Chapter 3 logistic regression

linear regression $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i$

Assume : $e_i \overset{iid}{\sim} N(0, \sigma^2)$

$$\mu_i = E(y_i) = \underbrace{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}_{\text{linear predictor}} + e_i$$

logistic regression ~~for~~ **binary data**

① **Ungrouped data** $y_i = 1, 0$ $i = 1, \text{—}, n$

y	x_1	x_2	$\cdots$	x_p
1 or 0				

$y_i \sim \text{Bernoulli}(p_i)$

② **grouped data** ~~y_1~~ ~~n_1~~

n	r	x_1,	---,	x_p
~~100~~ 100	60			
500	85			
⋮	⋮			

$r_i \sim \text{Binomial}(n_i, p_i)$

e.g. MATH 3424

p_i – depend on $\underset{\sim}{x}$

x_1 – assignment score

x_2 – quiz score

x_3 – ~~f~~final exam score

⋮

y – ~~grade~~ ~~grade~~ $<\,^{1}_{0}$

$$\underline{y}\quad 100\left\{\begin{matrix}\left.\begin{matrix}1\\ \vdots\\ 1\end{matrix}\right\}60\\ \left.\begin{matrix}0\\ \vdots\\ 0\end{matrix}\right\}40\end{matrix}\right.$$

linear regression model

① Normality ✗ C.L.T.

$r_i \sim \text{Binomial}(n_i, p_i)$

$n_i \to \infty$, $r_i \mathrel{\dot\sim} N(n_i p_i, n_i p_i q_i)$

② constant variance ? NO

$$\boxed{y_i} = \underbrace{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}} + e_i \qquad y_i \sim \text{Bernoulli}(P_i)$$

$\|$

$1, 0 \qquad E(y_i) = P_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + E(e_i)$

Assume

$\Rightarrow e_i = y_i - P_i$

$$= \begin{cases} 1 - P_i & \text{when } y_i = 1 \\ -P_i & \text{when } y_i = 0 \end{cases}$$

$$Var(e_i) = (1-P_i)^2 * Pr(y_i = 1) + (-P_i)^2 * Pr(y_i = 0)$$

$$= (1-P_i)^2 * P_i + (-P_i)^2 * (1-P_i)$$

$$= P_i Q_i \qquad P_i \text{ — diff. for diff } i$$

$\neq$ constant

③ linearity X

$$\hat{P_i} = \underbrace{\hat\beta_0 + \hat\beta_1 X_{i1} + \cdots + \hat\beta_p X_{ip}}_{\text{give any real no.}}$$

$0 < P_i < 1$

~~$\Rightarrow$ make transformation of P_i~~ ?

$$P_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

$\Rightarrow$ make transformation of μ_i ?

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

link function

e.g. ① $Y_i \sim N \Rightarrow$ linear regression

$$\mu_i = E(y_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

$$\Downarrow$$

$g(\mu_i) = \mu_i$ — link - identity

② $Y_i \sim$ Bernoulli $(P_i) \Rightarrow$ logistic regression

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

$$\uparrow$$

$0 < P_i < 1$ *

- logit function

$$g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$$

$$\Rightarrow \cancel{g(\mu_i)} \log_e\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

$$\Rightarrow \frac{P_i}{1-P_i} = \exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip})$$

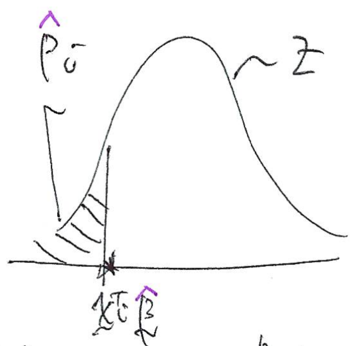
$$\Rightarrow P_i = \frac{\overset{+ve}{\exp}(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip})}{\boxed{1} + \exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip})}$$

$0 < P_i < 1$

- Probit function

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

c.d.f. of Z



$$\Rightarrow g(P_i) = \Phi^{-1}(P_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

$$\Rightarrow P_i = \Phi(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip})$$

