805305A JOHDATUS REGRESSIO- JA VARIANSSIANALYYSIIN, sl 2022
Harjoitus 6, viikko 40: mikroluokkatehtävät

1. Jatkoa harjoituksen 6 kotitehtäviin 2, 3, 4 ja 5. Analysoidaan koulutuksen ja tulojen välistä yhteyttä R:n avulla.

(a) Sijoita yksilökohtaiset kouluvuosien määrät vektoriin `koulu` ja kuukausitulot (markkoina) vektoriin `tulot`.

```
> koulu <- c(6, 12, 10, 8, 9)
> tulot <- c(10, 20, 17, 12, 11)
```

(b) Piirrä kouluvuosien ja tulojen välinen sirontakuvio merkiten havaintopisteet umpipallolla (graafisen parametrin `pch` arvo 16) ja varautuen kummallakin akselilla havaittua hieman leveämpään vaihteluväliin (parametrit `xlim` ja `ylim`):

```
> par(mfrow=c(1,1))
> plot( tulot ~ koulu, pch=16, xlim=c(5,13), ylim=c(8,22) )
```

(c) Laske funktioiden `mean()`, `sd()` ja `cor()` avulla kummankin muuttujan keskiarvo ja keskihajonta sekä niiden välinen korrelaatiokerroin. Saitko samat arvot kuin kotitehtävässä 2(c)? Moodle 1

(d) Sovita kotitehtävässä 2 määritelty regressiomalli vasteena `tulot` ja selittäjänä `koulu` käyttäen funktiota `lm()` (= *linear model*) ja sijoita sovitustulokset malliolioon `m1`:

```
> m1 <- lm( tulot ~ koulu )
```

Tästä eteenpäin sovitustuloksista voidaan tulostaa monia erilaisia tunnuslukuja kutsuen vastaavia funktioita pääargumentilla `m1`.

(e) Tulosta mallisovituksen päätulokset malliobjektista `m1` funktiolla `summary()` ja Anova-taulun rivit funktiolla `anova()` ja tutki, mitä informaatiota nämä antavat.

```
> summary(m1)
> anova(m1)
```

Saitko samat estimaatit ja keskivirheet regressiokertoimelle β_1 kuin kotitehtävässä 2(d)? Moodle 2

Tarkista myös, että komennon `summary(m1)` tulostuksessa suureen "`Residual standard error`" arvo on sama kuin Anova-taulun jäännöskeskineliösumman 4.25 neliöjuuri eli jäännöskeskihajontaa S. Moodle 3

(f) Kertoimien piste-estimaatit ja luottamusvälit saadaan tulostetuksi samanaikaisesti yhdistämällä funktioiden `coef()` ja `confint()` tuottamat sarakevektorit:

```
> round( cbind( coef(m1), confint(m1) ), 2)
```

Vertaa tulostusta kotitehtävässä 3(c) laskettuun luottamusväliin. Moodle 4

2. Jatkoa edelliseen tehtävään. Tutkitaan vielä mallituksesta saatavia sovitteita ja jäännöstermejä.

(a) Sijoita vasteen sovitteet vektoriin `yhat` funktiolla `fitted()` ja jäännöstermit vektoriin `res` funktiolla `resid()`. Listaa tämän jälkeen rinnakkain havainnot, sovitteet ja jäännöstermit:

```
> yhat <- fitted(m1)
> res <- resid(m1)
> data.frame(koulu, tulot, yhat, res)
```

Ovatko samat kuin kotitehtävässä 2? Moodle 5

(b) Lisää jo aiemmin piirrettyyn sirontakuvioon kullekin havaintoyksikölle sovitetut arvot, piirrä sovitettu suora ja havainnollista jäännöstermejä yhdistämällä havaitut ja sovitetut arvot pisteviivoin:

```
> points( yhat ~ koulu)  # lisätään kuvaan sovitteet
> abline(m1)             # mallin m1 mukainen sovitettu regressiosuora
> segments( koulu, tulot, koulu, yhat, lty=3 )  # pystysuorat etäisyydet
```

Moodle 6 Moodle 7

(c) Muodosta potentiaalisten uusien x-muuttujan arvojen hila vaihteluvälille [6, 12] puolikkaan välein:

```
> xnew <- data.frame( koulu = seq(6,12, by=0.5) ) ; xnew
```

