Generalized Linear Models

Jonathan Moyer and Heather Weaver, based on Agresti Ch 4

September 19, 2017

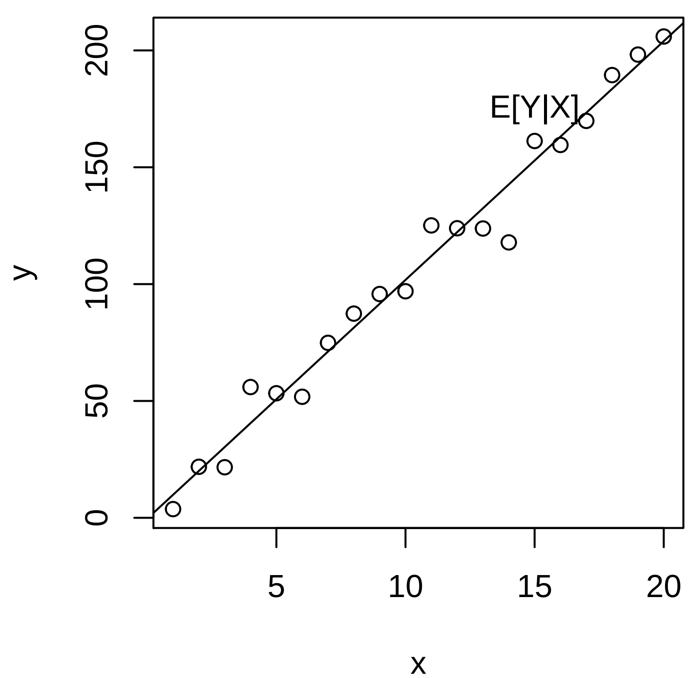
Generalized Linear Models (GLMS)

- Extensions to linear models (LMs).
- Outcomes are can be non-normal, expressed as functions on the mean.
- Example for ordinary LM:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

$$\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- The best fit line on the following plot represents the estimated value of Y given the value of X.



History and Overview of GLMs

- 1920s: RA Fisher develops score equations.
- 1970s: GLMs unified by Nelder and Wedderburn in the 1970s.

- Fairly flexible parametric framework (parametric still implies some rigidity).
- Good at describing relationships and associations, interpretable.
- Predictions: GLMs are an "outdated" way to make predictions, more popular models tend to be nonparametric, require fewer assumptions.

Notation

There are three components to a GLM:

Random Component

- Response/outcome variable Y, N observations:
- $f(y_i|\theta_i) = a(\theta_i)b(y_i)\exp\{y_i Q(\theta_i)\}$
- $Q(\theta_i)$ is the **natural parameter**

Systematic Component

- The "linear model" part.
- $\eta_i = X_i\beta$

Link Function

- Connects the mean of the original response scale to the systematic component
- $\mu_i = E[Y_i|X_i]$
- $\eta_i = g(\mu_i) = X_i\beta$
- $g(\mu_i)$ is the link function

Example 1: Binomial Logit (ie, Logistic Regression)

- For binary outcome data
- $Pr(Y_i = 1) = \pi_i = E(Y_i|X_i)$
- $f(y_i|\theta_i) = \pi^{y_i}(1-\pi_i)^{1-y_i} = (1-\pi_i)\left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} = (1-\pi_i)\exp\left\{y_i \log\frac{\pi_i}{1-\pi_i}\right\}$
- Where:
  - $\theta = \pi_i$
  - $a(\pi_i) = 1 - \pi_i$
  - $b(y_i) = 1$
  - $Q(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$
- The natural parameter $Q(\pi_i)$ implies the canonical link function: $\text{logit}(\pi) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$

Example 2: Poisson

- For count outcome data
- $Y_i \sim Pois(\pi_i)$
- $f(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = e^{-\mu_i}\left(\frac{1}{y_i}\right)\exp\{y_i \log\mu_i\}$
- Where:

