Lecture 6. GLM for Binary Response

Nan Ye

School of Mathematics and Physics
University of Queensland

Examples of Binary Responses

Medical trials

Predict whether a patient will recover or not after a treatment.

Spam filtering

Predict whether an email is a spam or not.

Information retrieval

Predict whether a document is relevant.

Credit decisions

Predict whether a loan applicant is credible.

This Lecture

- Model choices
- Logistic regression
- Binomial data
- Prospective vs. retrospective sampling
- The glm function in R

Models for Binary Responses

Structure

- A GLM for binary response data has the following form

$$\begin{aligned} \text{(systematic)} \quad & \mu = \mathbb{E}(Y \mid \mathbf{x}) = g^{-1}(\beta^\top \mathbf{x}). \\ \text{(random)} \quad & Y \mid \mathbf{x} \sim B(\mu). \end{aligned}$$

- The exponential family has to be a Bernoulli distribution.
- The link function $g : [0, 1] \to (-\infty, +\infty)$ is bijective.

Link functions

- Logit

$$g(\mu) = \text{logit}(\mu) = \ln \frac{\mu}{1-\mu}.$$

- Probit or inverse Normal function

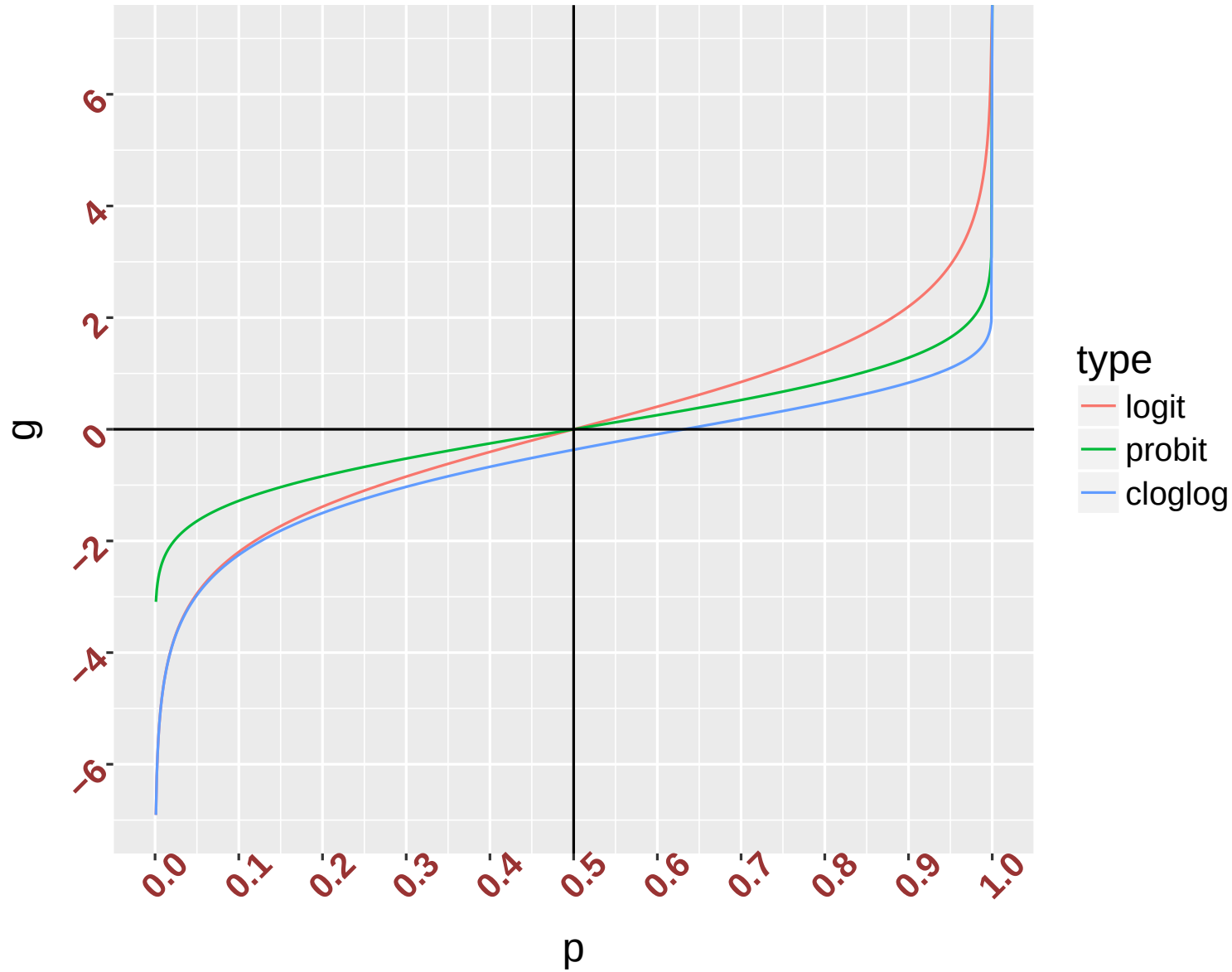
$$g(\mu) = \Phi^{-1}(\mu),$$

  where Φ is the normal cumulative distribution function.

- Complementary log-log

$$g(\mu) = \ln(-\ln(1-\mu)).$$

Plot of the link functions



Comparison of the link functions

- Logit and probit are almost linearly related when $\mu \in [0.1, 0.9]$.
- Logit and complementary log-log are both close to $\ln \mu$ for small μ.
- Logit leads to an easily interpretable model, and is suitable for data collected retrospectively.

We will focus on the logit link.

Logistic Regression

Recall

- When Y takes value 0 or 1, we can use the logistic function to squash $\mathbf{x}^\top \beta$ to $[0, 1]$, and use the Bernoulli distribution to model $Y \mid \mathbf{x}$, as follows.

$$\begin{aligned} \text{(systematic)} \quad & \mathbb{E}(Y \mid \mathbf{x}) = \mathit{logistic}(\beta^\top \mathbf{x}) = \frac{1}{1 + e^{-\beta^\top \mathbf{x}}}. \\ \text{(random)} \quad & Y \mid \mathbf{x} \text{ is Bernoulli distributed.} \end{aligned}$$

- Or more compactly,

