TILASTOLLISTEN MENETELMIEN PERUSTEET. 10.1.–4.3.2022. Tampereen yliopisto. Yliopisto-opettaja Pekka Pere.

Koe 4.3.2021

Koe alkaa 9.00 ja päättyy 12.20.

Esitä kaikkien laskujesi välivaiheet, ja perustele kaikki vastauksesi yksityiskohtaisesti. Pelkkä oikea vastaus on nollan pisteen arvoinen. Tehtävät ovat kuuden pisteen arvoisia paitsi ensimmäinen, jossa on ylimääräinen lisäpistekohta. Lisäpistekohdalla voit parantaa arvosanaasi; arvostelua ei alenna, jos et vastaa lisäpistekohtaan. **Palauta vastauksesi neljään kysymykseen. Riippumatta siitä, kuinka monta tehtävää olet palauttanut, vastaa Moodlessa 6. kysymykseen, mitkä tehtävät tulee arvostella.** Jos et vastaa, arvostelu ei välttämättä perustu parhaisiin tai jättämiisi vastauksiisi.

Vastausten tulee olla käsinkirjoitettuja. Myös lyhyet R-koodit, jos sellaisia esität, tulee olla käsinkirjoitettuja. Vastauksissa tulee käyttää kaavoja ja perustella laskut ja päätelmät huolellisesti. (Esim. pelkkä R-käsky ja oikea vastaus eivät ole riittävä selitys.)

Skannatkaa kunkin tehtävän vastauksenne pdf-tiedostoksi ja palauttakaa se Moodlessa kyseisen tehtävän kohtaan. Lähettäkää vastauksia kysymyksiin pitkin koeaikaa. Jos kaikki palauttaisivat vastauksensa viime tingassa, Moodle saattaisi tukkeutua, ettekä saisi lähetettyä vastauksianne. Ennen jättämistä tarkistakaa vastauksenne huolellisesti. Ja vielä toisen kerran!

Kaikissa pdf-tiedostoissa tulee heti alussa näkyä selkeästi nimenne ja opiskelijanumeronne. Huom! Skannaussovellus saattaa laittaa pdf:ään logon tai vastaavan. Se ei saa peittää vastauksenne osaa. Tehtävästä saa pisteitä vain vastauksen mukaan, joka näkyy pdf-tiedostossa.

Voitte käyttää Moodle-sivun luentomonistetta ja taulukoita, mitä tahansa tilastotieteellistä kirjallisuutta, Googlea, Internettiä, kuinka tahansa kehittyneitä laskimia, tietokonetta, tilasto-ohjelmia jne. apuna vastaamisessanne. Toista ihmistä ette saa millään tavalla käyttää apunanne. Ette saa pohtia vastauksia tai vastata ryhmissä, soittaa neuvoja, s-postitse, tekstiviestitse tai millään muulla tavalla olla kokeen aikana kehenkään yhteydessä, joka voisi auttaa teitä.

Saatan liittää kokeeseen suullisen kuulustelun joillekin opiskelijoille, jos hahmotan sen heidän osaltaan tarpeelliseksi. Suullisen kuulustelun läpäisy on heille edellytys tentin läpäisemiselle. Ilmoitan tällaisesta tarpeesta kyseisille opiskelijoille kokeiden tarkastamisen jälkeen.

Tarpeen vaatiessa lähetän s-postitse lisäohjeita kokeen aikana. Teidän tulee palauttaa vastauksenne ennen koeajan loppumista. Vakavassa ongelmatilanteessa minulle voi s-postittaa (pekka.pere@tuni.fi) tai soittaa (050 437 7568). Hätätilanteessa palauttakaa vastauksenne minulle s-postitse.

Parhainta koemenestystä!

1. Maailman talousfoorumi julkaisee vuosittain raportin sukupuolten tasa-arvosta eri maissa. Yksi raportin tuotos on sukupuolten tasa-arvoindeksi *Global Gender Gap Index* (GGGI eli GGG-indeksi). Se saa lukuja väliltä $[0, 1]$, jossa 1 tarkoittaa täydellistä sukupuolten tasa-arvoa. Vuoden 2021 raportin mukaan Suomi on toiseksi tasa-arvoisin tutkituista 156 maasta.

