Models for Additive Response Components

Gerhard Neubauer

JOANNEUM RESEARCH
Forschungsgesellschaft mbH

Institute of
Applied Statistics
and System Analysis

Gerhard Neubauer

Steyrergasse 25a
A-8010 Graz, Austria

Phone: +43 316 876-15 57
Fax: +43 316 876-915 57

gerhard.neubauer@joanneum.at
http://www.joanneum.at/sta

ISO 9001 certified

Motivation

Regression with Gaussian errors can be seen as additive decomposition of the expected response

$$E(Y) = \mu = \Sigma_j \mu_j \quad \text{with} \quad \mu_j = f_j(X_j)$$
$$\text{or} \quad \mu_j = \Sigma_j X_j \beta_j$$

Components μ_j refer to the predictor variables X_j.

Does not hold for GLM/GAM

McCullagh & Nelder, 1983; Hastie & Tibshirani, 1990

$$E(Y) = \mu = h(\eta) \qquad \eta = g(\mu) = \Sigma_j \eta_j$$
$$g(.) = h^{-1}(.)$$
$$\eta_j = X_j \beta_j \text{ for GLM}$$
$$\eta_j = f_j(X_j) \text{ for GAM}$$

The response function $h(.)$ is nonlinear in general and we have

$$\mu \neq \Sigma_j \mu_j.$$

Additive Response Components

Requiring $\mu = \Sigma_j \mu_j$ for other that normal errors implies non-additivity in η

$$E(Y) = \mu = h(\eta) = \Sigma_j \mu_j$$
$$\text{with } \mu_j = v[f_j(X_j)]$$
$$\eta = g(\mu) = g\{\Sigma_j v[f_j(X_j)]\}$$

Poisson Model with Additive Response Components

$Y \sim P(\lambda)$ with canonical link function $g(.) = \ln(.)$ and $h(.) = \exp(.)$

GLM/GAM

The model for the mean

$$\mu = \exp(\eta) = \exp[\Sigma_j f_j(X_j)] \geq 0$$

Additive Response Components

The model for the mean

$$\mu = \Sigma_j \mu_j = \Sigma_j \exp[f_j(X_j)] \geq 0$$

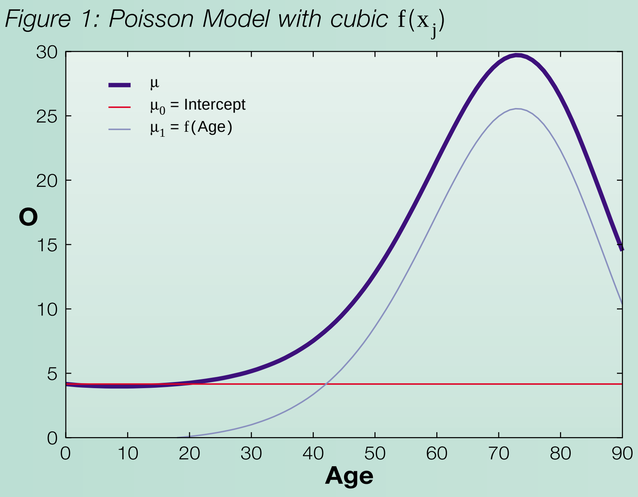
The additive components

$$\mu_j = \exp[f_j(X_j)] \geq 0$$

The nonlinear predictor

$$\eta = \ln(\mu) = \ln\{\Sigma_j \exp[f_j(X_j)]\}$$

Figure 1: Poisson Model with cubic $f(x_j)$



Binomial Model with Additive Response Components

$Y \sim B(n, p)$ with canonical link function $g(\mu) = \text{logit}(\mu)$ and $h(\eta) = \exp(\eta)/[1 + \exp(\eta)]$

GLM/GAM

The model for the mean

$$0 \leq \frac{\mu}{n} = \frac{\exp(\eta)}{1 + \exp(\eta)} = \frac{\exp[\sum_j f_j(X_j)]}{1 + \exp[\sum_j f_j(X_j)]} \leq 1$$

Additive Response Components

The model for the mean

$$0 \leq \frac{\mu}{n} = \frac{1}{n}\Sigma\mu_j = \frac{\sum_j \exp[f_j(X_j)]}{1 + \sum_j \exp[f_j(X_j)]} \leq 1$$

