Harjoitusten 4 vastaukset

1. Merkitään havaintojen lukumäärää n:llä ja ykkösistä koostuvaa $n \times 1$ -vektoria ι:lla.

a)

$$\begin{aligned} \mathbf{P_x} &\overset{\text{1 selittäjä}}{=} \mathbf{x}_1(\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1' \\ &\overset{\text{sel.=vakio}}{=} \iota(\iota'\iota)^{-1}\iota' \\ &= n^{-1}\iota\iota' \\ &= n^{-1}\{1\}_{ij}, \end{aligned}$$

jossa $i, j = 1, \ldots, n$ ja $\{1\}_{ij}$ on matriisi, jonka jokainen elementti on 1. $\mathbf{P_x}$ on $n \times n$ -matriisi, joka on astetta yksi, koska selittäjiä on vain yksi. (HT 2.1 b) yllä.) Aliavaruus, johon $\mathbf{P_X}$ projisoi, on vakiovektorin ι virittämä.

$$\begin{aligned} \mathbf{P_x y} &= n^{-1}\{1\}_{ij}\mathbf{y} \\ &= [\bar{y} \ldots \bar{y}]'. \end{aligned}$$

Toinen muotoilu:

$$\iota\hat{\beta}_1 = \iota\bar{y}.$$

Vektorin $\mathbf{y}$ elementtien paras ennuste on keskiarvo $\bar{y}$, kun mallissa on vain vakio.

b) Valtiot. yo. Antti Luukkosen ytimekäs vastaus, jota olen terävöittänyt:

$$R_u^2 \overset{\text{yltä}}{=} \frac{n\bar{y}^2}{\sum_{i=1}^n y_i^2}.$$

i) Valitaan $\mathbf{y} = [-1 \quad 1]'$, jolloin $n = 2$, $\bar{y} = 0$ ja $R_u^2 = 0$.

ii) Lisätään $\mathbf{y}$:n komponentteihin 1. Saadaan vektori ja suureet $\mathbf{y} = [0 \quad 2]'$, $\bar{y} = 1$ ja $R_u^2 = 2/4 = 0.5$.

iii) Lisätään $\mathbf{y}$:n komponentteihin 100. Saadaan vektori ja suureet $\mathbf{y} = [99 \quad 101]'$, $\bar{y} = 100$ ja $R_u^2 = 20000/20002 \approx 0.999$.

Esimerkki on helppo tulkita geometrisesti. $R_u^2 = \cos^2\phi$, jossa ϕ on vekto reiden $\mathbf{y}$ ja vakiovektorin $[1 \quad 1]'$ välinen kulma.

1. Vektorit $[-1 \quad 1]'$ ja $[1 \quad 1]'$ ovat ortogonaalisia ($\phi = \pi/2$), joten $R_u^2 = [(\cos\phi)|_{\phi=\pi/2}]^2 = 0$.

2. Vektoreiden $[0 \quad 2]'$ ja $[1 \quad 1]'$ välinen kulma on $\pi/4$, joten $R_u^2 = [(\cos\phi)|_{\phi=\pi/4}]^2 = 0.5$.

3. Vektoreiden $[99 \quad 101]'$ ja $[1 \quad 1]'$ välinen kulma on lähes 0, joten $R_u^2 = [(\cos\phi)|_{\phi\approx 0}]^2 \approx 1$.

Havaintovektori kiertää (y_1, y_2)-koordinaatistossa "luoteesta" lähes vakiovektorin $[1 \quad 1]'$ osoittamaan "koilliseen" edettäessä 1. tilanteesta 3. tilanteeseen.

