Generalized linear Models

Main point: $Y = X\beta$

potentially inappropriate

1) Y not continuous

2) Y not linearly related to

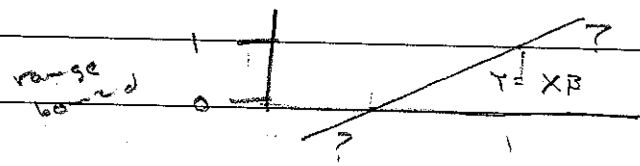
X and/or β

a) what about transformations

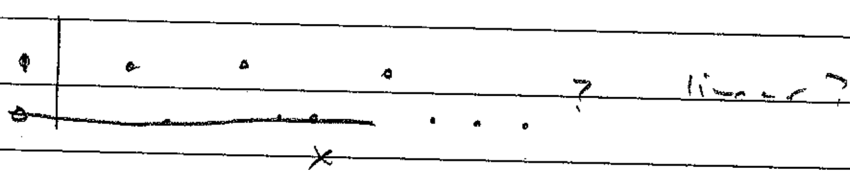
b) GLM's more flexible

Examples

1): probability of default / election



2): outcomes of sae



Another example

$Y \geq 0$

$Y = e^{Z}$, $Y = e^{X\beta}$

$\log Y = X\beta$

GLM recipe:

1) Linear predictor $\eta = X\beta$
2) Link function $g(\cdot)$
that relates the linear predictor to the mean of the outcome variable

$$\mu = g^{-1}(\eta) = g^{-1}(X\beta)$$

3) The random component specifying the distribution of the outcome variable y with mean $E(y|X) = \mu$. In general the dist of y may depend on a dispersion param ϕ

I.E. $E[y|X] = g^{-1}(X\beta)$

$$\text{where } X = \begin{array}{cc} & p \\ n & \begin{bmatrix} & \\ & \end{bmatrix} \end{array}$$

The likelihood of the data, $(X\beta)_i$ being the i^{th} linear predictor, is given by

$$p(y \mid X, \beta, \phi) = \prod_{i=1}^{n} p(y_i \mid (X\beta)_i, \phi)$$

The most common likelihoods, Poisson and Binomial do not use dispersion params.

Continuous data:

1) Normal linear model is a GLM with link func $g(\mu) = \mu = g^{-1}(\mu)$.

For all positive cont. data we can use the normal model on log data (for instance) or Gamma or Weibull.

Count data:

1) Poisson: The Poisson GLM is sometimes called the Poisson regression model.

Assume $y \mid \mu \sim \text{Poisson}(\mu)$

i.e. $p(y \mid \mu) = \frac{1}{y!} \mu^{y} e^{-\mu}$

A common (but not necessarily ideal) link function is $\log(\cdot)$, i.e.

$\log \mu = X\beta$, i.e. $\mu = e^{X\beta}$

The likelihood of the data is the

$$p(y \mid \beta) = \prod_{i=1}^{n} \frac{1}{y_i!} e^{-\exp(\eta_i)} (\exp(\eta_i))^{y_i}$$

where $\eta_i = (X\beta)_i$ is as before.

Can do Bayes estimation, ML or least squares

Binary or Probability data...

Suppose that $y_i \sim Bin(n_i, \mu_i)$ with n_i known

Then $$P(Y|B) = \prod_{i=1}^{n} \binom{n_i}{y_i} \mu_i^{y_i} (1-\mu_i)^{n_i - y_i}$$

Where μ_i must be between 0 and 1.