Tehtävä perustuu Stoetin ja Gearyn (2018) artikkeliin ja siinä käytettyyn vuoden 2015 GGG-indeksiin.[1] Toinen tutkittava muuttuja on naisten osuus (%) korkeakoulututkinnon suorittaneista luonnontieteissä, insinööritieteissä ja matematiikassa eli STEM-aineissa (*Science, Technology, Engineering, Mathematics*).

Ylempään sirontakuvioon on piirretty vastakkain STEM-osuus ja GGG-indeksi 50 maassa. Kuviosta nähdään, että Suomi ja Norja ovat olleet GGG-indeksin arvolla 0.85 tasa-arvoisimmat maat 2015. Molemmissa maissa STEM-osuus on pieni; Suomessa vain 20 %.

Alla on tiivistettyä R:n palautetta regressiosta, jossa on selitetty STEM-osuutta GGG-indeksillä (`data$GGGI`):

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.702  -4.111   0.498   3.439  10.098
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   70.037      9.867   7.098 5.21e-09 ***
## data$GGGI    -59.335     13.597  -4.364 6.76e-05 ***
## Residual standard error: 5.047 on 48 degrees of freedom
## Multiple R-squared:  0.284,     Adjusted R-squared:  0.2691
## F-statistic: 19.04 on 1 and 48 DF,  p-value: 6.76e-05
```

a) Mitä malli ennustaa STEM-osuudeksi, jos GGG-indeksi saa arvon 0? Onko tällainen ennuste luotettava? Perustele.

b) Jos GGGI-indeksi kasvaa yksiköllä, miten STEM-osuus muuttuu mallin mukaan? Entä jos GGGI-indeksi kasvaa 0.1 yksiköllä? Kumpi lasku on mielekkäämpi?

c) Miten t-arvo -4.364 voidaan laskea R-palautteen muista tunnusluvuista?

d) Selitä yksityiskohtaisesti, miksi t- ja F-testin p-arvo on sama. (Vihje: oppimateriaalin 17.2.-version s. 270.)

e) Selitetäänkin GGG-indeksiä STEM-osuudella (`data$STEM`) — kuten Stoet

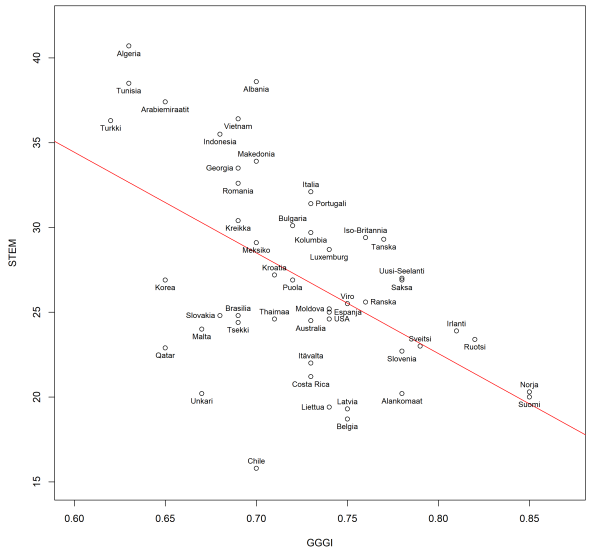
[1] `https://www.weforum.org/reports/global-gender-gap-report-2021` (haettu 25.2.2022). G. Stoet ja D.C. Geary (2018): The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education. *Psychological Science*, 29, 581–593. Aineisto: `https://journals.sagepub.com/doi/suppl/10.1177/0956797617741719`. Kiitän Kimmo Vattulaista aineiston muokkaamisesta ja kuvien piirtämisestä.

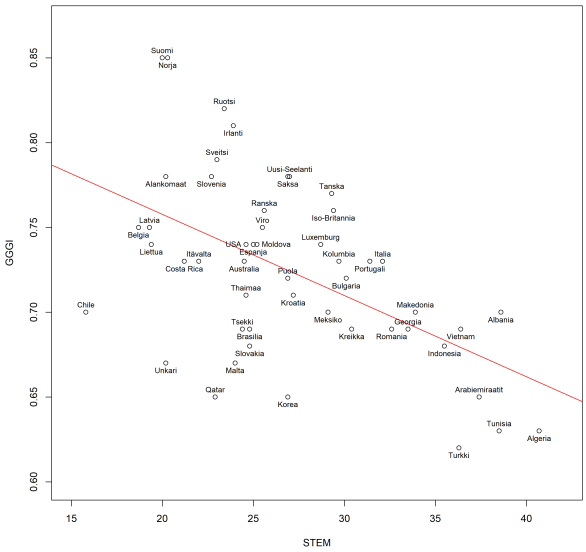
ja Geary (2018) tekevät artikkelissaan. Regressiosuora on alemmassa sirontakuviossa. Nyt R palauttaa tulokset alla:

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.093858 -0.023833 -0.003328  0.028509  0.093696
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.853482   0.030401  28.074  < 2e-16 ***
## data$STEM   -0.004787   0.001097  -4.364 6.76e-05 ***
## Residual standard error: 0.04533 on 48 degrees of freedom
## Multiple R-squared:  0.284,     Adjusted R-squared:  0.2691
## F-statistic: 19.04 on 1 and 48 DF,  p-value: 6.76e-05
```

