SOSIAALITUTKIMUKSEN TILASTOLLISET MENETELMÄT 17.1.–28.2.2012. Jakso Regressioanalyysi ja trendimäisten muuttujien muunnokset. Luennoi: yliopistonlehtori Pekka Pere.

Logaritmin muutos ja suhteellinen muutos

Lähtökohta on approksimaatio

$$\log(1 + \delta) \approx \delta,$$

jossa $|\delta|$ ("deltan itseisarvo") on "pieni" ja $1 + \delta > 0$. (Approksimaatio on huono, jos ensimmäinen ehto ei päde.) Esimerkki ($\delta = 0,1$): $\log(1 + 0,1) = \log(1,1) = 0,0953 \approx 0,1$.

Palautetaan mieliin laskusääntö

$$\log(xy) = \log(x) + \log(y),$$

jossa x ja y ovat positiivisia lukuja.

Approksimaatiosta ja laskusäännöstä seuraa, että

$$\log[x(1 + \delta)] = \log(x) + \log(1 + \delta) \approx \log(x) + \delta$$

ja

$$\log[x(1 + \delta)] - \log(x) \approx \delta.$$

Logaritmin argumentin muuttuessa $100 \times \delta$ prosentilla muuttuu logaritmi noin δ:lla. Logaritmin muutokset kerrottuna sadalla ovat siten approksimatiivisia prosenttimuutoksia.

Esimerkki ($x = 100$ ja $\delta = 0,05$):

- $100(1 + 0,05) = 105$ — logaritmin argumentti suurenee 5 prosenttia.
- $\log[100(1 + 0,05)] - \log(100) = \log(105) - \log(100) = 0,0488$ — logaritmi suurenee noin $0,05$:llä.[7]

[7] Lisää aiheesta on artikkelissa Leo Törnqvist, Pentti Vartia ja Yrjö Vartia (1985): How Should Relative Changes Be Measured? *The American Statistician*, 39, 43–46.

SOSIAALITUTKIMUKSEN TILASTOLLISET MENETELMÄT 17.1.–28.2.2012. Jakso Regressioanalyysi ja trendimäisten muuttujien muunnokset. Luennoi: yliopistonlehtori Pekka Pere.

Muuttujien logaritmoinnista lineaarisessa regressiossa

Saakoot muuttujat Y ja X vain positiivisia arvoja. Pohditaan lineaarisen mallin tulkintaa, kun jompikumpi tai molemmat muuttujista on logaritmoitu.

1. Logaritmoidaan vain selittävä muuttuja:

   $$Y_i = \beta_0 + \beta_1 \log(X_i) + \varepsilon_i.$$

   Tulkinta: X:n muuttuessa 1 prosentin, Y muuttuu $\beta_1 \times 0,01$:n verran. (Perustelu: Jakson "Logaritmin muutos ja suhteellinen muutos" mukaan logaritmin argumentin muuttuessa $100 \times \delta$ prosentilla muuttuu logaritmi noin δ:lla. Tässä $\delta = 0,01$.) X:n (absoluuttisen) muutoksen vaikutus Y:hyn pienenee X:n kasvaessa.

2. Logaritmoidaan vain selitettävä muuttuja:

   $$\log(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

   Tulkinta: X:n muuttuessa yksikön verran Y muuttuu $100 \times \beta_1$ prosenttia. (Perustelu kuten edellä.) X:n muutoksen vaikutus Y:hyn (absoluuttisesti) suurenee X:n kasvaessa.

3. Logaritmoidaan sekä selittävä että selitettävä muuttuja:

   $$\log(Y_i) = \beta_0 + \beta_1 \log(X_i) + \varepsilon_i.$$

   Tulkinta: X:n muuttuessa 1 prosentin, Y muuttuu β_1 prosenttia. β_1 on Y:n jousto X:n suhteen.[8]

Huom! Selitysastetta R^2 tilanteissa 2) ja 3) vertailemalla voidaan pohtia, tulisiko selittävä muuttuja logaritmoida vai ei. Vertailua ei voi laajentaa kattamaan tilanteen 1) mallia, sillä siinä selitettävä muuttuja on eri.[9] Selitettävän muuttujan logaritmointipäätös täytyy tehdä muilla perusteilla kuin selitysastetta (esimerkiksi tilanteissa 1) ja 3)) vertailemalla.

[8] Y:n jousto X:n suhteen on

$$\frac{Y\text{:n muutos prosenteissa}}{X\text{:n muutos prosenteissa}}.$$

[9] $\mathsf{R}^2 = 1 - \mathsf{SSE}/\mathsf{SST}$. Selitysasteiden vertailu perustuu SSE:n vaihteluun mallien välillä SST:n ollessa kiinteä. Jos selitettävä logaritmoidaan, muuttuu SST, eikä vertailu mallien välillä ole mahdollista.

SOSIAALITUTKIMUKSEN TILASTOLLISET MENETELMÄT 17.1.–28.2.2012. Jakso Regressioanalyysi ja trendimäisten muuttujien muunnokset. Luennoi: yliopistonlehtori Pekka Pere.

Logaritmoitujen muuttujien differenssointi

Monesti muuttujan arvo liittyy tiettyyn ajankohtaan (vuoteen tai kuukauteen jne.). Tällöin puhutaan aikasarjasta. Monet tällaiset muuttujat ovat trendimäisiä eli esimerkiksi kasvavat pitkällä aikavälillä. Useasti tällaiset muuttujat sekä logaritmoidaan että differenssoidaan.

Merkitään muuttujan ajankohdalla t (esim. $t = 1990, 1991, \ldots, 2011$) saamaa arvoa x_t:llä ja sen logaritmia $\log x_t$:llä. Uusi muuttuja

$$\log x_t - \log x_{t-1}$$

voi olla mielenkiintoisempi kuin alkuperäinen muuttuja. Esimerkiksi jos Y_t on hintataso, niin $100 \times (\log x_t - \log x_{t-1})$ on oleellisesti hintatason suhteellinen muutos prosenteissa (edellisen sivun perusteella) eli inflaatio.

Regressiomalleissa $\log x_t - \log x_{t-1}$ muotoa oleva selittäjä ei ole harvinainen. Siihen liittyvällä regressiokertoimella on luonteva tulkinta eritoten, kun myös selitettävä muuttuja on logaritmoitu: Selitettävä muuttuu (approksimatiivisesti) prosenteissa β_1 kertaa x_t:n prosenttimuutoksen $(\log x_t - \log x_{t-1})$ verran (x_t:n regressiokertoimen ollessa β_1).