DATA 606: Statistical Methods in Data Science

—— Generalized linear model

Wenjun Jiang

Department of Mathematics & Statistics
The University of Calgary

Lecture 10

Review: simple linear regression

In a simple linear regression model

$$Y = \alpha + \beta \cdot x + \epsilon,$$

usually, $\epsilon \sim N(0, \sigma^2)$.

- Y: response variable.
- x: covariate or explanatory variable.
- β catches the linear relationship between X and Y.
- When $\beta = 0$, there is no linear relationship between X and Y.

Review: simple linear regression

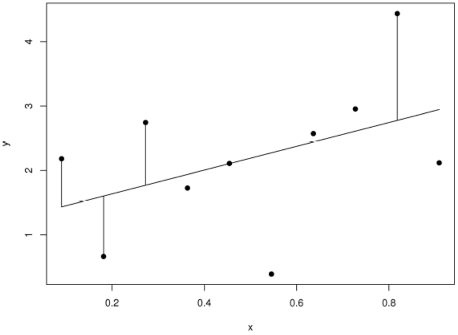
- Given data (x_i, y_i), $i = 1, 2, \ldots, n$, how to estimate α and β?

Review: simple linear regression

- Given data (x_i, y_i), $i = 1, 2, \ldots, n$, how to estimate α and β?
  We apply the so-called *least square* method:

$$\min \sum_{i=1}^{n} (y_i - \alpha - \beta \cdot x_i)^2 \implies (\hat{\alpha}, \hat{\beta}).$$

- An graphical illustration:



Review: simple linear regression

- Under the condition (or assumption) $\epsilon \sim N(0, \sigma^2)$, our linear model in fact can be rewritten as

$$Y \sim N(\alpha + \beta \cdot x, \sigma^2).$$

Review: simple linear regression

- Under the condition (or assumption) $\epsilon \sim N(0, \sigma^2)$, our linear model in fact can be rewritten as
$$Y \sim N(\alpha + \beta \cdot x, \sigma^2).$$

- The above distribution reminds us to use MLE (maximum likelihood estimation) to estimate α and β.

Review: simple linear regression

- Under the condition (or assumption) $\epsilon \sim N(0, \sigma^2)$, our linear model in fact can be rewritten as
$$Y \sim N(\alpha + \beta \cdot x, \sigma^2).$$
- The above distribution reminds us to use MLE (maximum likelihood estimation) to estimate α and β.
- The above simple linear model can be extended to multiple linear model
$$Y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots \beta_n \cdot x_n + \epsilon,$$
where $\epsilon \sim N(0, \sigma^2)$. Equivalently, we have
$$Y \sim N(\mu(x), \sigma^2), \quad \mu(x) = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots \beta_n \cdot x_n.$$

Basics about GLM

Three components of GLM: random component, systematic component and link function.

- Y: response variable; $\{x_i, i = 1, 2, \ldots, n\}$ are explanatory variables.
- *Random component:* Y is random.
- *Systematic component:* the predictive linear combination

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n.$$

Basics about GLM

Question: how we bridge Y and $\{x_i, i = 1, 2, \ldots, n\}$?

- Denote $\mu = \mathbf{E}[Y]$.
- With a function $g(\cdot)$, we relate μ and the *systematic component* via

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n.$$

- This function $g(\cdot)$ is called *link function*.
- In simple (or multiple) linear regression, the link function

$$g(\mu) = \mu.$$

Difference between GLM and data transformation

- *Data transformation*: in practice, in order to explore the relationship between Y and $\{x_i, i = 1, 2, \ldots, n\}$, sometimes we would apply

$$g(Y) = \alpha + \beta_1 x_1 + \cdots \beta_n x_n.$$

  This method transforms the response variable Y.

- DO note that the data transformation method is NOT generalized linear regression !

GLM for binary response

- When the response Y is binary (1/0, 1=success, 0=failure):

$$\mu = \mathbf{E}[Y] = 1 \times \mathbf{P}(Y = 1) + 0 \times \mathbf{P}(Y = 0) = \pi.$$

- With link function $g(\cdot)$, we have

$$g(\mu) = g(\pi) = \alpha + \beta x.$$

- Understandably, different g will result in different GLM.

