TAMPEREEN YLIOPISTO
Tilastollinen mallintaminen

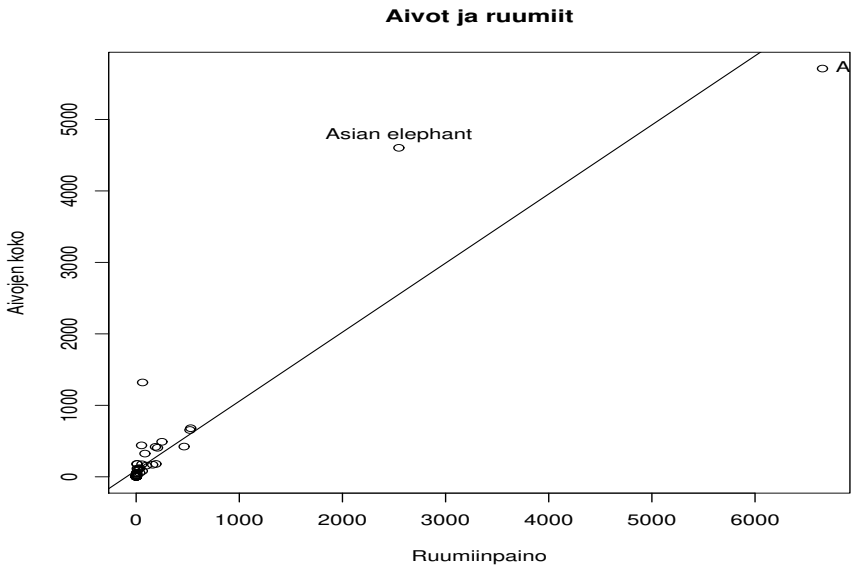
Kari Tokola ja Jouni Kirjola

Aivojen koon riippuvuus ruumiinpainosta

Matematiikan, tilastotieteen ja filosofian laitos
Tilastotiede
14. marraskuuta 2000

1 Johdanto

Aineistona meillä oli nisäkkäiden aivojen koko ja ruumiinpaino muuttujat. Tilastotyksiköitä aineistossa oli reilut kuusikymmentä. Eläinten koko vaihteli hiirestä norsuun. Tehtävänämme oli tutkia eläinten painon riippuvuutta aivojen kokoon.



Kuva 1: Aivojen koon riippuvuus ruumiinpainosta

2 Perusregressiomalli

Aloitimme tarkastelun yksinkertaisesta regressiomallista. Kuten kuvasta 1 on helppo havaita, että y (aivojen koko) kasvaa x:n (ruumiin paino) kasvaessa. Valitulla asteikolla muutamat havainnot erottuvat selvästi muusta aineistosta ja loput muodostavat yhden "möykyn". Nämä poikkeavat havainnot ovat afrikanelefantti ja intiannorsu. Niiden paino on aivan eri luokkaa vertattuna muiden eläinten painoon. Tämän vuoksi niillä on todella suuri vaikutus mallin parametreihin. Ja mallihan oli jokaiselle havainnolle y_i,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

ja virhetermit $\epsilon_i, i = 1, \ldots, n$ ovat toisistaan riippumattomia ja noudattavat normaalijakaumaa $N(0, \sigma^2)$. Ajamalla seuraava R-komento saadaan muodostettua malli (1). Jossa siis objektiin species.lm "sijoitetaan"malli.

```
species.lm <- lm(y ~ x)
```

Tulokseksi saadaan:

```
Residuals:
    Min      1Q  Median      3Q     Max
-810.07  -88.52  -79.64  -13.02 2050.33

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.00440   43.55258    2.09   0.0409 *
x            0.96650    0.04766   20.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 334.7 on 60 degrees of freedom
Multiple R-Squared: 0.8727,     Adjusted R-squared: 0.8705
F-statistic: 411.2 on 1 and 60 degrees of freedom,      p-value:     0
```

Aineestoon sovitettu suora on siis,

$$y_i = 91,0 + 0,97 * x_i, \tag{2}$$

Regressiokertoimen (β_1) estimaatin $\hat{\beta}_1$ t-arvo on erittäin merkitsevä (p-arvo lähellä nollaa). Suora näyttäisi sopivan aineistoon. Tämä ei kuitenkaan vaakuuttanut vielä meitä. Suoritimme graafisia diagnostisia tarkistuksia komenolla

```
plot(species.lm,ask=T)
```

Komento tulostaa seuraavat kuviot: *residuals vs. fitted*[1] , *Q-Q-diagramma*[2], *Scale-location plot*[3] ja *Cook's distance*[4] . Näistä enemmän seuraavassa alakappaleessa.

