TAMPEREEN YLIOPISTO
Tilastollisen mallintamisen harjoitustyö

Teemu Kivioja ja Mika Helminen

Epätasapainoisen koeasetelman analyysi

Worksheet 5

Matematiikan, tilastotieteen ja filosofian laitos
Tilastotiede
Lokakuu 2000

1 Kaksisuuntaisen epätasapainoisen koeasetelman analyysi

1.1 Lineaarinen malli

Aineistossa "bookpr.txt"oli kymmenen kirjan hinta viidessä maassa. Tutkimme kuinka kirjan hinta riippuu maa- ja kirjailijamuuttujista. Keskitymme lähinnä maamuuttujan tarkasteluun. Tutkimustuloksia saadaan pääasiallisesti kaksisuuntaisen varianssianalyysin avulla, jossa tarkasti ottaen yhtäaikaisesti tutkitaan muiden maa- ja kirjailijamuuttujien vaikutusta mallin "vakioparametriin"jonka lähtökohdaksi on otettu aineiston ensimmäinen kirjailija sekä maa. Kyseiset muuttujat ovat "Britannia"(UK) sekä "S.Hawking, A brief history of time". Aineistossa oli puuttuvia arvoja, jolloin poistamalla nämä puuttuvat arvot saatiin epätasapainoinen asetelma, minkä merkityksen selvitämme luvussa kaksi. "NA"tarkoittaa aineistossa puuttuvaa arvoa.

Sarakkeet ovat UK, Germany, France, US ja Austria

```
14.99  12.68   9.00  11.00  15.95 S.Hawking,"A brief history of time"
14.95  17.53  13.60  13.35  15.95 U.Eco,"Foucault's Pendulum"
12.95  14.01  11.60  11.60  13.60 J.Le Carre,"The Russia House"
14.95  12.00   8.45   NA      NA  J.Archer,"Kane & Abel"
12.95  15.90  15.10   NA    16.00 S.Rushdie,"The Satanic Verses"
12.95  13.40  12.10  11.00  13.60 J.Barnes"History of the world in ..."
17.95  30.01   NA    14.50  22.80 R.Ellman,"Oscar Wilde"
13.99   NA     NA    12.50  13.60 J.Updike,"Rabbit at Rest"
 9.95  10.50   NA     9.85   NA   P.Suskind,"Perfume"
 7.95   9.85   5.65   6.95   NA   M.Duras,"The Lover"
```

Taulukko 1: Aineisto

Aluksi sovitimme lineaarista mallia jossa kirjan hintaa selitetään maa- ja kirjailijamuuttujilla. Malli on siis muotoa:

$y_{ijk}=\mu+\alpha_i+\beta_j+\epsilon_{ijk}$, $\epsilon_{ijk} \sim \mathrm{NID}(0,\sigma^2)$
missä α on maa, eli $1\leq$ i ≤ 5 ja β on kirjailija, eli $1\leq$ j$\leq$ 10

R-koodi:

```
lmunb <- lm(p ~ country + author, na.action=na.omit); summary(lmunb)
```

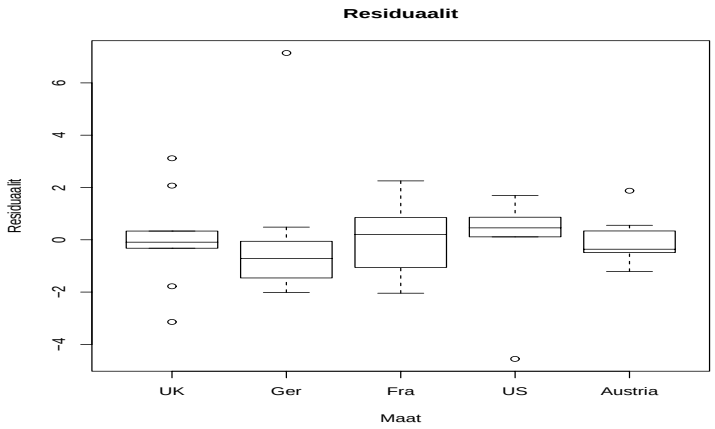
Koodissa "na.action=na.omit"optio poistaa aineistosta puuttuvat arvot, jolloin saadaan epätasapainoinen asetelma.

```
Residuals:
      Min        1Q    Median        3Q       Max
-4.558021 -0.711918  0.001043  0.554468  7.145296

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     12.9210     1.1957  10.806 2.62e-11 ***
countryGer       1.7730     1.0361   1.711  0.09850 .
countryFra      -1.8778     1.1312  -1.660  0.10847
countryUS       -2.0337     1.0778  -1.887  0.06998 .
countryAustria   1.1538     1.1308   1.020  0.31661
author2          2.3520     1.4134   1.664  0.10766
author3          0.0280     1.4134   0.020  0.98434
author4         -1.0860     1.6607  -0.654  0.51868
author5          1.8043     1.5102   1.195  0.24258
author6         -0.1140     1.4134  -0.081  0.93631
author7          8.1708     1.5116   5.405 1.03e-05 ***
author8          0.7357     1.6602   0.443  0.66120
author9         -2.7340     1.6603  -1.647  0.11120
author10        -4.7863     1.5115  -3.167  0.00380 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.235 on 27 degrees of freedom
Multiple R-Squared: 0.803,      Adjusted R-squared: 0.7082
F-statistic: 8.467 on 13 and 27 degrees of freedom,     p-value: 1.796e-006
```

