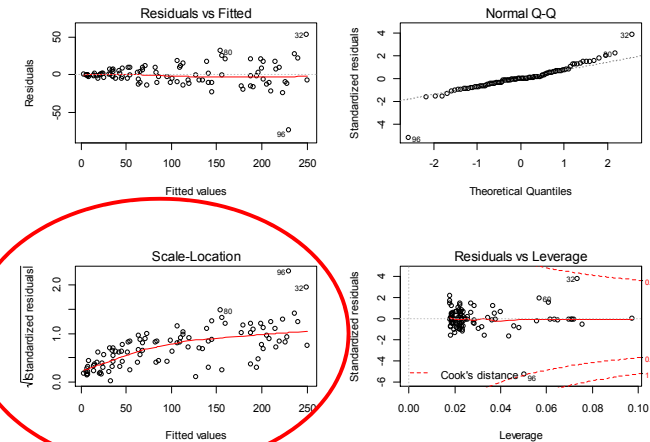


Eeldustest ja probleemide vältimisest

Lineaarsed mudelid

Konstantse hajuvuse eeldusest
(homoscedasticity)



Lineaarsed mudelid

Mis juhtub, kui eeldus on rikutud?

$$\begin{aligned} E(\mathbf{X}\hat{\beta}) &= \mathbf{P}_\mathbf{X}E(\mathbf{y}) \\ &= \mathbf{P}_\mathbf{X}\mathbf{X}\beta \\ &= \mathbf{X}\beta \end{aligned}$$

hinnang ikka nihketa

$$\begin{aligned} D(\mathbf{X}\hat{\beta}) &= \mathbf{P}_\mathbf{X}D(\mathbf{y})\mathbf{P}_\mathbf{X} \\ &= \mathbf{P}_\mathbf{X}\mathbf{V}\mathbf{P}_\mathbf{X} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &\neq \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2 \end{aligned}$$

hinnangu dispersioon
hinnatakse valesti...

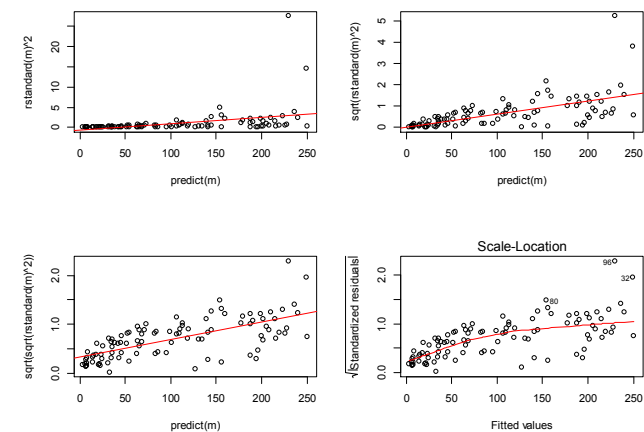
probleem vaid siis, kui
 $\mathbf{X}^T\mathbf{V}\mathbf{X} \neq \mathbf{X}^T\mathbf{X}\sigma^2$

leidub täpsemaid
hinnanguid

testide võimsus
kannatab...

Lineaarsed mudelid

Testimine – joonise abil



Lineaarsed mudelid

Paganatest

(Breusch-Pagan test)

```
> library(lmtest)
> bptest(mudel)
```

studentized Breusch-Pagan test

data: mudel
BP = 17.049, df = 2, p-value = 0.0001985

Hinda mudel $y = X\beta + \varepsilon$

Leia (tavalised) jäägid $e = y - \hat{y}$

Hinda mudel $e^2 = X\beta^* + \varepsilon^*$

Idee: võrdle, kas parem kui lihtsalt konstantne mudel...

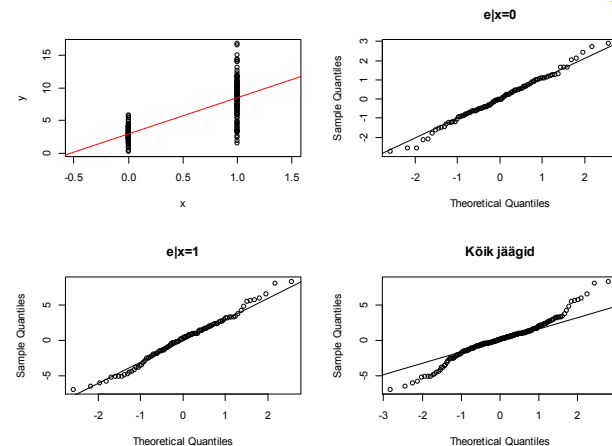
Aga: normaaljaotuse eeldusega veidi pahasti...

Osaline lahendus:

Teststatistik = $R^2 \cdot n$

H_0 kehtides: $R^2 \cdot n \sim \chi^2_{df=n-p+1}$

Kui jäägid ei ole sama hajuvusega, siis eksivad normaaljaotuse eeldust kontrollivad meetodid/testid



Miks asjad on viltu?

Näide...

Äkki vead on multiplikatiivsed? $Y = f(x_1, x_2, \dots) \cdot \varepsilon$

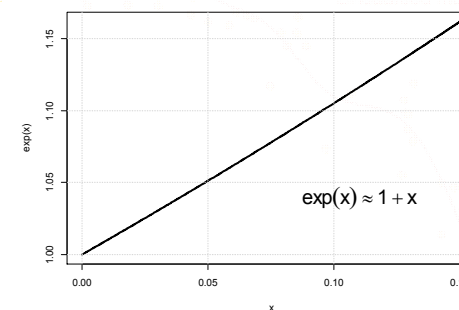
„laps võib kaaluda 10% rohkem või vähem kui tüüpkaal tema vanuses ette näeb...”

10% kõikumine funktsiooni $f(x_1, x_2, \dots)$ suure väärtuse korral viib suurema dispersioonini kui väikse funktsiooni f väärtuse korral.

$\log(Y) = \log(f(x_1, x_2, \dots)) + \log(\varepsilon)$

Miks asjad on viltu?

