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

Uusintakoe 14.5.2021

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 (päälle voi saada lisäpisteitä). **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 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. Päteköön yhden selittäjän regressiomalli ilman vakiotermiä:

$$Y = \beta_1 x + \varepsilon.$$

Käytettävissä on havaintoparit $(x_1, y_1), \ldots, (x_n, y_n)$. Estimoidaan malli PNS-menetelmällä. Regressiokertoimen β_1 estimaatti ja sen keskivirheen estimaatti ovat

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \quad \text{ja} \quad s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^n x_i^2}}.$$

Yllä $s^2 = \sum_{i=1}^n (y_i - \hat{\beta}_1 x_i)^2/(n-1)$. Regression selitysaste on

$$\mathsf{R}^2 = 1 - \frac{\mathsf{JNS}}{\mathsf{KNS}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Kerrotaan y-vasteet vakiolla $a \neq 0$. Uusi aineisto on $(x_1, ay_1), \ldots, (x_n, ay_n)$.

a) Merkitään uudesta aineistosta laskettua β_1:n estimaattia $\hat{\beta}_{1*}$:lla. Osoita, että $\hat{\beta}_{1*} = a\hat{\beta}_1$.

b) Osoita, että uudesta aineistosta estimoidun mallin jäännös on $a\hat{\varepsilon}_i$, kun $\hat{\varepsilon}_i$ on alkuperäisen mallin jäännös ($i = 1, \ldots, n$). Osoita, että uudesta aineistosta estimoitu satunnaistermin varianssin estimaatti on $a^2 s^2$.

c) Merkitään uudesta aineistosta laskettua selitysastetta R_*^2:llä. Osoita, että $\mathsf{R}_*^2 = \mathsf{R}^2$. (Vihje: Vasteen keskiarvo $\bar{y}$ kertaantuu $a\bar{y}$:ksi vakiolla kertomisen jälkeen.)

d) Onko regressiokertoimen estimaatin t-arvo (H_0: $\beta_1 = 0$) aina sama alkuperäisestä ja uudesta aineistosta laskettuna? Perustele huolellisesti.

2. Gustaf Gabriel Hällström (1775–1844) oli 1838 perustetun Suomen tiedeseuran ensimmäinen puheenjohtaja. Hän julkaisi seuran lehdessä 1842 tutkimuksen maannoususta Suomessa.[1] Artikkelissa on taulukoitu muun muassa maannousua Hangossa Gäddtarmshamnissa, jota Hällströmin mukaan kutsutaan "nykyään" Carlshamniksi. Maannousutiedot perustuvat kenttämarsalkka, kreivi A. Ehrensvärdin 1754 Gäddtarmshamnin kallioon merkitsemään merenpinnan korkeuteen. Maannousua on sen jälkeen voitu mitata vertaamalla Ehrensvärdin merkintää merenpinnan korkeuteen. Ehrensvärd muistetaan parhaiten Suomenlinnan suunnittelijana ja rakentajana Augustin Ehrensvärdinä (1710–1772).

Maannousu on ollut Carlshamnissa 0.92, 0,92, 1.25 ja 1.67 Ruotsin jalkaa (29.69 cm) vuosina 1754–1796, 1754–1800, 1754–1821 ja 1754–1837. Selitetään

[1] G.G. Hällström (1842): Ny mätning af Åbo slotts höjd öfver hafsytan, jemte slutsatser om södra Finlands höjning öfver hafvet. *Acta Societatis Scientiarum Fennicæ*, I, 519–525.

