San José State University

Math 261A: Regression Theory & Methods

Generalized Linear Models (GLMs)

Dr. Guangliang Chen

This lecture is based on the following textbook sections:

- Chapter 13: 13.1 – 13.3

Outline of this presentation:

- What is a GLM?
- Logistic regression
- Poisson regression

What is a GLM?

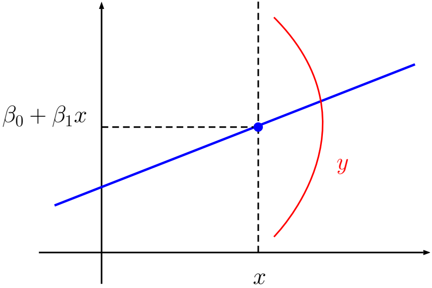
In **ordinary linear regression**, we assume that the response is a **linear** function of the regressors plus **Gaussian** noise:

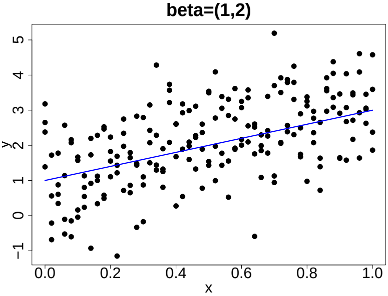
$$y = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\text{linear form } \mathbf{x}'\boldsymbol{\beta}} + \underbrace{\epsilon}_{N(0,\sigma^2) \text{ noise}} \sim N(\mathbf{x}'\boldsymbol{\beta}, \sigma^2)$$

The model can be reformulate in terms of

- distribution of the response: $y \mid \mathbf{x} \sim N(\mu, \sigma^2)$, and
- dependence of the mean on the predictors: $\mu = \mathrm{E}(y \mid \mathbf{x}) = \mathbf{x}'\beta$

Generalized Linear Models (GLMs)





Generalized linear models (GLM) extend linear regression by allowing the response variable to have

- a general distribution (with mean $\mu = \mathrm{E}(y \mid \mathbf{x})$) and
- a mean that depends on the predictors through a link function g:

That is,

$$g(\mu) = \boldsymbol{\beta}'\mathbf{x}$$

or equivalently,

$$\mu = g^{-1}(\boldsymbol{\beta}'\mathbf{x})$$

In GLM, the response is typically assumed to have a distribution in the **exponential family**, which is a large class of probability distributions that have pdfs of the form $f(x \mid \theta) = a(x)b(\theta)\exp(c(\theta) \cdot T(x))$, including

- **Normal** - ordinary linear regression
- **Bernoulli** - Logistic regression, modeling binary data
- **Binomial** - Multinomial logistic regression, modeling general categorical data
- **Poisson** - Poisson regression, modeling count data
- **Exponential, Gamma** - survival analysis

In theory, any combination of the response distribution and link function (that relates the mean response to a linear combination of the predictors) specifies a generalized linear model.

Some combinations turn out to be much more useful and mathematically more tractable than others in practice.

Response distribution	**Link function**	$g(\mu)$	**Use**
Normal	Identity	μ	OLS
Bernoulli	Logit	$\log\left(\frac{\mu}{1-\mu}\right)$	Logistic regression
Poisson	Log	$\log(\mu)$	Poisson regression
Exponential/Gamma	Inverse	$-1/\mu$	Survival analysis

Applications:

- **Logistic Regression**: Predict the likelihood that a consumer of an online shopping website will buy a specific item (say, a camera) within the next month based on the consumer's purchase history.
- **Poisson regression**: Modeling the number of children a couple has as a function of their ages, numbers of siblings, income, education levels, etc.
- **Exponential**: Modeling the survival time (time until death) of patients in a clinical study as a function of disease, age, gender, type of treatment etc.