2.1 Mallin (1) diagnostiset tarkastelut

Mallin oletusten mukaan virhetermit $\epsilon_i, i = 1, \ldots, n$ noudattavat normaalijakaumaa. Koska virhetermejä ei voida havaita, niiden käyttäytymistä tutkimme residuaalien $y - \hat{y}$ (havaittu arvo - sovite) avulla. $Q - Q$-diagramma antoikin viitteitä, etteivät residuaalit noudattaisi normaalijakaumaa. Lisäksi ulkoisesti Studentoidut residuaalit, jotka saadaa kaavalla,

$$t_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}, \qquad h_{ii} = \frac{1}{n} + \frac{(x_i + \bar{x})^2}{SS_x} \tag{3}$$

ovat testisuureita havaintojen vierauden (Outlier) tilastolliseksi testaamiseksi. Sillä ne noudattavat t-jakaumaa vapausastein 59.

[1] Kuva residuaaleista ja sovitetuista arvoista
[2] Residuaalien normaalisuus
[3] Studentoitujen residuaalien itseisarvojen neliöjuuri ja sovite
[4] Cookin etäisyys

Laskemalla Studentoidut residuaalit käskyllä

```
rstudent(species.lm)
```

saadaan jokaiselle havainnolle p-arvo (t-jakaumasta) hypoteesille

$$H_0 : "Havainto\ ei\ ole\ vieras"$$

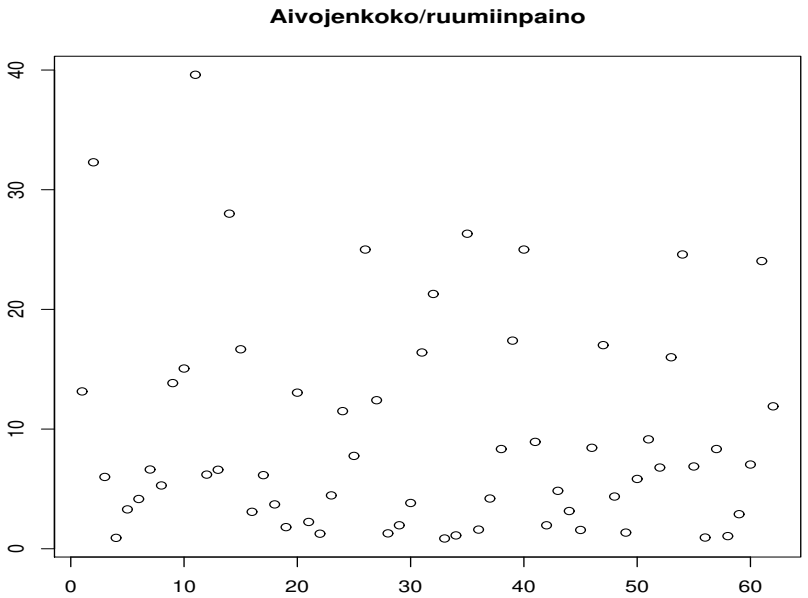
Nämä arvot laskemalla saadaan, että havainnot 19,32 ja 33 (intiannorsu, ihminen ja afrikanelefantti) ovat vieraita havaintoja. Lisäksi samaan tulokseen pääsimme käyttämällä komentoa,

```
influence.measures(species.lm)
```

joka paljasti saman tuloksen. Kyseinen komento laskee muunmuassa Cookin etäisyydet, joihin jo aikaisemmin viittasimme.

