78185 Yleistetyt lineaariset mallit

Harjoitus 4, syksy 2014

Esimerkkiratkaisut

1. Tarkastellaan havaintoaineistoa `esoph.dat`. Kyseessä on käytännössä sama aineisto kuin harjoituksen 3 tehtävässä 3, ainoastaan muuttujien arvot on muutettu numeerisiksi.

   a) Muuta muuttujat `agegp`, `alcgp` ja `tobgp` kaksiluokkaisiksi 0/1-muuttujiksi ja tutki ruokatorvensyöpään sairastumisen vaaraa logistisen regressioanalyysin avulla, kun selittävinä muuttujina ovat 0/1-muuttujat ikä, tupakointi ja alkoholinkäyttö.

   b) Plottaa residuaalidevianssit (tai Pearsonin residuaalit) lineaarisen prediktorin estimaatteja ($\hat{\eta}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$) vastaan. Jos logistinen malli on kunnossa, niin suurin osa pisteistä pitäisi sijaita tasaisesti vaaka-akselin molemmilla puolilla rajojen -3 ja $+3$ välissä. Huom! Tämä plotti ei ole kovin hyödyllinen, jos n_i:t ovat liian pieniä.

   c) Testaa linkkifunktion hyvyyttä seuraavan yksinkertaisen testin avulla (ks. esim. McCullagh & Nelder, s. 401). Sovita logistinen regressiomalli ja tallenna estimoidut lineaariset prediktorit $\hat{\eta}_i$. Lisää muuttuja $\hat{\eta}_i^2$ malliin. Jos $\hat{\eta}_i^2$ on tilastollisesti merkitsevä, niin linkkifunktio on väärin määritelty.

   Avuksi:

```
> mod1 <- glm(cbind(n1,n0)~x1+x2*x3, data=tied, family=binomial)
> res1<-residuals(mod1)  #residuaalidevianssit
> res2<-residuals(mod1, type = "pearson") #Pearsonin residuaalit
> pred<-mod1$linear.predictors #lineaaristen prediktorien estimaatit
```

   Ratkaisu:

   a) Muodostetaan uudet dikotomiset muuttujat:

$$age2 = \begin{cases} 0, & agegp \leq 3 \text{ eli 25-54 vuotta} \\ 1, & agegp \geq 4 \text{ eli 55+ vuotta} \end{cases}$$

$$alc2 = \begin{cases} 0, & alcgp \leq 2 \text{ eli } 0-79 \text{ g/vrk} \\ 1, & alcgp \geq 3 \text{ eli 80+ g/vrk} \end{cases}$$

$$tob2 = \begin{cases} 0, & tobgp \leq 2 \text{ eli } 0-19 \text{ g/vrk} \\ 1, & tobgp \geq 3 \text{ eli 20+ g/vrk} \end{cases}$$

   Jos käytetään havaintoaineistoa `esoph.dat`, niin uudet muuttujat saataisiin seuraavasti:

```
esophb<-read.table("esoph.dat",header=TRUE)
age2<- (esophb$agegp>=4)
alc2<- (esophb$agegp>=3)
tob2<- (esophb$tobgp>=3)
```

Käytetään harjoituksen 3 tehtävässä 3 käytettyä alkuperäistä aineistoa `esoph`. Tällöin muuttujat pitää ensin muuttaa takaisin numeeriseen muotoon komennolla `as.numeric`:

```
> age2<- (as.numeric(esoph$agegp)>=4)
> alc2<- (as.numeric(esoph$agegp)>=3)
> tob2<- (as.numeric(esoph$tobgp)>=3)
```

```
> mod1<-glm(cbind(ncases,ncontrols)~age2+alc2+tob2,
+ family=binomial,data=esoph)
> summary(mod1)

Call:
glm(formula = cbind(ncases, ncontrols) ~ age2 + alc2 + tob2,
    family = binomial, data = esoph)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6668  -0.7361   0.0364   0.8290   2.8197

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.6030     0.3268 -11.025  < 2e-16 ***
age2TRUE      0.4336     0.1902   2.280   0.0226 *
alc2TRUE      1.8992     0.3607   5.265  1.4e-07 ***
tob2TRUE      0.5523     0.1769   3.121   0.0018 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 227.24  on 87  degrees of freedom
Residual deviance: 133.62  on 84  degrees of freedom
AIC: 289.1

Number of Fisher Scoring iterations: 5
```

