskväärtustega. Vahel võib esineda olukordi, kus y-tunnuse jaotus iga konkreetse x-tunnuse väärtuse korral on normaaljaotusega, kuid y-tunnuse enda jaotus (üle erinevate x-tunnuse väärtuste) siiski pole normaaljaotusega. Ka sellisel juhul räägime, et regressioonanalüüsi jäägid on normaaljaotusega ja normaaljaotuse eeldus on täidetud. Taolist võimalikku olukorda kirjaldab joonis 8.3.

Tuleb silmas pidada, et regressioonmudel (8.4) ei kirjelda ära mitte ainult seda, kuidas y-tunnuse keskväärtus x-tunnuse väärtuste muutudes muutub, vaid kirjeldab ära ka selle, kui kaugele keskväärtusest üksikvaatlused sattuda võivad.

Nagu iga mudel, on ka kaalu ja pikkuse omavahelist seost kirjeldav regressioonanalüüsi mudel (veidi) vale. See ei tähenda, et antud mudelist ei võiks

Joonis 8.3: Normaaljaotusega jääkidega regressioonanalüüsi mudel 8.4



olla kasu pikkuse ja kaalu omavahelise seose mõistmiseks või uute, valimis mitteesindatud tudengite kaalu prognoosimisel pikkuse abil. Samas tuleks siiski kontrollida, kas lineaarne seos vähemalt ligikaudseltki sobib kirjeldama seost y ja x tunnuste vahel. Mudeli sobivuse kontrollimise võimalusi on lähemalt kirjeldatud alapeatükis **??**.

8.3 Hinnang ja tema täpsus

Uuritavas populatsioonis võib ju y-tunnuse keskväärtus sõltuda x-tunnuse väärtustest lineaarselt, $\mathrm{E}Y = c_0 + c_1 x$, kuid tavaliselt, vähemalt eluteadustes, pole võimalik kordajate c_0 ja c_1 väärtust võimalik leida teooriale toetudes. Nimetatud tundmatud kordajate c_0 ja c_1 hinnangud $\hat{c}_0$ ja $\hat{c}_1$ tuleb leida valimi põhjal. Valimi põhjal tehtud hinnang on aga paratamatult ekslik, seega tuleb kirjeldada ka hinnangu täpsust.

Esmalt hinnangust endast. Tavaliselt hinnatakse parameetrid c_0 ja c_1 vähimruutude meetodil, st minimiseeritakse olemasolevate vaatluste "prog-

noosimisel" tekkivate prognoosivigade ruutude summa:

$$\hat{c}_0, \hat{c}_1 = \underset{c_0,c_1}{\operatorname{argmin}} \left(y_i - (c_0 + c_1 x)\right)^2 .$$

Miks eelistatakse vähimruutude meetodit? Saksa teadlane Gauss tõestas nimelt paar sajandit tagasi teoreemi, mida tänapäeval tuntakse Gauss-Markovi teoreemi nime all. Antud teoreem väidab järgmist:
Kui

- tegemist on juhusliku valimiga;
- tegelikult kehtib seos $\mathrm{E}Y = c_0 + c_1 x$;
- kui jäägid on sama dispersiooniga iga x väärtuse korral, $\mathrm{D}\varepsilon = {\sigma_\varepsilon}^2 \ \forall x$

siis
on vähimruutude meetodil saadud hinnangud $\hat{c}_0$ ja $\hat{c}_1$ kõige täpsemad nihketa hinnangud parameetritele c_0 ja c_1. Lisaks on parameetrite c_0 ja c_1 mistahes lineaarkombinatsioonile (näiteks suurusele $c_0 + c_1 x$) kõige täpsemaks nihketa hinnanguks vähimruutude meetodil saadud hinnanguid kasutav lineaarkombinatsioon ($\hat{c}_0 + \hat{c}_1 x$). Märkus: inglise keeles tuntakse sellist hinnangut nime all *BLUE - Best Linear Unbiased Estimator*.

Järgnevalt mõned selgitused teoreemi sõnastuse kohta.

Nihketa hinnang tähendab keskmiselt õiget hinnangut, st ühe valimi korral võime saada kordaja c_1 hinnangu liiga suure ja mõne teise valimi korral liiga väikese, aga keskmiselt, üle kõikmõeldavate valimite, annab hinnangute keskmine kokku parameetri õige väärtuse.

Mida tähendab täpsem antud teoreemi seisukohast? See ei tähenda, et mõni teine meetod ei võiks mõne konkreetse valimi korral anda täpsemat tulemust. See tähendab seda, et kui võtaksime palju juhuslikke valimeid, ning iga valimi põhjal hindaksime meid huvitavad parameetrid, siis ükski teine meetod ei anna väiksema dispersiooniga (väiksema varieeruvusega) hinnanguid. Kuna nihketuse nõudest tulenevalt peavad kõik hinnangud kõikuma õige parameetri väärtuse ümber, siis väike hinnangute varieeruvus tähendab ühtlasi seda, et nad peavad (enamasti) olema lähedal hinnatava suuruse tõelisele väärtusele.

Tavaliselt leitakse regressioonmudeli parameetrite hinnangud arvuti abil. Samas pole nende leidmine ka eriti keeruline, sestap toome siin igaks juhuks ära ka nn käsitsi arvutamiseks kõlvulikud valemid ja vihje nende tuletuskäigule.

Mudeli jääkide (ehk prognoosivigade) ruutude summa (mida minimiseeritakse) on

$$f(c_0, c_1) = \sum_{i=1}^{n} (y_i - c_0 - c_1 x_i)^2.$$

Seda summat käsitletakse tundmatute parameetrite c_0 ja c_1 funktsioonina. Selleks, et leida, millised c_0 ja c_1 väärtused minimiserivad jääkide ruutude summa, tuleb leida funktsiooni $f(c_0, c_1)$ tuletised c_0 ja c_1 järgi ja võrdsustada need nulliga. Saadud võrrandisüsteemi

$$\begin{cases} \frac{\partial f(c_0,c_1)}{\partial c_0} = 0 \\ \\ \frac{\partial f(c_0,c_1)}{\partial c_1} = 0 \end{cases}$$

lahendiks (vajaik algebra pole keeruline, võid ise ka proovida tuletuskäiku läbi teha!) on

$$\begin{aligned} \hat{c}_1 &= \frac{n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n {x_i}^2 - \left(\sum_{i=1}^n x_i\right)^2} \\ \hat{c}_0 &= \frac{\sum_{i=1}^n y_i - \hat{c}_1 \sum_{i=1}^n x_i}{n} \end{aligned}$$

Muidugi, tõeline matemaatik kontrolliks täiendavalt, kas vaadeldaval funktsioonil on kohas $\hat{c}_0, \hat{c}_1$ ikkagi miinimum (seda ta on) ja mitte näiteks maksimum (kui võrdsustame funktsiooni tuletise nulliga, võime tulemuseks saada kas funktsiooni maksimumi, miinimumi või käänukoha...).

