Lecture 6: GLMs

Author: Nicholas Reich Transcribed by Nutcha Wattanachit/Edited by Bianca Doone

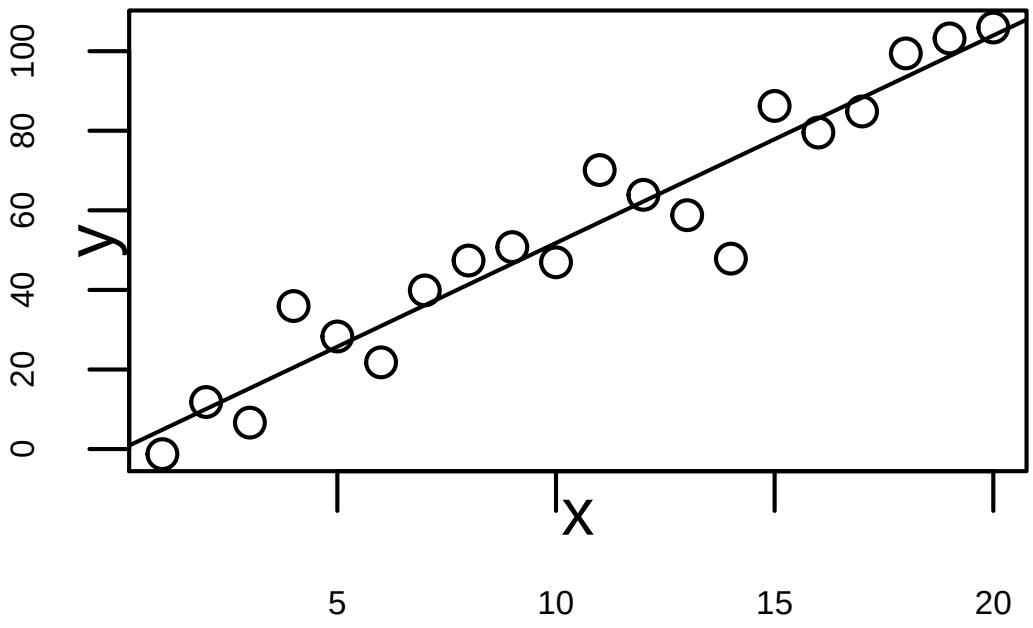
Course: Categorical Data Analysis (BIOSTATS 743)

Generalized Linear Models (GLMS)

GLMs are extensions or generalization of linear regression models to encompass non-normal response distribution and modeling functions of the mean . - Example for ordinary LM:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \qquad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

The best fit line on the following plot represents $E(Y|X)$.



Overview of GLMs

- Early versions of GLM were used by Fisher in 1920s and GLM theories were unified in 1970s.
- Fairly flexible parametric framework, good at describing relationships and associations between variables
- Fairly simple ('transparent') and interpretable, but basic GLMs are not generally seen as the best approach for predictions.
- Both frequentist and Bayesian methods can be used for parametric and nonparametric models.

GLMs: Parametric vs. Nonparametric Models

- **Parametric models**: Assumes data follow a fixed distribution defined by some parameters. GLMs are examples of parametric models. If assumed model is "close to" the truth, these methods are both accurate and precise.
- **Nonparametric models**: Does not assume data follow a fixed distribution, thus could be a better approach if assumptions are violated.

Components of GLMs

1. Random Component: Response variable Y with N observations from a distribution in the exponential family:

- One parameter: $f(y_i|\theta_i) = a(\theta_i)b(y_i)\exp\{y_i Q(\theta_i)\}$
- Two parameters: $f(y_i|\theta_i, \Phi) = exp\{\frac{y_i\theta_i - b(\theta_i)}{a(\Phi)} + c(y_i, \Phi)\}$, where Φ is fixed for all observations
- $Q(\theta_i)$ is the **natural parameter**

2. Systematic Component: The linear predictor relating η_i to X_i:

- $\eta_i = X_i\beta$

3. Link Function: Connects random and systematic components

- $\mu_i = E(Y_i)$
- $\eta_i = g(\mu_i) = g(E(Y_i|X_i)) = X_i\beta$
- $g(\mu_i)$ is the link function of μ_i

$g(\mu) = \mu$, called the identity link, has $\eta_i = \mu_i$ (a linear model for a mean itself).

