Tilastollinen mallintaminen
Harjoitustyö

Solja Malinen ja Hanna Virkkunen

WS4
sitä sun tätä

Matematiikan, tilastotieteen ja filosofian laitos
Tilastotiede
Marraskuu 2000

1 Miehet, naiset ja ammatissa pärjääminen

1.1 Aineiston kuvaus

Ensimmäisessä osassa työtämme käytämme aineistona The Independent lehden otsikolla "Irlantilaiset ja italialaiset ovat Euroopan seksistejä" 8.10.1992 julkaisemaa aineistoa. Aineistossa on havainnollistettu prosenttiluvuin, montako prosenttia maan kansalaisista (erään tutkimuksen mukaan) uskoo miesten ja naisten selviytyvän yhtä hyvin bussin/junankuljettajan, kirurgin, asianajajan ja kansanedustajan ammateissa. Maista mukana ovat Tanska, Hollanti, Ranska, Englanti, Belgia, Espanja, Portugali, Saksa, Luxemburg, Kreikka, Italia ja Irlanti. Aineisto on seuraavanlainen:

```
86 85 82 86  Denmark
75 83 75 79  Netherlands
77 70 70 68  France
61 70 66 75  UK
67 66 64 67  Belgium
56 65 69 67  Spain
52 67 65 63  Portugal
57 55 59 64  W. Germany
47 58 60 62  Luxembourg
52 56 61 58  Greece
54 56 55 59  Italy
43 51 50 61  Ireland
```

1.2 Tilastollinen analyysi

Kaksisuuntaisessa varianssianalyysissä testataan molemmille muuttujille hypoteeseja:

H_0: Maalla/ammatilla ei ole omavaikutusta
H_1: Maalla/ammatilla on omavaikutusta

```
Response: p
          Df Sum Sq Mean Sq F value    Pr(>F)
COUNTRY   11 4243.9   385.8 22.8885 2.438e-12 ***
OCC        3  291.7    97.2  5.7694  0.002752 **
Residuals 33  556.3    16.9
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Taulukko 1 *Varianssianalyysin tulostus*

Tulostuksesta nähdään, että niin ammatilla kuin maallakin on omavaikutusta 5 prosentin riskitasolla, koska H_0 hylätään molemman muuttujan tapauksessa. Eli suhtautuminen on erilaista eri ammatteihin ja eri maissa ajatellaan eri tavalla.

Kokeillaan seuraavaksi aineistoon erilaisia malleja niin, että prosentteja selitetään joko maalla tai ammatilla tai sitten molemmilla yhtä aikaa. Tarkastellaan ensin mallia

$$y_{ijk}) = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \tag{1}$$

missä $i = 1, \ldots, 12$; $j = 1, \ldots, 4$; $k = 1$ ja $\epsilon_{ijk} \sim NID(0, \sigma^2)$. Mallissa μ on yleiskeskiarvo (intercept), α on maa ja β on ammatti.

```
Residuals:
    Min      1Q  Median      3Q     Max
-6.2083 -1.9583 -0.7083  2.2917  9.6250

