Mathematics of Generalized Linear Models

Felix Michael Clark

February 16, 2026

Contents

1 Introduction

A generalized linear model (GLM) describes the expected value of a response variable y with a link function g and a linear combination of $K-1$ feature variables x_k (for K parameters including the intercept term).

$$g(E[y]) \sim \beta_0 + \beta_1 x_1 + \ldots + \beta_{K-1} x_{K-1}$$

The definition $x_0 \equiv 1$ is used so that with $\mathbf{x}^\intercal \equiv [1, x_1, \ldots, x_{K-1}]$ and $\boldsymbol{\beta}^\intercal \equiv [\beta_0, \beta_1, \ldots \beta_{K-1}]$ the above equation is written as the following.

$$g(E[y]) = \mathbf{x}^\intercal \boldsymbol{\beta}$$

The *linear predictor* is defined as $\omega = \mathbf{x}^\intercal \boldsymbol{\beta}$.

For N observations the response variables can be expressed as a vector $\mathbf{y}^\intercal \equiv [y^{(1)}, \ldots, y^{(N)}]$ and the data matrix is

$$\mathbf{X} \equiv \begin{pmatrix} 1 & x_1^{(1)} & \ldots & x_{K-1}^{(1)} \\ 1 & x_1^{(2)} & \ldots & x_{K-1}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(N)} & \ldots & x_{K-1}^{(N)} \end{pmatrix} \tag{1}$$

so that the matrix components are $X_{ik} = x_k^{(i)}$.

2 Exponential family of distributions

The exponential family is a useful class that includes many common elementary distributions. With a parameterization $\boldsymbol{\theta}$ and sufficient statistics $\mathbf{T}(y)$, the density function of each can be written in a canonical form with a function $\boldsymbol{\eta}(\boldsymbol{\theta})$ where $\boldsymbol{\eta}$ are the *natural parameters* of the distribution.

$$f(y;\boldsymbol{\theta}) = \exp\left[\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \mathbf{T}(y) - A(\boldsymbol{\theta}) + B(y)\right]$$

The function A is the log partition function and is particularly useful when expressed in terms of the natural parameters.

$$A(\boldsymbol{\eta}) = \log\left[\int dy\, e^{\boldsymbol{\eta}\cdot\mathbf{T}(y)+B(y)}\right]$$

The expectation value of the sufficient statistics is given by the gradient of A with respect to the natural parameters.

$$\mathrm{E}[\mathbf{T}(y)] = \nabla A(\boldsymbol{\eta})$$

The covariance matrix is given by the Hessian of the log partition function.

$$\mathrm{Cov}[\mathbf{T}(y)] = \nabla\nabla^{\intercal} A(\boldsymbol{\eta})$$

When the sufficient statistic T and the function η are both the identity, the distribution is said to be in the natural exponential family.

$$f(y;\theta) = \exp[\theta y - A(\theta) + B(y)]$$

If the canonical link function is used so that $\theta = \mathbf{x}\cdot\boldsymbol{\beta}$ then the link function is determined by the derivative of the log-partition function.

$$g^{-1}(\mathbf{x}\cdot\boldsymbol{\beta}) = \mathrm{E}[y|\mathbf{x},\boldsymbol{\beta}] = \frac{\partial A}{\partial\theta}$$

3 Likelihood function

For a dataset of N observations $\{(y^{(i)},\mathbf{x}^{(i)})|i \in [1,\ldots,N]\}$ the total log-likelihood function using a canonical-form exponential distribution for y is

$$l(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) = \sum_{i=1}^{N} \boldsymbol{\eta}(\omega^{(i)}) \cdot \mathbf{T}(y^{(i)}) - A\left(\boldsymbol{\eta}(\omega^{(i)})\right)$$

where the $B(y)$ terms have been excluded as they do not affect the dependency of the likelihood on $\boldsymbol{\beta}$.

When multiple sufficient statistics are used $\mathbf{T}^{\intercal}(y) = [T^1(y), T^2(y), \ldots]$ the natural approach is to make each natural parameter a function of a separate set of regression parameters $\eta^a = \eta^a(\mathbf{x}^{\intercal}\boldsymbol{\beta}^a)$. The gradient with respect to $\boldsymbol{\beta}^a$ is

$$\nabla_{\beta^a} l = \sum_{i=1}^{N} \mathbf{x}^{(i)} \left(T^a(y^{(i)}) - \mathrm{E}[T^a(y^{(i)})|\boldsymbol{\eta}]\right) \eta^{a\prime} \tag{2}$$

where the fact that $\nabla_\eta A(\boldsymbol{\eta}) = E[\mathbf{T}(y)]$ is used. By definition, the canonical link function is the one that results from using $\eta = \mathbf{x} \cdot \boldsymbol{\beta}$ which results in $\eta' = 1$.

The Hessian is given by

$$\begin{aligned} \nabla_{\beta^a} \nabla^{\mathsf{T}}_{\beta^b} l = \sum_{i=1}^{N} \mathbf{x}^{(i)} \mathbf{x}^{\mathsf{T}(i)} \Big\{ & -\eta^{a\prime} \mathrm{Cov}\left[T^{ab}(y^{(i)}) \middle| \boldsymbol{\eta} \right] \eta^{b\prime} \\ & + \delta_{ab} \eta^{a\prime\prime} \left(T^b(y^{(i)}) - \mathrm{E}\left[T^b(y^{(i)}) \middle| \boldsymbol{\eta} \right] \right) \Big\} \end{aligned} \tag{3}$$

where δ_{ab} is the Kronecker delta and there is no implicit summation over indices a, b in the term in the curly brackets.

Often the sufficient statistic is just the response variable itself $\mathbf{T}(y) = y$. This allows for some simplification of the above equations for the log-likelihood and its derivatives, while still allowing for a general link function.

$$l(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \eta(\mathbf{x}^{\mathsf{T}(i)} \boldsymbol{\beta}) y^{(i)} - A\left(\eta(\mathbf{x}^{\mathsf{T}(i)} \boldsymbol{\beta}) \right) \tag{4}$$

$$\nabla_\beta l = \sum_{i=1}^{N} \mathbf{x}^{(i)} \eta' \left(y^{(i)} - \mathrm{E}[y^{(i)}|\eta] \right) \tag{5}$$

$$\nabla_\beta \nabla^{\mathsf{T}}_\beta l = \sum_{i=1}^{N} \mathbf{x}^{(i)} \mathbf{x}^{\mathsf{T}(i)} \left\{ -(\eta')^2 \mathrm{Var}\left[y^{(i)} \middle| \eta \right] + \eta'' \left(y^{(i)} - \mathrm{E}\left[y^{(i)} \middle| \eta \right] \right) \right\} \tag{6}$$

