Lineaarsed mudelid
9. praktikum

Prognoosiintervall

Beebid ei oska arstile kaevata, kui nendega midagi lahti on. Samuti on nad väga kohanemisvõimelised ja võivad kiiresti õppida toime tulema mitmete ka tõsiste terviseprobleemidega. Võib kergesti juhtuda, et mõndagi väikelast või beebit vaevavat tõsist terviseprobleemi lapse ema või perearst ei märka.

Kui aga laps on haige, siis haigusega võitlemine võib nõuda beebi organismilt palju jõudu ja seetõttu lapse kasvamine ja areng aeglustub. Sestap jälgivad perearstid laste kasvu – kui see pole ootuspärane, siis võib tegemist olla mõne märkamata jäänud haiguse või terviseprobleemiga (ja vastavat last tasuks põhjalikumalt uurima hakata). Selleks, et perearst oskaks aga öelda, kas lapse suurus on eakohane, läheb tal tarvis abivahendit – kasvukõverat. Kasvukõvera pealt saab välja lugeda mingis vanuses lapse ootuspärase suuruse (näiteks kaalu ja pikkuse) ning ka selle, millises vahemikus nii vanade tervete laste suurused võiksid olla.

Vaata näiteks praktikumi materjali lõpus toodud Eesti poiste kasvukõverat (mille ma kunagi noore statistikuna Eesti arstide jaoks joonistasin – tehtud veel SASiga, sest R leiutati hiljem).

Kasvukõverate loomiseks on aja jooksul mõõdetud uskumatult paljude erinevas vanuses väikelaste pikkust. Loeme sisse neid mõõtmiseid sisaldava andmestiku:

```
andmed=read.csv2("http://www.ms.ut.ee/mart/linmud2023/lapsed2.csv",
      header=T)
head(andmed)
```

Soovime saada laste keskmise pikkuse muutumist kirjeldavat kõverat koos piiridega, mis näitaks kuhu vahele jääb 50%; 80% ja 95% normaalsete (tervete) laste pikkus (0,5- ;0,8- ja 0,95-prognoosiintervallid). Mõistlik on teha eraldi graafikud poistele ja tüdrukutele. Käesolevas praktikumis üritame esmalt luua mudelit, mis kirjeldaks poisslaste kasvamist ajas.

```
m1=lm(pikkus~vanus, data=andmed[sugu=="M",])
```

Piilume ka andmeid:

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch= ".")
```

või

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch=20,
      col=rgb(1, 0, 0, 0.01), cex=1.5)
```

ja lisame joonisele oma mudeli prognoosijoone koos 95%-prognoosiintervalliga

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch= ".", col="gray70")
x=seq(0,2.5, length=100)
y=predict(m1, data.frame(vanus=x), interval="prediction")
lines(x,y[,1], col=2, lwd=3)
lines(x,y[,2], col=2, lwd=2, lty=2)
lines(x,y[,3], col=2, lwd=2, lty=2)
```

Peaksime ka sellelt jooniselt märkama, et meil esinevad teatavad probleemid mudeliga. Seos vanuse ja pikkuse vahel pole lihtsalt lineaarne.

Proovi mudelit parandada, kasutades kõrgema astme polünoomi...

```
m2=lm(pikkus~poly(vanus, 15), data=andmed[sugu=="M",])
summary(m2)
```

Kui soovime leida sobivat järku polünoomi, siis võiksime kasutada poly-käsku ilma raw=TRUE parameetrita. Näeme, et kuni 10. polünoomi astmeni on kõik kordajad statistiliselt olulised, kõrgemad astmed (aga enamasti) pole statistiliselt olulised. Võime kontrolliks proovida, kas 10-järku polünoom võiks meie andmeid samahästi kirjeldada kui kõrgemat järku polünoom:

```
m3=lm(pikkus~poly(vanus, 10), data=andmed[sugu=="M",])
anova(m2, m3)
```

Näeme, et erinevus pole statistiliselt oluline. Seega eelistame lihtsamat mudelit – 10-järku polünoomi.

Ülesanne 1

Joonista välja antud mudelile (m3) vastav regressioonsirge koos algselt soovitud prognoosiintervallidega.

Normaaljaotuse eeldus pole enamasti väga oluline eeldus – sageli võime suurte valimite korral saada korrektseid tulemusi ka siis, kui uuritava tunnuse jaotuseks pole normaaljaotus. Paraku prognoosiintervallid on üks vähestest asjadest mille arvutus võib viia väga eksitavate tulemusteni kui uuritava tunnuse (prognoosijääkide) jaotuseks pole normaaljaotus (isegi suur valim ei aita siin). Kuidas normaaljaotuse eeldusega on lood antud andmestiku korral?

R-i on käsk mis joonistab neli kõige sagedamini kasutamist leidvat diagnostilist graafikut (järgmise graafiku saamiseks näiteks kliki hiirega eelneval joonisel):

```
plot(m3)
```

Üks neist graafikutest (*Normal Q-Q plot*) on mõeldud normaaljaotuse eelduse kontrollimiseks. Kui vaatlused on normaaljaotusega (y-tunnuse tinglik jaotus tingimusel et x-tunnuste väärtused on fikseeritud on normaaljaotus) siis peaks vaatluseid iseloomustavad punktid sellel graafikul paiknema enam-vähem sirge peal. Mida näed? Kuidas on lood normaaljaotuse eeldusega?

Seda joonist vaadates peame tõdema, et mudeli jäägid pole normaaljaotusega. Punktid graafiku mõlemas servas hälbivad selgelt sirgest. Samas ei pruugi olla kõik lootusetult hukas. Graafiku keskosas püsivad punktid kenasti sirgel – mis viitab sellele, et 0,5-kvantiilile lähedased kvantiilid võiksid käituda sarnaselt normaaljaotusele. Seega võiksime oletada, et probleeme prognoosiintervallidega võiks esineda vaid väga äärmusliku katvusega prognoosiintervallide korral. Uurime seda võimalust.

Katse 1. Kas normaaljaotuse eeldusel leitud kvantiilid on sarnased mitteparameetrilisel meetodil (normaaljaotuse eeldust pole kasutatud) hinnatud kvantiilidega?

```
# Jääkide kvantiilid, hinnatud ilma normaaljaotuse eeldust kasutamata:
quantile(residuals(m3), c(0.001, 0.025, 0.05, 0.25))

# Millised on jaotusega N(0; MSE) juhuslike suuruste samad kvantiilid:
qnorm(c(0.001, 0.025, 0.05, 0.25), sd=sd(residuals(m3)))
```

Näeme, et 0,001 kvantiili puhul on erinevus suur (üle ühe sentimeetri), teiste kvantiilide puhul jääb aga erinevus väiksemaks kui 1mm (viitab võimalusele, et normaaljaotuse eeldust kasutades leitud 0,95-prognoosiintervalli piirid võiksid paikneda graafikul vähem kui 1mm jagu valesti – mis on arvatavasti praktikas ignoreeritav viga – sest vaevalt et laste pikkuseid nagunii enam kui 1mm täpsusega mõõdetakse). Samas normaaljaotuse eeldusel leitud 0,998 –prognoosiintervall võib olla väga vigane.

Kindluse mõttes proovime veel ühte võrdlust. Leiame ühe aasta vanusele lapsele 0,998- ja 0,95-prognoosiintervallid nii eeldades normaaljaotust kui ka Olive meetodit kasutades (mis võiks anda ligikaudselt õigeid tulemusi ka siis, kui vaatluste jaotuseks pole normaaljaotus). Normaaljaotuse eeldusel leitud prognoosipiirid tulevad järgmised:

0,95-prognoosiintervalliks tuleb 70,9cm … 82,6cm :

```
predict(m3, interval="prediction", level=0.95, data.frame(vanus=1))
```

ja 0,998-prognoosiintervalliks tuleb 67,5cm … 85,9cm :