③ $Y_i \sim \text{Poisson}(\mu_i)$ $\Rightarrow$ Poisson Regression

$\mu_i > 0$

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

Choose $g(\mu_i) = \log_e(\mu_i)$ — Link function = log function

$$\Rightarrow \mu_i = \exp(\underbrace{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}})$$

Generalized linear regression

↑

DISTRIBUTION & LINK

- Estimation / Hypothesis Testing

Estimation

grouped data ① $Y_i \sim \text{Binomial}(n_i, p_i)$

$n_i \to \infty$

② Link function $g(\mu_i) = g(p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$

③ ~~Cont~~ Constant variance

~~Ca~~ If we ~~fit~~ fit

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) \text{ on } X_1, \text{—}, X_p$$

Is $Var\left(\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)\right)$ constant ?

Example 1 – one independent variable

Data

population prop. ← Sample proportion

x n r $\log(x)$

	Obs	load	speci	fail	logload
P_1	1	5	600	13	1.60944
P_2	2	35	500	95	3.55535
P_3	3	70	600	189	4.24850
P_4	4	80	300	95	4.38203
P_5	5	90	300	130	4.49981

P_i depends on $\underset{\sim}{x}$

$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 * \text{logload}_i$

$\Rightarrow$ grouped data

$r_i \sim \text{Binomial}(n_i, P_i)$

n_i ~~ex~~ large $\forall i$

Fit $\ln\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right)$ on ~~x~~logload$_i$

$\hat{P}_i = \frac{r_i}{n_i}$ $\quad i = 1, 2, \cdots, 5$

~~too~~ Model : $\ln\left(\frac{P_i}{1-P_i}\right) = \boxed{\beta_0} + \boxed{\beta_1} X_{i1}$ $\quad i = 1, \cdots, 5$

Obs	n	r	$\hat{P}$	y	Logload
1	600	13	$13/600$	$\ln\left(\frac{13/600}{1-13/600}\right) = -3.81008$	
2	500	95	$95/500$	$\ln\left(\frac{95/500}{1-95/500}\right) = -1.45001$	
3	600	189		-0.77685	
4	300	95		-0.76913	
5	300	130		-0.26826	

$\Rightarrow$ Fit a model of y on X

y: $\ln\left(\frac{\frac{r}{n}}{1-\frac{r}{n}}\right)$ $\quad$ X: log load

Is $\mathrm{Var}\left(\ln\left(\frac{\hat{P}}{1-\hat{P}}\right)\right)$ constant ?

3423 Delta Method

$$\mathrm{Var}\Big(\underbrace{\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)}_{f(\hat{p}_i)}\Big) \overset{?}{=} \text{constant}$$

$$f(\hat{p}_i) = \underbrace{f(p_i)}_{\text{constant}} + f'(p_i) * (\hat{p}_i - p_i) + \cdots$$

$$\mathrm{Var}(f(\hat{p}_i)) = (f'(p_i))^2\,\mathrm{Var}(\hat{p}_i) \qquad *$$

$$f(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) \qquad \hat{p}_i = \frac{r_i}{n_i}$$

$$f'(p_i) = \frac{1}{p_i(1-p_i)}$$

Assume $r_i \sim \text{Binomial}(n_i, p_i)$

$$\Rightarrow \mathrm{Var}(r_i) = n_i p_i (1-p_i)$$

$$\mathrm{Var}(\hat{p}_i) = \frac{1}{n_i^2}\mathrm{Var}(r_i) = \frac{1}{n_i^2} n_i p_i (1-p_i) = \frac{p_i(1-p_i)}{n_i}$$

$$\mathrm{Var}\left(\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)\right) = \left(\frac{1}{p_i(1-p_i)}\right)^2 * \frac{p_i(1-p_i)}{n_i}$$

$$= \boxed{\frac{1}{n_i p_i (1-p_i)}} \sim \sigma_i^2 \neq \text{constant}$$

depends on $\underset{\sim}{X}$ $\qquad i = 1, \ldots, s$

<u>Weighted least squares</u>

<u>Method of least squares</u> from Chapter 1 $\mathrm{Var}(\underset{\sim}{e}) = \sigma^2 \underset{\sim}{I}$

$$\text{Min} \sum_{i=1}^{n} \hat{e}_i^2 \quad \text{OR Min Res S.S.}$$

$$\Rightarrow \text{Min} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Rightarrow \text{Min } (\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta})^T(\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta})$$

$$\Rightarrow \hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{X}^T \underset{\sim}{Y} \quad \& \quad \mathrm{Var}(\hat{\underset{\sim}{\beta}}) = (\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\sigma^2$$

