

Ch 9: MAXIMUM LIKELIHOOD ANALYSIS AND ROBUST REGRESSION

9.1 Fitting expectations using iterative re-weighting

1. (Def) Let y_1, \dots, y_n be independent observations,

$$E y_i = \mu_i(\theta)$$

$$\text{var } y_i = \sigma_i^2(\theta)$$

Fit the $\mu_i(\theta)$ to the y_i with weights $w = \sigma_i^{-2}(\theta)$. Specifically, apply the GN algorithm with:

- (a) $f_i(\theta) = \mu_i(\theta)$ and $w_i = \sigma_i^{-2}(\theta)$
- (b) Up-date the weights on each iteration

Maximum likelihood analysis (Review)

Given a family of densities

$$f(y|\theta) \ , \ \theta \in \Theta$$

for a random vector y let

$$L(\theta) = f(y|\theta) = \text{likelihood function given } y$$

$$s(\theta) = \nabla \log L(\theta) = \text{score vector}$$

$$\mathcal{I}(\theta) = \text{cov } s(\theta) = \text{Fisher information matrix}$$

$$\hat{\theta} = \text{value of } \theta \text{ that maximizes } L(\theta) = \text{max. lik. est.}$$

$$\nabla \theta = \mathcal{I}^{-1}(\theta)s(\theta) = \text{Fisher scoring algorithm}$$

Under appropriate assumptions

$$\hat{\theta} \stackrel{\mathcal{Q}}{\sim} N(\theta, \mathcal{I}^{-1}(\theta))$$

Motivated by this let

$$\widehat{\text{cov}} \hat{\theta} = \mathcal{I}^{-1}(\hat{\theta})$$

The exponential family

1. (Def) $f(y|\theta)$, $\theta \in \Theta$ is an exponential family if

$$f(y|\theta) = g(\theta)h(y)e^{\pi(\theta)y}$$

2. (Poisson)

$$\begin{aligned} f(y|\mu) &= e^{-\mu} \mu^y / y! \\ &= e^{-\mu} (1/y!) e^{\log(\mu)y} \\ &= g(\mu)h(y)e^{\pi(\mu)y} \\ &= \text{exp family} \end{aligned}$$

3. (Binomial)

$$f(y|p) = \binom{n}{y} p^y (1-p)^{n-y} = \text{exp fmy (Prob 7)}$$

4. (Th) If $f_1(y_1|\theta)$ and $f_2(y_2|\theta)$ are exponential families, then

$$f(y_1, y_2|\theta) = f_1(y_1|\theta)f_2(y_2|\theta)$$

is an exponential family.

5. (Ex) If y_1, \dots, y_n are independent Poisson variables, then

$$f(y_1, \dots, y_n) = \text{exp fmy}$$

General results for the exponential family

1. (Notation)

$f(y|\theta)$, $\theta \in \Theta$ is an exponential family

$$\mu = \mu(\theta) = E y$$

$$\Sigma = \Sigma(\theta) = \text{cov } y$$

2. (Th)

$$s(\theta) = \frac{d\mu^T}{d\theta} \Sigma^- (y - \mu)$$

$$\mathcal{I}(\theta) = \frac{d\mu^T}{d\theta} \Sigma^- \frac{d\mu}{d\theta}$$

where Σ^- is any generalized inverse of Σ .

Pf: Jennrich and Moore (1975)

3. (Def) Σ^- is a generalized inverse of Σ if

$$\Sigma \Sigma^- \Sigma = \Sigma$$

If Σ is invertible, then $\Sigma^- = \Sigma^{-1}$

4. (Note) With $f = \mu$ and $W = \Sigma^-$,

GN becomes FS

Pf:

$$\begin{aligned}\nabla\theta_{\text{GN}} &= \left(\frac{d\mu^T}{d\theta} \Sigma^{-} \frac{d\mu}{d\theta} \right)^{-1} \frac{d\mu^T}{d\theta} \Sigma^{-} (y - \mu) \\ &= \mathcal{I}^{-1}(\theta) s(\theta) = \nabla\theta_{\text{FS}}\end{aligned}$$

5. (Note) Also

$$\widehat{\text{cov}}_{\text{GN}}(\hat{\theta}) = \hat{\sigma}^2 \widehat{\text{cov}}_{\text{ML}}(\hat{\theta})$$

Pf:

$$\begin{aligned}\widehat{\text{cov}}_{\text{GN}}(\hat{\theta}) &= \hat{\sigma}^2 \left(\frac{d\mu^T}{d\theta} \Sigma^{-} \frac{d\mu}{d\theta} \right)^{-1}_{\hat{\theta}} \\ &= \hat{\sigma}^2 \mathcal{I}^{-1}(\hat{\theta}) = \hat{\sigma}^2 \widehat{\text{cov}}_{\text{ML}}(\hat{\theta})\end{aligned}$$

6. (What do these mean?) If one:

- (a) Iteratively re-weights using $W = \Sigma^{-}$
- (b) Sets $\hat{\sigma}^2 = 1$

then

- (a) The GN and ML estimates are the same.
- (b) The GN standard errors are the ML theory standard errors.

