Random Effect Modelling for Regression Models with Gamma-Distributed Response

Jochen Einbeck and John Hinde

National University of Ireland, Galway

CASI, 19th May 2006

Motivating Example: Pennsylvanian Hospital-Stay-Data

Duration stay, age, and temperature at admission (Rosner, 2000)

ID	duration	age	temp1	ID	duration	age	temp1	ID	duration	age	temp1
1	5	30	99.0	10	3	50	98.0	18	4	69	98.0
2	10	73	98.0	11	9	59	97.6	19	3	47	97.0
3	6	40	99.0	12	3	4	97.8	20	7	22	98.2
4	11	47	98.2	13	8	22	99.5	21	9	11	98.2
5	5	25	98.5	14	8	33	98.4	22	11	19	98.6
6	14	82	96.8	15	5	20	98.4	23	11	67	97.6
7	30	60	99.5	16	5	32	99.0	24	9	43	98.6
8	11	56	98.6	17	7	36	99.2	25	4	41	98.0
9	17	43	98.0								

Aim: Modelling (Prediction) of duration given temperature and age.

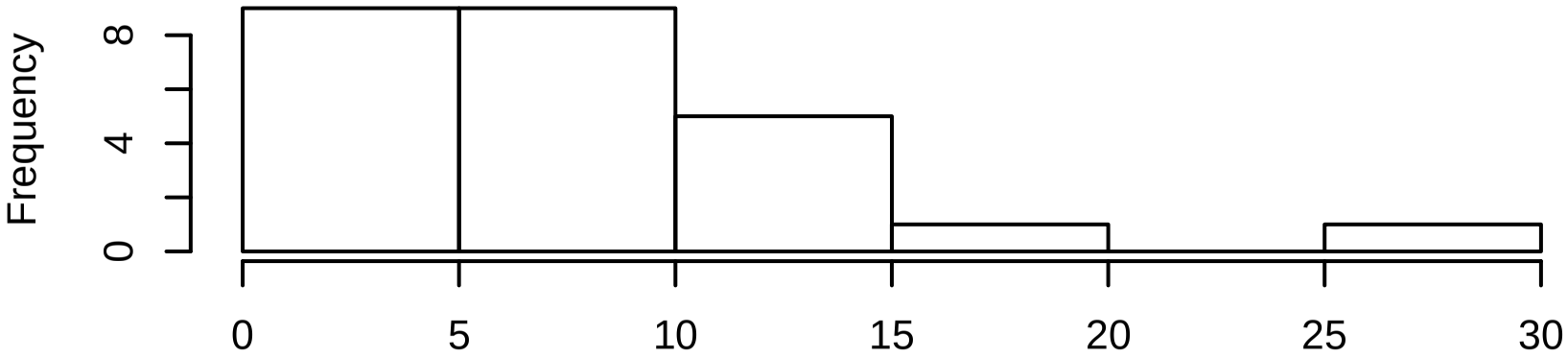
General framework: Generalized linear model

$$g(\mu_i) \equiv g(E(y_i|\beta, \phi)) = \eta_i \equiv x_i'\beta$$

with response y_i (here: duration stay), explanatory vector x_i' (temperature, age), dispersion ϕ, and link function $g(\cdot)$.

Need to specify the response distribution!

- waiting time/duration problems: Gamma (exponential) distribution.
- assumption supported by histogram of duration:



Initial fit with Gamma-model

```
Call: glm(formula = duration ~ age + temp1, family = Gamma(link

=log), data = hosp)

Coefficients:

             Estimate Std. Error t value Pr(>|t|)

(Intercept) -28.654096  16.621034  -1.724   0.0987 .

 age          0.014900   0.005698   2.615   0.0158 *

 temp1        0.306624   0.168141   1.824   0.0818 .

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.2690239)

                        -2 log L:    135.2
```

Overdispersion? Fit generalized linear model with random effect

$$\log(\mu_i) \equiv \log(E(y_i|z_i, \beta, \phi)) = \eta_i \equiv x_i'\beta + z_i,$$

- $y_i|z_i$ Gamma distributed,
- z_i is a random effect with unknown (!) density $f(\cdot)$, accounting for
  - unobserved covariates
  - model misspecification
  - individual unit variability

⋮

Fitting random effect models: Nonparametric Maximum Likelihood

(NPML, Aitkin, 1996).