(d) Käyttäen funktiota `predict()` luo 3-sarakkeinen matriisi `yfit`, jossa on edellä muodostetuille x-arvoille $x_k \in \{6, 6.5, \ldots, 11.5, 12\}$, lasketut vasteen *sovitteet* $\widehat{y}_k$ eli piste-estimaatit vasteen odotusarvoille $\mu_k = \beta_0 + \beta_1 x_k$ ja niiden 95 % *luottamusvälien* ala- ja ylärajat, ja tulosta tämä matriisi yhdessä x-arvojen kanssa:

```
> yfit <- predict(m1, newdata= xnew, interval="confidence", level=0.95)
> round(cbind(xnew, yfit), 2)
```

Saatko samat luottamusvälit muuttujan `koulu` arvoilla 9 ja 12 kuin kotitehtävässä 3(d)? Moodle 8 Moodle 9

(e) Vasteen odotusarvojen luottamusvälien ala- ja ylärajat ovat siis `yfit`-matriisin 2. ja 3. sarakkeella. Piirrä sirontakuvioon näiden luottamusvälien vyöhyke katkoviivoin:

```
> lines( xnew[, 1], yfit[, 2], lty=2)
> lines( xnew[, 1], yfit[, 3], lty=2)
```

Mitä havaintoja teet? Moodle 10

(f) Luo samantyyppinen matriisi `ypred` edelleen funktiolla `predict()` mutta käyttäen nyt optiota `interval='predict'`, joka laskee kullekin uudelle x-arvolle x_k vasteen *ennusteen* $\widetilde{y}_k$ (joka on siis käytännössä sama kuin sovite) sekä 95 % *ennustevälin*. Listaa tämän matriisin sisältö yhdessä `xnew`-arvojen sekä `yfit`-matriisin kanssa kuten edellä. Piirrä ennustevälien vyöhyke sirontakuvioon kuten sovitteelle edellä mutta käyttäen pisteviivaa (`lty=3`).

```
> ypred <- predict(m1, newdata= xnew, interval="predict", level=0.95)
> round(cbind(xnew, ypred), 2)
> lines( xnew[, 1], ypred[, 2], lty=2)
> lines( xnew[, 1], ypred[, 3], lty=2)
```

Mitä havaintoja teet? Vertaa saatuja ennustevälejä muuttujan `koulu` arvoilla 9 ja 12 (d)-kohdassa laskettuihin vastaaviin vasteen odotusarvojen luottamusväleihin. Moodle 11

3. Joukolta vapaaehtoisia miehiä ($n = 20$) ja naisia ($n = 20$) mitattiin heidän älykkyysosamääränsä (IQ), aivojen koko (MRI) magneettikuvauslaitteen avulla pikseleinä ($\times$100000, 18 MRI-kuvasta) sekä mm. pituus (cm). Halutaan selvittää kummallakin sukupuolella erikseen, missä määrin älykkyysosamäärä on yhteydessä aivojen kokoon. Havaintoaineisto on talletettu Moodleen tiedostonimellä `brain.txt`.

(a) Kopioi aineisto Moodlesta R-istunnon kotihakemistoosi, lue aineisto R:n muistiin ja tutki aineiston rakennetta funktiolla `str()`.

```
> brain <- read.table("brain.txt", header=T)
> str(brain)
```

(b) Sukupuolimuuttuja `suku` on koodattu seuraavasti: 0 = "mies" ja 1 = "nainen". Muodosta tästä luokitteluasteikollinen tekijä `sukup`. Laske sen jälkeen sukupuolittain perustunnuslukuja muuttujista IQ, MRI ja pituus.

```
> brain$sukup <- factor(brain$suku, levels=c(0,1), labels = c("mies", "nainen") )
> attach(brain)
> source("Esanfunktiot.r")   # funktio kopioitu Moodlesta R-harjoituksessa 2
> tunnus.taulu(IQ, sukup, 1)
> tunnus.taulu(MRI, sukup, 2)
> tunnus.taulu(pituus, sukup, 1)
```

Miehillä ja naisilla näyttäisi olevan eroa MRI-muuttujan jakauman sijainnissa, kun taas ero IQ-muuttujan keskiarvojen välillä on melko vaatimaton. Moodle 12