If in addition the canonical link function is used, $\eta = \mathbf{x} \cdot \boldsymbol{\beta}$ and $g(\mathrm{E}[y]) = \eta$. The above equations simplify further and can be expressed easily in matrix form to include the sum over observations,

$$l = \mathbf{y}^{\mathsf{T}} \mathbf{X} \boldsymbol{\beta} - \sum_{i=1}^{N} A\left(\omega^{(i)} \right) \tag{7}$$

$$\nabla_\beta l = \mathbf{X}^{\mathsf{T}} \left[\mathbf{y} - g^{-1}(\mathbf{X}\boldsymbol{\beta}) \right] \tag{8}$$

$$\nabla_\beta \nabla^{\mathsf{T}}_\beta l = -\mathbf{X}^{\mathsf{T}} \mathbf{S} \mathbf{X} \tag{9}$$

where the inverse link function g^{-1} is applied element-wise and the variance matrix is diagonal with $S_{ii} = \mathrm{Var}[y^{(i)}|\eta]$. Even when a non-canonical link function is used, the term with the 2nd derivative of η is often dropped, as it can lead to numerical issues if the Hessian is negative. There is a more robust justification for this involving e.g. Fisher information and expectations, but and a reference here would be good to add.

4 Dispersion parameter

A useful generalization is to use an *overdispersed* exponential family. The so-called dispersion is parameterized by the variable ϕ.

$$\log f(y; \boldsymbol{\theta}, \phi) = \frac{\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \mathbf{T}(y) - A(\boldsymbol{\theta})}{\phi} + B(y, \phi)$$

One useful feature of this formalism is that the dispersion parameter can sometimes absorb on of the parameters of a multi-parameter distribution, such as the variance in a normal distribution, allowing the regression to be agnostic of the absorbed parameter.

The first and second moments of $\mathbf{T}(y)$ become the following.

$$\mathrm{E}[\mathbf{T}(y)] = \nabla A(\boldsymbol{\eta}) \tag{10}$$

$$\mathrm{Cov}[\mathbf{T}(y)] = \phi \nabla \nabla^{\intercal} A(\boldsymbol{\eta}) \tag{11}$$

Note that the expression for the expectation value has not changed, but the variance picks up a factor of ϕ. In the natural case where $\mathbf{T}(y) = y$, the second derivative of A with respect to the natural parameter η is called the *variance function* $V(\mu)$ when it can be written as a function of the predicted mean of the response $\mu = g^{-1}(\mathbf{x} \cdot \boldsymbol{\beta})$.

$$V(\mu) = \left. \frac{\partial^2 A(\eta)}{\partial \eta^2} \right|_{\eta = \eta(g(\mu))}$$

Expressing the variance as a function of the mean is useful because it is immune to changing the link function. The variance function, like the dispersion parameter, is unique only up to a constant due to the following relationship.

$$\mathrm{Var}[y] = \phi V(\mu)$$

Allowing the dispersion parameter to be different for each observation also provides a consistent approach to give each observation separate weights. The likelihood and its derivatives are adjusted by a factor of $1/\phi^{(i)}$ for each sample. With the weight matrix $\mathbf{W}$ defined as a diagonal matrix such that $W_{ii} = 1/\phi^{(i)}$, the gradient and Hessian of the likelihood becomes in the natural case

$$\nabla_\beta l = \mathbf{X}^{\intercal} \mathbf{W} \left[\mathbf{y} - g^{-1}(\mathbf{X}\boldsymbol{\beta}) \right] \tag{12}$$

$$\nabla_\beta \nabla_\beta^{\intercal} l = -\mathbf{X}^{\intercal} \mathbf{W} \mathbf{S} \mathbf{X} \tag{13}$$

Correlations between different observations $y^{(i)}$ could be induced by allowing off-diagonal terms in $\mathbf{W}$. Note that $\mathbf{W}$ is reminiscent of the inverse covariance matrix, at least in the multivariate normal distribution. The Hessian is manifestly symmetric so long as $\mathbf{W}$ is.

In practice for GLMs, individual weights for observations can be used but are not typically discussed as observation-dependent dispersions. Instead, the dispersion parameter is either fixed to one (logistic, exponential) or considered a free parameter (OLS, gamma). It does not affect the best parameters, but it does impact other quantities like variance or standardized residuals.

When not fixed to 1, the dispersion parameter is estimated via the total deviance:

$$\phi = \frac{D}{N - K}$$

where K is the number of free parameters in the model. When variance weights[1] w_i are non-uniform, the unbiased estimator is

$$\phi_{\mathbf{w}} = \frac{1}{\left(1 - \frac{K}{n_{\text{eff}}}\right)} \frac{\Sigma_i w_i D_i}{\Sigma_i w_i}$$

where $n_{\text{eff}} \equiv \frac{(\Sigma w)^2}{\Sigma w^2}$ is Kish's *effective sample size.* The condition for this estimator to be defined changes from $K < N$ in the unweighted case to $K < n_{\text{eff}}$ in the weighted case.

5 Comments on computing the link and variance functions

In the standard description the link function generally relates the expected value of the response variable to the linear predictor.

$$g(\mathrm{E}[y|\omega]) = \omega$$

Once the link function is known, its inverse can be used to quickly get the expected value given the linear predictor.

$$\mathrm{E}[y|\omega] = g^{-1}(\omega)$$

The expected response value, or mean μ, is known either from knowledge of the standard parameters $\boldsymbol{\theta}$ or, assuming a single sufficient statistic y, through the first derivative of the log-partition function as a function of the natural parameter η.

$$\mu = \frac{\partial A(\eta)}{\partial \eta} \text{ for } \mathbf{T}(y) = [y]$$

This expression can be inverted to get the natural parameter as a function of the mean $\eta(\mu) = g_0(\mu) = (A')^{-1}(\mu)$. To use the canonical link function, the natural parameter is equal to the linear predictor so that $\eta_0(\omega) = \omega$. A different parameterization can be used for η, which implies a change in the link function. By comparing the expectation value for two different link functions, the natural parameter is found for an arbitrary link function g in terms of the canonical link function g_0 and linear predictor ω.

$$\eta(\omega) = g_0\left(g^{-1}(\omega)\right)$$

The same function $\eta(\mu)$ can be used to compute the variance function $V(\mu)$.

$$V(\mu) = \left.\frac{\partial^2 A(\eta)}{\partial \eta^2}\right|_{\eta=\eta(\mu)}$$

[1] also known as analytic weights or reliability weights

Note that it is unaffected by any transformation s. In fact, by the inverse function theorem, if $\eta(\mu) = t(\mu)$, then the variance function is the reciprocal of the derivative of t.

$$V(\mu) = \frac{1}{\frac{\partial t}{\partial \mu}}$$

If there are multiple sufficient statistics, these statements should all be generalizable in the obvious way. Instead of a parameter μ we have a vector of expectation values, and the generalized link function maps these multiple values to the set of linear predictors, which have the same cardinality. The multiple natural parameters can still be expressed as functions of these expected values. The variance function is now a symmetric matrix but there typically should be nothing stopping it from being expressible in terms of the expected values of the sufficient statistics; indeed, the multivariate version of the inverse function theorem should still apply.