Näide...



```
> summary(mudel)
```

Call:

```
lm(formula = log(y) ~ vanus + I(vanus^2) + I(vanus^3) + factor(sugu))
```

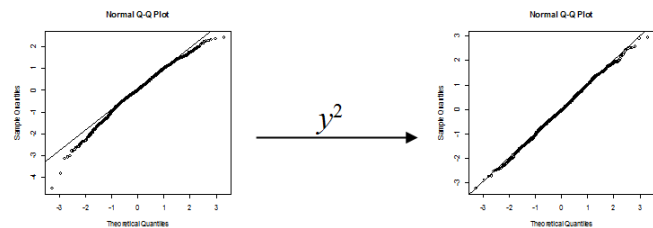
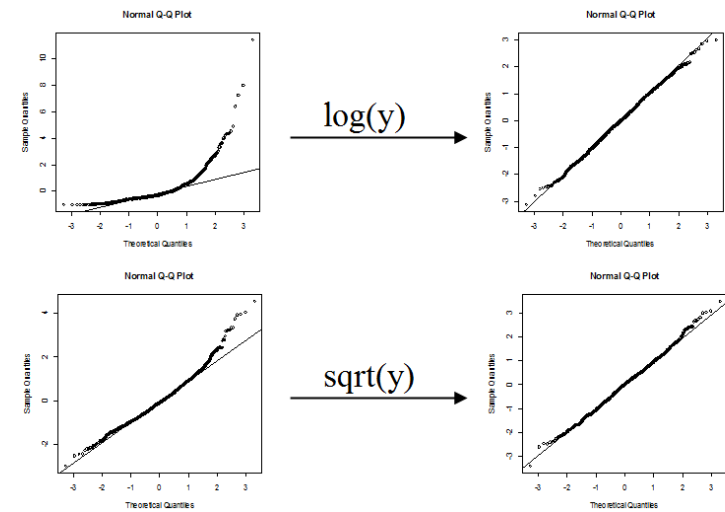
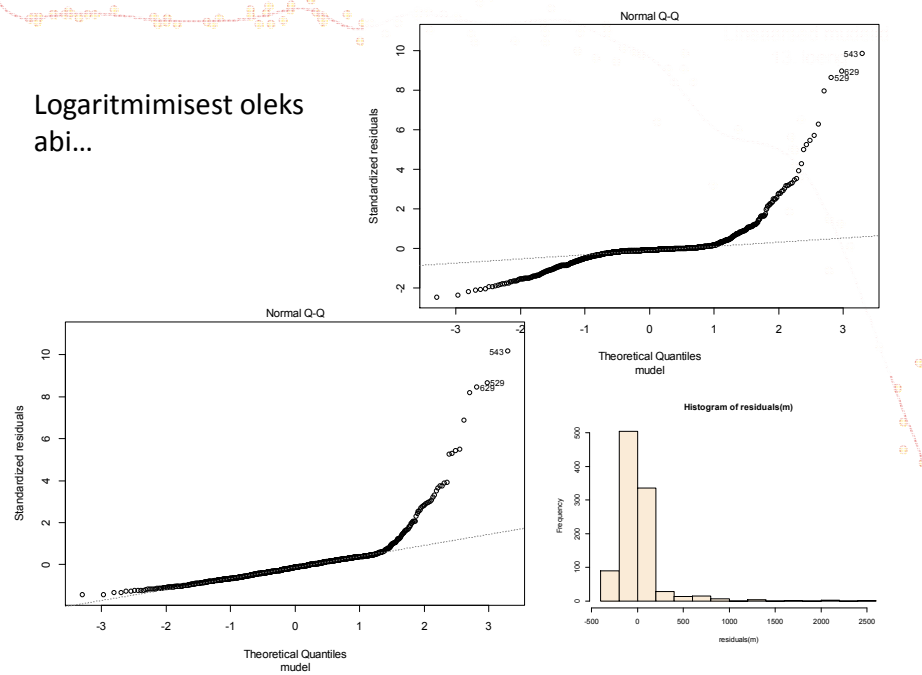
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-9.9071484	0.1157983	-85.56	< 2e-16 ***
vanus	0.7676614	0.0597481	12.85	< 2e-16 ***
I(vanus^2)	0.1025314	0.0091886	11.16	< 2e-16 ***
I(vanus^3)	0.0499722	0.0004321	115.65	< 2e-16 ***
factor(sugu) tüdruk	0.0552633	0.0161596	3.42	0.000652 ***

Tüdrukutel on Y-tunnuse väärtused 5,5% suuremad kui sama vanadel poistel...

Kuidas tõlgendada?

Logaritmisest oleks
abi...