Devianssi jaettuna vapausasteilla on $133.62/84 \approx 1.59$, joten mallissa on hieman ylihajontaa. Kerrotaan keskivirhe-estimaatit luvulla $\sqrt{133.62/84} \approx 1.26$ ja lasketaan uudestaan p-arvot:

```
> summary(mod1)$coeff
```

```
              Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -3.6030153  0.3267901 -11.025471 2.880064e-28
age2TRUE     0.4335569  0.1901599   2.279960 2.261006e-02
alc2TRUE     1.8991829  0.3606851   5.265487 1.398180e-07
tob2TRUE     0.5522725  0.1769284   3.121447 1.799646e-03
> est<-summary(mod1)$coeff[,1]
> est
(Intercept)    age2TRUE    alc2TRUE    tob2TRUE
 -3.6030153   0.4335569   1.8991829   0.5522725
> se<-summary(mod1)$coeff[,2]
> se
(Intercept)    age2TRUE    alc2TRUE    tob2TRUE
  0.3267901   0.1901599   0.3606851   0.1769284
> tt<-est/(sqrt(133.62/84)*se)   #z-testisuureen arvo
> tt
(Intercept)    age2TRUE    alc2TRUE    tob2TRUE
  -8.741804    1.807720    4.174865    2.474913
> 2*(1-pnorm(abs(tt)))   #kaksisuunt. testin p-arvot
 (Intercept)     age2TRUE     alc2TRUE     tob2TRUE
0.000000e+00 7.065011e-02 2.981622e-05 1.332687e-02
```

Kerrottaessa keskivirhe-estimaatit luvulla $D/(r-p)$ p-arvoiksi saadaan 0.07 (age2), 0.00002 (alc2) ja 0.01 (tob2).

Lasketaan OR-estimaatit:

```
> exp(coefficients(mod1))
(Intercept)    age2TRUE    alc2TRUE    tob2TRUE
 0.02724146  1.54273507  6.68043388  1.73719632
> exp(confint.default(mod1))
                2.5 %      97.5 %
(Intercept) 0.01435708  0.05168858
age2TRUE    1.06273990  2.23952399
alc2TRUE    3.29449565 13.54629103
tob2TRUE    1.22813812  2.45725704
```

Ruokatorvensyöpään sairastumisen odds on 6.7-kertainen alkoholiryhmässä "80+ g/vrk" verrattuna ryhmään "0-79 g/vrk". Likimääräinen 95% luottamusväli ristitulosuhteelle (OR) on $(3.3, 13.5)$.

Ruokatorvensyöpään sairastumisen odds on 1.7-kertainen tupakointiryhmässä "20+ g/vrk" verrattuna ryhmään "0-19 g/vrk". Likimääräinen 95% luottamusväli ristitulosuhteelle (OR) on $(1.2, 2.5)$.

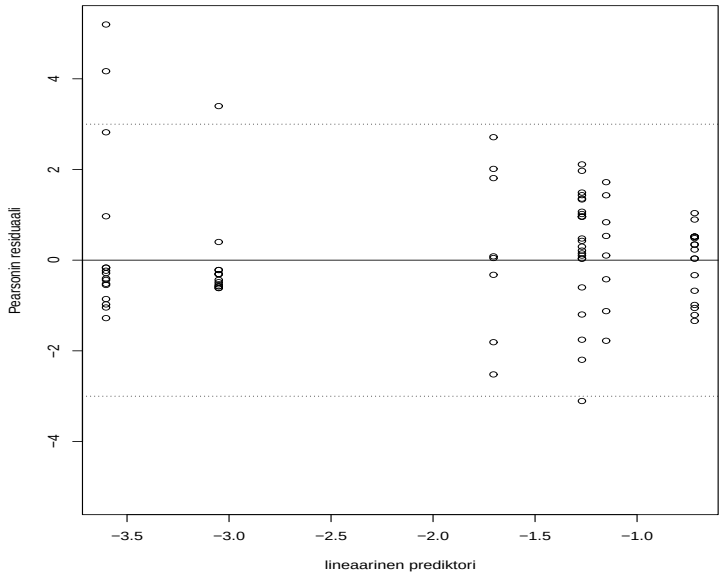
b) Lasketaan residuaalidevianssit, Pearsonin residuaalit ja lineaarisen prediktorin estimaatit:

```
> res.deviance<-residuals(mod1)
> res.pearson<-residuals(mod1, type = "pearson")
> sum(res.deviance^2)  #Devianssi (Residual deviance)
[1] 133.6174
> sum(res.pearson^2)   #Pearsonin chi^2-testisuureen arvo
[1] 158.5782
> pred<-predict(mod1) #lineaaristen prediktorien estimaatit
> pred<-mod1$linear.predictors #sama kuin edella
```