Weighted least squares

$$\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{e}$$

If $Var(\underset{\sim}{e}) = \underset{\sim}{V} \Rightarrow Var(\underset{\sim}{Y}) = \underset{\sim}{V}$

Find $\underset{\sim}{\tilde\beta}_V$ s.t.

~~Min~~ $SS_{res,V} = (\underset{\sim}{Y} - \underset{\sim}{X}\tilde{\beta}_V)^T \underset{\sim}{V}^{-1} (\underset{\sim}{Y} - \underset{\sim}{X}\tilde{\beta}_V)$

$$= (\underset{\sim}{Y}^T - \tilde{\beta}_V^T \underset{\sim}{X}^T)\underset{\sim}{V}^{-1}(\underset{\sim}{Y} - \underset{\sim}{X}\tilde{\beta}_V)$$

$$= (\underset{\sim}{Y}^T \underset{\sim}{V}^{-1} - \tilde{\beta}_V^T \underset{\sim}{X}^T \underset{\sim}{V}^{-1})(\underset{\sim}{Y} - \underset{\sim}{X}\tilde{\beta}_V)$$

$$= \underset{\sim}{Y}^T \underset{\sim}{V}^{-1} \underset{\sim}{Y} - \cdots\cdots \quad \text{is minized}$$

$$\Rightarrow \tilde{\beta}_V = \boxed{(\underset{\sim}{X}^T \underset{\sim}{V}^{-1} \underset{\sim}{X})^{-1}(\underset{\sim}{X}^T \underset{\sim}{V}^{-1}}\,\underset{\sim}{Y})$$

$\underset{\sim}{c}$

If $\underset{\sim}{V} = \sigma^2 \underset{\sim}{I} \Rightarrow \hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{X}^T\underset{\sim}{Y}$

From Theorem 3.2 in Chapter 1

$$Var(\underset{\sim}{c}\,\underset{\sim}{Y}) = \underset{\sim}{c}\,Var(\underset{\sim}{Y})\,\underset{\sim}{c}^T$$

$$\Rightarrow Var(\tilde{\beta}_V) = \cancel{(\underset{\sim}{X}^T\underset{\sim}{V})}(\underset{\sim}{X}^T\underset{\sim}{V}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}^T\underset{\sim}{V}^{-1}\underset{\sim}{V}$$

$$\underset{\sim}{V}^{-1}\underset{\sim}{X}(\underset{\sim}{X}^T\underset{\sim}{V}^{-1}\underset{\sim}{X})^{-1}$$

$$= (\underset{\sim}{X}^T \boxed{\underset{\sim}{V}^{-1}} \underset{\sim}{X})^{-1}$$

$Var(\hat{\underset{\sim}{\beta}}) = (\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\sigma^2$

Consider $\underset{\sim}{V} = \sigma^2 \underset{\sim}{I}$

Let $\underset{\sim}{V} = \begin{pmatrix} \sigma_1^2 & & \underset{\sim}{0} \\ & \ddots & \\ \underset{\sim}{0} & & \sigma_s^2 \end{pmatrix}$ obs. are indep.