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       80.875      2.295  35.238  < 2e-16 ***
COUNTRYHollanti   -6.750      2.903  -2.325 0.026363 *
COUNTRYRanska    -13.500      2.903  -4.650 5.15e-05 ***
COUNTRYEnglanti  -16.750      2.903  -5.770 1.90e-06 ***
COUNTRYBelgia    -18.750      2.903  -6.459 2.52e-07 ***
COUNTRYEspanja   -20.500      2.903  -7.061 4.41e-08 ***
COUNTRYPortug    -23.000      2.903  -7.923 3.89e-09 ***
COUNTRYSaksa     -26.000      2.903  -8.956 2.38e-10 ***
COUNTRYLuxemb    -28.000      2.903  -9.645 3.97e-11 ***
COUNTRYKreikka   -28.000      2.903  -9.645 3.97e-11 ***
COUNTRYItalia    -28.750      2.903  -9.903 2.06e-11 ***
COUNTRYIrlanti   -33.500      2.903 -11.539 3.99e-13 ***
OCCkirurgi         4.583      1.676   2.735 0.009970 **
OCCasianajaja      4.083      1.676   2.436 0.020405 *
OCCkans.ed.        6.833      1.676   4.077 0.000270 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.106 on 33 degrees of freedom
Multiple R-Squared: 0.8908,     Adjusted R-squared: 0.8444
F-statistic: 19.22 on 14 and 33 degrees of freedom,
p-value: 6.162e-012
```

Taulukko 2 *Mallin (1) kertoimet*

Tulostuksesta nähdään, että jos valitaan kaksi pahinta, niin italialaiset ja irlantilaiset ovat Euroopan seksistejä, koska maa-selittäjän estimaatit poikkeavat eniten alaspäin perustasosta. Perustasona tässä pidetään bussin/ junankuljettajan arvostusta Tanskassa, koska nämä ovat ensimmäiset muuttujat listassa ja siten määritellään vertailukohteiksi. Tämä siis saa nyt arvokseen pelkän vakiokertoimen 80.875. Mitä pienempi maan estimaatti on, sitä seksistisempää kansa on. Tanskalaiset uskovat eniten naisten ja miesten tasavertaiseen selviytymiseen. Maasta riippumatta naisten uskotaan selviävän ko. ammateista parhaiten kansanadustajan toimessa. F-arvo on 19.22 ja tätä vastaava p-arvo on 6.162e-012 eli H_0=Muuttujilla ei ole yhdysvaikutusta hylätään, ja siispä muuttujilla on yhdysvaikutusta. Selitysaste on 0.89, joten mallin avulla saadaan melkoisen hyvin selitettyä aineistoa, ja tuloksiin on näin ollen luottamista.

Samaan tulokseen päästään myös muuttujien keskiarvojen perusteella:

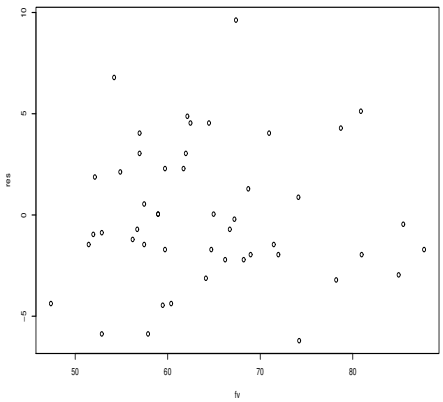
```
    bussik    kirurgi   asianajaja   kans.ed.
  60.58333   65.16667   64.66667    67.41667
```

Kansanedustajalla prosenttien keskiarvo on suurin ja niinpä ko. ammatissa uskotaan miesten ja naisten selviytyvän yhtä hyvin. Koska bussinkuljettajan kohdalla keskiarvo on pienin on se viittaus siihen, että bussinkuljettajan ammatissa oletettavasti mies selviää paremmin.

