week 13 Logistic Reg
14 Thanksgiving
15 Review.
16 Final (Linear / Logistic Reg) ← Quiz, programing.

odds $\triangleq \dfrac{p(Y=1)}{1-p(Y=1)}$ $= 1$: $p(Y=1) = 0.5$

> 1 higher chance of event

< 1 lower chance of event

* Odds Ratio (OR)

↳ $\dfrac{\text{odds}(Y=1 \mid \text{female})}{\text{odds}(Y=1 \mid \text{male})} = 1 \Rightarrow$ Female, male: equal chance

∴ no gender effect.

$> 1 \Rightarrow \text{odds}(Y=1 \mid \underline{\text{Female}}) > \text{odds}(Y=1 \mid \underline{\text{Male}})$.

$< 1 \Rightarrow$ " $<$ "

Linear Regression

continuous Y ; categorical / continuous: $X_1 \; X_2 \cdots X_p$

any number

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}_{} + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

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

$Y \leftrightarrow X$'s

Logistic Regression

binary Y: $0/1$; categorical / contin: $X_1 \cdots X_p$

$$0 \le P(Y=1) \le 1 \longleftrightarrow X_1 \cdots X_p$$

$$-\infty < \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) < \infty$$

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

$$P(Y=1) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

(one X)

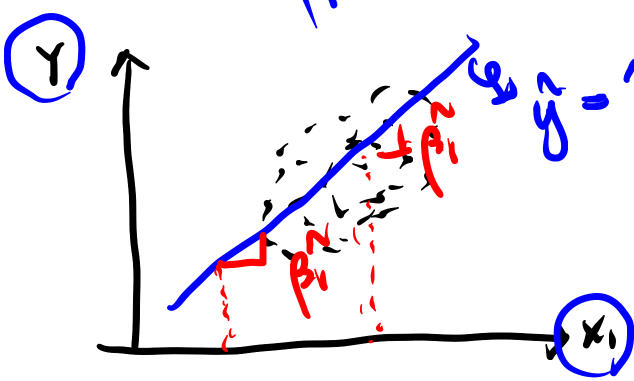
logistic fun

Goal : Inference / Prediction

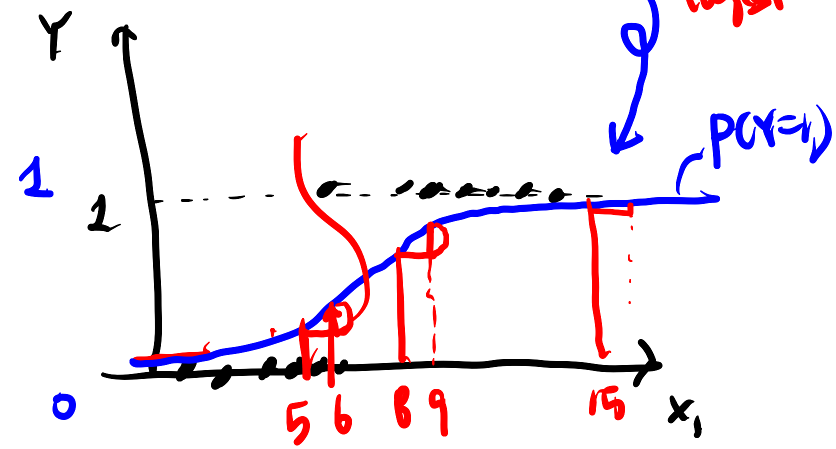
H_0: $\beta_1 = 0$ vs H_a: $\beta_1 \neq 0$

$\vdots$

$\beta_p = 0$ $\quad$ $\beta_p \neq 0$



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



Interpretation of β_1

$$\log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = \beta_0 + \beta_1 X$$

- X : categorical X
  - reference / base group : 0 male
  - comparison group : 1 Female

Fem / Male

$\beta_1 = 0.3$ ⟶

$$\frac{\text{odds}(Y=1 \mid \text{Female})}{\text{odds}(Y=1 \mid \text{male})} = OR(\text{Female vs. male}) = e^{0.3} = 1.35 > 1$$

base denominator

$\beta_1 > 0$ <0

- X : continuous

$\hat\beta > 1$

$\hat\beta > 0 \cdots e$ $\hat\beta < 1$

$\hat\beta < 0 \cdots e$

$$\frac{\text{odds}(Y=1 \mid \text{fib} = x+1)}{\text{odds}(Y=1 \mid \text{fib} = x)} = e^{1.91} > 1$$

Fib ↑ odds ↑ $p(Y=1)$ ↑

$$\frac{\text{odds}(Y=1 \mid \text{gama} = x+1)}{\text{odds}(Y=1 \mid \text{gama} = x)} = e^{0.15} > 1$$

gama ↑ $p(Y=1)$ ↑

higher fib / gama → higher chance to be in unhealthy status

Model Selection

$X_1, \dots, X_p$

Stepwise (Forward, Backward) selection + (AIC, BIC)

