WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Generalized Linear Models

Rainer Hirk

2021-10-11

Motivation

Assuming normality, the linear model $y = X\beta + e$ has

$$y_i = \beta' x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

such that

$$y_i \sim N(\mu_i, \sigma^2), \qquad \mathbb{E}(y_i) = \mu_i = \beta' x_i.$$

Various generalizations, including *general linear model* $Y = XB + E$ (with E normal with flexible error covariance structures)

But what if normality is not appropriate (e.g., skewed, bounded, discrete)? Transformations or *generalized linear models*.

Exponential Dispersion Models

Densities with respect to reference measure m of the form

$$f(y|\theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

(alternatively, write $a(\phi)$ instead of ϕ in the denominator).

For *fixed* ϕ, this is an exponential family in θ.

Exponential Dispersion Models

Differentiate $\int f(y|\theta,\phi)\,dm(y) = 1$ with respect to θ (and assume interchanging integration and differentiation is justified):

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int f(y|\theta,\phi)\,dm(y) &&= \int \frac{\partial f(y|\theta,\phi)}{\partial\theta}\,dm(y) \\ &= \int \frac{y - b'(\theta)}{\phi} f(y|\theta,\phi)\,dm(y) &&= \frac{E_{\theta,\phi}(y) - b'(\theta)}{\phi} \end{aligned}$$

so that

$$E_{\theta,\phi}(y) = b'(\theta)$$

(which does not depend on ϕ!).

Differentiate once more:

$$\begin{aligned} 0 &= \int \left(-\frac{b''(\theta)}{\phi} + \left(\frac{y - b'(\theta)}{\phi} \right)^2 \right) f(y|\theta, \phi)\, dm(y) \\ &= -\frac{b''(\theta)}{\phi} + \frac{V_{\theta,\phi}(y)}{\phi^2} \end{aligned}$$

so that

$$V_{\theta,\phi}(y) = \phi b''(\theta).$$

(which shows that ϕ is a dispersion parameter).

Exponential Dispersion Models

We can thus write

$$\mathbb{E}(y) = \mu = b'(\theta).$$

If $\mu = b'(\theta)$ defines a one-to-one relation between μ and θ (which it does: can be shown using convex analysis), we can write $b''(\theta) = V(\mu)$, formally

$$V(\mu) = b''((b')^{-1}(\mu))$$

where V is the *variance function* of the family. Thus:

$$\mathrm{var}(y) = \phi b''(\theta) = \phi V(\mu).$$

Example: Bernoulli Family

Take y binary with $\mathbb{P}(y = 1) = p$. With m counting measure (on $\{0, 1\}$),

$$
\begin{aligned}
f(y) &= p^y (1 - p)^{1-y} \\
&= \left(\frac{p}{1 - p} \right)^y (1 - p) \\
&= \exp \left(y \log \left(\frac{p}{1 - p} \right) + \log(1 - p) \right).
\end{aligned}
$$

I.e., exponential dispersion model with $\phi = 1$ (hence in fact, exponential family) and

$$
\theta = \log \left(\frac{p}{1 - p} \right) = \mathrm{logit}(p)
$$

(quantile function of standard logistic distribution).

Example: Bernoulli Family

Inverting $\theta = \mathrm{logit}(p)$ gives

$$p = \frac{e^\theta}{1 + e^\theta}$$

(probability function of standard logistic distribution) and hence

$$1 - p = \frac{1}{1 + e^\theta}$$

so that

$$b(\theta) = -\log(1 - p) = \log(1 + e^\theta).$$

Altogether (note that there is a problem for $p \in \{0, 1\}$):

$$f(y|\theta) = \exp(y\theta - \log(1 + e^\theta)), \qquad \theta = \mathrm{logit}(p).$$

Example: Bernoulli Family

Differentiation gives:

$$b'(\theta) = \frac{1}{1+e^\theta} e^\theta = p = \mu$$

and

$$b''(\theta) = \frac{e^\theta(1+e^\theta) - e^\theta e^\theta}{(1+e^\theta)^2} = \frac{e^\theta}{1+e^\theta} \frac{1}{1+e^\theta} = p(1-p) = \mu(1-\mu).$$

Necessary? We *know* that $\mathbb{E}(y) = p$ and $\mathrm{var}(y) = p(1-p)$. Hence,

$$b'(\theta) = p = \frac{e^\theta}{1+e^\theta} \implies b(\theta) = \log(1+e^\theta).$$

Generalized Linear Models

For $i = 1, \dots, n$ have responses y_i from an exponential dispersion family with the same b and covariates x_i such that for $\mathbb{E}(y_i) = \mu_i = b'(\theta_i)$ we have

$$g(\mu_i) = \beta' x_i = \eta_i,$$

where g is the *link function* and η_i is the *linear predictor*. Alternatively,

$$\mu_i = h(\beta' x_i) = h(\eta_i),$$

where h is the *response function* (and g and h are inverses of each other if invertible).

Why useful? General conceptual framework for estimation and inference.

