Generalized Linear Models

- Extension of linear models to the case where relationship between $E(y|X)$ and X is not linear or normal assumption is not appropriate.
- Sometimes a transformation works. Consider multiplicative model

$$y_i = x_{i1}^{b1} x_{i2}^{b2} x_{i3}^{b3} \epsilon_i$$

A simple log transformation leads to

$$\log(y_i) = b_1 \log(x_{i1}) + b_2 \log(x_{i2}) + b_3 \log(x_{i3}) + e_i$$

- When simple approaches do not work, we use GLIMs.

Generalized Linear Models (cont'd)

- There are three main components in the model:
  1. Linear predictor $\eta = X\beta$
  2. Link function $g(.)$ relating linear predictor to mean of outcome variable: $E(y|X) = \mu = g^{-1}(\eta) = g^{-1}(X\beta)$
  3. Distribution of outcome variable y with mean $\mu = E(y|X)$. Distribution can also depend on a *dispersion parameter* ϕ:

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

- In standard GLIMs for Poisson and binomial data, $\phi = 1$.
- In many applications, however, excess dispersion is present.

Some standard GLIM

- **Linear model**:
  - Simplest GLIM, with identity link function $g(\mu) = \mu$.
- **Poisson model**:
  - Mean and variance μ and link function $\log(\mu) = X\beta$, so that

$$\mu = \exp(X\beta) = \exp(\eta)$$

  - For $y = (y_1, ..., y_n)$:

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

with $\eta_i = (X\beta)_i$.

Some standard GLIM (cont'd)

- **Binomial model**: Suppose that $y_i \sim \text{Bin}(n_i, \mu_i)$, n_i known. Standard link function is logit of probability of success μ:

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = (X\beta)_i = \eta_i$$

- For a vector of data y:

$$p(y|\beta) = \Pi_{i=1}^n \binom{n_i}{y_i} \left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right)^{y_i} \left(\frac{1}{1+\exp(\eta_i)}\right)^{n_i - y_i}$$

- Another link used in econometrics is the ***probit*** link:

$$\Phi^{-1}(\mu_i) = \eta_i$$

with $\Phi(.)$ the normal cdf.
- In practice, inference from logit and probit very similar, except in extremes of the tails of the distribution.

Overdispersion

- In many applications, model can be formulated to allow for extra variability or *overdispersion*.
- E.g. in Poisson model, variance constrained to be equal to mean.
- As an example, suppose that data are the number of fatal car accidents at K intersections over T years. Covariates might include intersection characteristics and traffic control devices (stop lights, etc).
- To accommodate overdispersion: add a random effect for intersection with its own population distribution.

Setting up GLIMs

- **Canonical link functions:** Canonical link is function of mean that appears in exponent of exponential family form of sampling distribution.
- All links so far are canonical except probit.
- Can use any link in model.
- **Offset:** Arises when counts are obtained from different population sizes or volumes or time periods and we need to use an exposure. Offset is a covariate with a known coefficient.
- Example: Number of incidents in a given exposure time T are Poisson with rate μ per unit of time. Mean number of incidents is μT.
- Link function would be $\log(\mu) = \eta_i$, but here mean of y is not μ but μT.

- To apply the Poisson GLIM, add a column to X with values $\log(T)$ and fix the coefficient to 1. This is an offset.

Interpreting GLIMs

- In linear models, β_j is change in outcome when x_j is changed by one unit.
- Here, β_j reflects changes in $g(\mu)$ when x_j is changed.
- Effect of changing x_j depends of current value of x.
- To translate effects into the scale of y, measure changes relative to a baseline

$$y_0 = g^{-1}(x_0\beta).$$

- A change in x of Δx takes outcome from y_0 to y where

$$g(y_0) = x_0\beta \longrightarrow y_0 = g^{-1}(x_0\beta)$$

and

$$y = g^{-1}(g(y_0) + (\Delta x)\beta)$$

Priors in GLIM

- Focus on β although sometimes ϕ is present and has its own prior.
- *Non-informative prior for* β:
  - With $p(\beta) \propto 1$, posterior mode = MLE for β
  - Approximate posterior inference can be based on normal approximation to posterior at mode.
- *Conjugate prior for* β:
  - As in regression, express prior information about β in terms of hypothetical data obtained under same model.
  - Augment data vector and model matrix with y_0 hypothetical observations and $X_{0_{n_0 \times k}}$ hypothetical predictors.
  - Non-informative prior for β in augmented model.

Priors in GLIM (cont'd)

- *Non-conjugate priors:*
  - Often more natural to model $p(\beta|\beta_0, \Sigma_0) = N(\beta_0, \Sigma_0)$ with (β_0, Σ_0) known.
  - Approximate computation based on normal approximation (see next) particularly suitable.
- *Hierarchical GLIM*:
  - Same approach as in linear models.
  - Model some of the β as exchangeable with common population distribution with unknown parameters. Hyperpriors for parameters.

