78185 Yleistetyt lineaariset mallit

Harjoitus 1, syksy 2014

Esimerkkiratkaisut

1. Bernoulli-jakauman pistetodennäköisyysfunktio on

$$f(y;p) = p^y(1-p)^{1-y}, \quad y \in \{0,1\},\ p \in (0,1).$$

   a) Osoita, että Bernoulli-jakauma kuuluu eksponentiaaliseen jakaumaperheeseen

$$f(y;\theta,\phi) = \exp\Big\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\Big\},$$

   b) Laske jakauman odotusarvo ja varianssi funktioiden $a(\phi)$ ja $b(\theta)$ avulla.

   Ratkaisu:

   a)

$$\begin{aligned} f(y;\theta,\phi) &= p^y(1-p)^{1-y} = \exp\left\{\log(p^y) + \log\left[(1-p)^{1-y}\right]\right\} \\ &= \exp\left\{y\log(p) + (1-y)\log(1-p)\right\} \\ &= \exp\left\{y\log\left(\frac{p}{1-p}\right) + \log(1-p)\right\} \end{aligned}$$

   Kun asetetaan

$$\theta = \log\left(\frac{p}{1-p}\right) \Leftrightarrow p = \frac{e^\theta}{1+e^\theta},$$

$$b(\theta) = -\log(1-p) = \log(1+e^\theta),$$

$$a(\phi) = \phi = 1 \quad \text{ja} \quad c(y,\phi) = 0,$$

   niin huomataan, että Bernoulli-jakauma kuuluu kyseiseen eksponentiaaliseen jakaumaperheeseen.

   b)

$$\begin{aligned} E(y) &= b'(\theta) = \frac{e^\theta}{1+e^\theta} = p, \\ Var(y) &= a(\phi)b''(\theta) = 1 \cdot \frac{e^\theta(1+e^\theta) - e^\theta e^\theta}{(1+e^\theta)^2} = \frac{e^\theta}{(1+e^\theta)^2} \\ &= \frac{e^\theta}{(1+e^\theta)} \cdot \frac{1}{(1+e^\theta)} = p(1-p). \end{aligned}$$

2. Eksponenttijakauman tiheysfunktio on

$$f(y;\lambda) = \lambda e^{-\lambda y}, \quad y > 0,\ \lambda > 0.$$

   a) Osoita, että eksponenttijakauma kuuluu eksponentiaaliseen jakaumaperheeseen.

   b) Laske jakauman odotusarvo ja varianssi funktioiden $a(\phi)$ ja $b(\theta)$ avulla.

Ratkaisu:

a)

$$f(y;\lambda) = \lambda e^{-\lambda y} = \exp\{y(-\lambda) + \log(\lambda)\}.$$

Kun asetetaan

$$-\lambda = \theta \Leftrightarrow \lambda = -\theta,$$
$$b(\theta) = -\log(\lambda) = -\log(-\theta),$$
$$a(\phi) = \phi = 1 \quad \text{ja} \quad c(y,\phi) = 0,$$

niin huomataan, että eksponenttijakauma kuuluu kyseiseen eksponentiaaliseen jakaumaperheeseen.

b)

$$E(y) = b'(\theta) = -\frac{1}{-\theta}\cdot(-1) = -\frac{1}{\theta} = \frac{1}{\lambda},$$
$$Var(y) = a(\phi)b''(\theta) = \frac{1}{\theta^2} = \frac{1}{\lambda^2}.$$

3. Gammajakauman tiheysfunktio on muotoa

$$f(y;\mu,\nu) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}\exp\left(-\frac{\nu}{\mu}y\right),$$

jossa $y > 0$, $\mu > 0$, $\nu > 0$.

a) Osoita, että gammajakauma kuuluu eksponentiaaliseen jakaumaperheeseen.

b) Laske jakauman odotusarvo ja varianssi funktioiden $a(\phi)$ ja $b(\theta)$ avulla.