$$Y \mid \mathbf{x} \sim B\left(\frac{1}{1 + e^{-\beta^\top \mathbf{x}}}\right),$$

where $B(p)$ is the Bernoulli distribution with parameter p.

- The logistic regression can be written explicitly as

$$p(y \mid x, \beta) = \frac{e^{y\beta^\top \mathbf{x}}}{1 + e^{\beta^\top \mathbf{x}}}$$

- Given $\mathbf{x}$, we can predict Y as

$$\arg\max_y p(y \mid \mathbf{x}, \beta) = \begin{cases} 1, & \mathbf{x}^\top \beta > 0. \\ 0, & \mathbf{x}^\top \beta \leq 0. \end{cases}$$

Parameter interpretation

- The log-odds is

$$\ln \frac{p}{1-p} = \beta^\top \mathbf{x},$$

  where $p = p(y = 1 \mid x, \beta)$.
- A unit increase in x_i changes the odds by a factor of e^{β_i}.

Fisher scoring

- Let $\mathbf{X}$ be the design matrix, and

$$\mathbf{p} = (p_1, \ldots, p_n) \text{ with } p_i = \mathbb{E}(Y_i \mid, \mathbf{x}_i, \beta),$$
$$W = \operatorname{diag}\left(p_1(1-p_1), \ldots, p_n(1-p_n)\right).$$

- Then the gradient and the Fisher information are

$$\nabla \ell(\beta) = \mathbf{X}^\top(\mathbf{y} - \mathbf{p}),$$
$$I(\beta) = \mathbf{X}^\top W \mathbf{X},$$

- Fisher scoring updates β to

$$\beta' = \beta + I(\beta)^{-1} \nabla \ell(\beta).$$

Binomial Data

- In binomial data, for each $\mathbf{x}$, we perform some number of t trials, and observe some number s of successes.
- We want to model the success probability.
- Essentially, each binomial example is a set of binary data.
- Specifically, given $\mathbf{x}$, if we observe s successes among t trials, then we can think of the data as having s $(\mathbf{x}, 1)$ pairs, and $t - s$ $(\mathbf{x}, 0)$ pairs.

Prospective vs. Retrospective Sampling

Example

- Consider a study on the effect of exposure to a toxin on the incidence of a disease.
- Prospective sampling
  - Sample a group of exposed subjects, together with a comparable group of non-exposed, and monitor the progress of each group.
  - We may end up having too few diseased subjects to draw any meaning conclusion...
- Retrospective sampling
  - Sample diseased and disease-free individuals, and then identify at their exposure status.
  - We often end up with a sample with a much higher disease rate than the actual rate...