```
  Tanska Hollanti   Ranska Englanti   Belgia  Espanja   Portug
   84.75    78.00    71.25    68.00    66.00    64.25    61.75
  Saksa    Luxemb  Kreikka   Italia  Irlanti
   58.75    56.75    56.75    56.00    51.25
```

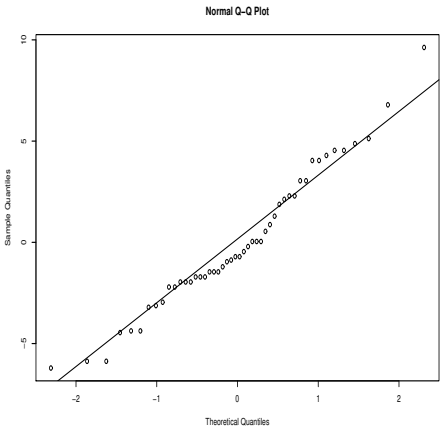
Maiden keskiarvojen perusteella päädymme puolestaan samaan tulokseen, että Irlannissa ja Italiassa asuvat Euroopan seksistit.

Tarkastellaan seuraavaksi pisteparvea, johon on piirretty estimoidun mallin residuaalit ja sovitteet. Koska pisteparvi on tasaisesti hajallaan nollan ympäristössä, on mallin valinta onnistunut.



Kuva 1: Pisteparvi

Entäs kun tarkastellaan Q-Q-kuvaajaa? Pisteet eivät asetu täydellisesti suoralle, mutta kuvaaja kertoo, että havainnot ovat normaalijakaumasta.



Kuva 2: Q-Q-kuvaaja

1.3 Muuttujien ortogonaalisuus

Yleisesti järjestys, jossa muuttujia lisätään malliin, on tärkeä ja vaikuttaa tuloksiin. Kun kokeilimme aineistoon malleja, joissa oli vain toinen faktori mukana, olivat kertoimet yhtä suuria verrattuna täyteen malliin. Toki vakiokerroin muuttui, mutta tämä johtui siitä, että täydessä mallissa vakio kuvaa bussin/junankuljettajaa Tanskassa kun taas yhden faktorin malleissa vakio kuvaa vain joko bussin/junankuljettajaa tai tanskalaisten yleistä

mielipidettä. Järjestyksellä ei ole väliä, koska muuttujat ovat ortogonaalisia. Ortogonaalisuus tarkoittaa nyt siis sitä, että selittäjän OCC (ammatti) estimaatit ovat samoja olipa muuttuja COUNTRY (maa) mallissa tai ei. Siispä muuttujien järjestyksen vaihtamisellakaan ei ole merkitystä tuloksiin.

2 Lomakohteiden hintavertailua

2.1 Aineiston kuvaus

Toisessa osiossa aineistona on samaisen lehden otsikolla "Tourists get hidden costs warnings" 16.6.1999 julkaisema aineisto. Tässä aineistossa muuttujina ovat kohde sekä hinnat (puntina) seuraaville tuotteille: kolmen ruokalajin ateria, pullo olutta, Suntan aurinkoöljy, taksi (5km), filmirulla (24) ja auton vuokra (viikko). Aineisto alla.

Algarve	8.00	0.50	3.50	3.00	4.00	100.00
CostaDelSol	6.95	1.30	4.10	12.30	4.10	130.85
Majorca	10.25	1.45	5.35	6.15	3.30	122.20
Tenerife	12.30	1.25	4.90	3.70	2.90	130.85
Florida	15.60	1.90	5.05	5.00	2.50	114.00
Tunisia	10.90	1.40	5.45	1.90	2.75	218.10
Cyprus	11.60	1.20	5.95	3.00	3.60	149.45
Turkey	6.50	1.05	6.50	4.90	2.85	263.00
Corfu	5.20	1.05	3.75	4.20	2.50	137.60
Sorrento	7.70	1.40	6.30	8.75	4.75	215.40
Malta	11.20	0.70	4.55	8.00	4.80	87.85
Rhodes	6.30	1.05	5.20	3.15	2.70	261.30
Sicily	13.25	1.75	4.20	7.00	3.85	174.40
Madeira	10.25	0.70	5.10	6.85	6.85	153.70

Tässä aineistossa tutkitaan siis eri lomapaikkojen hintatasoa. Kun tarkastellaan alkuperäisestä aineistosta esim. Turkin hintoja huomataan, että siellä on muuten suhteellisen "normaali" hintataso muiden tuotteiden osalta, mutta autonvuokraus on kalliimpaa kuin missään muualla.

2.2 Tilastollinen analyysi

Estimoidaan malli

$$\log(y_{ijk}) = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \tag{2}$$

missä $i = 1, \ldots, 6$; $j = 1, \ldots, 14$; $k = 1$ ja $\epsilon_{ijk} \sim NID(0, \sigma^2)$. Mallissa μ on yleiskeskiarvo, α on tuote ja β on kohteen vaikutus.

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.895833 -0.173510  0.004155  0.192365  0.814000

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.8778     0.1629  11.527  < 2e-16 ***
PRODBeer            -2.1084     0.1295 -16.286  < 2e-16 ***
PRODSunoil          -0.6343     0.1295  -4.900 6.69e-06 ***
PRODTaxi            -0.6284     0.1295  -4.854 7.92e-06 ***
PRODFilm            -0.9679     0.1295  -7.476 2.49e-10 ***
PRODCarRent          2.8016     0.1295  21.640  < 2e-16 ***
RESORTSCostaDelSol   0.4463     0.1978   2.257  0.02740 *
RESORTSMajorca       0.4105     0.1978   2.076  0.04189 *
RESORTSTenerife      0.3067     0.1978   1.551  0.12584
RESORTSFlorida       0.4235     0.1978   2.142  0.03597 *
RESORTSTunisia       0.2884     0.1978   1.458  0.14963
RESORTSCyprus        0.3457     0.1978   1.748  0.08519 .
RESORTSTurkey        0.3787     0.1978   1.915  0.05993 .
RESORTSCorfu         0.0943     0.1978   0.477  0.63508
RESORTSSorrento      0.5981     0.1978   3.025  0.00356 **
RESORTSMalta         0.3282     0.1978   1.659  0.10187
RESORTSRhodes        0.2525     0.1978   1.277  0.20616
RESORTSSicily        0.5508     0.1978   2.785  0.00700 **
RESORTSMadeira       0.4590     0.1978   2.321  0.02343 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3425 on 65 degrees of freedom
Multiple R-Squared: 0.962,      Adjusted R-squared: 0.9514
F-statistic: 91.34 on 18 and 65 degrees of freedom, p-value: 0
```

