Motivation for nonlinear models

The key properties of a linear model are that

$$E(Y|X) = \beta' X \quad \text{and} \quad \text{var}(Y|X) \propto I.$$

In some cases where these conditions are not met, we can transform Y so that the linear model assumptions are approximately satisfied.

However it is often difficult to find a transformation that simultaneously linearizes the mean and gives constant variance.

If Y lies in a restricted domain (e.g. $Y = 0, 1$), parameterizing $E(Y|X)$ as a linear function of X violates the domain restriction.

Generalized linear models (GLM's) are a class of nonlinear regression models that can be used in certain cases where linear models are not appropriate.

Logistic regression

Logistic regression is a specific type of GLM. We will develop logistic regression from first principals before discussing GLM's in general.

Logistic regression is used for binary outcome data, where $Y = 0$ or $Y = 1$. It is defined by the probability mass function

$$P(Y = 1|X = x) = \frac{\exp(\beta' x)}{1 + \exp(\beta' x)} = \frac{1}{1 + \exp(-\beta' x)},$$

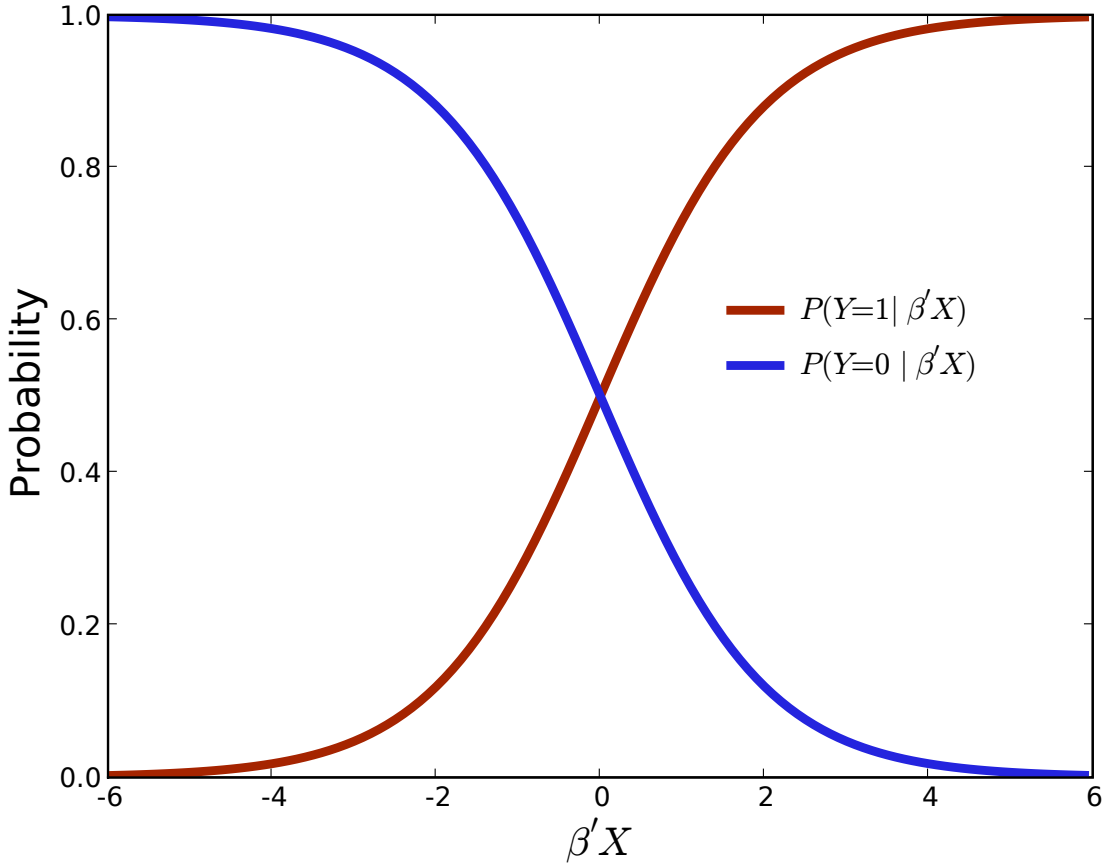
which implies that

$$P(Y = 0|X = x) = 1 - P(Y = 1|X = x) = \frac{1}{1 + \exp(\beta' x)},$$

where $x_0 \equiv 1$ so β_0 is the intercept.