```
predict(m3, interval="prediction", level=0.998, data.frame(vanus=1))
```

Millised tulevad samad prognoosiintervallid kasutades normaaljaotuse eeldust mittekasutava ligikaudse Olive meetodi korral (vt ülesanne 2)?

Olive poolt väljapakutud prognoosiintervalli saad leida järgmise arvutuseeskirja alusel:

$$[\hat{Y}_{uus} + a_n\hat{q}_{\alpha/2}; \hat{Y}_{uus} + a_n\hat{q}_{1-\alpha/2}],$$

kus

$$a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}} \cdot \sqrt{1 + \mathbf{x}_{uus}^T(\mathbf{X}^T\mathbf{X})^{-}\mathbf{x}_{uus}}$$

ja $\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}$ on hinnangud regressioonmudeli jääkide $(\alpha/2)$ ja $(1-\alpha/2)$-kvantiilile.

Ülesanne 2

Leia 0,95-prognoosiintervall ja 0,998-prognoosiintervall ühe aasta vanuste poisslaste pikkusele kasutades Olive meetodit. Võrdle ja kommenteeri saadud prognoosiintervalle normaaljaotuse eeldusel leitud intervallidega!

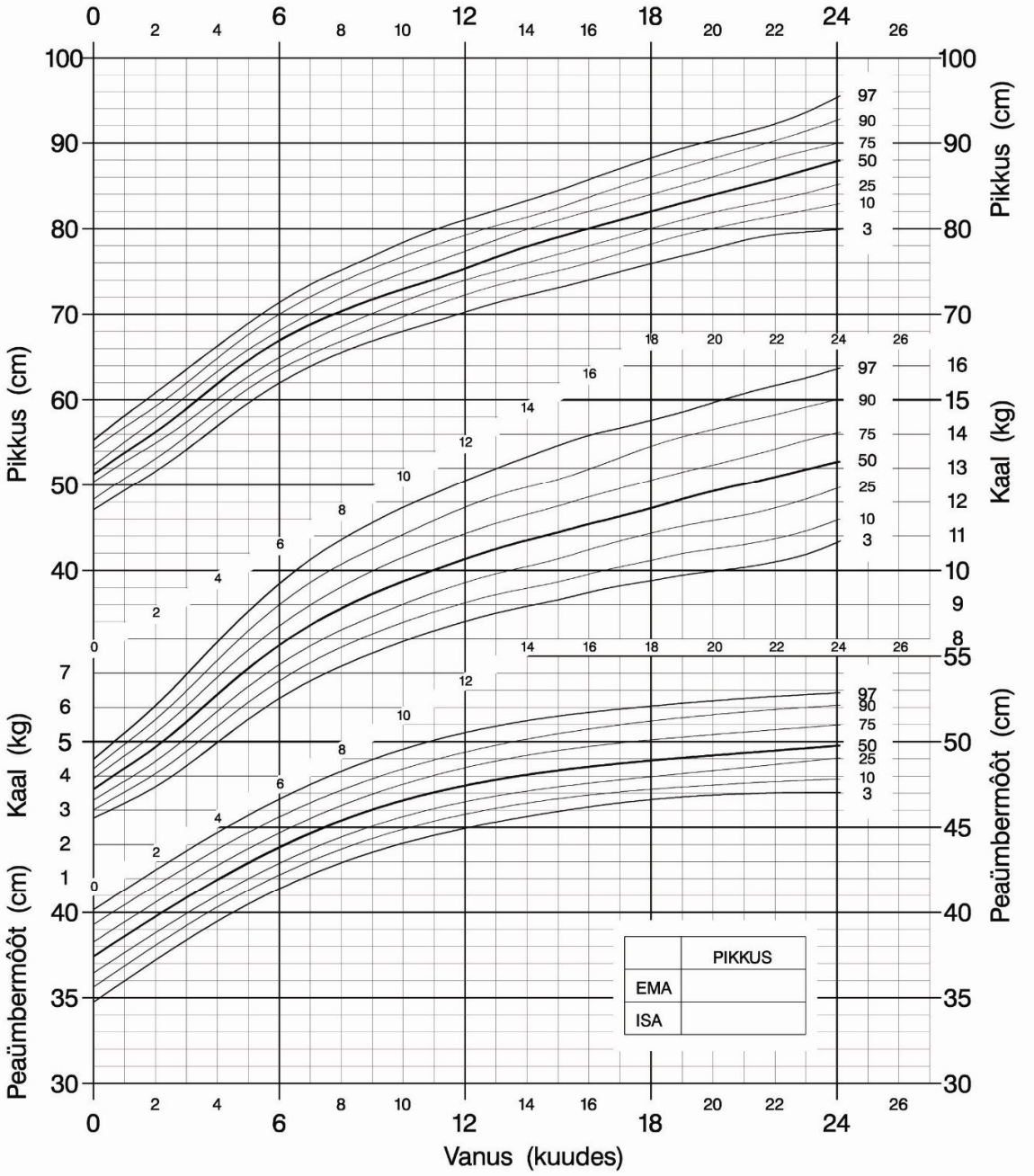
Näide prognoosiintervallide kasutamisest praktikas – kasvukõverad Eesti poisslaste jaoks.

Nimi: _____

Poeglaps

Sünniaeg: _ _._ _._ _ _ _ Isikukood: _ _ _ _ _ _ _ _ _ _ _

Kuupäev																
Kaal (kg)																
Pikkus (cm)																
Peaümbermõõt(cm)																



Lineaarsed mudelid
9. praktikum

Prognoosiintervall

Beebid ei oska arstile kaevata, kui nendega midagi lahti on. Samuti on nad väga kohanemisvõimelised ja võivad kiiresti õppida toime tulema mitmete ka tõsiste terviseprobleemidega. Võib kergesti juhtuda, et mõndagi väikelast või beebit vaevavat tõsist terviseprobleemi lapse ema või perearst ei märka.

Kui aga laps on haige, siis haigusega võitlemine võib nõuda beebi organismilt palju jõudu ja seetõttu lapse kasvamine ja areng aeglustub. Sestap jälgivad perearstid laste kasvu – kui see pole ootuspärane, siis võib tegemist olla mõne märkamata jäänud haiguse või terviseprobleemiga (ja vastavat last tasuks põhjalikumalt uurima hakata). Selleks, et perearst oskaks aga öelda, kas lapse suurus on eakohane, läheb tal tarvis abivahendit – kasvukõverat. Kasvukõvera pealt saab välja lugeda mingis vanuses lapse ootuspärase suuruse (näiteks kaalu ja pikkuse) ning ka selle, millises vahemikus nii vanade tervete laste suurused võiksid olla.

Vaata näiteks praktikumi materjali lõpus toodud Eesti poiste kasvukõverat (mille ma kunagi noore statistikuna Eesti arstide jaoks joonistasin – tehtud veel SASiga, sest R leiutati hiljem).

Kasvukõverate loomiseks on aja jooksul mõõdetud uskumatult paljude erinevas vanuses väikelaste pikkust. Loeme sisse neid mõõtmiseid sisaldava andmestiku:

```
andmed=read.csv2("http://www.ms.ut.ee/mart/linmud2023/lapsed2.csv",
      header=T)
head(andmed)
```

Soovime saada laste keskmise pikkuse muutumist kirjeldavat kõverat koos piiridega, mis näitaks kuhu vahele jääb 50%; 80% ja 95% normaalsete (tervete) laste pikkus (0,5- ;0,8- ja 0,95-prognoosiintervallid). Mõistlik on teha eraldi graafikud poistele ja tüdrukutele. Käesolevas praktikumis üritame esmalt luua mudelit, mis kirjeldaks poisslaste kasvamist ajas.