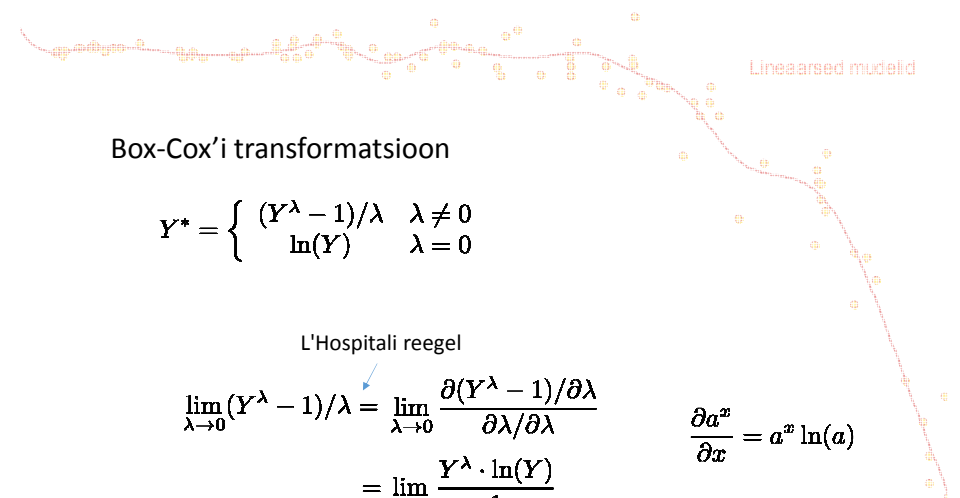
Box-Cox'i transformatsioon

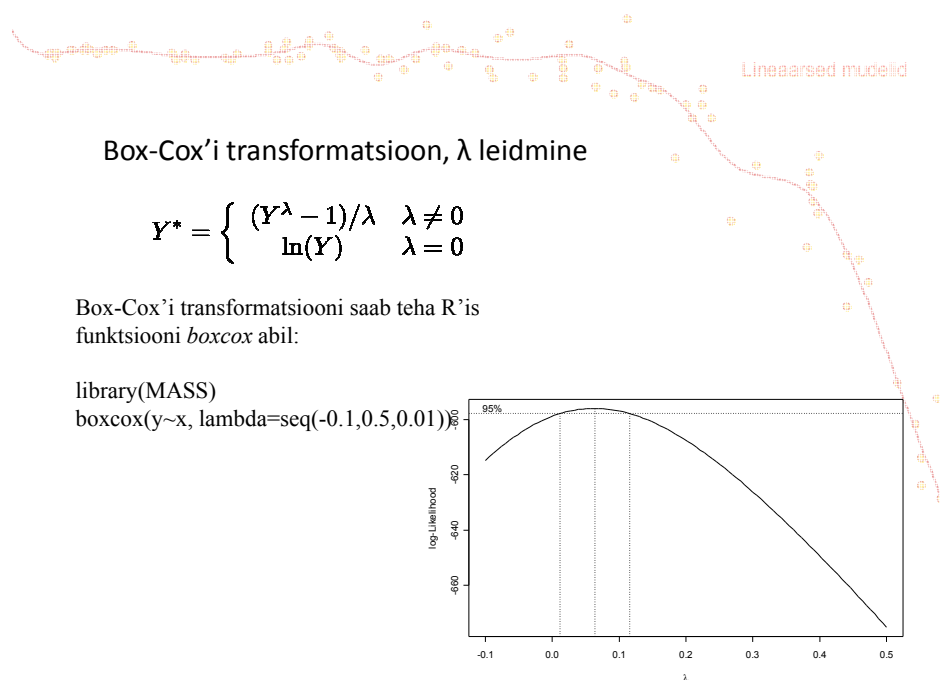
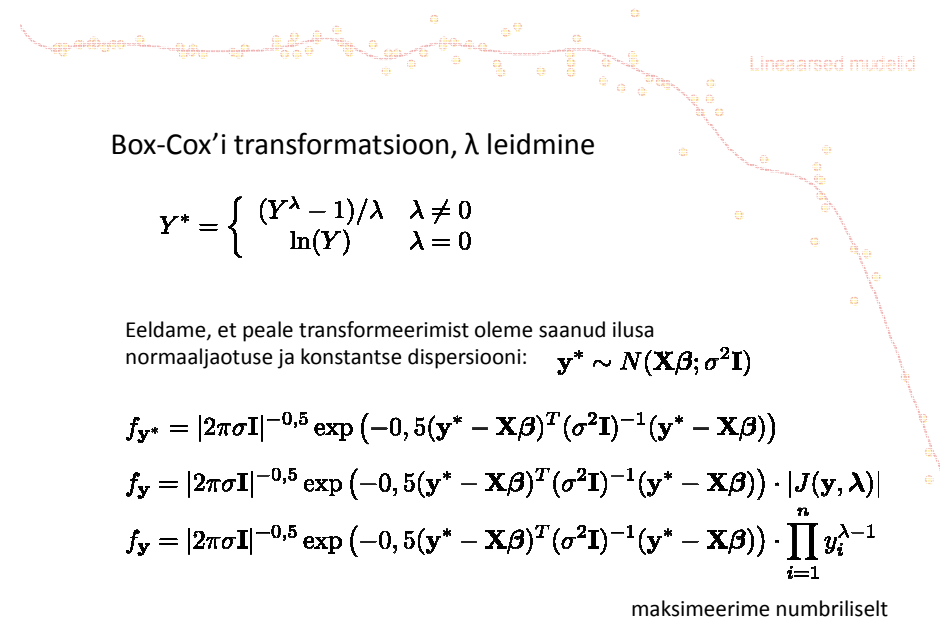
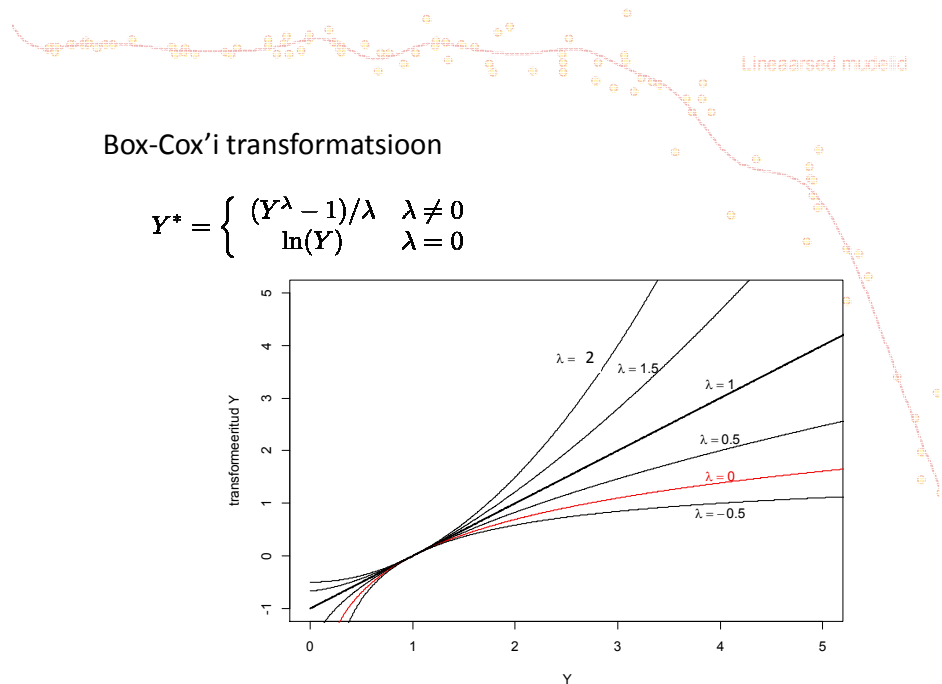
$$Y^* = \begin{cases} (Y^\lambda - 1)/\lambda & \lambda \neq 0 \\ \ln(Y) & \lambda = 0 \end{cases}$$