Logistic regression

This plot shows $P(Y = 1|X)$ and $P(Y = 0|X)$, plotted as functions of $\beta' X$:



Logistic regression

The logit function

$$\text{logit}(x) = \log(x/(1-x))$$

maps the unit interval onto the real line. The inverse logit function, or expit function

$$\text{expit}(x) = \text{logit}^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$$

maps the real line onto the unit interval.

In logistic regression, the logit function is used to map the linear predictor $\beta' X$ to a probability.

Logistic regression

The linear predictor in logistic regression is the conditional log odds:

$$\log\left[\frac{P(Y=1|X)}{P(Y=0|X)}\right] = \beta' X.$$

Thus one way to interpret a logistic regression model is that a one unit increase in X_j results in a change of β_j in the conditional log odds.

Or, a one unit increase in X_j results in a multiplicative change of $\exp(\beta_j)$ in the conditional odds.

Latent variable model for logistic regression

It may make sense to view the binary outcome Y as being a dichotomization of a latent continuous outcome Y_c,

$$Y = \mathcal{I}(Y_c \geq 0).$$

Suppose $Y_c|X$ follows a logistic distribution, with CDF

$$F(Y_c|X) = \frac{\exp(Y_c - \beta' X)}{1 + \exp(Y_c - \beta' X)}.$$

In this case, $Y|X$ follows the logistic regression model:

$$P(Y = 1|X) = P(Y_c \geq 0|X) = 1 - \frac{\exp(0 - \beta' X)}{1 + \exp(0 - \beta' X)} = \frac{\exp(\beta' X)}{1 + \exp(\beta' X)}.$$

Mean/variance relationship for logistic regression

Since the mean and variance of a Bernoulli trial are linked, the mean structure

$$E(Y|X) = P(Y = 1|X) = \frac{\exp(\beta' X)}{1 + \exp(\beta' X)}$$

also determines the variances

$$\mathrm{var}(Y|X) = P(Y = 1|X) \cdot P(Y = 0|X) = \frac{1}{2 + \exp(\beta' X) + \exp(-\beta' X)}.$$

Sicne the variance depends on X, logistic regression models are always heteroscedastic.

Logistic regression and case-control studies

Suppose we sample people based on their disease status D ($D = 1$ is a case, $D = 0$ is a control).

We are interested in a binary marker $M \in \{0, 1\}$ that may predict a person's disease status.

The prospective log odds

$$\log\left[\frac{P(D=1|M=m)}{P(D=0|M=m)}\right] = \log\left[\frac{P(M=m|D=1)P(D=1)}{P(M=m|D=0)P(D=0)}\right]$$

measures how informative the marker is for the disease.

Logistic regression and case-control studies

Suppose we model $M|D$ using logistic regression, so

$$P(M=1|D) = \frac{\exp(\alpha+\beta D)}{1+\exp(\alpha+\beta D)} \qquad P(M=0|D) = \frac{1}{1+\exp(\alpha+\beta D)}.$$

The prospective log odds can be written

$$\log\left[\frac{\exp(M\cdot(\alpha+\beta))/(1+\exp(\alpha+\beta))}{\exp(M\cdot\alpha)/(1+\exp(\alpha))}\cdot\frac{P(D=1)}{P(D=0)}\right]$$

which equals

$$\beta M + \log\left[\frac{1+\exp(\alpha)}{1+\exp(\alpha+\beta)}\cdot\frac{P(D=1)}{P(D=0)}\right].$$

Logistic regression and case-control studies

If we had prospective data and used logistic regression to model the prosepctive relationship $D|M$, the log odds would have the form

$$\theta + \beta M.$$

Therefore we have shown that the coefficient β when we use logistic regression to regress M on D using case-control data is the same coefficient (in the population sense) as we would obtain from regressing D on M in a prospective study.

Note that the intercepts are not the same in general.

Estimation and inference for logistic regression

Assuming independent cases, the log-likelihood for logistic regression is

$$\begin{aligned} L(\beta|Y,X) &= \log \prod_i \frac{\exp(Y_i \cdot \beta' X_i)}{1+\exp(\beta' X_i)} \\ &= \sum_{i:Y_i=1} \beta' X_i - \sum_i \log(1+\exp(\beta' X_i)). \end{aligned}$$

This likelihood is for the conditional distribution of Y given X.

As in linear regression, we do not model the marginal distribution of X.

Estimation and inference for logistic regression

Logistic regression models are usually fit using maximum likelihood estimation.

This means that the parametric likelihood above is maximized as a function of β.