```
m1=lm(pikkus~vanus, data=andmed[sugu=="M",])
```

Piilume ka andmeid:

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch= ".")
```

või

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch=20,
      col=rgb(1, 0, 0, 0.01), cex=1.5)
```

ja lisame joonisele oma mudeli prognoosijoone koos 95%-prognoosiintervalliga

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch= ".", col="gray70")
x=seq(0,2.5, length=100)
y=predict(m1, data.frame(vanus=x), interval="prediction")
lines(x,y[,1], col=2, lwd=3)
lines(x,y[,2], col=2, lwd=2, lty=2)
lines(x,y[,3], col=2, lwd=2, lty=2)
```

Peaksime ka sellelt jooniselt märkama, et meil esinevad teatavad probleemid mudeliga. Seos vanuse ja pikkuse vahel pole lihtsalt lineaarne.

Proovi mudelit parandada, kasutades kõrgema astme polünoomi...

```
m2=lm(pikkus~poly(vanus, 15), data=andmed[sugu=="M",])
summary(m2)
```

Kui soovime leida sobivat järku polünoomi, siis võiksime kasutada poly-käsku ilma raw=TRUE parameetrita. Näeme, et kuni 10. polünoomi astmeni on kõik kordajad statistiliselt olulised, kõrgemad astmed (aga enamasti) pole statistiliselt olulised. Võime kontrolliks proovida, kas 10-järku polünoom võiks meie andmeid samahästi kirjeldada kui kõrgemat järku polünoom:

```
m3=lm(pikkus~poly(vanus, 10), data=andmed[sugu=="M",])
anova(m2, m3)
```

Näeme, et erinevus pole statistiliselt oluline. Seega eelistame lihtsamat mudelit – 10-järku polünoomi.

Ülesanne 1

Joonista välja antud mudelile (m3) vastav regressioonsirge koos algselt soovitud prognoosiintervallidega.

Normaaljaotuse eeldus pole enamasti väga oluline eeldus – sageli võime suurte valimite korral saada korrektseid tulemusi ka siis, kui uuritava tunnuse jaotuseks pole normaaljaotus. Paraku prognoosiintervallid on üks vähestest asjadest mille arvutus võib viia väga eksitavate tulemusteni kui uuritava tunnuse (prognoosijääkide) jaotuseks pole normaaljaotus (isegi suur valim ei aita siin). Kuidas normaaljaotuse eeldusega on lood antud andmestiku korral?

R-i on käsk mis joonistab neli kõige sagedamini kasutamist leidvat diagnostilist graafikut (järgmise graafiku saamiseks näiteks kliki hiirega eelneval joonisel):

```
plot(m3)
```

Üks neist graafikutest (*Normal Q-Q plot*) on mõeldud normaaljaotuse eelduse kontrollimiseks. Kui vaatlused on normaaljaotusega (y-tunnuse tinglik jaotus tingimusel et x-tunnuste väärtused on fikseeritud on normaaljaotus) siis peaks vaatluseid iseloomustavad punktid sellel graafikul paiknema enam-vähem sirge peal. Mida näed? Kuidas on lood normaaljaotuse eeldusega?

Seda joonist vaadates peame tõdema, et mudeli jäägid pole normaaljaotusega. Punktid graafiku mõlemas servas hälbivad selgelt sirgest. Samas ei pruugi olla kõik lootusetult hukas. Graafiku keskosas püsivad punktid kenasti sirgel – mis viitab sellele, et 0,5-kvantiilile lähedased kvantiilid võiksid käituda sarnaselt normaaljaotusele. Seega võiksime oletada, et probleeme prognoosiintervallidega võiks esineda vaid väga äärmusliku katvusega prognoosiintervallide korral. Uurime seda võimalust.

Katse 1. Kas normaaljaotuse eeldusel leitud kvantiilid on sarnased mitteparameetrilisel meetodil (normaaljaotuse eeldust pole kasutatud) hinnatud kvantiilidega?

```
# Jääkide kvantiilid, hinnatud ilma normaaljaotuse eeldust kasutamata:
quantile(residuals(m3), c(0.001, 0.025, 0.05, 0.25))

# Millised on jaotusega N(0; MSE) juhuslike suuruste samad kvantiilid:
qnorm(c(0.001, 0.025, 0.05, 0.25), sd=sd(residuals(m3)))
```

Näeme, et 0,001 kvantiili puhul on erinevus suur (üle ühe sentimeetri), teiste kvantiilide puhul jääb aga erinevus väiksemaks kui 1mm (viitab võimalusele, et normaaljaotuse eeldust kasutades leitud 0,95-prognoosiintervalli piirid võiksid paikneda graafikul vähem kui 1mm jagu valesti – mis on arvatavasti praktikas ignoreeritav viga – sest vaevalt et laste pikkuseid nagunii enam kui 1mm täpsusega mõõdetakse). Samas normaaljaotuse eeldusel leitud 0,998 –prognoosiintervall võib olla väga vigane.

Kindluse mõttes proovime veel ühte võrdlust. Leiame ühe aasta vanusele lapsele 0,998- ja 0,95-prognoosiintervallid nii eeldades normaaljaotust kui ka Olive meetodit kasutades (mis võiks anda ligikaudselt õigeid tulemusi ka siis, kui vaatluste jaotuseks pole normaaljaotus). Normaaljaotuse eeldusel leitud prognoosipiirid tulevad järgmised:

0,95-prognoosiintervalliks tuleb 70,9cm … 82,6cm :

```
predict(m3, interval="prediction", level=0.95, data.frame(vanus=1))
```

ja 0,998-prognoosiintervalliks tuleb 67,5cm … 85,9cm :

```
predict(m3, interval="prediction", level=0.998, data.frame(vanus=1))
```

Millised tulevad samad prognoosiintervallid kasutades normaaljaotuse eeldust mittekasutava ligikaudse Olive meetodi korral (vt ülesanne 2)?

Olive poolt väljapakutud prognoosiintervalli saad leida järgmise arvutuseeskirja alusel:

$$[\hat{Y}_{uus} + a_n\hat{q}_{\alpha/2}; \hat{Y}_{uus} + a_n\hat{q}_{1-\alpha/2}],$$

kus

$$a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}} \cdot \sqrt{1 + \mathbf{x}_{uus}^T(\mathbf{X}^T\mathbf{X})^{-}\mathbf{x}_{uus}}$$

ja $\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}$ on hinnangud regressioonmudeli jääkide $(\alpha/2)$ ja $(1-\alpha/2)$-kvantiilile.

Ülesanne 2

Leia 0,95-prognoosiintervall ja 0,998-prognoosiintervall ühe aasta vanuste poisslaste pikkusele kasutades Olive meetodit. Võrdle ja kommenteeri saadud prognoosiintervalle normaaljaotuse eeldusel leitud intervallidega!