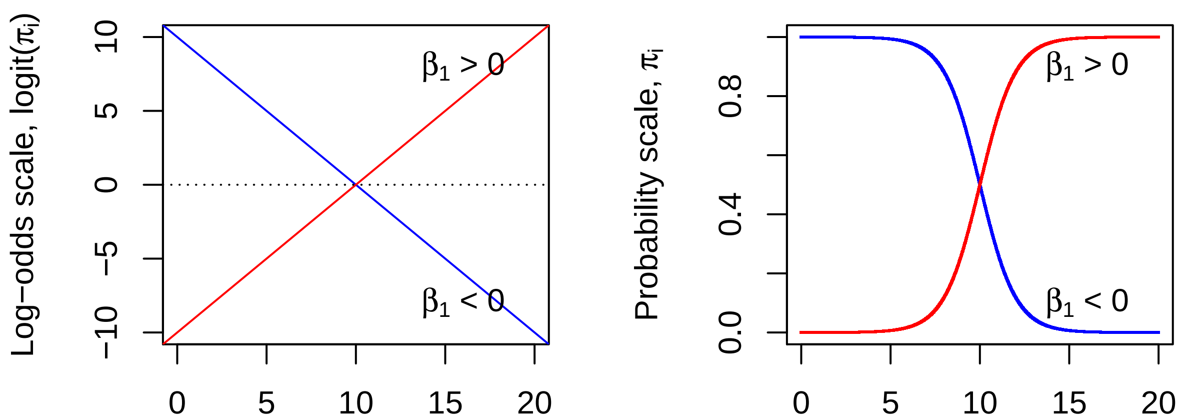
- $\theta = \mu_i$
- $a(\mu_i) = e^{-\mu_i}$
- $b(y_i) = \left(\frac{1}{y_i}\right)$
- $Q(\mu_i) = \log \mu_i$

Back to Logistic Regression

- For "simple" one predictor case:

$$\text{logit}(\pi_i) = \text{logit}[Pr(Y_i = 1|X_i)] = \beta_0 + \beta_1 x_{i1}$$

- The graphs below depict the correspondence between the linear systematic component and the logit link



Coefficient Interpretation

- A useful exercise everyone should do is write out what it means for their to be a one-unit change.
- This will be done for logistic regression.
- The value of the logit at a value $X_i = k$ is given by:

$$\text{logit}(Pr(Y_i = 1|X_i = k)) = \beta_0 + \beta_1 k$$

- A one unit increase $k + 1$ is:

$$\text{logit}(Pr(Y_i = 1|X_i = k+1)) = \beta_0 + \beta_1(k+1)$$

- Subtracting the first equation from the second:

$$\text{logit}(Pr(Y_i = 1|X_i = k+1)) - \text{logit}(Pr(Y_i = 1|X_i = k)) = \beta 1$$

- The difference of logits can be expressed as follows:

$$\log\left\{\frac{odds(Pr(Y_i|X_i=k+1))}{odds(Pr(Y_i|X_i=k))}\right\}=\beta 1$$

- The argument of the log function in the preceding equation is the **odds ratio**. - So we can write $\log OR=\beta 1$ or $\log OR=e^{\beta 1}$. - A one-unit increased in X_i implies the OR changes by a factor of $e^{\beta 1}$.

Note: this type of question is bread and butter for the midterm!

GLM Likelihood

- The GLM likelihood function is given as follows:

$$f(y|\theta,\phi)=\exp\left\{\frac{y\theta-b(\theta)}{a(\phi)}+C(y,\phi)\right\}$$

- Here ϕ is a dispersion parameter, allows for more than one parameter.

$$L(\overrightarrow{\beta})=\Sigma_i L_i=\Sigma_i\log f(y_i|\theta_i,\phi)=\sum_i\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+\Sigma_i C(y_i,\phi)$$

- θ_i is the βX_i from η_i
- ϕ not indexed by i, assumed to be fixed.

Likelihood/Score Equations

The maximum of the multivariate log-likelihood can be found by solving the **score equations** below:

$$\frac{\partial L(\overrightarrow{\beta})}{\partial\beta_j}=\sum_i\frac{\partial L_i}{\partial\beta_j}=0,\forall j$$

$$\sum_{i=1}^{N}\frac{(y_i-\mu_i)x_{ij}}{Var(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}=0$$

Where $\mu_i=E[Y_i|x_i]=g^{-1}(X\beta)$.