See this post for a helpful summary of non-canonical link functions, although be wary of the differing notation. Another note: When the canonical link function is used, $\mathbf{X}^\intercal \mathbf{y}$ is a sufficient statistic for the whole data set.

NOTE: When the canonical link function allows un-physical values, the Jeffreys prior may be useful in determining a natural link function. For instance, the Jeffrey's prior for the variance of a normal distribution imposes that the logarithm of the variance is uniformly distributed. This might suggest that the linear predictor for the variance should be associated with the logarithm of the variance, i.e. the variance expressed as an exponential function of the linear predictor.

6 Iteratively re-weighted least squares

Iteratively re-weighted least squares (IRLS) is a useful tool for fitting GLMs because it is typically relatively straightforward to compute the Jacobian and Hessian of the likelihood function. The step $\Delta\boldsymbol{\beta}$ in the space of parameters $\boldsymbol{\beta}$ is given by the solution to

$$-\mathbf{H}(\boldsymbol{\beta}) \cdot \Delta\boldsymbol{\beta} = \mathbf{J}(\boldsymbol{\beta})$$

where $\mathbf{J}$ and $\mathbf{H}$ are the Jacobian (gradient) and Hessian of the log-likelihood function, respectively. The Hessian does not have to be inverted completely; efficient linear algebra procedures exist to solve symmetric matrix equations of this form.

This procedure is similar to Newton's method for the gradient, and (up to potential numerical issues) it will move towards the point with zero gradient so long as the likelihood is a concave function. Fortunately in GLM applications this condition typically holds, especially if the canonical link function is used.

In the case of a canonical link function with scalar sufficient statistic y, this update rule for $\Delta\boldsymbol{\beta}$ becomes the following.

$$\left(\mathbf{X}^\intercal \mathbf{S} \mathbf{X}\right) \Delta\boldsymbol{\beta} = \mathbf{X}^\intercal \left[\mathbf{y} - g^{-1}(\mathbf{X}\boldsymbol{\beta})\right] \tag{14}$$

When a weight matrix $\mathbf{W}$ is included to include the observations with different weights, a simple adjustment is needed.

$$\left(\mathbf{X}^\intercal \mathbf{W} \mathbf{S} \mathbf{X}\right) \Delta\boldsymbol{\beta} = \mathbf{X}^\intercal \mathbf{W} \left[\mathbf{y} - g^{-1}(\mathbf{X}\boldsymbol{\beta})\right] \tag{15}$$

A Hermitian-solve algorithm can be applied to generate successive guess for $\boldsymbol{\beta}$ based on the previous guess until a desired tolerance is reached:

$$\left(\mathbf{X}^\intercal \mathbf{W} \mathbf{S} \mathbf{X}\right) \boldsymbol{\beta}_{l+1} = \mathbf{X}^\intercal \mathbf{W} \left[\mathbf{y} - g^{-1}\left(\mathbf{X}\boldsymbol{\beta}_l\right) + \mathbf{S}\mathbf{X}\boldsymbol{\beta}_l\right] \tag{16}$$

7 Control variables

In practical applications we often want to control for additional variables where the effect on the response is known. This can be expressed as an adjustment to the linear predictors for each observed $y^{(i)}$.

$$g(\mathrm{E}[y^{(i)}]) = \omega_0^{(i)} + \mathbf{x}^{(i)\intercal}\boldsymbol{\beta}$$

This adjusts the IRLS step to the following.

$$\left(\mathbf{X}^\intercal \mathbf{W} \mathbf{S} \mathbf{X}\right) \boldsymbol{\beta}_{l+1} = \mathbf{X}^\intercal \mathbf{W} \left[\mathbf{y} - g^{-1}\left(\boldsymbol{\omega}_\mathbf{0} + \mathbf{X}\boldsymbol{\beta}_l\right) + \mathbf{S}\mathbf{X}\boldsymbol{\beta}_l\right] \tag{17}$$

8 Regularization

Consider the case where the response variable is linearly separable by the predictor variables in logistic regression. The magnitude of the regression parameters will increase without bound. This situation can be avoided by penalizing large values of the parameters in the likelihood function.

With all forms of regularization the scaling symmetry of each x_k is broken, so it is likely useful to center and standardize the data.

Further reading: Regularization paths for GLMs via Coordinate Descent

8.1 Ridge regression (L2 regularization)

The parameters can be discouraged from taking very large values by penalizing the likelihood by their squares.

$$l = l_0 - \frac{\lambda_2}{2} \sum_{k=1}^{K} \left|\beta_k\right|^2$$

This can be thought of as imposing a Gaussian prior distribution on the β_k parameters. Note that the intercept $k = 0$ term is left out of the regularization. Most regression applications will not want to induce a bias in this term, and there should be no risk of failing to converge so long as there is more than a single value in the set of response variables $\mathbf{y}$.

This form of regression is particularly attractive in GLMs because the Jacobian and Hessian are easily adjusted:

$$\mathbf{J} \rightarrow \mathbf{J} - \lambda_2 \tilde{\boldsymbol{\beta}} \tag{18}$$

$$\mathbf{H} \rightarrow \mathbf{H} - \lambda_2 \tilde{\mathbf{I}} \tag{19}$$

where the tildes are a (sloppy) way of indicating zeros in the 0 index. Note that the Hessian is negative-definite, so adding the regularization term does not risk making it degenerate. In fact, it can help condition for small eigenvalues of the Hessian.

Another way of denoting the above is with a Tikhonov matrix $\boldsymbol{\Gamma} = \text{diag}(0, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_2})$ which penalizes the likelihood by $\left|\boldsymbol{\Gamma}\boldsymbol{\beta}\right|^2/2$. This makes the changes to the Jacobian and Hessian easier to express.

$$\mathbf{J} \rightarrow \mathbf{J} - \boldsymbol{\Gamma}^{\intercal}\boldsymbol{\Gamma}\boldsymbol{\beta} \tag{20}$$

$$\mathbf{H} \rightarrow \mathbf{H} - \boldsymbol{\Gamma}^{\intercal}\boldsymbol{\Gamma} \tag{21}$$

The IRLS update equation is still quite simple if written in terms of solving a new guess $\boldsymbol{\beta}'$ in terms of the previous $\boldsymbol{\beta}_{\text{last}}$.

$$\left(\mathbf{X}^{\intercal}\mathbf{S}\mathbf{X} + \boldsymbol{\Gamma}^2\right)\boldsymbol{\beta}' = \mathbf{X}^{\intercal}\left(\mathbf{S}\mathbf{X}\boldsymbol{\beta}_{\text{last}} + \mathbf{y} - g^{-1}(\mathbf{X}\boldsymbol{\beta}_{\text{last}})\right) \tag{22}$$

Note that the regularization matrix only appears once, as a correction to the diagonals of the Hessian.