The gradient of the log-likelihood function (the score function) is

$$G(\beta|Y,X) = \frac{\partial}{\partial\beta}L(\beta|Y,X) = \sum_{i:Y_i=1} X_i - \sum_i \frac{\exp(\beta' X_i)}{1+\exp(\beta' X_i)} X_i.$$

Estimation and inference for logistic regression

The Hessian of the log-likelihood is

$$H(\beta|Y,X) = \frac{\partial^2}{\partial\beta\beta'}L(\beta|Y,X) = -\sum_i \frac{\exp(\beta' X_i)}{(1+\exp(\beta' X_i))^2} X_i X_i'.$$

The Hessian is strictly negative definite as long as the design matrix has independent columns. Therefore $L(\beta|Y,X)$ is a concave function of β, so has a unique maximizer, and hence the MLE is unique.

Estimation and inference for logistic regression

From general theory about the MLE, the Fisher information

$$I(\beta) = -[EH(\beta|Y, X)|X]^{-1}$$

is the asymptotic sampling covariance matrix of the MLE $\hat{\beta}$. Since $H(\beta|Y, X)$ does not depend on Y, $I(\beta) = -H(\beta|Y, X)^{-1}$.

Since $\hat{\beta}$ is an MLE for a regular problem, it is consistent, asymptotically unbiased, and asymptotically normal if the model is correctly specified.

General development of GLM's

The modeling assumptions for a GLM are

- The Y_i are conditionally independent given X.
- The probability mass function or density can be written

$$\log p(Y_i|\theta_i, \phi, X_i) = w_i(Y_i\theta_i - \gamma(\theta_i))/\phi + \tau(Y_i, \phi/w_i),$$

where w_i is a known weight, $\theta_i = g(\beta' X_i)$ for an unknown vector of regression slopes β, $g(\cdot)$ and $\gamma(\cdot)$ are smooth functions, ϕ is the "scale parameter" (which may be either known or unknown), and $\tau(\cdot)$ is a known function.

General development of GLM's

The log-likelihood function is

$$L(\beta, \phi | Y, X) = \sum_i w_i (Y_i \theta_i - \gamma(\theta_i))/\phi + \tau(Y_i, \phi/w_i).$$

The score function with respect to θ_i is

$$w_i (Y_i - \gamma'(\theta_i))/\phi.$$

General development of GLM's

Next we need a fundamental fact about score functions.

Let $f_\theta(Y)$ be a density in Y with parameter θ. The score function is

$$\frac{\partial}{\partial\theta}\log f_\theta(Y) = f_\theta(Y)^{-1}\frac{\partial}{\partial\theta}f_\theta(Y).$$

The expected value of the score function is

$$\begin{aligned} E\frac{\partial}{\partial\theta}\log f_\theta(Y) &= \int f_\theta(Y)^{-1}\left(\frac{\partial}{\partial\theta}f_\theta(Y)\right)f_\theta(Y)dY \\ &= \frac{\partial}{\partial\theta}\int f_\theta(Y)dY \\ &= 0. \end{aligned}$$

Thus the score function has expected value 0 when θ is at its true value.

General development of GLM's

Since the expected value of the score function is zero, we can conclude that

$$E(w_i(Y_i - \gamma'(\theta_i))/\phi|X) = 0,$$

so

$$E(Y_i|X) = \gamma'(\theta_i) = \gamma'(g(\beta' X_i)).$$

Note that this relationship does not depend on ϕ or τ.

General development of GLM's

Using a similar approach, we can relate the variance to w_i, ϕ, and γ'. By direct calculation,

$$\partial^2 L(\theta_i | Y_i, X_i, \phi)/\partial\theta_i^2 = -w_i\gamma''(\theta_i)/\phi.$$

Returning to the general density $f_\theta(Y)$, we can write the Hessian as

$$\frac{\partial}{\partial\theta\theta'} \log f_\theta(Y) = f_\theta(Y)^{-2} \left(f_\theta(Y) \frac{\partial^2}{\partial\theta\theta'} f_\theta(Y) - \partial f_\theta(Y)/\partial\theta \cdot \partial f_\theta(Y)/\partial\theta' \right)$$

General development of GLM's

The expected value of the Hessian is

$$
\begin{aligned}
E\frac{\partial}{\partial\theta\theta'}\log f_\theta(Y) &= \int \frac{\partial}{\partial\theta\theta'} f_\theta(Y)\cdot f_\theta(Y)dY \\
&= \frac{\partial}{\partial\theta\theta'}\int f_\theta(Y)dY - \int\left(\frac{\partial f_\theta(Y)/\partial\theta}{f_\theta(Y)}\cdot\frac{\partial f_\theta(Y)/\partial\theta'}{f_\theta(Y)}\right) \\
&= -\mathrm{cov}\left(\frac{\partial}{\partial\theta}\log f_\theta(Y)|X\right).
\end{aligned}
$$

Therefore

$$
w_i\gamma''(\theta_i)/\phi = \mathrm{var}\left(w_i(Y_i-\gamma'(\theta_i))/\phi|X\right)
$$

so $\mathrm{var}(Y_i|X) = \phi\gamma''(\theta_i)/w_i$.