Computation

- Posterior distributions of parameters can be estimated using MCMC methods in WinBUGS or other software.
- Metropolis within Gibbs will often be necessary: in GLIM, most often full conditionals do not have standard form.
- An alternative is to **approximate** the sampling distribution with a **cleverly chosen** approximation.
- **Idea**:
  - Find mode of likelihood $(\hat{\beta}, \hat{\phi})$ perhaps conditional on hyperparameters
  - Create *pseudo-data* with their *pseudo-variances* (see later)
  - Model pseudo-data as normal with known (pseudo-)variances.

Normal approximation to likelihood

- Objective: find z_i and σ_i^2 such that normal likelihood
$$N(z_i|(X\beta)_i, \sigma_i^2)$$
is good approximation to GLIM likelihood $p(y_i|(X\beta)_i, \phi)$.
- Let $(\hat{\beta}, \hat{\phi})$ be mode of (β, ϕ) so that $\hat{\eta}_i$ is the mode of η_i.
- For L the loglikelihood, write
$$\begin{aligned} p(y_1, ..., y_n) &= \Pi_i p(y_i|\eta_i, \phi) \\ &= \Pi_i \exp(L(y_i|\eta_i, \phi)) \end{aligned}$$
- Approximate factor in exponent by normal density in η_i:
$$L(y_i|\eta_i, \phi) \approx -\frac{1}{2\sigma_i^2}(z_i - \eta_i)^2,$$
where (z_i, σ_i^2) depend on (y_i, η_i, ϕ).
- Now need to find expressions for (z_i, σ_i^2).

Normal approximation (cont'd)

- To get (z_i, σ_i^2), match first and second order terms in Taylor approx around $\hat{\eta}_i$ to (η_i, σ_i^2) and solve for z_i and for σ_i^2.
- Let $L' = \delta L / \delta \eta_i$:
$$L' = \frac{1}{\sigma_i^2}(z_i - \eta_i)$$
- Let $L'' = \delta^2 L / \delta \eta_i^2$:
$$L'' = -\frac{1}{\sigma_i^2}$$
- Then
$$\begin{aligned} z_i &= \hat{\eta}_i - \frac{L'(y_i|\hat{\eta}_i, \hat{\phi})}{L''(y_i|\hat{\eta}_i, \hat{\phi})} \\ \sigma_i^2 &= -\frac{1}{L''(y_i|\hat{\eta}_i, \hat{\phi})} \end{aligned}$$

Normal approximation (cont'd)

- Example: binomial model with logit link:
$$\begin{aligned} L(y_i, |\eta_i) &= y_i \log\left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right) \\ &\quad + (n_i - y_i)\log\left(\frac{1}{1+\exp(\eta_i)}\right) \\ &= y_i\eta_i - n_i \log(1+\exp(\eta_i)) \end{aligned}$$
- Then
$$\begin{aligned} L' &= y_i - n_i \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \\ L'' &= -n_i \frac{\exp(\eta_i)}{(1+\exp(\eta_i))^2} \end{aligned}$$
- Pseudo-data and pseudo-variances:
$$\begin{aligned} z_i &= \hat{\eta}_i + \frac{(1+\exp(\hat{\eta}_i))^2}{\exp(\hat{\eta}_i)}\left(\frac{y_i}{n_i} - \frac{\exp(\hat{\eta}_i)}{1+\exp(\hat{\eta}_i)}\right) \\ \sigma_i^2 &= \frac{1}{n_i}\frac{(1+\exp(\hat{\eta}_i))^2}{\exp(\hat{\eta}_i)} \end{aligned}$$

Models for multinomial responses

- Multinomial data: outcomes $y = (y_i, ..., y_K)$ are counts in K categories.
- Examples:
  - Number of students receiving grades A, B, C, D or F
  - Number of alligators that prefer to eat reptiles, birds, fish, invertebrate animals, or other (see example later)
  - Number of survey respondents who prefer Coke, Pepsi or tap water.
- In Chapter 3, we saw non-hierarchical multinomial models:
$$p(y|\alpha) \propto \Pi_{j=1}^k \alpha_j^{y_j}$$
with α_j: probability of jth outcome and $\Sigma_{j=1}^k \alpha_j = 1$ and $\Sigma_{j=1}^k y_j = n$.

Multinomial responses (cont'd)

- Here: we model α_j as a function of covariates (or predictors) X with corresponding regression coefficients β_j.
- For full hierarchical structure, the β_j are modeled as exchangeable with some common population distribution $p(\beta|\mu, \tau)$.
- Model can be developed as extension of either binomial or Poisson models.

