8 Sept

Chapter 1 Simple & multiple linear regression

$\underset{\sim}{Y}$ $\quad$ $\underset{\sim}{X}$ $\quad (X_1, \ldots, X_p)$

$m \times 1$ $\quad$ $p \times 1$

Model $\underset{\sim}{Y}$ on $\underset{\sim}{X}$

- response — predictor
- outcome — covariate
- dependent variable — independent variable

$\underset{\sim}{Y}$ ($m \times 1$): $m = 1$ MATH 3423; $m > 1$ MATH 4424 multivariate statistical Analysis

$\underset{\sim}{X}$ ($p \times 1$): $p = 1$ simple; $p > 1$ multiple

linear $\quad f(x) = \alpha + \beta X \overset{\text{OR}}{=} \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

linear in parameter, no linear in X

e.g. $f(x) = \alpha + \beta X^2$

X'

$f(x) = \alpha * \beta^X$ ← non-linear

$\log f(x) = \log \alpha + x \log \beta$

Linear regression $\quad y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e$

r.v.

⇒ ① y & $(X_1, \ldots, X_p)$ not prefectly linear ~~regr~~ related

② y is observed with error

$y_{obs} = y_{true} + \boxed{e_y}$ ← Classical measurement error model

$\Rightarrow y_{true} = \gamma_0 + \gamma_1 X + \varepsilon_y$

(y_{true} & X) are not ~~pe~~ prefectly linear related

Linear regression model

$$y_i = \boxed{\beta_0} + \boxed{\beta_1} X_{i1} + \cdots + \boxed{\beta_p} X_{ip} + e_i \quad i=1, \ldots, n$$

unknown

Assume

r.v. $\sim N(0, \sigma^2)$

Purpose

- do prediction
- find the ~~si~~ ~~importion~~ important factors among $\underset{\sim}{X}$ (varable selection)

X - fixed ✓ ~~[illegible]~~ $y|\underset{\sim}{x} \sim N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \sigma^2)$

X - random e.g. $X \sim N(\mu_x, \sigma_x^2)$

$y = \beta_0 + \beta_1 X + e \Rightarrow y \sim N(\beta_0 + \beta_1 X, \sigma^2)$

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

$\beta_0, \beta_1, \sigma^2 \Rightarrow$ dist. of (y, x)

$\Rightarrow$ bivariate normal

$Var(Y)$ $Cov(X,Y)$

$$\begin{pmatrix} y \\ x \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0 + \beta_1 \mu_x \\ \mu_x \end{pmatrix}, \begin{pmatrix} \boxed{\beta_1^2 \sigma_x^2 + \sigma^2} & \beta_1 \sigma_x^2 \\ & \boxed{\sigma_x^2} \end{pmatrix} \right)$$

$Var(X)$

$$\Rightarrow y|x \sim N\left(\beta_0 + \beta_1 \mu_x + \frac{\beta_1 \sigma_x^2}{\sigma_x^2}(x - \mu_x), \beta_1^2 \sigma_x^2 + \sigma^2 - \frac{\beta_1^2 \sigma_x^4}{\sigma_x^2}\right)$$

$$\sim N(\beta_0 + \beta_1 x, \sigma^2)$$

Except If X - random & measured ~~or~~ with error

e.g. $X_{obs} = X + \varepsilon_x$

X_{obs}

~~Write down the joint distribution~~

~~of y & x $\Rightarrow$ $y|x_{obs}$~~

$\cancel{f(y|x) =}$ $y|x \sim N(\beta_0 + \beta_1 X, \sigma^2)$

$$X \sim N(\mu_x, \sigma_x^2)$$

$$X_{obs} = X + \varepsilon_x$$

$\Rightarrow y \mid X_{obs} \Rightarrow$ likelihood function

$\Rightarrow$ m.l.e.

