Week 12

Y ← $(X_1 \; X_2 \cdots X_p)$

interest — conti

covariates

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X_1 + \cdots + \hat{\beta_p} X_p$$

Y

$[\beta_0 + \beta_1 X_1 + \cdots \beta_p X_p]$

Linear Model

Y: conti — Linear Reg

X: cate/

Y: conti

X:

ANOVA

Logistic Reg

Y: binary (0/1)

[dise / no dise
pass/fail

$p(Y=1 \mid \;) = 0.7$

no accident : 0

< at : 1

$Y \sim x_1 + x_2 + x_3$

$[\hat{p}(Y=1 \mid x_1, x_2, x_3)]$

Y Car Accident	X_1 # of tickets	X_2 gender	X_3 age
0	3	M	26
0	1	F	53
1	5	F	32
1	2	M	47
⋮	⋮		⋮
1	3	M	

<0
1

$= 0.9$

near opp

$\hat{p}(Y=1 \mid 0)$ $x \cdots$

$= 0.4 - 1$

goal : Inference / prediction

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

Y and X

significant relationship?

(eg)

if yes, quantify

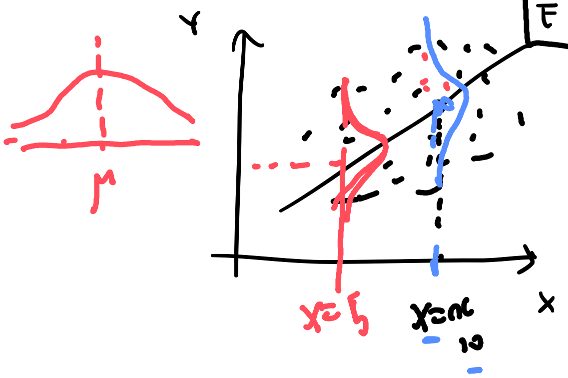
(0, 0, ...

Parametric model

linear reg model

Y : conti

$E(Y) = \beta_0 + \beta_1 x$



$Y \mid X = x \sim N(\beta_0 + \beta_1 x, \sigma^2)$

1 : car acc
0 : not

Y : binary

→ Logistic Regression

$\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

$P(Y=1)$

$P(Y=\text{head}) = 0.2$

$P(Y=\text{head}) = 0.8$

Y

H T

$Y \mid X = x \sim \text{Bernoulli}(P(x))$

(eg) coin toss $Y \sim \text{Bernoulli}(1/2)$

head $Y = 1$ (Head)

tail $Y = 0$

H H H T

specifically.

$P(x)$

odds : $\frac{p}{1-p}$: relative probability

$\frac{p(\text{car accident})}{1 - p(\text{car accident})}$ $= 1$, < 1, > 1

$Y = \begin{cases} 1 \\ 0 \end{cases}$

no event

- odds = 1 $\quad p(Y=1) = \frac{1}{2}$
- odds > 1 $\quad$ higher chance $Y=1$ $\quad p(Y=1) > \frac{1}{2}$
- odds < 1 $\quad$ less chance $Y=1$ $\quad p(Y=1) < \frac{1}{2}$

$\frac{p(Y=1)}{p(Y=0) + p(Y=1) = 1} = \frac{p(Y=1)}{1 - p(Y=1)} = \frac{p(Y=1)}{p(Y=0)}$

odds Ratio

$$OR(\text{male vs. Female}) = \frac{\text{odds}(\text{car accident} \mid \text{Male})}{\text{odds}(\text{car accident} \mid \text{Female})} \quad =1, \ <1$$

- OR = 1 : no gender effect
- OR > 1 Male > Female.
- OR < 1 Male < Female

$$\begin{array}{ccc} Y & x_1 \cdots & x \\ \begin{bmatrix} 0 & \cdots & \\ 0 & \vdots & \\ 1 & & \\ 0 & & \\ \vdots & & \end{bmatrix} \end{array}$$

lm(Y ~ age)

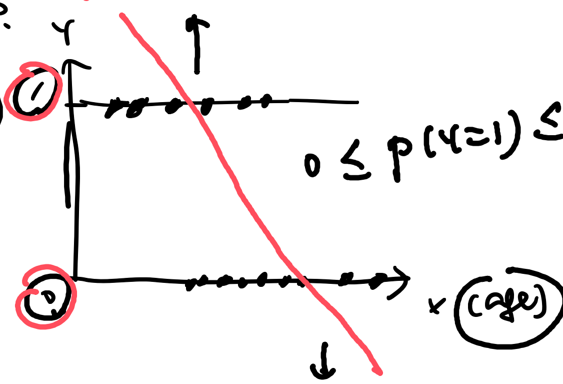
$\hat{y} < 0$

$\hat{y} > 1$

goal $P(Y=1 \mid X=x)$

why not linear regression?

