REGRESSIOANALYYSIN JATKOKURSSI, 5–10 OP (aine- ja syventävät opinnot). 15.9.–16.12.2011. Kirjallisuus: Russell Davidson ja James MacKinnon: Econometric Theory and Methods. Luennoi: yliopistonlehtori Pekka Pere.

Harjoitukset 2 (pe 23.9.)

1.

a) Osoita, että $\mathsf{tr}(\mathbf{AB}) = \mathsf{tr}(\mathbf{BA})$, jossa $\mathsf{tr}(\mathbf{X})$ on matriisin $\mathbf{X}$ jälki (trace) eli matriisin diagonaalialkioiden summa ja matriisit $\mathbf{A}$ ja $\mathbf{B}$ ovat $n \times k$- ja $k \times n$-matriiseja.

b) Osoita edellisen tuloksen avulla, että $\mathsf{tr}(\mathbf{P_X}) \equiv \mathsf{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = k$, kun $\mathbf{X}$ on $n \times k$ -matriisi. Tulkitse tulos. (Vihjeitä: Symmetrisen matriisin ominaisarvojen summa on matriisin jälki. Projektiomatriisin ominaisarvot ovat ykkösiä tai nollia. Symmetrisen matriisin aste on sen nollasta poikkeavien ominaisarvojen lukumäärä.)

c) Osoita edelleen, että $\mathsf{tr}(\mathbf{M_X}) \equiv \mathsf{tr}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = n - k$. Tulkitse tulos.

2. Merkitään $m \times n$ -matriisin $\mathbf{A}$ i:nnettä riviä $\mathbf{a}_i'$:lla $(1 \times n)$ ja j:nnettä saraketta $\mathbf{a}_{(j)}$:llä $(m \times 1)$, eli

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} \mathbf{a}_1' \\ \vdots \\ \mathbf{a}_m' \end{bmatrix} \\ &= [\mathbf{a}_{(1)} \cdots \mathbf{a}_{(n)}]. \end{aligned}$$

Todista *r,s-sääntö* eli väitteet alla.

a) Matriisin $\mathbf{A}$ kertominen vasemmalta $m \times m$ -diagonaalimatriisilla $\mathbf{D}_v = \lceil d_1 \dots d_m \rfloor$ tuottaa matriisin, jonka i. *rivi* on kerrottu d_i:llä:

$$\mathbf{D}_v\mathbf{A} = \begin{bmatrix} d_1\mathbf{a}_1' \\ \vdots \\ d_m\mathbf{a}_m' \end{bmatrix}.$$

b) Matriisin $\mathbf{A}$ kertominen oikealta $n \times n$ -diagonaalimatriisilla $\mathbf{D}_o = \lceil d_1 \dots d_n \rfloor$ tuottaa matriisin, jonka i. *sarake* on kerrottu d_i:llä:

$$\mathbf{A}\mathbf{D}_o = [d_1\mathbf{a}_{(1)} \cdots d_n\mathbf{a}_{(n)}].$$

3. Olkoon $\mathbf{X}$ täysiasteinen $n \times k$ -matriisi ($n \geq k$). Kerrotaan se oikealta singulaarisella $k \times k$ -matriisilla $\mathbf{A}$.

a) Todista, että matriisin $\mathbf{XA}$ sarakkeet ovat lineaarisesti riippuvia.

b) Todista, että $\mathsf{S}(\mathbf{XA}) \subset \mathsf{S}(\mathbf{X})$ (kirjan s:n 49 merkinnöin).

c) Oletetaan, että regression selittäjämatriisi on $\mathbf{X}$. Ovatko sovitteet ja residuaalit ylipäänsä samoja, jos selittäjämatriisi on $\mathbf{XA}$? Ovatko poikkeukset edelliseen vastaukseesi mahdollisia? Perustele.

4. Oletaan, että regressiomallissa on yksi selittäjä:

$$y_t = \beta_1 + \beta_2 x_t + u_t.$$

a) Osoita, että

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

ja

$$\hat{\beta}_2 = r \times s_y / s_x,$$

joissa "$-$" viittaa keskiarvoon, "$\hat{}$" pienimmän neliösumman (PNS) estimaattiin, r on muuttujien y ja x välinen otoskorrelaatio, $s_y^2 = (n-2)^{-1} \sum_{t=1}^{n} (y_t - \bar{y})^2$ ja s_x^2 on määritelty vastaavasti.

b) Osoita, että

$$\frac{\hat{y}_t - \bar{y}}{s_y} = r \frac{x_t - \bar{x}}{s_x},$$

jossa $\hat{y}_t$ on y_t:n PNS-sovite.

5. Jatketaan edellisen tehtävän regressiomallin pohtimista.

a) Onko ehdollinen odotusarvo $\mathsf{E}(y_t | x_t)$ ylipäänsä sama kaikille selitettäville havainnoille y_t?

b) Oheiset kuviot havainnollistavat regressiota odotusarvoa kohti (regression to the mean). Ensimmäinen kuvio on Francis Galtonin ensimmäinen regressio vuodelta 1877 ("vanhempi-pavut" ja "jälkipolvi-pavut") . Toinen kuvio on Galtonin havaitsema vastaava yhteys vanhempien pituuksien (mahdollisesti painotetun) keskiarvon (mid-parent[1]; jatkossa "keskipituus") ja heidän lastensa pituuden (children) välillä.[2] Mitä tarkoitetaan regressiolla odotusarvoa kohti?

c) Tapahtuuko regressiomallissa aina regressiota odotusarvoa kohti?

