806109 TILASTOTIETEEN PERUSMENETELMÄT I
Muut kuin taloustieteiden tiedekunnan opiskelijat
Harjoitus 6, syksy 2007

36. a) Kahdeksan johtamiskoulutukseen osallistunutta (A, B, C, D, E, F, G ja H) asetettiin ennen koulutusta paremmuusjärjestykseen aiempien opintojensa, työtehtäviensä ja muun toimintansa perusteella ja järjestys oli seuraava (paras ensin):

D, F, E, A, C, H, B, G

Koulutuksen jälkeen järjestetyssä testissä osallistujat saivat pisteitä (mitä suurempi, sitä parempi) seuraavasti:

A: 15, B: 11, C: 13, D: 20, E: 22, F: 20, G: 16, H:10

Laske ennen ja jälkeen koulutuksen tehtyjen arviointien välistä riippuvuutta kuvaava tunnusluku ja tulkitse se.

b) Eräälle opiskelijajoukolle tehdyssä kyselyssä muuttujina oli mm. sukupuoli ja kiinnostus opiskelijapolitiikkaan (vastausluokkina 1=ei kiinnosta lainkaan, 2=kiinnostaa jonkin verran, 3=kiinnostaa paljon).
Opiskelija NN sai tehtäväkseen tutkia saadusta aineistosta em. muuttujien välistä riippuvuutta sopivan riippuvuusluvun avulla. NN päätti suorittaa tehtävän R:llä ja sai aikaiseksi seuraavan tulostuksen:

```
> .Table  # Counts

       ei lainkaan jonkin verran paljon
nainen          72            56     32
mies            82            37     21

> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 5.5052, df = 2, p-value = 0.06376


> .Test$expected # Expected Counts

       ei lainkaan jonkin verran   paljon
nainen    82.13333          49.6 28.26667
mies      71.86667          43.4 24.73333
```

Eteenpäin NN ei osannut jatkaa.
Suorita tehtävä loppuun ja tulkitse tulos.

b2) Myöhemmin havaittiin, että NN oli ollut huolimaton ja unohtanut ottaa mukaan aineistoonsa kyselyyn vastanneista 60 (kaikki miehiä). Miten näiden 60 miehen vastausten on täytynyt jakautua, kun koko aineistossa sukupuoli ja kiinnostus opiskelijapolitiikkaan olivat täysin riippumattomia toisistaan.

37. Aineistossa on havaintoja sadasta tilastoyksiköstä. Kovarianssimatriisi on

$$\begin{array}{c} \\ x \\ y \\ z \end{array} \begin{array}{c} \begin{array}{ccc} x & y & z \end{array} \\ \begin{pmatrix} 25 & 16 & 9 \\ 16 & 16 & 36 \\ 9 & 36 & 25 \end{pmatrix} \end{array}$$

ja lisäksi tiedetään, että $\sum_{i=1}^{100} x_i = 1500$, $\sum_{i=1}^{100} y_i = 1200$ ja $\sum_{i=1}^{100} z_i = 1400$.

a) Sovita aineistoon regressiosuora $y = a + bx$.

b) Laske muuttujien x ja z välinen korrelaatiokerroin r_{xz}.

38. Eräässä kunnassa selvitettiin 12 kunnallisvaaliehdokkaan ikä (ika, vuosina), vaalimainontaan sijoittama rahamäärä (mainoskust, 1000 euroina) ja vaaleissa saatu äänimäärä (aanet). Tästä aineistosta R:llä laskettuja tuloksia on esitetty liitteessä 1.

a) Laske korrelaatiomatriisin peitetty arvo.

b) Tulkitse ehdokkaan

b1) iän ja mainontaan sijoittaman rahamäärän,

b2) mainontaan sijoittaman rahamäärän ja äänimäärän välistä riippuvuutta liitteestä löytyvien kuvioiden ja riippuvuuslukujen avulla.

c) Tulkitse liitteessä 1 esitetyn regressioanalyysin tulokset (regressioyhtälö ja sen kertoimien tulkinnat, determinaatiokerroin eli selitysaste ja sen tulkinta).

