TILASTOLLISTEN MENETELMIEN PERUSTEET. 8.3.–23.4.2021. Tampereen yliopisto. Yliopistonlehtori Pekka Pere.

Uusintakoe 7.6.2021

Koe alkaa 17.00 ja päättyy 21.00.

Esitä laskujesi välivaiheet, ja perustele vastauksesi yksityiskohtaisesti. Pelkkä oikea vastaus on nollan pisteen arvoinen. Kaikki tehtävät ovat samanarvoisia. **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 vastauksiisi.

Vastausten tulee olla käsinkirjoitettuja. Myös lyhyet R-koodit, jos sellaisia esitätte vastauksissanne, tulee olla käsinkirjoitettuja. Vastauksissa tulee käyttää kaavoja ja perustella laskut ja päätelmät huolella. (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 jättäisivät vastausten jättämisen viime tinkaan, Moodle saattaisi tukkeutua. Ennen jättämistä tarkistakaa vastauksenne huolellisesti. Ja vielä toisen kerran!

Kaikissa pdf-tiedostoissa tulee heti alussa näkyä selkeästi nimenne ja opiskelijanumeronne. Tarkistakaa ennen vastauksen lähettämistä, että se on tallentunut kokonaisena ja selkeänä. 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 luentomonistetta ja taulukoita, mitä tahansa tilastotieteellistä kirjallisuutta, 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 vastauksissa.

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. Vakavassa ongelmatilanteessa minulle voi s-postittaa (pekka.pere@tuni.fi) tai soittaa (050 437 7568). Koemenestystä!

1. Tutkitaan yhdysvaltalaista aineistoa nuorten työntekijöiden (jatkossa "miesten" ja "naisten") tuntipalkoista vuonna 1987 (*US National Longitudinal Survey*; 3 294 havaintoa, joista 1 698 on miehiä).[1] Tutkitaan miesten ja naisten palkkaeroja. Aineistossa muuttuja m (*male*) saa arvon 1, jos työntekijä on mies ja arvon 0, jos työntekijä on nainen. Selitetään palkkoja (w; *wage*) sukupuoliosoittimella m:

$$\begin{array}{rcccccc} w_i & = & 5.15 & + & 1.17 m_i & + & \hat{\varepsilon}_i, \\ & & (0.081) & & (0.112) & & \end{array} \tag{1}$$

$$s = 3.217,\ \mathsf{R}^2 = 0.032,\ F = 107.93.$$

Lisätään malliin koulutustasoa mittava muuttuja sc (*schooling*; vuosia) sekä työkokemusta mittaava muuttuja e (*experience*; vuosia):

$$\begin{array}{rcccccccc} w_i & = & -3.38 & + & 1.34 m_i & + & 0.64 sc_i & + & 0.12 e_i & + & \hat{\varepsilon}_i, \\ & & (0.465) & & (0.108) & & (0.033) & & (0.024) & & \end{array} \tag{2}$$

$$s = 3.046, \mathsf{R}^2 = 0.133, F = 167.63.$$

Yhtälöt on estimoitu pienimmän neliösumman menetelmällä. Luvut suluissa ovat estimoituja keskivirheitä.

a) Miten regressioissa raportoidut F- ja R^2-suureet liittyvät toisiinsa?

b) Keskitytään tutkimaan mallia (2). Ovatko sukupuolta, koulutusta ja työkokemusta mittaavat muuttujat yhdessä tilastollisesti merkitseviä selittäjiä 5 prosentin merkitsevyystasolla? Selitä huolella, mitä päättelet. Mikä on testisuureen p-arvo? Ovatko kaikki selittäjät sukupuoli, koulutus ja työkokemus yksinään tilastollisesti merkitseviä selittäjiä 5 prosentin merkitsevyystasolla? Selitä kaikkien testien nolla- ja vastahypoteesit huolella. Mitä päättelet?

c) Mallissa (2) suure

- s saa pienemmän
- R^2 saa suuremman
- F saa suuremman

[1] Tehtävä perustuu M. Verbeekin (2008) kirjassaan (*A Guide to Modern Econometrics*, 3. laitos. Wiley) raportoimiin estimointeihin.

arvon kuin mallissa (1). Käykö kaikkien näiden suureiden kohdalla näin väistämättä, kun malliin lisätään uusia muuttujia? Perustele huolellisesti.

d) Mallissa (2) sukupuoliosoittimen kertoimen estimaatti on itseisarvoltaan suurempi kuin mallissa (1). Kasvavatko kertoimien itseisarvojen estimaatit aina, kun malliin lisätään uusia muuttujia? Vapaaehtoinen lisäpistekohta: Mistä suureneminen johtuu? Keksi selitys. Perustele huolellisesti.

