Generalized Linear Regression: An Overview

- Binary data response regression
- Count data response regression
- Exponential family data response regression

Outline

Generalized linear models

- All models we have seen so far:
  - outcome variable is continuous with no restriction on their expectation
  - the mean and variance for the outcome is unrelated (i.e. variance is a constant)
- Many outcomes of interest do not satisfy this.
- Example: binary outcomes, poisson count outcomes.
- A Generalized Linear Model (GLM) is a model with two ingredients: a link function and a variance function.
  - The link function relates the mean of the observations to predictors
  - The variance function relates the means to the variances.

Example: disease outbreak

- A health study aims to investigate an epidemic outbreak of a disease. We collect 98 random individuals within two sectors in a city.
- The response variable Y was coded 1 for individual with disease, and 0 if not.
- Three predictors were included: age, socioeconomic status of household, and sector within city.
  1. Age (X_1) is a quantitative variable.
  2. Socioeconomic status (X_2, X_3) is a categorical variable with three levels ('Upper', 'Middle', 'Lower'). (how to code them?)
  3. City sector X_4 is also a categorical variable: $X_4 = 0$ for sector 1 and $X_4 = 1$ for sector 2.

Example: disease outbreak

```
> mydata = read.table("disease.txt", header=T); attach(mydata)
> head(mydata)
  Case X1 X2 X3 X4 Y
1    1 33  0  0  0 0
2    2 35  0  0  0 0
3    3  6  0  0  0 0
4    4 60  0  0  0 0
5    5 18  0  1  0 1
6    6 26  0  1  0 0
> table(Y)
Y
0  1
67 31
> table(Y)/length(Y)
Y
0         1
0.6836735 0.3163265
```

Outline

Binary Response Model

- Consider a binary response variable Y_i, taking on the values 0 and 1.
- Let $\pi_i = \mathbb{E}(Y_i) = \mathbb{P}(Y_i = 1)$ denote the "success" probability.
- Variance function

$$\text{Var}(Y_i) = \pi_i(1 - \pi_i).$$

  Variance is related to mean!
- A convenient way to model the dependence of $\mathbb{E}(Y_i)$ on covariates $X_{i,1}, \ldots, X_{i,p-1}$ is through the logit transformation:

$$\text{logit}(\pi_i) = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij}$$

Logit transform

- Logit transform:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \in (-\infty, +\infty)$$

- Inverse:

$$\text{logit}^{-1}(x) = \frac{e^x}{1+e^x} \in (0, 1).$$

- Derivative:

$$\frac{d}{d\pi}\text{logit}(\pi) = \frac{1}{\pi(1-\pi)} \equiv \frac{1}{V(\pi)},$$

  where $V(\pi) = \pi(1-\pi)$ is called the variance function for Bernoulli r.v.

- Note that the special relation between derivative and variance function — more on this next lecture.

Binary regression model set-up

1. Specify the type of distribution: assume Y_i are independent Bernoulli r.v. with mean π_i; i.e.

$$Y_i \sim \text{Ber}(\pi_i), \quad \text{independent;y.}$$

2. Specify the model on the $\pi_i = \mathbb{E}(Y_i)$:
   - Logistic regression:

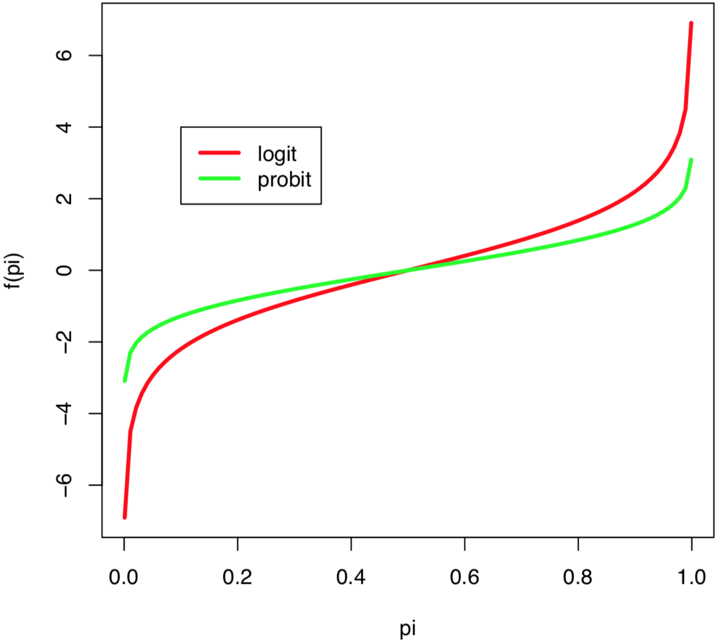
$$\text{logit}(\pi_i) = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij}.$$