Let $g(\mu_i) = \log(\mu_i/(1-\mu_i))$ ← logit transform

then $g^{-1}(\cdot) = \frac{\exp(\cdot)}{1+\exp(\cdot)}$ } logistic sigmoid

$$g^{-1}(g(\mu_i)) = \frac{\exp(\log(\mu_i/(1-\mu_i)))}{1+\exp(\log(\mu_i/(1-\mu_i)))}$$

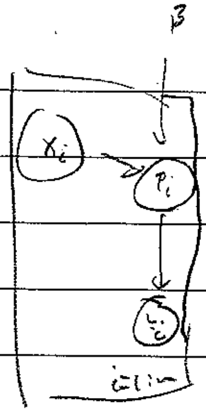
$$= \frac{\frac{\mu_i}{1-\mu_i}}{1+\frac{\mu_i}{1-\mu_i}} = \frac{\frac{\mu_i}{1-\mu_i}}{\frac{1-\mu_i}{1-\mu_i}+\frac{\mu_i}{1-\mu_i}} = \mu_i$$

Remember $\mu = g^{-1}(\eta) = g^{-1}(XB)$

so $$\mu_i = \frac{\exp(XB)}{1+\exp(x_i B)}$$ and

$$1-\mu_i = \frac{1}{1+\exp(-XB)}$$

Another interpretation of Logistic regression ...



$$p_i \sim \text{Beta}(\mu_i, 1), \quad \mu_i = \exp(x_i\beta)$$
$$Y_i \sim \text{Bernoulli}(p_i)$$

$$p(y_i \mid x_i, \beta) = \int p(Y_i \mid p_i)\, p(p_i \mid x_i, \beta)\, \partial p_i$$

$$p(Y_i \mid p_i) = p_i^{Y_i} (1-p_i)^{1-Y_i}$$

$$p(p_i \mid Y_i) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{\alpha-1} (1-p_i)^{\beta-1}$$

where $\alpha = \mu_i$ and $\beta = 1$

$$p(p_i \mid Y_i) = \frac{\Gamma(\mu_i+1)}{\Gamma(\mu_i)\Gamma(1)} p_i^{\mu_i-1} (1-p_i)^{1-1}$$

$$p(Y_i \mid x_i, \beta) = \int p_i^{Y_i} (1-p_i)^{1-Y_i} \frac{\Gamma(\mu_i+1)}{\Gamma(\mu_i)\Gamma(1)} p_i^{\mu_i-1} (1-p_i)^{1-1} dp_i$$

$$= \frac{\Gamma(\mu_i+1)}{\Gamma(\mu_i)\Gamma(1)} \int p_i^{Y_i+\mu_i-1} (1-p_i)^{1-Y_i} dp_i$$

$\beta' - 1 = 1 - Y_i$

$\beta' = 2 - Y_i$

$$= \frac{\Gamma(\mu_i+1)}{\Gamma(\mu_i)\Gamma(1)} \frac{\Gamma(Y_i+\mu_i)\Gamma(2-Y_i)}{\Gamma(\mu_i+2)}$$

$$= \begin{cases} Y_i = 1 & \dfrac{\Gamma(\mu_i+1)}{\Gamma(\mu_i)} \dfrac{\Gamma(1+\mu_i)\Gamma(1)}{\Gamma(\mu_i+2)} \\ & \to \dfrac{\mu_i}{1+\mu_i} \end{cases}$$

$$Y_i = 0 \quad \frac{\Gamma(\mu_i+1)\cancel{\Gamma(\mu_i)}\Gamma(2)}{\Gamma(\mu_i)\Gamma(\mu_i+2)}$$

$$= \frac{1}{\mu_i+1}$$

$\frac{\mu_i}{1+\mu_i} = p_i$

$1 - p_i = 1 - \frac{\mu_i}{1+\mu_i}$

$\frac{1+\mu_i}{1+\mu_i} - \frac{\mu_i}{1+\mu_i}$

$= \frac{1}{1+\mu_i}$

So $$p(\vec{Y} \mid \vec{X}, \beta) = \prod_{i=1}^{n} p(Y_i \mid X_i, \beta)$$

$$= \prod_{i=1}^{n} \left(\frac{\mu_i}{1+\mu_i}\right)^{Y_i} \left(\frac{1}{1+\mu_i}\right)^{1-Y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right)^{Y_i} \left(\frac{1}{1+e^{X_i\beta}}\right)^{1-Y_i}$$

Determining parameters

We can do gradient ascent directly on β by taking derivatives and doing gradient ascent.

