Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede

806112P DATA-ANALYYSIN PERUSMENETELMÄT (Jari Päkkilä & Esa Läärä)

Välikoe 2, ma 9.1.2012 klo 14-18

Mukana saa olla yksi molemmin puolin täytetty A4-kokoinen lunttilappu sekä tavanomainen funktiolaskin tai graafinen laskin. Tarvittavat taulukot ovat liitteinä.

Vastaa kaikkiin neljään kysymykseen.

1. Liite 1 sisältää tuloksia analyyseistä, joissa tutkimuskohteena oli älykkyysosamäärän (muuttuja `IQ`) riippuvuus sukupuolesta (`suku`, arvot 0 = mies, 1 = nainen), pituudesta (`pituus`, cm) ja aivojen koosta (`MRI`). Tutkimusjoukon muodosti 20 mies- ja 20 naispuolista psykologian opiskelijaa eräässä amerikkalaisessa yliopistossa.

Aivojen kokoa mitattiin magneettikuvaustekniikkaa (*magnetic resonance imaging* – siitä muuttujan nimi) käyttäen. `MRI`-tulos ilmaistaan tässä pikselimääränä, jonka verran aivokudos vie kuvassa tilaa, käyttäen mittayksikkönä 100000 pikseliä. Tutustu liitteen tulostukseen ja kuvioihin huolellisesti ja vastaa seuraaviin kysymyksiin perustellen päätelmäsi.

(a) Muuttujaan `MRI.c` liittyvän regressiokertoimen estimaatti on erilainen mallissa 1 (`malli1`) kuin mallissa 2 (`malli2`). Miten selität tämän eroavuuden? Millainen tulkinta ao. estimaattiin liittyy mallissa 2?

(b) Laske mallissa 2 muuttujaan `MRI.c` liittyvän regressiokertoimen 95% luottamusväli ja tulkitse tulos.

(c) Selitä lyhyesti mitä tarkoitetaan muuttujien välisellä (multi)kollineaarisuudella. Mitä ilmeisiä seurauksia kollineaarisuudesta on eri selittäjien vaikutusten arvioinnin kannalta?

(d) Kuinka merkittäväksi ongelmaksi arvioisit (multi)kollineaarisuuden tässä aineistossa, kun tarkastellaan erityisesti mallia 2? Kuinka suuria ovat kollineaarisuutta kuvaavan tunnusluvun arvot ko. mallissa? *(6 p)*

2. Jatketaan tehtävässä 1 esitellyn aineiston analysointia, joskin tämä tehtävä voidaan ratkaista edellisestä riippumatta. Vastaa alla oleviin kysymyksiin käyttämällä hyväksi liitteen 2 (ja mahdollisesti liitteen 1) R-ohjelman tulostusta.

(a) Millaisella tilastollisella mallilla aineistoa on analysoitu mallin 3 (`malli3`) avulla? Entä minkä niminen käytetty analyysimenetelmä on? Tulkitse analyysin keskeiset tulokset lyhyesti perustellen.

(b) Piirrä mallin 3 tulosten pohjalta sovitetut regressiosuorat naisille ja miehille kuvioon 2. Miten piirrosta voi käyttää mallin 3 tulosten havainnollistamiseen? **PALAUTA TÄYTETTY KUVIO 2 VASTAUSPAPERISI MUKANA!**

(c) Muuttuvatko (ja jos muuttuvat niin miten?) (a)-kohdassa tekemäsi johtopäätökset, kun tutkimusongelmaan liittyvät päätelmät perustuvat mallin 4 (`malli4`) tuloksiin? *(6 p)*

3. Tutkittaessa simuloidulla ajokokeella autonrenkaan kulutuskestävyyttä rengasta "ajetaan koepenkissä" vakio-olosuhteissa, kunnes renkaan kulutuspinta laskee alle sallitun minimimäärän. Tässä vaiheessa renkaalla ajettujen ajokilometrien määrä merkitään muistiin. Kyseisellä menetelmällä kerätty koeaineisto sisältää mittaustiedot kolmesta satunnaisesti valitusta merkin A renkaasta sekä viidestä satunnaisesti valitusta merkin B renkaasta. Koeaineisto (ajokilometrit on esitetty tuhansina kilometreinä) on seuraava:

Rengasmerkki	Ajetut kilometrit				
A:	47	44	39		
B:	38	37	41	34	36

Tehtävänä on vertailla rengasmerkkien A ja B kulutuskestävyyttä määräämällä kiinnostavan vertailuparametrin Δ

(a) piste-estimaatti,

(b) p-arvo kaksisuuntaiseen merkitsevyystestiin, jossa testataan nollahypoteesia $H_0 : \Delta = 0$,

(c) sekä likimääräinen 95 % luottamusväli,

kun estimointikriteerinä on järjestysluvuin painotettu itseisarvopoikkeamien summa. Tulkitse tulokset lyhyesti. Laskelmissa saattaa auttaa liitteen 3 tiedot.

(6 p)

4. Eräässä tutkimuksessa seurattiin kahtatoista lasta syntymästä alkaen. Vastemuuttujana oli ensimmäiseen hampaan paikkaamiseen (karioitumiseen) kulunut aika. Osalla koehenkilöistä vaste jäi havaitsematta, ja heidän kohdallaan tiedetään vain, kuinka kauan heillä ainakin oli täysin paikkaamattomat hampaat ennen kuin seuruu lopetettiin ja tulokset analysoitiin. Näiden henkilöiden seuruuaikoja merkitään seuraavassa $^+$:lla. Seuruuajat (vuosina) olivat seuraavat:

$$6,\ 8,\ 9^+,\ 9^+,\ 10^+,\ 11,\ 14^+,\ 15,\ 17^+,\ 18,\ 20^+,\ 22.$$

(a) Tee tarvittavat laskelmat ja piirrä elinaikojen jakauman kuvaamisessa käytettävä Kaplan–Meier -menetelmällä estimoitu vastemuuttujaan liittyvä välttökäyrä.

(b) Tarkastellaan vasteen jakauman alakvartiilia, mediaania ja yläkvartiilia. Mitkä ko. tunnusluvuista voidaan estimoida, kun aineiston analysoinnissa käytetään Kaplan–Meier -menetelmää? Arvioi ko. parametrien arvot.

(c) Estimoi vasteen jakauman odotusarvo, kun aineiston analysoinnissa, käytetään

(c1) Kaplan–Meier -menetelmää, (c2) eksponenttimallia.

(d) Mikä on eksponenttimallin antama estimaatti todennäköisyydelle, että lapsen hampaat säilyvät täysin paikkaamattomina korkeintaan 12 vuotta?

(6 p)