Milleks selline piinarikas kõrvalepõige matemaatika valda? Selleks, et mõista: matemaatikud/statistikud muretsevad tundmatute parameetrite c_0 ja c_1 väärtuste leidmise pärast. Regressioonmudel (8.1) on otsitavate parameetrite c_0 ja c_1 lineaarne funktsioon. Sellepärast kutsutakse antud regressioonmudelit ka lineaarseks (regressioon)mudeliks. See, kas seos tunnuste X ja Y vahel on lineaarne, ei oma tegelikult mingit tähendust. Näiteks mudeli

$$\mathrm{E}Y = c_0 + c_1 x^3$$

puhul on tegemist lineaarse (regressioonanalüüsi) mudeliga, mudeli

$$\mathrm{E}Y = c_0 + c_1^2 x$$

puhul aga pole tegemist lineaarse regressioonimudeliga (sest tegemist on parameetri c_1 ruutfunktsiooniga).

Leitud hinnangud $\hat{c}_0, \hat{c}_1$ on paraku justnimelt hinnangud, ehk veidi rohkem või vähem valed. Kui erinevaid hinnanguid me erinevates (sama suurusega) juhuslikes valimites võiksime näha, seda kirjeldab (hinnangu) standardviga ehk hinnangu standardhälve.

Kuna lineaarse regressioonmudeli parameetrite hinnangute ja nende standardvigade arvutus on juba üsnagi arvutusmahukas tegevus, siis kasutatakse nende leidmiseks enamasti arvutite abi. Alljärgnevalt toetume siingi ühe statistikapaketi (R) väljundile edaspidistes arutlustes. Ka teiste statistikapakettide abil on võimalik teha sarnaseid arvutusi ja jõuda samasuguste tulemusteni.

Alljärgnevalt vaatame väljavõtet statistikapaketi R poolt lineaarse regressioonimudeli kohta trükitavast väljundist:

```
> mudel=lm(kaal~pikkus); summary(mudel)
[...]
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -78.45948    9.42448  -8.325 7.92e-16 ***
pikkus        0.81972    0.05609  14.615  < 2e-16 ***

Residual standard error: 7.298 on 506 degrees of freedom
[...]
```

Programmi R abil leitud hinnang — kuidas tudengi kaal sõltub pikkusest — on samasugune nagu juba varem sai kirja pandud: $kaal = -78{,}46 + 0{,}82 \cdot pikkus + \varepsilon$. Näeme, et tunnuse *pikkus* ees olev kordaja hinnangu $\hat{c}_1 = 0{,}8197$ standardvea hinnang on 0,056, ehk $s(\hat{c}_1) = 0{,}056$. Kui palju erinevaid teadlaseid võtaksid täpselt samasuured juhuslikud valimid Tartu Ülikooli (nais)tudengitest, ning kui kõik need teadlased hindaksid lineaarse regressioonmudeli abil, kuidas sõltub tudengite keskmine kaal tudengi pikkusest, siis nende teadlaste poolt nähtud regressioonsirgete tõusude (hinnangute $\hat{c}_1$) standardhälve oleks ligikaudu 0,056. Saab näidata, et (erinevate teadlaste poolt leitud) hinnangute $\hat{c}_1$ jaotuseks on ligikaudselt normaaljaotus (kui tegemist on kas suure valimiga või kui regressioonmudeli jäägid on normaaljaotusega). Normaaljaotuse puhul jääb aga 95% juhusliku suuruse väärtustest keskväärtusest vähem kui kahe standardhälbe kaugusele. Kuna vähimruutude meetod tagab regressiooniparameetritele nihketa hinnangu, siis saame rääkida, et 95% teadlastel on nende poolt leitud regressioonisirge tõusu hinnang $\hat{c}_1$ vähem kui kahe standardvea kaugusel tõelisest väärtusest c_1. Ka meie poolt vaadeldud hinnangu (0,82) kohta ei oska me öelda, kas

ta üle- või alahindab tegelikku regressioonsirge tõusu (seda, mida näeksime, kui küsitleksime lõpmatult palju tudengeid), kuid võime olla üsna kindlad, et erinevus tegeliku ja meie poolt nähtud numbri vahel ei tohiks olla suurem kahest standardveast. Samasuguse arutluskäiku võime soovi korral korrata ka vabaliikme c_0 kohta.

Kui soovime matemaatiliselt korrektselt leida 95% usaldusintervalli parameetri c_1 (või c_0) tegelikule väärtusele, peame arvestama, et kasutatav standardviga on samuti hinnang (ning võib olla seetõttu ka ise veidi vigane). Korrektse usaldusintervalli leidmiseks peame seetõttu kasutama t-jaotuse kvantiile normaaljaotuse kvantiilide asemel (mistõttu venib usaldusintervall veidi laiemaks). Matemaatiliselt korrektne arvutusvalem leidmaks $(1-\alpha)$-usaldusintervalli parameetrile c_1 on

$$\hat{c}_1 + t_{\alpha/2;df}\, s(\hat{c}_1) \ldots \hat{c}_1 + t_{1-\alpha/2;df}\, s(\hat{c}_1),$$

kus df, t-jaotuse vabadusastmete arv (regressioonmudeli kontekstis tuntud ka kui jääkide vabadusastmete arv), leitakse valemiga:

$$\begin{aligned} df &= \text{vaatluste arv} - \text{regressioonmudeli parameetrite arv} \\ &= n - 2 \end{aligned}$$

Esitatud valemit tuleb veidi kommenteerida ja täpsustada: lihtsa regressioonmudeli korral on tegemist kahe hinnatava parameetriga, c_0 ja c_1, seega lahutatakse vaatluste arvust maha 2. Keerukamate mudelite korral, kus uuritava tunnuse (kesk)väärtuse muutumist kirjeldatakse rohkemate tunnuste ja enamate parameetrite abil kasvab ka mahalahutatav number. Kuigi enamasti hinnatakse regressioonmudeli hindamisel ka jääkide hajuvus σ_ε, siis seda (hajuvus)parameetrit vabadusastmete arvutamisel arvesse ei võeta — arvesse lähevad vaid keskväärtuse muutumist kirjeldavad ja vaatluste põhjal hinnatud parameetrid.

Enamasti trükivad statistikapaketid ka regressioonmudeli väljundisse vajaliku vabadusastmete arvu ära (*error degrees of freedom*). Viimast numbrit on mõistlik vaadata kontrollimaks, ega arvuti ja kasutaja vahel pole tekkinud mittemõistmist — kas vabadusastmete arv ikka on kooskõlas vaatluste arvuga, kas ülaltoodud vastavus ikka kehtib?

Näide: Leiame täpse 95% usaldusintervalli pikkuse ees olevale kordajale. Antud regressioonmudeli hinadamisel on kasutatud 508 naistudengi pikkuseid ja kaale. Seega $df = 506$ ja tabelist (või arvutist) tuleb leida t-jaotuse kvantiilid $t_{0{,}025;df=506} = -1{,}96466...$ ja $t_{0{,}975;df=506} = 1{,}96466....$ Edasi läheb

juba lihtsalt:

$$\begin{aligned} 0{,}81972 - 1{,}96466 * 0{,}05609 \quad &\ldots \quad 0.81972 + 1{,}96466 * 0{,}05609 \\ 0{,}7095 \quad &\ldots \quad 0{,}9299 \end{aligned}$$

Ehk 95% kindlusega võime väita, et tegelik regressioonsirge tõus (mille saaksime, kui oleksime mõõtnud üle kõigi Tartu Ülikoolis õppivate tudengineiude kaalud-pikkused) on vahemikus $0{,}7095 \ldots 0{,}9299$.