8.2 Lasso (L1 regularization)

The lasso penalizes the likelihood by the sum of the absolute values of the coefficients.

$$l = l_0 - \lambda_1 \sum_{k=1}^{K} |\beta_k|$$

This corresponds to a prior Laplace distribution for each parameter with width λ_1^{-1}. It tends to set small coefficients to exactly zero, in contrast to ridge regression which mostly makes large coefficients smaller.

There is difficulty in naively adjusting the Hessian and Jacobian using IRLS because the likelihood is non-differentiable when $\beta_k = 0$ for any $k \geq 1$.

In the simplest case of OLS with orthogonal covariates (i.e. $\mathbf{X}^{\intercal}\mathbf{X} = \mathbf{I}$) then the magnitude of each parameter is reduced by λ_1; if the absolute value is already less than λ_1 then it is set to zero.

$$\beta_k \rightarrow S_{\lambda_1}(\beta_k) \equiv \text{sgn}(\beta_k) \max(0, |\beta_k| - \lambda_1)$$

The Alternating Direction Method of Multipliers (ADMM) can be used for L1 regularization. This introduces additional variables to be updated in exchange for separating the non-smooth piece of the likelihood.

$$
\begin{aligned}
\boldsymbol{\beta}^{(k+1)} &= \text{argmin}_{\boldsymbol{\beta}} \left(-l(\boldsymbol{\beta}) + \frac{\rho}{2} \left| \boldsymbol{\beta} - \boldsymbol{\gamma}^{(k)} + \mathbf{u}^{(k)} \right|_2^2 \right) && (23) \\
\boldsymbol{\gamma}^{(k+1)} &= \text{argmin}_{\boldsymbol{\gamma}} \left(\lambda_1 \left| \boldsymbol{\gamma} \right|_1 + \frac{\rho}{2} \left| \boldsymbol{\beta}^{(k+1)} - \boldsymbol{\gamma} + \mathbf{u}^{(k)} \right|_2^2 \right) && (24) \\
&= S_{\lambda_1/\rho} \left(\boldsymbol{\beta}^{(k+1)} + \mathbf{u}^{(k)} \right) && (25) \\
\mathbf{u}^{(k+1)} &= \mathbf{u}^{(k)} + \boldsymbol{\beta}^{(k+1)} - \boldsymbol{\gamma}^{(k+1)} && (26)
\end{aligned}
$$

Presumably the iteration can start with $\gamma^{(0)}(\beta^{(0)})$ and $u = 0$, and ρ should be able to be anything, so $\rho = 1$ should work. The first equation can be solved by IRLS, and in the quadratic approximation, only one step could be taken here. The u variable is the cumulative sum of the residuals between β and γ. As long as the difference is added to u,k β and γ don't necessarily have to be tracked completely separately because each relies only on the previous iteration of the other one.

In practice, the minimization over β in the first equation does not need to performed to convergence. A single IRLS step can be executed instead, which makes the L1 ADMM implementation simpler - just include the augmented terms in the Hessian and Jacobian and run the γ and u updates in each iteration.

Another consideration is that ρ is often varied over time to promote convergence, a practice called *redisual balance*. This also requires a rescaling of u for each step that ρ is changed.

Note that in the case of completely correlated covariates, lasso does not uniquely determine the coefficients. An L2 penalty breaks the likelihood symmetry, so if this is a possibility then the elastic net is recommended.

Further reading:

- Efficient L1-regularized logistic regression
- SO summary with useful links, in particular:
- ... Statistical Learning via the ADMM in particular Section 6.3
- glmnet paper

8.3 Elastic net (L1 + L2 regularization)

The elastic net includes both L1 and L2 terms. The L2 term can be included by adding a constant diagonal to the Hessian, but it can also be included in the ADMM update by modifying the γ update slightly:

$$
\boldsymbol{\gamma}^{(k+1)} = \frac{1}{1 + \lambda_2/\rho} S_{\lambda_1/\rho}(\boldsymbol{\beta}^{(k+1)} + \mathbf{u}^{(k)})
$$

This can be derived by adding $\lambda_2|\gamma|_2^2$ to the function in the argmin over $\boldsymbol{\gamma}$.

8.4 TODO Effective degrees of freedom under regularization

See wikipedia and this paper.

9 Saturated likelihood and deviance

The saturated likelihood is the likelihood of a model with enough free parameters to fit every y_i exactly. It represents the likelihood of the best possible fit, which is often zero (OLS, binary logistic) but not always. As a function only of y_i, computing the saturated likelihood is tantamount to replacing $\mathbf{x}_i \cdot \boldsymbol{\beta}$ with $g(y_i)$.

$$l_{\text{sat}}(\mathbf{y}) = \sum_{i=1}^{N} \boldsymbol{\eta}(g(y^{(i)})) \cdot \mathbf{T}(y^{(i)}) - A\left(\boldsymbol{\eta}(g(y^{(i)}))\right)$$

The deviance, which measures the difference between the observations and the model's predictions for each data point, is given by[2]

$$D = -2\left[l(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) - l_{\text{sat}}(\mathbf{y})\right]$$

In OLS, this deviance is equal to the sum of squares of the residuals, and it can be interpreted as a generalization of that quantity.

We also consider the contribution from a single observation

$$D_i = -2\left[l(y^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\beta}) - l_{\text{sat}}(y^{(i)})\right]$$

so that $D = \Sigma_i w_i D_i$ where in these expressions D_i is the unweighted component of the deviance for observation i.

The *deviance residuals* are a useful measurement of the discrepancy between the response data and the model's predictions, that aren't sensitive to assumptions about the link function,

$$d_i = \text{sign}(e^{(i)})\sqrt{D_i}$$

where $e^{(i)} = y^{(i)} - \hat{y}^{(i)}$ are the *response residuals*.

10 Null likelihood

The null likelihood is the likelihood of a model where none of the covariates are used. An intercept term is allowed, unless the original model explicitly fixed the intercept $\beta_0 = 0$. In other words, the null model is a model in which $\mathbf{x} \cdot \boldsymbol{\beta} = \beta$. It can be shown (by setting the gradient to zero) that the maximum likelihood is found by $E[y|\beta] = \bar{y}$ (even for non-canonical link functions), and thus $\beta = g(\bar{y})$.