Regressiokerroin on muuttunut. Miten sen t-arvo voi olla sama kuin d)-kohdassa?

f) Merkitään $\hat{\beta}_1$:llä tai $\hat{\beta}_1^*$:lla regressiokertoimen PNS-estimaattia, kun selitettävä on y tai x (tavanomaisesta merkintätavasta poiketen). Osoita, että $\hat{\beta}_1^* = (s_x^2/s_y^2) \times \hat{\beta}_1$. R-komennot `var(data$STEM)` ja `var(data$GGGI)` laskevat otosvarianssit 34.85194 ja 0.002811796. Osoita, että yhtälö $\hat{\beta}_1^* = (s_x^2/s_y^2) \times \hat{\beta}_1$ pätee numeerisesti tehtävän tilanteessa.





2. Nuoren tytön itsevarmuus (jatkoa). Mallitetaan nuoren tytön itsevarmuutta osoitinmuuttujilla äidin (x) ja isän (z) kasvatustyyleistä (laiminlyövä = 1, autoritaarinen = 2, salliva = 3 ja auktoritatiivinen = 4). Osoitinmuuttujat saavat arvon 1 vanhemman kasvatustyylin voimassaollessa ja 0 muuten. Estimoidaan PNS-menetelmällä mallit alla (keskivirheet ovat suluissa):

$$\begin{array}{rcccccccccc} \hat{y} & = & 3.252 & - & 0.414x_1 & - & 0.135x_4 & + & 0.131z_4 & + & 0.548x_4z_4. \\ & & (0.122) & & (0.191) & & (0.174) & & (0.265) & & (0.311) \end{array}$$

$$s = 0.796,\ \mathsf{R}^2 = 0.1614,\ F = 8.275,\ n = 177.$$

ja

$$\begin{array}{rcccccc} \hat{y} & = & 3.206 & - & 0.364x_1 & + & 0.590x_4z_4. \\ & & (0.082) & & (0.171) & & (0.134) \end{array}$$

$$s = 0.794,\ \mathsf{R}^2 = 0.1557,\ F = 16.040,\ n = 177.$$

a) Mikä suure R^2 on? Miksi se on pienempi alemmassa mallissa? Mikä suure s on? Miksi se on pienempi alemmassa mallissa? (Vihje: Pohdi, miten s^2 määritellään ja miksi se voi käyttäytyä tehtävässä esitetyllä tavalla.)

b) Mitä hypoteesia F-testisuure testaa? Mikä on sen p-arvo alemmassa mallissa? Mitä päättelet sen perusteella?

c) Testaa alemman mallin yksittäisten regressiokertoimien estimaattien tilastollista merkitsevyyttä merkitsevyystasolla 0.05 (kaksisuuntaiset testit). Mitä päättelet?

d) Tulkitse alempi malli huolellisesti (vakiotermin ja regressiokertoimien estimaatit ja mihin verrataan, minkä arvon itsevarmuus kussakin tilanteessa saa).