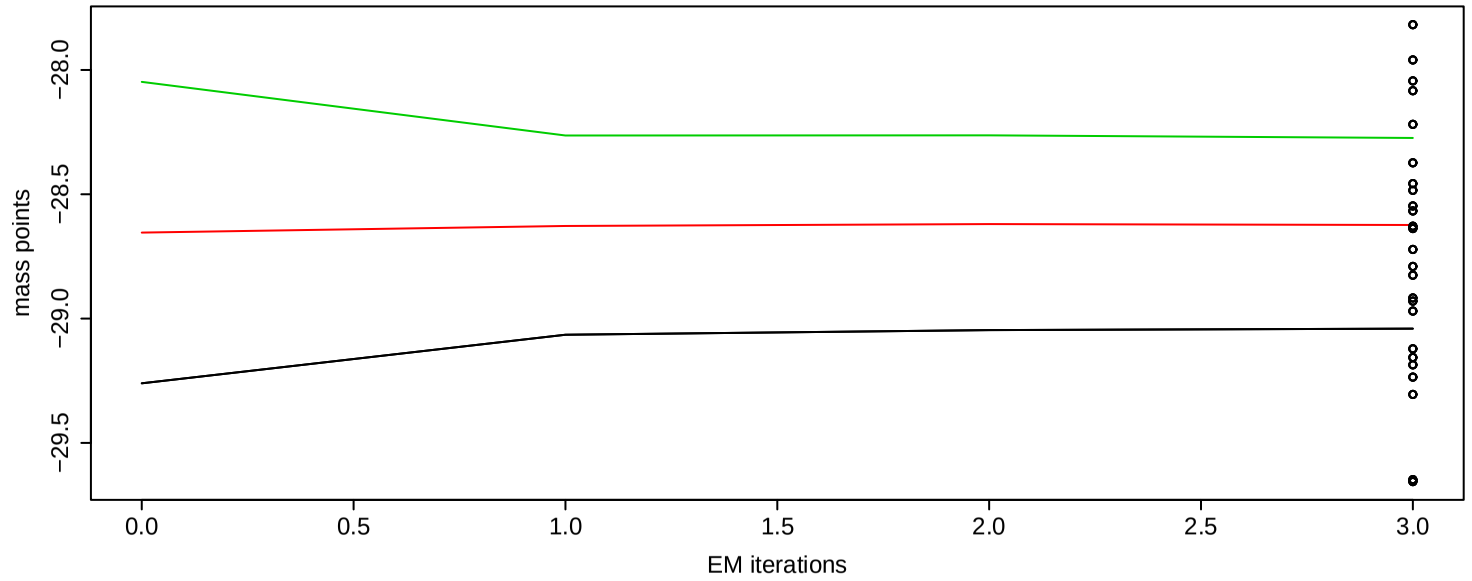
- No parametric assumption about the random effect distribution $f(\cdot)$.
- Approximate $f(\cdot)$ by a discrete mixture (Laird, 1978)

  mass points $\{z_k\}$ with masses $\{\pi_k\}$.
- Fitted model is a K component mixture model.
- uses EM algorithm (missing information = component membership)

E-Step Adjust weights $w_{ik} = P(\text{obs. } i \text{ comes from comp. } k)$

M-Step Update parameter estimates fitting a weighted GLM with weights w_{ik}.

EM Trajectories for 'Classical' NPML



What is the problem?

A high working dispersion parameter (here: inverse shape parameter) in the initial EM cycles blurs the mixture components. As a consequence, the posterior weights w_{ik} lose their discriminatory power, and the EM algorithm either behaves erratically (Einbeck & Hinde, 2005) or gets stuck (as above).