L'Hospitali reegel

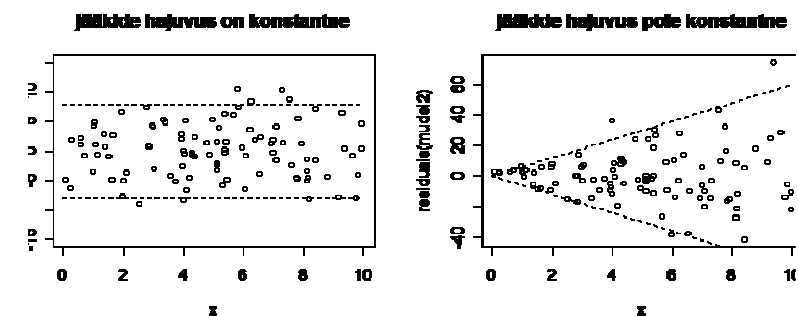
$$\begin{aligned} \lim_{\lambda \rightarrow 0} (Y^\lambda - 1)/\lambda &= \lim_{\lambda \rightarrow 0} \frac{\partial(Y^\lambda - 1)/\partial \lambda}{\partial \lambda / \partial \lambda} \\ &= \lim_{\lambda \rightarrow 0} \frac{Y^\lambda \cdot \ln(Y)}{1} \\ &= \ln(Y) \end{aligned}$$

$$\frac{\partial a^x}{\partial x} = a^x \ln(a)$$



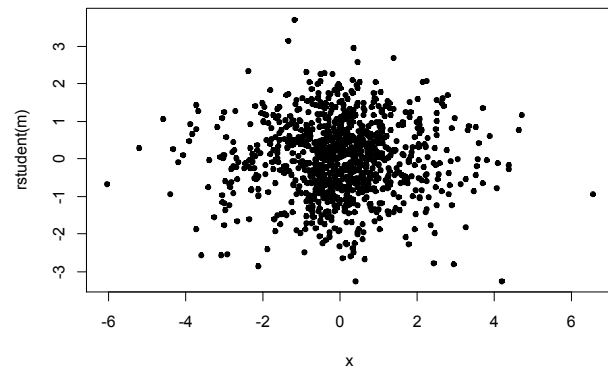


Õpikunäide tundub lihtne...



Praktikatikas...

Näide: jääkide hajuvus on konstantne

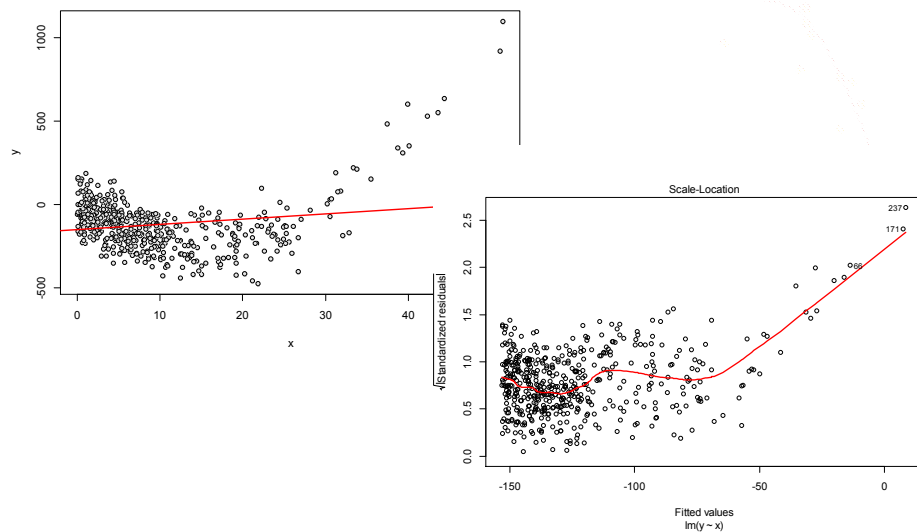


Valimi miinimum on (enamasti) seda väiksem, mida suurem on valim

Valimi maksimum on (enamasti) seda suurem, mida suurem on valim

Piirkonnas, kus on palju vaatluseid, on hajuvusgraafiku ulatus – maksimumi ja miinimumi vahe – suurem. Sestap tundub inimsilmale ka seal piirkonnas jääkide varieeruvus olevat suurem...

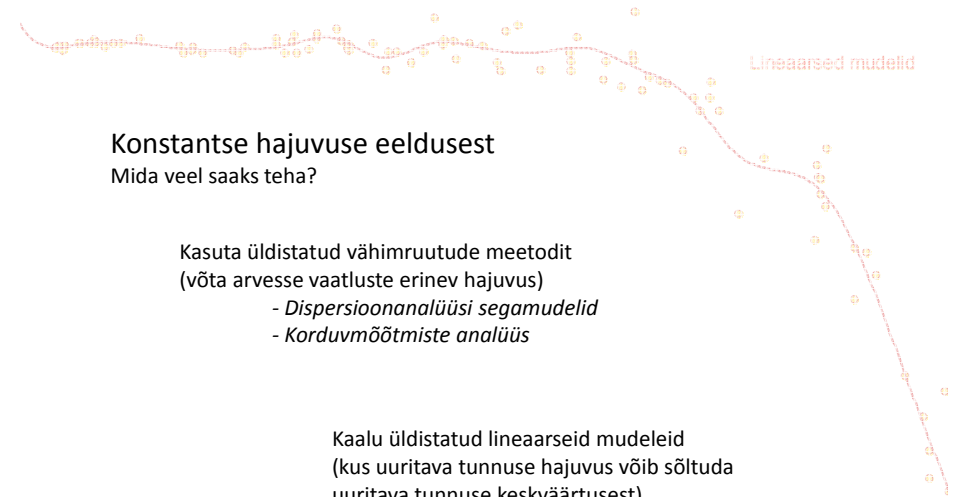
Vale mudel võib viia arvamuseni, et konstantse hajuvuse eeldust on riknud...



Konstantse hajuvuse eeldusest
Mida veel saaks teha?

Kasuta üldistatud vähimruutude meetodit
(võta arvesse vaatluste erinev hajuvus)
- Dispersioonanalüüsi segamudelid
- Korduvmõõtmiste analüüs

Kaalu üldistatud lineaarseid mudeleid
(kus uuritava tunnuse hajuvus võib sõltuda uuritava tunnuse keskväärtusest)
- Üldistatud lineaarsed mudelid