GLM for binary response

Linear probability model

- If we choose the link function g to be identity function: $g(\pi) = \pi$, then

$$\pi = \alpha + \beta x.$$

GLM for binary response

Linear probability model

- If we choose the link function g to be identity function: $g(\pi) = \pi$, then

$$\pi = \alpha + \beta x.$$

- NOTE: linear probability model is reasonable only if $\alpha + \beta x \in [0, 1]$

- If we choose the link function g to be identity function: $g(\pi) = \pi$, then

$$\pi = \alpha + \beta x.$$

- NOTE: linear probability model is reasonable only if $\alpha + \beta x \in [0, 1]$
- In the linear probability model, the coefficient β has a nice interpretation:

$$\beta = \pi(x + 1) - \pi(x).$$

- Inference for the risk difference in a 2×2 table can be achieved using the linear probability model

$$
\begin{array}{cc|c|c|c}
 & & \multicolumn{2}{c}{Y} & \\
 & & 1 & 0 & \\
\hline
X & 1 & y_1 & n_1 - y_1 & n_1 \\
\hline
 & 0 & y_2 & n_2 - y_2 & n_2 \\
\hline
\end{array}
$$

- Let $\pi_1 = \mathbf{P}(Y = 1|x = 1)$ and $\pi_0 = \mathbf{P}(Y = 1|x = 0)$ and we would like to make inference of $\phi = \pi_1 - \pi_0$.
- We can fit the linear probability model to the above table

$$\pi = \alpha + \beta x$$

and $\beta = \phi$.

An example

- Snoring and heart disease example

		Heart disease		
	x	Yes	No	n
Snoring	0 (never)	24	1355	1379
	2 (occasionally)	35	605	640
	4 (nearly every night)	21	192	213
	5 (every night)	30	224	254

An example

- Snoring and heart disease example

		Heart disease		
	x	Yes	No	n
Snoring	0 (never)	24	1355	1379
	2 (occasionally)	35	605	640
	4 (nearly every night)	21	192	213
	5 (every night)	30	224	254

- After assigning scores x_i: 0,2,4,5 to snoring, we can calculate the sample proportions p_i for each snoring level and plot p_i against x_i to check whether the linearity relationship is significant.

GLM for binary response

Log linear probability model

- For binary response, if we take the link function to be

$$g(\pi) = \log(\pi) = \alpha + \beta x.$$

- Given x and α, β, we have

$$\pi = e^{\alpha+\beta x}.$$

- The model is reasonable if the model produces a π which is between 0 and 1.

GLM for binary response

Log linear probability model

Interpretation of β:

- $\log \pi(x) = \alpha + \beta x$,
- $\log \pi(x+1) = \alpha + \beta(x+1)$,
- $\beta = \log \pi(x+1) - \log \pi(x) = \log \frac{\pi(x+1)}{\pi(x)}$.

β is the logarithm of relative risk.

GLM for binary response

Log linear probability model

- Inference for the risk difference in a 2×2 table can be achieved using the linear probability model

$$
\begin{array}{cc|c|c|c}
 & & \multicolumn{2}{c|}{Y} & \\
 & & 1 & 0 & \\
\hline
X & 1 & y_1 & n_1 - y_1 & n_1 \\
\hline
 & 0 & y_2 & n_2 - y_2 & n_2 \\
\hline
\end{array}
$$

- Let $\pi_1 = \mathbf{P}(Y = 1|x = 1)$ and $\pi_0 = \mathbf{P}(Y = 1|x = 0)$ and we would like to make inference of $RR = \pi_1/\pi_0$.
- We could fit the following log-linear model

$$\log \pi = \alpha + \beta x.$$

- Test $H_0 : \beta = 0$ is equivalent to $H_0 : X$ and Y are independent.

GLM for binary response

- For binary response, if we take the link function g to be

$$g(\pi) = \text{logit}(\pi) = \log(\frac{\pi}{1-\pi}) = \alpha + \beta x,$$

then we have a *logistic regression model*.

GLM for binary response

Logistic regression

- For binary response, if we take the link function g to be

$$g(\pi) = \text{logit}(\pi) = \log(\frac{\pi}{1-\pi}) = \alpha + \beta x,$$

then we have a *logistic regression model*.

- Now the probability π is of form

$$\pi = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}.$$

GLM for binary response

Logistic regression

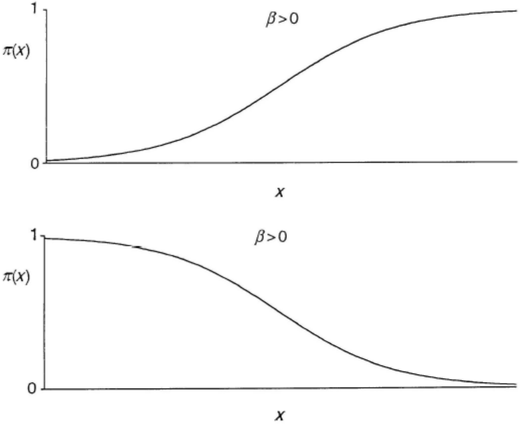


Figure 3.2. Logistic regression functions.