Solution

Damping of the EM algorithm (Einbeck & Hinde, 2006):

In the $j-$th EM cycle, apply the working dispersion

$$\phi_j = d_j^2 \hat{\phi}_j,$$

or, in other words, a working shape parameter

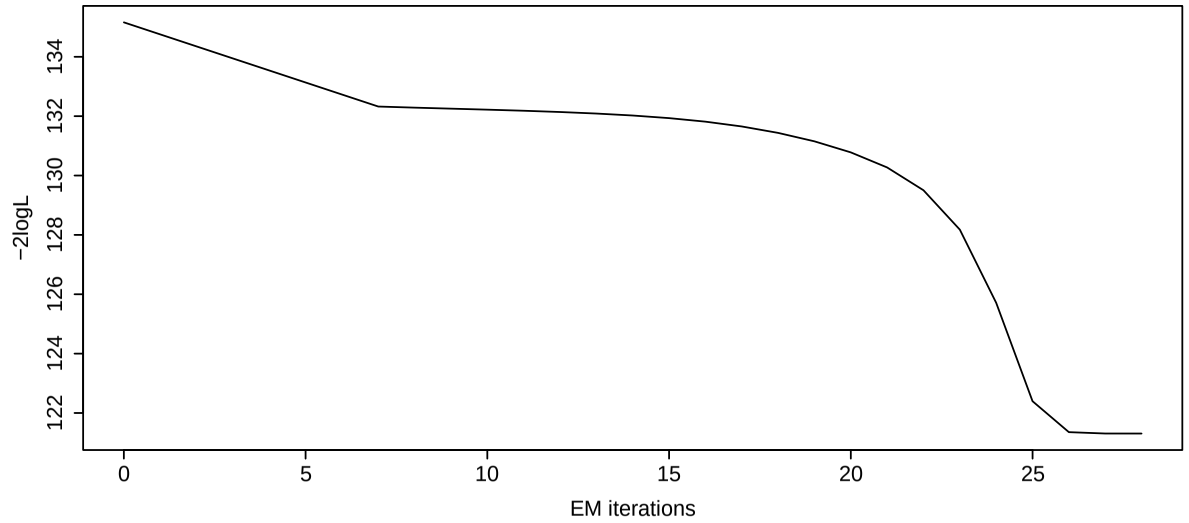
$$\nu_j = \frac{1}{d_j^2} \hat{\nu}_j,$$

employing a sequence d_j such that $d_0 < 1$ and $d_j \longrightarrow 1$ for $j \longrightarrow \infty$.

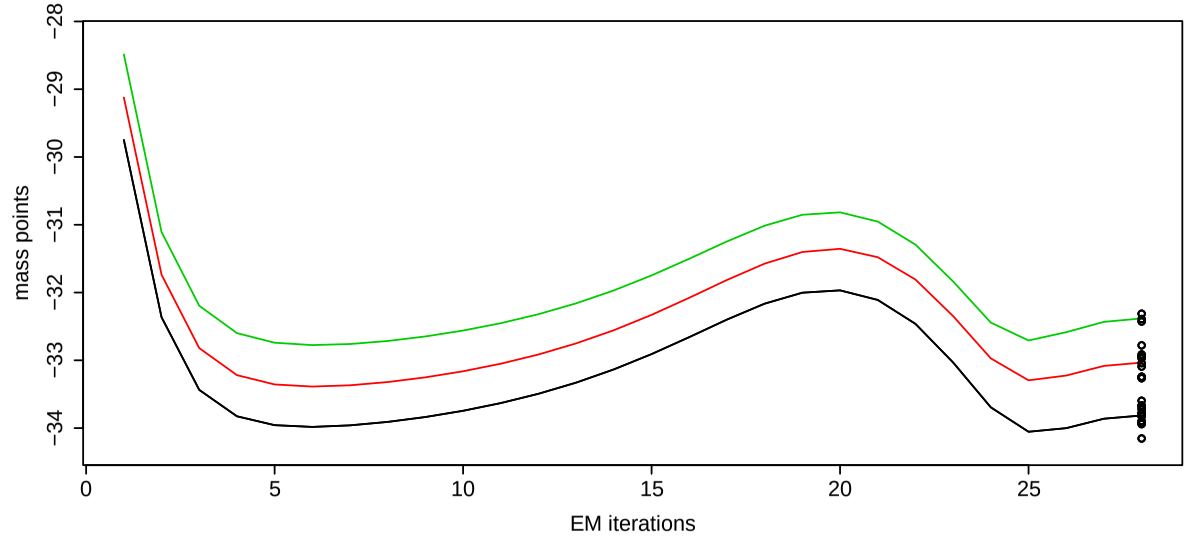
- EM Damping has its main effect in the first iterations.
- Damped EM is only 'asymptotical EM'
- Damping reduces sensitivity to starting points and the number of EM iterations.