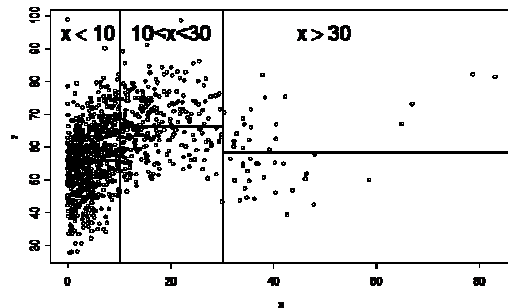


Mudeli kuju valikust – tundmatu seose lähendamine

Toivo Leiger,
Matemaatiline analüüs III

Lause 3.18 Olgu f lõigus $[a, b]$ pidev funktsioon. Siis iga $\varepsilon > 0$ jaoks leidub selline treppfunktsioon $h : [a, b] \rightarrow \mathbb{R}$, mis rahuldab tingimust

$$|f(x) - h(x)| < \varepsilon \text{ iga } x \in [a, b] \text{ korral.}$$

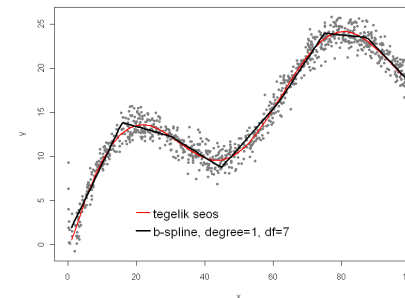


Mudeli kuju valikust – tundmatu seose lähendamine

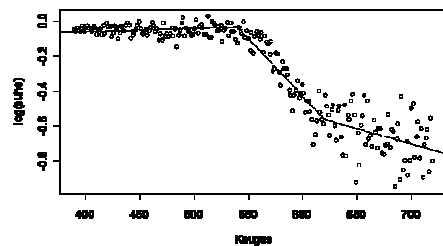
Toivo Leiger,
Matemaatiline analüüs III

Lause 3.19 Olgu f lõigus $[a, b]$ pidev funktsioon. Siis leidub iga $\varepsilon > 0$ jaoks lõigus $[a, b]$ tükati lineaarne funktsioon h , mis rahuldab tingimust

$$|f(x) - h(x)| < \varepsilon \text{ iga } x \in [a, b] \text{ korral.}$$



B-splainid

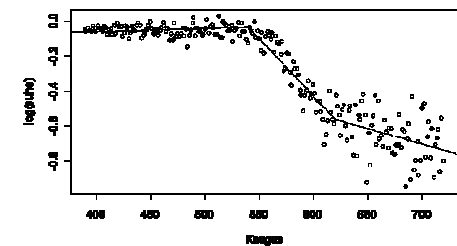


$$\begin{aligned} \text{abi1} &= I(\text{Kaugus} > 540) * (\text{Kaugus} - 540) \\ \text{abi2} &= I(\text{Kaugus} > 620) * (\text{Kaugus} - 620) \end{aligned}$$

ja lasta seejärel hinnata järgmise lineaarse mudeli parameetrid:

$$\log(\text{suhe}) = c_0 + c_1 \text{Kaugus} + c_2 \text{abi1} + c_3 \text{abi2} + e.$$

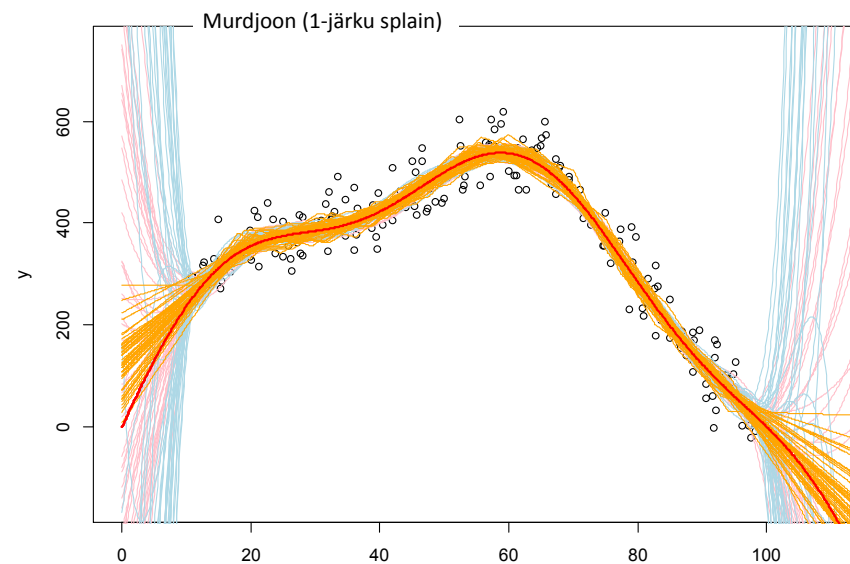
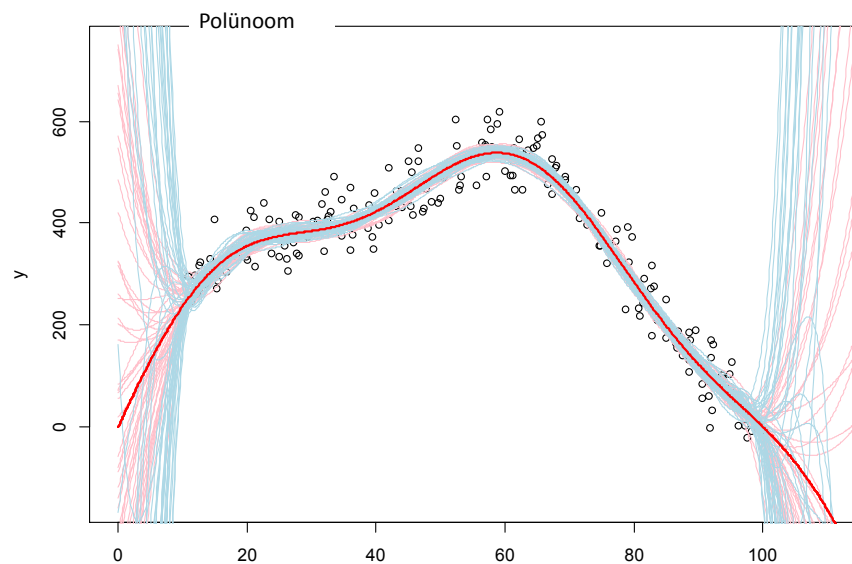
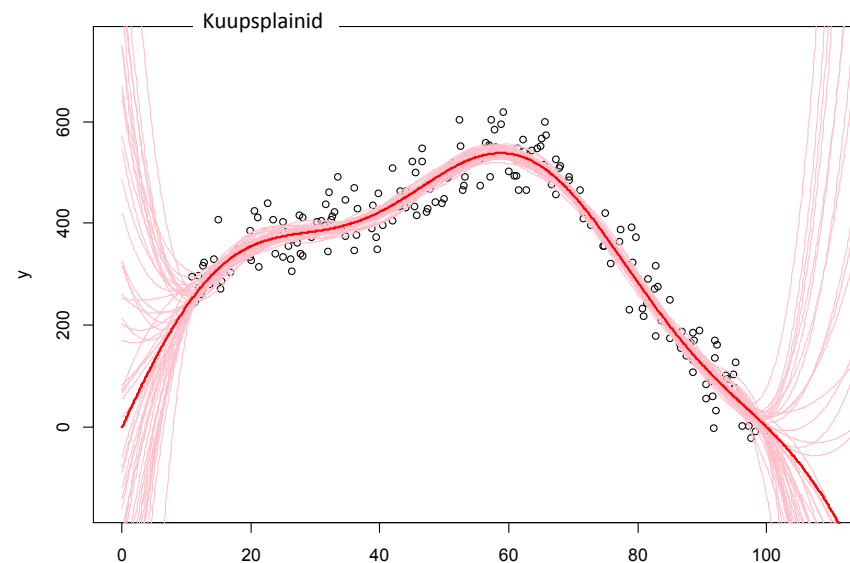
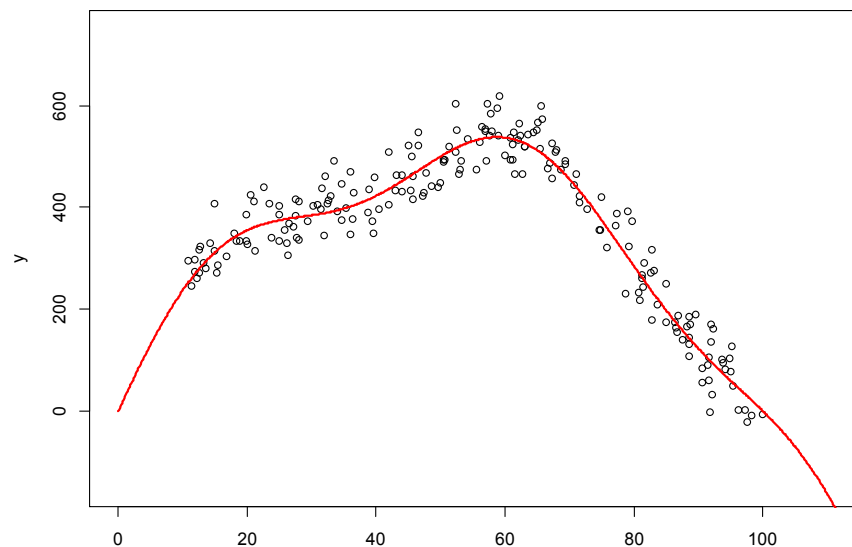
B-spliiinid (ruutsplain)