Example 1: Normal Distribution (with fixed variance)

Suppose y_i follows a normal distribution with

- mean $\mu_i = \hat{y}_i = E(Y_i|X_i)$
- fixed variance σ^2. The pdf is defined as

$$f(y_i|\mu_i, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)}} exp\{\frac{-(y_i - \mu_i)^2}{2\sigma^2}\}$$
$$= \frac{1}{\sqrt{(2\pi\sigma^2)}} exp\{\frac{-y_i^2}{2\sigma^2}\} exp\{\frac{2y_i\mu_i}{2\sigma^2}\} exp\{\frac{-\mu_i^2}{2\sigma^2}\}$$

- Where:
  - $\theta = \mu_i$
  - $a(\mu_i) = exp\{\frac{-\mu_i^2}{2\sigma^2}\}$
  - $b(y_i) = exp\{\frac{-y_i^2}{2\sigma^2}\}$
  - $Q(\mu_i) = exp\{\frac{\mu_i}{\sigma^2}\}$

Example 2: Binomial Logit for binary outcome data

- $Pr(Y_i = 1) = \pi_i = E(Y_i|X_i)$
- $$f(y_i|\theta_i) = \pi^{y_i}(1-\pi_i)^{1-y_i} = (1-\pi_i)\Big(\frac{\pi_i}{1-\pi_i}\Big)^{y_i}$$
$$= (1-\pi_i)\exp\Big\{y_i \log\frac{\pi_i}{1-\pi_i}\Big\}$$
- 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 3: Poisson for count outcome data

- $Y_i \sim Pois(\mu_i)$
- $$f(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$
$$= e^{-\mu_i}\left(\frac{1}{y_i}\right)\exp\{y_i \log \mu_i\}$$
- Where:
  - $\theta = \mu_i$
  - $a(\mu_i) = e^{-\mu_i}$
  - $b(y_i) = \left(\frac{1}{y_i}\right)$
  - $Q(\mu_i) = \log \mu_i$

Deviance

For a particular GLM for observations $y = (y_1, ..., y_N)$, let $L(\mu|y)$ denote the log-likelihood function expressed in terms of the means $\mu = (\mu_1, ..., \mu_N)$. The deviance of a Poisson or binomial GLM is

$$D = -2[L(\hat{\mu}|y) - L(y|y)]$$

- $L(\hat{\mu}|y)$ denotes the maximum of the log likelihood for $y_1, ..., y_n$ expressed in terms of $\hat{\mu}_1, ..., \hat{\mu}_n$
- $L(y|y)$ is called a saturated model (a perfect fit where $\hat{\mu}_i = y_i$, representing "best case" scenario). This model is not useful, since it does not provide data reduction. However, it serves as a baseline for comparison with other model fits.
- Relationship with LRTs: This is the likelihood-ratio statistic for testing the null hypothesis that the model holds against the general alternative (i.e., the saturated model)

Logistic Regression

For "simple" one predictor case where $Y_i \sim Bernoulli(\pi_i)$ and $Pr(Y_i = 1) = \pi_i$:

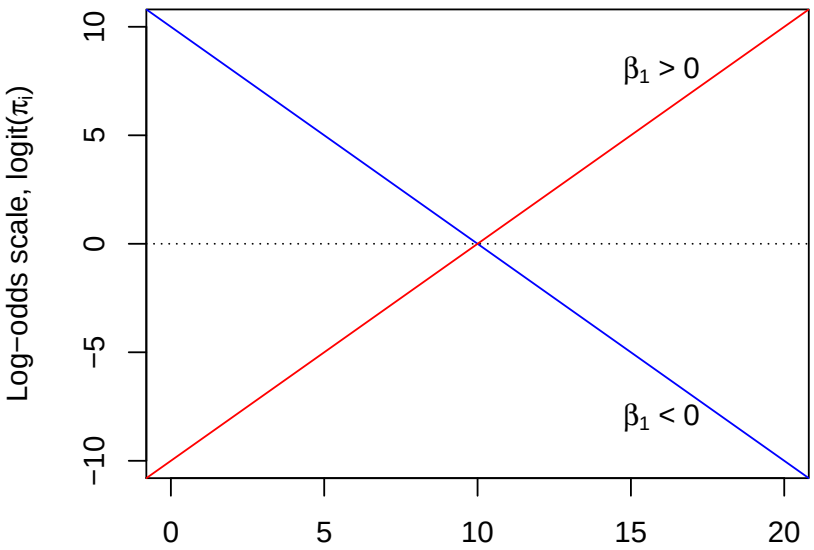
$$\begin{aligned} logit(\pi_i) &= log\left(\frac{\pi_i}{1-\pi_i}\right) \\ &= logit(Pr(Y_i = 1)) \\ &= logit(E[Y_i]) \\ &= g(E[Y_i]) \\ &= X\beta \\ &= \beta_0 + \beta_i x_i, \end{aligned}$$

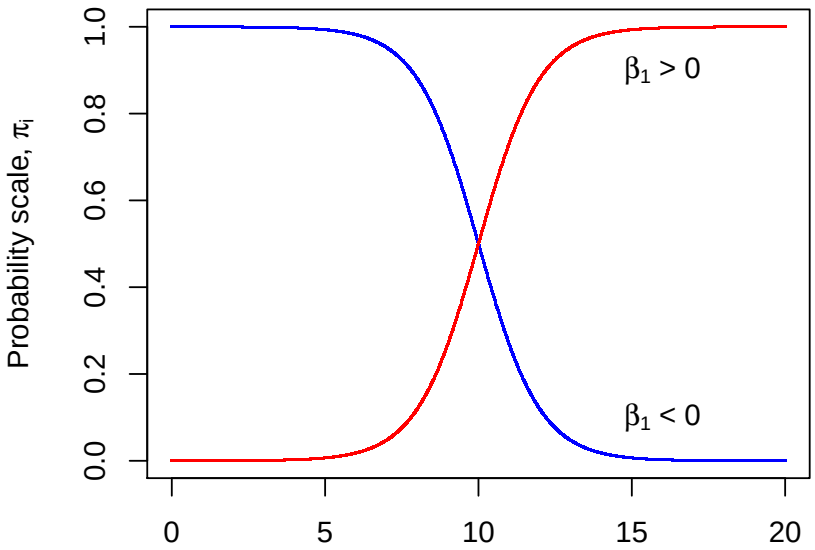
which implies $Pr(Y_i = 1) = \frac{e^{X\beta}}{1+e^{X\beta}}$.

- g does not have to be a linear function (linear model means linear with respect to β).

Logistic Regression (Cont.)

The graphs below illustrate the correspondence between the linear systematic component and the logit link. The logit transformation restricts the range Y_i to be between 0 and 1.





Example: Linear Probability vs. Logistic Regression Models

- For a binary response, the linear probability model $\pi(x) = \alpha + \beta_1 X_1 + ... + \beta_p X_p$ with independent observations is a GLM with binomial random component and identity link function
- Logistic regression model is a GLM with binomial random component and logit link function

Example: Linear Probability vs. Logistic Regression Models (Cont.)