3. Maiden rahallisen panostuksen huippu-urheiluun 2020 ja voitettujen mitalien määrää Tokion olympialaisissa 2020 tutkittiin aineistolla alla.[2]

maa	euro	mitali
Japani	229.030309	58
IsoBritannia	202.000000	65
Kanada	134.865699	24
Brasilia	117.119140	21
Sveitsi	107.608811	13
Puola	73.951093	14
Alankomaat	73.778223	36
UusiSeelanti	41.965039	20
Tanska	40.665183	11
BelgiaF	30.512558	5
Portugali	30.651000	4
Suomi	29.049400	2
Ruotsi	26.059694	9
BelgiaV	14.269597	2

Tallenna aineisto nimellä Panostuotos.txt sopivaan hakemistopolkuun. (Tehtävänlaatijalla polku on "`F:\\Aineisto\\Tokio_2020\\Panostuotos.txt`" R:n merkinnöillä. Huom! Tiedosto ei saa päättyä numeroon 2. Paina sen jälkeen syöttönäppäintä ↲. Se tuottaa tiedoston lopetusmerkin. Lopuksi tallenna tiedosto.) Iso-Britannian ja Uuden-Seelannin nimistä on taulukossa poistettu yhdysviivat R:ään lukemisen helpottamiseksi. "BelgiaF" ja "BelgiaV" ovat Belgian

[2] V. De Bosscher, S. Shibli ym. (2021): Tokyo 2020 Evaluation of the Elite Sport Expenditures and Success in 14 Nations. Vrije Universiteit Brussel. SPLISS. Kiitän Aline var Roey'jea aineiston luovuttamisesta 14.2.2022. Tokion olympialaiset 2020 järjestettiin vasta 2021 COVID-19-pandemian takia.

Flanderi ja Vallonia. “Euro” on maan rahallinen panostus miljoonissa euroissa huippu-urheiluun 2020 ja “mitali” on maan urheilijoiden voittamien mitaleiden lukumäärä Tokion olympialaisissa 2020.

Lue aineisto R:ään alla olevan tapaisella komennolla:

```
Tokio <- read.table("f:\\aineisto\\Tokio_2020\\Panostuotos.txt",header=T)
```

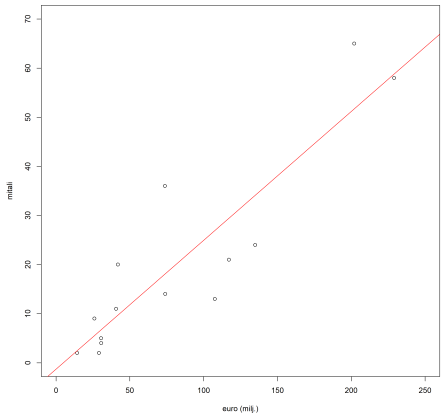
Sirontakuvio havainnollistaa aineistoa. Vastaa kysymyksiin alla olettaen, että kaikki vaaditut oletukset lineaarisesta regressiosta pätevät kohdissa a–c).

a) Aja R:ssä regressio, jossa selität mitalien lukumäärää huippu-urheiluun panostetulla rahamäärällä. (Sinun tulee osata itsenäisesti antaa sopivat R-komennot.) Mikä on regression selitysaste? Mikä on huippu-urheiluun panostetun rahamäärän ja mitalien lukumäärän korrelaatio?

b) Onko huippu-urheiluun panostetun rahamäärän ja mitalien lukumäärän yhteys tilastollisesti merkitsevä merkitsevyystasolla 0.001?

c) Onko Suomi saanut hyvän mitalivastineen rahapanostukselleen regression mukaan? Perustele.

d) Pätevätkö kaikki lineaariseen regressiomalliin oppimateriaalissa liitetyt oletukset tehtävän tilanteessa? Perustele. (Vihje: Pohdi vastetta. Lyhyt vastaus riittää.)



4. Dahl ja Moretti (2008) estimoivat lineaarisia todennäköisyysmalleja.[3] Niissä on monia selittäviä tekijöitä (vanhemman koulutus, ihonväri jne.), jotka ja joiden merkitys sivuutetaan artikkelissa ja tehtävässä. Molemmissa keskitytään mallien ydinsanomaan. Kaikki alla viitatut estimoidut kertoimet ovat tilastollisesti merkitseviä merkitsevyystasolla 0.05.

a) Havainnot ovat yhdysvaltalaiset kotitaloudet, joissa lapsi tai useampia asuu 18–40-vuotiaan isän, äidin tai molempien kanssa ($n = 4\,681\,967$; 1960–2000). Mallilla pyritään ymmärtämään, millaisissa kotitalouksissa ei ole isää ($Y = 1$) tai on ($Y = 0$). Mallin estimoitu vakio on 0.162. Mallin osoitinmuuttuja saa arvon 1, jos ensimmäinen lapsi kotitaloudessa on tyttö. Osoitinmuuttujan estimoitu kerroin on 0.0050. Mallissa lähtökohta (vertailuluokka) on kotitalous, jossa perheen ensimmäinen lapsi on poika. Tulkitse estimoitu malli huolellisesti.

