STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2019

Dr. Shivon Sue-Chee

February 4-8, 2019

STA 303/1002: Week 5-Generalized Linear Models

- Case Study III: The Donner Party Example
- Generalized Linear Models
  - What is a Generalized Linear Model?
  - Common link functions
  - What is a Binary Logistic Regression Model?
  - Maximum likelihood estimation of β's
- Case Study III Example
  - Data and Questions
  - Estimated Model
  - Interpretations

Case Study III: The Donner Party Example

- Background: (D.K. Grayson, Journal of Anthropological Research, 1990: 223-42)
  - In mid 19th century, a group of 86 American pioneers headed out from Missouri toward California in a wagon train.
  - Due to a combination of harsh weather, unsuitable travel equipment and divisions with the party, the group got stuck in the Sierra Nevada mountain range.
  - They had planned to arrive safe and sound in September but those who survived did not make it there until the following March.
- Question: Who survived?- Men? -Older pioneers?
- Data:
  - age
  - sex
  - outcome: survived or not
- AIM: Study the odds of survival

Case Study III: Model

- Response: Y_i- a binary variable (eg., survived or died)
- Predictor: X_i- eg., age, sex of ith pioneer
- Model: BINARY LOGISTIC REGRESSION

1 0

$$Y_i|X_i = \begin{cases} 1 & \text{if response is in category of interest} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_i|X_i \sim \text{Bernoulli}(\pi_i)$$

$0 \leq \pi \leq 1$

Then:

- $E[Y_i|X_i] = \pi_i$ and $\text{Var}(Y_i|X_i) = \pi_i(1 - \pi_i)$

- A logistic regression model is an example of a **Generalized Linear Model**.

Generalized Linear Models

- Have: · response, Y and
  · a set of explanatory variables $X_1, \ldots, X_p$
- Want: Model $E(Y)$ as a linear function in the parameters, ie.,

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = \mathbf{X}\beta$$

- Key idea: Choice of the link function, g such that

$$g(E(Y)) = \mathbf{X}\beta$$

transformation of y

$E[g(y)]$

$\neq g[E(y)]$

Some Link Functions

Let $E(Y) = \mu$.

Link	Function	Usual distribution of $Y \mid X$
Identity	$g(\mu) = \mu$	Normal — General Linear Model
Log	$g(\mu) = \log \mu, \mu > 0$	Poisson (count data)
Logit	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right), 0 < \mu < 1$	Bernoulli (binary), Binomial

Note: Link function, $g(\cdot)$ is a function of $\mu = E(Y)$, the mean of Y, and not a transformation of the data.

$$\log\left(\frac{\pi}{1-\pi}\right)$$

GLMs vs Transforming the data

- Transform Y so it has an approximate normal distribution with constant variance. Common variance stabilizing transformations (Weisberg, 3rd ed, p. 179):
  - $\sqrt{Y}$: mild transformation; used when $Var(Y|X) \propto E(Y|X)$ as for Poisson data
  - $\log(Y)$: most common; if $Var(Y|X) \propto [E(Y|X)]^2$ or errors behave like percentage of Y.
  - $1/Y$: used when responses are mostly close to 0, but some large values occur.

Box-Cox

- As GLM (Agresti, p. 117):
  - distribution of Y not restricted to Normal
  - model parameters describe g[E(Y)] rather than E(g(Y)) as in transformed data approach
  - GLMs provide a unified theory of modelling that encompasses the most important models for continuous and discrete variables.

LOG ODDS, ODDS, ODDS RATIO

$\frac{1}{3}$

$\frac{5}{1}$

- Let $\underline{\pi} = P(\text{"}success\text{"})$, $0 < \pi < 1$.
- The ODDS in favour of "success" is:

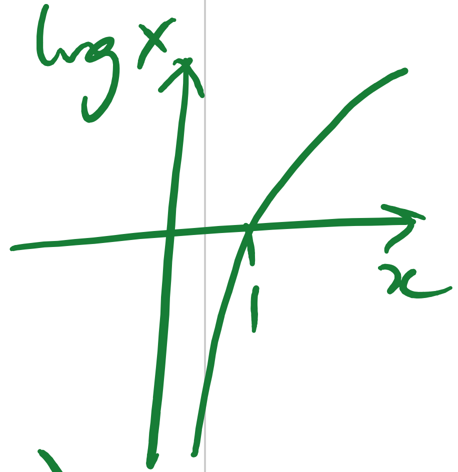
$(0, \infty)$

$$\frac{\pi}{1-\pi} = \frac{P(\text{"success"})}{P(\text{"failure"})}$$

- Then the LOG ODDS is:

$(-\infty, \infty)$

$$\log\left(\frac{\pi}{1-\pi}\right)$$



- An ODDS RATIO is a ratio of ODDS.

