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 4 (pe 7.10.)

1. Olkoon lineaarisessa mallissa ainoa selittäjä vakio.

a) Laske $\mathbf{P_x}$ ja $\mathbf{P_X y}$ ja tulkitse ne sanoin.

b) Esitä numeerinen esimerkki, jossa R_u^2 muuttuu, kun $\mathbf{y}$:hyn lisätään vakio. (Muutokset nollasta mielivaltaisen lähelle ykköstä ovat mahdollisia!)

2. Tutkitaan kahden muuttujan regressiota ilman vakiota ($y_t = x_t\beta + jäännös$). Aineisto on seuraava (neljä havaintoa): $\mathbf{y}' = [7.4\ 6.7\ 7.7\ 8.2]$ ja $\mathbf{X}' = \mathbf{x}_1' = [0.04\ 0.24\ -0.60\ 0.32]$.

a) Laske PNS-estimaatti β:lle. Piirrä havainnot ja regressiosuora samaan kuvaan. Kulkeeko regressiosuora havaintojen lävitse? Tulkitse tilanne. (Miksi PNS-estimaatti β:lle on tehtävän tilanteessa sama, jos malliin sisällytetään vakio?!)

b) Laske R_u^2 ja R_c^2 (kirjan s. 74–75).

Voit tehdä laskut käsin tai käyttää Sinulle tuttua regressio-ohjelmaa. Jos käytät Survo-ohjelmistoa, niin seuraavat komentorivit ovat ehkä avuksi:

```
* GPLOT HT23,X,Y / XSCALE=-1(0.5)0.5 YSCALE=-2(2)10
*
* REGDIAG HT23,CUR+9 / CONSTANT=0
* DATA HT23,A,B,C,D
C   Y     X
D YYY XXXXX
A 7.4  0.04
* 6.7  0.24
* 7.7 -0.60
B 8.2  0.32
```

3. Tutkitaan neljän havaintoparin (x_i, y_i) aineistoa $(10.00, 8.65)$, $(8.65, 8.90)$, $(8.90, 8.01)$ ja $(8.01, 7.81)$. Sovita tuntemallasi regressio-ohjelmistolla aineistoon PNS-menetelmällä mallit

$$y_t = \beta x_t + u_t$$

ja

$$y_t = \beta_1 + \beta_2 x_t + u_t^*,$$

jossa u_t ja u_t^* ovat tarkemmin määrittelemättömiä jäännöksiä. Vertaa käyttämäsi ohjelmistojen raportoimia R^2-suureita edellisille regressioille ja kommentoi tuloksiasi. Pieneneekö käyttämäsi regressio-ohjelmiston tulostama R^2, kun malliin lisätään vakio?! Selitä tilanne.

Jos käytät Survo-ohjelmistoa, niin seuraavat komentorivit ovat ehkä avuksi:

```
* GPLOT HT52,X,Y
* XSCALE=0(3)12 YSCALE=0(3)12
*
* REGDIAG HT52,CUR+9 / CONSTANT=0
* DATA HT52,A,B,C,D
D  YYYY   XXXXX
C     Y       X
A  8.65   10.00
*  8.90    8.65
*  8.01    8.90
B  7.81    8.01
```

4. Pohditaan selitysasteen mittaamista ja mallinvalintaa, kun selitettävänä muuttujana voi olla alkuperäisen selitettävän muuttujan y_i epälineaarinen muunnos $f(y_i)$. Tällöin R_c^2:ten vertailu ei ole järkevää, koska selitettävät muuttujat ovat eri mittayksikössä. Eri mallien selityskykyä voidaan verrata esimerkiksi suureella

$$\mathsf{R}_1^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

jossa $\hat{y}_i$ on $f(y_i)$:n sovitetta vastaava arvo y_i:lle ja $\bar{y} = \sum_{i=1}^n y_i/n$. Jos estimoidun mallin selitettävä muuttuja on y_i ($f(y_i) = y_i$), niin $\mathsf{R}_1^2 = \mathsf{R}_\mathsf{c}^2$. Muulloin näin ei ylipäänsä ole.

