Week 13. Logistic Reg
14. Thanksgiving
15. Review
16. Final

$$\text{odds} \triangleq \frac{p(Y=1)}{1-p(Y=1)} = 1 \qquad p(Y=1) = \tfrac{1}{2}$$

> 1 :

< 1

$Y=1$

$$[\text{Odds Ratio}] = \frac{\text{odds}(Y=1 \mid \text{Female})}{\text{odds}(Y=1 \mid \text{Male})} = 1$$

> 1

< 1

x gender effe

F > M

F < M

Linear Reg. $\longleftrightarrow$ Logistic Reg $\hat{p}(Y=1 \mid X=x) < 1$

conti

$Y \quad X_1 \; X_2 \cdots X_p$

$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$

$Y \sim N(\beta_0 + \beta_1 X_1 + \cdots \beta_p X_p, \sigma^2)$

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 - \hat{\beta}_p X_p$

$E(X)$

1, 0 > $Y \quad X_1 \; X_2 \cdots X_p$

link

$-\infty <$ $\quad < \infty$

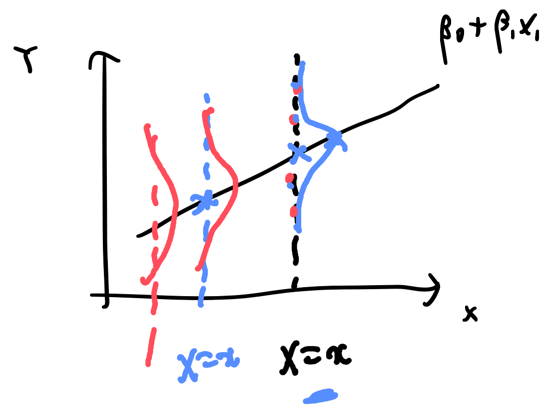
$$\log\left[\frac{p(Y=1)}{1-p(Y=1)}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

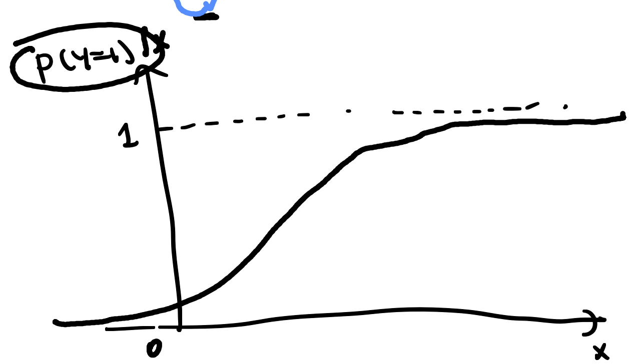
$-\infty < \quad < \infty$

$$p(Y=1 \mid X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots \beta_p X_p)}$$

$p(Y=1 \mid x) = 0.8$

$Y \sim \text{Bernoulli}(p(Y=1 \mid x))$





Interpretation of β_1 $\qquad \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \underline{\beta_1} x$

male

▪ x is categorical $\begin{cases} \underline{\text{Ref}} & 0 \\ \underline{\text{comp.} \; 1} & \end{cases}$

fem

Female

$$\frac{\text{odds}(Y=1 \mid \text{comp})}{\text{odds}(Y=1 \mid \text{ref})} = \text{OR}(\text{comp vs. ref}) = e^{\beta_1}$$

$\beta_1 = 0$

fem $\qquad$ male

- $\beta_1 = 0 \quad e^{\beta_1} = 1$
- $\beta_1 > 0 \quad e^{\beta_1} > 1$
- $\beta_1 < 0 \quad < 1$

▪ x is continuous $\quad x \rightarrow x+1$

$$\frac{\text{odds}(Y=1 \mid fib = x+1)}{\text{odds}(Y=1 \mid fib = x)} = \text{OR}(fib = x+1 \text{ vs. } fib = x)$$

$$= e^{\hat{\beta}_1} = e^{1.91} = > 1$$

fib↑ p↑

$$\frac{\text{odds}(Y=1 \mid gaw = x+1)}{\text{odds}(Y=1 \mid gaw = x)} = \text{OR}(g = x+1 \text{ vs. } g = x)$$

$$= e^{\hat{\beta}_2} = e^{0.155} > 1$$

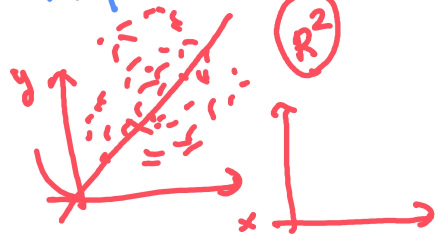
gaw↑ p↑.

Model Selection

$X_1 \cdots X_p$

[stepwise, forward, backward] selection + AIC, BIC

$H_0: \beta_1 = 0$ $\quad \neq 0$



R^2

β

$\underline{R}^2 \approx 0.1$

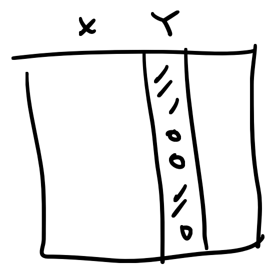
Model Fitting

- significance of X_j
- Interpretation of $\hat{\beta}_j$
- Goodness-of-fit
  - pseudo R^2
  - H_0: model fits the data well
  - H_a: " doesn't