Vaihtoehtoinen esimerkki: Valitaan $n = 4$ ja $\mathbf{y} = [1\ 2\ 3\ 4]'$. Tällöin $\bar{y} = 2.5$ ja $\sum_{i=1}^{n} y_i^2 = 30$. Lasketaan R_u^2:

$$\begin{aligned} R_u^2 &\overset{\text{yltä}}{=} \frac{n\bar{y}^2}{\sum_{i=1}^{n} y_i^2} \\ &= \frac{4 \times (2.5)^2}{30} \\ &\approx 0.8333. \end{aligned}$$

Lisätään $\mathbf{y}$:n elementteihin 100 ja merkitään uutta vektoria $\mathbf{y}^*$:llä: $\mathbf{y}^* = [101\ 102\ 103\ 104]'$. Nyt $\bar{y}^* = 102.5$, $\sum_{i=1}^{n} (y_i^*)^2 = 42030$ (ilmeisin merkinnöin) ja R_u^2 on

$$\begin{aligned} R_u^2 &\overset{\text{yltä}}{=} \frac{4 \times (102.5)^2}{42030} \\ &\approx 0.9999. \end{aligned}$$

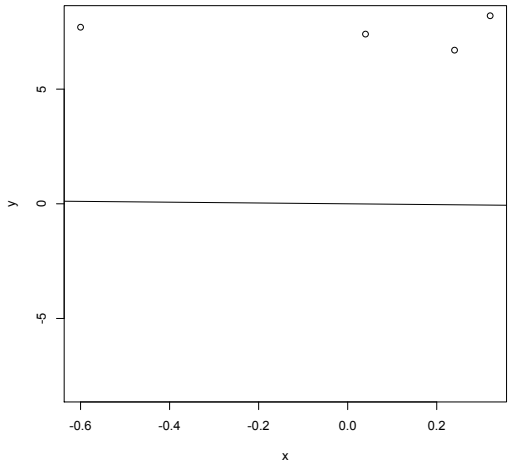
2

a) PNS-estimaatti on (kahden muuttujan regressio; merkinnät ilmeiset):

$$\sum_{i=1}^{4} y_i x_i / \sum_{i=1}^{4} x_i^2 = -0.18.$$

```
Regressiotulokset:
estimaatti keskivirhe t-arvo
-0.18 12.0 -0.01
Selitett\U{e4}v\U{e4}n varianssi: 0.40 (vap.asteita 3)
J\U{e4}\U{e4}nn\U{f6}svarianssi:=75.0 (vap.asteita 3)
R^{2}=0.001
```

Havainnot näyttävät tältä:



Kuvaan on mys piirretty sovite. (Esim. kolmannen ja neljännen havainnon kohdalla regressiosuoran arvot ovat $-0.27 * (-0.60) = 0.16$ ja $-0.27 * (0.32) = -0.09$.)

Estimaatti olisi niin ikään -0.18, jos mallissa olisi vakio!

Selitys I: Vakiovektori ι ja selittäjä $\mathbf{x}$ ovat ortogonaalisia:

$$\iota'\mathbf{x} = \sum_{i=1}^{4} x_i = 0.04 + 0.24 - 0.60 + 0.32 = 0.$$

Aiempiin selittäjiin (tässä vain $\mathbf{x}$) nähden ortogonaalisen selittäjän lisääminen malliin ei muuta aiempien selittäjien estimoituja kertoimia (kirjan s. 66). Näin ollen vakion lisääminen malliin ei muuta alkuperäisen regressiosuoran kertoimen estimaattia. Vakion lisääminen siirtää regressiosuoraa (ylöspäin) muuttamatta sen kulmakerrointa.

Selitys II: PNS-estimaatti voidaa laskea keskistetyistä muuttujista. Tehtävän tilanteessa selittävien muuttujien keskiarvo on nolla, joten PNS estimaatti olisi (ilmeisin merkinnöin)

$$\frac{\sum_{i=1}^4 (y_i - \bar{y})x_i}{\sum_{i=1}^4 x_i^2} = \frac{\sum_{i=1}^4 y_i x_i}{\sum_{i=1}^4 x_i^2} - \bar{y}\frac{\sum_{i=1}^4 x_i}{\sum_{i=1}^4 x_i^2} = \frac{\sum_{i=1}^4 y_i x_i}{\sum_{i=1}^4 x_i^2} \overset{\text{yltä}}{=} -0.18.$$