Examples of GLM's

Gaussian linear model: The density of $Y|X$ can be written

$$\begin{aligned}\log p(Y_i|X_i) &= -\log(2\pi\sigma^2)/2 - \frac{1}{2\sigma^2}(Y_i - \beta' X_i)^2 \\ &= -\log(2\pi\sigma^2)/2 - Y_i^2/2\sigma^2 + (Y_i\beta' X_i - (\beta' X_i)^2/2)/\sigma^2.\end{aligned}$$

This can be put into GLM form by setting $g(x) = x$, $\gamma(x) = x^2/2$, $w_i = 1$, $\phi = \sigma^2$, and $\tau(Y_i, \phi) = -\log(2\pi\phi)/2 - Y_i^2/2\phi$.

Examples of GLM's

Logistic regression: The mass function of $Y|X$ can be written

$$\begin{aligned} \log p(Y_i|X_i) &= Y_i \log(p_i) + (1 - Y_i) \log(1 - p_i) \\ &= Y_i \log(p_i/(1 - p_i)) + \log(1 - p_i), \end{aligned}$$

where

$$p_i = \text{logit}^{-1}(\beta' X_i) = \frac{\exp(\beta' X_i)}{1 + \exp(\beta' X_i)}.$$

Since $\log(p_i/(1 - p_i)) = \beta' X$, this can be put into GLM form by setting $g(x) = x$, $\gamma(x) = -\log(1 - \text{logit}^{-1}(x)) = \log(1 + \exp(x))$, $\tau(Y_i, \phi) \equiv 0$, $w = 1$, and $\phi = 1$.

Examples of GLM's

Poisson regression: In Poisson regression, the distribution of $Y|X$ follows a Poisson distribution, with the mean response related to the covariates via

$$\log E(Y|X) = \beta' X.$$

It follows that $\log \mathrm{var}(Y|X) = \beta' X$ as well. The mass function can be written

$$\log p(Y_i|X_i) \quad = \quad Y_i \beta' X_i - \exp(\beta' X_i) - \log(Y_i!),$$

so in GLM form, $g(x) = x$, $\gamma(x) = \exp(x)$, $w = 1$, $\tau(Y_i) = -\log(Y_i!)$, and $\phi = 1$.

Examples of GLM's

Negative binomial regression: In negative binomial regression, the probability mass function for the dependent variable Y is

$$P(Y_i = y|X) = \frac{\Gamma(y + 1/\alpha)}{\Gamma(y + 1)\Gamma(1/\alpha)} \left(\frac{1}{1 + \alpha\mu_i}\right)^{1/\alpha} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y}.$$

The mean of this distribution is μ_i and the variance is $\mu_i + \alpha\mu_i^2$. If $\alpha = 0$ we get the same mean/variance relationship as the Poisson model. As α increases, we get increasingly more overdispersion.

Examples of GLM's

Negative binomial regression (continued):

The log-likelihood (dropping terms that do not involve μ) is

$$\log P(Y_i = y|X) = y \log(\frac{\alpha\mu_i}{1+\alpha\mu_i}) - \alpha^{-1}\log(1+\alpha\mu_i)$$

Suppose we model the mean as $\mu_i = \exp(\beta' X_i)$. Then in the standard GLM notation, we have

$$\theta_i = \log\left(\frac{\alpha\exp(\beta' X_i)}{1+\alpha\exp(\beta' X_i)}\right),$$

so $g(x) = \log(\alpha) + x - \log(1+\alpha\exp(x))$, and
$\gamma(x) = -\alpha^{-1}\log(1-\exp(x))$.

Link functions

In a GLM, the link function maps the mean to the linear predictor $\eta_i = X_i'\beta$. Since

$$E[Y_i|X] = \gamma'(g(\eta)),$$

it follows that the link function is the inverse of $\gamma' \circ g$.

For example, in the case of logistic regression,

$$\gamma'(g(\eta)) = \exp(\eta)/(1+\exp(\eta)),$$

which is the expit function. The inverse of this function is the logit function $\log(p/(1-p))$, so the logit function is the link in this case.