<u>linear regression model</u>

$$y_i = \underbrace{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}_{} + e_i \qquad i = 1, -, n$$

fixed

Assume

$$\boxed{\begin{array}{l} \sim N(0, \sigma^2) \\ \text{cov}(e_i, e_j) = 0 \end{array}} \quad i \neq j$$

(obs. are indep)

$$\underset{n\times 1}{\underset{\sim}{Y}} = \underset{n\times(p+1)}{\underset{\sim}{X}} \; \underset{(p+1)\times 1}{\underset{\sim}{\beta}} + \underset{n\times 1}{\underset{\sim}{e}}$$

intercept

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \boxed{\beta_0} \\ \boxed{\begin{matrix} \beta_1 \\ \vdots \\ \beta_p \end{matrix}} \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \rightarrow \underset{\sim}{e} \sim MN\left(\underset{\sim}{0}, \boxed{\begin{pmatrix} \sigma^2 & & \underset{\sim}{0} \\ & \ddots & \\ \underset{\sim}{0} & & \sigma^2 \end{pmatrix}}\right)$$

$\sigma^2 \underset{\sim}{I}$

regression coefficient of X_j $\quad j = 1, -, p$

column with all elements equal to 1

<u>MATH 3423</u> $\quad y_i \overset{iid}{\sim} N(\mu, \sigma^2) \Leftarrow$

$$E\left(\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}\right) = \sigma^2$$

<u>MATH 3424</u> $\quad y_i = \beta_0 + \beta_1 X_{i1} \cancel{+ \cdots + \beta_p X_{ip}} + e_i$

diff. mean $\qquad \sim N(0, \sigma^2)$

$$y_i \sim N(\underline{\beta_0 + \beta_1 X_{i1}}, \sigma^2) \Leftarrow$$

Is $\dfrac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$ unbiased est. of σ^2 ?

$\Rightarrow$ Is $E\left(\dfrac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}\right)$ equal to σ^2 ? NO

$$y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i \sim N(0, \sigma^2)$$

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

$$\mu_i = \underset{\sim}{X}_i^T \underset{\sim}{\beta}$$

Define $g(\mu) = \underset{\sim}{X}^T \underset{\sim}{\beta}$

link function

~~linear~~ ① $g(\mu) = \mu$ — identity function

② y ~ binary data Y / N

Assume $f(y|x) \sim$ Bernoulli (P)

$$E(y) = P \leftarrow \mu = P$$

$$0 < P < 1$$

$$P = \frac{\exp(\underset{\sim}{X}^T \underset{\sim}{\beta})}{1 + \exp(\underset{\sim}{X}^T \underset{\sim}{\beta})} \Rightarrow \frac{P}{1-P} = \exp(\underset{\sim}{X}^T \underset{\sim}{\beta})$$

$$\Rightarrow \boxed{\ln\left(\frac{P}{1-P}\right)} = \underset{\sim}{X}^T \underset{\sim}{\beta}$$

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$$ — logit function

③ y — count data e.g.

$\sim$ Poisson (μ)

$$E(y) = \mu \qquad \mu = \exp(\underset{\sim}{X}^T \underset{\sim}{\beta})$$

$$\Rightarrow \log_e(\mu) = \underset{\sim}{X}^T \underset{\sim}{\beta}$$

$g(\mu) = \ln(\mu)$ — link function

Section 2 Estimation

Unknown parameters: $\boxed{\boxed{\beta_0, \beta_1, \cdots \beta_p}, \sigma^2}$ — method of maximum likelihood

Model: $y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i$

Assume $N(0, \sigma^2)$

Method of least squares

$$y_i = \beta_0 + \beta_1 X_{i1} + \boxed{e_i} \qquad i = 1, \ldots, n$$

~~unobservation~~

unobservable

$$\hat{e}_i = y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 X_{i1})}_{\hat{y}_i \text{ — fitted value of } y_i} \quad \text{— residual}$$

observable

Method of least squares

Find $\hat{\beta}_0, \hat{\beta}_1$ such that

$$\boxed{\sum_{i=1}^{n} (\boxed{y_i - \hat{y}_i})^2}$$

or $\sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{i1}))^2$

is minimized.

Residual ~~SS~~. Sum of Squares = Res. S.S.