Ratkaisu:

a)

$$\begin{aligned} f(y;\mu,\nu) &= \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}\exp\left(-\frac{\nu}{\mu}y\right) \\ &= \exp\{-\log(\Gamma(\nu)) + \nu\log(\nu/\mu) + (\nu-1)\log(y) - (\nu/\mu)y\} \\ &= \exp\left\{\frac{y(-1/\mu)}{1/\nu} - \log(\Gamma(\nu)) + \nu\log(\nu) - \nu\log(\mu) + (\nu-1)\log(y)\right\} \\ &= \exp\left\{\frac{y(-1/\mu) - \log(\mu)}{1/\nu} - \log(\Gamma(\nu)) + \nu\log(\nu) + (\nu-1)\log(y)\right\} \end{aligned}$$

Kun asetetaan

$$-1/\mu = \theta \Leftrightarrow \mu = -1/\theta,$$
$$b(\theta) = \log(\mu) = \log(-1/\theta),$$
$$a(\phi) = 1/\phi, \quad \phi = \nu$$

ja

$$\begin{aligned} c(y,\phi) &= -\log(\Gamma(\nu)) + \nu\log(\nu) + (\nu-1)\log(y) \\ &= -\log(\Gamma(\phi)) + \phi\log(\phi) + (\phi-1)\log(y), \end{aligned}$$

niin huomataan, että gammajakauma kuuluu kyseiseen eksponentiaaliseen jakaumaperheeseen.

b)

$$\begin{aligned} E(y) &= b'(\theta) = \frac{1/\theta^2}{-1/\theta} = -\frac{1}{\theta} = \mu, \\ Var(y) &= a(\phi)b''(\theta) = (1/\phi)(1/\theta^2) = (1/\nu)\mu^2 = \mu^2/\nu. \end{aligned}$$

4. Aineistossa `lapset85.dat` on tietoja Oulun läänissä 1985-86 syntyneistä lapsista ja heidän vanhemmistaan. Muuttujankuvaukset ovat tiedostossa `muuttujat.pdf`.

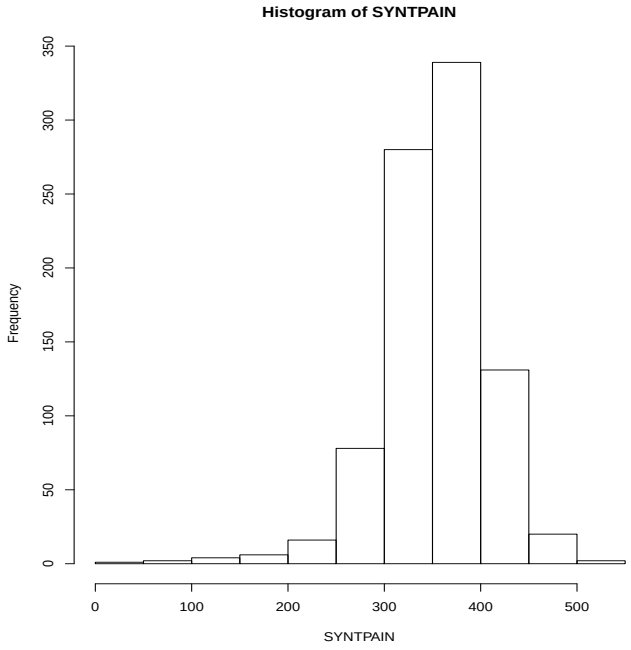
   a) Lue aineisto R-ohjelmistoon (voit käyttää myös muita tilasto-ohjelmistoja, esim. SPSS, SAS,...) ja kuvaile muuttujia sopivien tunnuslukujen ja graafisten esitysten avulla.

   b) Halutaan tutkia äidin raskaudenaikaisen tupakoinnin vaikutusta lapsen syntymäpainoon, kun mahdolliset analyysiä sekoittavat tekijät aineistossa ovat äidin ikä, äidin sosiaaliluokka, pariteetti, lapsen gestaatioikä ja sukupuoli. Sovita aineistoon erilaisia lineaarisia regressiomalleja ja mieti näiden mallien regressiokertoimien tulkintaa.