Taulukko 3 *Mallin (2) kertoimet*

Vakio kuvaa kolmen ruokalajin ateriaa Algarvessa. Koska oluen estimaatti on pienin, on se halvinta ko. tuotteista. Kalleinta on odotetusti autonvuok-

raus ja matkakohteista puolestaan kallein on Sorrento. Jokaisen tuotteen estimaatti on merkitsevä 5%:n riskitasolla. Algarve on halvin lomakohteista, koska muiden maiden estimaatit ovat nollaa suurempia.

Kuten ensimmäisessä aineistossa, myös tässä huomasimme, että yhden selittäjän malleissa vain vakion arvo muuttuu jonkin verran ja muiden muuttujien estimaatit pysyvät samoina. Kun logaritmoidun hinnan selittäjänä käytetään vain tuotetta, niin selityskerroin on 0.95 ja kun selittäjänä on pelkästään matkakohde on selityskerroin vain 0.01. Hintaa ei siis voida selittää pelkästään matkakohteen avulla.

Estimoidaan seuraavaksi malli niin, että hinta y_{ijk} ei ole logaritmoitu.

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.5315    11.1499  -0.137   0.8912
PRODBeer            -8.5214     8.8610  -0.962   0.3398
PRODSunoil          -4.7214     8.8610  -0.533   0.5960
PRODTaxi            -4.1500     8.8610  -0.468   0.6411
PRODFilm            -6.0393     8.8610  -0.682   0.4979
PRODCarRent        151.6214     8.8610  17.111   <2e-16 ***
RESORTSCostaDelSol   6.7667    13.5354   0.500   0.6188
RESORTSMajorca       4.9500    13.5354   0.366   0.7158
RESORTSTenerife      6.1500    13.5354   0.454   0.6511
RESORTSFlorida       4.1750    13.5354   0.308   0.7587
RESORTSTunisia      20.2500    13.5354   1.496   0.1395
RESORTSCyprus        9.3000    13.5354   0.687   0.4945
RESORTSTurkey       27.6333    13.5354   2.042   0.0453 *
RESORTSCorfu         5.8833    13.5354   0.435   0.6652
RESORTSSorrento     20.8833    13.5354   1.543   0.1277
RESORTSMalta        -0.3167    13.5354  -0.023   0.9814
RESORTSRhodes       26.7833    13.5354   1.979   0.0521 .
RESORTSSicily       14.2417    13.5354   1.052   0.2966
RESORTSMadeira      10.7417    13.5354   0.794   0.4303
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.44 on 65 degrees of freedom
Multiple R-Squared: 0.8911,     Adjusted R-squared: 0.8609
F-statistic: 29.55 on 18 and 65 degrees of freedom, p-value: 0
```

Taulukko 4 *Logaritmoimattoman mallin kertoimet*

Selitysaste ensinnäkin hieman pienenee, mutta voidaan edelleen puhua mallin hyvästä selityskyvystä. Edelleen kuitenkin olut on halvinta ja autonvuok-

raus kalleinta, mutta matkakohteiden kohdalla tapahtuu muutos: Malta on halvin ja Turkki kallein.

Keskiarvojen perusteellakin nähdään, että matkakohteista Malta on halvin ja Turkki kallein. Turkin suureen keskiarvoon vaikuttaa juuri autonvuokrauksen korkea hinta.