$(0, \infty)$

$$\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1}{\pi_2}\left(\frac{1-\pi_2}{1-\pi_1}\right)$$

Binary Logistic Regression

AIM
- $\boxed{E(Y|X) = \pi}$
- $\text{Var}(Y|X) = \pi(1-\pi)$. *Notice that variance is not constant!*
- Logistic regression model:

mean kernel function
linear in the parameters

$$\underbrace{\left[\log\left(\frac{\pi}{1-\pi}\right)\right]}_{\text{logit}} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \qquad (1)$$

- Linear predictor:

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- LOGISTIC FUNCTION: Find by inverting equation (1)

$$\boxed{\pi(\eta) = \frac{e^{\eta}}{1+e^{\eta}}} = \frac{1}{1+e^{-\eta}}$$

$$= \frac{e^{\eta}/e^{\eta}}{1/e^{\eta} + \frac{e^{\eta}}{e^{\eta}}}$$

$$\log\left(\frac{\pi}{1-\pi}\right) = \eta$$

$$\frac{\pi}{1-\pi} = e^{\eta}$$

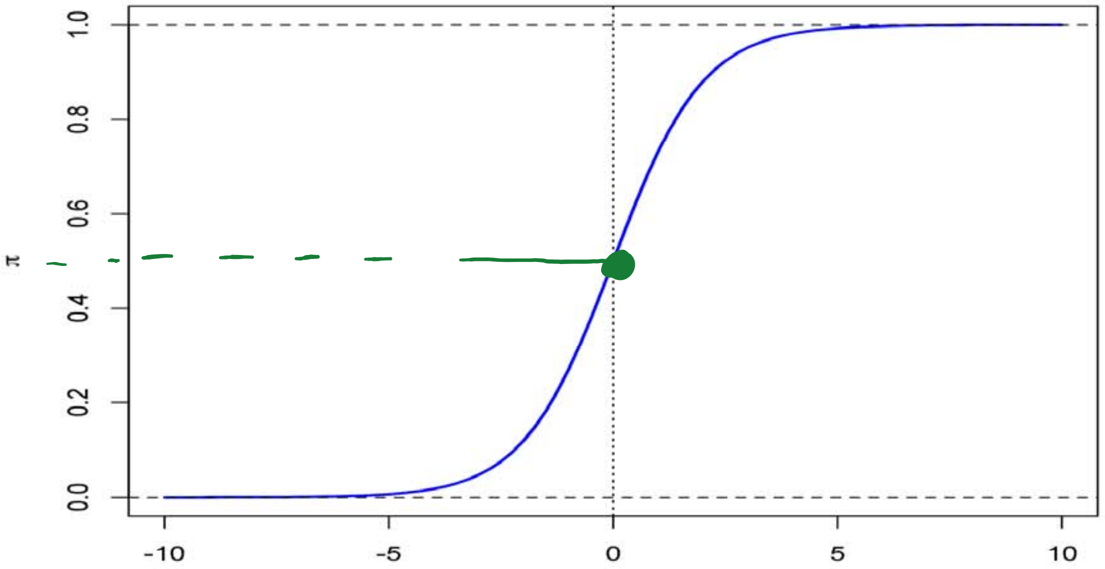
$$\pi = e^{\eta} - \pi e^{\eta}$$

What does the logistic function look like?

- LOGISTIC FUNCTION: $\pi = \frac{e^{\eta}}{1+e^{\eta}}$ $= \frac{1}{1+e^{-\eta}}$

π

0.5



$-\infty \leftarrow \eta \rightarrow +\infty$

$\eta \to +\infty$

$\pi \to 1$

$\eta \to -\infty$

$\pi \to 0$

$0 < \pi < 1$

- S-shaped; sigmoid function
- Horizontal asymptotes at 0 and 1; the logistic function, $\pi(\eta)$ varies between 0 and 1

Binary Logistic Regression Model

$\log_e$

$\ln$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}, \quad i = 1, \ldots, n$$

- Log-odds, $\log(\pi/(1-\pi))$ are between $-\infty$ and ∞ (good characteristic of a link function)
- As π_i (the probability of "success") increases, odds of success and log-odds increase
- Predicts the natural log of the odds for a subject being in one category or another
- Regression coefficients can be used to estimate odds ratio for each of the independent variables
- Tells which predictors can be used to determine if a subject was in a category of interest

How to estimate the parameter coefficients?