For canonical link functions, $\eta = \beta = g(\bar{y})$, but in general

[2] This is a different D than the one used elsewhere for index of dispersion.

$$l_0(\mathbf{y}) = \sum_{i=1}^{N} \boldsymbol{\eta}(g(\bar{y})) \cdot \mathbf{T}(y^{(i)}) - A\left(\boldsymbol{\eta}(g(\bar{y}))\right)$$

In the (unlikely) case of $\beta = 0$ being fixed in the model, then instead:

$$l_0(\mathbf{y}) = \sum_{i=1}^{N} \boldsymbol{\eta}(0) \cdot \mathbf{T}(y^{(i)}) - A\left(\boldsymbol{\eta}(0)\right)$$

When linear offset control factors are included, β cannot be solved analytically in general. In this case, a model can be re-fit using only the constant intercept term.

11 Projection matrix and leverage

In OLS, the projection matrix $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal$ (also known as the influence matrix or hat matrix) maps the vector of response values to the vector of predicted values $\mathbf{Py} = E[\mathbf{y}|\mathbf{X}\boldsymbol{\beta}] \equiv \hat{\mathbf{y}}$. While this precise property cannot be satisfied for all GLMs due to the non-linearities, an analogue can still be defined that retains useful properties.

$$P_{ij} \equiv \frac{\partial \hat{y}_i}{\partial y_j}$$

There are a few possible conventions for this generalization, but all result in the same diagonal. We'll start with the fundamental-looking definition and expand the total derivative in partials with respect to the fit parameters:

$$\begin{aligned} P_{ij} &\equiv \frac{\partial \hat{y}_i}{\partial y_j} && (27) \\ &= \left[\nabla_\beta g^{-1}\left(\mathbf{x}^{(i)} \cdot \boldsymbol{\beta}\right)\right] \cdot \frac{\partial \boldsymbol{\beta}}{\partial y_j} && (28) \end{aligned}$$

This is not symmetric, but it will turn out to be orthogonal to the response residuals $\mathbf{y} - \hat{\mathbf{y}}$. The first term is a straightforward application of the chain rule and derivatives of the partition function, and is equal to $\mathbf{x}^{(i)} S_{ii} \eta'^{(i)}$. The second term can be derived by inspecting the IRLS step equation and perturbing it around the minimum. Altogether this yields

$$\mathbf{P} = \boldsymbol{\eta}'\mathbf{S}\mathbf{X}\left(\mathbf{X}^\intercal\mathbf{W}\boldsymbol{\eta}'^2\mathbf{S}\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\mathbf{W}\boldsymbol{\eta}'$$

where $\mathbf{S}, \mathbf{W}, \boldsymbol{\eta}'$ are treated as diagonal matrices with elements as defined above. Equivalently, the components of the projection matrix are (for the canonical case)

$$P_{ij} = -S_{ii}\mathbf{x}^{(i)\intercal}\mathbf{H}^{-1}\mathbf{x}^{(j)}w_j$$

where $\mathbf{H}$ is the Hessian matrix of the likelihood. Note that at the fit minimum, this satisfies $\mathbf{P}^2 = \mathbf{P}$ exactly so it is a projection matrix[3] and is orthogonal to the response residuals $\mathbf{P} \cdot (\mathbf{y} - \hat{\mathbf{y}}) = 0$ as can be observed from the IRLS step equation and the convergence of the parameters such that $\Delta\boldsymbol{\beta} = 0$.

11.1 Pearson residuals

The exact projection $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$ such that the response residuals $\mathbf{e} = (\mathbf{I} - \mathbf{P})\mathbf{y}$ only holds exactly for OLS. However, for sufficiently small perturbations

$$\delta\mathbf{e} = \delta\mathbf{y} - \frac{\partial\hat{\mathbf{y}}}{\partial\mathbf{y}}\delta\mathbf{y} = (\mathbf{I} - \mathbf{P})\delta\mathbf{y}$$

To the extent that $\mathrm{Cov}[\hat{y}^{(i)}, y^{(i)}] = P_{ii}\mathrm{Var}[y^{(i)}]$ (exact for OLS and approximate/asymtotic otherwise),

$$\mathrm{Var}[e^{(i)}] = \mathrm{Var}[y^{(i)}](1 - h_i) = \frac{\phi S_{ii}}{w_i}(1 - h_i)$$

where ϕ is the dispersion parameter (often fixed to 1). This motivates the definition of the studentized Pearson residuals, which are expected to have a mean of 0 and a variance of 1.

$$\tilde{r}^{(i)} \equiv \sqrt{\frac{w_i}{\phi S_{ii}(1 - h_i)}} e^{(i)}$$

11.2 Leave-one-out (LOO) corrections

It can be shown[4] that the change in the parameters $\boldsymbol{\beta}^{(-i)}$ resulting from excluding an observation i is, using a single-step approximation (for the unweighted, canonical case):

$$(\mathbf{X}^\intercal \mathbf{S} \mathbf{X})\, \Delta\boldsymbol{\beta}^{(-i)} = \frac{\mathbf{x}^{(i)} e^{(i)}}{1 - h_i}$$

where $e^{(i)} \equiv y^{(i)} - \hat{y}^{(i)}$ are the response residuals. [5] The factor $h_i \equiv P_{ii}$ is called the leverage of the ith observation, and is a measure of how strongly a given observation affects its corresponding prediction. The more general version of this expression has additional factors of the ith weight and eta-derivative multiplying the RHS.

[3] However, regularization seems to break this property as $\mathbf{P}^2 = \mathbf{P} + \mathcal{O}(\lambda)$.

[4] This derivation uses the Sherman-Morrison formula

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^\intercal)^{-1} = \mathbf{A}^{-1} - \frac{\left(\mathbf{A}^{-1}\mathbf{u}\right)\left(\mathbf{v}^\intercal \mathbf{A}^{-1}\right)}{1 + \mathbf{v}^\intercal \mathbf{A}^{-1}\mathbf{u}}$$

[5] In the case of L1/L2 regularization, the Hessian is adjusted as normal and the expression becomes $-\mathbf{H}\Delta\boldsymbol{\beta}^{(-i)} = \frac{1}{1-h_i}\left(\mathbf{x}^{(i)} e^{(i)} - \lambda_1 \mathrm{sign}(\boldsymbol{\beta})\right)$. Perhaps a full ADMM iteration (without IRLS steps) would be required to solve this exactly. However, if a quadratic likelihood in $\Delta\boldsymbol{\beta}$ is considered, perhaps a modification of the exact OLS L1 adjustment can be applied.