39. (jatkoa tehtävään 35)

a) Sovita aineistoon regressiosuora $y = a + bx$, missä y=aamulla käytettävän pitkävaikutteisen insuliinin määrä ja x=diabeteksen kesto. Tulkitse kertoimet. Määrää myös regressioyhtälön determinaatiokerroin eli selitysaste ja tulkitse se.

b) Lisää a)-kohdan regressiosuora tehtävän 35 korrelaatiodiagrammiin. Paljonko regresioyhtälö ennustaa insuliinin määrän olevan, jos diabeteksen kesto on 10.0 vuotta?

c) Määrää havaintoaineiston viimeisen tilastoyksikön (jolla muuttujien havaitut arvot ovat x=1.3 ja y=4) ennustearvo $\hat{y}_i$ ja residuaali $e_i = y_i - \hat{y}_i$.

d) Liitteessä 2 on esitetty R-ohjelman tulostus a)-kohtaan liittyen. Vertaa R:n tulostuksen antamaa informaatiota a)-kohdassa saamiisi tuloksiin.

40. Pienen ala-asteen kakkosluokan liikuntaryhmän pojat (n=11) ottelivat neliottelun, jossa lajeina olivat 60 metrin juoksu, 100 metrin juoksu, pallonheitto ja pituushyppy.

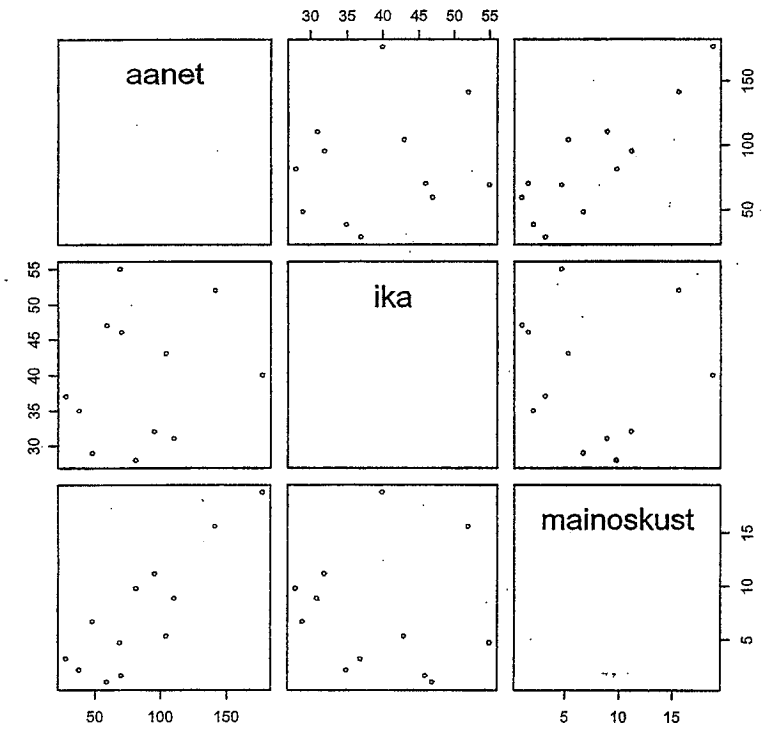
a) Pallonheitossa lopputulokset (metreinä) olivat
18.7 23.4 15.9 19.6 21.0 19.7 18.5 19.7 21.8 18.4 20.1
Laske pallonheiton lopputulosten aritmeettinen keskiarvo ja keskihajonta.

b) Liitteessä 3 on esitetty tarkasteltavasta aineistosta saatu R:n tulostus. Kommentoi 60 metrin juoksun ja 100 metrin juoksun välistä riippuvuutta korrelaatiokertoimen r perusteella. Missä korrelaatiodiagrammin kuvassa/kuvissa (1-12) kuvataan 60 metrin juoksun ja 100 metrin juoksun välistä riippuvuutta?

c) Määrää regressioyhtälö $y = a + bx$ ja tulkitse kertoimet, kun vastemuuttujana on 100 metrin juoksu ja selittävänä muuttujana 60 metrin juoksu. Määrää lisäksi regressioyhtälön determinaatiokerroin (eli selitysaste) ja tulkitse se.

Matin tulokset neliottelussa olivat: 60 metrin juoksu = 12.5 sekuntia, 100 metrin juoksu = 23.2 sekuntia, pallonheitto = 19.6 metriä ja pituushyppy 2.20 metriä. Ennusta Matin vastemuuttujan arvo regressioyhtälön avulla.