b)

$$\begin{aligned}
R_u^2 &= \frac{\|\mathbf{P_X y}\|^2}{\|\mathbf{y}\|^2} \\
&\overset{\text{vain 1 selittäjä}}{=} \frac{\left\|\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}\right\|^2}{\|\mathbf{y}\|^2} \\
&\approx \frac{\left\|[0.04\ 0.24\ -0.60\ 0.32]'(0.5126)^{-1}(-0.092)\right\|^2}{[7.4\ 6.7\ 7.7\ 8.2][7.4\ 6.7\ 7.7\ 8.2]'} \\
&\approx \frac{\|[0.04\ 0.24\ -0.60\ 0.32]'(-0.179)\|^2}{226.18} \\
&\approx \frac{0.016}{226.18} \\
&\approx 0.00007.
\end{aligned}$$

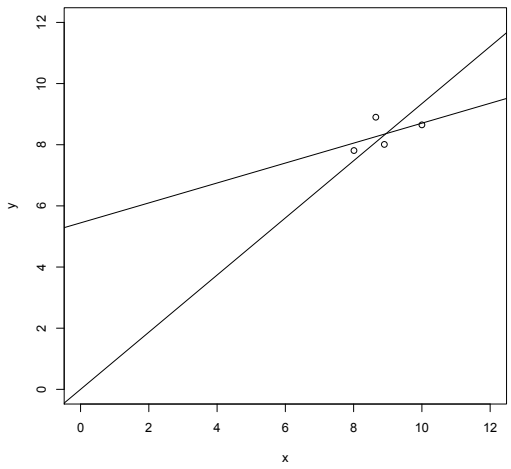
Lasketaan R_c^2 (kaava $\|\mathbf{P_X M}_\iota \mathbf{y}\|^2 / \|\mathbf{M}_\iota \mathbf{y}\|^2$ ei päde, koska mallissa ei ole vakiota):

$$\begin{aligned}
\mathsf{R}_\mathsf{c}^2 &= 1 - \frac{\|\mathbf{M_X y}\|^2}{\|\mathbf{M}_\iota \mathbf{y}\|^2} \\
&\overset{\text{vain 1 selittäjä}}{=} 1 - \frac{\|(\mathbf{I}_4 - \mathbf{P_X})\mathbf{y}\|^2}{\|(\mathbf{I}_4 - 4^{-1}\iota\iota')\mathbf{y}\|^2} \\
&\overset{\bar{y}=7.5}{\approx} 1 - \frac{\|\mathbf{y} - \mathbf{P_X y}\|^2}{1.18} \\
&\overset{\mathbf{P_X}=\mathbf{P_X}\mathbf{P_X}}{=} 1 - \frac{\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{P_X y} - y'P_X y + y'P_X y}{1.18} \\
&\overset{\text{yltä}}{\approx} 1 - \frac{226.18 - 0.016}{1.18} \\
&\approx -190.66.
\end{aligned}$$

Tälle aineistolle R_c^2 on negatiivinen. Intuitiivinen selitys: $\|\mathbf{M_X y}\|^2$ ja $\|\mathbf{M}_\iota \mathbf{y}\|^2$ovat residuaalien neliösummat, kun $\mathbf{y}$:tä selitetään $\mathbf{x}_1$:llä tai vakiolla. Jos vakio on selityskyvyltään parempi, niin $\mathsf{R}_\mathsf{c}^2 = 1 - \|\mathbf{M_X y}\|^2 / \|\mathbf{M}_\iota \mathbf{y}\|^2$ on negatiivinen. Johtopäätös: Jos mallissa ei ole vakiota, niin R_c^2 ei ole tulkittavissa tavanomaiseen tapaan.

3. Seuraavat laskut on laskettu SURVO MM:llä (versio 1.25). (SURVO 98 laskee samat tulokset.)

Katsotaan aineistoa. Kuvaan on piirretty sovitteet molemmista regressiomalleista.