Damped EM algorithm for hospital stay data

Disparity trend



EM trajectories



Fitted model

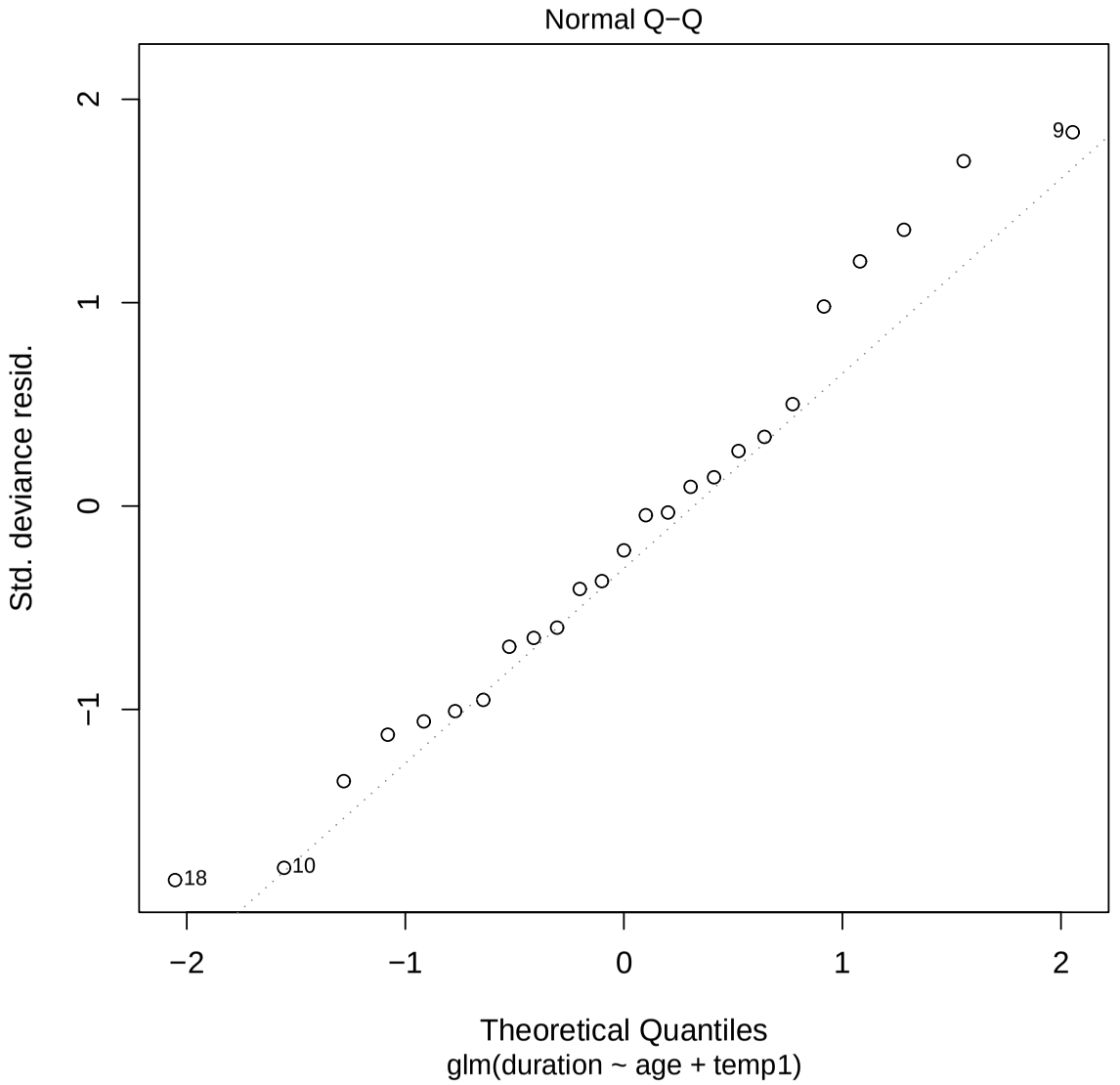
```
Coefficients:
       age        temp1        MASS1        MASS2        MASS3
  0.004024     0.357661  -33.813698  -33.033934  -32.382143


MLE of shape parameter:    50.74
Mixture proportions:
    MASS1       MASS2       MASS3
0.4798529  0.3979878  0.1221593          -2 log L:  121.3
```