Maximum Likelihood Estimation

- Data: $Y_i = \begin{cases} 1 & \text{if response is in category of interest} \\ 0 & \text{otherwise} \end{cases}$
- Model: $P(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$
- Assume: The n observations are independent
- Joint density:

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \Pi_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

where

$\pi_i(\mu_i)$

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip})} = \frac{e^{\mu}}{1+e^{\mu}}$$

and $1 - \pi_i = 1 - \frac{e^{\mu}}{1+e^{\mu}} = \frac{1+e^{\mu}-e^{\mu}}{1+e^{\mu}} = \frac{1}{1+e^{\mu}} = (1+e^{\mu})^{-1}$

Maximum Likelihood Estimation

y_i

$\log ab = \log a + \log b$

- Likelihood function: Plug in observed data and think of the joint density as a function of β's-

$$\mathcal{L}(\beta_0, \ldots, \beta_p) = \Pi_{i=1}^{n} \pi_i(\boldsymbol{\beta})^{y_i}(1 - \pi_i(\boldsymbol{\beta}))^{1-y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{M_i}}{1+e^{M_i}}\right)^{y_i} \left[(1+e^{M})^{-1}\right]^{1-y_i}$$

- Log-likelihood function:

$$\log \mathcal{L}(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}) \\ - y_i \log(1 + \exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip})) \\ - (1 - y_i) \log(1 + \exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}))]$$

- Maximize the log-likelihood:

$$(\widehat{\beta}_0, \ldots, \widehat{\beta}_p) = \arg\max \{\log \mathcal{L}(\beta_0, \ldots, \beta_p)\}$$

MLE solution methods

- No explicit expression exists for the maximum likelihood estimators -$(\widehat{\beta}_0, \ldots, \widehat{\beta}_{\boldsymbol{p}})$.
- Two iterative numerical solution methods are:
  (1) Newton-Raphson algorithm
  (2) Fisher scoring or Iteratively Re-weighted Least Squares (IWLS). This is done in `glm()`.

Large-sample properties of MLEs

If model is correct, and sample size is large enough, as $n \to \infty$

1. MLEs are nearly unbiased
2. MLEs have minimum variance
3. MLEs are Normally distributed
4. Formulas for standard errors of MLEs are well-known. Estimates of standard errors are available as by-product of numerical optimization (maximization) procedures.

Case Study III: The Donner Party Example

Case Study III: The Data

- Data: n=45 pioneers

AGE	SEX	STATUS
23	MALE	DIED
40	FEMALE	SURVIVED
40	MALE	SURVIVED
30	MALE	DIED
28	MALE	DIED
40	MALE	DIED
	...	

  - AGE: Adults, 15-65 yrs old
  - SEX: 15 Females, 30 Males
  - BINARY OUTCOME: 25 Died, 20 Survived

- Questions: What are the odds of survival for a 20-yr old female? Compare the odds of survival to that of a male of the same age.

for females

Case Study III: Binary Logistic Regression Additive Model

(15, 65) (0, 1)

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \overset{(15,65)}{Age_{i1}} + \beta_2 \overset{(0,1)}{Sex_{i2}}, \quad i = 1, \ldots, 45$$

$0 < \pi < 1$

- ~~Cannot predict survival ($\pi = 1$) or death ($\pi = 0$) of a pioneer~~ X

Descriptive

- Can estimate:
  - π_i (the probability of survival)
  - odds of survival and
  - log-odds of survival based on *Age* and *Sex* of a pioneer

Inferential

- Can be used to get point and interval estimates of odds ratios
- Can test which predictors are relevant to determine odds of survival

Using R for fitting GLMs

- fitting function:

```
glm(formula, family, data)
```

- family: link function, distribution of Y.
  Examples include `binomial`, `gaussian`, `poisson`, `Gamma`

  logit identity log

- complementary functions:
  - `coefficients()`: coefficient estimates
  - `summary()`: prints a summary of results
  - `anova()`: produces an analysis of variance table
  - `residuals`
  - `deviance`
- Optimization technique: Fisher Scoring / IWLS

Case Study III: Fitted equations

Using *defaults*, $\pi = P(SURVIVED)$:

(1) $$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 3.23 - 0.078 Age_i - 1.60 \mathbb{1}_{Male,i}$$

$X_2 = 1$ if Males

$$\log\left(\frac{\pi_1}{1-\pi_1}\right) = -\log\left(\frac{1-\pi_1}{\pi_1}\right)$$
$$= \log\left(\frac{\pi_1}{1-\pi_1}\right)^{-1}$$

Using other reference status, $\pi = P(DIED)$:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -3.23 + 0.078 Age_i + 1.60 \mathbb{1}_{Male,i}$$

Using sex reference group as Males, $\pi = P(SURVIVED)$:

(3) $$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 1.63 - 0.078 Age_i + 1.60 \mathbb{1}_{Female,i}$$

$X_2 = 1$ if Female

Using sex reference group as Males, $\pi = P(DIED)$:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -1.63 + 0.078 Age_i - 1.60 \mathbb{1}_{Female,i}$$