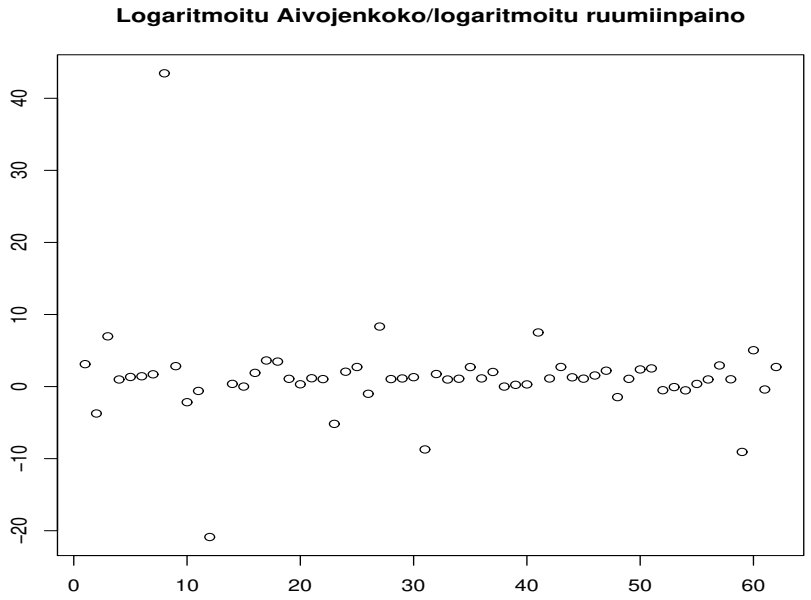
3 Logaritmoitu malli

Alkuperäisessä mallissa tarkastelu kohdistuu lähinnä muuttujien suhteeseen. Tällöin havaintojen suhteet eivät ole keskenään verrattavissa. Kuten kuvasta 2 huomaamme, muuttujien suhde ei ole vakio. Logaritmoimalla muuttujat,



Kuva 2: Alkuperäisten muuttujien suhteet

saamme muuttujien suhteen lähes vakioksi. Kuten kuvasta 3 voimme huomata. Suhteen vakioinnilla päästään eroon alussa mainitsemastamme "möykystä. Jonka edut selviävät myöhemmin.



Kuva 3: Logaritmoitujen muuttujien suhteet

Logaritmointi muuttaa havaintojen suhteet lineaarisemmiksi seuraavalla tavalla.

$$\frac{y_i}{x_i} \qquad \frac{ln(y_i)}{ln(x_i)} = ln(y_i) - ln(x_i) \tag{4}$$

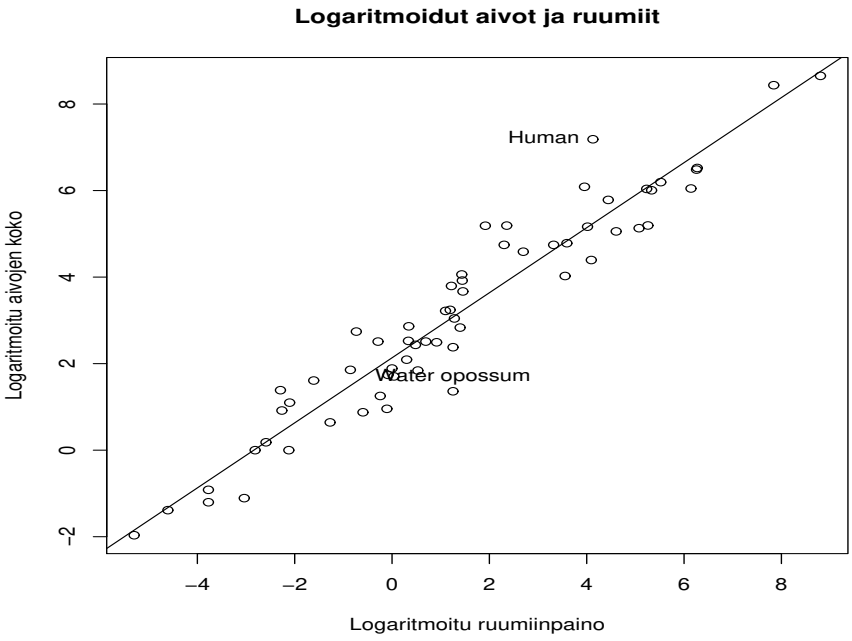
Saatu logaritmoitu teoreettinen malli on siis seuraavanlainen,

