Generalized Linear Models

- Last time: definition of exponential family, derivation of mean and variance (*memorize*)
- Today: definition of GLM, maximum likelihood estimation
  - Include predictors $\mathbf{x}_i$ through a regression model for θ_i
  - Involves choice of a link function (*systematic component*)
  - Examples for counts, binomial data
  - Algorithm for maximizing likelihood

Systematic Component, Link Functions

Instead of modeling the mean, μ_i, as a linear function of predictors, $\mathbf{x}_i$, we introduce on one-to-one continuously differentiable transformation $g(\cdot)$ and focus on

$$\eta_i = g(\mu_i),$$

where $g(\cdot)$ will be called the <u>link function</u> and η_i the <u>linear predictor</u>.

We assume that the transformed mean follows a linear model,

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

Since the link function is invertible and one-to-one, we have

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}).$$

Note that we are transforming the expected value, μ_i, instead of the raw data, y_i.

For classical linear models, the mean is the linear predictor.

In this case, the identity link is reasonable since both μ_i and η_i can take any value on the real line.

This is not the case in general.

Link Functions for Poisson Data

For example, if $Y_i \sim \text{Poi}(\mu_i)$ then μ_i must be > 0. In this case, a linear model is not reasonable since for some values of $\mathbf{x}_i$ $\mu_i \leq 0$.

By using the model, $\eta_i = \log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$, we are guaranteed to have $\mu_i > 0$ for all $\boldsymbol{\beta} \in \Re^p$ and all values of $\mathbf{x}_i$.

In general, a link function for count data should map the interval $(0, \infty) \rightarrow \Re$ (i.e., from the + real numbers to the entire real line).

The log link is a natural choice

Link Functions for Binomial Data

For the binomial distribution, $0 < \mu_i < 1$ (mean of y_i is $n_i\mu_i$)

Therefore, the link function should map from $(0, 1) \rightarrow \Re$

Standard choices:

1. *logit*: $\eta_i = \log\{\mu_i/(1 - \mu_i)\}$.
2. *probit*: $\eta_i = \Phi^{-1}(\mu_i)$, where $\Phi(\cdot)$ is the $N(0, 1)$ cdf.
3. *complementary log-log*: $\eta_i = \log\{-\log(1 - \mu_i)\}$.

Each of these choices is important in applications & will be considered in detail later in the course

Recall that the exponential family density has the following form:

$$f(y_i; \theta_i, \phi) = \exp\Big\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\Big\}.$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions.

Specifying the GLM involves choosing $a(\cdot), b(\cdot), c(\cdot)$:

1. Specify $a(\cdot), c(\dot{)}$ to correspond to particular distribution (e.g., Binomial, Poisson)

2. Specify $b(\cdot)$ to correspond to a particular link function

Recall that mean & variance are

$$\mu_i = b'(\theta_i) \quad \text{and} \quad \sigma^2 = b''(\theta_i)\phi.$$

Using $b'(\theta_i) = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$, we can express the density as $f(y_i; \mathbf{x}_i, \boldsymbol{\beta}, \phi)$, so that the conditional likelihood of y_i given $\mathbf{x}_i$ depends on parameters $\boldsymbol{\beta}$ and ϕ.

It would seem that a natural choice for $b(\cdot)$ and hence $g(\cdot)$, would correspond to $\theta_i = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}$, so that $b'(\cdot)$ is the inverse link

Canonical Links and Sufficient Statistics

Each of the distributions we have considered has a special, canonical, link function for which there exists a sufficient statistic equal in dimension to $\boldsymbol{\beta}$.

Canonical links occur when $\theta_i = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}$, with θ_i the canonical parameter

As a homework exercise, please show for next Thursday that the following distributions are in the exponential family and have the listed canonical links:

Normal	$\eta_i = \mu_i$
Poisson	$\eta_i = \log \mu_i$
binomial	$\eta_i = \log\{\mu_i/(1-\mu_i)\}$
gamma	$\eta_i = \mu_i^{-1}$

For the canonical links, the sufficient statistic is $\mathbf{X}'\mathbf{y}$, with components $\sum_i x_{ij} y_i$, for $j = 1, \ldots, p$.

Although canonical links often nice properties, selection of the link function should be based on prior expectation and model fit

<u>Example: Logistic Regression</u>

Suppose $y_i \sim \text{Bin}(1, p_i)$, for $i = 1, \dots, n$, are independent 0/1 indicator variables of an adverse response (e.g., preterm birth) and let $\mathbf{x}_i$ denote a $p \times 1$ vector of predictors for individual i (e.g., dose of dde exposure, race, age, etc).