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



$0 \le P(Y=1) \le 1$

Logistic Regression

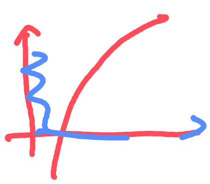
$0 \le P(Y=1) \le 1$

$$-\infty < \underbrace{\log\left(\frac{p}{1-p}\right)}_{\text{odds}} < \infty = \left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right), \quad -\infty < \quad < \infty$$

$p(Y=1 \mid X)$

$0 \le p \le 1 \qquad 0 \le \frac{p}{1-p} < \infty$

$0 \le 1-p \le 1 \qquad -\infty < \log\frac{p}{1-p} < \infty$



Bernoulli $Y \sim Bern(0.8)$

prediction

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

$$\left[\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right]$$

→ MLE

$x_1 \cdots x_p$

$$p(Y=1) = \left(\frac{e^{\beta_0 + \cdots \beta_p x_p}}{1+e^{\beta_0 \cdots \beta_p x_p}}\right) = 0.8$$

-12.29 $\quad 5.7 \quad 0.158$

Given data, $\quad \cancel{\log}\ \text{exp}\left(\frac{\hat{p}(Y=1\mid x)}{1-\hat{p}(Y=1\mid x)}\right) = \text{exp}\ \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$

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

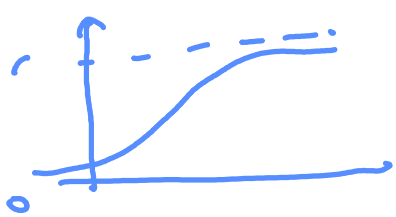
$$\frac{\hat{p}(Y=1\mid X)}{1-\hat{p}(Y=1\mid x)} = e^{[\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots \hat{\beta}_p x_p]}$$

$$\hat{p}(Y=1\mid x) = \{1-\hat{p}(Y=1\mid x)\}\cdot e^{\hat{\beta}_0 + \hat{\beta}_1 \cdots \hat{\beta}_p x_p}$$

$$\hat{p}(Y=1\mid x) + \left(e^{\hat{\beta}_0 + \cdots \hat{\beta}_p x_p}\right)\hat{p}(Y=1) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 \cdots \hat{\beta}_p x_p}$$

$$0 \le \hat{p}(Y=1\mid x) = \left[\frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 \cdots \hat{\beta}_p x_p}}{1+e^{\hat{\beta}_0 + \hat{\beta}_1 \cdots \hat{\beta}_p x_p}}\right] \le 1$$

$$\frac{e^x}{1+e^x}$$: logistic fn,



Interpretation of β $\hat{\beta}_1$

$$\left[\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \right]$$

(i) X is categorical
- 0 : male
- 1 : female

$x=0$

$\log\left(\frac{p}{1-p} \mid \text{male}\right) = \beta_0$ (exp, odds)

$x=1$

$\log\left(\frac{p}{1-p} \mid \text{female}\right) = \beta_0+\beta_1$ (exp)

odds (event | male) $= e^{\beta_0}$

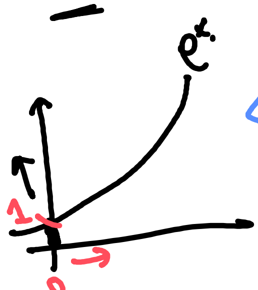
odds (event | female) $= e^{(\beta_0+\beta_1)}$

$$\text{OR(female vs. male)} = \left[\frac{\text{odds(event | female)}}{\text{odds(event | male)}}\right] = \frac{e^{\beta_0}\cdot e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