2. Keväällä 2020 ihmeteltiin, miksi Afrikassa SARS-CoV-2:n (koronaviruksen) aiheuttama COVID-19-tauti ei ole levinnyt tulipalon lailla huolimatta köyhyydestä, huonosta hygieniasta, paikallisesti suurista väestötiheyksistä ja muista sentapaisista viruksille yleensä edullisista olosuhteista. Olisiko selitys alla?

Wang ym. (2020) mallittivat COVID-19-taudin uusiutumislukua (*reproductive number R*) sadassa kiinalaisessa kaupungissa 21.–23.1.2020.[2] Mitä suurempi uusiutumisluku on, sitä useamman ihmisen sairastunut tartuttaa. Uusiutumisluku tulisi saada alle yhden, jotta epidemia päättyisi. Selittäjät ovat ilman lämpötila (c celsiusastetta) ja suhteellinen kosteus (k %). Yksi Wangin ym:iden PNS-menetelmällä estimoimista malleista on

$$\begin{array}{rcccccccc} R & = & 2.802 & - & 0.0287c & - & 0.00892k & + & \hat{\varepsilon} \\ & & (6.66) & & (-4.25) & & (-1.74) & & \end{array}$$

$$n = 100, \quad \mathsf{R}^2 = 0.22.$$

Luvut suluissa ovat t-arvoja.

Aineistossa lämpötila on vaihdellut noin −21:n ja 23:n asteen ja suhteellinen ilmankosteus noin 45 ja 100 %:n välillä. Lämpötilan keskiarvo ja -hajonta ovat 6.183 ja 7.283. Suhteellisen kosteuden keskiarvo ja -hajonta ovat 83.340 ja 9.567.

a) Laske F-testisuure ja sen p-arvo nollahypoteesille, että selittäjien kertoimet ovat nollia. Laske t-arvojen p-arvot (kaksisuuntaiset testit). Ovatko selittäjien kertoimien estimaatit tilastollisesti merkitseviä?

b) Tulkitse malli sanoin, eli miten lämpötila ja kosteus ovat yhteydessä uusiutumislukuun. Pohdi vakiotermin tulkintaa.

c) Wang ym. esittävät useita vertailevia arvioita lämpötilan ja kosteuden vaikutuksesta uusiutumislukuun. Voitko arvioida, kumpi selittäjä vaikuttaa enemmän uusiutumislukuun? Perustele. Jos voit, arvioi.

[2] J. Wang, K. Tang, K. Feng ja W. Lv (2020): High Temperature and High Humidity Reduce the Transmission of COVID-19. `https://ssrn.com/abstract=3551767` (haettu 22.3.2020). Käsikirjoituksesta julkaistaan 2021 kehittyneempiä menetelmiä soveltava versio lehdessä *BMJ Open*, 11, e043863.

d) Tokion olympialaiset 24.7.–9.8.2020 siirrettiin tapahtuvaksi 23.7.–8.8.2021 COVID-19-pandemian takia. Keskimäärin heinäkuisessa Tokiossa lämpötila on 28 astetta ja ilmankosteus 85 %. Millaista uusiutumislukua malli ennustaa Tokioon suunniteltujen olympialaisten aikaan?

3. Nuoren tytön itsevarmuutta (indeksi saa arvoja välillä $[1, 5]$) selitetään luokitteluasteikollisella muuttujalla äidin kasvatustyyli (AVT). Kasvatustyylejä on neljä: laiminlyövä, autoritaarinen (AVT.aut), salliva (AVT.sal) ja auktoritatiivinen (AVT.auk). Kolmea viimeksi mainittua kasvatustyyliä varten on luotu osoitinmuuttuja. PNS-estimointi tuottaa (karsitun) R-palautteen alla.