$\Rightarrow$ ~~Ro~~ $SS_{res,\,weighted} = (\underset{\sim}{Y} - \underset{\sim}{X}\tilde{\beta}_V)^T \underset{\sim}{V}^{-1}(\underset{\sim}{Y} - \underset{\sim}{X}\tilde{\beta}_V)$

$$= \sum_{i=1}^{s} \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2}$$

$$= \sum_{i=1}^{s} \boxed{\frac{1}{\sigma_i^2}} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{s} w_i (y_i - \hat{y}_i)^2$$

$$\hat{\sigma}_i^2 = \widehat{Var}\left(\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)\right) = \frac{1}{n_i \hat{p}_i(1-\hat{p}_i)}$$

$$w_i = \frac{1}{\hat{\sigma}_i^2} = n_i \hat{p}_i (1-\hat{p}_i) = n_i \frac{r_i}{n_i}\left(1 - \frac{r_i}{n_i}\right) = \frac{r_i(n_i - r_i)}{n_i}$$

w_i

Example 1 – one independent variable

Data

$$y = \ln\left(\frac{\frac{r_i}{n_i}}{1-\frac{r_i}{n_i}}\right)$$

$$w = n_i\left(\frac{r_i}{n_i}\right)\left(1-\frac{r_i}{n_i}\right) = \frac{r_i(n_i-r_i)}{n_i}$$

	x	n	r	log(x)		
Obs	**load**	**speci**	**fail**	**logload**	y	w
1	5	600	13	1.60944	-3.81608	12.718
2	35	500	95	3.55535	-1.45001	76.950
3	70	600	189	4.24850	-0.77685	129.465
4	80	300	95	4.38203	-0.76913	64.917
5	90	300	130	4.49981	-0.26826	73.667

x =logload

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	5.3883	3	1.7961	0.1455
Pearson	5.3792	3	1.7931	0.1460

Parameter estimates with confidence interval

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.5784	0.3682	229.4877	<.0001
logload	1	1.1400	0.0893	163.0932	<.0001

Covariance matrix

Estimated Covariance Matrix

Parameter	Intercept	logload
Intercept	0.1356	-0.03253
logload	-0.03253	0.007968

8

$$\underset{\sim}{X} = \begin{pmatrix} 1 & 1.60944 \\ 1 & 3.55535 \\ 1 & 4.24850 \\ 1 & 4.38203 \\ 1 & 4.49981 \end{pmatrix}$$

$$\underset{\sim}{X} = \begin{pmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{s1} \end{pmatrix}$$

$$\underset{\sim}{V}^{-1} = \begin{pmatrix} 12.718 & & & & \underset{\sim}{0} \\ & 76.950 & & & \\ & & 129.465 & & \\ & \underset{\sim}{0} & & 64.917 & \\ & & & & 73.667 \end{pmatrix}$$

$$\underset{\sim}{V}^{-1} = \begin{pmatrix} w_1 & & \underset{\sim}{0} \\ & \ddots & \\ \underset{\sim}{0} & & w_s \end{pmatrix}$$

$$\underset{\sim}{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_s \end{pmatrix}$$

$$\underset{\sim}{Y} = \begin{pmatrix} -3.81008 \\ -1.45001 \\ -0.77685 \\ -0.76913 \\ -0.26826 \end{pmatrix}$$

$$\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}^T \underset{\sim}{V}^{-1} \underset{\sim}{X})^{-1} (\underset{\sim}{X}^T \underset{\sim}{V}^{-1} \underset{\sim}{X})$$

$$= \begin{pmatrix} \sum_{i=1}^{s} w_i & \sum_{i=1}^{s} w_i x_{i1} \\ \sum_{i=1}^{s} w_i x_{i1} & \sum_{i=1}^{s} w_i x_{i1}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{s} w_i y_i \\ \sum_{i=1}^{s} w_i x_{i1} y_i \end{pmatrix}$$

$$= \begin{pmatrix} 357.717 & 1460.041 \\ 1460.041 & 6080.621 \end{pmatrix}^{-1} \begin{pmatrix} -330.31 \\ -1209.7 \end{pmatrix}$$

$$= \begin{pmatrix} 0.146036 & -0.0336246 \\ -0.0336246 & 0.00823819 \end{pmatrix} \begin{pmatrix} -336.31 \\ -1209.7 \end{pmatrix}$$

$$= \begin{pmatrix} -5.5784 \\ \cancel{0}1.1405 \end{pmatrix}$$