An epidemiological survey of 2484 subjects to investigate snoring as a risk factor for heart disease.

```
n<-c(1379, 638, 213, 254)
snoring<-rep(c(0,2,4,5),n)
y<-rep(rep(c(1,0),4),c(24,1355,35,603,21,192,30,224))
```

TABLE 4.2 Relationship between Snoring and Heart Disease

Snoring	Heart Disease		Proportion	Linear	Logit
	Yes	No	Yes	Fit[a]	Fit[a]
Never	24	1355	0.017	0.017	0.021
Occasionally	35	603	0.055	0.057	0.044
Nearly every night	21	192	0.099	0.096	0.093
Every night	30	224	0.118	0.116	0.132

[a]Model fits refer to proportion of yes responses.

Source: P. G. Norton and E. V. Dunn, *British Med. J.* **291**: 630–632 (1985), BMJ Publishing Group.

Example: Linear Probability vs. Logistic Regression Models (Cont.)

```
library(MASS)
logitreg <- function(x, y, wt = rep(1, length(y)),
                     intercept = T, start = rep(0, p), ...)
{
  fmin <- function(beta, X, y, w) {
    p <- plogis(X %*% beta)
    -sum(2 * w * ifelse(y, log(p), log(1-p)))
  }
  gmin <- function(beta, X, y, w) {
    eta <- X %*% beta; p <- plogis(eta)
    t(-2 * (w *dlogis(eta) * ifelse(y, 1/p, -1/(1-p))))%*% X
  }
  if(is.null(dim(x))) dim(x) <- c(length(x), 1)
  dn <- dimnames(x)[[2]]
  if(!length(dn)) dn <- paste("Var", 1:ncol(x), sep="")
  p <- ncol(x) + intercept
  if(intercept) {x <- cbind(1, x); dn <- c("(Intercept)", dn)}
  if(is.factor(y)) y <- (unclass(y) != 1)
  fit <- optim(start, fmin, gmin, X = x, y = y, w = wt, ...)
  names(fit$par) <- dn
  invisible(fit)
}
logit.fit<-logitreg(x=snoring, y=y, hessian=T, method="BFGS")
logit.fit$par
```

```
## (Intercept)        Var1
##   -3.866245    0.397335
```