Link functions

When $g(x) = x$, the resulting link function is called the canonical link function.

In the examples above, linear regression, logistic regression, and Poisson regression all used the canonical link function, but negative binomial regression did not.

The canonical link function for negative binomial regression is $1/x$, but this does not respect the domain and is harder to interpret than the usual log link.

Another setting where non-canonical links arise is the use of the log link function for logistic regression. In this case, the coefficients β are related to the log relative risk rather than to the log odds.

Overdispersion

Under the Poisson model, $\mathrm{var}[Y|X] = E[Y|X]$. A Poisson model results from using the Poisson GLM with the scale parameter ϕ fixed at 1.

The quasi-Poisson model is the Poisson model with a scale parameter that may be any non-negative value. Under the quasi-Poisson model, $\mathrm{var}[Y|X] \propto E[Y|X]$.

The negative binomial GLM allows the variance to be non-proportional to the mean.

Any situation in which $\mathrm{var}[Y|X] > E[Y|X]$ is called overdispersion. Overdispersion is often seen in practice.

One mechanism that may give rise to overdispersion is heterogeneity. Suppose we have a hierarchical model in which λ follows a Γ distribution, and $Y|\lambda$ is Poisson with mean parameter λ. Then marginally, Y is negative binomial.

Model comparison for GLM's

If ϕ is held fixed across models, then twice the log-likelihood ratio between two nested models $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ is

$$L \equiv 2\sum_i (Y_i\hat{\theta}_i^{(1)} - \gamma(\hat{\theta}_i^{(1)}))/\phi - 2\sum_i (Y_i\hat{\theta}_i^{(2)} - \gamma(\hat{\theta}_i^{(2)}))/\phi,$$

where $\hat{\theta}^{(2)}$ is nested within $\hat{\theta}^{(1)}$, so $L \geq 0$. This is called the scaled deviance.

The statistic $D = \phi L$, which does not depend explicitly on ϕ, is called the deviance.

Model comparison for GLM's

Suppose that $\hat{\theta}^{(1)}$ is the saturated model, in which $\theta_i = Y_i$. If the GLM is Gaussian and $g(x) \equiv x$, as discussed above, the deviance is

$$\begin{aligned} D &= 2\sum_i (Y_i^2 - Y_i^2/2) - 2\sum_i (Y_i \hat{\theta}_i^{(2)} - \hat{\theta}_i^{(2)\,2}/2) \\ &= \sum_i Y_i^2 - 2Y_i\hat{\theta}_i^{(2)} + \hat{\theta}_i^{(2)\,2} \\ &= \sum_i (Y_i - \hat{\theta}_i^{(2)})^2. \end{aligned}$$

Model comparison for GLM's

Thus in the Gaussian case, the deviance is the residual sum of squares for the smaller model ($\hat{\theta}^{(2)}$).

In the Gaussian case, $D/\phi = L \sim \chi^2_{n-p-1}$.

When ϕ is unknown, we can turn this around to produce an estimate of the scale parameter

$$\hat{\phi} = \frac{D}{n-p-1}.$$

This is an unbiased estimate in the Gaussian case, but is useful for any GLM.

Model comparison for GLM's

Now suppose we want to compare two nested generalized linear models with deviances $D_1 < D_2$. Let $p_1 > p_2$ be the number of covariates in each model. The likelihood ratio test statistic is

$$L_2 - L_1 = \frac{D_2 - D_1}{\phi}$$

which asymptotically has a $\chi^2_{p_1-p_2}$ distribution.

If ϕ is unknown, we can estimate it as described above (using the larger of the two models).

The "plug-in" likelihood ratio statistic $(D_2 - D_1)/\hat{\phi}$ is still asymptotically $\chi^2_{p_1-p_2}$, as long as $\hat{\phi}$ is consistent.

The finite sample distribution may be better approximated using

$$\frac{D_2 - D_1}{\hat{\phi}(p_1 - p_2)} \approx F_{p_1-p_2, n-p_1},$$

Model comparison for GLM's

We can compare any two fitted GLM's using model selection statistics like AIC or BIC.

AIC favors models having small values of $L_{\text{opt}} - \text{df}$, where L_{opt} is the maximized log-likelihood, and df is the degrees of freedom. Equivalently, the AIC can be expressed

$$-D/2\hat{\phi} - p - 1.$$

The same $\hat{\phi}$ value should be used for all models being compared (i.e. by using the one from the largest model).