Exponential family tests

1. (Pearson's χ^2 statistic)

$$\chi^2 = \sum (y_i - \hat{\mu}_i)^2 / \hat{\sigma}_i^2$$

Note

$$\chi^2 = \text{RSS}$$

where RSS is from the iteratively re-weighted Gauss-Newton algorithm.

2. (Pearson's χ^2 test) If χ^2 and χ_0^2 are from full and restricted models with p and q free parameters, in many situations

$$\chi_0^2 - \chi^2 \stackrel{0}{\sim} \chi^2(p - q)$$

Pf: Asymptotics.

3. (Note) This is the fundamental χ^2 test. Unlike the fundamental F test there is no denominator because there is no scale parameter σ .

Poisson models

1. (Note) These models arise in for example tracer experiments, fish catches, and disease counts.
2. (Model) y_i independent $P(\mu_i(\theta))$, $i = 1, \dots, k$
3. (Gauss-Newton set-up) $f_i = \mu_i$, $w_i = \mu_i^{-1}$
4. (Note) Using μ_i^{-1} and up-dating rather than using \hat{y}_i^{-1} (from a least squares fit) gives true MLE.
5. (Goodness of fit) The Pearson χ^2 test for $\mu = \mu(\theta)$ verses μ free is:

$$\text{RSS} \stackrel{\circ}{\sim} \chi^2(k - p)$$

provided the model $\mu = \mu(\theta)$ is correct.

Pf: The test is

$$\chi_0^2 - \chi^2 \stackrel{\circ}{\sim} \chi^2(k - p)$$

For the unrestricted μ free model $\chi^2 = 0$. For the restricted model $\chi_0^2 = \text{RSS}$. Thus

$$\text{RSS} = \chi_0^2 - \chi^2 \stackrel{\circ}{\sim} \chi^2(k - p)$$

6. (Note) The previous result becomes exact as all $\mu_i \rightarrow \infty$ with k fixed and holds well already when all $\mu_i \geq 5$.

7. (Ex 9.1, p289) Formula data Table 9.1

$x = \#$ pages

$y = \#$ formulas

$y \sim P(\mu)$

$\mu = \beta_0 + \beta_1 x + \beta_2 x^2$

GN: $f = \mu$, $w = \mu^{-1}$

PROC NLIN details (Fig 9.1)

From Fig 9.2:

$$\text{RSS} = 12.50 < 15.5 = \chi_{.05}^2(8) \quad , \quad (\text{accept model})$$

Could the model actually be a Poisson process?

$$\mu = \beta x$$

$$\text{RSS}_0 - \text{RSS} \stackrel{0}{\sim} \chi^2(3 - 1)$$

This is Prob 1.

8. (Warning) In PROC NLIN the confidence intervals are incorrect when σ^2 is known. Then

$$df(\hat{\sigma}^2) = \infty$$

and not $k - p$ as NLIN assumes. One needs to use the reference distribution $N(0, 1)$ and not $t(k - p)$. This makes a substantial difference when k is small as happens fairly often in Poisson applications.

Log-linear Poisson models

1. (Def) A Poisson model with

$$\mu_i = e^{x_i\beta}$$

$$x_i\beta = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p$$

This is called a log-linear model because $\log \mu_i = x_i\beta$.

2. The log-linear model corresponding to the formula data model is

$$\mu = e^{\beta_0 + \beta_1 x + \beta_2 x^2}$$

In terms of the general model

$$x_1 = 1, \quad x_2 = x, \quad x_3 = x^2$$

3. (Fig 9.3) This is code for the general log-linear Poisson model.

The user modifies only lines 2 & 3 and of course the data.

Binomial models

1. (Note) The primary application here is to logistic regression. Consider for example the number y_i of rats in a sample of n_i that develop tumors when exposed to a carcinogen at intensity x_i .