The likelihood is as follows:

$$
\begin{aligned}
f(\mathbf{y} \,|\, \boldsymbol{\beta}) &= \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = \prod_{i=1}^{n} \Big(\frac{p_i}{1-p_i}\Big)^{y_i}(1-p_i) \\
&= \prod_{i=1}^{n} \exp\Big\{y_i \log\Big(\frac{p_i}{1-p_i}\Big) - \log\Big(\frac{1}{1-p_i}\Big)\Big\} \\
&= \exp\Big[\sum_{i=1}^{n} \{y_i\theta_i - \log(1+e^{\theta_i})\}\Big].
\end{aligned}
$$

Choosing the canonical link,

$$\theta_i = \log\Big(\frac{p_i}{1-p_i}\Big) = \mathbf{x}_i'\boldsymbol{\beta},$$

the likelihood has the following form:

$$f(\mathbf{y} \mid \boldsymbol{\beta}) = \exp[\sum_{i=1}^{n} \{y_i \mathbf{x}_i'\boldsymbol{\beta} - \log(1 + e^{\mathbf{x}_i'\boldsymbol{\beta}})\}].$$

This is logistic regression, which is widely used in epidemiology and other applications for modeling of binary response data.

In general, if $f(y_i; \theta_i, \phi)$ is in the exponential family and $\theta_i = \theta(\eta_i)$, $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$, then the model is called a generalized linear model (GLM)

Model fitting

- Choosing a GLM results in a likelihood function:

$$L(\mathbf{y};\boldsymbol{\beta},\phi,\mathbf{x}) = \prod_{i=1}^{n} \exp\Big\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\Big\},$$

where θ_i is a function of $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$.

- The maximum likelihood estimate is defined as

$$\widehat{\boldsymbol{\beta}} = \sup_{\boldsymbol{\beta}} L(\mathbf{y};\boldsymbol{\beta},\phi,\mathbf{x}),$$

with ϕ initially assumed to be known

- Frequentist inferences for GLMs typically rely on $\widehat{\boldsymbol{\beta}}$ and asymptotic approximations.

- In the normal linear model special case, the MLE corresponds to the least squares estimator

- In general, there is no closed form expression so we need an algorithm to calculate $\widehat{\boldsymbol{\beta}}$.

Maximum Likelihood Estimation of GLMs

All GLMs can be fit using the same algorithm, a form of iteratively re-weighted least squares:

1. Given an initial value for $\widehat{\boldsymbol{\beta}}$, calculate the estimated linear predictor $\hat{\eta}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ and use that to obtain the fitted values $\widehat{\mu}_i = g^{-1}(\hat{\eta}_i)$. Calculate the adjusted dependent variable,

$$z_i = \hat{\eta}_i + (y_i - \widehat{\mu}_i)\Big(\frac{d\eta_i}{d\mu_i}\Big)_0,$$

where the derivative is evaluated at $\widehat{\mu}_i$.

2. Calculate the iterative weights

$$W_i^{-1} = \Big(\frac{d\eta_i}{d\mu_i}\Big)_0 V_i.$$

where V_i is the variance function evaluated at $\widehat{\mu}_i$.

3. Regress z_i on $\mathbf{x}_i$ with weight W_i to give new estimates of $\boldsymbol{\beta}$

Justification for the IWLS procedure

Note that the log-likelihood can be expressed as

$$l = \sum_{i=1}^{n} \{y_i\theta_i - b(\theta_i)\}/a(\phi) + c(y_i, \phi).$$

To maximize this log-likelihood we need $\partial l/\partial \beta_j$,

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{a(\phi)} \frac{1}{V_i} \frac{d\mu_i}{d\eta_i} x_{ij}, \\ &= \sum_{i=1}^{n} (y_i - \mu_i) \frac{W_i}{a(\phi)} \frac{d\eta_i}{d\mu_i} x_{ij} \end{aligned}$$

since $\mu_i = b'(\theta_i)$ and $b''(\theta_i) = V_i$ implies $d\mu_i/d\theta_i = V_i$.

With constant dispersion ($a(\phi) = \phi$), the MLE equations for β_j:

$$\sum_{i=1}^{n} W_i(y_i - \mu_i) \frac{d\eta_i}{d\mu_i} x_{ij} = 0.$$

Fisher's scoring method uses the gradient vector, $\partial l/\partial \boldsymbol{\beta} = \mathbf{u}$, and minus the expected value of the Hessian matrix

$$-E\Big(\frac{\partial^2 l}{\partial \beta_r \partial \beta_s}\Big) = \mathbf{A}.$$

Given the current estimate $\mathbf{b}$ of $\boldsymbol{\beta}$, choose the adjustment $\delta\mathbf{b}$ so

$$\mathbf{A}\delta\mathbf{b} = \mathbf{u}.$$

Excluding ϕ, the components of $\mathbf{u}$ are

$$u_r = \sum_{i=1}^{n} W_i(y_i - \mu_i)\, \frac{d\eta_i}{d\mu_i}\, x_{ir},$$

so we have $A_{rs} = -E(\partial u_r/\partial \beta_s) =$

$$-E \sum_{i=1}^{n} \Big[(y_i - \mu_i)\frac{\partial}{\partial \beta_s}\Big\{W_i \frac{d\eta_i}{d\mu_i} x_{ir}\Big\} + W_i \frac{d\eta_i}{d\mu_i} x_{ir} \frac{\partial}{\partial \beta_s}(y_i - \mu_i)\Big].$$