Paljud statistikapaketid oskavad muidugi ka vastavaid arvutusi automaatselt teha, vaata näiteks statistikapaketi R'i poolt saadavat tulemust. Samas võib käsitsi arvutamise oskus osutuda vajalikuks, kui soovitakse usaldusintervalli leida vaid kirjanduses/artiklis avaldatud informatsiooni põhjal.

Leiame hinnatud parameetritele 95%-usaldusintervalli tarkvarapaketi R abil:

```
> confint(mudel)
                 2.5 %      97.5 %
(Intercept) -96.9754049 -59.9435493
pikkus        0.7095254   0.9299087
```

Näeme juba usaldusintervalle vaadates, et pikkuse ees olev kordaja võib tegelikkuses olla valimi põhjal arvutatud hinnangust $\hat{c}_1 = 0{,}82$ mõnevõrra suurem (kuni 0,9299) või väiksem (kuni 0,7095), kuid kindlasti on tegemist positiivse kordajaga — pikemate tudengite keskmine kaal on suurem kui lühemate tudengite keskmine kaal.

Sarnane küsimus — kas regressioonsirge tõus c_1 võib tegelikkuses olla ka 0 — tekib üllatavalt sageli. Sestap testitakse ka sageli hüpoteesipaare $H_0 : c_1 = 0$ vs $H_1 : c_1 \neq 0$. Vastavat hüpoteesi saab kergesti kontrollida t-testi abil:

$$t := \frac{\hat{c}_1}{s(\hat{c}_1)} \overset{H_0}{\sim} t_{df}$$

ehk parameetri hinnang jagatud tema standardveaga peab nullhüpoteesi kehtides olema t-jaotusega juhuslik suurus. T-jaotuse vabadusastmete arv on jälle see nn jääkide vabadusastmete arv, mida saab leida valemiga (8.5). Edasi saab jätkata nii nagu hüpoteeside statistilisel testimisel tavaks — vaatame, kas saadud t-statistiku väärtus on selline, mida võiksime näha nullhüpoteesi kehtides (kas jääb t-jaotuse 0,025- ja 0,975-kvantiili vahele), või arvutame p-väärtuse ehk tõenäosuse näha sedavõrd ekstreemset (või veel ekstreemsemat) t-statistiku väärtust nullhüpoteesi kehtides. Enamik statistikapakette teeb ülaltoodud hüpoteesidepaari kontrolli ära automaatselt ja

väljastab kohe ka p-väärtuse. Varemtoodud R-programmi väljundist võime näha, et hüpoteesipaari $H_0 : c_1 = 0$ vs $H_1 : c_1 \neq 0$ kontrollimisel saadi p-väärtuseks suurus, mis oli väiksem kui 0,0000000000000002 (<2e-16) ja hüpoteesipaari $H_0 : c_0 = 0$ vs $H_1 : c_0 \neq 0$ kontrollimisel saadi p-väärtuseks 7,92e-16 (0,000000000000000792). Igal juhul on p-väärtused väiksemad kui 0,05 (tüüpiliselt kasutatav olulisustõenäosus) ja mõlemad nullhüpoteesid tuleb kummutada (pole võimalik, et $c_1 = 0$, samuti pole võimalik, et $c_0 = 0$.

Kui regressioonsirge tõus võib uuritavas populatsioonis olla ka 0, siis on võimalik, et Y-tunnuse väärtuste prognoosimiseks kasutatav tunnus X on tegelikult täiesti kasutu, ei sisalda tegelikult informatsiooni Y-tunnuse kohta.

Vahel võib esile kerkida soov kontrollida veidi keerukamaid küsimusi. Näiteks võib mõnikord tekkida vajadus kontrollida, kas kehtib hüpotees $H_0 : c_1 = 1$, või veel üldisemal kujul kirjapandult, võime soovida kontrollida hüpoteesipaari $H_0 : c_1 = c_{H_0}$ (vs $H_1 : c_1 \neq c_{H_0}$), kus c_{H_0} on mõni meie uurimistöö jaoks oluline number (näiteks 1). Ka selliseid hüpoteese saab kontrollida analoogse t-testi abil, kasutades teststatistikut

$$t := \frac{\hat{c}_1 - c_{H_0}}{s(\hat{c}_1)},$$

mis taas nullhüpoteesi kehtides peab olema t-jaotusega, $t \overset{H_0}{\sim} t_{df}$. Selliseid teste arvutitarkvara juba automaatselt ei tee ja vajalikud arvutused tuleb teha ise, kasutades arvuti poolt leitud regressioonparameetrite hinnanguid ja nende standardvigu.

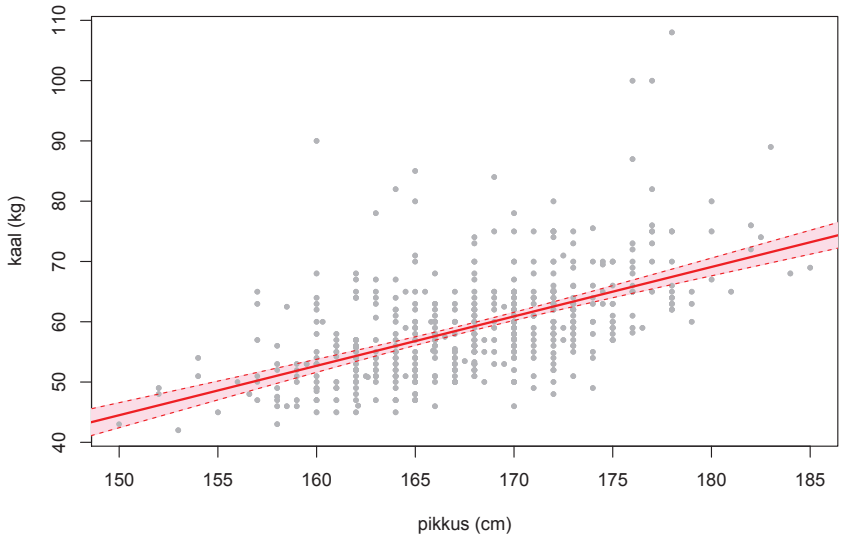
Muidugi tuntakse huvi regressioonmudeli parameetrite ja nende hinnangute täpsuse vastu. Tuleb siiski meeles pidada, et regressioonmudeli parameetrid omavad tähendust vaid seepärast, et ütlevad midagi y-tunnuse kohta, aitavad paremini ja täpsemalt kirjeldada y-tunnuse võimalikke väärtuseid. Sestap on ka tähtis mõista, kui täpselt oskame hinnata y-tunnuse keskväärtust ühel või teisel juhul. Kui täpselt teame, milline on y tunnuse keskväärtus siis, kui vaatleme vaid uuritavaid, kelle x tunnuse väärtuseks on näiteks 165? Vajalikku arvutust pole kahjuks võimalik teha kasutades vaid parameetrite hinnanguid ja nende standardvigu (vaja läheb parameetrite hinnangute kovariatsioonimaatriksit). Õnneks paljud statistikapaketid võimaldavad vajalikke arvutusi teha, nii ka R. Näiteks on võimalik arvutada, kui täpselt me teame 165cm pikkuste tudengineiude kaalu keskväärtust. 95%-usaldusintervalli 165cm pikkuste tudengineiude kaalu keskväärtusele saab statistikapaketis R küsida järgmiselt (kasutame varem hinnatud regressioonmudelit *mudel*):

```
> predict(mudel, data.frame(pikkus=165), interval="confidence")
       fit      lwr      upr
1 56.79383 56.08023 57.50743
```