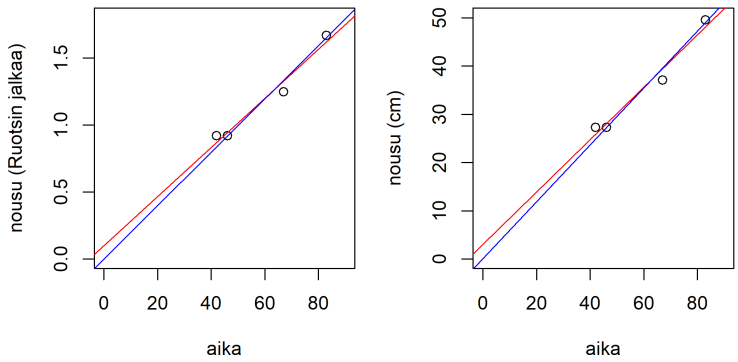
maannousua vuosilla Ehrensvärdin merkinnästä yhden selittäjän regressiomallilla vakiotermin kanssa ja ilman. Mallit estimoidaan PNS-menetelmällä. R-koodi ja osa sen palautteesta on alla. Regressiosuorat ovat vasemmanpuoleisessa kuvassa.

```
nousu <- c(0.92, 0.92, 1.25, 1.67)
aika <- c(1796-1754, 1800-1754, 1821-1754, 1837-1754)
aika # katsotaan aika-selittäjää
## [1] 42 46 67 83

summary(lm(nousu~aika)) # malli vakiotermin kanssa
## Residuals:
##        1        2        3        4
##  0.05065 -0.02264 -0.07742  0.04942
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.099799   0.141485   0.705   0.5537
## aika        0.018323   0.002291   7.998   0.0153 *
## Residual standard error: 0.07587 on 2 degrees of freedom
## Multiple R-squared:  0.9697,    Adjusted R-squared:  0.9545
## F-statistic: 63.97 on 1 and 2 DF,  p-value: 0.01527

summary(lm(nousu~0+aika)) # malli ilman vakiotermiä
## Residuals:
##         1         2         3         4
##  0.085065  0.005547 -0.081920  0.020009
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## aika 0.0198794  0.0005605   35.47 4.93e-05 ***
## Residual standard error: 0.06923 on 3 degrees of freedom
## Multiple R-squared:  0.9976,    Adjusted R-squared:  0.9968
## F-statistic:  1258 on 1 and 3 DF,  p-value: 4.928e-05
```

a) Voidaanko nollahypoteesi H_0: $\beta_1 = 0$ hylätä vakiotermin kanssa estimoidun mallin estimointitulosten perusteella 1 %:n merkitsevyystasolla (kaksisuuntainen testi)? Entä ilman vakiotermiä estimoidun mallin estimointitulosten perusteella? Sanoita testien tulos (mitä päättelet).

b) Selitä, miten on mahdollista, että R:n palautteen mukaan regressiossa vakiotermin kanssa satunnaistermin varianssin estimaatti on suurempi ja selitysaste "`Multiple R-squared`" pienempi kuin mallissa ilman vakiotermiä ($0.07587 > 0.06923$ ja $0.9697 < 0.9976$). (Vihje1: Raportoiko R tavanomaisen selitysasteen, jos mallissa ei ole vakiotermiä? Vihje2: Harjoitustehtävä 3.3–3.4.)

c) Muutetaan vasteen mittayksikkö Ruotsin jaloista senttimetreiksi kertomalla alkuperäiset havainnot vasteesta 29.69:llä. Muuttuvatko — ja miten, jos muuttuvat — regressiokertoimen estimaatti, jäännökset, selitysaste ja regressio-



kertoimen estimaatin t-arvo, jos mallissa on vakiotermi? Entä jos ei ole? Perustele huolellisesti. Regressiosuorat ovat oikeanpuoleisessa kuvassa. (Vihje1: Voit vedota tehtävän 1 tehtävänannon tietoihin tai estimoida mallit uudella vasteella ja toteamalla muuttuminen tai ei. Vihje2: Komento `nousu_cm <- 29.69*nousu`, ja toista analyysit uudella vasteella. Vihje3: Kuvioon vetoaminen ei ole riittävä perustelu!)

d) Vapaaehtoinen lisäpistekysymys: Onko malli vakion kanssa vai ilman perustellumpi tehtävän tilanteessa? Perustele huolellisesti.