Logistic regression

Logistic regression is a GLM that combines the Bernoulli distribution (for the response) and the logit link function (relating the mean response to predictors):

$$\log\left(\frac{\mu}{1-\mu}\right) = \boldsymbol{\beta}'\mathbf{x} \qquad (y \sim \text{Bernoulli}(p))$$

Remark. Since $\mu = \mathrm{E}(y \mid \mathbf{x}) = p$, we have

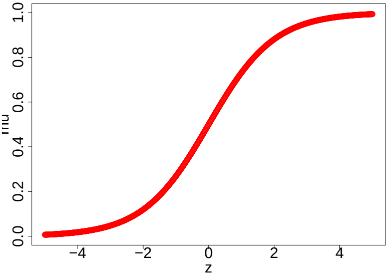
$$\log\left(\frac{p}{1-p}\right) = \boldsymbol{\beta}'\mathbf{x} \qquad (y \sim \text{Bernoulli}(p))$$

where p: probability of success, $\frac{p}{1-p}$: odds, $\log(\frac{p}{1-p})$: log-odds.

Solving for μ (and also p), we obtain that

$$\mu = \frac{1}{1+e^{-\boldsymbol{\beta}'\mathbf{x}}} = \sigma(\boldsymbol{\beta}'\mathbf{x}), \qquad s(z) = \frac{1}{1+e^{-z}},$$

where $s(\cdot)$ is the **sigmoid** function, also called the **logistic** function.



Properties of the sigmoid function:

- $s(0) = 0.5$
- $0 < s(z) < 1$ for all z
- $s(z)$ monotonically increases as z goes from $-\infty$ to $+\infty$

For fixed $\boldsymbol{\beta}$ (model parameter) and each given $\mathbf{x}$ (sampled location),

$$\mu = p = s(z), \quad z = \boldsymbol{\beta}'\mathbf{x}$$

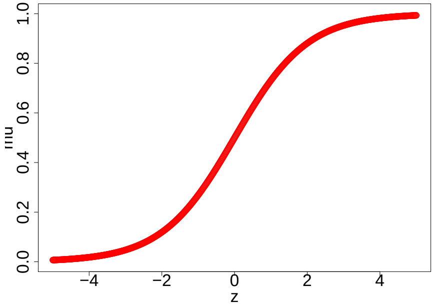
has the following interpretations:

- mean response

$$\mathrm{E}(y \mid \mathbf{x}, \boldsymbol{\beta}) = s(z)$$

- probability of success:

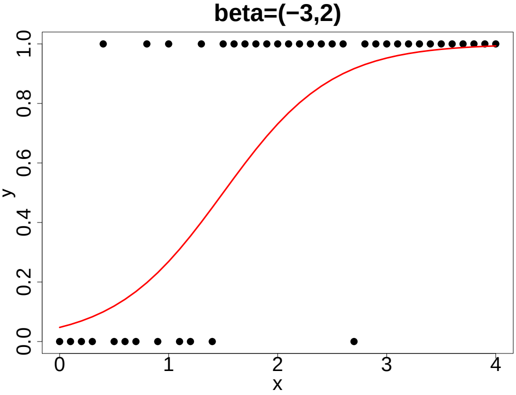
$$P(y = 1 \mid \mathbf{x}, \boldsymbol{\beta}) = s(z)$$



Population model:

$$y \mid \mathbf{x}, \boldsymbol{\beta} \sim \textbf{Bernoulli}(p = s(\boldsymbol{\beta}'\mathbf{x}))$$

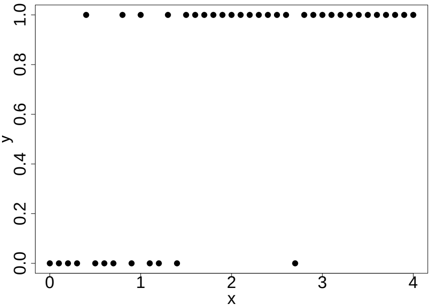
A sample from the logistic regression model, with $p = s(-3 + 2x)$



Parameter estimation via MLE

Given a data set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, fitting a logistic regression model is equivalent to choosing the value of $\boldsymbol{\beta}$ such that the mean response

$$\mu = s(\boldsymbol{\beta}'\mathbf{x})$$



matches the sample as "closely" as possible.