2. (Binomial model)

$$y_i \sim B(n_i, \pi_i(\theta)) \quad , \quad \text{independent} \quad i = 1, \dots, k$$

3. (GN set-up) $f_i = \mu_i$, $w_i = \sigma_i^{-2}$

$$\mu_i = n_i \pi_i \quad , \quad \sigma_i^2 = n_i \pi_i (1 - \pi_i)$$

4. (Logistic model for π)

$$\pi_i(\theta) = \frac{e^{\alpha + \beta x_i}}{1 + e^{\alpha + \beta x_i}} \quad , \quad \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \quad , \quad x_i = \text{ind. var.}$$

5. (Picture)

6. (Note) The value that gives a 50% probability of a positive response is

$$\text{ED50} = -\alpha/\beta$$

$$\text{Pf: } x = \text{ED50} \Rightarrow \alpha + \beta x = 0 \Rightarrow x = -\alpha/\beta$$

7. (Note)

$$\alpha + \beta x = \log \left(\frac{\pi}{1 - \pi} \right) = \log \text{ odds ratio}$$

Pf: Let $u = \alpha + \beta x$. Then

$$\pi = \frac{e^u}{1 + e^u}$$

$$e^u = \frac{\pi}{1 - \pi}$$

$$\alpha + \beta x = u = \log \frac{\pi}{1 - \pi}$$

8. (Goodness of fit) $\pi = \pi(\theta)$ verses π free.

$$\text{RSS} \stackrel{o}{\sim} \chi^2(k - p)$$

This becomes exact as all $n_i \rightarrow \infty$ and is a good approximation when all

$$n_i \pi_i \geq 5 \quad , \quad n_i(1 - \pi_i) \geq 5$$

That is when the expected numbers of successes and failures are all at least 5.

9. (Ex 9.2, p295) Frog data, Table 9.2

From the output (not shown)

$$\text{RSS} = 4.09 < 7.82 = \chi_{.05}^2(3)$$

The logistic model fits

$$\hat{\alpha} = -10.66 \quad , \quad \hat{\beta} = 7.30$$

$$\widehat{\text{ED}}50 = -\hat{\alpha}/\hat{\beta} = 1.460$$

For $\widehat{\text{std}}(\widehat{\text{ED}}50)$ add data (0,1,..) and add code:

```
IF Y=. THEN F=-A/B;
```

```
MODEL Y=F;
```

Multinomial models

1. (Note) These arise when the response is categorical as for example when the response is blood type or race.

2. (Model)

$$(y_1, \dots, y_k) \sim M(n, (\pi_1, \dots, \pi_k))$$

Here

$$y_1 + \dots + y_k = n$$

$$\pi_1 + \dots + \pi_k = 1$$

3. (Recall)

$$\mu_i = E y_i = n\pi_i$$

$$\sigma_{ij} = \text{cov}(y_i, y_j) = n\pi_i(\delta_{ij} - \pi_j)$$

4. (Note) $\Sigma = (\sigma_{ij}) = \text{cov } y$ is singular.

5. (Th) If

$$W = \begin{pmatrix} \mu_1^{-1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mu_k^{-1} \end{pmatrix}$$

then W is a generalized inverse of Σ and is diagonal.

Pf: Using matrix multiplication and some effort one can show $\Sigma W \Sigma = \Sigma$.

6. (GN set-up) Using the previous result the GN set-up is

$$f_i = n\pi_i$$

$$w_i = f_i^{-1} \quad , \quad \text{as for Poisson}$$

7. (Goodness of fit) $\pi = \pi(\theta)$ verses π free.

$$RSS \stackrel{0}{\sim} \chi^2(k - 1 - p)$$

This becomes exact as $n \rightarrow \infty$ and is a good approximation when all $n\pi_i \geq 5$.

8. (Ex 9.3, p298) Blood type data, Table 9.3

$p, q, r = A, B, O$ gene frequencies

$$p + q + r = 1$$

$p, q =$ free parameters

$$r = 1 - p - q$$

$\pi_1, \dots, \pi_4 = O, A, B, AB$ blood type frequencies

$y_1, \dots, y_4 =$ blood type counts

Output:

$$\hat{p} = .2544$$

$$\hat{q} = .0932$$

$$\text{RSS} = 1.376 < \chi_{.05}^2(1)$$

The standard genetic model fits.

Log-linear multinomial models

1. (Def)

$$\pi_i(\beta) = e^{\nu + x_i\beta}$$

$$\nu = -\log \sum_{i=1}^k e^{x_i\beta}$$

The second condition is required because the π_i must sum to one.

