Tampereen yliopisto
Tilastollinen mallintaminen

Mikko Alivuotila ja Anne Puustelli

Lentokoneiden rakennuksessa käytettävien metallinkiinnittimien puristuskestävyys

Matematiikan, tilastotieteen ja filosofian laitos
Tilastotiede
Marraskuu 2000

1 Aineiston kuvaus

Aineistomme on kirjasta 'Modelling Binary Data' (D. Collett, 1991) ja siinä tutkitaan lentokoneiden rakennuksessa käytettävien metallinkiinnittimien puristuskestävyyttä. Aineistossa on yhteensä 690 metallinkiinnitintä ja ne on testattu eri painekuormituksilla. Painekuormitus kasvaa 2500:sta 4300:aan 200:n yksikön (psi) välein, eli yhteensä havaintoja on tutkittu kymmenellä eri painekuormituksella. Aineistosta on pudotettu jokaisesta painekuormituksesta kaksi viimeistä nollaa pois mukavuussyistä. Aineisto on seuraava:

Taulukko 1 Aineisto

	Kuormitus ($\times 100$)	n	r	$p = \frac{r}{n}$
1	25	50	10	0.20
2	27	70	17	0.24
3	29	100	30	0.30
4	31	60	21	0.35
5	33	40	18	0.45
6	35	85	43	0.51
7	37	90	54	0.60
8	39	50	33	0.66
9	41	80	60	0.75
10	43	65	51	0.78

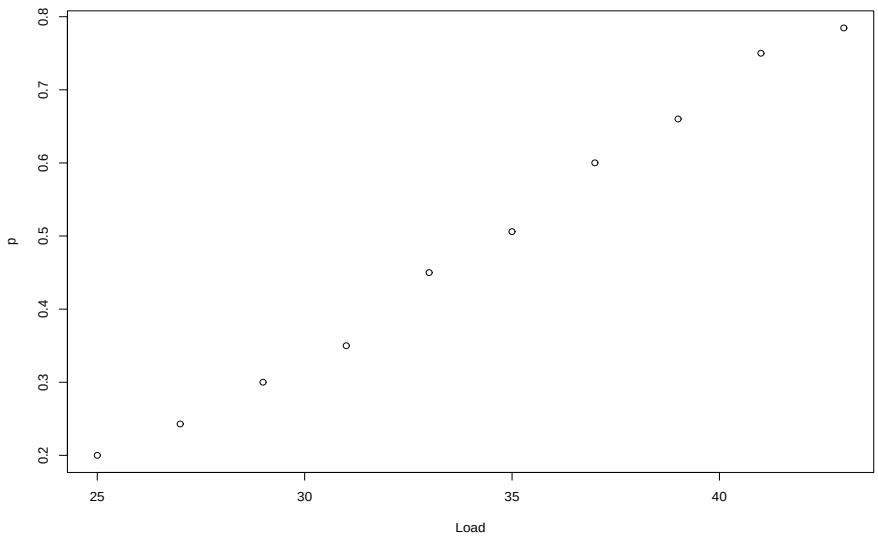
Aineistossa

n = testattujen kiinnittimien lukumäärä kussakin kuormituksessa

r = särkyneiden kiinnittimien lukumäärä kussakin kuormituksessa

p = särkyneiden kiinnittimien lukumäärä verrattuna testattujen kiinnittimien lukumäärään

Aineistossamme tutkittava muuttuja noudattaa binomijakaumaa; metallinkiinnitin joko kestää tai rikkoutuu. Kiinnostuksemme kohde on, vaikuttaako paineen vaihtelu metallinkiinnittimen rikkoutumiseen ja jos vaikuttaa, niin miten. Tätä tutkiaksemme meidän tulee rakentaa tilastollinen malli, joka kertoo miten kuormitus vaikuttaa todennäköisyyteen, että metallinkiinnitin



Kuva 1 Painekuormituksen vaikutus metallinkiinnittimen rikkoutumistodennäköisyyteen

rikkoutuu. Tällaista mallia kutsutaan lineaariseksi todennäköisyysmalliksi. Piirretään ensin kuva, jossa x-akselilla on kuormitus ja y-akselilla todennäköisyys, että metallinkiinnitin rikkoutuu.

Silmämääräisesti voimme nähdä, että matalalla paineella kiinnittimen rikkoutumisen todennäköisyys on pienempi verrattuna korkeisiin paineisiin. Todennäköisyys, että metallinkiinnitin rikkoutuu, voi saada arvoja nollasta yhteen, ja siitä syystä ei ole järkevää rakentaa lineaarista mallia. Lineaarisella mallilla todennäköisyydet saisivat arvoja $[-\infty, \infty]$, joka ei ole tulkinnallisesti järkevää. Tästä syystä päätämme mallintaa aineistoa logistisella regressiomallilla.