Comparing the two sampling schemes

- Prospective sampling
  - Sample $\mathbf{x}$, then sample y.
  - The sampling distribution is designed to faithful to actual joint distribution $P(\mathbf{x}, y)$.
- Retrospective sampling
  - Sample y, then sample $\mathbf{x}$.
  - y is usually not randomly sampled from the true marginal $P(y)$.
  - The sampling distribution may be very different from $P(\mathbf{x}, y)$.

When $P(y \mid \mathbf{x})$ is logistic regression...

- Assume that $P(y \mid \mathbf{x})$ is a logistic regression model $p(y \mid \mathbf{x}, \beta)$.
- Retrospective sampling is sampling from a distribution $\hat{P}(\mathbf{x}, y)$ that is generally different from $P(\mathbf{x}, y)$.
- However, if the probability of sampling $\mathbf{x}$ depends only on y, then

$$\hat{P}(y \mid \mathbf{x}) = \frac{e^{y(\alpha + \mathbf{x}^\top \beta)}}{1 + e^{y(\alpha + \mathbf{x}^\top \beta)}},$$

- That is, $\hat{P}(\mathbf{x}, y)$ is the same as $p(y \mid \mathbf{x}, \beta)$ except that the intercept may be different.

Notation: P denotes a data distribution, and p denotes a model.

Justification

- Introduce the dummy variable Z indicating whether $\mathbf{x}$ is sampled.
- Our assumption is that

$$P(Z = 1 \mid Y = 0, \mathbf{x}) = \pi_0, \qquad P(Z = 1 \mid Y = 1, \mathbf{x}) = \pi_1,$$

  where π_0 and π_1 are independent of $\mathbf{x}$.
- Using Bayes rule, we have

$$\begin{aligned}
&\hat{P}(y \mid \mathbf{x}) \\
&= P(y \mid z = 1, \mathbf{x}) \\
&= \frac{P(y \mid \mathbf{x}) P(z = 1 \mid \mathbf{x}, y)}{P(y = 1 \mid \mathbf{x}) P(z = 1 \mid \mathbf{x}, y = 1) + P(y = 0 \mid \mathbf{x}) P(z = 1 \mid \mathbf{x}, y = 0)} \\
&= \frac{e^{y(\alpha + \mathbf{x}^\top \beta)}}{1 + e^{\alpha + \mathbf{x}^\top \beta}},
\end{aligned}$$

where $\alpha = \ln(\pi_1 / \pi_0)$.