2. (Note) $\log \pi_i = \nu + x_i\beta =$ almost a linear model.

Almost because ν is actually a non-linear function of β .

3. (SAS problem) We can't use PROC NLIN because of the sum in the definition of ν . The general code on page 302 no longer works because RETAIN has been re-defined by SAS.

4. (Poisson trick) Pretend

$$y_i \sim P(\mu_i) \quad , \quad \text{indep} \quad i = 1, \dots, k$$

$$\mu_i = e^{\alpha + x_i\beta} \quad , \quad \alpha, \beta \text{ free}$$

Relative to the multinomial model:

- (a) The values of $\hat{\beta}$, $\hat{\mu}$, and \hat{e} are unchanged

(b) RSS is unchanged

(c) All standard errors are correct except those involving $\hat{\alpha}$, e.g. those for \hat{y}_i and \hat{e}_i

Pf: A course on log-linear models.

5. (Ex) Two-way tables of multinomial counts

$$(y_{ij}) \sim M(n, (\pi_{ij}))$$

$$i, j \text{ independent} \Leftrightarrow \pi_{ij} = e^{\nu + \alpha_i + \beta_j}$$

$$\text{w.l.g. } \alpha_1 = \beta_1 = 0 \quad (\text{makes } \alpha_i \text{ and } \beta_j \text{ unique})$$

6. (Ex 9.4, p301) Cholesterol data, independence model

$$\pi_{ij} = e^{\gamma + \alpha_i + \beta_j}$$

With the Poisson trick

$$\mu_{ij} = e^{\gamma + \alpha_i + \beta_j}, \quad \gamma \text{ free}$$

$$\theta = (\gamma, \alpha_2, \alpha_3, \alpha_4, \beta_2, \beta_3, \beta_4)$$

To use our general program we must write

$$\begin{aligned} \gamma + \alpha_i + \beta_j &= x_{ij}\theta \\ &= x_{ij}^{(1)}\gamma + x_{ij}^{(2)}\alpha_2 + \cdots + x_{ij}^{(7)}\beta_4 \end{aligned}$$

The following table shows how to generate values for $x^{(1)}, \dots, x^{(7)}$

i	j	$\gamma + \alpha_i + \beta_j$	γ	α_2	α_3	α_4	β_2	β_3	β_4
1	1	γ	1	0	0	0	0	0	0
1	2	$\gamma + \beta_2$	1	0	0	0	1	0	0
		\vdots							
2	3	$\gamma + \alpha_2 + \beta_3$	1	1	0	0	0	1	0
		\vdots							
4	4	$\gamma + \alpha_4 + \beta_4$	1	0	0	1	0	0	1

- Input: Fig 9.7, p305
- Output: (Not shown)

$$RSS = 20.85 > 16.92 = \chi_{.05}^2(16 - 1 - 6)$$

Reject independence of cholesterol and blood pressure.

- Quasi-independence: The residual for cell (1,1) is large.
With this cell removed $k = 15$ and

$$RSS = 12.29 < 15.51 = \chi^2(15 - 1 - 6)$$

Accept quasi-independence.

- Only the second application requires the computer because for the complete model there is a closed form formula

$$\hat{y}_{ij} = y_{i+}y_{+j}/y_{++}$$

- Closed form formulas are rare.

Exponential survival models

1. (Def) Let $\text{Ex}(\lambda)$ denote the distribution with density

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

Here

$$\lambda = \text{hazard rate} = \text{risk per unit time}$$

2. (Model)

$$y_i \sim \text{Ex}(\lambda_i(\theta)) \text{ indep } i = 1, \dots, k$$

Recall

$$\mu_i = \lambda_i^{-1} \quad , \quad \sigma_i^2 = \mu_i^2 = \lambda_i^{-2}$$

3. (GN set-up)

$$f_i = \lambda_i^{-1} \quad , \quad w_i = \lambda_i^2 = f_i^{-1}$$

4. (Note) There is no goodness of fit test.

$$RSS \stackrel{o}{\not\sim} \chi^2(k-p)$$

but

$$\hat{\theta} \stackrel{o}{\sim} N(\theta, \mathcal{I}^{-1}(\theta))$$

when k is large.

5. (Ex, Prob 17) Leukemia data

$$e y_i = \alpha e^{\beta(x_i - \bar{x})}$$

Go over.

Robust estimation

We are concerned with departures from normality, primarily outliers and long tailed residual distributions.