Regressiossa ilman vakiota R^2 on 0.99:

```
Selittj Kerroin Kertoimen SD         t
        X    0.93        0.038     24.85
Y:n otosvarianssi: 0.27 (vapausasteita 3)
Estimoitu jnnsvarianssi: 0.45 (vapausasteita 3)
R^2=0.99
```

Estimoitu jäännösvarianssi on suurempi kuin selitettävän otosvarianssi! Tällöin regressiosuora on huonompi sovite kuin vakio ja R_c^2 on negatiivinen. Raportoitu R^2 on positiivinen, joten käytetty ohjelmisto laskee jonkin toisen suureen kuin R_c^2:n.

Kun malliin lisätään vakio, näyttää selitysaste pienenevän 0.29:ään(!):

```
Selittj Kerroin  Kertoimen SD     t
    Vakio    5.44          3.34  1.63
        X    0.33          0.37  0.87
Y:n otosvarianssi: 0.27 (vapausasteita 3)
Estimoitu jnnsvarianssi: 0.29 (vapausasteita 2)
R^2=0.27
```

Vakion lisääminen malliin ei voi huonontaa mallin sovitetta, joten sovitteen "huonontumisen" täytyy johtua käytetystä kaavasta R^2:lle. Estimoitu jäännösvarianssi on tämänkin mallin kohdalla suurempi kuin selitettävän otosvarianssi. Se johtuu kuitenkin edellisen vapausastekorjauksesta. Ilman vapausastekorjausta estimoitu jäännösvarianssi on

$$(4 - 2) * 0.29/3 = 0.19,$$

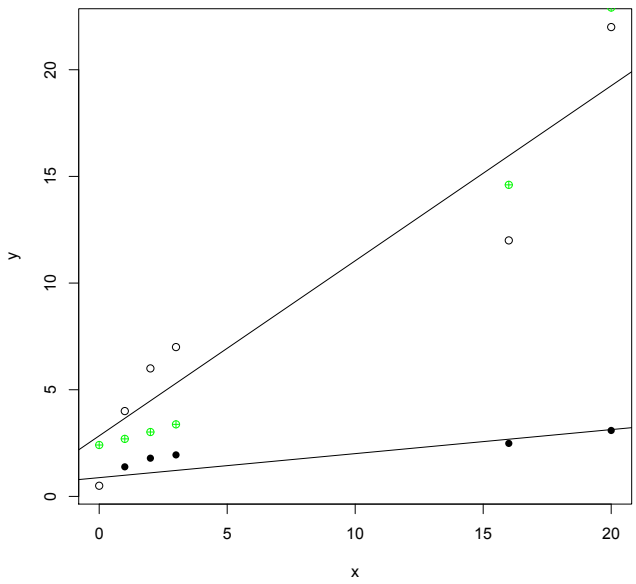
joka on pienempi kuin Y:n otosvarianssi. Näin täytyykin olla, koska vakion ja regressiosuoran täytyy tuottaa vähintään yhtä hyvä sovite kuin pelkän vakion.

Harjoitustehtävä on empiirisesti relevantti. Tiedän tapauksen, jossa empiirikot ilakoivat, kun olivat saaneet lähellä yhtä olevan selitysasteen mallille, jossa ei ole vakiota.

Jäännösvarianssin suuruus on luotettavampi kriteeri mallin hyvyydelle kuin tilasto-ohjelmiston raportoima R^2. Esimerkin varoitus on R^2:n laskutapaa yleisempi eli että *tilasto-ohjelmistoihin ei pidä luottaa sokeasti*! (Ks. esim. B.D. McCullough ja H.D. Vinod (1999): The Numerical Reliability of Econometric Software, *J. of Economic Literature*, XXXVII, 633–65.)