$\hat{p}(Y=1|x)$

Model Diagnostics

- Residual plot
  - deviance: $\text{std}\left(\boxed{Y} - \boxed{\hat{p}(Y=1|x)}\right) \sim N(0,1)$
  - pearson
  - ± 2
- Cook's D



prediction

$\boxed{\hat{p}(Y=1|X)}$

$\hat{\beta}_0 \ \hat{\beta}_1 \cdots \hat{\beta}_p$

$$\hat{p}(Y=1) = \frac{\exp(\cdots)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1 \cdots)}$$

$[X_1 \quad X_2 \quad X_3 \quad X_4]$

Backward selection + AIC / BIC

AIC : the smaller error + $(2 * p)$ model penalty w

① Full Model glm$(Y \sim X_1 + \cdots + X_4)$ AIC : 500

②
(i) w/o X_1
$Y \sim X_2 + X_3 + X_4$
AIC = 450

(ii) w/o X_2
$Y \sim X_1 + X_3 + X_4$
AIC = 470

(iii) w/o X_3
$Y \sim X_1 + X_2 + X_4$
AIC = 499

(iv) w/o X_4
$Y \sim X_1 + X_2 + X_3$
AIC = 480

Remove X_1

$[X_2 \quad X_3 \quad X_4]$

③
(i) w/o $X_1 . X_2$
$Y \sim X_3 + X_4$
AIC = 620

(ii) w/o $X_1 . X_3$
$Y \sim X_2 + X_4$
AIC = 510

(iii) w/o $X_1 . X_4$
$Y \sim X_2 + X_3$
AIC = ~~505~~ 455

~~X_4~~

$[Y \sim X_2 + X_3 + X_4]$

Forward selection + AIC

① (i) $Y \sim X_1$ (ii) $Y \sim X_2$ (iii) $Y \sim X_3$ (iv) $Y \sim X_4$

AIC = 700 AIC = 800 AIC = 900 AIC = 750

Add X_1

$(X_2 \quad X_3 \quad X_4)$

② (i) $Y \sim X_1 + X_2$ (ii) $Y \sim X_1 + X_3$ (iii) $Y \sim X_1 + X_4$

AIC = 650 AIC = 600 AIC = 700

Add X_3

AIC error + 2p

BIC log n p

log 300

p ↑ : comp ↑

n

Amputation Data   significance level $\alpha = 0.1$

amputation ~ illness_severity + diabetes + ulcers

* who will have the highest chance?

$\hat{p}(Y=1 \mid x=x)$

$OR(\cancel{L\ vs.\ M})$
L vs. 0

- Illness_severity (L M H) ref

$$\left[\begin{array}{l} OR(L\ vs.\ H) = \dfrac{odds(Y=1 \mid L)}{odds(Y=1 \mid H)} = e^{-2.19} = 0.11 < 1 \\ OR(M\ vs.\ H) = \dfrac{odds(Y=1 \mid M)}{odds(Y=1 \mid H)} = e^{-0.67} < 1 \end{array}\right.$$

$\left[\begin{array}{l} L < H \\ M < H \end{array}\right.$

- Diabetes (Uncontrolled vs. Controlled)

$$OR(un\ vs.\ con) = \frac{odds(Y=1 \mid un)}{odds(Y=1 \mid con)} = e^{1.83} > 1$$

Uncon > Con

- Ulcers (1 vs. 0) ref

$$OR(1\ vs.\ 0) = e^{2.18}$$

1 > 0

$\alpha = 0.1$

(High. Uncon | -)

if using significance level $\alpha = 0.05$

$\left[\begin{array}{l} L < H \\ M < H \end{array}\right]$ → sig.

→ insig.

$H_0: \beta_{ZM} = 0$ ✓

✓ L ✓ $M = H$

((High or Mode)), uncont u/w = 1

$Y \sim$ illness severity

↓

prediction (plasma data)

cut off 0.35

$$\log\left[\frac{\hat{p}(Y=1)}{1-\hat{p}(Y=1)}\right] = -6.84 + 1.83 * Fib$$

$$\hat{p}(Y=1) = \frac{\exp(-6.84 + 1.83 * Fib)}{1+\exp(-6.84+1.83 * Fib)}$$

Fib	esr (Y)	$\hat{p}$(esr=1)	$\widehat{esr}$ ($\hat{Y}$)	$\widehat{esr}$
2.52	0	0.3	0	0
2.56	0	0.2	0	0
2.19	0	0.6	1	1
2.18	0	0.4	0	1
3.41	0	0.3	0	0
⋮				

$\hat{p} \le 1$

↗ 1

0.5

↘ 0

Y=0

1

① 0.5 cut off

misclassificatio

Y. $\hat{Y}$

1/5 20%

② prop $Y=1$

sample = 0.35

misclass

2/5

n=50,000

Y=0 49,990

Y=1 10

$\hat{p}$

$Y=1$ $\hat{Y}=0$ → False negative

$Y=0$ $\hat{Y}=1$ → False positive

− :

+ :

cut-off 0.5

cut-off : sample proportion

rare

50,000

Y	$\hat{Y}$	
0	0	0
0	0	/
0	0	0
0	0	/
/	0	/
/	0	/

misc $10/50{,}000$

0.001

50,000

10

10

$Y=1$

1 : disease

0 : No-disease

False Negative $= 1$

$P(\hat{Y}=0 \mid Y=1)$

$= 10/10 = 1$

Y	$\hat{Y}$
/	
/	
/	
0	
0	
/	
/	

False Positive

$P(\hat{Y}=1 \mid Y=0)$

How to choose the optimal cut-off?

(1) ROC curve

(2) Cross-validation

10%

n=500

90%

450

test

cut-off

machine

train

pred

future

fit by

$\hat{\beta}_0 \cdot \hat{\beta}_p$ $\quad \hat{p}(Y=1 \mid 0)$

cut-off tuning