$$ln(y_i) = \beta_0 + \beta_1 ln(x_i) + \epsilon_i \tag{5}$$

jossa on siis samat oletukset kuin mallissa (1). Tässä kohtaa on siis hyvä huomata, ettei logaritmin kannalla ole väliä. Tämän varmistaminen jätetään lukijan tehtäväksi. Nyt jääkin siis muodostettavaksi malli R:llä. Annamme mallille (objektille) nimeksi species.llm, kuten seuraavasta komennosta huomaamme.

```
species.llm <- lm(ly ~ lx)
```

Kuvasta 4 huomaamme havainoillisesti kuinka "möykky"häviää. Vertaa kuvaa 1.



Kuva 4: Logaritmoidut muuttujat

Huom. muutttujille on tehty logaritmointi ennen mallin luomista. Ajettaessa yhteenveto mallista species.llm saadaan seuraavanlaiset tulokset.

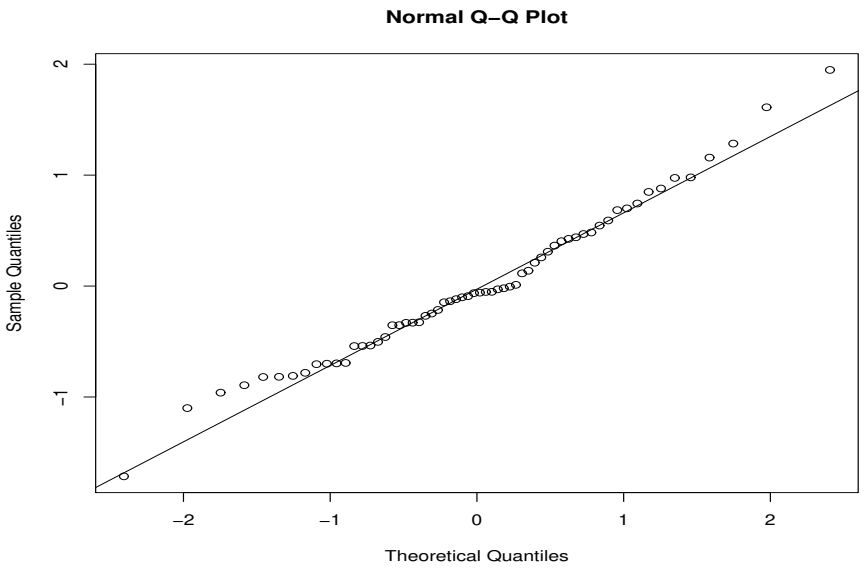
```
Residuals:
     Min       1Q   Median       3Q      Max
-1.71550 -0.49228 -0.06162  0.43597  1.94829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.13479    0.09604   22.23   <2e-16 ***
lx           0.75169    0.02846   26.41   <2e-16 ***
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-Squared: 0.9208,     Adjusted R-squared: 0.9195
F-statistic: 697.4 on 1 and 60 degrees of freedom,      p-value:     0
```

Vertaamalla näitä tuloksia mallin (1) antamiin tuloksiin huomasimme seuraavaa. Ensinnäkin huomattava muutos on tapahtunut estimaattorien hajonnoissa $var(\hat{\beta})$. Mallissa (1) $\hat{\beta}_0$:n hajonta on huomattavasti suurempi kuin mallissa (4). Toinen huomattava asia on kohonnut selitysaste. Syy tähän löytynee tarkemmista tarkasteluista.

3.1 Mallin (4) tarkemmat tarkastelut

Lähdimme tarkastelemaan "uuden"mallin hyvyyttä kuten aikaisemman mallin. Piirsimmekin ensin $Q-Q$-diagramman, joka antoi viitteitä että residuaalit olisivat jakautuneet normaalimmin kuin mallissa (1). Kuten huomaamme kuvasta 5.