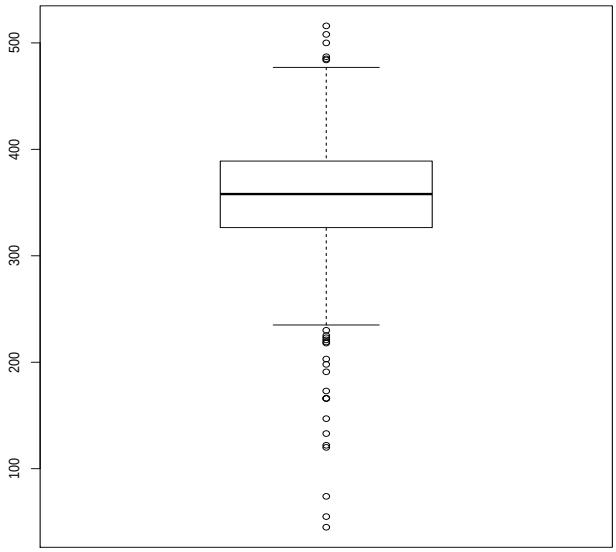
Ratkaisu:

a)
```
> lapset85 <- read.table("lapset85.dat", header=TRUE)
> attach(lapset85)
The following objects are masked from lapset85 (pos = 3):

    AIDINIKA, AIDINPAI, AIDINPIT, AIDINTUP, KAKSOSTU, KOHORTTI, KT2,
    KUOIKA, KUUKVIIK, NEUVVIIK, PARITEET, PERINKUO, RASKHIST, SIVSAATY,
    SOSLK, SUKUPUOL, SYNTPAIN, TUNNUS

> names(lapset85)
 [1] "TUNNUS"   "AIDINIKA" "AIDINPIT" "AIDINPAI" "AIDINTUP" "SIVSAATY"
 [7] "SOSLK"    "KOHORTTI" "RASKHIST" "NEUVVIIK" "KUUKVIIK" "KAKSOSTU"
[13] "KT2"      "PARITEET" "SUKUPUOL" "SYNTPAIN" "PERINKUO" "KUOIKA"
> summary(lapset85)
     TUNNUS        AIDINIKA        AIDINPIT        AIDINPAI
 Min.   :1224   Min.   :15.00   Min.   :143.0   Min.   :42.00
 1st Qu.:1462   1st Qu.:24.00   1st Qu.:160.0   1st Qu.:53.00
 Median :1692   Median :27.00   Median :163.0   Median :58.00
 Mean   :1692   Mean   :27.67   Mean   :163.2   Mean   :59.26
 3rd Qu.:1922   3rd Qu.:31.00   3rd Qu.:167.0   3rd Qu.:64.00
 Max.   :2167   Max.   :46.00   Max.   :187.0   Max.   :99.00
    AIDINTUP         SIVSAATY         SOSLK          KOHORTTI    RASKHIST
 Min.   :0.0000   Min.   :1.000   Min.   :1.000   Min.   :2   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2   1st Qu.:0.0000
 Median :0.0000   Median :1.000   Median :3.000   Median :2   Median :1.0000
 Mean   :0.3436   Mean   :1.064   Mean   :2.678   Mean   :2   Mean   :0.9215
 3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:3.000   3rd Qu.:2   3rd Qu.:1.0000
 Max.   :2.0000   Max.   :4.000   Max.   :4.000   Max.   :2   Max.   :2.0000
    NEUVVIIK        KUUKVIIK        KAKSOSTU          KT2
 Min.   :26.00   Min.   :27.00   Min.   :1.000   Min.   :1.000
 1st Qu.:39.00   1st Qu.:39.00   1st Qu.:1.000   1st Qu.:1.000
```

