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 6 (pe 21.10.)

1. Seuraavat kaksi puheluiden kysyntään liittyvää mallia on estimoitu PNS:llä Telecom Canadan neljännesvuosiaineistosta 1975–1985. (Mallit eivät ole realistisia; ne on laadittu havainnollistamaan tehtävän ideaa.) Ensimmäinen estimoitu malli on

$$\log(P_t) = -40.51 - 0.58\log(H_t/KHI_t) + 1.64 D_t \log(H_t/KHI_t) + \mathbf{x}_t'\hat{\boldsymbol{\gamma}}_{1t} + \hat{\varepsilon}_{1t}. \quad (1)$$

Yllä "log" on luonnollinen logaritmi, P on puheluiden määrä, H on puheluiden hinta, KHI on Kanadan kuluttajahintaindeksi, jonka arvo on normeerattu sadaksi vuonna 1981[5], D on dummy-muuttuja, joka saa arvon nolla vuoteen 1983 asti ja sen jälkeen arvon yksi, $\hat{\varepsilon}_{1t}$ on residuaali ja alaindeksi t viittaa ajankohtaan. Malliin kuuluu muita tehtävän kannalta epäoleellisia muuttujia ja parametreja, joita on merkitty $\mathbf{x}_t'\hat{\boldsymbol{\gamma}}_{1t}$:llä. Mallin mukaan puheluiden määrä on riippunut negatiivisesti puheluiden reaalisesta hinnasta (hinnasta suhteessa kuluttajahintaindeksiin) mutta vuoden 1983 jälkeen riippuvuutta osoittava kerroin muuttuu.

Jos KHI jaetaan sadalla — KHI:n tulkinnan kannalta täysin merkityksetön muutos — lasketaan uusi H/KHI^*-muuttuja ja estimoidaan malli (1) ("vapailla" parametreilla) uudestaan, niin saadaan

$$\log(P_t) = -40.32 - 0.50\log(H_t/KHI_t^*) - 0.02 D_t \log(H_t/KHI_t^*) + \mathbf{x}_t'\hat{\boldsymbol{\gamma}}_{1t}^* + \hat{\varepsilon}_{1t}^* \quad (1^*)$$

ilmeisin merkinnöin. Mallien (1) ja (1*) kertoimet (ja residuaalit) poikkeavat!

a) Selitä, miksi estimaatit ovat eri suuria yhtälöissä (1) ja (1*). (Vihje: Sijoita ensimmäiseen yhtälöön KHI:n paikalle $KHI^*=KHI/100$, muokkaa yhtälö yhtäpitäväksi mallin (1) kanssa ja pohdi, miten laatimasi yhtälö eroaa mallista (1*).)

b) Perustele, miksi β_i-kertoimien estimaatit eivät muuttuisi edellä kuvatunlaisen normeerauksen jälkeen, jos alunperin estimoitu malli olisi ollut

$$\log(P_t) = \alpha_1 + \alpha_2 D_t + \beta_1\log(H_t/KHI_t) + \beta_2 D_t\log(H_t/KHI_t) + x_t'\gamma_{2t} + \varepsilon_{2t}. \quad (2)$$

(Vihje: Sijoita $a \times KHI$ yhtälöön (2). Edellä a on mielivaltainen kerroin.)

[5] Kuluttajahintaindeksi mittaa kuluttajan kannalta tärkeiden tavaroiden ja palvelusten hintaa — pähkinänkuoressa yleistä hintatasoa. Kun yksittäisen tuotteen hinta eri ajankohtina jaetaan kuluttajahintaindeksin vastaavalla lukemalla, saadaan selville, miten tuotteen reaalinen hinta on kehittynyt.

2. Testaan F-testillä HT 3.2:n malleissa (1), (2) ja (3) kertoimien nolluutta (pois lukien vakiovektorinkertoimen). Poikkeavatko näistä malleista lasketut F-testisuureet nollahypoteeseille H_0: $\beta_2 = \ldots \beta_k = 0$ ja H_0^*: $\beta_2^* = \ldots \beta_k^* = 0$? Perustele.

3. Regressiomalli on

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}_1,$$

jossa $\mathbf{u} \sim \mathsf{IID}(\mathbf{0}, \sigma^2\mathbf{I})$, $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^{'} \quad \boldsymbol{\beta}_2^{'}]'$, $\boldsymbol{\beta}_1$ ja $\boldsymbol{\beta}_2$ ovat $k_1 \times 1$ ja $k_2 \times 1$ -vektoreita, $\mathbf{X}_1$ ja $\mathbf{X}_2$ ovat $n \times k_1$ ja -$n \times k_2$-matriiseja (k_1,$k_2 \geq 1$) ja $\mathbf{X}$ on täysiasteinen $n \times k$ -matriisi ($k_1 + k_2 = k$). Merkitään $r = k_2$. Testataan F-testillä nollahypoteesia H_0: $\boldsymbol{\beta}_2 = \mathbf{0}$ (r lineaarisesti riippumatonta rajoitusta).