Taulukko 2: Täyden mallin parametrit

Taulukosta 2 nähdään, että selitysaste on mainio. Ainoastaan kolme parametriä on tilastollisesti merkitseviä. Arvoista näkee mm. sen, että Yhdysvalloissa ja Ranskassa kirjojen hinnat ovat halvempia verrattuna muihin maihin. Kokeilimme selittää mallia pelkän maamuuttujan avulla, mutta kyseisen mallin selitysaste R^2 oli kovin alhainen (0.2224). Läheisemmin tarkastelemme taulukossa 2 olevaa mallia.

Boxplot-kuviosta nähdään miten kirjojen hinnat vaihtelevat maittain.



Kuva 1: Mallin residuaalit

2 Varianssianalyysi

2.1 Yksisuuntainen varianssianalyysi

Tutkimme miten maa- ja kirjailijamuuttujat vaikuttavat yksinään kirjojen hintaan yksisuuntaisella varianssianalyysillä. R-koodi:

```
anova(lm(p ~ author, na.action=na.omit)))
```

Vastaavasti suoritimme myös pelkälle maalle varianssianalyysin.

```
Analysis of Variance Table
Response: p
          Df Sum Sq Mean Sq F value    Pr(>F)
author     9 457.02   50.78  6.9181 2.101e-05 ***
Residuals 31 227.54    7.34
---
Response: p
          Df Sum Sq Mean Sq F value  Pr(>F)
country    4 152.27   38.07  2.5745 0.05403 .
Residuals 36 532.29   14.79
```

Taulukko 3: Kirjailijoiden ja maiden yksisuuntaiset varianssianalyysin tulokset

Havaitsimme, että kirjailijoiden vaikutus kirjan hintaan on selvempi kuin maiden, sillä maamuuttujan tapauksessa p-arvo on suurempi. Luonnollisesti kirjailijamuuttujalla on vaikutusta, sillä eri kirjat ovat eri hintaisia jo painokustannuksiltaan.

2.2 Kaksisuuntaisen varianssianalyysin malli

Kaksisuuntaisen varianssianalyysin avulla pyrimme tutkimaan "yhtaikaisesti"kahden selittäjän vaikutusta hintaan. Kaksisuuntaisen varianssianalyysin malli voidaan kirjoittaa muodossa
y_{ijk}=μ +α_i+β_j+ϵ_{ijk}, $\epsilon_{ijk} \sim$ NID(0,σ^2)

Parametri α_i kuvaa sarakevaikutusta ja β_j rivivaikutusta. Taulukosta 4 näemme että maalla on vaikutusta kirjojen hintoihin, kuten myös kirjailijalla. R-koodi:

```
unbaov <- anova(lmunb)
print(unbaov)
```

Tulokset ovat nähtävissä taulukossa 4.

```
Analysis of Variance Table

Response: p
          Df Sum Sq Mean Sq F value    Pr(>F)
country    4 152.27   38.07  7.6223 0.0003025 ***
author     9 397.45   44.16  8.8428 4.608e-06 ***
Residuals 27 134.84    4.99
---

Response: p
               Df Sum Sq Mean Sq F value Pr(>F)
country:author 40 684.56   17.11
Residuals       0   0.00
```

Taulukko 4: Kaksisuuntainen varianssianalyysi

Muuttujilla on yhteisvaikutusta, mikä voidaan huomata käyttämällä hyväksi edellisen taulukon arvoja. Asetetaan H_0 :ksi "maalla ja kirjailijalla ei ole yhdysvaikutusta". Jos H_0 on tosi, niin $F_{\alpha\beta} = MS_{\alpha\beta}/MSE \sim F_{df_{\alpha\beta},df_{SSE}}$
Nyt H_0 hylätään riskitasolla γ, jos aineistosta laskettu $F_{\alpha\beta}$:n arvo $> F_{\gamma;df_{\alpha\beta},df_{SSE}}$
Näin laskettu F-arvo on 3.42886 ja "R" :llä laskettu p-arvo saadaan seuraavasti:

```
df(3.42886, 40, 27, log = FALSE)
0.001517231
```

Siis H_0 hylätään koska p-arvo (0.0015) on hyvin pieni. Siis maalla ja kirjailijalla on yhdysvaikutusta eli interaktiota. Eli kirjailijalla on vaikutusta hintaan eri maaryhmissä.