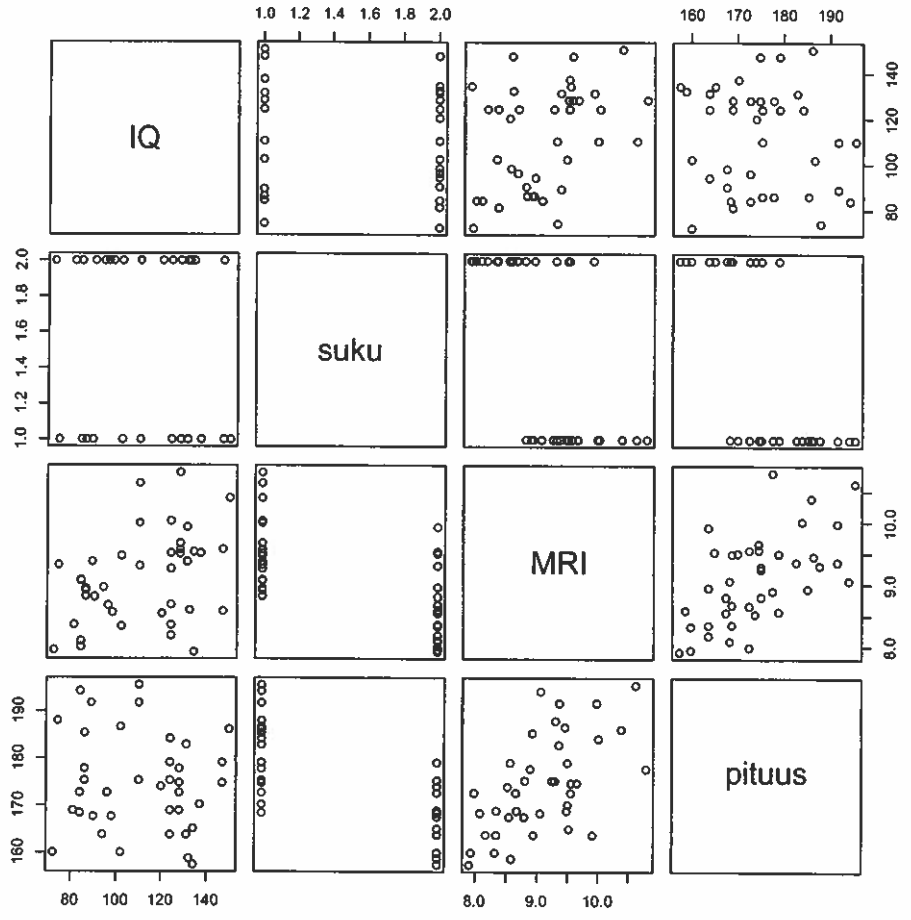
LIITE 1

```
> dats
    IQ suku    MRI pituus
1  124    1  8.169  163.8
2  124    0 10.011  184.2
3  150    0 10.384  186.2
4  128    0  9.654  174.8
5  134    1  9.515  165.1
6  110    1  9.288  175.3
7  131    1  9.913  163.8
8   98    1  8.543  167.6
9   84    0  9.049  168.4
10 147    0  9.555  174.8
11 124    1  8.339  163.8
12 128    0 10.795  177.8
13 124    0  9.241  175.3
14 147    1  8.565  179.1
15  90    1  8.789  167.6
16  96    1  8.654  172.7
17 120    1  8.522  174.0
18 102    0  9.451  186.7
19  84    1  8.080  168.4
20  86    0  8.891  177.8
21  86    0  8.924  185.4
22  84    0  9.059  194.3
23 134    1  7.906  157.5
24 128    0  9.550  172.7
25 102    1  8.318  160.0
26 131    0  9.355  182.9
27  84    1  7.986  172.7
28 110    0 10.625  195.6
29  72    1  7.935  160.0
30 124    1  8.667  168.9
31 132    1  8.578  158.8
32 137    0  9.496  170.2
33 110    0  9.979  191.8
34  86    0  8.800  175.3
35  81    1  8.343  168.9
36 128    1  9.481  168.9
37 124    0  9.494  179.1
38  94    1  8.940  163.8
39  74    0  9.300  188.0
40  89    0  9.359  191.8
> plot(dats)
> mean(dats)
     IQ    suku     MRI  pituus
111.025   0.500   9.088 174.345
> sd(dats)
     IQ    suku     MRI  pituus
22.4710  0.5064  0.7228 10.1867
```

```
> cor(dats)
             IQ     suku     MRI  pituus
IQ       1.00000 -0.02591  0.3869 -0.1052
suku    -0.02591  1.00000 -0.6459 -0.7267
MRI      0.38686 -0.64594  1.0000  0.5857
pituus -0.10517 -0.72674  0.5857  1.0000
>
> tapply(MRI,suku,mean)
    0     1
9.549 8.627
> tapply(IQ,suku,mean)
    0     1
111.6 110.5
> tapply(pituus,suku,mean)
    0     1
181.7 167.0
>
> MRI.c <- MRI - mean(MRI)
> pit.c <- pituus - mean(pituus)
> malli1 <- lm(IQ ~ MRI.c)
> summary(malli1)

Call:
lm(formula = IQ ~ MRI.c)

Residuals:
   Min     1Q Median     3Q    Max
-39.58 -17.95  -1.60  17.03  42.26

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   111.02       3.32   33.45   <2e-16
MRI.c          12.03       4.65    2.59    0.014
---
Residual standard error: 21 on 38 degrees of freedom
Multiple R-squared: 0.15,       Adjusted R-squared: 0.127
F-statistic: 6.69 on 1 and 38 DF,  p-value: 0.0137

> malli2 <-lm(IQ ~ MRI.c + pit.c)
> summary(malli2)

Call:
lm(formula = IQ ~ MRI.c + pit.c)

Residuals:
   Min     1Q Median     3Q    Max
-32.83 -10.55  -3.58  13.14  52.36
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   111.03       3.01   36.83  < 2e-16
MRI.c          21.22       x.xx    4.07  0.00024
pit.c          -1.11       0.37   -3.01  0.00465
---
Residual standard error: 19.1 on 37 degrees of freedom
Multiple R-squared: 0.317,      Adjusted R-squared: 0.28
F-statistic: 8.59 on 2 and 37 DF,  p-value: 0.00086

>
```