---

[1] Midparent-käsitteen määritelmä löytyy Wikipediasta (http://en.wikipedia.org/wiki/Midparent; viitattu 16.4.2011).

[2] Kuviot ovat kirjoista Stephen M. Stigler (1999): *Statistics on the Table: The History of Statistical Concepts and Methods* ja Michael O. Finkelstein (2009): *Basic Concepts of Probability and Statistics in the Law.* Springer (s. 128). Jälkimmäisessä kuviossa kahdesti esiintyvä "mild-parent" on painovirhe.

6. Ohessa on kolme kuviota kolmeen eri PNS:llä estimoituun regressiomalliin liittyen. Estimoidut mallit ja niiden sovitteet on esitetty kuvioissa. Vaikuttavatko mallit kelpoisilta (residuaalit riippumattomilta jne.)? Onko niissä mitään erityistä kommentointia herättävää? Perustele yksityiskohtaisesti. Alla on lisätietoja malleista — lähinnä kiinnostavuutta lisäämään.[3]

Ensimmäinen malli on estimoitu 1990-luvulla Yhdysvalloissa ympäristönsuojeluvirastossa (Environmental Protection Agency). Malli liittyy sähkön tuottamiseen hiilikäyttöisissä höyrykattiloissa. Höyrykattiloihin määrättiin asennettavaksi uutta typpioksideja polttavaa (low NO_x burner, LNB) teknologiaa. Ympäristövirasto argumentoi estimoimiensa regressiomallien perusteella, että tulevat kiristykset typpioksideja rajoittaviin säädöksiin tulisi sitoa höyrykattiloiden kokoon, koska regressiomallien mukaan isoissa hörykattiloissa voidaan päästä uuden teknologian avulla suhteellisesti suurempiin piennyksiin typpioksidipäästöissä kuin pienemmissä. Oheinen malli on yksi ympäristönsuojeluviraston estimoimista. Siinä selitettävä muuttuja on typpioksidien vähennys prosenteissa LNB-teknologian käyttöönoton jälkeen ja selittävänä muuttujana on höyrykattilan typpioksidipäästöt ennen LNB-teknologian käyttöönottoa.

Toinen malli on laadittu Yhdysvalloissa kuljetusturvahallinnossa (Federal Motor Carrier Safety Administration) vuonna 2005. Selitettävä muuttuja on rekan todennäköisyys joutua kuskin väsymyksestä johtuvaan kuolettavaan kolariin. Selittävä muuttuja on kuskin ajotunnit päivän aikana sekä tämän muuttujan toinen ja kolmas potenssi. Esimerkiksi 10-tuntisen ajon ajaneista kuskeista 4,4 % (496:sta rekasta 22 kolaroi) ja 11-tuntisen päivän ajaneista 9,6 % (94:stä rekasta 9 kolaroi) joutui kuolettavaan kolariin, ja näitä havaintoja on selitetty regressiomallilla.[4] Oikeanpuoleisin havainto 17-tuntista päivää koskien on laskettu keskiarvona yli 12-tuntista päivää ajaneiden kuskien kolaririskistä (olivat ajaneet keskimäärin 17-tuntisen päivän ja heistä n. 25 % oli joutunut kuolettavaan kolariin).

Kolmas malli on Vanhasen (2010). Mallissa selitetään etnistä väkivaltaa etnisellä heterogeenisuudella eri maissa. Etnisellä väkivallalla Vanhanen tarkoittaa etnisten ryhmien välistä väkivaltaa. Vanhanen mittasi sen määrää eri maissa vuosina 2003–2008 indeksillä, joka saa arvoja 1–5. Esimerkiksi Suomi sai arvon 1 (pienin mahdollinen määrä etnistä väkivaltaa) ja Somalia arvon 5 (suurin mahdollinen määrä etnistä väkivaltaa). Maan etnistä heterogeenisuutta Vanhanen mittasi (ilmeisesti) muuttujalla $(1/x - 1) \times 100$, jossa x on maan suurimman etnisen ryhmän osuus. Suomessa muuttuja sai arvon 7, koska Vanhasen mukaan suomenkielinen väestö on Suomen suurin etninen ryhmä, sen osuus väestöstämme on $0,93$ ja $(1/0,93 - 1) \times 100 \approx 7,53$ (ilmeisesti $0,93$ on pyöristetty alaspäin; suurempi osuus tuottaisi 7:ään pyöristyvän desimaaliluvun).

---

[3] Ensimmäinen ja toinen kuvio ovat kirjasta Michael O. Finkelstein (2009): *Basic Concepts of Probability and Statistics in the Law* (s:t 131–134). Kolmas kuvio on artikkelista Tatu Vanhanen (2010): Globaalit haasteet: demokratia, etninen väkivalta ja eriarvoisuus. *Suomen tilastoseuran vuosikirja 2010*, s:t 104–115.

[4] Prosentit on raportoitu Finkelsteinin (mt.) kirjassa näin. Osuudet viitannevat kolaririskiin jatkuvassa esimerkiksi 10-tuntisessa päiväajossa — eivät esimerkiksi yhden 10-tuntisen päivän ajon aikana.