4. Seuraavat laskut on laskettu SURVO MM:llä (versio 1.25). (SURVO 98 laskee samat tulokset.)

Katsotaan aineistoa:



Kuvassa ylempi viiva on sovite regressiosta (1). Alempi viiva on sovite regressiosta jossa selitetn $log(y_i)$:tä. Valkoiset pallot kuvaavat havaintoja y_i ja mustat pallot havaintoja $log(y_i)$. Vihreät ristit ovat regressiosta (2) lasketut ennusteet y_i:lle, $exp\left(\hat{logy_i}\right)$

R_c^2 on 0.881 regressiossa (1) ja 0.565 regressiossa (2). Regressio (2) näyttää R_c^2:n perusteella selvästi huonommalta mallillta. R_1^2:n mukaan mallit ovat kuitenkin lähes yhtä hyviä, kun vertaillaan mallien kykyä selittää alkuperäistä aineistoa: R_1^2 on 0.881 regressiolle (1) (jolloin (R_1^2 =R_c^2) ja 0.878 regressiolle (2)!

Aineisto on Scottin ja Wildin (1991) artikkelista *American Statistician*'issä. He esittävät lisäksi esimerkin, jossa R_c^2 on hieman parempi regressiolle (2) (R_c^2 = 0.94) kuin regressiolle (1) ($R_c^2 = 0.92$), mutta regressiosta (2) lasketut sovitteet $\exp(\hat{\log y_i})$ ovat niin huonot, että R_1^2 on negatiivinen!

Huomautus: Ei ole selvää, että tehtävänkaltainen yksinkertainen tapa siirtyä $\log y$-asteikon sovitteista y-asteikon sovitteisiin on paras mahdollinen. Ks. esim. pohdinta ja viitteet kirjassa T.C. Mills: (1987): Time Series Techniques for Economists, s:t 337–339. Davidson ja MacKinnon käsittelevät samaa aihepiiriä 1993-kirjan jaksossa 14.3. Pääviesti kuitenkin säilyy: R_c^2:ten vertailu mallien välillä, joissa on eri selitettävä, ei ole järkevää.

5. FWL-lauseen (s. 68) mukaan

$$\begin{aligned}
\hat{\beta}_1 &= (\iota'\mathbf{M}_{\mathbf{y}_1}\iota)^{-1}\iota'\mathbf{M}_{\mathbf{y}_1}\mathbf{y} \\
&\overset{(3.11)}{=} (\iota'\mathbf{M}_{\mathbf{y}_1}\iota)^{-1}\iota'\mathbf{M}_{\mathbf{y}_1}(\beta_1\iota + \beta_2\mathbf{y}_{-1} + \mathbf{u}) \\
&\overset{\mathbf{M}_{\mathbf{y}_1}\mathbf{y}_1=\mathbf{0}}{=} \beta_1 + (\iota'\mathbf{M}_{\mathbf{y}_1}\iota)^{-1}\iota'\mathbf{M}_{\mathbf{y}_1}\mathbf{u}.
\end{aligned}$$

Näin ollen

$$\begin{aligned}
\mathsf{E}(\hat{\beta}_1) &= \mathsf{E}[\beta_1 + (\iota'\mathbf{M}_{\mathbf{y}_1}\iota)^{-1}\iota'\mathbf{M}_{\mathbf{y}_1}\mathbf{u}] \\
&= \beta_1 + \mathsf{E}[(\iota'\mathbf{M}_{\mathbf{y}_1}\iota)^{-1}\iota'\mathbf{M}_{\mathbf{y}_1}\mathbf{u}].
\end{aligned}$$

Odotusarvo-operaattoria ei voi viedä lausekkeen $\iota'\mathbf{M}_{\mathbf{y}_1}\iota)^{-1}\iota'\mathbf{M}_{\mathbf{y}_1}\mathbf{u}$ sisälle vektorin $\mathbf{u}$ eteen, koska vektorit $\mathbf{u}$ ja $\mathbf{y}_1$ ja siten suureet $\mathbf{u}$ ja $\mathbf{M}_{\mathbf{y}_1}$ eivät ole riippumattomia (tehtävän kaava (1)). Näin ollen (ilmeisesti) $\mathsf{E}(\hat{\beta}_1) \neq \hat{\beta}_1$. (Harhaisuuden voi todentaa esimerkiksi simulointikokeella.)