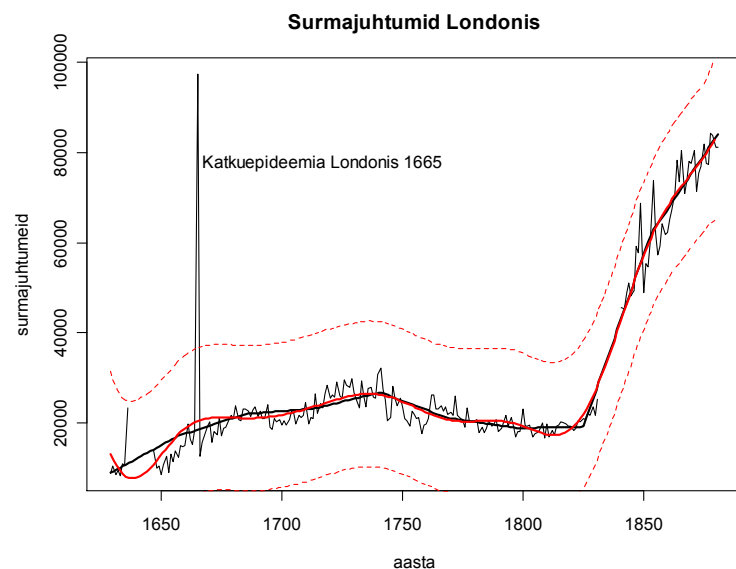


$$\begin{aligned} \text{abi1} &= I(\text{Kaugus} > 540) * (\text{Kaugus} - 540)^2 \\ \text{abi2} &= I(\text{Kaugus} > 620) * (\text{Kaugus} - 620)^2 \end{aligned}$$

ja lasta seejärel hinnata järgmise lineaarse mudeli parameetrid:

$$\log(\text{suhe}) = c_0 + c_1 \text{Kaugus} + c_2 \text{Kaugus}^2 + c_3 \text{abi1} + c_4 \text{abi2} + e.$$

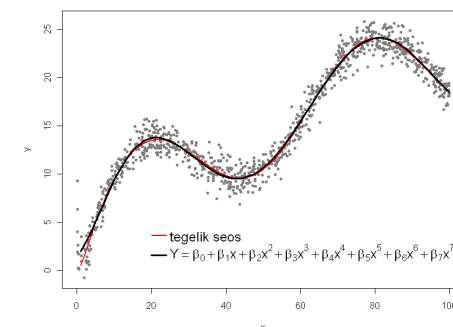




Mudeli kuju valikust – tundmatu seose lähendamine

Toivo Leiger,
Matemaatiline analüüs III

Teoreem 6.26 (Weierstrassi lähendusteoreem). Iga lõigus $[a, b]$ pideva funktsiooni f korral leidub selline polünoomide jada (P_n) , mis koondub funktsiooniks f ühtlaselt lõigus $[a, b]$.



Mudeli kuju valikust – tundmatu seose lähendamine Sessoonsuse modelleerimine

Olgu tunnus X näiteks päev aastas (1..365). Tahame iseloomustada mingi näitaja Y sõltuvust aastaajast (sessoonset – aastaaegade tingitud - muutumist, mitte pikaajalist trendi). Tahame, et kohal $X=366$ oleks graafik sealsamas kus päeval $X=1$ (sest mõlemad näitavad sama kohta ajas). Seega $X+365=X$.