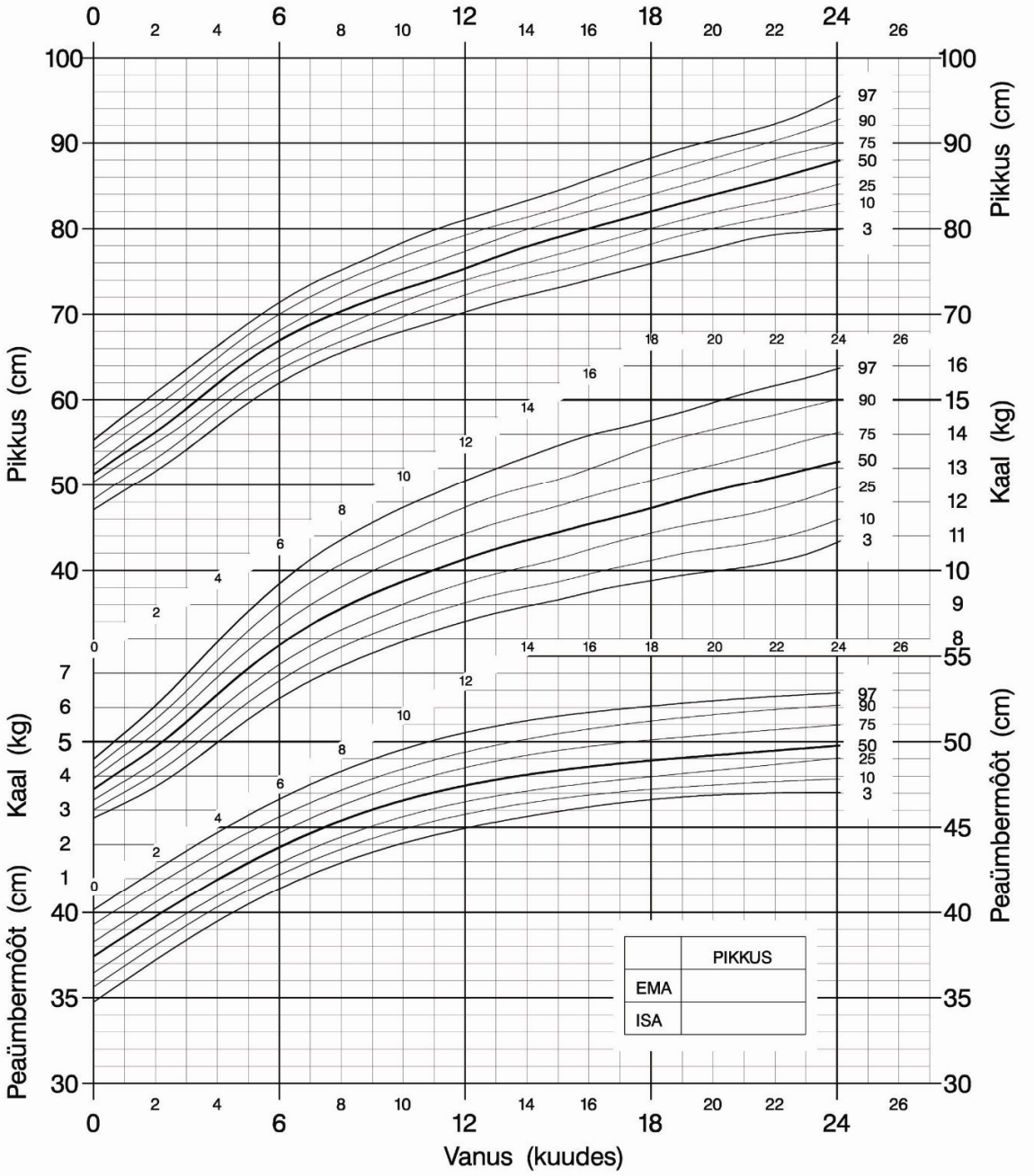
Näide prognoosiintervallide kasutamisest praktikas – kasvukõverad Eesti poisslaste jaoks.

Nimi: _____

Poeglaps

Sünniaeg: _ _._ _._ _ _ _ Isikukood: _ _ _ _ _ _ _ _ _ _ _

Kuupäev																
Kaal (kg)																
Pikkus (cm)																
Peaümbermõõt(cm)																



Lineaarsed mudelid
9. praktikum

Prognoosiintervall

Beebid ei oska arstile kaevata, kui nendega midagi lahti on. Samuti on nad väga kohanemisvõimelised ja võivad kiiresti õppida toime tulema mitmete ka tõsiste terviseprobleemidega. Võib kergesti juhtuda, et mõndagi väikelast või beebit vaevavat tõsist terviseprobleemi lapse ema või perearst ei märka.

Kui aga laps on haige, siis haigusega võitlemine võib nõuda beebi organismilt palju jõudu ja seetõttu lapse kasvamine ja areng aeglustub. Sestap jälgivad perearstid laste kasvu – kui see pole ootuspärane, siis võib tegemist olla mõne märkamata jäänud haiguse või terviseprobleemiga (ja vastavat last tasuks põhjalikumalt uurima hakata). Selleks, et perearst oskaks aga öelda, kas lapse suurus on eakohane, läheb tal tarvis abivahendit – kasvukõverat. Kasvukõvera pealt saab välja lugeda mingis vanuses lapse ootuspärase suuruse (näiteks kaalu ja pikkuse) ning ka selle, millises vahemikus nii vanade tervete laste suurused võiksid olla.

Vaata näiteks praktikumi materjali lõpus toodud Eesti poiste kasvukõverat (mille ma kunagi noore statistikuna Eesti arstide jaoks joonistasin – tehtud veel SASiga, sest R leiutati hiljem).

Kasvukõverate loomiseks on aja jooksul mõõdetud uskumatult paljude erinevas vanuses väikelaste pikkust. Loeme sisse neid mõõtmiseid sisaldava andmestiku:

```
andmed=read.csv2("http://www.ms.ut.ee/mart/linmud2023/lapsed2.csv",
      header=T)
head(andmed)
```

Soovime saada laste keskmise pikkuse muutumist kirjeldavat kõverat koos piiridega, mis näitaks kuhu vahele jääb 50%; 80% ja 95% normaalsete (tervete) laste pikkus (0,5- ;0,8- ja 0,95-prognoosiintervallid). Mõistlik on teha eraldi graafikud poistele ja tüdrukutele. Käesolevas praktikumis üritame esmalt luua mudelit, mis kirjeldaks poisslaste kasvamist ajas.

```
m1=lm(pikkus~vanus, data=andmed[sugu=="M",])
```

Piilume ka andmeid:

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch= ".")
```

või

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch=20,
      col=rgb(1, 0, 0, 0.01), cex=1.5)
```

ja lisame joonisele oma mudeli prognoosijoone koos 95%-prognoosiintervalliga

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch= ".", col="gray70")
x=seq(0,2.5, length=100)
y=predict(m1, data.frame(vanus=x), interval="prediction")
lines(x,y[,1], col=2, lwd=3)
lines(x,y[,2], col=2, lwd=2, lty=2)
lines(x,y[,3], col=2, lwd=2, lty=2)
```

Peaksime ka sellelt jooniselt märkama, et meil esinevad teatavad probleemid mudeliga. Seos vanuse ja pikkuse vahel pole lihtsalt lineaarne.

Proovi mudelit parandada, kasutades kõrgema astme polünoomi...

```
m2=lm(pikkus~poly(vanus, 15), data=andmed[sugu=="M",])
summary(m2)
```

Kui soovime leida sobivat järku polünoomi, siis võiksime kasutada poly-käsku ilma raw=TRUE parameetrita. Näeme, et kuni 10. polünoomi astmeni on kõik kordajad statistiliselt olulised, kõrgemad astmed (aga enamasti) pole statistiliselt olulised. Võime kontrolliks proovida, kas 10-järku polünoom võiks meie andmeid samahästi kirjeldada kui kõrgemat järku polünoom:

```
m3=lm(pikkus~poly(vanus, 10), data=andmed[sugu=="M",])
anova(m2, m3)
```

Näeme, et erinevus pole statistiliselt oluline. Seega eelistame lihtsamat mudelit – 10-järku polünoomi.

Ülesanne 1

Joonista välja antud mudelile (m3) vastav regressioonsirge koos algselt soovitud prognoosiintervallidega.

Normaaljaotuse eeldus pole enamasti väga oluline eeldus – sageli võime suurte valimite korral saada korrektseid tulemusi ka siis, kui uuritava tunnuse jaotuseks pole normaaljaotus. Paraku prognoosiintervallid on üks vähestest asjadest mille arvutus võib viia väga eksitavate tulemusteni kui uuritava tunnuse (prognoosijääkide) jaotuseks pole normaaljaotus (isegi suur valim ei aita siin). Kuidas normaaljaotuse eeldusega on lood antud andmestiku korral?

R-i on käsk mis joonistab neli kõige sagedamini kasutamist leidvat diagnostilist graafikut (järgmise graafiku saamiseks näiteks kliki hiirega eelneval joonisel):