KUVIO 1. Muuttujien välinen sirontakuviomatriisi.

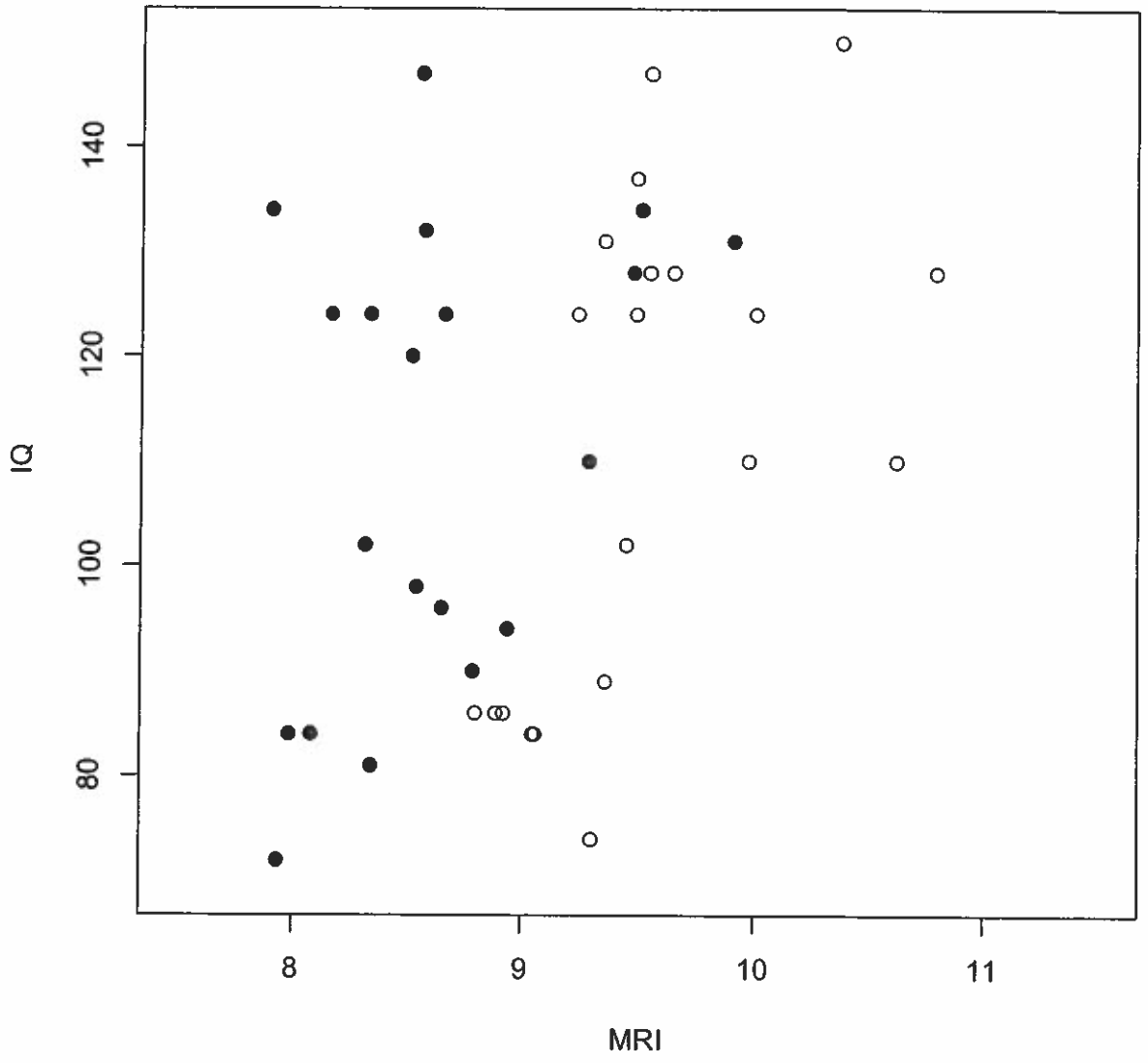


LIITE 2

```
> plot(MRI,IQ,pch=15*suku+1,xlim=c(7.5,11.5),ylim=c(70,150))
> sex <- factor(suku, labels = c("   mies", " nainen"))
> malli3 <-lm(IQ ~ sex + MRI.c)
> summary(malli3) ; confint(malli3)

Call:
lm(formula = IQ ~ sex + MRI.c)

Residuals:
   Min     1Q Median     3Q    Max
-32.69 -15.16  -6.26  16.90  37.78

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   102.50       5.26   19.50   <2e-16
sex nainen     17.06       8.35    2.04   0.0484
MRI.c          19.74       5.85    3.37   0.0018
---
Residual standard error: 20.2 on 37 degrees of freedom
Multiple R-squared: 0.236,      Adjusted R-squared: 0.194
F-statistic: 5.71 on 2 and 37 DF,  p-value: 0.00692

              2.5 % 97.5 %
(Intercept) 91.8490 113.15
sex nainen   0.1269  33.98
MRI.c        7.8852  31.60
>
>
> malli4 <- lm(IQ ~ suku + MRI.c + pit.c)
> summary(malli4); confint(malli4)

Call:
lm(formula = IQ ~ suku + MRI.c + pit.c)

Residuals:
   Min     1Q Median     3Q    Max
-29.75 -12.38  -3.39  14.50  49.67

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  108.348      5.703   19.00  < 2e-16
suku           5.355      9.646    0.56  0.58227
MRI.c         22.480      5.727    3.93  0.00037