- Logistic regression model fit: $logit[\hat{\pi}(x)] = -3.87 + 0.40x$

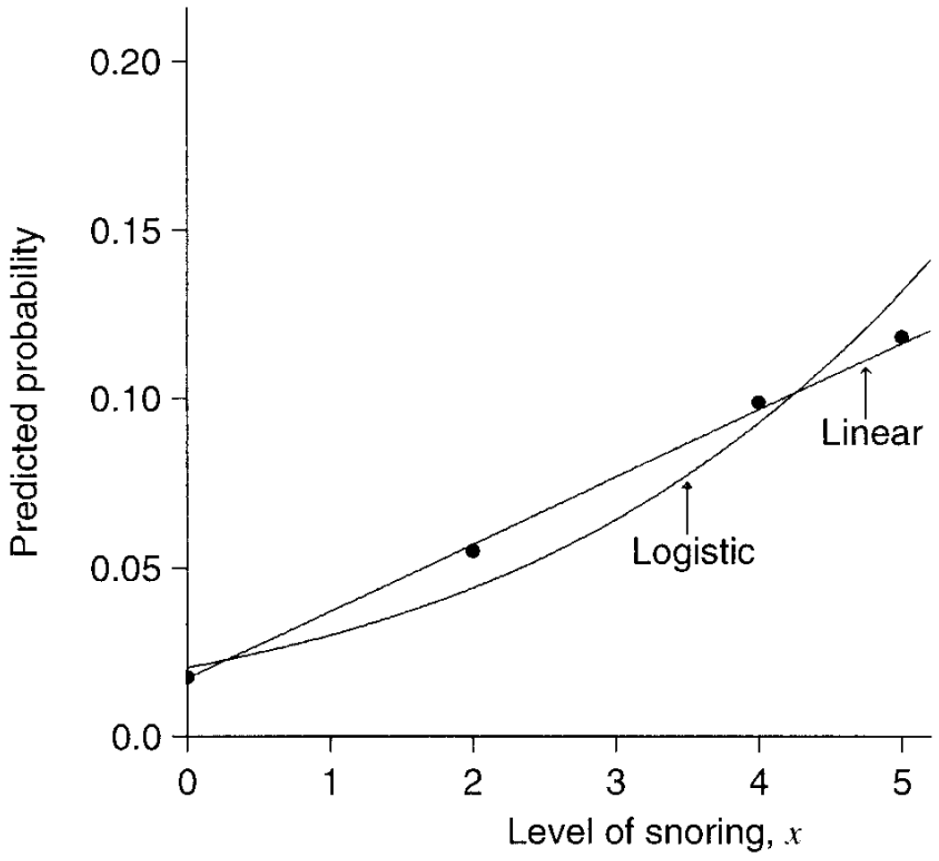
Example: Linear Probability vs. Logistic Regression Models (Cont.)

```
lpmreg <- function(x, y, wt = rep(1, length(y)),
                     intercept = T, start = rep(0, p), ...)
{
  fmin <- function(beta, X, y, w) {
    p <- X %*% beta
    -sum(2 * w * ifelse(y, log(p), log(1-p)))
  }
  gmin <- function(beta, X, y, w) {
    p <- X %*% beta;
    t(-2 * (w * ifelse(y, 1/p, -1/(1-p))))%*% X
  }
  if(is.null(dim(x))) dim(x) <- c(length(x), 1)
  dn <- dimnames(x)[[2]]
  if(!length(dn)) dn <- paste("Var", 1:ncol(x), sep="")
  p <- ncol(x) + intercept
  if(intercept) {x <- cbind(1, x); dn <- c("(Intercept)", dn)}
  if(is.factor(y)) y <- (unclass(y) != 1)
  fit <- optim(start, fmin, gmin, X = x, y = y, w = wt, ...)
  names(fit$par) <- dn
  invisible(fit)
}
lpm.fit<-lpmreg(x=snoring, y=y, start=c(.05,.05), hessian=T, method="BFGS")
lpm.fit$par
```

```
## (Intercept)        Var1
##  0.01724645  0.01977784
```

- Linear probability model fit: $\hat{\pi}(x) = 0.0172 + 0.0198x$

Example: Linear Probability vs. Logistic Regression Models (Cont.)



Coefficient Interpretation in Logistic Regression

Our goal is to say in words what β_j is. Consider

$$logit(Pr(Y_i = 1)) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ...$$

- The logit function at $X_i = k$ and at one-unit increase $k + 1$ are given by:

$$\text{logit}(Pr(Y_i = 1|X_1 = k, X_2 = z)) = \beta_0 + \beta_1 k + \beta_2 z$$

$$\text{logit}(Pr(Y_i = 1|X_1 = k + 1, X_2 = z)) = \beta_0 + \beta_1 (k + 1) + \beta_2 z$$

Coefficient Interpretation in Logistic Regression (Cont.)

- Subtracting the first equation from the second:

$$log[odds(\pi_i|X_1 = k+1, X_2 = z)] - log[odds(\pi_i|X_1 = k, X_2 = z)] = \beta 1$$

- The difference can be expressed as

$$\log\left[\frac{odds(\pi_i|X_i = k+1, X_2 = z)}{odds(\pi_i|X_i = k, X_2 = z)}\right] = \text{log odds ratio}$$

- We can write $\log OR = \beta_1$ or $\log OR = e^{\beta_1}$.

Coefficient Interpretation in Logistic Regression (Cont.)

- For continuous X_i: For every one-unit increase in X_i, the estimated odds of outcome changes by a factor of e^{β_1} or by $[(e^{\beta_1} - 1) \times 100]\%$, controlling for other variables
- For categorical X_i: Group X_i has e^{β_1} times the odds of outcome compared to group X_j, controlling for other variables