By Taylor expanding to first order in $\Delta\boldsymbol{\beta}^{(-i)}$, the leave-out residual that corrects for the influence of the observation itself is given by $e^{(-i)} \equiv y^{(i)} - \hat{y}^{(-i)} = \frac{e^{(i)}}{1-h_i}$. This is useful for outlier detection.

11.2.1 Likelihood

Since the gradient of the likelihood function is zero at the minimum, the 2nd-order Taylor expansion gives the leading term for the change in the total likelihood.

$$\begin{align}
\Delta l^{(-i)} &= \frac{1}{2}\Delta\boldsymbol{\beta}^{(-i)\intercal}\mathbf{H}\Delta\boldsymbol{\beta}^{(-i)} \tag{29}\\
&= \frac{1}{2}\frac{w_i\mathbf{x}^{(i)\intercal}e^{(i)}}{1-h_i}\mathbf{H}^{-1}\frac{w_i\mathbf{x}^{(i)}e^{(i)}}{1-h_i} \tag{30}\\
&= -\frac{1}{2}\frac{e^{(i)2}P_{ii}w_i}{(1-h_i)^2 S_{ii}} \tag{31}\\
&= -\frac{1}{2}\frac{h_i w_i}{1-h_i}\frac{e^{(i)2}}{(1-h_i)S_{ii}} \tag{32}\\
& \tag{33}
\end{align}$$

The effect on the total deviance is:

$$\begin{align}
D^{(-i)} &= -2\left[l^{(-i)} - l_{\text{sat}}\right] \tag{34}\\
&= D - 2\Delta l^{(-i)} \tag{35}\\
&= D + \frac{h_i}{(1-h_i)^2}r^{(i)2} \tag{36}
\end{align}$$

where $r^{(i)} \equiv \sqrt{\frac{w_i}{S_{ii}}}e^{(i)}$ are the (non-studentized) Pearson residuals. For a single term, the change works out to

$$D_i^{(-j)} - D_i = \frac{1}{(1-h_j)S_{ii}}\left(2P_{ij}e^{(i)}e^{(j)} + \frac{P_{ij}^2}{1-h_j}e^{(j)2}\right)$$

which for the self-change (i.e. $j = i$)

$$D_i^{(-i)} - D_i = \frac{h_i r^{(i)2}}{1-h_i}\left(2 + \frac{h_i}{1-h_i}\right)$$

11.2.2 Dispersion parameter

When the dispersion parameter ϕ is free, it is typically estimated via the deviance.

$$\phi = \frac{D}{N-K}$$

When leaving one observation out, the estimation becomes

$$\phi^{(-i)} = \frac{1}{N-K-1}\left(\Sigma_{j\neq i} D_j^{(-i)}\right) \tag{37}$$

$$= \frac{1}{N-K-1}\left(D^{(-i)} - D_i^{(-i)}\right) \tag{38}$$

$$= \frac{1}{N-K-1}\left[D - D_i - \frac{h_i}{1-h_i} r^{(i)2}\right] \tag{39}$$

in terms of the Pearson results $r^{(i)}$.

With variance weights, the denominator has a more complicated form to correct. Defining moments of the variance weights $v_1 = \Sigma_i w_i$ and $v_2 = \Sigma_i w_i^2$,

$$\phi^{(-i)} = \frac{D^{(-i)} - D_i^{(-i)}}{v_1^{(-i)} - K\frac{v_2^{(-i)}}{v_1^{(-i)}}} \tag{40}$$

$$= \frac{D - D_i - \frac{h_i}{1-h_i} r^{(i)2}}{v_1 - w_i - K\frac{v_2 - w_i^2}{v_1 - w_i}} \tag{41}$$

Note that both the deviance terms D_i and the squared pearson residuals $r^{(i)2}$ include a factor of a variance weight w_i in this formulation.

`http://people.stat.sfu.ca/~raltman/stat402/402L25.pdf`

`https://www.stats.ox.ac.uk/~steffen/teaching/bs2HT9/glim.pdf`

See numpy.cov documentation for possible guidelines on how to handle frequency and variance weights.

11.2.3 Studentized deviance residuals

The externally studentized residuals are, under certain conditions, t-distributed with $N-K-1$ degrees of freedom. They are constructed by removing the effect of each data point on the fit itself when computing the deviance residuals.

$$\tilde{t}_i = \operatorname{sign}(e^{(i)})\sqrt{\frac{D_i^{(-i)}}{\phi^{(-i)}}}$$

11.2.4 Cook's distance

The Cook's distance is another measurement of how much the fit is impacted by each observation. For GLM's, it's defined as

$$C_i = \frac{\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(i)}\right)^\intercal \operatorname{Cov}[\boldsymbol{\beta}]^{-1}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(i)}\right)}{K\phi} \tag{42}$$

$$= \frac{\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(i)}\right)^\intercal \left(\mathbf{X}^\intercal \mathbf{W}\mathbf{S}\mathbf{X}\right)\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(i)}\right)}{K\phi} \tag{43}$$

$$= \frac{h_i}{(1-h_i)^2}\frac{r^{(i)2}}{K\phi} \tag{44}$$

where $r^{(i)} = \sqrt{\frac{w_i}{S_{ii}}} e^{(i)}$ are the Pearson residuals.

12 TODO Goodness of fit

- Compare log-likelihoods of fit model to saturated model. The total deviance should be Chi-squared distributed with $N - K$ d.o.f. under sufficiently normal conditions. See these notes on computing the deviance.
- Aikaike and Bayesian information criteria
- Generalized R2

13 TODO Significance of individual parameters

The difference between the likelihood at the fitted values and the likelihood with one parameter fixed to zero should follow a χ^2 distribution with 1 degree of freedom and implies the Z-score of the parameter. The likelihood should possibly be re-minimized over the other parameters to allow them to describe some of what the parameter of interest captured; however, this could cause to correlated parameters to both appear insignificant when at least one of them is highly relevant.