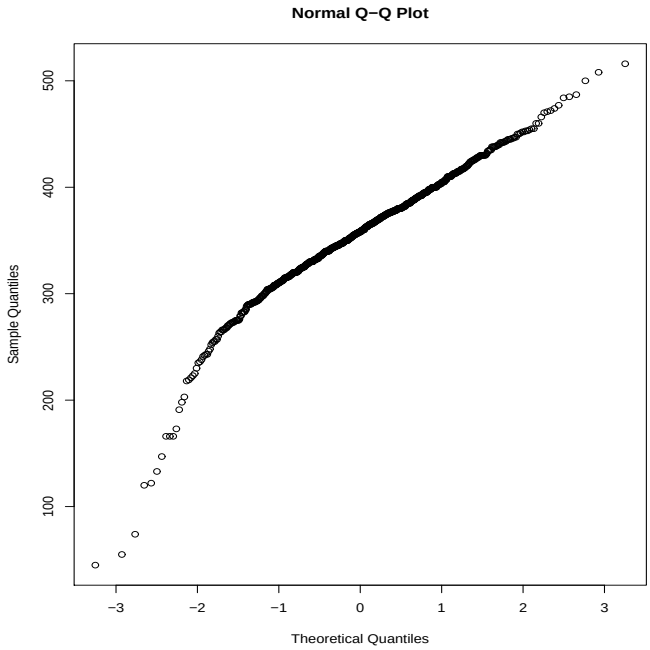
```
 Median :40.00   Median :40.00   Median :1.000   Median :1.000
 Mean   :39.36   Mean   :39.53   Mean   :1.028   Mean   :1.019
 3rd Qu.:40.00   3rd Qu.:41.00   3rd Qu.:1.000   3rd Qu.:1.000
 Max.   :43.00   Max.   :48.00   Max.   :2.000   Max.   :2.000
    PARITEET         SUKUPUOL        SYNTPAIN        PERINKUO
 Min.   : 0.000   Min.   :1.000   Min.   : 45.0   Min.   :0.00000
 1st Qu.: 0.000   1st Qu.:1.000   1st Qu.:326.5   1st Qu.:0.00000
 Median : 1.000   Median :1.000   Median :358.0   Median :0.00000
 Mean   : 1.344   Mean   :1.494   Mean   :355.4   Mean   :0.01024
 3rd Qu.: 2.000   3rd Qu.:2.000   3rd Qu.:389.0   3rd Qu.:0.00000
 Max.   :20.000   Max.   :2.000   Max.   :516.0   Max.   :1.00000
     KUOIKA
 Min.   :  -1
 1st Qu.:9999
 Median :9999
 Mean   :9885
 3rd Qu.:9999
 Max.   :9999
>
> hist(SYNTPAIN)
> boxplot(SYNTPAIN)
> qqnorm(SYNTPAIN)
```



Kuva 1: Histogrammi muuttujasta SYNTPAIN



Kuva 2: Box-plot-kuvio muuttujasta SYNTPAIN



Kuva 3: Kvantiilikuvio muuttujasta SYNTPAIN

Kuvioista huomataan, että lapsen syntymäpainon jakauma on selvästi vasemmalle vino. Logaritmimuunnoksella $\log(c - SYNTPAIN)$ tai neliöjuurimuunnoksella $\sqrt{c - SYNTPAIN}$ voisi mahdollisesti saada symmetrisemmän jakauman. Täytyy kuitenkin muistaa, että muuttujanmuun-

nokset muuttavat aina mallin tulkintaa.

b) Jätetään pois moniraskaudet ja perinataalikuolleet:

```
> lapset85<-lapset85[(KAKSOSTU==1)&(PERINKUO==0),]
```

Äidin tupakoinnista faktorimuuttuja:

```
> aidintup<-factor(AIDINTUP)
```

Sosiaaliluokasta kaksiluokkainen ("Alempi sos.luokka"= TRUE, "Ylempi sosiaaliluokka"= FALSE):

```
soslk34<-(SOSLK>=3)
```

Pariteettimuuttujasta kaksiluokkainen ("Ei aikaisempia synnytyksiä"= TRUE, "On aikaisempia synnytyksiä"= FALSE):

```
parity0<-(PARITEET==0)
```