3. Taloustieteessä tutkitaan ihmisten käyttäytymistä tyypillisesti olettaen heidän päätöksensä rationaalisiksi. Gary Becker (1930–2014) sai 1992 taloustieteen palkinnon Alfred Nobelin muistoksi ja oli taloustieteen ja sosiologian professori Chicagon yliopistossa. Becker sovelsi rationaalisuusoletusta mitä erilaisimpiin elämänalueisiin kuten puolison valintaan, lasten hankkimiseen, koulutukseen, rikollisuuteen, syrjintään ja elinsiirtoihin.[2]

Beckerin ja Murphyn (1988) mukaan ihmisen päätös ryhtyä käyttämään riippuvuutta aiheuttavia tuotteita (esim. tupakkaa, alkoholia tai huumeita) on ra-

[2] `https://en.wikipedia.org/wiki/Gary_Becker` (viitattu 14.5.2021).

tionaalinen.[3] Teorian (*rational addiction*) mukaan ihmiset huomioivat riippuvuutta aiheuttavan tuotteen käytön aloittamispäätöksen tehdessään tuotteen käytön aiheuttamat muutokset heidän tarpeisiinsa ja hyvinvointiinsa tulevaisuudessa. Teorialle on esitetty empiiristä tukea, mutta tehtyjä tutkimuksia on myös kritisoitu ekonometriselta kannalta.[4]

Gruben ja Köszegi (2001) testaavat teoriaa odottavien äitien tupakoimisaineistolla vuosilta 1989–1996.[5] He estimoivat yhtälön

$$\begin{array}{rcccccccl} y & = & - & 0.215x_1 & - & 0.344x_2 & - & 0.118x_3 & + \text{ muita tekijöitä.} \\ & & & (0.226) & & (0.142) & & (0.180) & \end{array}$$

Aineisto on kuukausiaineistoa Yhdysvaltojen osavaltioista. Osavaltio- ja kuukausikohtaisia havaintoja on 4 341 (havaintoja ei ollut saatavilla kaikista osavaltiosta kaikilta kuukausilta). Selitettävä muuttuja on äitien päivittäin tupakoimien savukkeiden keskimääräinen lukumäärä, x_1 on tupakkaveroaste, x_2 on kyseisessä osavaltiossa ja kuussa päätetty tuleva tupakkaveroaste, joka ei ole vielä astunut voimaan ja x_3 on tupakkaveroaste kahdella kuukaudella viivästettynä. Muut tekijät ovat muuttujia, joiden estimoituja kertoimia ei artikkelissa raportoida (vakiotermi sekä osavaltio- ja kuukausi-osoittimet). Kertoimien estimaattien estimoidut keskivirheet ovat suluissa. Tupakkaveron korotuksen vaikutuksen tupakointiin oletetaan mallissa jakautuvan kahdelle ajanjaksolle (x_1- ja x_3-selittäjät — tämä oletus ei liity Beckerin ja Murphyn teoriaan).

Teorian mukaan tupakan hinnan tulevan (veronnoususta aiheutuvan) korotuksen tulisi vähentää tupakointia jo aiemmin, koska rationaaliset ihmiset huomioivat, että riippuvuuden kustannus on suurempi tulevaisuudessa.

a) Onko estimoitu kerroin teorian mukainen? Onko estimoitu vaikutus tilastollisesti merkitsevä 1 %:n merkitsevyystasolla? Tee yksi- tai kaksisuuntainen testi sen mukaan, kumpi on tehtävän tilanteessa perustellumpi. Perustele valintasi huolellisesti.

b) Ovatko kaikki selittäjät tilastollisesti merkitseviä 1 %:n merkitsevyystasolla? Tee yksi- tai kaksisuuntaiset testit sen mukaan, kumpi on tehtävän tilanteessa perustellumpi. Perustele valintasi huolellisesti. Lisäpistekysymys: Entä jos selitettävä muuttuja olisi myytyjen eikä poltettujen savukkeiden määrä?!