```
plot(m3)
```

Üks neist graafikutest (*Normal Q-Q plot*) on mõeldud normaaljaotuse eelduse kontrollimiseks. Kui vaatlused on normaaljaotusega (y-tunnuse tinglik jaotus tingimusel et x-tunnuste väärtused on fikseeritud on normaaljaotus) siis peaks vaatluseid iseloomustavad punktid sellel graafikul paiknema enam-vähem sirge peal. Mida näed? Kuidas on lood normaaljaotuse eeldusega?

Seda joonist vaadates peame tõdema, et mudeli jäägid pole normaaljaotusega. Punktid graafiku mõlemas servas hälbivad selgelt sirgest. Samas ei pruugi olla kõik lootusetult hukas. Graafiku keskosas püsivad punktid kenasti sirgel – mis viitab sellele, et 0,5-kvantiilile lähedased kvantiilid võiksid käituda sarnaselt normaaljaotusele. Seega võiksime oletada, et probleeme prognoosiintervallidega võiks esineda vaid väga äärmusliku katvusega prognoosiintervallide korral. Uurime seda võimalust.

Katse 1. Kas normaaljaotuse eeldusel leitud kvantiilid on sarnased mitteparameetrilisel meetodil (normaaljaotuse eeldust pole kasutatud) hinnatud kvantiilidega?

```
# Jääkide kvantiilid, hinnatud ilma normaaljaotuse eeldust kasutamata:
quantile(residuals(m3), c(0.001, 0.025, 0.05, 0.25))

# Millised on jaotusega N(0; MSE) juhuslike suuruste samad kvantiilid:
qnorm(c(0.001, 0.025, 0.05, 0.25), sd=sd(residuals(m3)))
```

Näeme, et 0,001 kvantiili puhul on erinevus suur (üle ühe sentimeetri), teiste kvantiilide puhul jääb aga erinevus väiksemaks kui 1mm (viitab võimalusele, et normaaljaotuse eeldust kasutades leitud 0,95-prognoosiintervalli piirid võiksid paikneda graafikul vähem kui 1mm jagu valesti – mis on arvatavasti praktikas ignoreeritav viga – sest vaevalt et laste pikkuseid nagunii enam kui 1mm täpsusega mõõdetakse). Samas normaaljaotuse eeldusel leitud 0,998 –prognoosiintervall võib olla väga vigane.

Kindluse mõttes proovime veel ühte võrdlust. Leiame ühe aasta vanusele lapsele 0,998- ja 0,95-prognoosiintervallid nii eeldades normaaljaotust kui ka Olive meetodit kasutades (mis võiks anda ligikaudselt õigeid tulemusi ka siis, kui vaatluste jaotuseks pole normaaljaotus). Normaaljaotuse eeldusel leitud prognoosipiirid tulevad järgmised:

0,95-prognoosiintervalliks tuleb 70,9cm … 82,6cm :

```
predict(m3, interval="prediction", level=0.95, data.frame(vanus=1))
```

ja 0,998-prognoosiintervalliks tuleb 67,5cm … 85,9cm :

```
predict(m3, interval="prediction", level=0.998, data.frame(vanus=1))
```

Millised tulevad samad prognoosiintervallid kasutades normaaljaotuse eeldust mittekasutava ligikaudse Olive meetodi korral (vt ülesanne 2)?

Olive poolt väljapakutud prognoosiintervalli saad leida järgmise arvutuseeskirja alusel:

$$[\hat{Y}_{uus} + a_n\hat{q}_{\alpha/2}; \hat{Y}_{uus} + a_n\hat{q}_{1-\alpha/2}],$$

kus

$$a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}} \cdot \sqrt{1 + \mathbf{x}_{uus}^T(\mathbf{X}^T\mathbf{X})^{-}\mathbf{x}_{uus}}$$

ja $\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}$ on hinnangud regressioonmudeli jääkide $(\alpha/2)$ ja $(1-\alpha/2)$-kvantiilile.

Ülesanne 2

Leia 0,95-prognoosiintervall ja 0,998-prognoosiintervall ühe aasta vanuste poisslaste pikkusele kasutades Olive meetodit. Võrdle ja kommenteeri saadud prognoosiintervalle normaaljaotuse eeldusel leitud intervallidega!

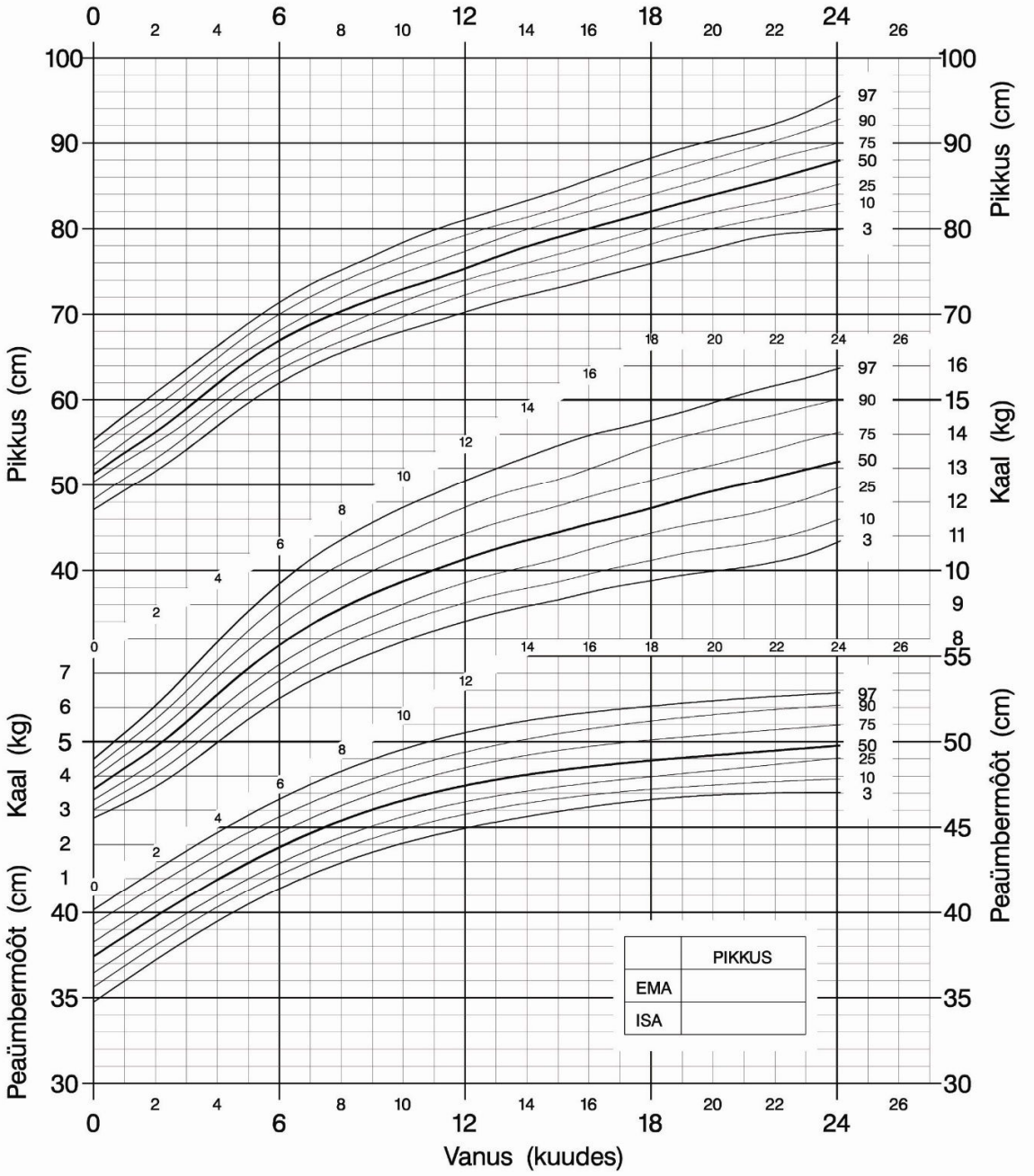
Näide prognoosiintervallide kasutamisest praktikas – kasvukõverad Eesti poisslaste jaoks.

Nimi: _____

Poeglaps

Sünniaeg: _ _._ _._ _ _ _ Isikukood: _ _ _ _ _ _ _ _ _ _ _

Kuupäev																
Kaal (kg)																
Pikkus (cm)																
Peaümbermõõt(cm)																



Lineaarsed mudelid
9. praktikum

Prognoosiintervall

Beebid ei oska arstile kaevata, kui nendega midagi lahti on. Samuti on nad väga kohanemisvõimelised ja võivad kiiresti õppida toime tulema mitmete ka tõsiste terviseprobleemidega. Võib kergesti juhtuda, et mõndagi väikelast või beebit vaevavat tõsist terviseprobleemi lapse ema või perearst ei märka.

Kui aga laps on haige, siis haigusega võitlemine võib nõuda beebi organismilt palju jõudu ja seetõttu lapse kasvamine ja areng aeglustub. Sestap jälgivad perearstid laste kasvu – kui see pole ootuspärane, siis võib tegemist olla mõne märkamata jäänud haiguse või terviseprobleemiga (ja vastavat last tasuks põhjalikumalt uurima hakata). Selleks, et perearst oskaks aga öelda, kas lapse suurus on eakohane, läheb tal tarvis abivahendit – kasvukõverat. Kasvukõvera pealt saab välja lugeda mingis vanuses lapse ootuspärase suuruse (näiteks kaalu ja pikkuse) ning ka selle, millises vahemikus nii vanade tervete laste suurused võiksid olla.

Vaata näiteks praktikumi materjali lõpus toodud Eesti poiste kasvukõverat (mille ma kunagi noore statistikuna Eesti arstide jaoks joonistasin – tehtud veel SASiga, sest R leiutati hiljem).

Kasvukõverate loomiseks on aja jooksul mõõdetud uskumatult paljude erinevas vanuses väikelaste pikkust. Loeme sisse neid mõõtmiseid sisaldava andmestiku:

```
andmed=read.csv2("http://www.ms.ut.ee/mart/linmud2023/lapsed2.csv",
      header=T)
head(andmed)
```

Soovime saada laste keskmise pikkuse muutumist kirjeldavat kõverat koos piiridega, mis näitaks kuhu vahele jääb 50%; 80% ja 95% normaalsete (tervete) laste pikkus (0,5- ;0,8- ja 0,95-prognoosiintervallid). Mõistlik on teha eraldi graafikud poistele ja tüdrukutele. Käesolevas praktikumis üritame esmalt luua mudelit, mis kirjeldaks poisslaste kasvamist ajas.

```
m1=lm(pikkus~vanus, data=andmed[sugu=="M",])
```

Piilume ka andmeid:

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch= ".")
```

või

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch=20,
      col=rgb(1, 0, 0, 0.01), cex=1.5)
```

ja lisame joonisele oma mudeli prognoosijoone koos 95%-prognoosiintervalliga

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch= ".", col="gray70")
x=seq(0,2.5, length=100)
y=predict(m1, data.frame(vanus=x), interval="prediction")
lines(x,y[,1], col=2, lwd=3)
lines(x,y[,2], col=2, lwd=2, lty=2)
lines(x,y[,3], col=2, lwd=2, lty=2)
```

Peaksime ka sellelt jooniselt märkama, et meil esinevad teatavad probleemid mudeliga. Seos vanuse ja pikkuse vahel pole lihtsalt lineaarne.

Proovi mudelit parandada, kasutades kõrgema astme polünoomi...

```
m2=lm(pikkus~poly(vanus, 15), data=andmed[sugu=="M",])
summary(m2)
```

Kui soovime leida sobivat järku polünoomi, siis võiksime kasutada poly-käsku ilma raw=TRUE parameetrita. Näeme, et kuni 10. polünoomi astmeni on kõik kordajad statistiliselt olulised, kõrgemad astmed (aga enamasti) pole statistiliselt olulised. Võime kontrolliks proovida, kas 10-järku polünoom võiks meie andmeid samahästi kirjeldada kui kõrgemat järku polünoom:

```
m3=lm(pikkus~poly(vanus, 10), data=andmed[sugu=="M",])
anova(m2, m3)
```

Näeme, et erinevus pole statistiliselt oluline. Seega eelistame lihtsamat mudelit – 10-järku polünoomi.

Ülesanne 1

Joonista välja antud mudelile (m3) vastav regressioonsirge koos algselt soovitud prognoosiintervallidega.

Normaaljaotuse eeldus pole enamasti väga oluline eeldus – sageli võime suurte valimite korral saada korrektseid tulemusi ka siis, kui uuritava tunnuse jaotuseks pole normaaljaotus. Paraku prognoosiintervallid on üks vähestest asjadest mille arvutus võib viia väga eksitavate tulemusteni kui uuritava tunnuse (prognoosijääkide) jaotuseks pole normaaljaotus (isegi suur valim ei aita siin). Kuidas normaaljaotuse eeldusega on lood antud andmestiku korral?

R-i on käsk mis joonistab neli kõige sagedamini kasutamist leidvat diagnostilist graafikut (järgmise graafiku saamiseks näiteks kliki hiirega eelneval joonisel):

```
plot(m3)
```

Üks neist graafikutest (*Normal Q-Q plot*) on mõeldud normaaljaotuse eelduse kontrollimiseks. Kui vaatlused on normaaljaotusega (y-tunnuse tinglik jaotus tingimusel et x-tunnuste väärtused on fikseeritud on normaaljaotus) siis peaks vaatluseid iseloomustavad punktid sellel graafikul paiknema enam-vähem sirge peal. Mida näed? Kuidas on lood normaaljaotuse eeldusega?

Seda joonist vaadates peame tõdema, et mudeli jäägid pole normaaljaotusega. Punktid graafiku mõlemas servas hälbivad selgelt sirgest. Samas ei pruugi olla kõik lootusetult hukas. Graafiku keskosas püsivad punktid kenasti sirgel – mis viitab sellele, et 0,5-kvantiilile lähedased kvantiilid võiksid käituda sarnaselt normaaljaotusele. Seega võiksime oletada, et probleeme prognoosiintervallidega võiks esineda vaid väga äärmusliku katvusega prognoosiintervallide korral. Uurime seda võimalust.

Katse 1. Kas normaaljaotuse eeldusel leitud kvantiilid on sarnased mitteparameetrilisel meetodil (normaaljaotuse eeldust pole kasutatud) hinnatud kvantiilidega?

```
# Jääkide kvantiilid, hinnatud ilma normaaljaotuse eeldust kasutamata:
quantile(residuals(m3), c(0.001, 0.025, 0.05, 0.25))

# Millised on jaotusega N(0; MSE) juhuslike suuruste samad kvantiilid:
qnorm(c(0.001, 0.025, 0.05, 0.25), sd=sd(residuals(m3)))
```

Näeme, et 0,001 kvantiili puhul on erinevus suur (üle ühe sentimeetri), teiste kvantiilide puhul jääb aga erinevus väiksemaks kui 1mm (viitab võimalusele, et normaaljaotuse eeldust kasutades leitud 0,95-prognoosiintervalli piirid võiksid paikneda graafikul vähem kui 1mm jagu valesti – mis on arvatavasti praktikas ignoreeritav viga – sest vaevalt et laste pikkuseid nagunii enam kui 1mm täpsusega mõõdetakse). Samas normaaljaotuse eeldusel leitud 0,998 –prognoosiintervall võib olla väga vigane.