Vastuseks saame, et hinnang 165cm pikkuste tudengineiude kaalu keskväärtusele on 56,79 ($-78{,}46 + 0{,}8197 \cdot 165$), 95%-usaldusintervall keskväärtusele aga (56,08 . . . 57,51). Võime olla üsna kindlad, et isegi kui kaaluksime ära kõik 165cm pikkused tudengineiud, siis saadud kaalude keskmine jääks mainitud vahemikku.

Taolisi usaldusintervalle saab leida erinevate pikkuste, erinevate x-tunnuse väärtuste jaoks. Kui kannaksime leitud usaldusintervallid graafikule, tekiks regressioonsirge ümber tema täpsust kirjeldav "koridor". Inglise keeles tuntakse leitut kui *95% pointwise confidence interval*. Vaata ka joonist 8.4, kus hajuvusgraafikule on lisatud regressioonisirge koos usaldusintervalliga.

Joonis 8.4: Regressioonsirge koos 95%-usaldusintervalliga.



NB! Usaldusintervalli interpretatsioon (näitab, kui täpselt teame Y-tunnuse keskväärtust mingi konkreetse X-tunnuse väärtuse korral) kehtib ainult siis, kui Y-tunnuse keskväärtus tõepoolest muutub nii, nagu kasutatud regressioonmudel ütleb. Usaldusintervall arvestab/kirjeldab vaid parameetrite (ebatäpsest) hindamisest tingitud võimaliku vea suurust, mitte aga mudeli valest

valikust tingitud vea suurust!

Kui valitud mudel on aga vale (enamasti on, vähemalt veidikenegi), siis võib leitud usaldusintervallile siiski anda ligikaudse interpretatsiooni: kui hindaksime üldkogumis, kõikmõeldavate uurimisobjektide andmeid kasutades, oma (vigase) regressioonmudeli, siis millise sirgeni võiksime sellisel juhul jõuda? Lisanduvate andmete tõttu saaksime tulemuseks arvatavasti veidi teistsuguse regressioonisirge, kuid üsna kindlasti jääks ta siiski 95%-usaldusintervalli poolt määratud koridori.

8.4 Prognoos ja tema täpsus

Kui on tarvis prognoosida uue objekti Y-tunnuse väärtust, siis pakutakse enamasti prognoosiks Y-tunnuse keskväärtust (keskväärtuse hinnangut). Lisainformatsiooni olemasolul (teame, et X-tunnuse väärtus oli x) pakume prognoosiks nn tinglikku keskväärtus (tema hinnangut): hinnangut nende objektide Y-tunnuse keskväärtusele, kellel $X = x$. Taolise tingliku keskväärtuse saame lihtsalt arvutada enda poolt hinnatud regressioonmudeli abil. Kasutades tudengite kaalu ja pikkuse vahelist seost kirjeldavat regressioonmudelit võime hinnata näiteks 180cm pikkuste tudengite keskmist kaalu. Saadud keskmine sobiks ka ühe uue (varem mittemõõdetud) 180cm pikkuse tudengi kaalu mõistlikuks prognoosiks:

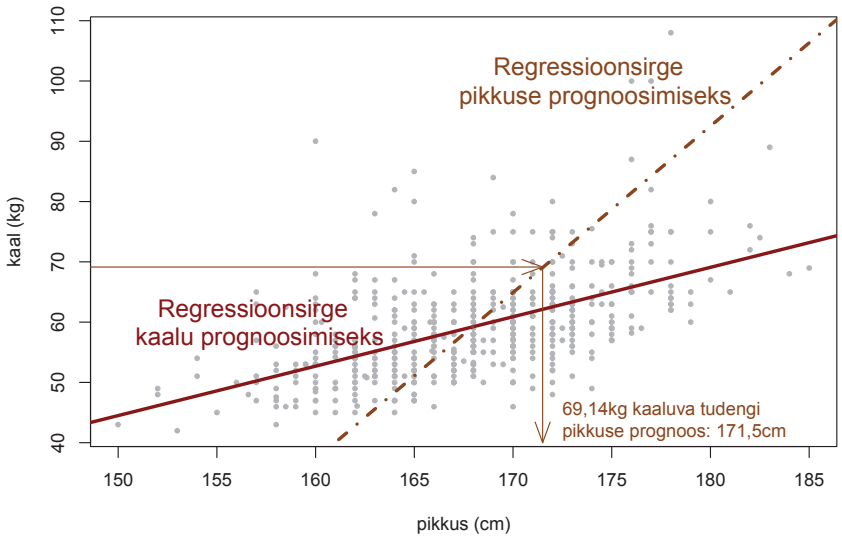
$$\hat{\mathrm{E}}(kaal|pikkus = 180cm) \quad = \quad -78{,}46 + 0{,}82 \cdot 180 = 69{,}14kg$$

Tasub ehk märkida, et 69,14kg raskuse tudengi pikkuse parim prognoos pole 180cm. Kui leiaksime regressioonmudeli, mis kirjaldab, kuidas pikkus sõltub kaalust (antud andmeid kasutades tuleb selleks regressioonmudel $\hat{\mathrm{E}}pikkus = 146{,}5 + 0{,}362 \cdot kaal$), siis saaksime 69,14kg kaaluva tudengi pikkuse prognoosiks hoopis 171,5cm. Kaalu prognoosib pikkuse järgi üks regressioonsirge, aga pikkuse prognoosimiseks kaalu järgi tuleb kasutada täiesti teistsugust regressioonsirget, vaata ka joonist 8.5. Tugevalt soovitatav on luua regressioonmudel just selle tunnuse jaoks, mille väärtuseid on tegelikult vaja prognoosida.

Saadud prognoos võib olla küll parim, mida olemasolevate andmete ja informatsiooni kasutades on võimalik pakkuda, kuid täiesti täpne ta (vähemalt enamasti) pole. Mitte kõik 180cm pikkused tudengid ei kaalu täpselt samapalju. Tahaksime teada, kui suur võib olla leitud prognoosi viga või võimalik eksimus.

Kui täpselt me keskväärtust tegelikult teame, võisime kirjeldada näiteks usaldusintervalli abil. Kuid isegi siis, kui keskväärtus oleks täpselt teada (kõik

Joonis 8.5: Pikkust kaalu abil prognoosiv regressioonsirge ja kaalu pikkuse järgi prognoosiv regressioonsirge



eestimaal elavad inimesed oleksid kaalutud-mõõdetud), poleks võimalik täpselt ennustada järgmise tänaval vastutuleva inimese kaalu. Iga inimene (või uuritav) on alati mõnevõrra individuaalne, tal on oma eripära, mistõttu tema uuritava tunnuse väärtus võib olla mõnevõrra erinev uuritava populatsiooni keskmisest.