Tutkitaan kuuden havaintoparin (x_i, y_i) aineistoa $(0, 0.5)$, $(1, 4)$, $(2, 6)$, $(3, 7)$, $(16, 12)$ ja $(20, 22)$. Sovita aineistoon PNS-menetelmällä lineaarinen malli

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

ja laske $\mathsf{R}_1^2 = \mathsf{R}_\mathsf{c}^2$. Korvaa regressiossa y_i-muuttujat $\log y_i$:llä ja sovita aineistoon PNS-menetelmällä lineaarinen malli

$$\log y_i = \beta_1^* + \beta_2^* x_i + u_i^*$$

ja laske R_c^2. Laske jälkimmäisen mallin sovitteita $\widehat{\log y_i}$ vastaavat sovitteet alkuperäiselle aineistolle eli $\exp(\widehat{\log y_i})$:t ja niiden avulla R_1^2. Vertaa ja kommentoi laskemiasi selitysasteita! Onko jompikumpi malleista selvästi parempi ja jos on niin kumpi?

Jos käytät Survo-ohjelmistoa, niin seuraavat komentorivit ovat ehkä avuksi:

```
*GPLOT HT53,X,Y / XSCALE=0(5)25 YSCALE=0(5)25
*
*REGDIAG HT53,CUR+14
*DATA HT53,A,B,C,D
D YYYY  XX  PP.PP
C    Y   X    FIT
A  0.5   0
*  4     1
*  6     2
*  7     3
* 12    16
B 22    20
*
* log(0.5)=-0.693
* exp(-0.693)=0.500
```

5. Tutkitaan kirjan mallia (3.11) eli niin sanottua 1. asteen autoregressiivistä (AR) mallia

$$\mathbf{y} = \beta_1 \boldsymbol{\iota} + \beta_2 \mathbf{y}_{-1} + \mathbf{u}.$$

Yllä vektori $\mathbf{y}$ koostuu aikajärjestyksessä olevista havainnoista y_t, $t = 1, \ldots, n$, siten, että ylimpänä on n. havainto, vektori $\mathbf{y}_{-1}$ on yhdellä ajanjaksolla viivästetty vektori $\mathbf{y}$ ($\mathbf{y}_{-1}$:n ylin komponentti on y_{n-1}) ja $\mathbf{u} \sim \mathsf{IID}(0,\sigma^2\mathbf{I})$. Kuten luennolla selitettiin, rekursiivisesti sijoittamalla ($y_t = \beta_1 + \beta_2 y_{t-1} + u_t$) saadaan

$$y_t = \beta_2^t y_0 + \beta_1 \sum_{i=0}^{t-1} \beta_2^i + \sum_{i=0}^{t-1} \beta_2^i u_{t-i}. \tag{1}$$

Kirjassa perusteltiin, että (ilmeisesti) $\mathsf{E}(\beta_2) \neq \beta_2$. Perustele, että vakiovektorin $\boldsymbol{\iota}$ kertoimen β_1 PNS-estimaattori on (ilmeisesti) myös harhainen eli että $\mathsf{E}(\beta_1) \neq \beta_1$.

6. Gauss-Markov -lauseen todistuksessa (kirjan s. 107) ei perusteltu viimeistä yhtäsuuruutta

$$\mathsf{Var}(\tilde{\boldsymbol{\beta}}) = \mathsf{Var}(\hat{\boldsymbol{\beta}}) + \mathsf{Var}(\mathbf{Cy}).$$

Todista yhtäsuuruus. (Vihje: Kovarianssimatriisin määritelmä (3.22) ja tulokset $\mathbf{Cy} = \mathbf{Cu}$ ja $\mathbf{CX} = \mathbf{0}$.)