Testattaessa Studentin residuaaleja kolme havaintoa tuli merkitseväksi. Tämä siis tarkoittaa, että havaintojen p-arvot olivat ≤ 0.05.



Kuva 5: Q-Q -diagramma mallissa (4)

4 Mallien vertailu

On itsestään selvää, että logaritmoitaessa $\hat{\beta}$-vektorin arvot muuttuvat. Myös estimaattoreiden hajonnat pienenevät huomattavasti logaritmoidessa. Tämä johtuu siitä, että havaintojen arvojen pienetessä huomattavasti pienenee myös hajonta. Logaritmoidussa mallissa jäännösneliösumma (SSE) on paljon

pienempi ja se johtaa seuraaviin tuloksiin. Logaritmoidun mallin estimaattorien p-arvot pienenevät. Myös selitysaste tulee korkeammaksi. Tehtäessä tämä muunnos, alkuperäinen y on seuraavaa muotoa:

$$y_i = \beta_0 + \beta_1 e^{x_i} \tag{6}$$

5 Yhteenveto

Lukijalle saattaa tulla käsitys, että logaritmointiin päättyminen on itsestäänselvyys. Näin ei kuitenkaan ole. Joskus saattaa muuttujien tulkinnan kannalta olla hankalaa logaritmoida ne. Toisin sanoen niiden merkitys saattaa muuttua. Meille autuaana tipahtanut logaritmointi, johti meidät parempaan malliin. Siis meidän mielestä logaritmoitu malli on todellakin parempi kuin alkuperäinen, verrattaessa esim. F-testien p-arvoja, estimaattorien hajontoja ja selitysateita.

5.1 Vieraista havainnoista (outlier)

Vieraiden havaintojen havaitsemiseen voidaan käyttää edellä mainitsemiamme testejä, esim. Studentoidut residuualit ja Cookin etäisyydet. Alkuperäisessä aineistossa havaittiin kolme vierasta havaintoja, kuten olemme kappalessa 2.1. todenneet. Syy miksi emme niitä poistaneet on se, että ne vetävät regressiosuoraa puoleensa. Logaritmoidussa mallissa vieraita havaintoja olivat ihminen ja *water opussum* [5] ja *rhesus monkey*. Näillä havoinnoilla oli itseisarvoiltaan suurimmat Studentoidut residuaalit. Haluttaessa voitaisiin näitä arvoja käyttää ns. *viisausmittarina*. Negatiivinen arvo vastaisi normaalia tyhmempää ja päinvastoin. Ihmisen ja *rhesus monkey'n* arvot olivat positiivisia. Sensijaan *water opussumin* arvo oli negatiivinen. Edellisen perusteella jokainen vetäköön johtopäätöksensä. Riippuen tutkimuksen tarkoituksesta jatkettaisiin vieraiden havaintojen poistamisella tai niiden tarkemmalla tutkimisella.

[5] Emme tienneet suomenkielisiä nimiä :)

6 R-koodi

```
library(mass)
data(mammals)
attach(mammals)
species <- row.names(mammals)
x <- body; y <- brain
plot(x,y,main="Aivot ja ruumiit",xlab="Ruumiinpaino",
ylab="Aivojen koko")
species.lm <- lm(y ~ x)
abline(species.lm)
identify(x,y,species)
summary(species.lm)
plot(species.lm,ask=T)
rstudent(species.lm)
influence.measures(species.lm)
yx <- y/x
ly <- log(y)
lx <- log(x)
plot(yx,main="Aivojenkoko/ruumiinpaino",xlab="",ylab="")
plot(lyx,main="Logaritmoitu Aivojenkoko/logaritmoitu
ruumiinpaino",xlab="",ylab="")
species.llm <- lm(ly ~ lx)
plot(log(x),log(y),main="Logaritmoidut aivot ja ruumiit",
xlab="Logaritmoitu ruumiinpaino",ylab="Logaritmoitu aivojen koko")
abline(species.llm)
identify(log(x),log(y),species)
summary(species.llm)
plot(species.llm,ask=T)
rstudent(species.llm)
influence.measures(species.llm)
```