Kui erinevad võivad uuritavad olla (tinglikust) keskmisest, seda saab mõõta/kirjeldada standardhälbe abil. Regressioonanalüüsi puhul huvitutakse enamasti regressioonmudeli jääkide standardhälbest (*residual standard error*). Jääkide standardhälve iseloomustab, kui kaugele (tinglikust) keskväärtusest võivad vaatlused tegelikult sattuda (enamasti jäävad vaatlused vähem kui kahe standardhälbe kaugusele keskväärtusest).

Juhul, kui regressioonmudeli jäägid — kõrvalekalded (tinglikust) keskväärtusest — on normaaljaotusega, on võimalik väga täpselt iseloomustada vahemikku, kuhu peaks jääma uue, prognoositava, objekti y-tunnuse väärtus. Seda saab teha prognoosiintervalli abil (peatükk 5.1). Prognoosiintervall iseloomustab vahemikku, kuhu uus, juhuslikult uuritavast populatsioonist valitud vaatlus, jääb soovitud tõenäosusega. Näiteks 95%-prognoosiintervall on vahemik, kuhu prognoositava suuruse tegelik väärtus jääb tõenäosuse-

ga 0,95. Kui regressioonmudeli jäägid on normaaljaotusega (ja valitud regressioonmudel sobib y-tunnuse keskväärtuse muutumise kirjeldamiseks), siis jääb uue, tulevase, vaatluse y tunnuse väärtus 95% tõenäosusega kahe jääkide standardhälbe σ_ε kaugusele regressioonsirgest ehk y-tunnuse tinglikust keskväärtusest. Prognoosiintervalli täpseks arvutamiseks tuleb arvesse võtta ka seda, et tegelikku jääkide standardhälvet pole teada (kasutada saab ainult jääkide standardhälbe hinnangut), samuti ka seda, et ka regressioonsirge on kõigest hinnang ning regressioonsirge parameetrid võivad olla hinnatud veaga. Regressioonmudelit kasutades leitud prognoosile prognoosiintervalli leidmine on enamasti arvutuslikult tülikas, sestap on mugavam lasta vajalikud arvutused teha arvutil. Enamikus statistikapakettides on vastav võimalus olemas, R-is saab prognoosiintervalli leida predict-käsu abil. Näiteks 95%-prognoosiintervalli uue tudengi kaalule, kelle pikkus on 165cm, saab leida nii:

```
> predict(mudel, data.frame(pikkus=165), interval="prediction")
       fit      lwr      upr
1 56.79383 42.43865 71.14901
```

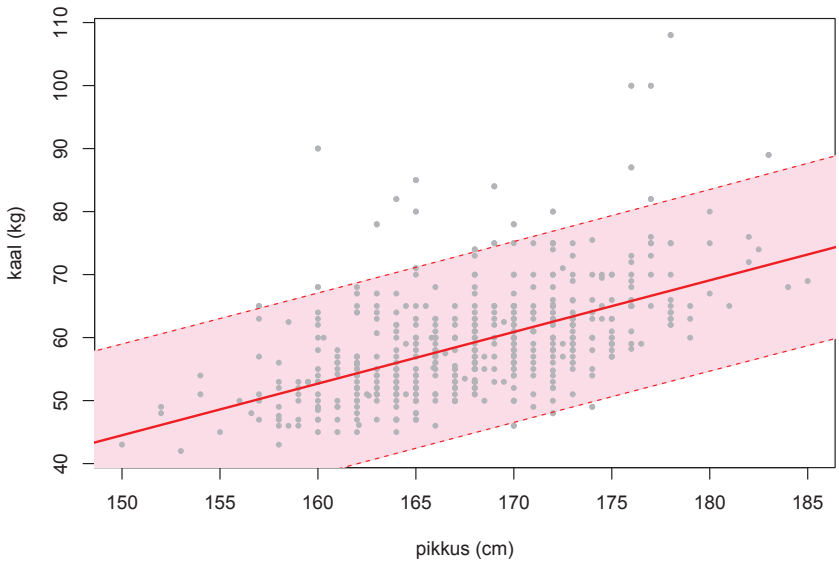
Selgub, et 165cm pikkuse tudengi kaaluks prognoosib kasutatav regressioonmudel 56,8kg, kuid 95% tõenäosusega jääb uue (valimisse mittekuulunud tudeng, selline tudeng, kelle pikkust ja kaalu regressioonmudeli hindamisel ei kasutatud) 165cm pikkuse tudengineiu kaal vahemikku 42,4kg kuni 71,2kg. Märkus: usaldusintervallide ja prognoosiintervallide puhul on heaks tavaks, et intervalli ümmardatakse laiemaks, näiteks: 42kg..72kg.

Taolisi prognoosiintervalle saab leida muidugi iga mõeldava pikkuse (x-tunnuse väärtuse) jaoks. Saadud prognoosiintervallid saab siis soovi korral ka hajuvusgraafikule kanda, saamaks koridori, kuhu järgmise tudengi kaal peaks suure tõenäosusega (95%) sattuma. Saadud graafikut iseloomustab joonis 8.6. Võrdle ka prognoosiintervalli iseloomustavat joonist usaldusintervalli kirjeldava joonisega (joonis 8.4). Võiks märgata, et usaldusintervall on märkimisväärselt kitsam prognoosiintervallist — kaalu keskväärtust on võimalik hinnata märksa täpsemalt, kui prognoosida üksiku juhuslikult valitud tudengi kaalu!

Mida prognoosiintervalli leidmisealgoritm (statistikapakett) arvestab ja mis jääb kasutaja mureks? Prognoosiintervalli arvutamisel võetakse enamasti arvesse, et regressioonsirge parameetrite hinnangud ei pruugi olla korrektsed ja täiesti õiged (arvestatakse, et $\hat{c}_0 \neq c_0$, $\hat{c}_1 \neq c_1$, $\hat{\sigma}_\varepsilon \neq \sigma_\varepsilon$), arvestatakse ka sellega, et y-tunnus on juhuslik, igal uurimisobjekt on eripärane.

Mis jääb prognoosiintervalli kasutaja vastutusele?

Joonis 8.6: 95%-prognoosiintervall.



- Kõigepealt, kasutaja peab vastutama, et tegelikult muutub Y tunnuse keskväärtus X tunnuse väärtuste muutudes tõepoolest lineaarset, ehk kasutaja peab garanteerima, et regressioonmudel 8.1 tõepoolest kehtib. Ei prognoosiintervall ega usaldusintervall ei üritagi kirjeldada võimalikku viga, mis tuleneb regressioonmudeli valest valikust (kui seos x-tunnuse ja y-tunnuse vahel pole lineaarne, vaid midagi muud, siis prognoosiintervalli ega usaldusintervalli arvutused pole enam korrektsed — kuigi kui mudel on ligikaudu õige, siis *võivad* ka prognoosi- ja usaldusintervallid *ligikaudu* paika pidada).