2 Johdatus yleistettyihin lineaarisiin malleihin

Logistinen regressiomalli kuuluu yleistettyjen lineaaristen mallien perheeseen. Yleistetyt lineaariset mallit muodostuvat kolmesta komponentista: satunnaiskomponentista, systemaattisesta komponentista ja linkkifunktiosta. Satunnaiskomponentissa määritetään selitettävä muuttuja Y_i ja tämän to-

dennäköisyysjakauma. Systemaattisessa komponentissa määritetään, millaisella funktiolla x_j:t selittävät selitettävää muuttujaa. Yksinkertaisin systemaattinen komponentti on muotoa

$$\alpha + \beta_1 x_1 + ... + \beta_k x_k. \tag{1}$$

Linkkifunktio määrittelee, miten $\mu = E(Y)$ on yhteydessä selittäviin muuttujiin. Käytämme aineistomme analysoimisessa kolmea linkkifunktiota. Yleisin näistä on logit linkki, joka määritetään

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right). \tag{2}$$

Sulkujen sisällä oleva murtoluku on onnistumisen mahdollisuus (odds) ja logit linkki on siten logaritmi onnistumisen mahdollisuudesta. Koska p saa arvoja väliltä (0,1), on helppo nähdä, että kun $p \to 0$, niin $\text{logit}(p) \to -\infty$. Samoin, kun $p \to 1$, niin $\text{logit}(p) \to \infty$. Logit($p$) funktio on muodoltaan hiukan tangentti- käyrän muotoinen, jossa p saa x-akselilla arvot väliltä (0,1). Pisteessä $p = 0.5$ $\text{logit}(p) = 0$ ja funktio on symmetrinen tämän pisteen suhteen.

Probit linkki määritellään seuraavasti

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\xi} \exp\left(-\frac{1}{2}u^2\right) du = p\,, \tag{3}$$

missä $U \sim N(0,1)$, ja siten $p = P(u \leq \xi)$. Standardoitua normaalijakaumaa merkitään usein $\Phi(\xi)$, joten ξ on sellainen, että $\Phi(\xi) = p$. Tällöin $\xi = \Phi^{-1}(p)$, missä käänteisfunktio $\Phi^{-1}(p)$ on probit transformaatio p:lle. Myös probit(p) saa arvoja $(-\infty, \infty)$ ja on symmetrinen pisteen $p = 0.5$ suhteen. Kyseisessä pisteessä $\text{probit}(p) = 0$, koska normaalijakauman kertymäfunktio saavuttaa pisteen 0, kun $p = 0.5$. Logit ja probit linkit antavat hyvin samanlaisia tuloksia ja on erittäin harvinaista, että toisella linkillä malli olisi hyvä ja toisella huono.

Komplementaarinen log-log linkki määritellään seuraavasti

$$\text{cloglog}(p_i) = \log[\log(1-p_i)] \tag{4}$$

Myös tämä linkki transformoi todennäköisyydet välille $(-\infty, \infty)$, mutta ei ole symmetrinen pisteen $p = 0.5$ suhteen kuten kaksi aiempaa linkkifunktiota ja $\text{cloglog}(0.5) < 0$. Pienillä p:n arvoilla komplementaarinen log-log linkki antaa lähes samat arvot kuin logit linkki.

3 Logistinen regressiomalli

Oletetaan, että aineistossamme

$$r_i \sim \text{Bin}(n_i, \pi_i), \tag{5}$$

missä $i = 1, \ldots, 10$ ja $r_1, \ldots, r_{10}$ ovat riippumattomia. Mallinnetaan parametrien π_i riippuvuus kuormituksesta, jota merkitään x_i. Käytetään ensin logit linkkiä, jolloin malli on muotoa

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \text{logit}(\pi_i) = \alpha + \beta x_i. \tag{6}$$

Ratkaistaan yhtlöstä π_i:n ja saadaan

$$\pi_i = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}. \tag{7}$$

Tämän avulla voimme siis määrittää π_i:t eri x_i:n arvoilla, kun olemme estimoineet α:n ja β:n.