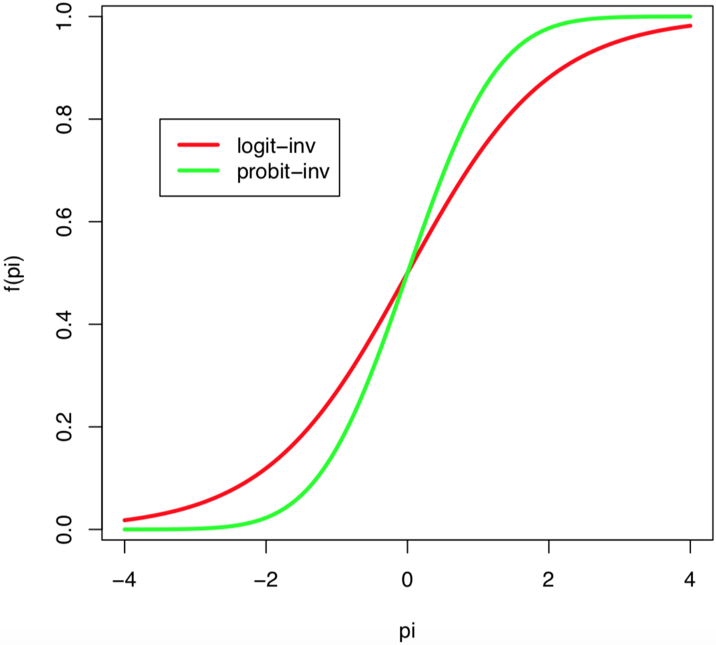
   - Probit regression:

$$\text{Probit}(\pi_i) = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij},$$

     where Probit is the inverse CDF for $N(0, 1)$, i.e. $\text{Probit}(z) = qnorm(z)$.

In each case, the variance model satisfies $\text{Var}(Y_i) = \pi_i(1 - \pi_i)$, but the mean model is different.


logit
probit
6
4
2
0
−2
−4
−6
f(pi)
0.0
0.2
0.4
0.6
0.8
1.0
pi


logit−inv
probit−inv
f(pi)
pi
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4

Link & variance function of a GLM

- If
$$g(\mathbb{E}(Y_i)) = g(\pi_i) = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij}$$
then $g(\cdot)$ is called the link function for the model.
- If
$$\mathrm{Var}(Y_i) = \phi V(\mathbb{E}(Y_i)) = \phi V(\pi_i)$$
for $\phi > 0$ and some function V. Then $V(\cdot)$ is called variance function and ϕ is the dispersion parameter.
- Standard reference: Generalized linear models, McCullagh and Nelder.

Binary (again)

- For a logistic model,

$$g(\mu) = \text{logit}(\mu), \quad V(\mu) = \mu(1-\mu)$$

- For a probit model,

$$g(\mu) = \Phi^{-1}(\mu), \quad V(\mu) = \mu(1-\mu)$$

where Φ is the CDF for $N(0, 1)$.

Other common example of GLMs

- Standard multiple linear regression: $g(\mu) = \mu$, $\mathrm{Var}(\mu) = 1$.
- Linear regression with variance tied to mean, for example: $g(\mu) = \mu$, $\mathrm{Var}(\mu) = \mu^2$.
- Poisson log-linear models: $g(\mu) = \log(\mu)$, $\mathrm{Var}(\mu) = \mu$.

Outline

Logistic regression model

- Model specification:

$$Y_i \sim \text{Ber}(\pi_i), \quad \text{independently},$$

$$\text{where } \pi_i = \mathbb{E}(Y_i) = \frac{\exp(\boldsymbol{X}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}_i'\boldsymbol{\beta})}.$$

- $\boldsymbol{X}_i = (1, X_{i,1}, \ldots, X_{i,p-1})'$ is the $p \times 1$ vector of explanatory variables of the ith observation.
- Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1})'$ denote the MLE of $\boldsymbol{\beta}$.
- Given $\hat{\boldsymbol{\beta}}$, compute the **fitted logistic response function**

$$\hat{\pi}_i = \widehat{\mathbb{E}(Y_i)} = \frac{\exp(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})}.$$

- Also the **fitted logit response function**

$$\log\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right) = \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}.$$

Odds rations & logistic regression

Definition

For any event A and any probability $\mathbb{P}$,

$$\text{Odds}(A) = \frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)}.$$

- In the logistic regression model with outcome Y

$$\frac{\text{Odds}(Y = 1 | \dots, X_j = x_j + 1, \dots)}{\text{Odds}(Y = 1 | \dots, X_j = x_j, \dots,)} = e^{\beta_j},$$

  is the multiplicative change in odds if variable X_j increases by 1.
- e^{β_j} is known as the odds ratio for X_j.
- β_j is also known as the log odds ratio for X_j.