b) Aineisto on ensimmäisen lapsensa synnyttäneiden kalifornialaisten äitien lasten syntymäkortit ($n = 1\,403\,601$; 1989–1994). Mallilla pyritään ymmärtämään, millaisissa olosuhteissa äiti on ($Y = 1$) tai ei ole ($Y = 0$) naimisissa ensimmäisen lapsensa synnyttäessään. Mallin estimoitu vakio on 0.084. Mallissa on kaksi osoitinmuuttujaa. Ensimmäinen saa arvon 1, jos äidin ensimmäinen lapsi on tyttö, ja arvon 0 muulloin. Tämän osoitinmuuttujan estimoitu kerroin on 0.0019. Toinen osoitinmuuttuja saa arvon 1, jos äidille on tehty ultraäänitutkimus raskauden aikana, ja arvon 0 muulloin. Sikiön sukupuoli selviää usein tutkimuksessa. Yksinkertaisuuden vuoksi tässä oletetaan, että sukupuoli selviää tutkimuksessa aina. Tämän osoitinmuuttujan estimoitu kerroin on 0.0303. Mallissa on myös yhdysvaikutusmuuttuja osoitinmuuttujien tulo. Sen kerroin on -0.0046. Mallissa lähtökohta (vertailuluokka) on äiti, joka ei ole naimisissa alkaessaan odottamaan lasta mutta on mennyt naimisiin ennen synnyttämistä ja että synnyttämänsä lapsi on poika. Tulkitse estimoitu malli huolellisesti.

5. Aineisto alla on Lainialan ja Säävälän (2013)[4] taulukosta 2. Tiedot monikulttuurisessa avioliitossa olevista suomea, ruotsia tai saamea äidinkielenään puhuvista (jatkossa suomalaistaustaisista tai suom.) 25–44-vuotiaista (lapsettomista tai yhden lapsen saaneista) miehistä ja naisista on kerätty 2012. Vastaavat tiedot yksikulttuurisessa avioliitossa olevista suomalaistaustaisista miehistä ja naisista on kerätty 2008. Taulukkoon alle on kirjattu osuudet ja lukumäärät suomalaistaustaisista miehistä ja naisista, jotka ovat pohtineet avioeroa edeltävän

[3] G.B. Dahl ja E. Moretti (2008): The Demand for Sons. *Review of Economics Studies*, 75, 1085–1120.

[4] L. Lainiala ja M. Säävälä (2013): Intercultural Marriages and Consideration of Divorce in Finland: Do Value Differences Matter? Väestöliiton Väestöntutkimuslaitoksen työpaperi 2013 (4). Tehtävän aineisto on osin päätelty Lainialan ja Säävälän tiedoista.

vuoden aikana eriteltynä sen mukaan, elävätkö he moni- vai yksikulttuurisessa avioliitossa.

			pohtinut avioeroa		
suom.	avioliitto	kyllä (%)	kyllä	ei	n
mies	monikulttuurinen	18.5	29	128	157
	yksikulttuurinen	12.3	48	342	389
nainen	monikulttuurinen	27.9	78	202	280
	yksikulttuurinen	24.1	127	399	526

Tutkitaan suomalaistaustaisten miesten avioeroajatuksia moni- ja yksikulttuurisissa avioliitoissa.

a) Laske riskisuhde (*risk ratio*) avioeroajatuksia omaaville miehille moni- ja yksikulttuurisissa avioliitoissa. Tulkitse laskemasi riskisuhde sanallisesti. Olisiko riskisuhde sama avioeroajatuksia omaamattomille miehille moni- ja yksikulttuurisissa avioliitoissa? Perustele.

b) Laske vastasuhde (*odds*) avioeroajatuksia omaaville miehille monikulttuurisissa avioliitoissa. Tulkitse laskemasi vastasuhde sanallisesti.

c) Laske ristisuhde (*odds ratio*) avioeroajatuksia omaaville miehille moni- ja yksikulttuurisissa avioliitoissa. Tulkitse laskemasi ristisuhde sanallisesti.