2.3 Ortogonaalisuuden tarkistus

Käytettäessä "na.action ...-optiota poistamaan puuttuvat arvot saimme epätasapainoisen (unbalanced) asetelman. Tällä seikalla on merkittäviä seurauksia: parametrit β_1 (maa) ja β_2 (kirjailija) eivät ole enää ortogonaalisia. β_1, β_2 ovat ortogonaalisia jos ja vain jos $\hat{\beta}_1$=$\tilde{\beta}_1$, missä $\hat{\beta}_1$ on pienimmän neliösumman estimaattorien arvot kun β_1 on sovitettu malliin Y=$\mathbf{X}_1\beta_1+\epsilon$ (eli oletetaan β_2=0) ja $\tilde{\beta}=\binom{\tilde{\beta}_1}{\tilde{\beta}_2}$ on pienimmän neliösumman estimaattorien arvot kun β sovitettu malliin Y=$\mathbf{X}\beta+\epsilon$ (eli oletetaan täysi malli). Taulukossa 2 on täyden mallin $\tilde{\beta}_1$:n arvot joita vertasimme malliin joka olettaa,että β_2=0. Tämä tulostus on taulukossa 5. Näemme heti että $\hat{\beta}_1 \neq \tilde{\beta}_1$, eli parametrit eivät ole ortogonaalisia. Siis tämän tutkimuksen tapauksessa maamuuttujan estimoidut arvot olisivat täsmälleen samat sekä taulukossa 2 että taulukossa 5 mikäli parametrit olisivat ortogonaalisia. Tämä ei selvästikään pidä paikkaansa.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      13.358      1.216  10.985 4.76e-13 ***
countryGer        1.740      1.767   0.985    0.331
countryFra       -2.572      1.895  -1.357    0.183
countryUS        -2.014      1.824  -1.104    0.277
countryAustria    2.571      1.895   1.357    0.183
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.845 on 36 degrees of freedom
Multiple R-Squared: 0.2224,     Adjusted R-squared: 0.136
F-statistic: 2.574 on 4 and 36 degrees of freedom,      p-value: 0.05403
```

Taulukko 5: Mallin parametrit jossa on muuttujana vain maa

3 Multiplikatiivinen malli ja logaritmointi

Tutkimme oletusta, että hinta muodostuu multiplikatiivisesta mallista,
$y = a{x_1}^{\alpha}{x_2}^{\beta}$ missä ${x_1}^{\alpha}$ on maan vaikutus ja ${x_2}^{\beta}$ kirjailijan vaikutus, koska multiplikatiivinen malli sopii usein hyvin jos selitettävänä on hinta. Logaritmointihan muuttaa tämän mallin additiiviseksi jolloin tarkastelu helpottuu. Näin saatu malli on muotoa:
logy= $loga + \alpha_l ogx_1 + \beta logx_2$

R-koodina tämä malli on muotoa:

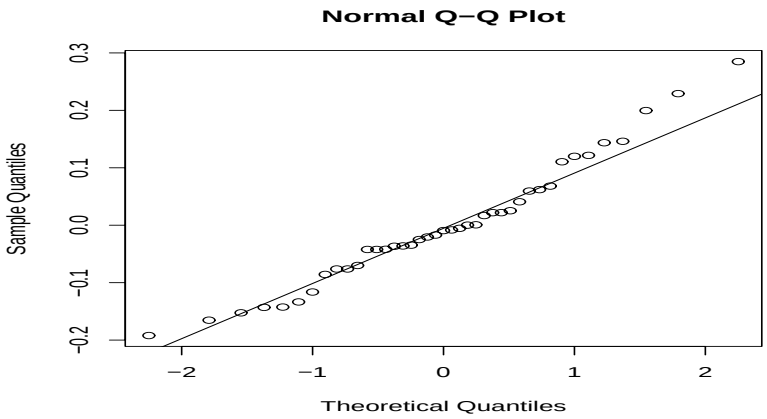
```
lmunblp <- lm(lp ~ country + author, na.action=na.omit); summary(lmunblp)
```

missä lp on logaritmoitu hinta. Saimme mallin joka sopii paremmin kuin aikaisempi additiivinen malli. Saadut tulostukset ovat taulukossa 6.

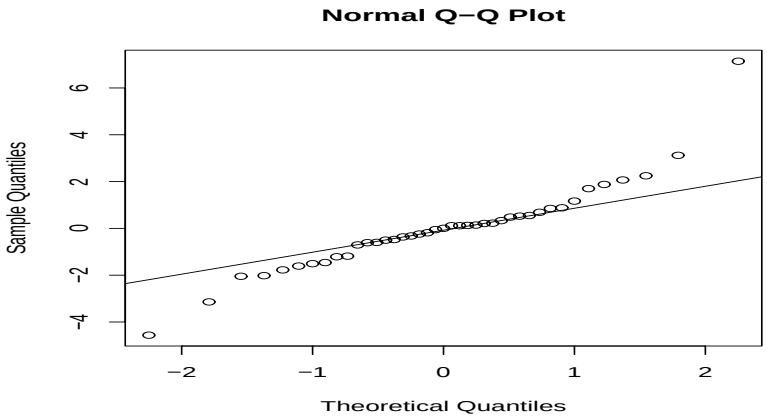
```
Residuals:
      Min        1Q    Median        3Q       Max
-0.191893 -0.070148 -0.009257  0.059484  0.285033