Binge drinker example

- The response variable Y_i was coded 1 if the ith student is a frequent binge drinker and 0 if not.
- We express gender numerically using an indicator variable,

$$X_i = \begin{cases} 1 & ; \quad \text{if the } i\text{th student is a man} \\ 0 & ; \quad \text{if the } i\text{th student is a woman} \end{cases}$$

- $Y_i \sim Ber(\pi_i)$ and $\pi_i = \frac{\exp(\beta_0+\beta_1 X_i)}{1+\exp(\beta_0+\beta_1 X_i)}$.
- There are two possible values for π. For men

$$\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} \tag{1}$$

and for women

$$\frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \tag{2}$$

Binge drinker example

- β_1 : the difference between the log(odds) for men and the log(odds) for women.
- $\exp(\beta_1)$: the ratio of the odds that a man is a frequent binge drinker to the odds that a woman is a frequent binge drinker.
- R command `glm` gives the estimated odds for women 0.2045; For men, the estimated odds are 0.2937; The odds that a man is a frequent drinker are 1.43 times the odds for women.
- Thus, the MLEs of β_0 and β_1 are $\hat{\beta}_0 = -1.59$ and $\hat{\beta}_1 = 0.36$. [Why?]

Outline

Logistic Regression

- Recall that the logistic regression specifies the model for a binary response variable Y_i as

$$Y_i \sim Ber(\pi_i), \quad \text{independently.}$$

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$ is the $p \times 1$ vector of logistic regression coefficients.
- $\boldsymbol{X}_i = (1, X_{i,1}, \ldots, X_{i,p-1})'$ is the $p \times 1$ vector of explanatory variables of the ith observation.
- The mean model for logistic regression is

$$\mathbb{E}(Y_i) = \pi_i = \frac{\exp(\boldsymbol{X}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}_i'\boldsymbol{\beta})}.$$

- We use maximum likelihood for parameter estimation.

Likelihood Function

- Since $Y_i \sim Ber(\pi_i)$, the probability density function is

$$f_i(Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i},$$

where $Y_i = 0$ or 1, $i = 1, \ldots, n$.

- Since Y_i's are independent, the joint probability density function is

$$f(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}.$$

Likelihood Function

- Take logarithm of $f(Y_1, \ldots, Y_n)$ and obtain

$$
\begin{aligned}
l(\boldsymbol{\beta}) &= \log f(Y_1, \ldots, Y_n) \\
&= \sum_{i=1}^{n} \{ Y_i \log(\pi_i) + (1 - Y_i) \log(1 - \pi_i) \} \\
&= \sum_{i=1}^{n} \left\{ Y_i \log\left(\frac{\pi_i}{1 - \pi_i} \right) + \log(1 - \pi_i) \right\} \\
&= \sum_{i=1}^{n} \left[Y_i(\boldsymbol{X}_i' \boldsymbol{\beta}) - \log\{1 + \exp(\boldsymbol{X}_i' \boldsymbol{\beta})\} \right].
\end{aligned}
$$

- Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1})'$ denote the MLE of $\boldsymbol{\beta}$.

Fitting a binary regression GLM: IRLS

- Algorithm:
  1. Initialize: set $\hat{\mu}_i = 0.999$ or 0.001 depending on whether $Y_i = 1$ or 0.
  2. Compute $Z_i \to g(\hat{\mu}_i) + g'(\hat{\mu}_i)(Y_i - \hat{\mu}_i)$.
  3. Obtain $\hat{\beta}$ by regressing $\boldsymbol{Z}$ onto $\boldsymbol{X}$ using WLS with weights $W_i^{-1} = g'(\hat{\mu}_i)^2 V(\hat{\mu}_i)$ to
  4. Compute $\hat{\mu}_i = g^{-1}(\boldsymbol{X}_i'\hat{\beta})$.
  5. Repeat steps 2–4 until convergence.
- If ϕ has to be estimated, a simple choice is Pearson's X^2:

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

- Approximate distribution of $\hat{\beta}$:

$$\hat{\beta} \sim N(\beta, \phi(\boldsymbol{X}^T\hat{W}\boldsymbol{X})^{-1}).$$

Large-Sample (Asymptotic) Properties of MLEs

- Inference about the logistic regression coefficients relies on asymptotic normality of the MLEs.
- Let β^0 denote the $p \times 1$ vector of true regression parameters.
- Let $\boldsymbol{H}$ denote the $p \times p$ **Hessian matrix** $\boldsymbol{H}(\beta) = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$.
- Let $\mathcal{I}(\beta^0)$ denote the $p \times p$ **Fisher information matrix** $\mathcal{I}(\beta) = -\mathbb{E}(\boldsymbol{H}(\beta))$ evaluated at β^0.

Approximate distribution of $\hat{\beta}$

Under suitable regularity conditions, as $n \to \infty$,

$$\hat{\beta} \approx N\left(\beta^0, \mathcal{I}(\beta^0)^{-1}\right), \quad \text{or} \quad \hat{\beta} \approx N\left(\beta^0, -\boldsymbol{H}(\hat{\beta})^{-1}\right)$$