Maximum Likelihood Estimation

Log-likelihood is

$$\ell = \ell(\beta) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i) \right),$$

where

$$g(\mu_i) = g(b'(\theta_i)) = \beta' x_i.$$

Differentiating the latter with respect to β_j:

$$x_{ij} = \frac{\partial \beta' x_i}{\partial \beta_j} = \frac{\partial g(b'(\theta_i))}{\partial \beta_j} = g'(b'(\theta_i)) b''(\theta_i) \frac{\partial \theta_i}{\partial \beta_j} = g'(\mu_i) V(\mu_i) \frac{\partial \theta_i}{\partial \beta_j}$$

Maximum Likelihood Estimation

Hence,

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{x_{ij}}{g'(\mu_i)V(\mu_i)}$$

and

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{\phi_i} \frac{x_{ij}}{g'(\mu_i)V(\mu_i)} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_i V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)}.$$

MLE typically performed by solving score equations $\partial \ell / \partial \beta_j = 0$. For Newton-type algorithms, need the Hessian $H(\beta) = [\partial^2 \ell / \partial \beta_i \partial \beta_j]$.

As $\mu_i = b'(\theta_i)$,

$$\frac{\partial \mu_i}{\partial \beta_j} = b''(\theta_i) \frac{\partial \theta_i}{\partial \beta_j} = V(\mu_i) \frac{x_{ij}}{g'(\mu_i)V(\mu_i)} = \frac{x_{ij}}{g'(\mu_i)}$$

Maximum Likelihood Estimation

Hence:

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial \beta_j \beta_k} &= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_k} \frac{y_i - \mu_i}{\phi_i V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)} \\
&= \sum_{i=1}^{n} \frac{x_{ij}}{\phi_i} \left(-\frac{\partial \mu_i}{\partial \beta_k} \frac{1}{V(\mu_i) g'(\mu_i)} - \frac{y_i - \mu_i}{(V(\mu_i) g'(\mu_i))^2} \frac{\partial (V(\mu_i) g'(\mu_i))}{\partial \beta_k} \right) \\
&= -\sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\phi_i V(\mu_i) g'(\mu_i)^2} \\
&\quad -\sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij} x_{ik}}{\phi_i V(\mu_i)^2 g'(\mu_i)^3} (V'(\mu_i) g'(\mu_i) + V(\mu_i) g''(\mu_i))
\end{aligned}
$$

Maximum Likelihood Estimation

Second term looks complicated, but has expectation zero.

Hence, drop and only use first term for "Newton-type" iteration: *Fisher scoring algorithm*.

Equivalently, replace observed information matrix (negative Hessian of log-likelihood) by its expectation (Fisher information matrix).

Next problem: what about ϕ_i? Assume that

$$\phi_i = \phi / a_i$$

with *known* case weights a_i.

Maximum Likelihood Estimation

Then Fisher information matrix is

$$\frac{1}{\phi}\sum_{i=1}^{n}\frac{a_i}{V(\mu_i)g'(\mu_i)^2}x_{ij}x_{ik} = \frac{X'W(\beta)X}{\phi},$$

where X is the usual regressor matrix (with x_i' as row i) and

$$W(\beta) = \mathrm{diag}\left(\frac{a_i}{V(\mu_i)g'(\mu_i)^2}\right), \qquad g(\mu_i) = x_i'\beta.$$

Similarly, score function is

$$\frac{1}{\phi}\sum_{i=1}^{n}\frac{a_i}{V(\mu_i)g'(\mu_i)^2}g'(\mu_i)(y_i-\mu_i)x_{ij} = \frac{X'W(\beta)r(\beta)}{\phi},$$

where $r(\beta)$ has elements $g'(\mu_i)(y_i-\mu_i)$: so-called *working residuals*.

Maximum Likelihood Estimation

Remember: Newton updates for minimizing $\ell(\beta)$ are $\beta_{\mathrm{new}} \leftarrow \beta - (H(\ell)(\beta))^{-1} \nabla \ell(\beta)$. Thus, Fisher scoring update (with approximation for H) uses

$$\begin{aligned} \beta_{\mathrm{new}} & \leftarrow \beta + (X' W(\beta) X)^{-1} X' W(\beta) r(\beta) \\ & = (X' W(\beta) X)^{-1} X' W(\beta) (X\beta + r(\beta)) \\ & = (X' W(\beta) X)^{-1} X' W(\beta) z(\beta) \end{aligned}$$

where *working response* $z(\beta)$ has elements $\beta' x_i + g'(\mu_i)(y_i - \mu_i)$, $g(\mu_i) = x_i' \beta$.

I.e., update computed by weighted least squares regression of $z(\beta)$ on X (weights: square roots of $W(\beta)$): Fisher scoring algorithm for obtaining the MLEs is an *iterative weighted least squares* (IWLS) algorithm.

Note: common dispersion parameter ϕ not used!