```
   Algarve CostaDelSol     Majorca    Tenerife     Florida
  19.83333    26.60000    24.78333    25.98333    24.00833
   Tunisia      Cyprus      Turkey       Corfu    Sorrento
  40.08333    29.13333    47.46667    25.71667    40.71667
     Malta      Rhodes      Sicily     Madeira
  19.51667    46.61667    34.07500    30.57500
```

Tuotteiden keskiarvoista voi päätellä, että autonvuokraus on kalleinta ja halvinta on olut.

```
     Meal       Beer     Sunoil       Taxi       Film    CarRent
 9.714286   1.192857   4.992857   5.564286   3.675000 161.335714
```

Koska autonvuokraus näyttää vaikuttavan niin paljon tuloksiin, poistetaan se aineistosta ja estimoidaan ensiksi malli

$$\log(y_{ijk}) = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \tag{3}$$

missä $i = 1, \ldots, 5$; $j = 1, \ldots, 14$; $k = 1$ ja $\epsilon_{ijk} \sim NID(0, \sigma^2)$.

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.812386 -0.162964  0.009351  0.194066  0.763684

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.89262    0.17032  11.112 2.44e-15 ***
PRODBeer           -2.10842    0.12695 -16.609  < 2e-16 ***
PRODSunoil         -0.63434    0.12695  -4.997 6.96e-06 ***
PRODTaxi           -0.62845    0.12695  -4.951 8.19e-06 ***
PRODFilm           -0.96792    0.12695  -7.625 5.01e-10 ***
RESORTSCostaDelSol  0.48174    0.21242   2.268   0.0275 *
RESORTSMajorca      0.45247    0.21242   2.130   0.0379 *
RESORTSTenerife     0.31421    0.21242   1.479   0.1451
RESORTSFlorida      0.48206    0.21242   2.269   0.0274 *
RESORTSTunisia      0.19007    0.21242   0.895   0.3750
RESORTSCyprus       0.33446    0.21242   1.575   0.1214
RESORTSTurkey       0.26100    0.21242   1.229   0.2247
```

```
RESORTSCorfu        0.04932    0.21242   0.232   0.8173
RESORTSSorrento     0.56430    0.21242   2.656   0.0105 *
RESORTSMalta        0.41969    0.21242   1.976   0.0535 .
RESORTSRhodes       0.11094    0.21242   0.522   0.6037
RESORTSSicily       0.54974    0.21242   2.588   0.0125 *
RESORTSMadeira      0.46488    0.21242   2.188   0.0332 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3359 on 52 degrees of freedom
Multiple R-Squared: 0.8601,     Adjusted R-squared: 0.8143
F-statistic:  18.8 on 17 and 52 degrees of freedom,
p-value: 2.22e-016
```

Taulukko 5 *Mallin (3) kertoimet*

Edelleenkin olut on tuotteista halvin, mutta nyt kallein on kolmen ruokalajin ateria. Algarve on edelleen halvin ja Sorrento kallein. Erot ovat kuitenkin tasaantuneet.