Pearsonin χ^2-testisuureen havaittu arvo on siis 158.5782 ja devianssi 133.6174. Tehdään residuaalikuviot:

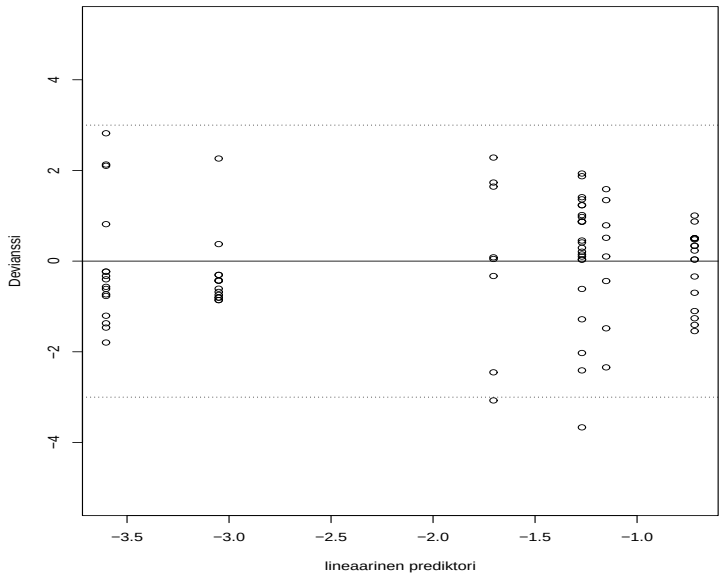
```
> plot(pred,res.pearson,xlab="lineaarinen prediktori",
+ ylab="Pearsonin residuaali")
> abline(h=0)
> abline(h=c(-3,3),lty=3)

> plot(pred,res.deviance,xlab="lineaarinen prediktori",
+ ylab="Devianssi")
> abline(h=0)
> abline(h=c(-3,3),lty=3)
```



Kuva 1: Pearsonin residuaali vs lineaarinen prediktori

Voidaan käyttää myös satunnaistettuja kvantiiliresiduaaleja (Dunn, K. P. & Smyth, G. K. (1996). Randomized quantile residuals. Journal of Com-



Kuva 2: Devianssi vs lineaarinen prediktori

putational and Graphical Statistics 5, 1-10). Satunnaistettuja kvantiiliresiduaaleja pystyy käyttämään vaikka lukumäärät n_i olisivat pieniä:

```
> library(statmod)
> res.quantile<-qresiduals(mod1)
> plot(pred,res.quantile,xlab="lineaarinen prediktori",
+ ylab="Satunnaistettu kvantiiliresiduaali")
> abline(h=0)
> abline(h=c(-3,3),lty=3)
```