pit.c         -0.973      0.452   -2.15  0.03800
---
Residual standard error: 19.2 on 36 degrees of freedom
Multiple R-squared: 0.323,      Adjusted R-squared: 0.267
F-statistic: 5.73 on 3 and 36 DF,  p-value: 0.00261
```

```
              2.5 %    97.5 %
(Intercept)  96.782 119.91344
suku        -14.209  24.91825
MRI.c        10.865  34.09546
pit.c        -1.889  -0.05691
```

KUVIO 2. Älykkyysosamäärän `IQ` riippuvuus aivojen koosta (`MRI`) miehillä (○) ja naisilla (●).



Täytä kuvio annetun tehtävän mukaisesti, kirjoita nimesi ja palauta vastauspaperin mukana!

LIITE 3

```
> rengas_A <- c(47,44,39)
> rengas_B <- c(38,37,41,34,36)
> y <- c(rengas_A,rengas_B)
> n1 <- length(rengas_A); n2 <- length(rengas_B)
> d <- c(matrix(rep(rengas_B,n1),n2)-t(matrix(rep(rengas_A,n2),n1)))
> d <- sort(d) ; d
 [1] -13 -11 -10 -10  -9  -8  -7  -6  -6  -5  -3  -3  -2  -1   2
> n <- length(y)
> y. <- sort(y) ; y.
[1] 34 36 37 38 39 41 44 47
> apu <- matrix(rep(y.,n),n)
> VV <- (apu+t(apu))/2
> V <- c(VV[row(VV) <= col(VV)])
> V <- sort(V) ; V
 [1] 34.0 35.0 35.5 36.0 36.0 36.5 36.5 37.0 37.0 37.5 37.5 37.5 38.0 38.0 38.5
[16] 38.5 39.0 39.0 39.0 39.5 40.0 40.0 40.5 40.5 41.0 41.0 41.5 41.5 42.0 42.5
[31] 42.5 43.0 44.0 44.0 45.5 47.0
>
```