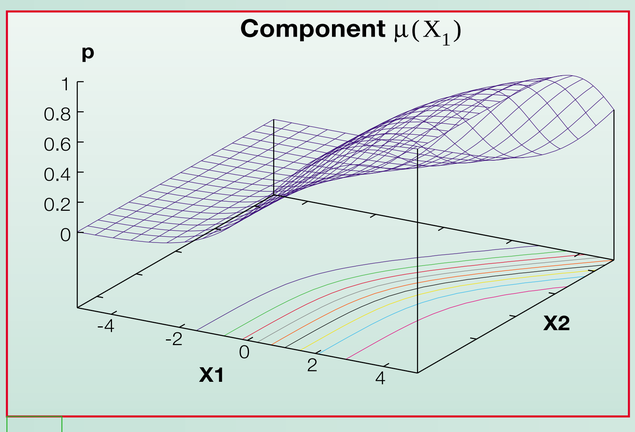
The additive components

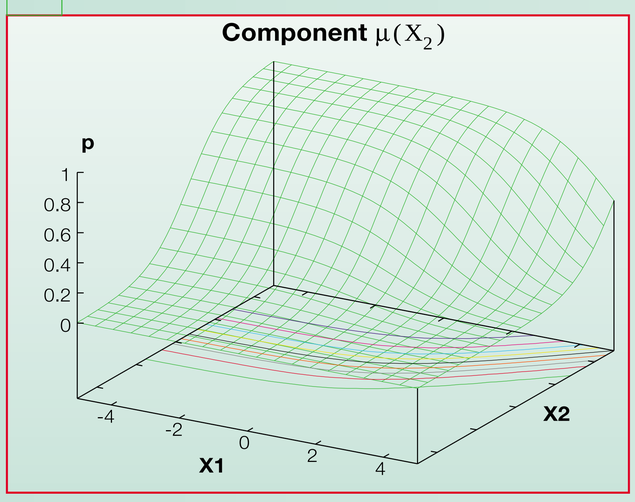
$$\mu_k = \frac{\exp[f_k(X_k)]}{1 + \sum_j \exp[f_j(X_j)]}$$

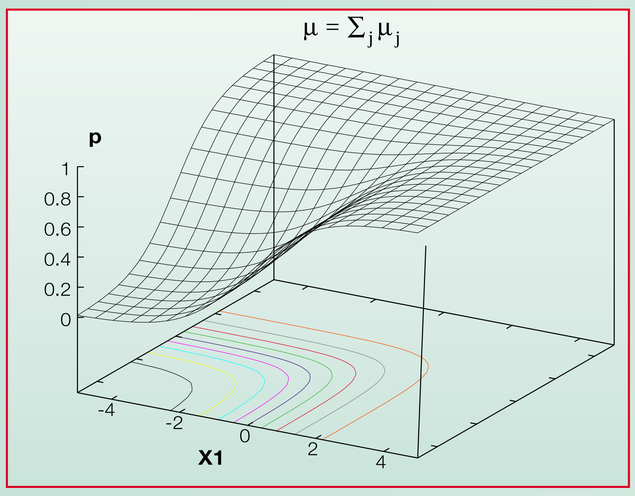
The nonlinear predictor

$$\eta = \ln[\mu/(1-\mu)] = \ln\{\Sigma_j \exp[f_j(X_j)]\}$$

Figure 2: Binominal Model for Rates with linear $f(x_j)$ *and* $\beta_1 = \beta_2 = 1$







ML Estimation

For exponential family the log likelihood as a function of the canonical parameter θ_i is given by

$$l(\theta_i) = \frac{y_i \theta_i - b(\theta_i)}{\phi} + \text{constant.}$$

Let $f_{ij} = f_j(X_{ij})$ be a polynominal of degree d_j, i. e. $f_{ij} = \sum_{k=1}^{d_j} x_{ij}^k \beta_{jk}$, then

$$\frac{\partial l_i}{\partial \beta_{ra}} = \frac{\partial l_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_{ra}} = \frac{y_i - \mu_i}{V_i \phi} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial f_{ir}} x_{ir}^a$$

For the Poisson and the Binominal model we have $\eta_i = \ln[\sum_j \exp(f_{ij})]$ and further using canonical links we obtain

$$\frac{\partial l_i}{\partial \beta_{ra}} = \frac{y_i - \mu_i}{V_i \phi} \mu_{ir} x_{ir}^a, \text{ and}$$

$$\frac{\partial l_i}{\partial \beta_{ra}} = \frac{y_i - \mu_i}{V_i \phi} (1 - \mu_i) \mu_{ir} x_{ir}^a$$

for the Poisson and the Binominal Model respectively. In matrix notation we have

$$l' = \begin{pmatrix} X_1^T Q_1 \\ \vdots \\ X_p^T Q_p \end{pmatrix} DW(y - \mu) \qquad (1)$$

- $n \times d_r$-matrix $X_r = (\partial f_r / \partial \beta_{ra})$
- $D = \text{diag}(\partial \mu_i / \partial \eta_i)$
- $W = \text{diag}[1/\text{var}(Y_i)]$
- $Qr = \text{diag}(\partial \eta_i / \partial f_{ir})$

Solving (1) by the Newton-Raphson algorithm requires the second derivates

$$l''_{rs} = X_r^T \left(\frac{\partial Q_r DW}{\partial f_s} (y - \mu) - Q_r DW Q_s \frac{\partial \mu}{\partial \eta} \right) X_s. \qquad (2)$$

For canonical links (1) and (2) simplify as now $DW = I$ and further using Fisher scoring l''_{rs} is replaced by its expectation

$$E(l''_{rs}) = - X_r^T Q_r Q_s \frac{\partial \mu}{\partial \eta} X_s.$$

in the estimation algorithm.