Kindluse mõttes proovime veel ühte võrdlust. Leiame ühe aasta vanusele lapsele 0,998- ja 0,95-prognoosiintervallid nii eeldades normaaljaotust kui ka Olive meetodit kasutades (mis võiks anda ligikaudselt õigeid tulemusi ka siis, kui vaatluste jaotuseks pole normaaljaotus). Normaaljaotuse eeldusel leitud prognoosipiirid tulevad järgmised:

0,95-prognoosiintervalliks tuleb 70,9cm … 82,6cm :

```
predict(m3, interval="prediction", level=0.95, data.frame(vanus=1))
```

ja 0,998-prognoosiintervalliks tuleb 67,5cm … 85,9cm :

```
predict(m3, interval="prediction", level=0.998, data.frame(vanus=1))
```

Millised tulevad samad prognoosiintervallid kasutades normaaljaotuse eeldust mittekasutava ligikaudse Olive meetodi korral (vt ülesanne 2)?

Olive poolt väljapakutud prognoosiintervalli saad leida järgmise arvutuseeskirja alusel:

$$[\hat{Y}_{uus} + a_n\hat{q}_{\alpha/2}; \hat{Y}_{uus} + a_n\hat{q}_{1-\alpha/2}],$$

kus

$$a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}} \cdot \sqrt{1 + \mathbf{x}_{uus}^T(\mathbf{X}^T\mathbf{X})^{-}\mathbf{x}_{uus}}$$

ja $\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}$ on hinnangud regressioonmudeli jääkide $(\alpha/2)$ ja $(1-\alpha/2)$-kvantiilile.

Ülesanne 2

Leia 0,95-prognoosiintervall ja 0,998-prognoosiintervall ühe aasta vanuste poisslaste pikkusele kasutades Olive meetodit. Võrdle ja kommenteeri saadud prognoosiintervalle normaaljaotuse eeldusel leitud intervallidega!

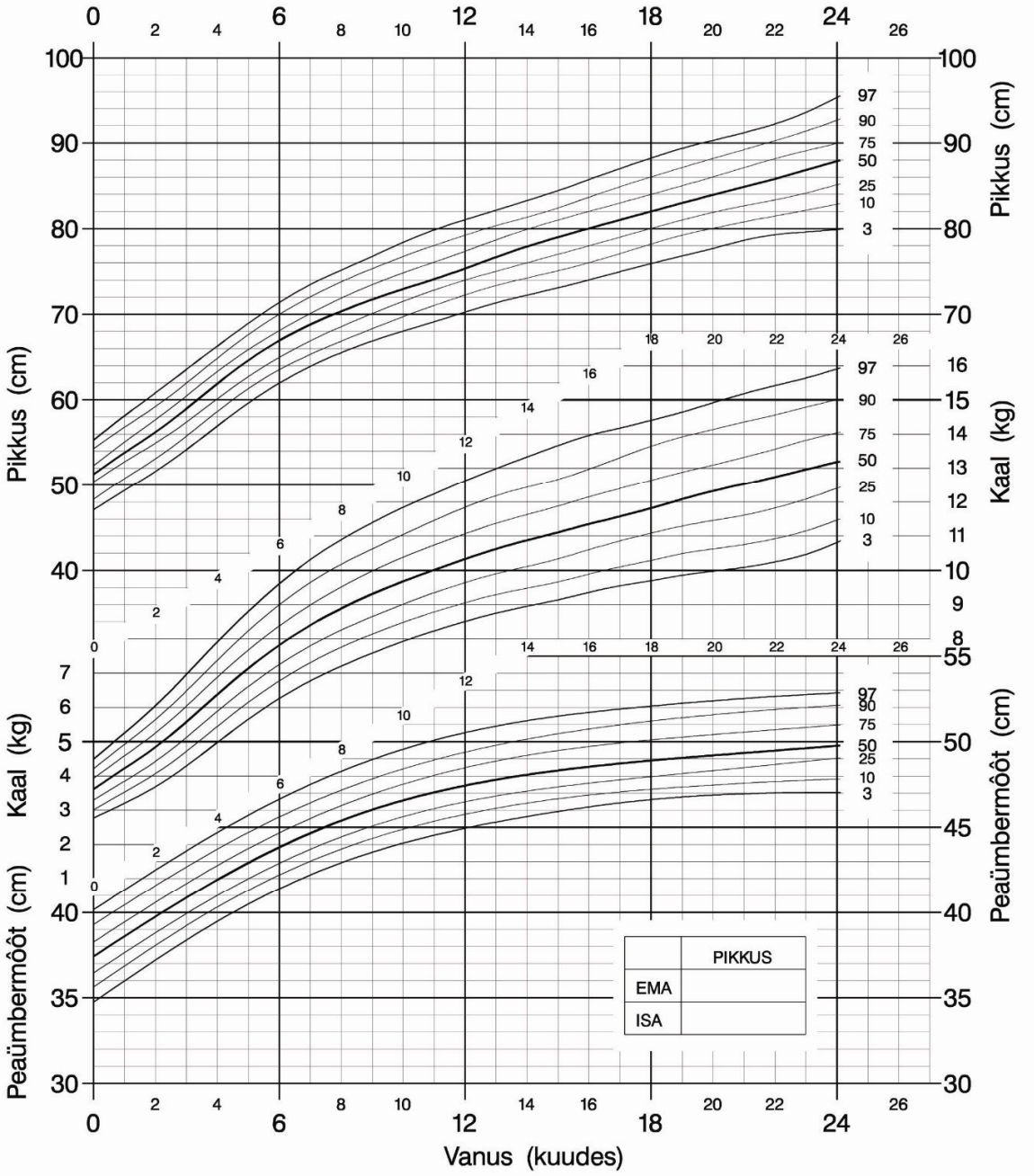
Näide prognoosiintervallide kasutamisest praktikas – kasvukõverad Eesti poisslaste jaoks.

Nimi: _____

Poeglaps

Sünniaeg: _ _._ _._ _ _ _ Isikukood: _ _ _ _ _ _ _ _ _ _ _

Kuupäev																
Kaal (kg)																
Pikkus (cm)																
Peaümbermõõt(cm)																



Lineaarsed mudelid
9. praktikum

Prognoosiintervall

Beebid ei oska arstile kaevata, kui nendega midagi lahti on. Samuti on nad väga kohanemisvõimelised ja võivad kiiresti õppida toime tulema mitmete ka tõsiste terviseprobleemidega. Võib kergesti juhtuda, et mõndagi väikelast või beebit vaevavat tõsist terviseprobleemi lapse ema või perearst ei märka.

Kui aga laps on haige, siis haigusega võitlemine võib nõuda beebi organismilt palju jõudu ja seetõttu lapse kasvamine ja areng aeglustub. Sestap jälgivad perearstid laste kasvu – kui see pole ootuspärane, siis võib tegemist olla mõne märkamata jäänud haiguse või terviseprobleemiga (ja vastavat last tasuks põhjalikumalt uurima hakata). Selleks, et perearst oskaks aga öelda, kas lapse suurus on eakohane, läheb tal tarvis abivahendit – kasvukõverat. Kasvukõvera pealt saab välja lugeda mingis vanuses lapse ootuspärase suuruse (näiteks kaalu ja pikkuse) ning ka selle, millises vahemikus nii vanade tervete laste suurused võiksid olla.

Vaata näiteks praktikumi materjali lõpus toodud Eesti poiste kasvukõverat (mille ma kunagi noore statistikuna Eesti arstide jaoks joonistasin – tehtud veel SASiga, sest R leiutati hiljem).