14 TODO Numerical considerations

15 Other references

- `https://www.stat.cmu.edu/~ryantibs/advmethods/notes/glm.pdf`
- `https://bwlewis.github.io/GLM/`
- `https://statmath.wu.ac.at/courses/heather_turner/glmCourse_001.pdf`
- Maalouf, M., & Siddiqi, M. (2014). Weighted logistic regression for large-scale imbalanced and rare events data. Knowledge-Based Systems, 59, 142–148.
- Convergence problems in GLMs

16 Case studies

16.1 Linear (Gaussian)

The probability distribution function (PDF) of a normally-distributed variable with the standard parameterization is

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

or equivalently:

$$\begin{aligned} \log f(y; \mu, \sigma^2) &= -\frac{(y-\mu)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) && (45) \\ &= \frac{\mu y - \frac{\mu^2}{2}}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) && (46) \end{aligned}$$

The common approach is to use a natural description that uses only the linear term in y, taking the dispersion parameter $\phi = \sigma^2$ so that $\eta = \mu$, $A(\eta) = \frac{\eta^2}{2}$, and $B(y, \phi) = -\frac{y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi)$.[6] The canonical link function is the identity because if $\eta = \mathbf{x}^\intercal \boldsymbol{\beta}$ we clearly have $\mu = \mathbf{x}^\intercal \boldsymbol{\beta}$. The variance function is $V(\mu) = 1$.

Even with variable weights, the log-likelihood function is simple.

$$l = -\frac{1}{2}(\mathbf{y} - \mathbf{X}^\intercal \boldsymbol{\beta})^\intercal \mathbf{W} (\mathbf{y} - \mathbf{X}^\intercal \boldsymbol{\beta})$$

This form shows the equivalence to weighted least squares (WLS), or ordinary least squares (OLS) when the weights are identical for each point. The correlated case is easily included by letting $\mathbf{W}$ be the inverse covariance matrix of the observed $y^{(i)}$. The $\boldsymbol{\beta}$ that maximizes the likelihood can be found analytically,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{W} \mathbf{y}$$

although if the estimated parameters are far from zero an additional IRLS step may be useful for improved numerical accuracy.

An alternative application is to include both y and y^2 in the sufficient statistics $\mathbf{T}^\intercal(y) = [y, y^2]$. This leads to natural parameters $\boldsymbol{\eta}^\intercal = \left[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right]$ with a trivial dispersion parameter $\phi = 1$. Up to a constant the log-partition function is $A(\eta_1, \eta_2) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-\eta_2)$. This is likely a good application for a non-canonical link function for η_2, for instance $\eta_2 = -\exp(\mathbf{x} \cdot \boldsymbol{\beta}_2)$, since $\sigma^2 > 0$ so $\eta_2 < 0$. Here the variance function $V(\mu)$ could possibly be replaced by a covariance function of the two-component expected values of μ and σ^2, but checking this should require some more careful analysis.

[6] Exercise: Check that the mean and variance of y are μ and σ^2, respectively, by taking first and second derivatives of $A(\eta)$ while making sure to include the dispersion parameter factor in the expression for the variance.

16.2 Logistic (Bernoulli)

The PDF of the Bernoulli distribution is

$$f(y;p) = p^y(1-p)^{1-y}$$

or equivalently

$$\begin{aligned}\log f(y;p) &= y\log p + (1-y)\log(1-p) && (47)\\ &= y\log\left(\frac{p}{1-p}\right) + \log(1-p) && (48)\end{aligned}$$

where y takes a value of either 0 or 1. The natural parameter is $\eta = \log\left(\frac{p}{1-p}\right) = \text{logit}(p)$ and as $p = \mu$ is the mean of the Bernoulli distribution the logit function is also the canonical link function. In terms of the natural parameter the log-partition function is $A(\eta) = \log(1+e^\eta)$. With the simple convention of $\phi = 1$, the variance function is $V(\mu) = \mu(1-\mu)$.

16.3 Poisson

The Poisson distribution for non-negative integer y is

$$f(y;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

or equivalently

$$\log f(y;\lambda) = y\log\lambda - \lambda - \log(y!)$$

The mean and variance are both equal to λ, so the natural parameter is $\eta = \log\lambda$ and the canonical link function is $g(\mu) = \log\mu$. The log-partition is $A(\eta) = e^\eta$ and $B(y) = -\log(y!)$. The dispersion parameter is taken to be $\phi = 1$ and the variance function is $V(\mu) = \mu$.

16.4 Exponential

While exponentially-distributed error terms are not common, this distribution is still a useful case study because the distribution itself is mathematically simple but the constraint on the distribution's parameter motivates a non-canonical link function. The PDF for an exponentially-distributed positive real value y is

$$f(y;\lambda) = \begin{cases}\lambda e^{-\lambda y} & \text{for } y \geq 0\\ 0 & \text{for } y < 0\end{cases}$$

with $\lambda > 0$ and the log-PDF is

$$\log f(y;\lambda) = -\lambda y + \log\lambda$$

where $y \geq 0$. The natural parameter is $\eta = -\lambda$ and the log-partition function is $A(\eta) = -\log(-\eta)$. Since the mean of the distribution is $\mu = -\frac{1}{\eta} = \frac{1}{\lambda}$, the canonical link function is the negative inverse $g_0(\mu) = -\frac{1}{\mu}$. The variance is $\mathrm{Var}(y|\eta) = \frac{1}{\eta^2}$ or $\mathrm{Var}(y|\lambda) = \frac{1}{\lambda^2}$ so with the simple choice of dispersion parameter $\phi = 1$ the variance function is $V(\mu) = \mu^2$.

The canonical link function $-\frac{1}{\mu} = \mathbf{x}^\intercal \boldsymbol{\beta}$ does not prohibit values of $\mu < 0$. The extent to which this is problematic will depend on the data points themselves, but we will show how to work around this as an example.

Consider the parameterization $\eta = -\exp(-\omega)$ where $\omega = \mathbf{x}^\intercal \boldsymbol{\beta}$. The mean and variance are e^ω and $e^{2\omega}$, respectively, so the implied link function is logarithmic $g(\mu) = \log \mu = \mathbf{x}^\intercal \boldsymbol{\beta}$. It is instructive to write out the likelihood and its first and second derivatives in this parameterization[7].

$$l = \sum_{i=1}^{N} \left[-\exp\left(-\omega^{(i)}\right) y^{(i)} - \omega^{(i)} \right] \tag{49}$$

$$\nabla_\beta l = \sum_{i=1}^{N} \mathbf{x}^{(i)} \left[\exp\left(-\omega^{(i)}\right) y^{(i)} - 1 \right] \tag{50}$$

$$\nabla_\beta \nabla_\beta^\intercal l = -\sum_{i=1}^{N} \mathbf{x}^{(i)} \mathbf{x}^{\intercal(i)} \exp\left(-\omega^{(i)}\right) y^{(i)} \tag{51}$$

Note that the exponential distribution is a special case of the gamma distribution with $\alpha = 1$. In practice using a gamma GLM may often be a better choice, but a similar issue arises with negative values so non-canonical link functions are often used there as well.