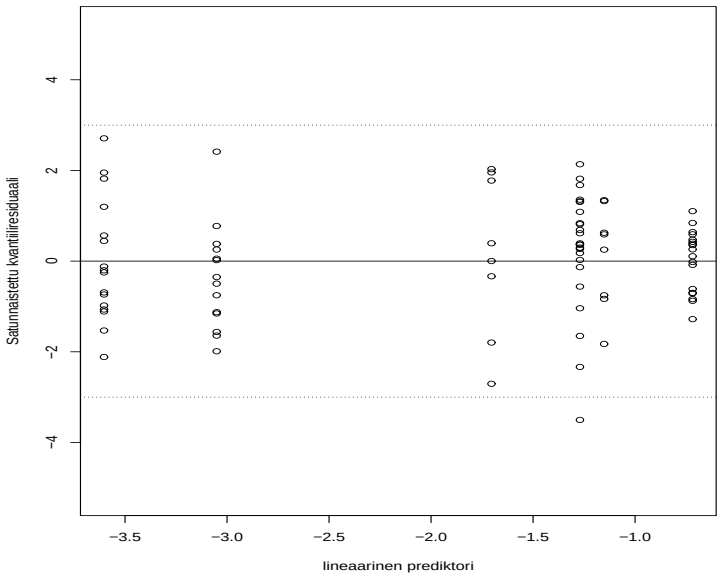
Residuaaliplottien (kuvat 1-3) perusteella näyttäisi siltä, että varianssi on likimain vakio lineaarisen prediktorin eri arvoilla.

c) Lisätään malliin selittäjäksi $\hat{\eta}_i^2 = (\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})^2$:

```
> mod2<-glm(cbind(ncases,ncontrols)~age2+alc2+tob2+I(pred^2),
+ family=binomial,data=esoph)
```

tai vaihtoehtoisesti

```
> pred2<-pred^2
> mod2<-glm(cbind(ncases,ncontrols)~age2+alc2+tob2+pred2,
+ family=binomial,data=esoph)
```



Kuva 3: Satunnaistettu kvantiiliresiduaali vs lineaarinen prediktori

Testataan onko uusi muuttuja tilastollisesti merkitsevä:

```
> anova(mod1,mod2,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(ncases, ncontrols) ~ age2 + alc2 + tob2
Model 2: cbind(ncases, ncontrols) ~ age2 + alc2 + tob2 + I(pred^2)
  Resid. Df Resid. Dev Df   Deviance Pr(>Chi)
1        84     133.62
2        83     133.62  1 0.00015947   0.9899
>
```

Testin perusteella (p-arvo=0.9899) $\hat{\eta}_i^2$ ei ole tilastollisesti merkitsevä, joten linkkifunktio vaikuttaisi olevan kunnossa.

2. Tuholaismyrkyn tehoa on tutkittu suihkuttamalla myrkkyä lehtikirvojen päälle ja laskemalla jälkeenpäin kuolleiden kirvojen lukumäärät. Kirvat jaettiin kymmeneen 200 yksilön ryhmään, joita käsiteltiin eri vahvuisilla myrkkyseoksilla. Aineisto löytyy tiedostosta `myrkky.dat`.

Annos (mg/l) x	Hyönteisten lkm n	Kuolleiden lkm r
1.0	200	2
2.0	200	6
3.0	200	33
4.0	200	68
5.0	200	108
6.0	200	137
7.0	200	152
8.0	200	162
9.0	200	178
10.0	200	186

a) Tutki graafisesti logistisen regressiomallin sopivuutta aineistoon. Paraneeko sopivuus, jos käytetään muuttujan x sijasta muuttujaa $logx = \log(x)$?
Vihje: Jos malli on kunnossa, niin pisteiden

$$\left(x_i, \log\left(\frac{p_i}{1-p_i}\right)\right)$$

pitäisi sijaita likimain suoralla. Todennäköisyydet p_i voi laskea suoraan havaintoaineistosta.

b) Tutki aineistoa logistisen regressiomallin avulla (esim. summary, anova).

c) Tutki b)-kohdassa saadun mallin sopivuutta graafisesti vertaamalla ennustettuja arvoja $\hat{p}_i$ todellisiin arvoihin $p_i = r_i/n_i$ (eri $logx$:n arvoilla).