Risk Estimation

Comparing population segments, e. g. regions, $s = 1, \ldots, S$

Crude Rate

$cr_s = O_s / N_s$
$O_s = \sum_c O_{cs}$ the observed disease frequency
$c = 1, \ldots, C$ age classes

Controlling for age effects

- Standardization is usually applied
- Relies on the choice of a standard population
- Results depend on this choice

Breslow & Day, 1980, 1987

Standardized Rate

$sr_s = E_s / N_s$
$E_s = \sum_c E_{cs}$ the expected disease frequency
$E_{cs} = w_{cs} O_{cs}, \quad 0 \leq w_{cs} \leq \infty$

The effect of age is not removed but equalized

Adjusted Rate

Neubauer, 2001

$ar_s = A_s / N_s$
$A_s = \sum_c \mu_0(X_{0cs})$ with $\mu_0(X_{0cs})$ estimated from the additive response component model

$$E(O_{cs}) = \mu_0(X_{0cs}) + \mu_1(age_{cs})$$

The effect of age is estimated and removed from the prediction

Application to Hospital Discharge Data

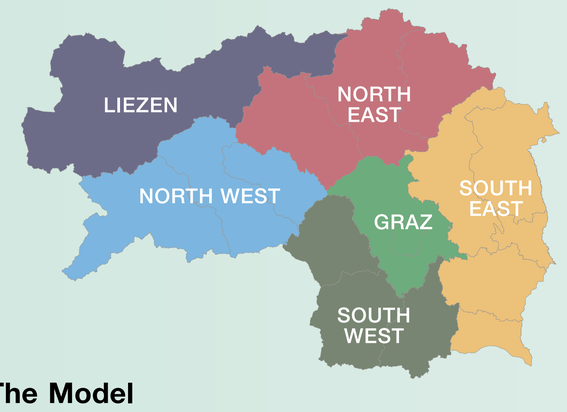
The Data

Hospital discharges with a ICD-9 diagnosis from the group Cardiovascular diseases in Styria 1995–2000

The Problem

Identify high-risk regions on the NUTS-3 level. The six NUTS-3 regions in Styria are: GRAZ, LIEZEN, NORTH WEST, NORTH EAST, SOUTH WEST, SOUTH EAST

Figure 3: The six NUTS-3 regions of Styria



The Model

Y_{ir} the number of discharges in region r

$$E(Y_{ir}) = \mu_{ir} = \exp[g_r(sex_{ir}, size_{ir})] + \exp[g_r(age_{ir})]$$

where $size_{ir}$ is the size of the population, i. e. the number of people at risk and $g(.)$ is a third order polynomial.

The Adjusted Rate

$$ar_s = n_s^{-1} \sum_i \exp[g_r(sex_{ir}, size_{ir})]$$

Results

Adjusted rates are much smaller than standardized rates, as for

- sr_s age effects are equalized, and for
- ar_s age effects are estimated and removed.

Rank Order of Regions

Crude and standardized rates:
North of Styria has highest risk,

Adjusted rates:
LIEZEN and SOUTH WEST have highest risk.

A recommendation for a health care intervention would differ depending on the chosen method.

Table 1: Rate estimates for the six NUTS-3 regions of Styria

NUTS 3 region	Crude rate Estimate	Crude rate Rank	Standardized rate Estimate	Standardized rate Rank	Adjusted rate Estimate	Adjusted rate Rank
GRAZ	0.0312	6	0.0231	6	0.0029	6
LIEZEN	0.0322	4	0.0242	5	**0.0097**	**2**
NORTH WEST	**0.0419**	**1**	**0.0292**	**1**	0.0030	5
NORTH EAST	**0.0357**	**2**	**0.0277**	**2**	0.0062	3
SOUTH WEST	0.0313	5	0.0262	4	**0.0101**	**1**
SOUTH EAST	0.0324	3	0.0269	3	0.0053	4

References

Breslow, N. E. & Day, N. E. (1980). *Statistical Methods in Cancer Research. Vol. I—The Analysis of Case-Control Studies*. Lyon: IARC.

Breslow, N. E. & Day, N. E. (1987). *Statistical Methods in Cancer Research. Vol. II—The Design and Analysis of Cohort Studies*. Lyon: IARC.

Hastie, T. J. & Tibshirani, R. J. (1990). ***Generalized Additive Models***. London: Chapman and Hall.

McCullagh, P. and Nelder, J. A. (1983). ***Generalized Linear Models***. London: Chapman and Hall.

Neubauer, G. (2001). Krankenhaus-Entlassungsdaten, Krankheitshäufigkeiten und statistische Modelle (Hospital discharge data, disease frequency and statistical models). Presentation at ROeS Seminar 2001, 24. 9.–27. 9. 2001, Mayerhofen i. Z., Austria.

Acknowledgements

I would like to thank Herwig Friedl for helpful discussions and comments throughout the work on this paper.

oefpos02043

a TRADITION of INNOVATION