Parametrien estimaatit johdetaan suurimman uskottavuuden menetelmällä. Olkoon k selittävien muuttujien lukumäärä. Uskottavuusfunktio on muotoa

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i} \tag{8}$$

Uskottavuusfunktion logaritmi on

$$\begin{aligned}
\log L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left\{ \log\binom{n_i}{y_i} + y_i \log p_i + (n_i - y_i)\log(1-p_i) \right\} \\
&= \sum_{i=1}^{n} \left\{ \log\binom{n_i}{y_i} + y_i \log\left(\frac{p_i}{1-p_i}\right) + n_i \log(1-p_i) \right\} \\
&= \sum_{i=1}^{n} \left\{ \log\binom{n_i}{y_i} + y_i\eta_i - n_i \log(1+e^{\eta_i}) \right\},
\end{aligned} \tag{9}$$

missä $\eta_j = \sum_{j=0}^{k} \beta_j x_{ij}$ ja $x_{0i} = 1$ kaikilla i:n arvoilla. Logaritmoidun uskottavuusfunktion osittaisderivaatta parametrin $\boldsymbol{\beta}$:n suhteen on

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j} = \sum y_i x_{ji} - \sum n_i x_{ji} e^{\eta_i}(1+e^{\eta_i})^{-1}, \tag{10}$$

missä $j = 0, 1, \ldots, k$. Merkitsemällä tämä yhtälö nollaksi ja ratkaisemalla $\hat{\boldsymbol{\beta}}$ saadaan $k+1$ epälineaarista yhtälöä, joissa on tuntematon parametri $\hat{\beta}_j$. Näille yhtälöille on olemassa ainoastaan numeerinen ratkaisu. Yleisesti käytetty algoritmi $\hat{\boldsymbol{\beta}}$ suurimman uskottavuuden estimaattien ratkaisuun on Fisherin menetelmä (Fisher's method of Scoring), joka vastaa iteratiivista painotettua pienimmän neliösumman menetelmää. Fisherin menetelmä on sama kuin yleisesti käytetty Newton-Rapson algoritmi, mutta lähestyy ongelmaa informaatiomatriisin avulla.

4 Mallin hyvyys

Mallin hyvyyttä mitataan devianssilla (D). Se saadaan estimoidun ja täyden mallin suurimman uskottavuuden suhteesta kaavalla

$$D = -2 \log \left(L_c / L_f \right), \tag{11}$$

missä c viittaa estimoituun malliin ja f viittaa täyteen malliin. D saa suuria arvoja kun L_c on suhteellisen pieni verrattuna L_f:ään. Tällöin malli on huono. Kun L_c lähenee L_f:ää, saa D pieniä arvoja ja malli on tällöin hyvä.

Havaintojen ollessa binomijakautuneita, devianssi lähenee asymptoottisesti χ^2 jakaumaa vapausastein $n - p$, missä n on havaintojen määrä ja p on mallissa olevien tuntemattomien parametrien määrä. Jatkuvien muuttujien tapauksessa, kun oletuksena on normaalijakuma, devianssi on jäännösneliösumma. Tässä tapauksessa devianssi on χ^2:n monikerta, riippumatta otoskoosta.

Yksittäisten havaintojen määrä vaikuttaa otoskoon suuruuden toteamiseen, ei niinkään eri binomitodennäköisyyksien määrä. Esimerkissämme metallikiinnittimien määrä ratkaisee, koska voidaan olettaa otoksen olevan tarpeeksi suuri χ^2 approksimointiin, eikä eri painekuomitusluokkien määrä. Eli vaikka luokkien määrä olisi pieni, devianssin jakaumaan voidaan käyttää χ^2 approksimointia, jos havaintojen määrä on riittävän suuri. "Riittävästä suuruudesta" voidaan sanoa, että se riippuu enemmän havaintojen määristä eri luokissa ja estimoitavista luokkatodennäköisyyksistä kuin havaintojen kokonaismäärästä.

Erityistapauksessa, jossa binomijakautuneita havaintoja ei ole luokiteltu, devianssi ei ole edes approksimoiden χ^2 jakautunut. Sama ongelma esiintyy, jos joissakin luokissa on vain muutamia havaintoja tai mallin sovitteet ovat lähellä nollaa tai ykköstä.

Nyrkkisääntönä voidaan sanoa, että jos devianssi on pienempi kuin vapausasteensa tai lähellä sitä, malli on tyydyttävä.

5 Logistisen regressiomallin estimointi

Estimoidaan logistinen regressiomalli komennolla

```
ex6 <- glm(p ~ Load,weights=n,family=binomial)
```

Tulokseksi saadaan

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.29475  -0.11129   0.04162   0.08847   0.35016

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.33971    0.54569  -9.785   <2e-16 ***
Load         0.15484    0.01575   9.829   <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: 112.83207  on 9  degrees of freedom
Residual deviance:   0.37192  on 8  degrees of freedom
AIC: 847.71