Since ordinary linear regression is very fast it makes sense to use related computational tools to fit the model (ie. learn β).

If we write the joint

$$p(\vec{Y} \mid \eta, \phi) = p(Y_1, \ldots, Y_n \mid \eta, \phi) = \prod_{i=1}^{n} p(Y_i \mid \eta_i, \phi)$$

$$= \prod_{i=1}^{n} \exp\left(L(Y_i \mid \eta_i, \phi)\right)$$

Where L is the log likelihood function for the individual observations.

$$\frac{\partial}{\partial \eta_i} \frac{1}{2\sigma_i^2}(z_i-\eta_i)^2 = -\frac{1}{\sigma_i^2}(z_i-\eta_i)$$

$$\frac{\partial}{\partial \eta_i} -\frac{1}{\sigma_i^2}(z_i-\eta_i) = +\frac{1}{\sigma_i^2}$$

Remember $\eta = X\beta$

We approximate each factor in the product by a normal density in η_i;

$$\text{i.e. } L(y_i \mid \eta_i, \phi) \approx -\frac{1}{2\sigma_i^2}(z_i-\eta_i)^2 + \text{const}$$

where z_i, σ_i^2 depend on y, $\hat{\eta}_i = (X\hat{\beta})_i$, and $\hat{\phi}$ ← when an overdispersion param is involved

A standard way to fit normal approximations to likelihoods is to use a "Laplace-like" approximation, namely, matching the first and second order terms of the Taylor expansion of the likelihood.

Remember

$$f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \dots$$

near a

or in this case

$$L(y_i \mid \eta_i, \phi) \approx L(y_i \mid \hat{\eta}_i, \phi) + \frac{L'(y_i \mid \hat{\eta}_i, \phi)}{1!}(\hat{\eta}_i - \eta_i) + \frac{L''(y_i \mid \hat{\eta}_i, \phi)}{2!}(\hat{\eta}_i - \eta_i)^2 + \dots$$

this is a func. of η_i which we approx. near $\hat{\eta}_i$ our current estimate.

and

$$L(y_i \mid \eta_i, \phi) \approx \frac{1}{2\sigma_i^2}(\hat{\eta}_i - z_i)^2$$

where $\eta_i = z_i$ in this expression because we're looking for a function of z.

which we also expand

$$\frac{1}{2\sigma_i^2}(\hat{\eta}_i - z_i)^2 \approx \frac{1}{2\sigma_i^2}(\hat{\eta}_i - z_i)^2 + \frac{1}{\sigma_i^2}(\hat{\eta}_i - z_i)(\eta_i - \hat{z}_i)$$

$$+ \frac{1}{\sigma_i^2 2}(\eta_i - z_i)^2$$

So $\frac{1}{2\sigma_i^2} = \frac{L''(Y_i|\hat{\eta}_i, \phi)}{2!}$

$\Rightarrow \sigma_i^2 = \frac{1}{L''(Y_i|\hat{\eta}_i, \phi)}$

and $\frac{1}{\sigma_i^2}(\hat{\eta}_i - z_i) = \frac{L'(Y_i|\hat{\eta}_i, \phi)}{1}$

$\Rightarrow z_i = \hat{\eta}_i - \frac{L'(Y_i|\hat{\eta}_i, \phi)}{L''(Y_i|\hat{\eta}_i, \hat{\phi})}$

So now one can iteratively fit z_i, σ_i^2 then run weighted least squares, then fit z_i, σ_i^2 run weighted LS, etc until convergence.

Example - Bernoulli: logistic

$$L(y_i \mid \eta_i) = y_i \log\left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right) + (1-y_i)\log\left(\frac{1}{1+e^{\eta_i}}\right)$$

$$= y_i \eta_i - y_i \log(1+e^{\eta_i}) + \cancel{(1-y_i)\log 1}$$
$$- (1-y_i)\log(1+e^{\eta_i})$$

$$= y_i \eta_i - \log(1+e^{\eta_i})$$

Now L' and L'' can be computed and used to solve for z_i and σ_b^2

And in turn to iteratively solve for model params.