Location

1. (Median)
 - (a) The median is a robust estimate of location.
 - (b) The mean is not.
2. (Ex) Let C denote the Cauchy distribution. The means and medians for 10 samples of 100 from $C(0, 1)$ were:

sample	mean	median
1	-.04	.24
2	.14	.20
3	-18.47	.20
4	-.66	-.12
5	-.77	.07
6	-10.08	.10
7	1.43	.14
8	3.90	-.30
9	108.11	-.06
10	.67	.03

When sampling from the Cauchy the mean is a very poor estimate of location. Recall that

$$X_1, \dots, X_n \text{ indep } C(0, 1) \Rightarrow \bar{X} \sim C(0, 1)$$

3. (Note) We need an approach better than throwing out outliers.

4. (M-estimation) $\hat{\theta}$ is a solution to

$$\sum_{i=1}^n \psi(y_i - \hat{\theta}) = 0$$

5. (Note) Formulas for ψ are given on page 308.
6. (Scale) To adjust for scale we re-define the M-estimate using

$$\sum \psi \left(\frac{y_i - \hat{\theta}}{ks} \right) = 0$$

where

s = a robust estimate of scale

= MAD = median $|y_i - \hat{\theta}|$

k = 1, 2, 3 for Huber

k = 6, 9 for Bi-square

Robust regression

1. (Model)

$$y_i = f_i(\theta) + e_i$$

$f_i(\theta)$ may be linear.

2. (Least squares) Recall that for least squares $\hat{\theta}$ satisfies

$$\sum_{i=1}^n \frac{\partial f_i}{\partial \theta_j} (y_i - f_i(\theta)) = 0$$

3. (M-estimation) $\hat{\theta}$ satisfies

$$\sum_{i=1}^n \frac{\partial f_i}{\partial \theta_j} \psi \left(\frac{y_i - f_i(\theta)}{ks} \right) = 0$$

4. (Iterative re-weighting) Solve with respect to θ

$$\sum_{i=1}^n \frac{\partial f_i}{\partial \theta_j} w_i (y_i - f_i(\theta)) = 0$$

$$w_i(u_i) = \psi(u_i)/u_i$$

$$u_i = (y_i - f_i(\theta))/ks(\theta)$$

using the GN algorithm in iteratively re-weighted form.

Why it works:

$$\begin{aligned}\sum_{i=1}^n \frac{\partial f_i}{\partial \theta_j} \psi \left(\frac{y_i - f_i(\theta)}{ks} \right) &= 0 \\ \sum_{i=1}^n \frac{\partial f_i}{\partial \theta_j} \frac{\psi \left(\frac{y_i - f_i(\theta)}{ks} \right)}{\frac{y_i - f_i(\theta)}{ks}} \frac{y_i - f_i(\theta)}{ks} &= 0 \\ \sum_{i=1}^n \frac{\partial f_i}{\partial \theta_j} \frac{\psi(u_i)}{u_i} (y_i - f_i(\theta)) &= 0 \\ \sum_{i=1}^n \frac{\partial f_i}{\partial \theta_j} w_i(\theta) (y_i - f_i(\theta)) &= 0\end{aligned}$$

5. (Down weighting) Recall

$$w(u) = \psi(u)/u$$

Huber:

Bisquare:

6. (Ex 9.6, p310) Stack loss data, Table 9.6

Model: $y = \alpha + \beta_1 x_1 + \cdots + \beta_3 x_3 + e$

$\psi =$ Huber

$$w(u) = \begin{cases} 1 & |u| \leq 1 \\ 1/|u| & |u| > 1 \end{cases}$$

$$u = \frac{y-f}{ks} \quad , \quad s = \text{MAD} \quad , \quad k = 1, 2, 3$$

First find a least squares fit. This gives starting values for

$$\alpha, \beta_1, \beta_2, \beta_3$$

MAD

Swindle for $s(\theta)$: Up-dating $s(\theta)$ on each iteration is difficult. Instead use:

Least squares MAD = 1.9 (Fig 9.10)

Fixed $ks =$ least squares MAD (k=1 at start)

Fig 9.11 Input p313

Fig 9.10 Note outliers, note bi-square

The four outliers are times when the tower was not lined-out.

Back to the swindle:

$$ks = 2$$

$$s = \text{MAD} = 1.3 \quad (\text{Fig 9.10})$$

Thus

$$k = 2/1.3 = 1.5 \quad (\text{at the end})$$

This is equivalent to using $k = 1.5$ and computing MAD every iteration. This is a reasonable value for k . Other values can be obtained by using other fixed values for ks .