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.8333     0.1502  18.866  < 2e-16 ***
tdata$AVT.aut   0.3078     0.2497   1.233 0.219281
tdata$AVT.sal   0.5095     0.2047   2.490 0.013704 *
tdata$AVT.auk   0.6696     0.1710   3.915 0.000128 ***
Residual standard error: 0.8226 on 179 degrees of freedom
Multiple R-squared:  0.08349,   Adjusted R-squared:  0.06813
F-statistic: 5.435 on 3 and 179 DF,  p-value: 0.001338
```

Luodaan laiminlyövälle kasvatustyylille osoitinmuuttuja (AUT.lai). Estimoidaan malli uudestaan asettamalla auktoritatiivinen kasvatustyyli oletusluokaksi. R:n (karsittu) palaute on alla.

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.50297    0.08185  42.797  < 2e-16 ***
tdata$AVT.lai -0.66964    0.17104  -3.915 0.000128 ***
tdata$AVT.aut -0.36179    0.21565  -1.678 0.095147 .
tdata$AVT.sal -0.16011    0.16135  -0.992 0.322365
Residual standard error: 0.8226 on 179 degrees of freedom
Multiple R-squared:  0.08349,   Adjusted R-squared:  0.06813
F-statistic: 5.435 on 3 and 179 DF,  p-value: 0.001338
```

Mitä vakio mittaa nyt? Miksi kertoimien estimaatit ovat muuttuneet? Miten on mahdollista, että salliva kasvatustyyli vaikutti ensimmäisessä estimoinnissa tilastollisesti merkitsevältä selittäjältä ($p < 0.05$) muttei jälkimmäisessä?

4. Arvioi lainausten alla paikkansapitävyyttä huolellisesti perustellen omaan oppimateriaalimme tukeutuen. Vastaa täsmälleen kolmeen kohtaan tehtävässä. Jos vastaat useampaan, kerro vastauksesi alussa, mitkä kohdat tulee arvostella.

Soveltajan kirjoittamassa tilastotieteen oppikirjassa (2004) opastetaan [hakasuluissa selvennyksiä]:

> a) Tarkastellaan seuraavaksi *logistisia regressiomalleja*, jotka ovat lineaarisen regressiomallin yleistyksiä tilanteeseen, jossa selitettävä muuttuja on kategorinen [luokitteluasteikollinen].
>
> b) *Logistisen regression* etuna [lineaariseen regressiomalliin verrattuna] on se, että selittävien muuttujien mitta-asteikoista ei tehdä mitään oletuksia, vaan ne voivat olla yhtä hyvin laatuero- [luokittelu-], järjestys-, välimatka- tai suhdeasteikollisia. – – Logististen regressiomallien suurin etu on se, että logistiset mallit tekevät huomattavasti vähemmän oletuksia kuin lineaariset mallit. Ensinnäkin logistinen regressioanalyysi ei tee mitään oletuksia mallissa käytettävien muuttujien jakaumista. Edelleen muuttujien välisten yhteyksien tyypistä ei oleteta mitään: yhteydet voivat olla lineaarisia, eksponentiaalisia tai vaikkapa logaritmisia.
>
> c) Logistisen regressiomallin lähtökohta on ns. *riskisuhde* (engl. *odds ratio* = OR), joka on tapahtumien 0 ja 1 todennäköisyyksien osamäärä. Riskisuhde määritellään seuraavasti:
>
> $$OR = \frac{P}{1-P}$$
>
> [jossa P on tapahtuman todennäköisyys].
>
> d) Regressiokertoimet (betat) ovat logistisessa regressiossa *riskisuhteita.* Kaikkein yksinkertaisinta on tarkastella riskisuhteiden muutoksia $\exp(B)$ [B on regressiokerroin yhden selittäjän logistisessa regressiossa]. Tarkastellaan edellen tutkimusta, jossa on ennustettu todennäköisyyttä sille, että lapsi on turvallisesti kiintynyt. Jos esimerkiksi äidin ikä-muuttujan $\exp(B)$-kerroin on 1.05, niin jokainen äidin ikävuosi kasvattaa 5 %:lla turvallisen kiintymyksen todennäköisyyttä. Jos selitettävä muuttuja on kategorinen [osoitinmuuttuja], $\exp(B)$ ilmaisee kuinka paljon selittävän muuttujan tiettyyn luokkaan kuuluminen kasvattaa todennäköisyyttä sille, että tutkittava kuuluu selitettävän muuttujan referenssikategoriaan.
>
> e) Logistinen regressioanalyysi on lineaarisen regressioanalyysin epäparametrinen vastine. – – Logistinen regressioanalyysi on *epäparametrinen* testi, joka ei tee jakaumaoletuksia.