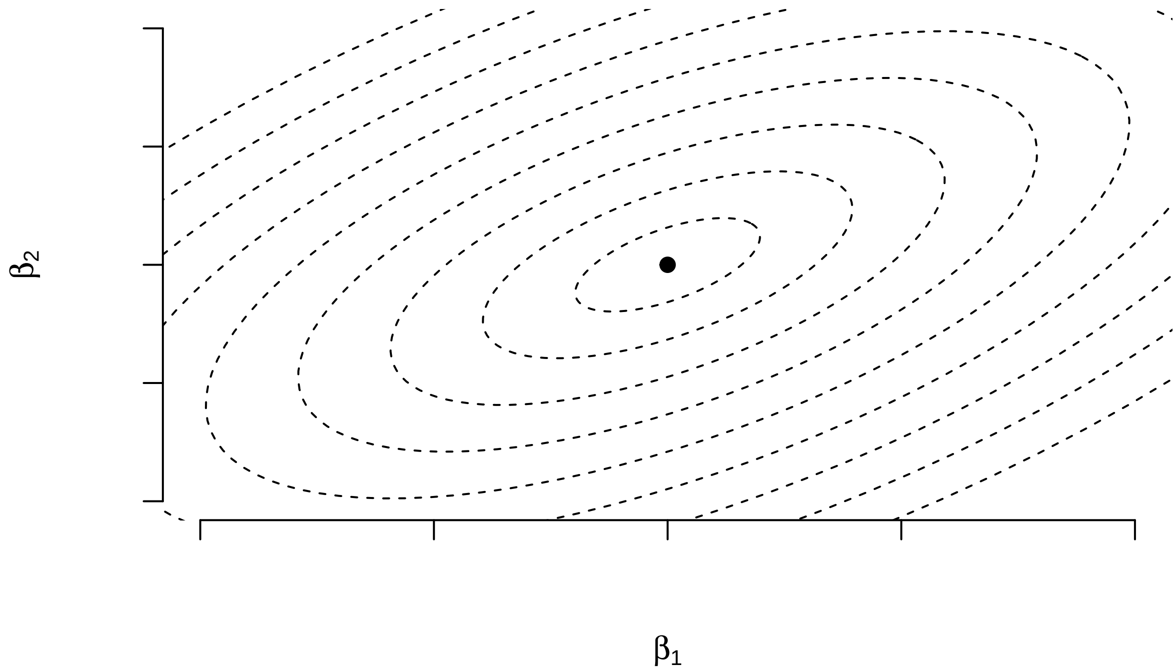
- There is no easy way to solve this!
- Statistical software does this by iteratively refining estimates of parameters until convergence is achieved.

Likelihood Example: Binomial

$$\frac{\partial L(\overrightarrow{\beta})}{\partial\beta_j}=\Sigma_i(y_i-n_i\pi_i)x_{ij}$$

Here, $\pi_i=\frac{e^{X_i\beta}}{1+e^{X_i\beta}}$.

A "top down" view of the likelihood function is sketched out below. The black dot in the center marks the "peak" that maximized the likelihood.



Asymptotic Convergence of the MLE Estimate $\hat{\beta}$

- The likelihood function determines the covergence of $\hat{\beta}$.
- The information matrix, $\mathcal{I}$, has hj elements

$$\mathcal{I} = E\Big[\frac{-\partial^2 L(\vec{\beta})}{\partial \beta_h \beta_j}\Big] = \sum_{i=1}^{N} \frac{x_{ih} x_{ij}}{Var(Y_i)} \Big(\frac{\partial \mu_i}{\partial \eta_i}\Big)^2 = \sum_{i=1}^{N} x_{ih} x_{ij} w_i$$

Where

$$w_i = \frac{1}{Var(Y_i)} \Big(\frac{\partial \mu_i}{\partial \eta_i}\Big)^2$$

Thus, if matrix W is a diagonal matrix with w_i as a diagonal element, $\mathcal{I} = X^T W X$. -Here X is a $p \times n$ and W is $n \times n$.

Properties of the $\hat{\beta}$ Estimators

- For simple linear regression: $Var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$
- Matrix notation: $Cov(\hat{\beta}_1) = \hat{\sigma}^2 (X^T X)^{-1}$
- For GLMs: $Cov(\hat{\beta}_1) = \hat{\mathcal{I}}^{-1} = (X^T W X)^{-1}$