805305A JOHDATUS REGRESSIO- JA VARIANSSIANALYYSIIN, sl 2022
Harjoitus 7, viikko 41: mikroluokkatehtävät

1. Aloitetaan R-osuus palkka-aineiston analysoinnilla ja sovitetaan aineistoon vielä kerran yhden selittävän muuttujan lineaarinen regressiomalli, jossa tuloja selitetään koulutuksen pituudella.

(a) Mallin sovitus ja keskeisimpien mallitustulosten poiminta.

```
> koulu <- c(6, 12, 10, 8, 9)
> tulot <- c(10, 20, 17, 12, 11)
> plot(tulot ~ koulu, pch=16, xlim=c(5,13), ylim=c(8, 22))
> m1 <- lm(tulot ~ koulu) ; summary(m1)
> cbind(coef(m1), round( confint(m1), 2) )
> anova(m1)
```
Moodle 1
```
> yhat <- fitted(m1) ; yhat
> res <- resid(m1) ; res
> abline(m1)
```

(b) `Summary()`-funktion tuottamassa tulostuksessa determinaatiokertoimen R^2 arvo löytyy kohdasta `Multiple R-squared`. Sama lopputulos saadaan myös Anova-taulun tulosten perusteella, sillä määritelmän mukaan $R^2 = \frac{\text{SSR}}{\text{SSY}} = \frac{\text{SSR}}{\text{SSR+SSE}}$. Moodle 2

Luentomonisteen sivulla 109 olevan tuloksen perusteella lineaarisen regressiomallin determinaatiokerroin voidaan laskea myös ns. *yhteiskorrelaatiokertoimen* $R_{Y\widehat{Y}}$ avulla. Piirretään seuraavaksi vasteen havaittujen arvojen Y_i ja sovitteiden $\widehat{Y}_i$ välinen sirontakuvio ja lasketaan havaittujen arvojen ja sovitteiden välinen yhteiskorrelaatiokerroin $R_{Y\widehat{Y}}$, jonka neliö mallin determinaatiokerroin on.

```
> par(mfrow=c(2,1))
> plot(tulot ~ yhat) ; plot(tulot ~ koulu)
> cor(tulot,yhat) ; cor(tulot,koulu)
> cor(tulot,yhat)^2 ; cor(tulot,koulu)^2
```

Millainen edellä piirretty sirontakuvio olisi, jos mallin determinaatiokertoimen arvo olisi ollut 1?
Moodle 3

2. Laajennetaan seuraavaksi tulojen mallitusta ottamalla mallin toiseksi selittäjäksi työntekijän ikä

```
> ika <- c( 28, 40, 32, 36, 34)
> m2 <- lm(tulot ~ koulu + ika)
> summary(m2)     # perustulostus malliobjektista
```
Moodle 4
```
> confint(m2)     # kertoimien luottamusvälit
```
Moodle 5
```
> anova(m2)       # ANOVA-taulu
> library(car);  vif(m2)  # ladataan vif-kertoimia varten paketti car
```
Moodle 6

(a) Huomioi, että R:n oletusarvoinen Anova-taulu poikkeaa jonkin verran harjoituksen 7 kotitehtävässä 2 muodostetusta Anova-taulusta, sillä R muodostaa useamman selittäjän lineaarimallin Anova-taulun ns. tyypin I lisäneliösummiin perustuvana tauluna.

Yhden selittäjän mallissa `m1` regressioneliösumma `SSR` saatiin luettua suoraan Anova-taulusta selittäjään `koulu` liittyvältä riviltä ja `SSR` oli suuruudeltaan 61.25. Kahden selittäjän mallissa koulumuuttuja tulee mallinmäärityskomennossa `lm()` ensimmäisenä selittäjänä, jolloin kyseiseen muuttujaan liittyvä lisäneliösumma `SSR(1)` vastaa yhden selittäjän regressioneliösummaa.

Toisena selittäjänä tulevalle ikä-muuttujalle laskettava lisäneliösumma `SSR(2|1)` kuvaa sitä regressioneliösumman lisäystä, joka iän lisäämisellä selittäjäksi saadaan aikaan, kun mallissa on jo entuudestaan selittäjänä koulumuuttuja. Mallissa `m2` ikä-muuttujaan liittyvä lisäneliösumma on 1.25 ja siten kahden selittäjän mallissa "perinteinen" regressioneliösumma `SSR` on $61.25 + 1.25 = 62.5$.