Canonical Links

The *canonical link* is given by $g = (b')^{-1}$ so that

$$\eta_i = g(\mu_i) = g(b'(\theta_i)) = \theta_i,$$

$$g'(\mu) = \frac{d}{d\mu}(b')^{-1}(\mu) = \frac{1}{b''((b')^{-1}(\mu))} = \frac{1}{V(\mu)},$$

so that $g'(\mu)V(\mu) \equiv 1$, and hence

$$\frac{\partial \ell}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n} a_i(y_i - \mu_i)x_{ij}, \qquad \frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = -\frac{1}{\phi}\sum_{i=1}^{n} a_i V(\mu_i) x_{ij} x_{ik}$$

Thus: observed and expected information coincide, IWLS Fisher scoring algorithm is the same as Newton's algorithm.

Under suitable conditions, MLE $\hat{\beta}$ asymptotically

$$N(\beta, I(\beta)^{-1})$$

with expected Fisher information matrix

$$I(\beta) = \frac{1}{\phi} X' W(\beta) X.$$

Thus, standard errors can be computed as square roots of diagonal elements of

$$\widehat{\mathrm{cov}}(\hat{\beta}) = \phi (X' W(\hat{\beta}) X)^{-1}$$

where $X' W(\hat{\beta}) X$ is a by-product of the final IWLS iteration.

Inference

This needs an estimate of ϕ (unless known).

Estimation by MLE is practically difficult: hence, usually estimated by method of moments.

Remember $\mathrm{var}(y_i) = \phi_i V(\mu_i) = \phi V(\mu_i)/a_i$.

Hence: if β was known, unbiased estimate of ϕ would be

$$\frac{1}{n} \sum_{i=1}^{n} \frac{a_i (y_i - \mu_i)^2}{V(\mu_i)}.$$

Taking into account that β is estimated, estimate is

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{a_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

(where p is the number of β parameters).

Deviance

A quality-of-fit statistic for model fitting achieved by ML, generalizing the idea of using the sum of squares of residuals in ordinary least squares:

$$D = 2\phi(\ell_{sat} - \ell_{mod})$$

(assuming a common ϕ, perhaps after taking out weights), where the *saturated model* uses separate parameters for each observation so that the data is fitted exactly. For GLMs: $y_i = \mu_i^* = b'(\theta_i^*)$ achieves zero scores.

Contribution of observation i to $\ell_{sat} - \ell_{mod}$ is

$$\frac{y_i\theta_i^* - b'(\theta_i^*)}{\phi_i} - \frac{y_i\hat{\theta}_i - b'(\hat{\theta}_i)}{\phi_i} = \left.\frac{y_i\theta - b(\theta)}{\phi_i}\right|_{\hat{\theta}_i}^{\theta_i^*},$$

where $\hat{\theta}_i$ is obtained from the fitted model (i.e., $g(b'(\hat{\theta}_i)) = \hat{\beta}' x_i$).

Deviance

We can write

$$(y_i\theta - b(\theta))\Big|_{\hat\theta_i}^{\theta_i^*} = \int_{\hat\theta_i}^{\theta_i^*} \frac{d}{d\theta}(y_i\theta - b(\theta))\,d\theta = \int_{\hat\theta_i}^{\theta_i^*} (y_i - b'(\theta))\,d\theta.$$

Substituting $\mu = b'(\theta)$: $d\mu = b''(\theta)\,d\theta$, i.e., $d\theta = V(\mu)^{-1}\,d\mu$, so that

$$\int_{\hat\theta_i}^{\theta_i^*} (y_i - b'(\theta))\,d\theta = \int_{\hat\mu_i}^{y_i} \frac{y_i - \mu}{V(\mu)}\,d\mu$$

and the deviance contribution of observation i is

$$2\phi_i\,\frac{y_i\theta - b(\theta)}{\phi_i}\Bigg|_{\hat\theta_i}^{\theta_i^*} = 2\int_{\hat\mu_i}^{y_i} \frac{y_i - \mu}{V(\mu)}\,d\mu.$$

Can be taken to define deviance and introduce quasi-likelihood models.

Residuals

Several kinds of residuals can be defined for GLMs:

response $y_i - \hat{\mu}_i$

working from working response in IWLS, i.e., $g'(\hat{\mu}_i)(y_i - \hat{\mu}_i)$

Pearson

$$r_i^P = \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)}$$

so that $\sum_i (r_i^P)^2$ equals the generalized Pearson statistic.

deviance so that $\sum_i (r_i^D)^2$ equals the deviance (see above).

(All definitions equivalent for the Gaussian family.)

Generalized Linear Mixed Models

Augment the linear predictor by (unknown) random effects b_i:

$$\eta_i = x_i'\beta + z_i'b_i$$

where the b_i come from a suitable family of distributions and the z_i (as well as the x_i, of course) are known covariates. Typically, $b_i \sim N(0, G(\vartheta))$. Conditionally on b_i, y_i is taken to follow an exponential dispersion model with

$$g(\mathbb{E}(y_i|b_i)) = \eta_i = x_i'\beta + z_i'b_i.$$

Marginal likelihood function is observed y_i obtained by integrating out the joint likelihood of the y_i and b_i with respect to the marginal distribution of the b_i. If b_i are independent across observation units,

$$L(\beta, \phi, \vartheta) = \prod_{i=1}^{n} \int f(y_i|\beta, \phi, \vartheta, b_i) f(b_i|\vartheta)\, db_i$$