Siis teeme abitunnused järgmise eeskirja alusel:

$$X_1 = \cos(2 \cdot \pi / 365 \cdot X); \quad X_2 = \sin(2 \cdot \pi / 365 \cdot X)$$

Vajadusel lisa veel täiendavaid tunnuseid:

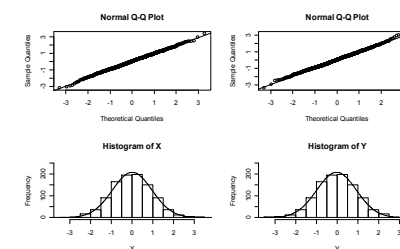
$$X_3 = \cos(2 \cdot 2 \cdot \pi / 365 \cdot X); \quad X_4 = \sin(2 \cdot 2 \cdot \pi / 365 \cdot X)$$

$$X_5 = \cos(3 \cdot 2 \cdot \pi / 365 \cdot X); \quad X_6 = \sin(3 \cdot 2 \cdot \pi / 365 \cdot X)$$

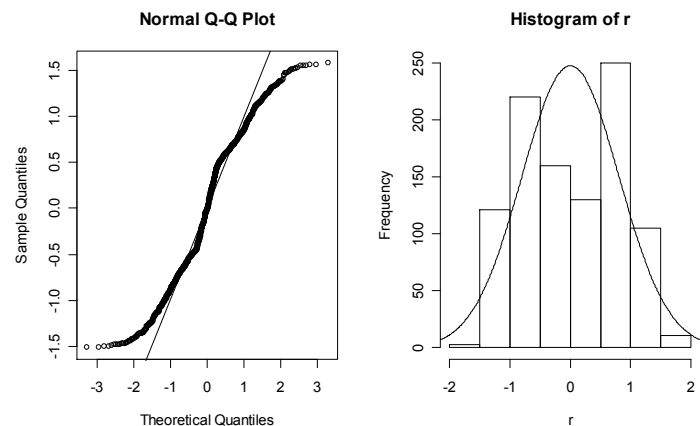
.....

Normaaljaotuse eelduse kontrollimisest

Regressioonanalüüsis rääkisime, et paljud arvutustulemused, nagu olulisustõenäosused (p-value) ja tolerantsi (või prognoosiintervallid) on õiged vaid siis, kui funktsioontunnuse jaotus on normaaljaotus. Vaatame näidet: soovime prognoosida tunnuse Y väärtuseid tunnuse X väärtuste abil lineaarse regressioonimudeli abil. Tunnuste X ja Y tõenäosuspaberid ja histogrammid on toodud all. Kas mudeli eeldused on rahuldatud?

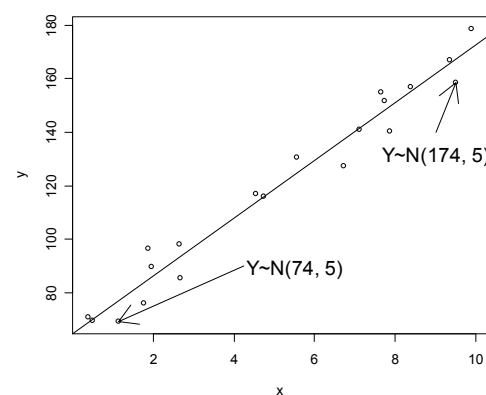


Ülaltoodud näites polnud normaaljaotuse nõue rahuldatud ja testitulemused osutuvad vägagi eksitavateks! Selle mõistmiseks vaatame mudeli jääkide graafikut – pole ju tegemist normaaljaotusega:



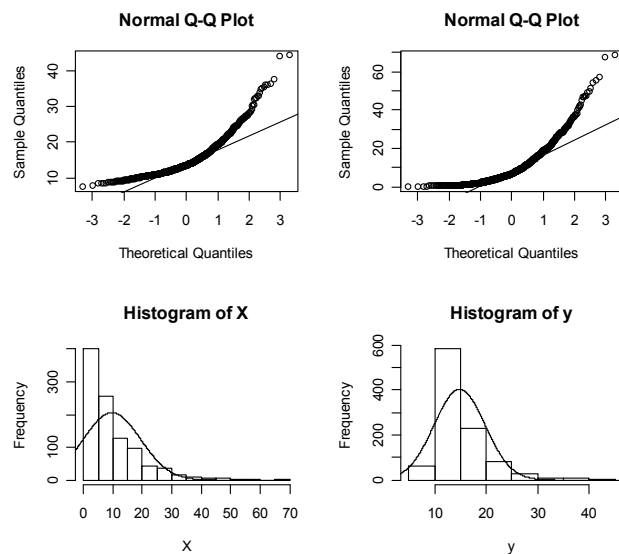
Milles asi?

Iga Y väärtus peab olema normaaljaotusega juhuslik suurus – aga need normaaljaotused on iga isendi/objekti puhul erinevad!

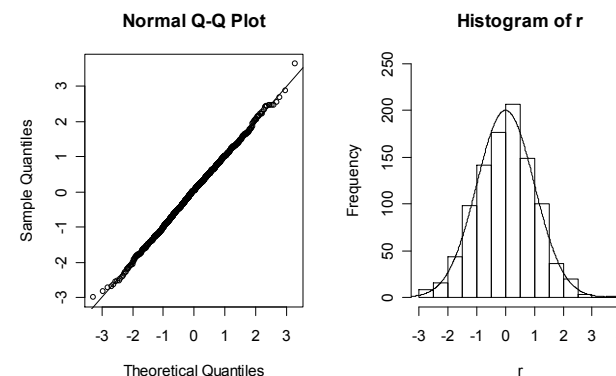


Ühe vaatluse puhul kontrollima, kas ta võiks olla pärit normaaljaotusest $N(74, 5)$ ja teise vaatluse puhul uurima, kas ta võiks olla pärit normaaljaotusest $N(174, 5)$! Pannes kõik Y -tunnuse väärtused ühte patta ja joonistades ühise histogrammi ei suuda me kuidagi kontrollida, kas üks vaatlus oli ühest ja teine teisest normaaljaotusest! Lahendus: Lahutame vaatlusest oodatava keskvväärtuse (saame mudeli jäägi) ja vaatame, kas sel viisil võrreldavaks muudetud väärtuste jaotuseks on normaaljaotus!

Näide 2 – funktsioontunnuse (Y) normaaljaotuse eeldus rahuldatud!



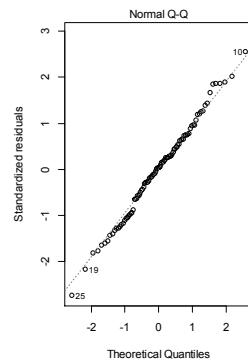
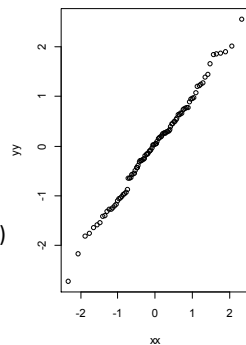
Kust näen nõude rahuldatust? Uurides mudeli jääke!