[3] G.S. Becker ja K.M. Murphy (1988): A Theory of Rational Addiction. *Journal of Political Economy*, XCVI, 675–700.

[4] Esim. M. Christopher Auld ja P. Grootendorst (2004): An Empirical Analysis of Milk Addiction. *Journal of Health Economics*, 23, 1117–1133.

[5] J. Gruben ja B. Köszegi (2001): Is Addiction "Rational"? Theory and Evidence. *Quarterly Journal of Economics*, 116, 1261–1303. Havaintoja on painotettu osavaltion koon mukaan, mikä seikka sivuutetaan tehtävässä.

4. Kannabiksen käytön (x_1) ja psykoosien yhteyttä tutkittiin logistisen regressiomallin avulla Di Fortin ym:iden (2019) tutkimuksessa.[6] Siinä vertailtiin kannabista päivittäin käyttävien ($x_1 = 1$) ja kannabista käyttämättömien ($x_1 = 0$) todennäköisyyksiä saada psykoosi. Raportoitu suure oli ristisuhde (*odds ratio*) 3.2 (4.8, jos kannabis oli ollut vahvaa (*high-potency type*)). Yksinkertaisuuden vuoksi vastaa kysymyksiin alla olettaen, että olisi estimoitu yhden selittäjän logistinen regressio.

a) Selitä, miten ristisuhde määräytyy logistisessa regressiossa.

b) Tulkitse ristisuhde 3.2 sanoin tehtävän yhteydessä.

c) Lisäpistekysymys: Mikä on selittäjän x_1 kerroin logistisessa regressiossa, jos ristisuhde on 3.2?

5. Robert Parker on monien mielestä maailman vaikutusvaltaisin viiniasiantuntija.[7] Parker arvottaa viinejä pistemäärillä $50 - 100$ (100 on täydellinen viini.) Parker arvioi 158 Bordeaux-viiniä vuonna 2001. Viinit saivat keskimäärin 88.52 pistettä (otoskeskihajonta 2.42). Vuonna 2002 Parkerin arvioimia Bordeaux-viinejä oli 121 keskimääräisen pistemäärän ollessa 89.40 (otoskeskihajonta 2.19). Viinien hintojen jakaumat olivat molempina vuosina hyvin vinot oikealle (hintojen otoskeskihajonnat olivat lähellä viinien keskihintoja).

a) Viinin Parkerilta saaman pistemäärän ja hinnan välinen Spearmanin järjestyskorrelaatiokerroin ja Pearsonin korrelaatiokerroin olivat 0.74 ja 0.64 vuonna 2001. Mikä voisi olla tilastotieteellinen syy korrelaatiokertoimien erolle? (Vihje: Viittaavatko tiedot lineaariseen riippuvuuteen viinin laadun ja hinnan välillä vuonna 2001?)

b) Viinin pistemäärän ja hinnan välinen Spearmanin järjestyskorrelaatiokerroin oli 0.64 vuonna 2002. Testaa 0.5 %:n merkitsevyystasolla, erosiko Spearmanin järjestyskorrelaatiokerroin 0.64 nollasta vuonna 2002 (yksisuuntainen testi).

---

[6] Di Forti ym. (2019): The Contribution of Cannabis Use to Variation in the Incidence of Psychotic Disorder across Europe (EU-GEI): Multicentre Case-Control Study. *Lancet*. `DOI: https://doi.org/10.1016/S2215-0366(19)30048-3`.

[7] `https://en.wikipedia.org/wiki/Robert_Parker_(wine_critic)` (viitattu 13.5.2021). Tehtävän tiedot ovat artikkelista H. Hadj Ali, S. Lecocq ja M. Visser (2008): The Impact of Gurus: Parker Grades and *En Primeur* Wine Prices. *Economic Journal*, 118, F158–F173.