Entäs kun estimoidaan malli niin, että y_{ijk} ei ole logaritmoitu.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          8.4864     0.9855   8.611 1.40e-11 ***
PRODBeer            -8.5214     0.7346 -11.601 4.44e-16 ***
PRODSunoil          -4.7214     0.7346  -6.427 4.04e-08 ***
PRODTaxi            -4.1500     0.7346  -5.650 6.85e-07 ***
PRODFilm            -6.0393     0.7346  -8.222 5.69e-11 ***
RESORTSCostaDelSol   1.9500     1.2292   1.586   0.1187
RESORTSMajorca       1.5000     1.2292   1.220   0.2278
RESORTSTenerife      1.2100     1.2292   0.984   0.3295
RESORTSFlorida       2.2100     1.2292   1.798   0.0780 .
RESORTSTunisia       0.6800     1.2292   0.553   0.5825
RESORTSCyprus        1.2700     1.2292   1.033   0.3063
RESORTSTurkey        0.5600     1.2292   0.456   0.6506
RESORTSCorfu        -0.4600     1.2292  -0.374   0.7097
RESORTSSorrento      1.9800     1.2292   1.611   0.1133
RESORTSMalta         2.0500     1.2292   1.668   0.1014
RESORTSRhodes       -0.1200     1.2292  -0.098   0.9226
RESORTSSicily        2.2100     1.2292   1.798   0.0780 .
RESORTSMadeira       2.1500     1.2292   1.749   0.0862 .
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.943 on 52 degrees of freedom
Multiple R-Squared: 0.7536,     Adjusted R-squared: 0.6731
F-statistic: 9.356 on 17 and 52 degrees of freedom,
p-value: 1.785e-010
```

Taulukko 6 *Logaritmoimattoman mallin kertoimet*

Nyt mallin selitysaste laskee taas kerran. Tuotteiden halvin/kallein -suhteissa ei tapahdu muutosta edelliseen malliin, mutta matkakohteiden hintataso on muuttunut - halvin paikka on nyt Corfu ja kalleimmat paikat ovat Sisilia ja Florida. Tosin 5%:n riskitasolla ei voida sanoa mitään varmaa matkakohteiden hintatasosta. Tulokset ovat aivan erisuuntaiset kuin autonvuokrauksen kanssa. Kun autonvuokraus oli mukana olivat Turkki ja Rhodos kalleimpia paikkoja ja halvimpia olivat Algarve ja Malta.

3 Yhteenveto

Ensimmäisen aineiston suhteen tulimme siis siihen tulokseen, että suhtautuminen naisten pärjäämiseen eri ammateissa vaihtelee sekä ammateittain että maittain. Naisten uskotaan siis pärjäävän toisissa ammateissa paremmin kuin toisissa. Eri maitten välillä on myöskin eroa. Tanskalaiset luottavat naisten pärjäämiseen eniten. Vähiten luottamusta löytyy Italiasta ja Irlannista. Kansanedustajan toimessa naisten uskotaan selviävän parhaiten.

Matkakohteita koskevassa osuudessa saimme neljänlaisia tuloksia. Oli malli, jossa ensin auton vuokrauksen hinta oli mukana ja sitten se poistettiin mallista. Näistä molemmista versioista tehtiin sekä logaritmoitu että logaritmoimaton versio. Auton vuokrauksen hinta poistettiin siksi, että se oli yksikköhinnaltaan kohtuuttoman paljon suurempi kuin muiden mukana olleiden tuotteiden hinnat.

Tarkastelimme ensin malleja, joissa auton hinta oli mukana. Logaritmoidussa mallissa tuotteista halvin oli olut ja kalleinta luonnollisesti autonvuokraus. Kohteista halvin oli Algarve ja kallein puolestaan Sorrento.

Logaritmoimattomassa mallissa saatiin hieman heikompi selitysaste kuin logaritmoidussa. Selitysaste oli kuitenkin edelleen mainio. Tuotteiden hintavuusjärjestykseen logaritmointi ei vaikuta, sillä edelleen olut on halvinta ja pisimmän pennin saa pulittaa autosta. Kohteisiin logaritmointi puolestaan vaikuttaa, sillä nyt Malta on halvin ja kallein on Turkki.

Kun autonvuokraus sitten poistettiin aineistosta ja estimoitiin uudelleen sekä logaritmoitu että logaritmoimaton malli, saatiin taas hieman eri tulokset. Logaritmoidussa mallissa kalleimmaksi tuotteeksi nousi nyt ateria, oluen

pidellessä edelleen paikkaansa halvimmassa päässä. Kohteiden puolesta kallein ja halvin on sama kuin silloin kun autonvuokraus oli mukana. Logaritmoimattoman mallin tulokset taasen erosivat tästä vain kohteen osalta, siinä kalleimmat kohteet olivat Sisilia ja Florida, kun taas keveimmän kukkaron osoite sijaitsi Corfulla.