Normaaljaotuse eelduse kontrollimisest – kvantiil-kvantiil graafik

```
n=length(y)
xx=qnorm( (1:n)/(n+1) )
yy=sort(rstandard(m))
```

```
par(mfrow=c(1,2))
plot(xx,yy)
plot(m, 2)
```



Normaaljaotuse eeldusest

Asymptotic Validity of F Tests for the Ordinary Linear Model and the Multiple Correlation Model

Author(s): Steven F. Arnold

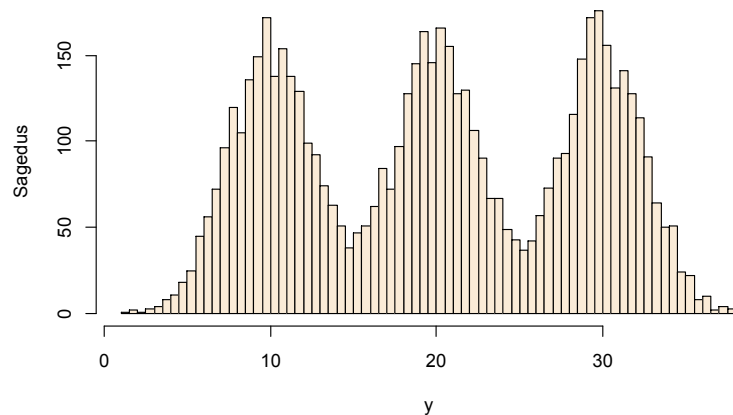
Source: Journal of the American Statistical Association, Vol. 75, No. 372 (Dec., 1980), pp.890-894

Artikli sisu:

Kui Hubberi tingimus on rahuldatud (maksimaalne mõjusus koondub 0-ks) ja kui $\text{rank}(X)/n$ koondub 0-ks, siis on F -test ja t -test asümptootiliselt korrektsed

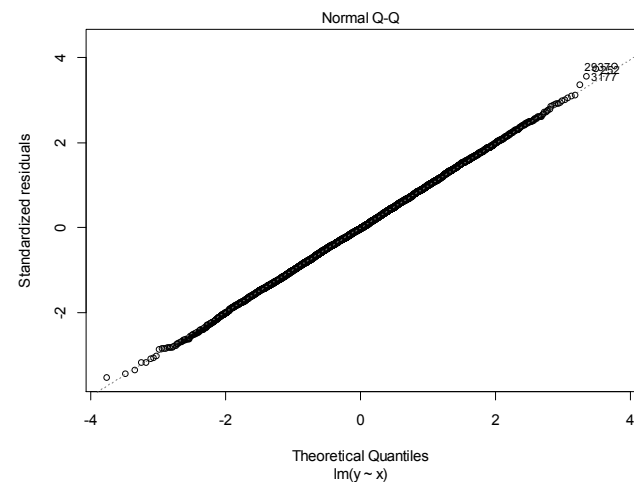
Kas normaaljaotuse nõue on rahuldatud?

Y-tunnuse histogramm



Kvantiil-kvantiil graafik (qq-plot)

Kui normaaljaotus, siis asuvad punktid enam-vähem sirgel...



Kvantiil-kvantiil graafik (qq-plot)

Mis on tegelikult graafikul kujutatud (peamine idee)?

Vaatlused (järjestatud):

1 2 3 4 5 6 7 8 9

Kvantiil-kvantiil graafik (qq-plot)

Mis on tegelikult graafikul kujutatud (peamine idee)?

Vaatlused (järjestatud):

1 2 3 4 5 6 7 8 9
 10% 10% 10% 10% 10% 10% 10% 10% 10%
 Valimi 0,1-kvantiil
 Valimi 0,2-kvantiil

Kvantiil-kvantiil graafik (qq-plot)

Mis on tegelikult graafikul kujutatud (idee)?

Vaatlused (järjestatud):

1 2 3 4 5 6 7 8 9
 10% 10% 10% 10% 10% 10% 10% 10% 10%

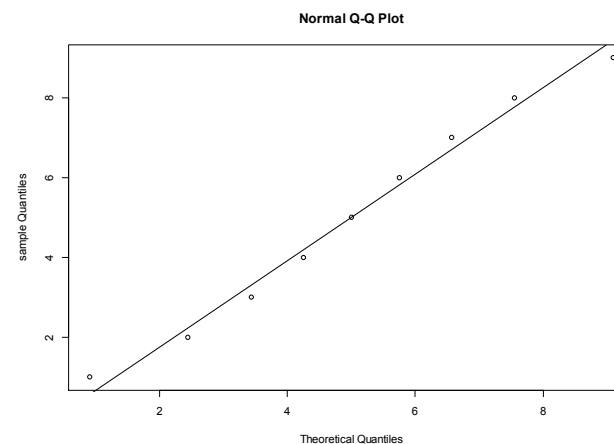
Kui vaatlused oleksid normaaljaotusega (sama hajuvuse ja keskväärtusega kui meie vaatlused), siis kui suur – ootuspäraselt - oleks pidanud olema 9-st vaatlusest väikseim?

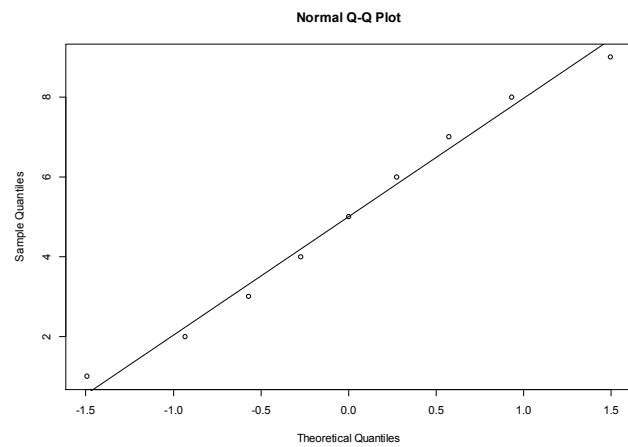
$$\begin{aligned} \text{Vastus: } 0.91 &\approx q(1/(n+1)) \\ &\approx q((1+3/8)/(n+1-2*3/8)) \\ &\approx q((1+1/2)/(n+1-2*1/2)) \end{aligned}$$

Millised oleksid pidanud olema – ootuspäraselt – 0,1-kvantiil, 0,2-kvantiil jne?

0.91 2.45 3.43 4.25 5.00 5.75 6.57 7.55 9.09

Kanname saadud arvud joonisele!





Andmetega kõige paremini sobituva normaaljaotuse asemel võime kasutada ka standardset normaaljaotust – graafiku kuju jääb samaks!