Luennolla todistettiin, että nollahypoteesin pätiessä F-testisuure voidaan esittää muodossa

$$F_{\boldsymbol{\beta}_2} = \frac{\boldsymbol{\varepsilon}'\mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2^{'}\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2^{'}\mathbf{M}_1\boldsymbol{\varepsilon}/r}{\boldsymbol{\varepsilon}'\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}/(n-k)} \tag{3}$$

(kaava (4.34) ja että se noudattaa asymptoottisesti $\chi^2(r)/r$-jakaumaa, jos selittäjät ovat eksogeenisia. Yllä $\boldsymbol{\varepsilon} = \sigma^{-1}\mathbf{u}$, joten $\boldsymbol{\varepsilon} \sim \mathsf{IID}(\mathbf{0}, \mathbf{I})$. Todista, että sama asymptoottinen jakauma pätee, jos selittäjät ovat ennaltamääräytyneitä. Voit olettaa ilman eri todistusta, että $F_{\boldsymbol{\beta}_2}$:n nimittäjä konvergoi stokastisesti 1:seen (perustele intuitiivisesti). (Vihje 1: Normeeraa $F_{\boldsymbol{\beta}_2}$:n termit kuten kirjan kaavassa (4.58). Vihje 2: Todistus aivan samaan tapaan kuin vastaavalle tulokselle t-testisuureelle kirjan sivuilla 152–154.)

4. Tutkitaan regressiomallia

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}_1 \tag{4}$$

ilmeisin merkinnöin. Sitokoon k-vektorin $\boldsymbol{\beta}$ parametreja r lineaarisesti toisistaan riippumatonta rajoitusta $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. Tässä $\mathbf{R}$ on astetta r oleva $r \times k$ -matriisi ($r \leq k$; r on myös vektorin $\mathbf{r}$ pituus eli rajoitusten lukumäärä).[1] Tehtävä liittyy kirjan sivuihin 138 ja 141.

a) Perustele, että selittäjät (matriisin $\mathbf{X}$ sarakkeet) ja niitä vastaavat parametrit (vektorissa $\boldsymbol{\beta}$) voidaan järjestää aina niin, että lineaariset rajoitukset voidaan kirjoittaa muodossa

$$\mathbf{R}_1\boldsymbol{\beta}_1 + \mathbf{R}_2\boldsymbol{\beta}_2 = \mathbf{r}, \tag{5}$$

jossa $\mathbf{R}_1$ on *täysiasteinen* $r \times r$ -matriisi ja $\mathbf{R}_2$ on $r \times (k - r)$ -matriisi ($\boldsymbol{\beta}' = [\boldsymbol{\beta}_1' \quad \boldsymbol{\beta}_2']$).

[1] Merkinnät $\mathbf{r}$ ja r voivat sekoittua, jos $r = 1$. Selkeämpää olisi merkitä rajoitusten lukumääräksi k_1. Tässä noudatetaan kuitenkin kirjan merkintää rajoitusten lukumäärälle ($= r$; s:t 141 ja 173).

b) Perustele, että yhtälö

$$\mathbf{y}^* = \mathbf{Z}_2\boldsymbol{\beta}_2 + \mathbf{u}_0 \tag{6}$$

on yhtäpitävä yhtälön (4) kanssa, kun rajoitukset (5) pätevät. Yllä $\mathbf{y}^* = \mathbf{y} - \mathbf{X}_1\mathbf{R}_1^{-1}\mathbf{r}$ ja $\mathbf{Z}_2 = \mathbf{X}_2 - \mathbf{X}_1\mathbf{R}_1^{-1}\mathbf{R}_2$. ($\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2]$). (Vihje: Ratkaise $\boldsymbol{\beta}_1$ yhtälöstä (5).)

c) Lisätään regressioon (6) selittäjiksi $\mathbf{X}_1$:

$$\mathbf{y}^* = \mathbf{Z}_1\boldsymbol{\gamma}_1 + \mathbf{Z}_2\boldsymbol{\gamma}_2 + \mathbf{u}_1, \tag{7}$$

jossa $\mathbf{Z}_1 = \mathbf{X}_1$ ja $\boldsymbol{\gamma}_2 = \boldsymbol{\beta}_2$. Perustele, että $S(\mathbf{Z}_1,\mathbf{Z}_2) = S(\mathbf{X}_1,\mathbf{Z}_2) = S(\mathbf{X})$.

d) Esimerkkinä olkoon regressio

$$\mathbf{y} = \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \mathbf{u}_1$$

rajoituksella $\beta_1+\beta_2 = 1$ ($\mathbf{x}_1$ ja $\mathbf{x}_2$ ovat n-vektoreita). Selitä miten se ja regressiot

$$\mathbf{y} - \mathbf{x}_1 = \beta_2(\mathbf{x}_2 - \mathbf{x}_1) + \mathbf{u}_0$$

ja

$$\mathbf{y}^* = \gamma_1\mathbf{z}_1 + \gamma_2\mathbf{z}_2 + \mathbf{u}_1$$

liittyvät regressioihin edellä. Yllä $\mathbf{y}^* = \mathbf{y} - \mathbf{x}_1$, $\mathbf{z}_1 = \mathbf{x}_1$ ja $\mathbf{z}_2 = \mathbf{x}_2 - \mathbf{x}_1$.

e) Selitä, miten oheiset kuviot liittyvät edellisiin kohtiin.

f) Perustele edellisten kohtien avulla (vrt. kirjan s:t 138, 141 ja 173), että regressioiden (4) ja (7) residuaalivektorit ovat samoja ja että nollahypoteesin $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ pätevyyttä mallille (4) voidaan tutkia testaamalla rajoitusta $\boldsymbol{\gamma}_1 = \mathbf{0}$ F-testisuureella regressiomallissa (7). Edellä $\boldsymbol{\gamma}_1$ on r-vektori.