(c) Poimitaan `brain`-aineistosta erilliset datakehikot miehille ja naisille funktiolla `subset()`. Sen avulla valitaan 1. argumenttina annettavasta alkuperäisestä datakehikosta osajoukko, jonka alkiot täyttävät 2. argumentin sisältämän loogisen ehdon.

```
> brain.m <- subset(brain, sukup == "mies") ; brain.m
> brain.n <- subset(brain, sukup == "nainen") ; brain.n
```

(d) Piirretään miehille ja naisille vierekkäiset sirontakuviot

```
> par(mfrow=c(1,2))
> with(brain.m, plot(IQ ~ MRI, pch = 16, main = "Miehet") )
> with(brain.n, plot(IQ ~ MRI, pch = 16, main = "Naiset") )
```

(e) Lasketaan seuraavaksi IQ:n ja MRI:n välinen korrelaatiokerroin erikseen miehille ja naisille

```
> with( brain.m, cor(MRI, IQ) )
> with( brain.n, cor(MRI, IQ) )
```

Moodle 13

Millä edellytyksillä korrelaatiokerroin on mielekäs riippuvuusluku? Mitä arvelet, kuinka hyvin nämä edellytykset on täytetty tässä aineistossa?

4. Regressio aivojen koon ja älykkyysosamäärän välillä.

(a) Sovitetaan **pelkästään naisten aineistoon** lineaarinen regressiomalli $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, jossa vastemuuttujana (Y) on älykkyysosamäärää kuvaava muuttuja IQ ja selittäjänä (X) aivojen kokoa kuvaava muuttuja MRI.

```
> malli.n <- lm( IQ ~ MRI, data = brain.n)
> summary(malli.n)
> round( cbind( coef(malli.n), confint(malli.n) ), 2)
> anova(malli.n)
```

Miten tulkitset estimoituja regressiokertoimia ja mitä päättelet aivojen koon ja älykkyysosamäärän välisestä riippuvuudesta tämän aineiston pohjalta? Moodle 14 Moodle 15 Moodle 16

(b) Piirretään uudelleen naisten sirontakuvio ja lisätään siihen edellä sovitettu regressiosuora funktiolla `abline()`.

```
> par(mfrow=c(1,1))
> with(brain.n, plot( IQ ~ MRI, pch = 16) )
> abline(malli.n)
```

(c) Täydennetään seuraavaksi naisten aineistoa uudella muuttujalla MRI.kesk, jossa alkuperäisen MRI-muuttujan arvot on keskistetty keskiarvoonsa

```
> brain.n$MRI.kesk <- brain.n$MRI-mean(brain.n$MRI)
> brain.n$MRI.kesk
```

Lasketaan muuttujan MRI sekä keskiarvoonsa keskistetyn muuttujan MRI.kesk keskiarvot ja keskihajonnat

```
> mean(brain.n$MRI); mean(brain.n$MRI.kesk)
> sd(brain.n$MRI); sd(brain.n$MRI.kesk)
```

Vertaile muuttujien keskiarvoja ja keskihajontoja. Mitä huomaat? Moodle 17

(d) Sovitetaan vielä naisten aineistoon lineaarinen regressiomalli $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, jossa vastemuuttujana (Y) on älykkyysosamäärää kuvaava muuttuja IQ ja selittäjänä (X) on aivojen kokoa kuvaava (keskiarvoonsa keskistetty) muuttuja MRI.kesk.

```
> malli2.n <- lm( IQ ~ MRI.kesk, data = brain.n)
> summary(malli2.n)
> round( cbind( coef(malli2.n), confint(malli2.n) ), 2)
```

Vertaa saatuja tuloksia kohdassa (a) saatuihin tuloksiin. Mitä huomaat? Moodle 18 Moodle 19

(e) Laske (a)-kohdassa muodostetun mallin avulla vasteen sovitteet ja niiden 95 % luottamusvälit, kun selittäjän MRI arvoina ovat 8, 8.5, 9, 9.5 ja 10. Ota mallia tarvittaviin R-komentoihin tehtävän 2 kohdista (c) ja (d).

(f) Laske (a)-kohdassa muodostetun mallin avulla vasteen ennusteet ja niiden 95 % ennustevälit, kun selittäjän MRI arvoina ovat 8, 8.5, 9, 9.5 ja 10. Ota mallia tarvittaviin R-komentoihin tehtävän 2 kohdasta (e). Moodle 20