Final model

Model Fitting

check significance of x

Interpretation of β

Goodness-of-fit test

Hosmer Lemeshow test

H_0: Model is adequate (Model fits data well)

H_1: " inadequate (Not fit ")

pseudo R^2

Model Diagnostics

Residual plot

deviance Residual (Like Name)

pearson Residuals

$y - \hat{p}(y=1)$

Residual

$y - \hat{p}(y=1)$

0 ≤ ... ≤ 1

1

1

1

1

0

0

0

0

1

0

> 0

< 0

Cook's D

4/n

# of samples

Prediction

$\hat{p}(Y=1) \to$ classificat

X_1 X_2 ... X_4

smaller the better

Backward Selection + AIC

① Full model $\underbrace{Y \sim X_1 + X_2 + \cdots + X_4}$ AIC: 500

② (i) w/o X_1
$\underbrace{Y \sim X_2 + X_3 + X_4 + \cdots}$
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 ∵ model (i) smallest AIC: 450 < 500

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

Conclusion: Remove X_1. Keep X_2 ... X_4

Forward selection + AIC X_1 -- X_4

① $\underbrace{Y \sim X_1}$ $\underbrace{Y \sim X_2}$ $\underbrace{Y \sim X_3}$ $\underbrace{Y \sim X_4}$

AIC : 700   AIC : 800   AIC : 900   AIC : 750

Enter X_1

② $Y \sim X_1 + X_2$ $Y \sim X_1 + X_3$ ✓ $Y \sim X_1 + X_4$

Amputation Data Example

Reference / Base group.

(1) $\alpha = 0.1$

Illness severity (L. M. High)

$$\frac{\text{odds}(Y=1 \mid \text{Low})}{\text{odds}(Y=1 \mid \text{High})} = e^{-2.19} < 1$$

$= 0.11$

$$\frac{\text{odds}(Y=1 \mid \text{Moderate})}{\text{odds}(Y=1 \mid \text{High})} = e^{-0.67} < 1$$

$= 0.59$

High > Moderate > Low

diabetes (Uncontrolled, controlled)

$$\frac{\text{odds}(Y=1 \mid \text{uncontrolled})}{\text{odds}(Y=1 \mid \text{controlled})} = e^{1.03} > 1$$

Uncontrolled > controlled.

ulcer (1, 0)

$$\frac{\text{odds}(Y=1 \mid \text{at least one ulcer})}{\text{odds}(Y=1 \mid \text{No ulcer})} = e^{2.18} > 1$$

(ulcers at least one > no ulcer

2.52

2.52

$$\hat{p}(Y=1) = \frac{\exp(-6.84 + 1.83 \cdot \text{f(x)})}{1 + \exp(-6.84 + 1.83 \cdot \text{f(x)})}$$

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

Prediction (Plasma Data)

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

	fib	esr (True Y)	$\hat{p}$(esr=1)	$\hat{esr}$ ($\hat{Y}$)	$\hat{Y}$	
obs1	2.52	0	0.3	0	0	correctly class
obs2	2.56	0	0.2	0	0	correctly class
obs3	2.19	0	0.6	1	1	incorrect / mis classified
4	2.18	0	0.4	0	1	corre
5	3.41	0	0.3	0	0	corr
	⋮		⋮			

given

① 0.5 cut-off → misclassification rate

rate of obs. misclassified $\hat{Y}$

0.35 cutoff → misclass

② sample proportion cut-off

# of 1's in the data

data w/ rare 1.

(eg) 0.35

50,000

10 positive $Y=1$

49,990 negative. $Y=0$

$\hat{Y}$ will be generally small

20%

0.6
0.5
0.4
0.3
0.2

assign to 1

cut-off

assign to 0

plasma example continued.

cut-off 0.5

12.5%

misclass.

$\hat{p}$:

cut-off sample proportion 0.18

28%

more $\hat{Y}=1$

more many $\hat{Y}=0$

High False Negative rate

False Positive

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

Relatively not danger

False Negative

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

danger

cut-off / threshold

cut-off 0.5 → tool 1

cut 0.4 → tool 2

0.3 → tool 3

0.2 →

train test

logistic Reg.

$$\log\left(\frac{\hat{p}(Y=1)}{1-\hat{p}(Y=1)}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots \hat{\beta}_p X_p$$

$\alpha = 0.05$

Low Moderate High

Low ↔ High: signifi diff

Moder ↔ High: Not sig.

- Illness severity significant
- Moderate High > Low

Moderate High / uncontrolled / ulcerus