- Prognoosiintervalli arvutusvalem on väga tundlik normaaljaotuse eelduse suhtes — kui regressioonmudeli jääkide jaotuseks pole normaaljaotus, siis on ka prognoosiintervall valesti arvutatud. Tõsi, kui jääkide jaotus on enam-vähem normaaljaotusega, siis võiks ka leitud prognoosiintervall olla enam-vähem korrektne.

- Prognoosiintervalli arvutamisel eeldatakse, et regressioonmudeli jääkide hajuvus on kogu aeg samasuur — olgu x-tunnuse väärtus kas suur või väike, ikka on y-tunnuse väärtuste võimalikud kõrvalekalded ligi-

kaudu samasuured ehk jääkide dispersioon peaks olema samasugune mistahes x-tunnuse väärtuse korral.

Antud eelduste kontrolli võimalusi tutvustame regressioonmudeli eeldustest rääkivas alampeatükis.

Kui regressioonmudeli jäägid pole normaaljaotusega, siis on võimalik siiski leida ligikaudselt korrektset prognoosiintervalli, kasutades näiteks D.J.Olive (Olive, 2006) poolt pakutud nn jaotusvaba prognoosiintervalli arvutusvalemit. Olive mitteparameetrilise usaldusintervalli arvutusvalem pole kahjuks statistikapaketis R valmiskujul realiseeritud, küll aga on võimalik vajalik arvutus ise väikese vaevaga läbi teha. Alljärgnevalt on ära toodud R-keeles kirja pandud programm, mis leiab ligikaudselt korrektse prognoosiintervalli ilma jääkide normaaljaotust eeldamata (prognoosiintervall uue, 165cm pikkuse, tudengi kaalule):

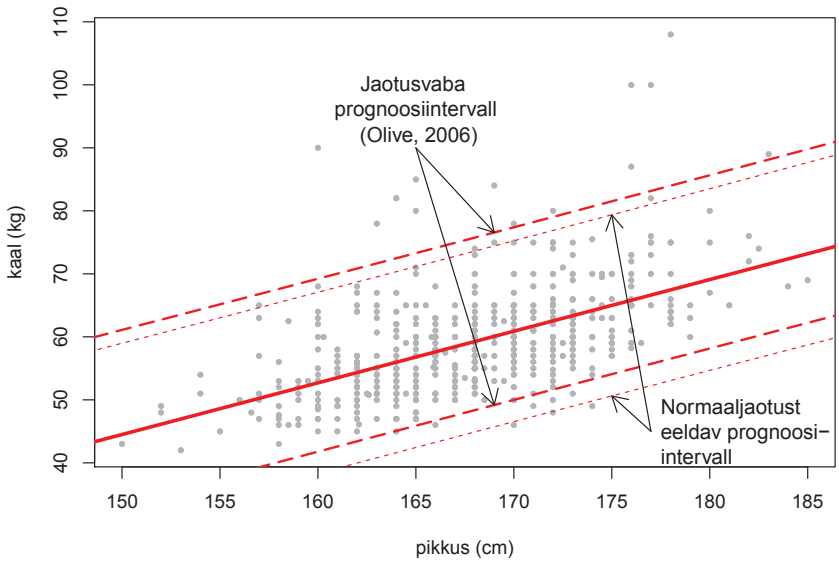
```
# Kasutatav regressioonmudel: prognoosime kaalu pikkuse abil
mudel=lm(kaal~pikkus)

# X-tunnuse väärtus, mille jaoks prognoosiintervalli leiame:
x=165

# Regressioonmudeli hinnatavate parameetrite arv;
# lihtsa regressioonmudeli korral p=2
p=length(coef(mudel))

# Regressioonmudeli hindamiseks kasutatud vaatluste arv
n=mudel$df+p

aa=predict(mudel, data.frame(pikkus=x), interval="confidence")
q_alumine = quantile(residuals(mudel), 0.025)
q_ylemine = quantile(residuals(mudel), 0.975)
se=summary(mudel)$sigma
konst=(1+15/n)*sqrt(n/(n-p)*(1+(aa[,3]-aa[,1])/(se**2*qt(0.975,mudel$df))))

alumine_PI=aa[,1]+konst*q_alumine
ylemine_PI=aa[,1]+konst*q_ylemine

alumine_PI; ylemine_PI
```

Antud programmi tulemusena saaksime 95%-prognoosiintervalliks 45,89...73,29, mis on märgatavalt erinev kui jääkide normaaljaotuse eeldusel leitud prog-

noosiintervall (42,43...71,15). Selline märkimisväärne erinevus erinevatel meetoditel leitud prognoosiintervallide vahel viitab võimalusele, et mõni varemtehtud eeldustest ei pea paika (äkki antud mudeli korral pole regressioonmudeli jäägid ikkagi normaaljaotusega?).

Normaaljaotuse eeldusest hoiduva Olive prognoosiintervalli ja klassikalise, jääkide normaaljaotust eeldava prognoosiintervalli võrdlus on toodud joonisel 8.7.

Joonis 8.7: 95%-prognoosiintervall normaaljaotuse eeldusel ja ilma selleta (nn jaotusvaba prognoosiintervall).

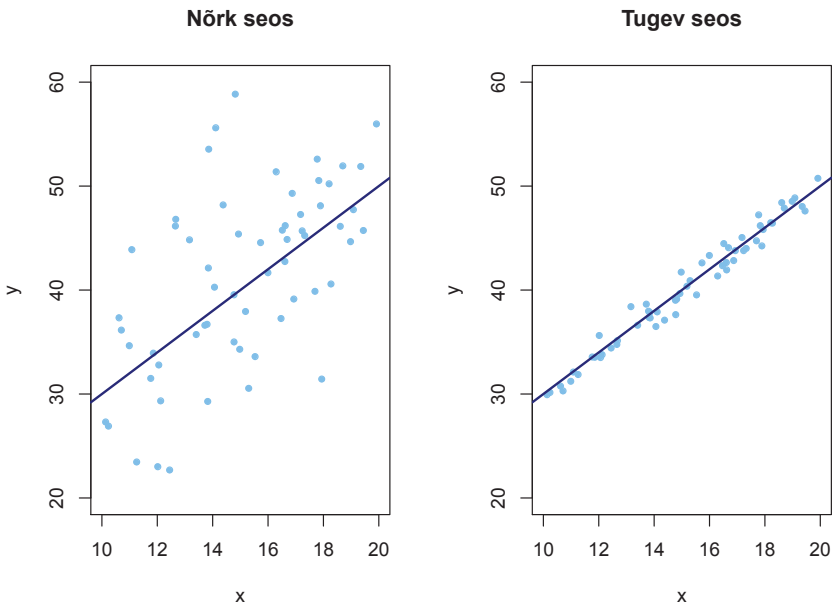


Tuleb siiski meeles pidada, et Olive jaotusvaba prognoosiintervall vabastab meid vaid ühe eelduse painest — jääkide jaotus ei pea enam olema normaaljaotus — kuid kaks järgijäänud eeldust (sama uuritava tunnuse hajuvus kõigi x-tunnuse väärtuste korral ja regressioonmudeli kuju sobivus) jäävad siiski alles.