Ratkaisu:

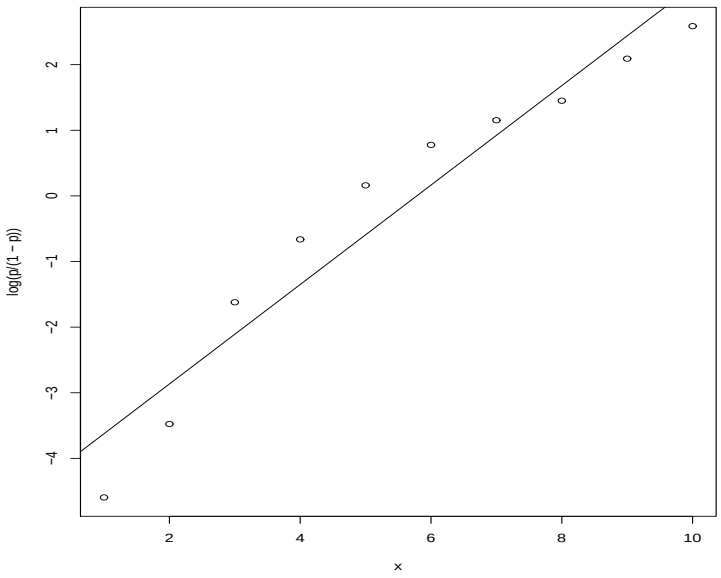
a) Luetaan havaintoaineisto:

```
> myrkky<-read.table("myrkky.dat",header=TRUE)
> myrkky
    x   n   r
1   1 200   2
2   2 200   6
3   3 200  33
4   4 200  68
5   5 200 108
6   6 200 137
```

```
7    7 200 152
8    8 200 162
9    9 200 178
10 10 200 186
```

Lasketaan kuolleiden suhteelliset osuudet $p_i = r_i/n_i$ aineistosta ja tehdään muuttujien x_i ja logit(p_i) välinen hajontakuvio. Lisätään lopuksi hajontakuvioon lineaarinen regressiosuora:

```
> attach(myrkky)
> p<-r/n
> plot(x,log(p/(1-p)))
> abline(lm(log(p/(1-p))~x))
```



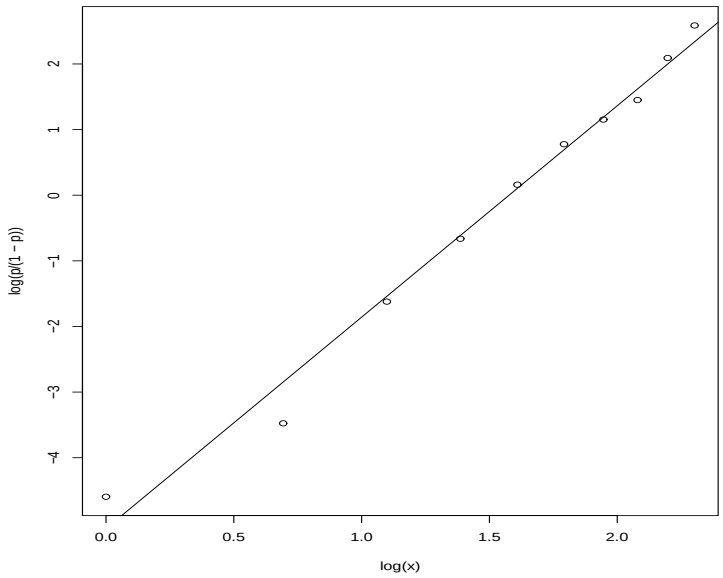
Kuva 4: Muuttujien x ja logit(p) välinen hajontakuvio

Kuviosta 4 nähdään, että logit(p) ei selvästikään riipu lineaarisesti selittävästä muuttujasta x. Tehdään muuttujalle x logaritmimuunnos ja katsotaan muuttuuko tilanne:

```
> plot(log(x),log(p/(1-p)))
> abline(lm(log(p/(1-p))~log(x)))
```

Kuviosta 5 nähdään, että logit(p) riippuu lineaarisesti muuttujasta $\log(x)$.

b) Sovitetaan aineistoon logistinen regressiomalli:



Kuva 5: Muuttujien $\log(x)$ ja $\text{logit}(p)$ välinen hajontakuvio

```
> mod1<-glm(cbind(r,n-r)~log(x),family=binomial,data=myrkky)
> summary(mod1)

Call:
glm(formula = cbind(r, n - r) ~ log(x), family = binomial, data = myrkky)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2236  -0.2805   0.1047   0.5405   0.9373

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.3425     0.2636  -20.27   <2e-16 ***
log(x)        3.3710     0.1521   22.16   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1078.5036  on 9  degrees of freedom
Residual deviance:    4.8924  on 8  degrees of freedom
AIC: 57.223
```

```
Number of Fisher Scoring iterations: 4
```

Funktiolla anova voi testata sisäkkäisiä malleja. Jos argumentissa annetaan vain yksi malli, niin anova testaa kuinka paljon malli paranee, kun selittäjiä lisätään malliin yksi kerrallaan. Muuttujat lisätään samassa järjestyksessä kuin muuttujat ovat glm-funktion kutsussa.

```
> anova(mod1,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(r, n - r)

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       9    1078.50
log(x)  1   1073.6         8       4.89 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Muuttuja $\log(x)$ on selvästi tilastollisesti merkitsevä (p-arvo$< 2.2 \cdot 10^{-16}$).

Lasketaan seuraavaksi estimaatit $\exp(\beta_1)$:lle ja $\exp(\beta_2)$:lle ja annetaan myös niiden likimääräiset 95% luottamusvälit:

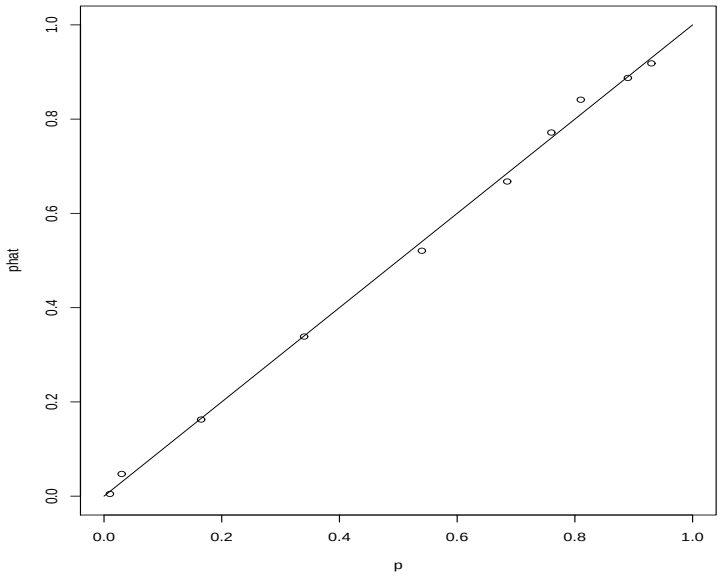
```
> exp(coefficients(mod1))
 (Intercept)       log(x)
 0.004783724 29.108350555
> exp(confint.default(mod1))
                   2.5 %       97.5 %
(Intercept)  0.002853685  0.008019113
log(x)      21.603212750 39.220836355
```

Seoksen vahvuuden logaritmin kasvaessa yhdellä yksiköllä kuolemisen odds kasvaa 29-kertaiseksi. Likimääräinen 95% luottamusväli on $(21.6, 39.2)$.

c) Tutkitaan b)-kohdassa saadun mallin sopivuutta graafisesti vertaamalla ennustettuja arvoja $\hat{p}_i$ todellisiin arvoihin $p_i = r_i/n_i$.

```
> phat<-predict(mod1,type="response")
> phat
          1           2           3           4           5           6
0.004760949 0.047159173 0.162589431 0.338652975 0.520718060 0.667639972
          7           8           9          10
0.771566209 0.841217345 0.887390926 0.918304961
```

```
> p
 [1] 0.010 0.030 0.165 0.340 0.540 0.685 0.760 0.810 0.890 0.930
> plot(p,phat,xlim=c(0,1),ylim=c(0,1))
> lines(c(0,1),c(0,1))
```



Kuva 6: Muuttujien p ja $\hat{p}$ välinen hajontakuvio

Kuvan 6 perusteella ennustetut arvot $\hat{p}_i$ ovat hyvin lähellä todellisia arvoja $p_i = r_i/n_i$.