Number of Fisher Scoring iterations: 3
```

Logistiseksi regressiomalliksi saadaan

$$\text{logit}\,(\hat{p}) = -5.340 + 0.1548\,Load \tag{12}$$

Tämän avulla voidaan määrittää kiinnittimen hajoamisen todennäköisyys eri painekuormituksilla.

$$\hat{p_i} = \frac{\exp\left(-5.340 + 0.1548 Load_i\right)}{1 + \exp\left(-5.340 + 0.1548 Load_i\right)} \tag{13}$$

Esimerkiksi kuormituksella 25 psi saadaan kiinnittimen hajoamisen todennäköisyydeksi $\hat{p} = 0.1877$. Tämä on hyvin lähellä havaittua arvoa kyseisellä kuormituksella, joka on $p = 0.20$.

Normaalijakaumaan perustuvalla testauksella voidaan todeta, että mallin selittäjän kerroin eroaa merkittävästi nollasta ($p < 2e - 16$). Tämä merkitsee sitä, että kiinnittimen hajoamistodennäköisyyden ja kuormituksen välillä on merkittävä yhteys. Mallin devianssi on 0.3719 ja sen vapausasteet ovat $10 - 2 = 8$. Testattaessa mallin hyvyyttä 0.5%:n riskitasolla verrataan devianssia χ^2-jakauman taulukkoarvoon $\chi^2_{8,\,0.005} = 21.96$. Taulukoitu arvo on huomattavasti suurempi kuin havaittu arvo, joten H_0 ='malli on hyvä' voidaan hyväksyä. Tämä tarkoittaa, että painekuormituksella voidaan estimoida kiinnittimen hajoamistodennäköisyys.

6 Vertailumallien estimointi

Estimoidaan seuraavaksi probit regressiomalli, jotta voidaan vertailla eri linkkifunktioiden vaikutusta.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.33473  -0.11319   0.01760   0.09600   0.36474

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.27121    0.32128  -10.18   <2e-16 ***
Load         0.09488    0.00928   10.23   <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: 112.83207  on 9  degrees of freedom
Residual deviance:   0.43735  on 8  degrees of freedom
AIC: 847.78

Number of Fisher Scoring iterations: 2
```

Probit regressiomalliksi saadaan

$$\text{probit}(\hat{p}) = -3.271 + 0.949\, Load \tag{14}$$

Mallin selittäjän kerroin eroaa merkittävästi nollasta ($p < 2e-16$), joka antaa saman päätelmän kuin edellisessä mallissa. Mallin devianssi on 0.4374 ja sen vapausasteet ovat $10 - 2 = 8$. Tämä tarkoittaa, että painekuormituksella voidaan estimoida kiinnittimen hajoamistodennäköisyys myös probit regressiomallia käyttäen.

Estimoidaan lopuksi komplementaarinen log-log malli.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.55154  -0.16444  -0.03201   0.19367   0.39434

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.19175    0.39453  -10.63   <2e-16 ***
Load         0.10941    0.01084   10.10   <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: 112.8321  on 9  degrees of freedom
Residual deviance:   0.6867  on 8  degrees of freedom
AIC: 848.03

Number of Fisher Scoring iterations: 3
```

Komplementaariseksi log-log malliksi saadaan

$$\text{cloglog}\,(\hat{p}) = -4.191 + 0.109\, Load \tag{15}$$

Mallin selittäjän kerroin eroaa merkittävästi nollasta ($p < 2e-16$), joka antaa saman päätelmän kuin edellisessä mallissa. Mallin devianssi on 0.6867 ja sen vapausasteet ovat $10 - 2 = 8$. Tämä tarkoittaa, että painekuormituksella voidaan estimoida kiinnittimen hajoamistodennäköisyys myös komplementaarista log-log mallia käyttäen.

7 Mallien vertailu

Verrataan estimoituja malleja devianssin ja Akaiken informaatiokriteerin avulla. Vertailu on mahdollista, koska kaikissa malleissa on yhtä paljon estimoitavia parametreja. Malleja vertailtaessa haetaan pienintä AIC:n arvoa. Pienin arvo saadaan logistisella regressiomallista (847.71), mutta probit ja komplementaarinen log-log malli antavat lähes samat AIC:n arvot. Mallien devianssit ovat myös hyvin lähellä toisiaan ja parhaimmaksi malliksi osoittautuu jälleen logistinen regressiomalli pienimmällä devianssilla 0.37192.

Tarkasteltaessa mallien parametrien estimaatteja huomaamme, että logistisella ja probit mallilla estimaattien suhde on sama. Siis

$$\frac{\alpha_1}{\beta_1} = \frac{\alpha_2}{\beta_2} = -34.48 \tag{16}$$

Tämä seuraa linkkifunktioiden ominaisuuksista, joita olemme esitelleet kappaleessa 2. Komplementaarisen log-log mallin estimoitujen parametrien suhde sitävastoin ei ole sama edellisten mallien kanssa. Tämä johtuu siitä, että linkki ei ole symmetrinen pisteen $p = 0.5$ suhteen ja $\text{cloglog}(0.5) < 0$. Tästä syystä komplementaarisen log-log mallin parametrien suhde on pienempi kuin edellisten mallien.