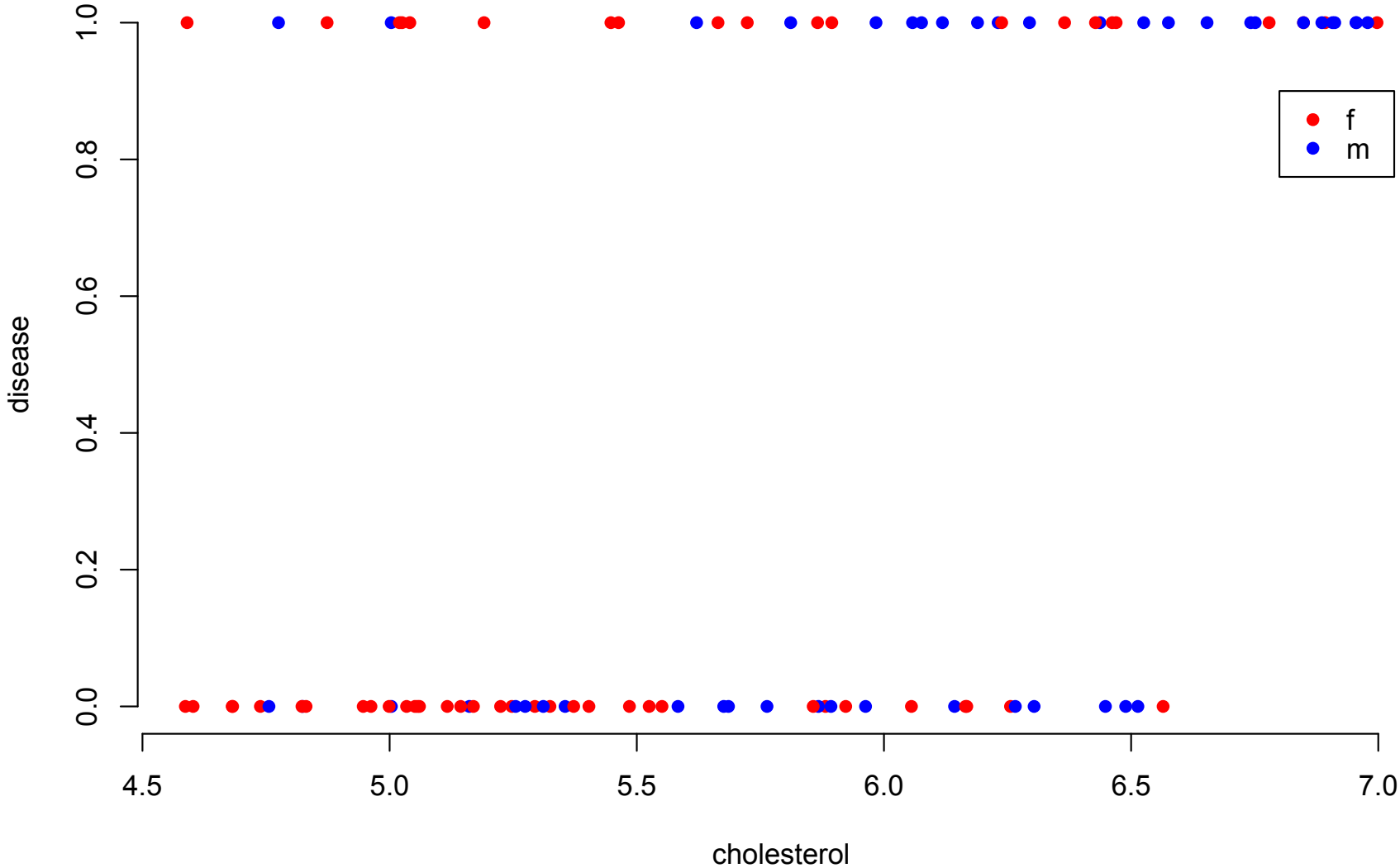
The glm Function in R

Data

```
> chol = read.csv("cholest.csv")
> head(chol)
  X cholesterol gender genderS disease
1 1    6.741923      1       m       1
2 2    5.675853      1       m       0
3 3    5.247094      0       f       0
4 4    5.034348      0       f       0
5 5    6.167538      0       f       0
6 6    5.025060      0       f       1
```

Plot

```
> # plot disease status against cholesterol level
> palette(c('red', 'blue'))
> plot(chol$cholesterol, chol$disease, xlab='cholesterol',
    ylab='disease', axes=F, col=chol$genderS, pch=16)
> # put a legend
> legend(6.8, 0.9, levels(chol$genderS), col=1:length(chol$genderS),
    pch=16)
> # manually label x and y axes
> axis(1, at = c(4.5,5,5.5,6,6.5,7))
> axis(2, at=c(0,0.2,0.4,0.6,0.8,1.0))
```


disease
0.0
0.2
0.4
0.6
0.8
1.0
4.5
5.0
5.5
6.0
6.5
7.0
cholesterol
f
m

Fit a model

```
> # fit a logistic regression model of disease against gender and
    cholesterol
> fit.bin = glm(disease ~ gender + cholesterol, data=chol,
    family=binomial)
> # same as the following
> fit.bin = glm(disease ~ gender + cholesterol, data=chol,
    family=binomial(link='logit'))
```

For more information...

- *glm: https://goo.gl/zYUs5U*
- *formula: https://goo.gl/aQyeU7*
- *family: https://goo.gl/ZXsbN4*

Predition

```
> # fitted link on the training data
> predict(fit.bin)
> # predict link on new data
> predict(fit.bin, newdata=chol)
> # same as above
> predict(fit.bin, newdata=chol, type='link')
> # predict probabilities on new data
> predict(fit.bin, newdata=chol, type='response')
> # predict classes on new data
> as.numeric(predict(fit.bin, newdata=chol) > 0)
```

Inspect a model

```
> fit.bin

Call:  glm(formula = disease ~ gender + cholesterol, family =
    binomial,
    data = chol)

Coefficients:
(Intercept)       gender  cholesterol
    -9.3203      -0.1094       1.5842

Degrees of Freedom: 99 Total (i.e. Null);  97 Residual
Null Deviance:     137.6
Residual Deviance: 114  AIC: 120
# also try this
> summary(fit.bin)
```

What You Need to Know

- Model choices

  Bernoulli for random component, several commonly used link functions

- Logistic regression

  $p(y \mid \mathbf{x}, \beta)$, prediction, parameter interpretation, Fisher scoring

- Binomial data
- Prospective vs. retrospective sampling
- The glm function in R