The expectation of the first term is 0 and the second term is

$$\sum_{i=1}^{n} W_i \frac{d\eta_i}{d\mu_i} x_{ir} \frac{\partial \mu_i}{\partial \beta_s} = \sum_{i=1}^{n} W_i \frac{d\eta_i}{d\mu_i} x_{ir} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_s} = \sum_{i=1}^{n} W_i x_{ir} x_{is}.$$

The new estimate $\mathbf{b}^* = \mathbf{b} + \delta\mathbf{b}$ of $\boldsymbol{\beta}$ thus satisfies

$$\mathbf{A}\mathbf{b}^* = \mathbf{A}\mathbf{b} + \mathbf{A}\delta\mathbf{b} = \mathbf{A}\mathbf{b} + \mathbf{u},$$

where $(\mathbf{A}\mathbf{b})_r = \Sigma_s A_{rs} b_s = \Sigma_{i=1}^n W_i x_{ir} \eta_i$.

Thus, the new estimate $\mathbf{b}^*$ satisfies

$$(\mathbf{A}\mathbf{b}^*)_r = \sum_{i=1}^n W_i x_{ir} \{\eta_i + (y_i - \mu_i) d\eta_i / d\mu_i\}.$$

These equations have the form of linear weighted least squares equation with weight W_i and dependent variable z_i.

Some Comments

- The IWLS procedure is simple to implement and converges rapidly in most cases
- Procedures are available to calculate MLEs and implement frequentist inferences for GLMs in most software packages.
- In R or S-PLUS the $glm(\cdot)$ function can be used - try help(glm)
- In Matlab the $glmfit(\cdot)$ function can be used

Example: Smoking and Obesity

- $y_i = 1$ if the child is obese and $y_i = 0$ otherwise, for $i = 1, \ldots, n$
- $\mathbf{x}_i = (1, age_i, smoke_i, age_i \times smoke_i)'$
- Bernoulli likelihood,

$$L(\mathbf{y}; \boldsymbol{\beta}, \mathbf{x}) = \prod_{i=1}^{n} \mu_i^{y_i}(1 - \mu_i)^{1-y_i},$$

where $\mu_i = \Pr(y_i = 1 \mid \mathbf{x}_i, \boldsymbol{\beta})$.

- Choosing the canonical link, $\mu_i = 1/\{1 + \exp(-\mathbf{x}_i'\boldsymbol{\beta})\}$, results in a logistic regression model:

$$\Pr(y_i = 1 \mid \mathbf{x}_i, \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})},$$

Hence, probability of obesity depends on age and smoking through a non-linear model

- Letting X = cbind(age,smoke,age*smoke) and Y = 0/1 obesity outcome in R, we use

```
fit<- glm(Y ~ age + smoke + age*smoke, family=binomial,
                data=obese)
```

to implement IWLS and fit the model

- Note that data are available on the web - try to replicate results (note children a year or younger have been discarded)

- The command summary(glm) yields the results:

```
Coefficients:
                   Value Std. Error    t value
(Intercept) -2.365173738 0.50112688 -4.7197104
        age -0.066204429 0.08957593 -0.7390873
      smoke -0.043079741 0.22375895 -0.1925275
  age:smoke -0.008448488 0.04010827 -0.2106420

    Null Deviance: 1580.905 on 3874 degrees of freedom
Residual Deviance: 1574.663 on 3871 degrees of freedom

Number of Fisher Scoring Iterations: 6

Correlation of Coefficients:
          (Intercept)        age      smoke
      age -0.9382877
    smoke -0.9067235   0.8520241
age:smoke  0.8496495  -0.9062117 -0.9391875
```

- Thus, the IWLS algorithm converged in 6 iterations to the MLE:

$$\widehat{\boldsymbol{\beta}} = (-2.365, -0.066, -0.043, -0.008)'$$

- For any value of the covariates we can calculate the probability of obesity

- For example, for non-smokers the age curves can be plotted by using:

```
beta<- fit$coef

## introduce grid spanning range of observed ages

x<- seq(min(obese$age),max(obese$age),length=100)

## calculate fitted probability of obesity

py<- 1/(1+exp(-beta[1]+beta[2]*x))

plot(x,py,xlab="age in years", ylab="Pr(obesity)")
```

- Meaning of the rest of the R/S-PLUS output will be clear after next class

Next Class

Topic: Frequentist inference for GLMs

Have homework exercise completed and written up for next Thursday

Complete the following exercise:

1. Write down generalized linear models for the Caesarian data (grouping the two different infection types) and the cellular differentiation data.

2. Show the different components of the GLM, expressing the likelihood in exponential family form & using a canonical link function

3. Fit the GLM using maximum likelihood and report the parameter estimates.