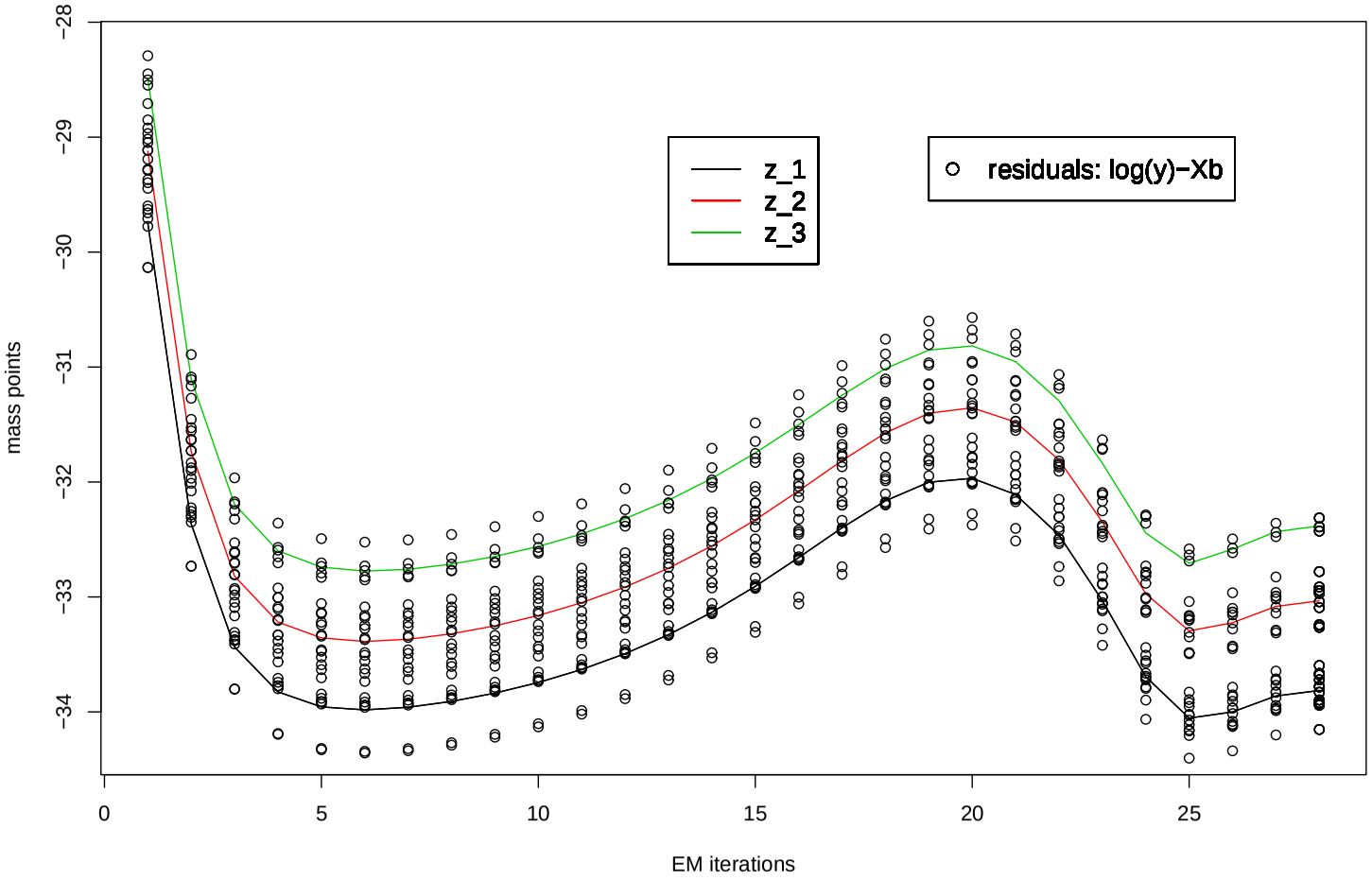
Accounting for overdispersion gives 14 points reduction in disparity!