Tee seuraavaksi kahden selittän mallitus uudelleen, mutta anna nyt selittäjät `lm()`-funktiossa järjestyksessä `ika, koulu`

```
> m2b <- lm(tulot ~ ika + koulu)
```

Tulosta muodostetusta malliobjektista `m2b` Anova-taulu ja tutki taulun sisältöä. Moodle 7

(b) Havaintoaineiston kuvaamiseen tarvitsemme nyt 3-ulotteista kuvaa, jonka piirtämiseen voidaan käyttää `car`-paketin `scatter3d()`-funktiota. Suorita alla oleva kuvanpiirtokomento ja suurenna avautuva RGL-ikkuna.

```
> scatter3d(tulot ~ koulu + ika, surface=FALSE, id.method="identify")
```

Voit halutessasi tarkastella 3-ulotteista kuvaa eri kulmista "pyörittämällä" kuvaa hiirellä pitämällä hiiren vasenta nappia pohjassa. Tutki kuvaa kaikessa rauhassa ja etsi kuvasta esim. viidennen havaintoyksikön tuottama piste, jossaa tulojen määrä on 11 tuhatta markkaa, koulusta 9 vuotta ja ikää 34 vuotta. Kun lopetat kuvan tarkastelun, pysäytä R:n interaktiivinen tila Rgui-ikkunan päävalikon alapuolella olevalla stop-painikkeella (näpäytä ensi Rgui-ikkuna aktiiviseksi, jotta saat käyttöösi tarvittavan painikkeen) tai sulje RGL-ikkuna ikkunan oikeasta yläkulmasta.

(c) Piirretään vielä äsken piirretty kuva uudelleen, mutta vaihdetaan `surface`-agumentin arvoksi TRUE ja jätetään yksittäisen havainnon identifiointimääre pois.

```
> scatter3d(tulot ~ koulu + ika, surface=TRUE)
```

Kuvaan saatiin nyt mukaan tehtävän alussa määritellyn regressioanalyysin (kaksi selittäjää) lopputulos 2-ulotteisena tasona. Tarkastele kuvaa jälleen eri kulmista ja kiinnitä huomiosi myös residuaaleihin, joita havainnollistetaan kuvassa apuviivojen avulla.

(d) Verrataan vielä havaintoyksiköiden potentiaalilukuja (vaikuttavuuden mittari) malleissa `m1` ja `m2`. Potentiaalit voidaan poimia muodostetuista malliobjekteista funktiolla `hatvalues()`:

```
> pot1 <- hatvalues(m1) ; pot2 <- hatvalues(m2)
> data.frame(pot1, pot2)
```

Havaintoyksikön potentiaali mittaa havaintoyksikön vaikuttavuutta regressiokertoimien (ja sitä myötä sovitteiden ja ennusteiden) estimoinnissa: mitä suurempi potentiaali havaintoyksiköllä on sitä enemmän kyseinen havaintoyksikkö vaikuttaa saatujen estimaattien arvoihin. Moodle 8

Lasketaan myös potentiaalien summa sekä mallissa `m1` että mallissa `m2`

```
> sum(pot1); sum(pot2)
```

Potentiaalien summa on yhtä suuri kuin mallin systemaattisen osan tuntemattomien parametrien lukumäärä. Vastemuuttujan ja sovitteiden välisen korrelaatiokertoimen (yhteiskorrelaatiokertoimen) neliön tulisi olla tässäkin mallissa yhtä suuri kuin mallin determinaatiokertoimen arvo. Tarkistetaan

```
> cor(tulot, fitted(m2))^2
```

3. Tarkastellaan seuraavaksi jäännöstermejä eli residuaaleja E_i tarkemmin. Standardoidut residuaalit R_i saadaan poimittua `lm()`-funktiolla luodusta malliobjektista funktiolla `rstandard()`. Ns. studentoidut residuaalit voidaan puolestaan poimia malliobjektista funktiolla `rstudent()`.

```
> res1 <- resid(m1) ; stand1 <- rstandard(m1) ; stud1 <- rstudent(m1)
> res2 <- resid(m2) ; stand2 <- rstandard(m2) ; stud2 <- rstudent(m2)
> data.frame(res1, stand1, stud1)
```

Moodle 9

```
> data.frame(res2, stand2, stud2)
```

Huomioi tulostuksessa kahden selittäjän mallissa havaintoyksikköön nro 5 liittyvä itseisarvoltaan suuri studentoidun residuaalin arvo. Kyseinen havaintoyksikkö havaittiin aiemmin vaikuttavuudeltaan vähäiseksi tilastoyksiköksi.