8.5 Regressioonseose tugevus

Kui eksisteerib statistiline seos tunnuste X ja Y vahel, siis tänu X-tunnuse väärtuse teadmisele oskame midagi täpsemalt öelda ka Y-tunnuse väärtuse kohta. Aga kui palju kasu ikkagi X-tunnuse väärtuse teadmisest on? Kui tugev on seos X-tunnuse ja Y-tunnuse vahel? Kui palju täpsema prognoosi me tänu X-tunnuse väärtuse teadmisele suudame anda Y-tunnuse väärtusele? Vaata näidet tugevast seosest ja nõrgast seosest joonisel 8.8.

Joonis 8.8: Regressioonseose tugevus. Nõrk seos ja tugev seos.



8.5.1 Determinatsioonikordaja R^2

Prognoosi täpsuse kirjeldamiseks on võimalik kasutada regressioonmudeli jääkide ehk prognoosivigade dispersiooni ($\sigma_\varepsilon{}^2$). Võrdluseks võime võtta olukorra, kus pole võimalik kasutada X-tunnuse väärtust prognoosimisel. Sellisel juhul oleks parimaks mõeldavaks uue uurimisobjekti Y-tunnuse prognoosiks lihtsalt Y-tunnuse keskväärtus $\mathrm{E}Y$. Taolise naiivse prognoosi puhul oleks prognoosivigade $\varepsilon_{naiivne} = Y - \mathrm{E}Y$ dispersioon võrdne Y-tunnuse dispersiooniga, $\mathrm{D}(\varepsilon_{naiivne}) = \mathrm{D}(Y - \mathrm{E}Y) = \mathrm{D}Y$ (meenuta dispersiooni 3. oma-

dust). Seega võit prognoosi täpsuses tänu X-tunnuse väärtuse teadmisele on $\mathrm{D}Y - {\sigma_\varepsilon}^2$. Üks võimalus olekski kasutada antud näitajat seose tugevuse iseloomustamiseks, aga... tema tõlgendamine osutub sageli ebamugavaks. Kui muudetaks mõõtühikut, milles mõõdetakse Y-tunnust, muutub ka võit täpsuses (kui kaalu mõõdetakse grammides kilogrammide asemel, muutub saavutatud võit täpsuses 1000000 korda "suuremaks". Seetõttu eelistatakse mõõta suhtelist võitu täpsuses, kui suure osa Y-tunnuse hajuvusest (dispersioonist) oli võimalik ära kirjeldada tänu X-tunnusele. Saadud tulemust esitatakse sageli ka protsendina (mitu protsenti prognoosivigade hajuvus väheneb tänu X-tunnuse väärtuste teadmisele):

$$\frac{\mathrm{D}Y - {\sigma_\varepsilon}^2}{\mathrm{D}Y} \ (\cdot 100\%).$$

Antud valem kirjeldab ideaalnäitajat, mille abil võiks regressioonseose tugevust kirjeldada. Paraku seda ideaalnäitajat pole võimalik niisama lihtsalt arvutada — tema leidmiseks oleks tarvis teada uuritava tunnuse hajuvust populatsioonis ($\mathrm{D}Y$), ja prognoosivigade tegelikku dispersiooni ${\sigma_\varepsilon}^2$. Nende näitajate väärtuseid on aga peaaegu alati teadmata, tuleb kasutada nimetatud näitajate hinnanguid. Kui suuruse $\mathrm{D}Y$ hindamises ollakse enamasti üksmeelel, tema hinnanguks kasutatakse tavalist valimidispersiooni $s^2(Y) = \frac{1}{n-1}\sum (y_i - \overline{y})^2$, siis regressioonijääkide dispersiooni hindamisel esineb kaks üsnagi erinevat võimalust. Ühel juhul hinnatakse regressioonijääkide dispersiooni vägagi klassikalisel viisil — leitakse olemasolevate vaatluste prognoosimisel tehtud prognoosivead ja leitakse siis nende prognoosivigade dispersiooni hinnang kasutades harilikku valimisdispersiooni arvutamise valemit:

$$\begin{aligned} e_i &= y_i - (\hat{c}_0 + \hat{c}_1 x_i) \\ {\tilde{\sigma}_\varepsilon}^2 &= \frac{1}{n-1}\sum (e_i - \overline{e})^2 \end{aligned}$$

Viimast valemit saab veel veidi lihtsustada, sest $\overline{e} = 0$.

Tulemuseks saadakse näitaja, mida kutsutakse determinatsioonikordajaks, R^2:

$$R^2 := \frac{{s_Y}^2 - {\tilde{\sigma}_\varepsilon}^2}{{s_Y}^2} \ (\cdot 100\%)$$

Kui determinatsioonikordajat hakati juba hoolega kasutama, märgati, et antud hinnanguga on seotud üks probleem. Nimelt on suuruse ${\sigma_\varepsilon}^2$ hinnang, mida determinatsioonikordaja arvutamisel kasutatakse (${\tilde{\sigma}_\varepsilon}^2$), nihkega hinnang — ta alahindab (keskmiselt) regressioonimudeli jääkide dispersiooni.

Probleemi alge peitub selles, et regressioonmudeli parameetrid pole täpselt teada, tegemist on kõigest hinnangutega. Kuna regressioonsirge paigutatakse nii, et ta kõige paremini kirjeldaks valimis olemasolevaid vaatluseid, siis saadaksegi sirge, mis väga hästi kirjeldab olemasolevaid, regressioonsirge hindamiseks kasutatud vaatluseid. Uusi, tulevasi vaatluseid ei pruugi aga regressioonsirge enam samavõrra hästi kirjeldada — tulevased vaatlused võivad vigaselt hinnatud regressioonsirgest paikneda kaugemal, kui olemasolevate vaatluste pealt võiks arvata.

Parem, nihketa hinnang jääkide dispersioonile (tulevaste vaatluste keskmine ruutkaugus regressioonisirgest) oleks suurus

$$\hat{\sigma_\varepsilon}^2 \quad = \quad \frac{1}{n-p}\sum (e_i - \overline{e})^2$$

Kus p on regressioonsirge kirjeldamiseks hinnatavate parameetrite arv (lihtsa regressioonmudeli korral $p = 2$, sest hinnatakse vabaliige c_0 ja sirge tõus c_1). Kui kasutatakse mõistlikumat hinnangut jääkide dispersioonile saadakse näitaja, mida kutsustakse kohandatud või parandatud determinatsioonikordajaks (*adjusted* R^2) R^2_{adj}:

$$R^2_{adj} := \frac{{s_Y}^2 - \hat{\sigma_\varepsilon}^2}{{s_Y}^2} \ (\cdot 100\%).$$