Observation

- The large reduction in disparity is surprising, as e.g. a QQ plot for the original GLM does not indicate a substantial overdispersion.



Look at 'residuals' $log(y_i) - x_i'\hat{\beta}_j$ in $j-$th iteration.



The residuals group themselves to clusters during the EM algorithm. In the residuals of the GLM fit, the cluster structure was hidden!

Summarizing remarks

- Overdispersion can be present even when the GLM residuals look 'fine'.
- NPML estimation for random effect models is well established for Gaussian, Poisson and Binomial response.
- Extension to the Gamma distribution requires the use of a damping procedure (which is also useful for Gaussian response, see Einbeck & Hinde, 2005, 2006).
- Yet, there did not exist an available implementation for Gamma distributed response (except for the exponential distribution - GLIM, C.A.MAN), possibly due to the computational problems mentioned above.

General R Package {npmlreg} on generalized linear random effect modelling,

available at

`www.nuigalway.ie/maths/je/npml.html`

(based on initial work by R. Darnell).

- Normal, Binomial, Poisson, Gamma - distributed response
- NPML and Gaussian Quadrature (Hinde, 1982)
- Random coefficient models
- Variance component models (Aitkin, 1999)

References

AITKIN, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing* **6**, 251–262.

AITKIN, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 117–128.

AITKIN, M. and FRANCIS, B. (1995). Fitting overdispersed generalized linear models by nonparametric maximum likelihood. *The GLIM Newsletter* **25**, 37–45.

BÖHNING D., SCHLATTMANN P., and LINDSEY, B. (1992). Computer-assisted analysis of mixtures (C.A.MAN): statistical algorithms. *Biometrics* **48**, 283–303.

EINBECK, J. and HINDE, J. (2005). Making the EM algorithm for NPML estimation less sensitive to tuning parameters. *CASI - 2005. Book of Abstracts*, 52–53.

EINBECK, J. and HINDE, J. (2006). A note on NPML estimation for exponential family regression models with unspecified dispersion parameter. *Austrian Journal of Statistics*, to appear.

HINDE, J. (1982): Compound Poisson regression models. *Lecture Notes in Statistics* **14**, 109-121.

LAIRD, N. M. (1978): Nonparametric maximum likelihood estimation of a mixing distribution. *JASA* **73**, 805–811.

ROSNER, B. (2000). *Fundamentals of Biostatistics.* Thomson Learning, Duxbury, CA, USA.