Logit model for multinomial data

- Here $i = 1, ..., I$ is number of covariate patters. E.g., in alligator example, 2 sizes × four lakes = 8 covariate categories.
- Let y_i be a multinomial random variable with sample size n_i and k possible outcomes. Then

$$y_i \sim \text{ Mult } (n_i; \alpha_{i1}, ..., \alpha_{ik})$$

with $\Sigma_i y_i = n_i$, and $\Sigma_j^k \alpha_{ij} = 1$.

- α_{ij} is the probability of jth outcome for ith covariate combination.
- Standard parametrization: log of the probability of jth outcome relative to baseline category $j = 1$:

$$\log\left(\frac{\alpha_{ij}}{\alpha_{i1}}\right) = \eta_{ij} = (X\beta_j)_i,$$

with β_j a vector of regression coefficients for jth category.

Logistic regression for multinomial data (cont'd)

- Sampling distribution:

$$p(y|\beta) \propto \Pi_{i=1}^I \Pi_{j=1}^k \left(\frac{\exp(\eta_{ij})}{\Sigma_{l=1}^k \exp(\eta_{il})}\right)^{y_{ij}}.$$

- For identifiability, $\beta_1 = 0$ and thus $\eta_{i1} = 0$ for all i.
- β_j is effect of changing X on probability of category j relative to category 1.
- Typically, indicators for each outcome category are added to predictors to indicate relative frequency of each category when $X = 0$. Then

$$\eta_{ij} = \delta_j + (X\beta_j)_i$$

with $\delta_1 = \beta_1 = 0$ typically.

Poisson model for multinomial responses

- If total number of observations is fixed by design, then we can use a Poisson model to analyze multinomial responses.
- Suppose that $y = (y_1, ..., y_k)$ are independent Poisson variables with means $\lambda = (\lambda_1, ..., \lambda_k)$.
- Conditional on $n = \Sigma_j y_j$, y is multinomial:

$$p(y|n, \alpha) = \text{ Mult } (y|n; \alpha_1, ..., \alpha_k)$$

with

$$\alpha_j = \lambda_j / \sum_{l=1}^k \lambda_l$$

Poisson for multinomial (cont'd)

- Let

$$y_{ij} \sim \text{ Poi } (\lambda_{ij})$$

$$\lambda_{ij} = \lambda_{i1} \exp(x_i'\beta_j).$$

- Likelihood:

$$p(y|\beta, \lambda) \propto \Pi_i \Pi_j \lambda_{ij}^{y_{ij}} \exp(-\lambda_{ij})$$

$$\propto \Pi_i \lambda_{i1}^{n_i} \exp(\sum_j y_{ij} x_i'\beta_j) \exp(-\lambda_{i1} \sum_j \exp(x_i'\beta_j))$$

- Suppose that

$$p(\lambda_{i1}|a, b) = \text{ Gamma } (a, b),\ i = 1, ..., I$$

- Integrating $p(y|\beta, \lambda)$ with respect to λ_{i1} gives marginal likelihood of β's.

Poisson for multinomial (cont'd)

- Marginal likelihood:

$$p(y|\beta) \propto \Pi_i \frac{\exp(\Sigma_j y_{ij} x_i'\beta_j)}{[\Sigma_j \exp(x_i'\beta_j) + b]^{n_i + a}}.$$

- When $(a, b) \to 0$, marginal likelihood looks like multinomial likelihood in earlier transparencies.
- Formally, Gamma(0,0) on λ_{i1} is equivalent to a uniform prior on $\log(\lambda_{i1})$.
- Then, can reformulate problem as:

$$\log(\lambda_{ij}) = \delta_i + (X\beta_j)_i,$$

with uniform prior on δ_i.

- To implement Poisson model for multinomial responses, just fit $Poi(\lambda_{ij})$, model λ_{ij} as above, and then back-transform to multinomial probabilities using expression on page 431 in text book.

Example from WinBUGS - Alligators

- Agresti (1990) analyzes feeding choices of 221 alligators.
- Response is one of five categories: fish, invertebrate, reptile, bird, other.
- Two covariates: length of alligator (less than 2.3 meters or larger than 2.3 meters) and lake (Hancock, Oklawaha, Trafford, George).
- $2 \times 4 = 8$ covariate combinations (see data)
- For i, j a combination of size and lake, we have counts in five possible categories $y_{ij} = (y_{ij1}, ..., y_{ij5})$.

Alligators from WinBUGS

- Model

$$p(y_{ij}|\alpha_{ij}, n_{ij}) = \text{ Mult } (y_{ij}|n_{ij}, \theta_{ij1}, ..., \theta_{ij5})$$

with

$$\theta_{ijk} = \frac{\exp(\eta_{ijk})}{\Sigma_{l=1}^k \exp(\eta_{ijl})},$$

and

$$\eta_{ijk} = \delta_k + \beta_{ik} + \gamma_{jk}.$$

- Here,
  - δ_k is baseline indicator for category k
  - β_{ik} is coefficient for indicator for lake
  - γjk is coefficient for indicator for size