16.5 Binomial (fixed n)

The binomial distribution describes a variable y that can take values in a finite range from 0 to n inclusive. It is in the exponential family for a fixed n, which is usually an obvious constraint based on the maximum logically possible value for y. The PDF is

$$f(y; n, p) = \binom{n}{y} p^y (1-p)^{n-y}$$

or

$$\begin{aligned} \log f(y; n, p) &= y \log p + (n - y) \log(1 - p) + \log n! - \log y! - \log[(n - y)!] \quad (52) \\ &= y \log \left(\frac{p}{1 - p} \right) + n \log(1 - p) + B(y, n) \quad (53) \end{aligned}$$

and up to factors of n the analysis is similar to the logistic (Bernoulli) case. The natural parameter is $\eta = \mathrm{logit}(p)$, the log-partition function is $A(\eta) = n \log(1 + e^\eta)$, the mean is $\mu = np$, and the variance is $np(1-p)$. With dispersion parameter $\phi = 1$, the variance function is $V(\mu) = \mu \left(1 - \frac{\mu}{n}\right)$. By writing η in terms of μ it is seen that the canonical link function is $g(\mu) = \log \left(\frac{\mu}{n - \mu} \right)$.

[7] Exercise: Verify that the expression holds both by direct derivation and by applying the equations for the derivatives as a function of η and its derivatives.

16.6 Gamma

The gamma distribution has two parameters, but it turns out that the shape parameter α is often treated as the same for every observation, allowing for the β parameter to be predicted by $\mathbf{x} \cdot \boldsymbol{\beta}$. See these notes. This seems analogous to the situations in OLS where the variance σ^2 is the same for each data point so the minimization is unaffected by its value.

The PDF of the gamma distribution with the α, β parameterization is

$$f(y; \alpha, \beta) = \frac{\beta^\alpha y^{\alpha-1} e^{-\beta y}}{\Gamma(\alpha)}$$

or equivalently

$$\log f(y; \alpha, \beta) = -\beta y + (\alpha - 1)\log y + \alpha \log \beta - \log \Gamma(\alpha)$$

where it is clear that the sufficient statistics should include at least one of $[y, \log y]$ with corresponding natural parameters $[-\beta, \alpha - 1]$.

We will start the analysis a bit differently than the other distributions we've looked at so far, in order to try to get to the simplest version as quickly as possible. The mean and variance of the gamma distribution are known to be α/β and α/β^2, respectively. The variance could thus be written as either[8] $\text{Var}[y] = \frac{1}{\beta}\mu = \frac{1}{\alpha}\mu^2$ with dispersion factor ϕ being the reciprocal of either β or α, respectively. It turns out that the second choice, with $\phi = \frac{1}{\alpha}$ and $V(\mu) = \mu^2$, is the most straightforward.

$$\log f(y; \mu, \phi) = \frac{-\frac{1}{\mu}y - \log \mu}{\phi} + B(y, \phi)$$

In fact, other than the $B(y, \phi)$ term this looks just like the case of the exponential distribution with $\lambda = 1/\mu$, but with a non-unity dispersion parameter that does not affect the regression itself.

16.7 Negative binomial (fixed r)

The negative binomial distribution is an over-dispersed Poisson distribution, in that its support is over the non-negative integers but its variance is greater than its mean.

$$f(y; r, p) = \frac{p^r (1-p)^y}{y B(r, y)} \tag{54}$$

$$\log f(y; r, p) = y \log(1-p) + r \log p - \log y - \log \text{Beta}(r, y) \tag{55}$$

Like the binomial distribution, the negative binomial distribution is in the exponential family if and only if one of its standard parameters (r) is held fixed.

[8] Actually by this logic the variance function could be any power of μ with the right powers of α and β as the dispersion factor, but for simplicity we'll focus on the options that allow ϕ to be written only as a function of one or the other.

However, while the choice of n in a binomial distribution is typically obvious, it is not usually so clear how to select a value for r in the negative binomial distribution. In terms of the index of dispersion $D = \frac{\mathrm{Var}[y]}{\mathrm{E}[y]}$ and the mean μ, r is given by[9]:

$$r = \frac{\mu}{D-1}$$

Assuming that a reasonable choice of r can be selected, the natural parameter is $\eta = \log(1-p)$, the log-partition function is $A(\eta) = -r\log(1-e^{\eta})$, the mean is $\mu = \frac{r(1-p)}{p}$, the canonical link function is $g(\mu) = \log\frac{\mu}{r+\mu}$. With the obvious dispersion parameter $\phi = 1$ the variance function is $V(\mu) = \mu\left(1+\frac{\mu}{r}\right)$. Since $0 < p < 1$, a non-canonical link function might be useful, such as the one implied by $\eta = -\exp(-\mathbf{x}^{\intercal}\boldsymbol{\beta})$.

16.8 Inverse Gaussian

The inverse Gaussian (also known as the Wald) distribution has a somewhat misleading name – it is NOT the distribution of the inverse of a Gaussian-distributed parameter. Rather it is the distribution of the time it takes for a Gaussian process to drift to a certain level.

$$\begin{aligned} f(y;\mu,\lambda) &= \sqrt{\frac{\lambda}{2\pi y^3}}e^{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}} && (56)\\ \log f(y;\mu,\lambda) &= -\frac{\lambda(y-\mu)^2}{2\mu^2 y} + \frac{1}{2}\log\frac{\lambda}{2\pi y^3} && (57)\\ &= -\frac{\lambda}{2\mu^2}y - \frac{\lambda}{2}\frac{1}{y} + \frac{\lambda}{\mu} + \frac{1}{2}\log\lambda + B(y) && (58) \end{aligned}$$

The sufficient statistics are therefore $\mathbf{T}(y)^{\intercal} = [y, 1/y]$ with corresponding natural parameters $\boldsymbol{\eta}^{\intercal} = [-\lambda/2\mu^2, -\lambda/2]$, and the log-partition function is $A(\boldsymbol{\eta}) = -2\sqrt{\eta_1\eta_2} - \frac{1}{2}\log(-2\eta_2)$. The expected values of these statistics are $\mathrm{E}[\mathbf{T}^{\intercal}(y);\mu,\lambda] = [\mu, 1/\mu + 1/\lambda]$ or $\mathrm{E}[\mathbf{T}^{\intercal}(y);\boldsymbol{\eta}] = \left[\sqrt{\frac{\eta_2}{\eta_1}}, \sqrt{\frac{\eta_1}{\eta_2}} - \frac{1}{2\eta_2}\right]$.

16.9 Others

Many other distributions are in the exponential family and should be amenable in principle to a GLM model.

- Chi-squared
- Inverse gamma

[9] It is a curious fact that the natural parameter η is a simple function of the index of dispersion:

$$\eta = -\log D$$

- Beta
- Categorical
- Multinomial
- Dirichlet
- Normal-gamma

Wikipedia has a summary table with many properties of specific distributions.