Interpretation of β:

- At $x = x_1$, we have π_1: $\log \frac{\pi_1}{1-\pi_1} = \alpha + \beta x_1$.
- At $x = x_1 + 1$, we have π_2: $\log \frac{\pi_2}{1-\pi_2} = \alpha + \beta(x_1 + 1)$.
- Now we take the difference of above two

$$\begin{aligned}\beta &= \log \frac{\pi_2}{1-\pi_2} - \log \frac{\pi_1}{1-\pi_1} \\ &= \log \left(\frac{\pi_2(1-\pi_1)}{\pi_1(1-\pi_2)} \right).\end{aligned}$$

That's the log of *odds ratio*!

GLM for binary response

- Inference for the risk difference in a 2×2 table can be achieved using the linear probability model

$$
\begin{array}{cc|c|c|c}
 & & \multicolumn{2}{c}{Y} & \\
 & & 1 & 0 & \\
\hline
X & 1 & y_1 & n_1 - y_1 & n_1 \\
\hline
 & 0 & y_2 & n_2 - y_2 & n_2 \\
\hline
\end{array}
$$

- We are interested in the odds ratio $\theta = \frac{\pi_2(1-\pi_1)}{\pi_1(1-\pi_2)}$. If $\theta = 1$ then Y and X are independent.

- We could use logistic regression

$$\log \frac{\pi}{1-\pi} = \alpha + \beta x,$$

we know $\beta = 0 \iff \theta = 1$.

Poisson regression

- The response Y follows Poisson distribution:

$$\mathbf{P}(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}.$$

- The mean of Y is $\mu = \mathbf{E}[Y] = \lambda$.

- Suppose x is explanatory variable, with link function g, we have

$$g(\mu) = g(\lambda) = \alpha + \beta x.$$

- Since $\lambda > 0$, we usually use log function as link function

$$\log \lambda = \alpha + \beta x.$$

An example

Horseshoe crabs and their satellites (see R notebook).

An example

The count data

Carapace width (x)	Num. of Obs.
≤ 23.25	14
$23.25 - 24.25$	14
$24.25 - 25.25$	28
$25.25 - 26.25$	39
$26.25 - 27.25$	22
$27.25 - 28.25$	24
$28.25 - 29.25$	18
> 29.25	14

For convenience, we use explanatory variable X: 22.125, 23.750, 24.750, 25.750, 26.750, 27.75, 28.750,31.375.

Negative binomial regression

- The response Y follows negative binomial distribution

$$\mathbf{P}(Y = y) = \binom{y + k - 1}{y} (1 - \pi)^y \pi^k.$$

- The mean and variance of Y are

$$\mathbf{E}[Y] = \frac{k(1-\pi)}{\pi} = \mu, \quad \mathrm{Var}(Y) = \frac{k(1-\pi)}{\pi^2} = \mu + \frac{\mu^2}{r}.$$

- Suppose x is explanatory variable, with link function g, we have

$$g(\mu) = \alpha + \beta x.$$

- Since $\mu > 0$, we usually use log function as the link function:

$$\log \mu = \alpha + \beta x.$$

GLM for rate data

- When the response Y represents the number of events over a time window with length T or over a population with size T. It may be more meaningful to model the rate data $R = \frac{Y}{T}$
- Let $\mu = \mathbf{E}[Y]$, then the expected rate $r = \mathbf{E}[R] = \frac{\mu}{T}$.
- Now we use a log-linear model for the rate

$$\log(r) = \alpha + \beta x,$$

which is equivalent to

$$\log(\mu) = \log(T) + \alpha + \beta x.$$

The term $\log(T)$ is called an *offset*.

GLM for rate data

Table 3.4. Collisions Involving Trains in Great Britain

Year	Train-km	Train Collisions	Train-road Collisions	Year	Train-km	Train Collisions	Train-road Collisions
2003	518	0	3	1988	443	2	4
2002	516	1	3	1987	397	1	6
2001	508	0	4	1986	414	2	13
2000	503	1	3	1985	418	0	5
1999	505	1	2	1984	389	5	3
1998	487	0	4	1983	401	2	7
1997	463	1	1	1982	372	2	3
1996	437	2	2	1981	417	2	2
1995	423	1	2	1980	430	2	2
1994	415	2	4	1979	426	3	3
1993	425	0	4	1978	430	2	4
1992	430	1	4	1977	425	1	8
1991	439	2	6	1976	426	2	12
1990	431	1	2	1975	436	5	2
1989	436	4	4				

Source: British Department of Transport.

Figure 1: British train accidents over time.

GLM for rate data

Now regarding this dataset, we consider

- y is yearly number of train accidents with road vehicles.
- T is the length of rail.
- x the number of years since 1975.

Consider the log-rate GLM

$$\log(\mu) = \log T + \alpha + \beta x.$$