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.561246   0.070331  36.417  < 2e-16 ***
countryGer      0.095050   0.060942   1.560  0.13048
countryFra     -0.198485   0.066535  -2.983  0.00599 **
countryUS      -0.155406   0.063399  -2.451  0.02099 *
countryAustria  0.064596   0.066511   0.971  0.34007
author2         0.185536   0.083135   2.232  0.03413 *
author3         0.020192   0.083135   0.243  0.80993
author4        -0.085507   0.097683  -0.875  0.38911
author5         0.152138   0.088831   1.713  0.09824 .
author6         0.009107   0.083135   0.110  0.91358
author7         0.460201   0.088911   5.176 1.90e-05 ***
author8         0.060404   0.097654   0.619  0.54140
author9        -0.228988   0.097658  -2.345  0.02664 *
author10       -0.488776   0.088908  -5.498 8.03e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1314 on 27 degrees of freedom
Multiple R-Squared: 0.8665,     Adjusted R-squared: 0.8023
F-statistic: 13.49 on 13 and 27 degrees of freedom,     p-value: 1.375e-008
```

Taulukko 6: Multiplikatiivisen täyden mallin parametrit

Havaitsimme, että mallin selitysaste on hivenen parempi, mutta mallin selittäjien p-arvot ovat selvästi merkitsevämpiä. Toisaalta suuri p-arvo merkitsee vain sitä, että kyseisen parametrin arvo on sama kuin vertailtava arvo. Eli taulukossa 6 author6:n arvo on p-arvoltaan merkityksetön. Tämä ei kuitenkaan tarkoita että sen voisi poistaa mallista, koska kyseisen kirjailijan kirja vain sattuu olemaan lähes saman hintainen kuin "S.Hawking, A brief history of time"kirja, joka on vertailukohtana.



Kuva 2: QQ-plot kun hinta logaritmoitu



Kuva 3: QQ-plot kun hintaa ei ole logaritmoitu

Kuten kuvioistakin helposti huomataan, suora sopii paljon paremmin kun hinta on logaritmoitu. QQ-plot tarkastelu liittyy mallin valintaan. Alemmassa kuvassa pisteet karkaavat pois suoralta alussa ja lopussa, eli "hännissä". Ylemmässä kuvassa "hännät"ovat suoralla paljon kauniimmin. Tämä malli sopii aineistoon paremmin.

4 Päätelmät

Tutkimuksen päällimmäisenä tarkoituksena oli tutkia ovatko kirjat kalliimpia muissa maissa Britanniaan verrattuna. Tulokseksi saatiin, että Yhdysvalloissa ja Ranskassa kirjat ovat kalliimpia, ja Saksassa sekä Itävallassa halvempia. Logaritmoitu malli sopii paremmin, joka voidaan todeta mm. QQ-plot kuvioista. Logaritmoidun mallin parametrien estimaateista nähdään suoraan kuinka monta prosenttia kallimpia kirjoja muissa maissa on verrattuna Britanniaan. Näin on myös kirjailijoiden tapauksessa. Esimerkiksi Ranskassa "S.Hawking, A brief history of time"kirja on 19,8 prosenttia halvempi kuin Britanniassa. Parametrien estimaattien arvot voidaan siis muuttaa prosenteiksi, ja etumerkki kertoo onko kirja kalliimpi vai halvempi.