Gestaatioikä kuukausina

```
gestika<-KUUKVIIK
```

Tehdään uusi sukupuolimuuttuja *poika* ("poika"=TRUE, "tyttö"=FALSE):

```
poika<-(SUKUPUOL==1)
```

```
> mod1<-lm(SYNTPAIN~aidintup+aidinika+soslk34+parity0+gestika+poika)
> summary(mod1)

Call:
lm(formula = SYNTPAIN ~ aidintup + aidinika + soslk34 + parity0 +
    gestika + poika)

Residuals:
     Min       1Q   Median       3Q      Max
-151.094  -29.183    1.558   28.710  133.023

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -232.333     33.198  -6.998 5.17e-12 ***
aidintup1    -14.580      5.401  -2.700 0.007074 **
aidintup2    -16.169      4.773  -3.387 0.000737 ***
aidinika       0.204      0.332   0.615 0.538998
soslk34TRUE    3.981      3.470   1.147 0.251613
parity0TRUE  -15.100      3.665  -4.120 4.14e-05 ***
gestika       14.725      0.779  18.903  < 2e-16 ***
poikaTRUE     11.571      3.099   3.734 0.000201 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 45.76 on 871 degrees of freedom
Multiple R-squared:  0.3185,    Adjusted R-squared:  0.313
F-statistic: 58.15 on 7 and 871 DF,  p-value: < 2.2e-16
```

Kun kolmiluokkainen faktorimuuttuja *aidintup* lisättiin malliin, niin funktio `lm` muodosti kaksi 0/1-muuttujaa, *aidintu*1 ja *aidintup*2:

$$\text{aidintup1} = \begin{cases} 1, & \text{kun äiti on tupakoinut vähän,} \\ 0, & \text{muulloin.} \end{cases}$$

$$\text{aidintup2} = \begin{cases} 1, & \text{kun äiti on tupakoinut paljon,} \\ 0, & \text{muulloin.} \end{cases}$$

Muuttujan *aidintup*1 regressiokertoimen mukaan äidin tupakoidessa vähän (äidin tupakoidessa paljon) raskauden aikana syntymäpäino on keskimäärin 146 grammaa pienempi (162 grammaa pienempi) kuin tupakoimattomilla äideillä, kun muiden muuttujien arvot pysyvät samana. Huom! Syntymäpainon mittayksikkö on 10 g. Muiden 0/1-muuttujien regressiokertoimet tulkitaan samalla tavalla. Gestaatioiän (jatkuva muuttuja *gestika*) kasvaessa yhdellä viikolla syntymäpäino kasvaa keskimäärin 147 grammaa, kun muiden muuttujien arvot pysyvät samana.

Vakiolla ei ole tässä mallissa mitään mielekästä tulkintaa. Keskistetään jatkuvat muuttujat *aidinika* ja *gestika* ja sovitetaan uudestaan lineaarinen regressiomalli:

```
> mean(aidinika)
[1] 27.67463
> aidika<-aidinika-28
> mean(gestika)
[1] 39.5256
> gest<-gestika-40
> mod1<-lm(SYNTPAIN~aidintup+aidika+soslk34+parity0+gest+poika)
> summary(mod1)

Call:
lm(formula = SYNTPAIN ~ aidintup + aidika + soslk34 + parity0 +
    gest + poika)

Residuals:
     Min       1Q   Median       3Q      Max
-151.094  -29.183    1.558   28.710  133.023

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  362.385      3.515 103.104  < 2e-16 ***
aidintup1    -14.580      5.401  -2.700 0.007074 **
aidintup2    -16.169      4.773  -3.387 0.000737 ***
aidika         0.204      0.332   0.615 0.538998
soslk34TRUE    3.981      3.470   1.147 0.251613
parity0TRUE  -15.100      3.665  -4.120 4.14e-05 ***
gest          14.725      0.779  18.903  < 2e-16 ***
poikaTRUE     11.571      3.099   3.734 0.000201 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 45.76 on 871 degrees of freedom
Multiple R-squared:  0.3185,    Adjusted R-squared:  0.313
F-statistic: 58.15 on 7 and 871 DF,  p-value: < 2.2e-16
```

Keskistämällä saatiin myös vakiolle järkevämpi tulkinta. Vakio kertoo nyt, että lapsen syntymäpaino on keskimäärin 3600 grammaa, kun äiti on tupakoimaton, äidin ikä on 28 vuotta, perhe kuuluu ylempään sosiaaliluokkaan, äidilllä on aikaisempia raskauksia, gestaatioikä on 40 viikkoa ja lapsi on tyttö.