HUOM! 1. välikoe pe 19.10. klo 8.00-12.00 salissa L5.

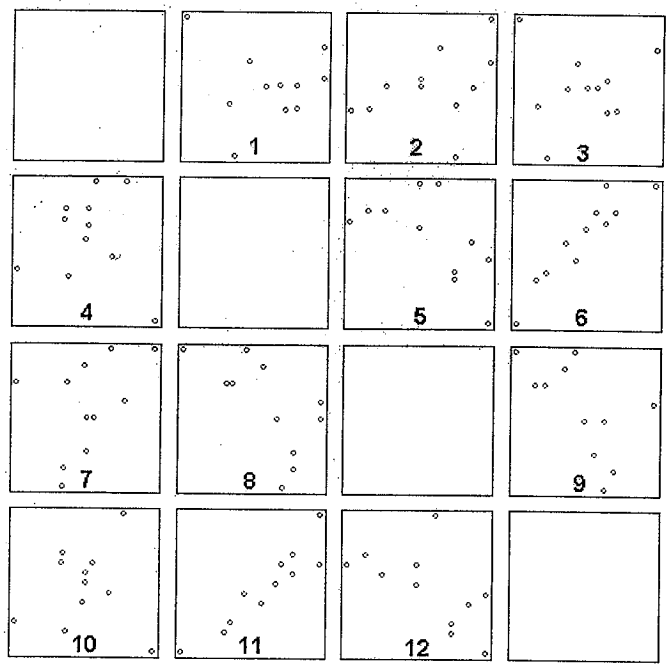



```
#Keskiarvot

> mean(ika)
[1] 39.58333

> mean(mainoskust)
[1] 7.408333

> mean(aanet)
[1] 85


#Korrelaatiomatriisi

                aanet       ika mainoskust
aanet       1.0000000           0.8688182
ika                     1.0000000 -0.0954054
mainoskust  0.8688182 -0.0954054  1.0000000


#Kovarianssimatriisi

                aanet       ika mainoskust
aanet       1880.54545 70.272727 213.181818
ika           70.27273 82.265152  -4.896212
mainoskust   213.18182 -4.896212  32.015379



> RegModel.1 <- lm(aanet~mainoskust, data=vaalit)

> summary(RegModel.1)

Call:
lm(formula = aanet ~ mainoskust, data = vaalit)

Residuals:
    Min      1Q  Median      3Q     Max
-32.283 -16.417   1.744  16.277  33.039

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   35.670     11.013   3.239 0.008888
mainoskust     6.659      1.200   5.549 0.000244
---


Residual standard error: 22.52 on 10 degrees of freedom
Multiple R-Squared: 0.7548,     Adjusted R-squared: 0.7303
F-statistic: 30.79 on 1 and 10 DF,  p-value: 0.0002445
```

LIITE 2

```
> RegModel.1 <- lm(insuliinin.maara~diabeteksen.kesto, data=diabetes)

> summary(RegModel.1)

Call:
lm(formula = insuliinin.maara ~ diabeteksen.kesto, data = diabetes)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0168 -2.8374 -0.5759  2.8743  5.5994

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         7.7599     2.1089   3.680  0.00622 **
diabeteksen.kesto   0.9668     0.3058   3.162  0.01336 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.124 on 8 degrees of freedom
Multiple R-Squared: 0.5554,   Adjusted R-squared: 0.4999
F-statistic: 9.995 on 1 and 8 DF,  p-value: 0.01336
```

LIITE 3



#keskiarvot:

```
> mean(juoksu.100m)
[1] 23.3182

> mean(juoksu.60m)
[1] 12.6636

> mean(pituushyppy)
[1] 2.1318
```

varianssit:

```
> var(juoksu.100m)
[1] 1.5796

> var(juoksu.60m)
[1] 0.1605

> var(pituushyppy)
[1] 0.0186
```

#korrelaatiomatriisi

```
> cor(neliottelu, use="complete.obs")
```

	juoksu.100m	juoksu.60m	pallonheitto	pituushyppy
juoksu.100m	1.0000	0.9328	-0.1067	-0.6302
juoksu.60m	0.9328	1.0000	0.0322	-0.5983
pallonheitto	-0.1067	0.0322	1.0000	0.3907
pituushyppy	-0.6302	-0.5983	0.3907	1.0000