Case Study III: Using Fitted equation

$\pi = \frac{e^{M}}{1+e^{M}}$

Using the fitted equation for $\pi = P(SURVIVED)$:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 1.63 - 0.078 Age_i + 1.60 \mathbb{1}_{Female,i},$$

$\frac{odds}{1+odds}$

$e^{\log odds}$

Q: Estimate the log odds, odds and probability of survival for a:

	Log odds, $\log(\frac{\hat{\pi}}{1-\hat{\pi}})$	Odds, $\frac{\hat{\pi}}{1-\hat{\pi}}$	$\hat{\pi}$
$\mathbb{1}_F = 1$ (i) 20-yr old Female	$1.63 - 0.078(20) + 1.6(1)$		
(ii) 40-yr old Female	$1.63 - 0.078(40) + 1.6$		
$\mathbb{1}_F = 0$ (iii) 20-yr old Male	$1.63 - 0.078(20)$	$e^{1.63-0.078(20)}$	
(iv) 40-yr old Male	$1.63 - 0.078(40)$		

Case Study III: Using Fitted equation

Using the fitted equation for $\pi = P(SURVIVED)$:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 1.63 - 0.078 Age_i + 1.60 \mathbb{1}_{Female,i},$$

Q: Estimate the log odds, odds and probability of survival for a:

	Log odds, $\log(\frac{\hat{\pi}}{1-\hat{\pi}})$	Odds, $\frac{\hat{\pi}}{1-\hat{\pi}}$	$\hat{\pi}$
(i) 20-yr old Female	1.67	5.31	0.84
(ii) 40-yr old Female	0.11	1.12	0.53
(iii) 20-yr old Male	0.07	1.07	0.52
(iv) 40-yr old Male	-1.49	0.225	0.18

Qs: Compare the odds of survival for a 40-yr old Female to that of a 20-yr old Female. Compare the odds of survival for a 20-yr old Female to that of a Male of the same age.

Case Study III: Odds Ratios

1. Compare the odds of survival for a 40-yr old Female to that of a 20-yr old Female.

$$\frac{1.12}{5.31} = 0.21 \approx \frac{1}{5}$$

Hence, the odds of survival for a 20-yr old Female are about 5 times the odds for a 40-yr old Female.

2. Compare the odds of survival for a 20-yr old Female to that of a Male of the same age. 40yr

$$\frac{5.31}{1.07} = 4.96 \approx 5$$

$$\frac{1.12}{0.225} = 4.96.$$

Hence, the odds of survival for a 20-yr old Female are about 5 times the odds for a Male of the same age.

Interpreting coefficients of a Binary Logistic model

"success"

For $\boxed{\pi = P(Y = 1)}$, we model

$$\log\left(\frac{\pi}{1-\pi}\right) = \log \text{odds}_{\{Y=1\}} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Let ω be the odds that $Y\!=\!1$ based on $X_1, \ldots, X_p$, then

$$\omega = \exp\{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\}.$$

Interpretation of β_1: Holding $X_2, \ldots, X_p$ fixed, the ratio of the odds ('ODDS RATIO') that $Y\!=\!1$ at $X_1\!=\!a$ to $X_1\!=\!b$ is

$$\frac{\omega_a}{\omega_b} = \exp\{\beta_1(a-b)\}. = \frac{e^{\beta_0+\beta_1 a+\beta_2 X_2+\cdots+\beta_p X_p}}{e^{\beta_0+\beta_1 b+\beta_2 X_2+\cdots+\beta_p X_p}}$$

* If X_1 increases by 1 unit, holding all other X's constant, the odds that $Y\!=\!1$ change by a multiplicative factor of e^{β_1}.

$$e^{2\beta_1} = e^{\beta_1} * e^{\beta_1}$$

Case Study III: Using coefficients to find Odds Ratios

$$\boxed{\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 = 1.63 - 0.078 Age + 1.60 \mathbb{1}_{Female}}$$

1. (Fixed Sex) Compare the odds of survival for a 40-yr old (Female/Male) to that of a 20-yr old (Female/Male).

   −0.078

   $$\exp\{-0.78(40-20)\} = 0.21 \approx \frac{1}{5}$$

   Hence, the odds of survival for a 20-yr old are about 5 times the odds for a 40-yr old of the same sex.

2. (Fixed Age) Compare the odds of survival for a 20-yr old Female to that of a Male of the same age.

   $$\exp\{1.60(1-0)\} = 4.95 \approx 5$$

   Hence, the odds of survival for a Female are about 5 times the odds for a Male of the same age.

Next

- Confidence interval for Odds Ratio
- Testing β's $\rightarrow$ Higher-order Models