>1

<1

e^x



$\Leftrightarrow$ odds (event | fem) $= e^{\beta_1} \cdot$ odds(event | male)

compar   1   ref

no gender

$e^0 = 1$

$\beta_1 = 0 \Leftrightarrow e^{\beta_1} = 1$

$\beta_1 > 0 \Leftrightarrow e^{\beta_1} > 1$ female > male

$\beta_1 < 0 \Leftrightarrow e^{\beta_1} < 1$ female < male

(ii) X is continuous

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

$x \to x+1$

age

$X = x$ $\qquad$ $X = x+1$

$$\exp \log\left(\frac{p}{1-p}\right) = \exp\, \beta_0 + \beta_1 x$$

$$\text{odds}(\cdot \mid x) = e^{\beta_0+\beta_1 x}$$

$$\exp \log\left(\frac{p}{1-p}\right) = \exp\, \beta_0 + \beta_1(x+1)$$

$$= \beta_0 + \beta_1 x + \beta_1$$

$$\text{odds}(\text{event} \mid x+1) = e^{\beta_0+\beta_1 x+\beta_1}$$

$$OR(x+1 \text{ vs. } x) = \frac{\text{odds}(\text{event} \mid x+1)}{\text{odds}(\text{event} \mid x)} = e^{\beta_1} = \boxed{}$$

>1 <1 $=1$

$$= \frac{\cancel{e^{\beta_0}} \cdot \cancel{e^{\beta_1 x}} \cdot e^{\beta_1}}{\cancel{e^{\beta_0}} \cdot \cancel{e^{\beta_1 x}}} = e^{\beta_1}$$

$\beta_1 = 0 \Leftrightarrow e^{\beta_1} = 1$ no X effect

age

$\beta_1 > 0 \Leftrightarrow e^{\beta_1} > 1$ $\quad$ ~~past~~ $X\uparrow$ $\quad$ 1 change $\uparrow$

$\beta_1 < 0 \Leftrightarrow e^{\beta_1} < 1$ $\quad$ $X\uparrow$ $\quad$ " $\downarrow$

$Y=1$ accident

X_1 : # of tickets

X_2 : gender (male=0 / Female=1)

X_3 : age

$\hat{\beta_1} = 2.3$

$\hat{\beta_2} = 1.5$

$\hat{\beta_3} = -1.3$

more tic

$$OR_{ticket} = \frac{odds(event \mid x+1)}{odds(event \mid x)} = e^{\hat{\beta_1}} = e^{2.3} = 9.97 > 1$$

young

$$OR_{age} = \frac{odds(event \mid x+1)}{odds(event \mid x)} = e^{\hat{\beta_3}} = e^{-1.3} = 0.27 < 1$$

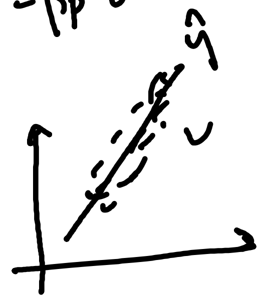
$$OR_{gender} = \frac{odds(event \mid female)}{odds(event \mid male)} = e^{1.5} = 4.48 > 1$$ F

who will have the highest chance of event?

Female. young. ticket ↑

Model Significance

- H_0: model is not useful $\beta_1 = \beta_2 \cdots = \beta_p = 0$
- H_a: model is useful



$\hat{y}$

sig.

→ Individual term significance

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

$H_0: \beta_g = 0$
$H_a: \beta_g \neq 0$

- Interpretation of β

$OR_f = \dfrac{odds(esr=1 \mid x+1)}{odds(esr=1 \mid x)} = e^{1.9} = 6.75 > 1$

$\hat{\beta} > 0$

$OR_g = \dfrac{odds(esr=1 \mid x+1)}{odds(esr=1 \mid x)} = e^{0.156} = 1.16 > 1$

Goodness-of-fit

- Pseudo R^2 SS ≈ 1
- H_0: model fits the data well
  H_a: " does "

⇒ $\hat{p}(Y=1 \mid x)$

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

Model diagnostics

- Assumption: Std res $\sim N(0,1)$
  fall betw $(-2, +2)$
- data
  cook's

Y	$\hat{p}(Y=1)$
0	0.3
0	0.4
1	0.5
1	0.6
0	
1	
0	
0	

$x_1, x_2 \ldots x_p$

CLT

$-1 <$ ρ < 1