5. Hugh-Jones (2016) raportoi Internetin joukkoistamispalvelusta rekrytoimillaan koehenkilöillä teettämistään rehellisyyskokeista. Lanttikokeessa koehenkilö heittää lanttia. Hänelle on kerrottu, että hän saa pienen rahasumman, jos tuli kruuna. Seuraavaksi koehenkilö kertoo, saiko kruunan vai klaavan. Tietovisassa koehenkilölle esitetään kuusi kysymystä musiikista ja kerrotaan pienestä rahapalkkiosta, jos osaa vastata oikein kaikkiin kysymyksiin. Kysymykset olivat sellaisia, että harva tietää oikean vastauksen kaikkiin mutta osaan vastaus löytyy helposti Internetistä. Lanttikokeessa 50 %:a suuremmat osuudet viittaavat epärehellisyyteen. Visassa ei ole selkeää rajaa rehellisyydelle mutta suuri pistemäärä vihjaa epärehellisyyttä.

Ilmoitettujen kruunujen osuuden lantinheittokokeessa ja kysymysten oikeiden vastausten keskiarvon maakohtainen sirontakuvio on ohessa. Testeissä alla oleta, että testien taustaoletukset pätevät ja sovella merkitsevyystasoa 0.05.

a) Jos lanttikoe ja visa mittaavat yhtälailla ja hyvin rehellisyyttä, miten havaintojen tulisi sijoittua sirontakuviossa?

b) Pearsonin korrelaatiokerroin, Spearmanin järjestyskorrelaatiokerroin ja Kendallin τ saavat aineistossa arvot 0.247 ($p = 0.187$), 0.261 ($p = 0.174$) ja 0.200 ($p = 0.149$). Suluissa on vastaavat p-arvot. Mitä testataan? Mitä päättelet? Perustele huolellisesti. (Vihje: Mitä ei päätellä?!)

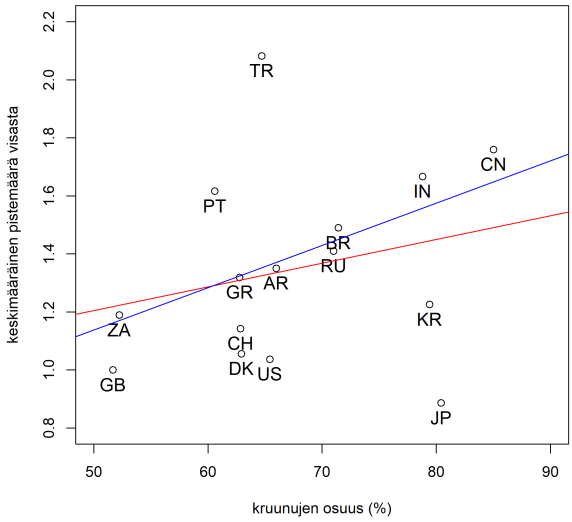
c) Sirontakuviossa pistää silmään Japani (JP). Lantinheittokokeessa kruunujen osuus on japanilaisilla toiseksi suurin. Visassa heidän pisteensä ovat pienimmät. Edellinen viittaa epärehellisyyteen; jälkimmäinen rehellisyyteen. Hugh-Jones pohtii, että uhkapelaaminen ylipäänsä (*most gambling*) on kiellettyä Japanissa, joten japanilaiset ovat saattaneet pitää lantinheittokoetta epäeettisenä ja ovat siksi valehdelleet siinä. Voi olla perusteltua poistaa Japani aineistosta.

Sirontakuviossa sininen viiva on koko aineistosta laskettu regressiosuora. Punainen viiva on regressiosuora, kun Japani on poistettu aineistosta. Suorat on estimoitu PNS-menetelmällä. Miksi sininen suora on punaisen suoran yläpuolella Japanin kohdalla? (Sanallinen selitys.)

d) Pearsonin korrelaatiokerroin, Spearmanin järjestyskorrelaatiokerroin ja Kendallin τ ovat 0.448 ($p = 0.054$), 0.495 ($p = 0.036$) ja 0.362 ($p = 0.035$) laskettuna aineistosta, josta on poistettu Japani. Mitä päättelet?

Turkki (TR) on aineistossa kaikista kauimpana regressiosuorista. Poistetaan myös se aineistosta. Pearsonin korrelaatiokerroin, Spearmanin järjestyskorrelaatiokerroin ja Kendallin τ ovat nyt 0.638 ($p = 0.010$), 0.577 ($p = 0.020$) ja 0.436 ($p = 0.019$). Mitä päättelet?

Onko perusteltua poistaa Japani ja Turkki aineistosta?


keskimääräinen pistemäärä visasta
kruunujen osuus (%)
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
50
60
70
80
90
TR
CN
IN
PT
BR
RU
GR
AR
ZA
KR
CH
DK
US
GB
JP