4. Piirretään seuraavaksi neljä diagnostista kuvaa, jotka saadaan geneerisen funktion `plot()` sillä erityisellä versiolla eli *metodilla* (nimeltään `plot.lm()`), joka tulee käyttöön, kun `plot()`-funktion pääargumentiksi annetaan `lm`-luokan malliobjekti; tässä tapauksessa malliobjektin nimi on `ma2`. Tarjolla on kaikkiaan 6 kuvaa, joista `which`-argumentilla valitsemme kuvat 1, 2, 3 ja 5. Ennen piirtämistä kuvaikkuna jaetaan neljään osaan (kahteen riviin ja kahteen sarakkeeseen).

```
> par(mfrow=c(2,2))
> plot(m2, which=c(1,2,3,5))
```

Jotta kuvien tulkinta olisi mielekästä, tulisi analysoitavan havaintoaineiston olla kooltaan suurempi. Diagnostiikkakuvien tulkinnasta löydät luentomateriaalin oheen lisätietoa esimerkiksi nettiosoitteesta `https://data.library.virginia.edu/diagnostic-plots/`.

5. Siirrytään seuraavaksi analysoimaan kuuluisaa *Minitab Tree Data* -aineistoa, joka löytyy Moodlesta tekstitiedostona `puut.txt`. Aineiston muuttujat ovat

`halk.m` = puun rungon halkaisija (m) mitattuna rinnan korkeudelta,
`kork.m` = puun rungon korkeus (m) ja
`tilav.m3` = puun runkotilavuus (m^3).

Kopioi aineisto koneellesi ja pidetään jatkossa mallituksissa vastemuuttujana tilavuutta ja kaksi muuta ovat selittäviä tai ennustavia muuttujia.

(a) Lue aineisto datakehikoksi `puut` ja listaa muodostetun datakehikon sisältö.

```
> puut <- read.table("puut.txt", header=TRUE)
> puut
> attach(puut)
```

(b) Piirretään havaintoaineiston muuttujien parittaiset sirontakuviot tavanomaisella piirtofunktiolla `plot()` ja lasketaan sirontakuvioihin liittyvät parittaiset korrelaatiokertoimet funktiolla `cor()`. Kuvataan asetelma lisäksi kolmiulotteisena kuvana

```
>    plot(puut)
>    cor(puut)
>    scatter3d(tilav.m3 ~ kork.m + halk.m surface=FALSE)
```

Moodle 10

(c) Lähdetään vasteen mallituksessa liikkeelle tilavuus- ja korkeusmuuttujilla. Piirrä sirontakuvio muuttujien `tilav.m3` ja `kork.m` välille uudestaan tavanomaiseen tapaan funktiolla plot() siten, että koko kuvaikkunan alue tulee hyödynnettyä kuvan piirrossa. Sovita tämän jälkeen aineistoon lineaarinen regressiomalli (`malli1`), jossa vasteena on muuttuja `tilav.m3` ja selittävänä muuttujana `kork.m`. Tulosta mallituksen keskeisimmät tulokset funktiolla `summary()` ja lisää sovitettu regressiosuora piirtämääsi sirontakuvioon funktiolla `abline()`.

```
> par(mfrow=c(1,1))
> plot(tilav.m3 ~ kork.m)
> malli1 <- lm(tilav.m3 ~ kork.m)
> summary(malli1)
> abline(malli1)
```

Saatujen tulosten perusteella nähdään mm., että korkeus-muuttujaan liittyvässä merkitsevyystestauksessa P-arvo on varsin pieni ja determinaatiokertoimen perusteella puun korkeudella voidaan selittää noin 35.6 % puun tilavuuden kokonaisvaihtelusta. Mallilla on siis ainakin jossain määrin selityskykyä vasteeseen nähden. Mutta ovatko havainnot sopusoinnussa mallitukseen liittyvien oletusten kanssa?

Piirrä mallioliosta `malli1` saatavat neljä diagnostista kuvaa komennolla

```
> par(mfrow=c(2,2)) ; plot(malli1, which=c(1:3, 5))
```

Näyttävätkö diagnostiikkakuvien perusteella malliin liittyvät oletukset realistisilta? Moodle 11

(d) Vakiovarianssiongelman poistamisen yhtenä keinona luentomonisteessa mainitaan erilaisten muunnosfunktioiden käyttö. Kokeillaan seuraavaksi käyttää edellisen mallituksen selittäjän eli puun korkeuden logaritmimuunnosta mallin ainoana selittäjänä. Tehdään tarvittava uusi muuttuja ja piirretään alkuperäisen vastemuuttujan (`tilav.m3`) ja logaritmoidun korkeusmuuttujan (`log.kork`) välinen sirontakuvio. Lisätään havaintoaineistoon samalla tulevia tarpeita varten myös muiden muuttujien logaritmoidut versiot.

```
> puut$log.kork <- log(puut$kork.m)
> puut$log.halk <- log(puut$halk.m)
> puut$log.til <- log(puut$tilav.m3)
> attach(puut)
> par(mfrow=c(1,1))
> plot(tilav.m3 ~ log.kork)
```

Tutki hetki piirrettyä sirontakuviota. Poistuiko vakiovarianssiongelma? Moodle 12