Mathematically, the best β is usually found by maximizing the likelihood of the sample:

$$L(\boldsymbol{\beta} \mid y_1, \ldots, y_n) = f(y_1, \ldots, y_n \mid \boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i \mid \boldsymbol{\beta})$$

where $f(y_i \mid \boldsymbol{\beta})$ is the probability function of the ith observation:

$$f(y_i \mid \boldsymbol{\beta}) = p_i^{y_i}(1-p_i)^{1-y_i} = \begin{cases} p_i, & y_i = 1 \\ 1-p_i & y_i = 0 \end{cases}$$

and

$$p_i = \frac{1}{1+e^{-\boldsymbol{\beta}'\mathbf{x}_i}}$$

However, there is no closed-form solution, and the optimal β has to be computed numerically.

Prediction by logistic regression

Once the optimal parameter $\hat{\boldsymbol{\beta}}$ is found, the mean response at a new location $\mathbf{x}_0$ is

$$\mathrm{E}(y \mid \mathbf{x}_0, \hat{\boldsymbol{\beta}}) = \frac{1}{1 + e^{-\hat{\boldsymbol{\beta}}'\mathbf{x}_0}}$$

Note that this would not be our exact prediction at $\mathbf{x}_0$ (why?).

To make a prediction at $\mathbf{x}_0$ based on the estimates $\hat{\boldsymbol{\beta}}$, consider

$$y_0 \mid \mathbf{x}_0, \hat{\boldsymbol{\beta}} \sim \text{Bernoulli}(\hat{p}_0), \quad \hat{p}_0 = \frac{1}{1 + e^{-\hat{\boldsymbol{\beta}}' \mathbf{x}_0}}.$$

The prediction at $\mathbf{x}_0$ is

$$\hat{y}_0 = \begin{cases} 1, & \text{if } \hat{p}_0 > 0.5 \\ 0, & \text{if } \hat{p}_0 < 0.5 \end{cases}$$

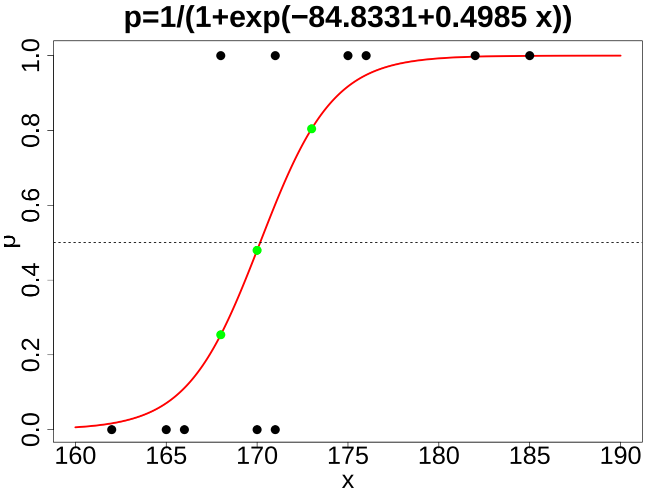
R scripts

```
x = c(162, 165, 166, 170, 171, 168, 171, 175, 176, 182, 185)
y = c(0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1)
model <- glm(y~x,family=binomial(link='logit'))

p = model$fitted.values
# p = [0.0168, 0.0708, 0.1114, 0.4795, 0.6026, 0.2537, 0.6026, 0.9176,
0.9483, 0.9973, 0.9994]

beta = model$coefficients        # beta = [-84.8331094 0.4985354]

fitted.prob <- predict(model,data.frame(x=c(168,170,173)),type='response')
# fitted.prob = [0.2537, 0.4795 0.8043 ]
```


p=1/(1+exp(−84.8331+0.4985 x))
p
x
0.0
0.2
0.4
0.6
0.8
1.0
160
165
170
175
180
185
190