Kasvukõverate loomiseks on aja jooksul mõõdetud uskumatult paljude erinevas vanuses väikelaste pikkust. Loeme sisse neid mõõtmiseid sisaldava andmestiku:

```
andmed=read.csv2("http://www.ms.ut.ee/mart/linmud2023/lapsed2.csv",
      header=T)
head(andmed)
```

Soovime saada laste keskmise pikkuse muutumist kirjeldavat kõverat koos piiridega, mis näitaks kuhu vahele jääb 50%; 80% ja 95% normaalsete (tervete) laste pikkus (0,5- ;0,8- ja 0,95-prognoosiintervallid). Mõistlik on teha eraldi graafikud poistele ja tüdrukutele. Käesolevas praktikumis üritame esmalt luua mudelit, mis kirjeldaks poisslaste kasvamist ajas.

```
m1=lm(pikkus~vanus, data=andmed[sugu=="M",])
```

Piilume ka andmeid:

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch= ".")
```

või

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch=20,
      col=rgb(1, 0, 0, 0.01), cex=1.5)
```

ja lisame joonisele oma mudeli prognoosijoone koos 95%-prognoosiintervalliga

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch= ".", col="gray70")
x=seq(0,2.5, length=100)
y=predict(m1, data.frame(vanus=x), interval="prediction")
lines(x,y[,1], col=2, lwd=3)
lines(x,y[,2], col=2, lwd=2, lty=2)
lines(x,y[,3], col=2, lwd=2, lty=2)
```

Peaksime ka sellelt jooniselt märkama, et meil esinevad teatavad probleemid mudeliga. Seos vanuse ja pikkuse vahel pole lihtsalt lineaarne.

Proovi mudelit parandada, kasutades kõrgema astme polünoomi...

```
m2=lm(pikkus~poly(vanus, 15), data=andmed[sugu=="M",])
summary(m2)
```

Kui soovime leida sobivat järku polünoomi, siis võiksime kasutada poly-käsku ilma raw=TRUE parameetrita. Näeme, et kuni 10. polünoomi astmeni on kõik kordajad statistiliselt olulised, kõrgemad astmed (aga enamasti) pole statistiliselt olulised. Võime kontrolliks proovida, kas 10-järku polünoom võiks meie andmeid samahästi kirjeldada kui kõrgemat järku polünoom:

```
m3=lm(pikkus~poly(vanus, 10), data=andmed[sugu=="M",])
anova(m2, m3)
```

Näeme, et erinevus pole statistiliselt oluline. Seega eelistame lihtsamat mudelit – 10-järku polünoomi.

Ülesanne 1

Joonista välja antud mudelile (m3) vastav regressioonsirge koos algselt soovitud prognoosiintervallidega.

Normaaljaotuse eeldus pole enamasti väga oluline eeldus – sageli võime suurte valimite korral saada korrektseid tulemusi ka siis, kui uuritava tunnuse jaotuseks pole normaaljaotus. Paraku prognoosiintervallid on üks vähestest asjadest mille arvutus võib viia väga eksitavate tulemusteni kui uuritava tunnuse (prognoosijääkide) jaotuseks pole normaaljaotus (isegi suur valim ei aita siin). Kuidas normaaljaotuse eeldusega on lood antud andmestiku korral?

R-i on käsk mis joonistab neli kõige sagedamini kasutamist leidvat diagnostilist graafikut (järgmise graafiku saamiseks näiteks kliki hiirega eelneval joonisel):

```
plot(m3)
```

Üks neist graafikutest (*Normal Q-Q plot*) on mõeldud normaaljaotuse eelduse kontrollimiseks. Kui vaatlused on normaaljaotusega (y-tunnuse tinglik jaotus tingimusel et x-tunnuste väärtused on fikseeritud on normaaljaotus) siis peaks vaatluseid iseloomustavad punktid sellel graafikul paiknema enam-vähem sirge peal. Mida näed? Kuidas on lood normaaljaotuse eeldusega?

Seda joonist vaadates peame tõdema, et mudeli jäägid pole normaaljaotusega. Punktid graafiku mõlemas servas hälbivad selgelt sirgest. Samas ei pruugi olla kõik lootusetult hukas. Graafiku keskosas püsivad punktid kenasti sirgel – mis viitab sellele, et 0,5-kvantiilile lähedased kvantiilid võiksid käituda sarnaselt normaaljaotusele. Seega võiksime oletada, et probleeme prognoosiintervallidega võiks esineda vaid väga äärmusliku katvusega prognoosiintervallide korral. Uurime seda võimalust.

Katse 1. Kas normaaljaotuse eeldusel leitud kvantiilid on sarnased mitteparameetrilisel meetodil (normaaljaotuse eeldust pole kasutatud) hinnatud kvantiilidega?

```
# Jääkide kvantiilid, hinnatud ilma normaaljaotuse eeldust kasutamata:
quantile(residuals(m3), c(0.001, 0.025, 0.05, 0.25))

# Millised on jaotusega N(0; MSE) juhuslike suuruste samad kvantiilid:
qnorm(c(0.001, 0.025, 0.05, 0.25), sd=sd(residuals(m3)))
```

Näeme, et 0,001 kvantiili puhul on erinevus suur (üle ühe sentimeetri), teiste kvantiilide puhul jääb aga erinevus väiksemaks kui 1mm (viitab võimalusele, et normaaljaotuse eeldust kasutades leitud 0,95-prognoosiintervalli piirid võiksid paikneda graafikul vähem kui 1mm jagu valesti – mis on arvatavasti praktikas ignoreeritav viga – sest vaevalt et laste pikkuseid nagunii enam kui 1mm täpsusega mõõdetakse). Samas normaaljaotuse eeldusel leitud 0,998 –prognoosiintervall võib olla väga vigane.

Kindluse mõttes proovime veel ühte võrdlust. Leiame ühe aasta vanusele lapsele 0,998- ja 0,95-prognoosiintervallid nii eeldades normaaljaotust kui ka Olive meetodit kasutades (mis võiks anda ligikaudselt õigeid tulemusi ka siis, kui vaatluste jaotuseks pole normaaljaotus). Normaaljaotuse eeldusel leitud prognoosipiirid tulevad järgmised:

0,95-prognoosiintervalliks tuleb 70,9cm … 82,6cm :

```
predict(m3, interval="prediction", level=0.95, data.frame(vanus=1))
```

ja 0,998-prognoosiintervalliks tuleb 67,5cm … 85,9cm :

```
predict(m3, interval="prediction", level=0.998, data.frame(vanus=1))
```

Millised tulevad samad prognoosiintervallid kasutades normaaljaotuse eeldust mittekasutava ligikaudse Olive meetodi korral (vt ülesanne 2)?

Olive poolt väljapakutud prognoosiintervalli saad leida järgmise arvutuseeskirja alusel:

$$[\hat{Y}_{uus} + a_n\hat{q}_{\alpha/2}; \hat{Y}_{uus} + a_n\hat{q}_{1-\alpha/2}],$$

kus

$$a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}} \cdot \sqrt{1 + \mathbf{x}_{uus}^T(\mathbf{X}^T\mathbf{X})^{-}\mathbf{x}_{uus}}$$

ja $\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}$ on hinnangud regressioonmudeli jääkide $(\alpha/2)$ ja $(1-\alpha/2)$-kvantiilile.

Ülesanne 2

Leia 0,95-prognoosiintervall ja 0,998-prognoosiintervall ühe aasta vanuste poisslaste pikkusele kasutades Olive meetodit. Võrdle ja kommenteeri saadud prognoosiintervalle normaaljaotuse eeldusel leitud intervallidega!

Näide prognoosiintervallide kasutamisest praktikas – kasvukõverad Eesti poisslaste jaoks.

Nimi: _____

Poeglaps

Sünniaeg: _ _._ _._ _ _ _ Isikukood: _ _ _ _ _ _ _ _ _ _ _

Kuupäev																
Kaal (kg)																
Pikkus (cm)																
Peaümbermõõt(cm)