6. Luennoilla perusteltiin yhtäsuuruus

$$\mathsf{Var}(\tilde{\beta}) = \mathsf{Var}(\hat{\beta} + \mathbf{C}\mathbf{y}).$$

Jatketaan kirjan sivun 107 todistusta tästä. Vihjeen mukaan $\mathbf{C}\mathbf{y} = \mathbf{C}\mathbf{u}$. Siitä ja PNS-estimaattorin harhattomuudesta (tehdyin oletuksin) seuraa, että $\mathsf{E}(\mathbf{C}\mathbf{y}) = \mathsf{E}(\mathbf{C}\mathbf{u}) = \mathbf{C}\mathsf{E}(\mathbf{u}) = \mathbf{0}$ ja $\mathsf{E}(\hat{\beta} + \mathbf{C}\mathbf{y}) = \beta + \mathbf{0} = \beta$. Näin ollen

$$\begin{aligned}
\mathsf{Var}(\hat{\beta} + \mathbf{C}\mathbf{y}) &= \mathsf{E}(\hat{\beta} + \mathbf{C}\mathbf{y} - \beta)(\hat{\beta} + \mathbf{C}\mathbf{y} - \beta)' \\
&= \mathsf{E}[(\hat{\beta} - \beta) + \mathbf{C}\mathbf{y}][(\hat{\beta} - \beta) + \mathbf{C}\mathbf{y}]' \\
&= \mathsf{E}(\hat{\beta} - \beta)(\hat{\beta} - \beta)' + \mathsf{E}(\hat{\beta} - \beta)(\mathbf{C}\mathbf{y})' + \mathsf{E}(\mathbf{C}\mathbf{y})(\hat{\beta} - \beta)' + \mathsf{E}(\mathbf{C}\mathbf{y})(\mathbf{C}\mathbf{y})'
\end{aligned}$$

Yhtälöstä (3.05) $\hat{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$, vihjeestä $\mathbf{C}\mathbf{y} = \mathbf{C}\mathbf{u}$, eksogeenisuusoletuksesta ja kirjassa perustellusta harhattomuusehdosta $\mathbf{C}\mathbf{X} = \mathbf{0}$ seuraa, että

$$\begin{aligned}
\mathsf{E}(\hat{\beta} - \beta)(\mathbf{C}\mathbf{y})' &= \mathsf{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}](\mathbf{C}\mathbf{u})' \\
&= \mathsf{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}](\mathbf{u}'\mathbf{C}') \\
&\overset{\text{eksog.}}{=} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathsf{E}(\mathbf{u}\mathbf{u}')\mathbf{C}' \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}' \\
&\overset{\mathbf{C}\mathbf{X}=\mathbf{0}}{=} \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{0}'_{k\times k} \\
&= \mathbf{0}_{k\times k}.
\end{aligned}$$

Yllä odotusarvot tulee tulkita ehdollisiksi odotusarvoiksi (kuten kirjassakin). Sijoitetaan se edelliseen yhtälöketjuun, jolloin saadaan kysytty yhtäsuuruus:

$$\begin{aligned}
\mathsf{Var}(\hat{\beta} + \mathbf{C}\mathbf{y}) &= \mathsf{E}(\hat{\beta} - \beta)(\hat{\beta} - \beta)' + \mathbf{0}_{k\times k} + \mathbf{0}_{k\times k} + \mathsf{E}(\mathbf{C}\mathbf{y} - \mathbf{0})(\mathbf{C}\mathbf{y} - \mathbf{0})' \\
&= \mathsf{Var}(\hat{\beta}) + \mathsf{Var}(\mathbf{C}\mathbf{y}).
\end{aligned}$$