Other models for binary response data

Instead of using the logit link function,

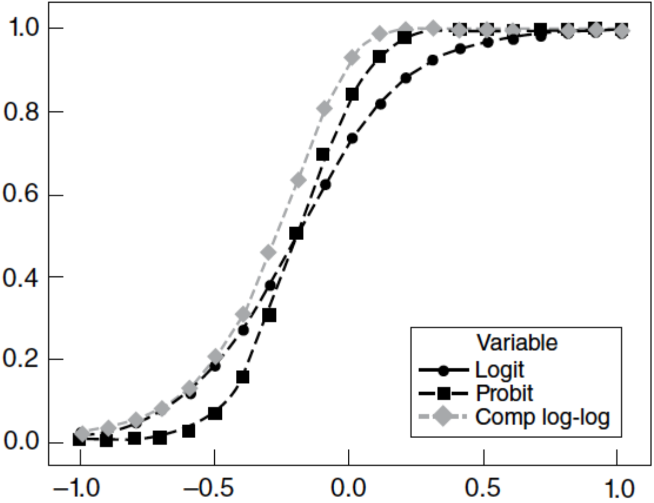
$$p = \frac{1}{1 + e^{-\boldsymbol{\beta}'\mathbf{x}}}$$

to force the estimated probabilities to lie between 0 and 1:

$$y \mid \mathbf{x}, \boldsymbol{\beta} \sim \text{Bernoulli}(p)$$

one could use

- **Probit**: $p = \Phi(\boldsymbol{\beta}'\mathbf{x})$, where Φ is the cdf of standard normal.
- **Complimentary log-log**: $p = 1 - \exp(-\exp(\boldsymbol{\beta}'\mathbf{x}))$


1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
Variable
Logit
Probit
Comp log-log

Poisson regression

Poisson regression is a GLM that combines the Poisson distribution (for the response) and the log link function (relating mean response to predictors):

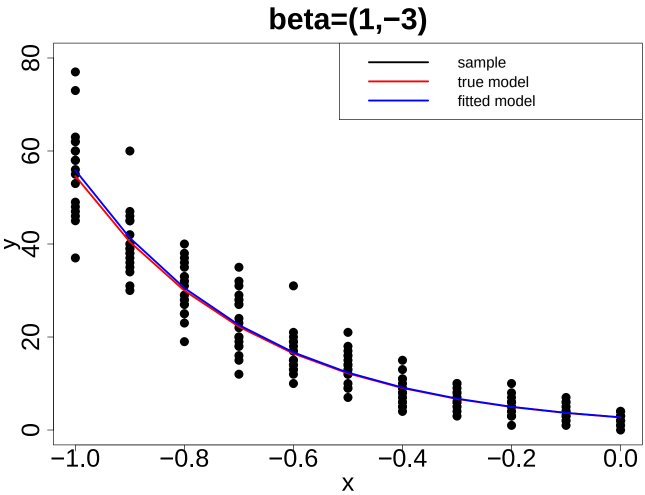
$$\log(\mu) = \boldsymbol{\beta}'\mathbf{x} \qquad (y \sim \text{Poisson}(\lambda))$$

Remark. Since $\mu = \mathrm{E}(y \mid \mathbf{x}) = \lambda$, we have

$$\log \lambda = \boldsymbol{\beta}'\mathbf{x} \qquad \text{or} \quad \lambda = e^{\boldsymbol{\beta}'\mathbf{x}}$$

That is,

$$y \mid \mathbf{x}, \boldsymbol{\beta} \sim \text{Poisson}(\lambda = e^{\boldsymbol{\beta}'\mathbf{x}})$$


beta=(1,−3)
sample
true model
fitted model
y
0
20
40
60
80
x
−1.0
−0.8
−0.6
−0.4
−0.2
0.0

R code

```
poisson.model ← glm(y∼x,family=poisson(link='log'))

poisson.model$coefficients

(Intercept)    x
1.003291       -3.019297
```

Summary and beyond

We talked about the concept of generalized linear models and its two special instances:

- **Logistic regression**: logit link function + Bernoulli distribution
- **Poisson regression**: log link function + Poisson distribution

Note that parameter estimation for GLM is through MLE; prediction is based on the mean (plus some necessary adjustments).

Further learning on logistic and multinomial regression:

http://www.sjsu.edu/faculty/guangliang.chen/Math251F18/lec5logistic.pdf