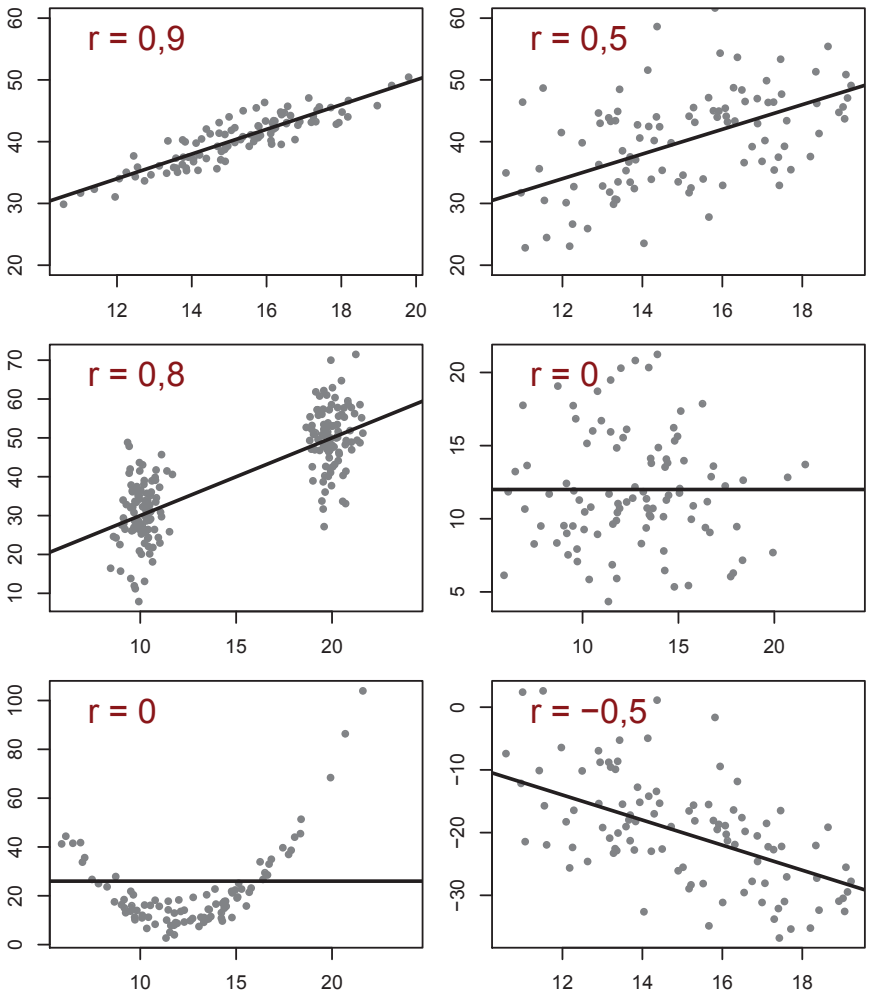
Matemaatiku vaatevinklist on kohandatud determinatsioonikordaja kasutamine paremini õigustatud kui tavalise determinatsioonikordaja kasutamine, ajaloolistel põhjustel kohtab kirjanduses siiski sageli ka tavalist determinatsioonikordajat. Kohandatud determinatsioonikordaja on alati veidi väiksem kui tavaline determinatsioonikordaja. Mõlema näitaja interpretatsioon on sama — mõlemad näitajad iseloomustavad, kui suure osa uuritava tunnuse hajuvusest on võimalik kirjeldada kasutades tunnust X ehk kui palju väheneb Y-tunnuse prognoosivigade dispersioon kui on võimalik kasutada prognoosimisel X-tunnuse väärtuseid (võrreldes olukorraga, kus X-tunnuse väärtuseid pole võimalik Y-tunnuse väärtuste prognoosimisel kasutada).

8.5.2 Lineaarne korrelatsioonikordaja r

Arvatavasti kõige sagedamini kasutatakse kahe tunnuse vahelise seose tugevuse iseloomustamiseks lineaarset korrelatsioonikordajat r (tuntud ka kui Pearsoni korrelatsioonikordaja). Korrelatsioonikordaja ruut on determinatsioonikordaja, kusjuures lineaarne korrelatsioonikordaja on positiivne, kui

ühe tunnuse väärtuste kasvades teise tunnuse väärtused kipuvad samuti kasvama. Kui aga ühe tunnuse väärtuste kasvades teise tunnuse väärtused pigem kahanevad, siis on korrelatsioonikordaja negatiivne. Näiteid erinevatest hajuvusgraafikutest ja nendele vastavatest korrelatsioonikordajatest r vaata jooniselt 8.9.

Joonis 8.9: Pearsoni lineaarne korrelatsioonikordaja. Näiteid.



Pearsoni korrelatsioonikordaja omadused:

- Kui tunnuste X ja Y vahel on lineaarne funktsionaalne seos $Y = c_0 +$

c_1X (ehk täpne lineaarne seos), siis on korrelatsioonikordaja väärtus kas 1 või -1 vastavalt kordaja c_1 märgile.

- Kui $r > 0$, siis ühe tunnuse suurenedes keskmiselt teine tunnus kasvab ja vastupidi — ühe vähenedes väheneb ka teine.
- Kui $r < 0$, siis ühe tunnuse väärtuste suurenedes keskmiselt teise tunnuse väärtused kahanevad ja vastupidi — ühe kahanedes teine kasvab.
- Kui tunnused on lineaarselt sõltumatud (tunnuste vahel võib aga olla mittelineaarne sõltuvus), siis on korrelatsioonikordaja null $r = 0$.
- Korrelatsioonikordaja ruut r^2 ehk determinatsioonikordaja $r^2 = R^2$ näitab, kui suur osa ühe tunnuse hajuvusest (dispersioonist) on kirjeldatud teise poolt (lineaarse regressioonmudeli abil).
- Mida suurem on korrelatsioonikordaja absoluutväärtus, seda tugevam on korrelatiivne seos tunnuste vahel.
- Mõõtühiku (lineaarne) vahetus ei muuda korrelatsioonikordaja suurust (Korrelatsioonikordaja ei muutu, kui mõõdame temperatuuri Celsiuse kraadide C^o asemel Farenheitides F^o, samuti võime pikkust mõõta sentimeetrites või meetrites- korrelatsioonikordaja jääb ikka samaks).

Pearsoni lineaarne korrelatsioonikordaja iseloomustab ainult lineaarse seose tugevust. Kui tunnuste X ja Y vahelist seost ei sobi kirjeldama sirge, siis võib korrelatsioonikordaja r väärtus osutuda ka nulliks või nullilähedaseks isegi siis, kui tegelikult eksisteerib tugev (mittelineaarne) statistiline seos tunnuste vahel.

Determinatsioonikordajat saab edukalt kasutada ka keerukamate mudelite juures (näiteks selliste mudelite juures, kus kasutatakse paljusid erinevaid tunnuseid Y-tunnuse prognoosimiseks) — on ju ikkagi võimalik mõõta, kui palju täpsemaks muutus prognoos tänu kasutada olevale lisainformatsioonile. Seega determinatsioonikordajat nähes võib tegemist olla nii lihtsa regressioonmudeli kui ka keeruka, paljusid erinevaid tunnuseid sisaldava regressioonmudeliga. Kui raporteeritakse Pearsoni lineaarset korrelatsioonikordajat r, siis võib olla üsna kindel, et kasutatud on kõige lihtsamat lineaarset regressioonmudelit.