(e) Joudumme siis muotoilemaan edelleen mallia. Kokeillaan seuraavaksi tehdä mallitus siten, että vastemuuttujana on alkuperäisen vasteen `tilav.m3` logaritmoitu versio `log.til` ja selittäjänä alkuperäinen korkeusmuuttuja `kork.m`

```
> plot(log.til ~ kork.m)
```

Tämän alustavan tarkastelun perusteella tilanne näyttäisi nyt paremmalta ainakin vakiovarianssioletuksen näkökulmasta. Jatketaan tällä perusteella analyysiä mallitusvaiheeseen

```
> malli2 <- lm(log.til ~ kork.m)
> summary(malli2)
> abline(malli2)
> par(mfrow=c(2,2)) ; plot(malli2, which=c(1:3, 5))
```

Mitä voimme nyt sanoa mallioletusten paikkansapitävyydestä diagnostiikkakuvien perusteella? Moodle 13 Entä miltä mallin ”hyvyys” näyttää determinaatiokertoimen perusteella? Moodle 14

6. Laajennetaan vielä edellä muodostettua `malli2`:ta kahden selittäjän malliksi ottamalla toiseksi selittäjäksi puun halkaisija `halk.m`.

(a) Suoritetaan tarvittavat mallituskomennot

```
> malli3 <- lm(log.til ~ halk.m + kork.m)
> summary(malli3)
> confint(malli3)
> plot(malli3)
```

Ensisilmäyksellä mallin näyttää varsin toimivalta, sillä determinaatiokertoimen arvo on korkea ja diagnostiikkakuvat eivät osoita mitään hälyttäviä seikkoja oletusten suhteen. Moodle 15 Miten tässä mallissa tulkitaan selittäjään `halk.m` liittyvä regressiokertoimen β_2 estimaatti 5.72? Moodle 16

(b) Edellinen malli ei ole vielä optimaalinen ns. ”lopulliseksi” malliksi mm. kertoimien tulkintaan liittyvien haasteiden takia. Viimeisen mallin muodostamisessa käytämme hyväksi pelkästään logaritmoituja muuttujia ja haemme tukea/perusteita valinnallemme matematiikasta ja kappaleiden tilavuuslaskentaan liittyvistä teorioista.

Jos oletamme puun rungon olevan muodoltaan kartion, saamme laskettua rungon tilavuuden V kaavalla

$$V = \frac{1}{3}\pi r^2 h$$

, missä r on pohjan säde ja h on kartion korkeus (`katso kuva tiedoston lopusta`). Aineistossamme puun rungosta oli mitattu pohjan säteen sijasta pohjan halkaisija d (`kork.m`) ja siten $r = d/2$.

Tilavuuden laskentakaava saadaan nyt muotoon $V = \frac{1}{3}\pi(\frac{d}{2})^2 h \quad \Leftrightarrow \quad V = \frac{\pi}{12}d^2 h$

Otetaan seuraavaksi logaritmi yhtälön molemmilta puolilta: $log(V) = log(\frac{\pi}{12}d^2 h)$

ja hyödynnetään logaritmiin liittyviä laskusääntöjä: $log(V) = log(\frac{\pi}{12}) + log(d^2) + log(h)$

$$\Leftrightarrow \quad log(V) = log(\tfrac{\pi}{12}) + 2\ log(d) + log(h).$$

Yllä esitetystä tilavuuden laskukaavasta voimme nyt päätellä, että jos puun rungon muoto oletetaan kartioksi ja mallituksessa käytetään muuttujien logaritmoituja versioita, päädymme malliin, jonka systemaattinen osa on muotoa

$$log(tilavuus) = \beta_0 + \beta_1 \cdot log(\text{halkaisija}) + \beta_2 \cdot log(\text{korkeus})$$

Edellä tehdyllä oletuksilla odotamme siis mallin parametrien estimaattien olevan (ainakin likimain) seuraavat:

$$\widehat{\beta_0} = log(\frac{\pi}{12}) \approx -1.34, \ \widehat{\beta_1} = 2 \text{ ja } \widehat{\beta_2} = 1$$

(c) Muodostetaan nyt "lopullinen" malli edellä esitetyllä periaatteella ja piirretään malliin liittyvät diagnostiikka kuvat

```
> malli.final <- lm(log.til ~ log.halk + log.kork )
> summary(malli.final)
> confint(malli.final)
```
Moodle 17
```
> plot(malli.final)
```
Moodle 18

(d) Tarkistetaan vielä, löytyykö aineistosta tämän mallin näkökulmasta katsottuna erityisen vaikuttavia havaintoja tai outlier-havaintoja

```
> data.frame(hatvalues(malli.final), rstandard(malli.final) )
```
Moodle 19 Moodle 20