- Review of Linear Regression Model

Multiple linear regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \varepsilon$$

Or in matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & & x_{nk} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$\mathbf{y}, \mathbf{X}$: observed

$\boldsymbol{\beta}$: fixed constant, unknown

$\boldsymbol{\varepsilon}$: unobservable, random

One seeks a least square estimator $\hat{\beta}$ that minimizes

$$S(\beta) = (y - X\beta)^T (y - X\beta)$$

$$\rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$$

fitted values for y are

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}Y = HY$$

- the hat matrix H is idempotent, which is the projection matrix unto the space $\text{span}(X) = \text{span}(1, X_1, \cdots X_k)$
- The estimated $\hat{\beta}$ is unbiased $E(\hat{\beta}) = \beta$ with covariance matrix $\sigma^2 (X^TX)^{-1}$
- Estimation of σ^2: an unbiased estimator of σ^2 is, $\hat{\sigma}^2 = \dfrac{SS_{Res}}{n-k-1}$

in which

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

$$= (\mathbf{y} - \mathbf{X}\hat{\beta})^T (\mathbf{y} - \mathbf{X}\hat{\beta})$$

$$= \mathbf{y}^T\mathbf{y} - (\mathbf{X}\hat{\beta})^T (\mathbf{X}\hat{\beta})$$

$$= \mathbf{y}^T\mathbf{y} - \hat{\beta}^T \mathbf{X}^T \mathbf{X}\hat{\beta} = \mathbf{y}^T\mathbf{y} - \hat{\mathbf{y}}^T\mathbf{y}$$

- Hypothesis testing in multple linear regression.

decide between the full model and a reduced model.

Partition the vector β as $\beta = (\beta_1^T, \underbrace{\beta_2^T}_{r\times 1})^T$, $r < k$

One can test against $H_0: \beta_2 = 0$

$SS_R(\beta)$, $SS_R(\beta_1)$, $SS_R(\beta_2 | \beta_1) = SS_R(\beta_1) - SS_R(\beta)$

- F test is constructed for testing H_0

$$\frac{SS_R(\beta_2 | \beta_1)/r}{\hat{\sigma}^2} \sim F_{r,\, n-k-1}$$

Confidence interval

C.I. for β_j

$$\hat{\beta}_j \pm t_{\frac{\alpha}{2}, n-k-1} \sqrt{\hat{\sigma}^2 C_{jj}}$$

$$(X^T X)^{-1} = \begin{pmatrix} C_{11} & \cdots & * \\ * & \ddots & \\ & & C_{k+1,k+1} \end{pmatrix}$$

C.I. for the mean response at $\mathbf{x}_0 = (1, x_{01}, \dots x_{0k})$ is

$$\hat{y}_0 \pm t_{\frac{\alpha}{2}, n-k-1} \sqrt{\hat{\sigma}^2 \mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0}$$

- Response variable (quantitative and Normally distributed)
- Two types of models where response variable is discrete and error terms don't follow a Normal distr.

  logistic and Poisson Regression.
- Both belong to a family of regression models called generalized linear models (GLM)

in R, use glm( ) to work with GLM.
include an "family" and a link function.
eg. glm(formular, family = binomial (probit)
data = mydata)

- GLM and exponential family.
  - GLM
    two Components.
    - Response $\sim$ exponential family distr.
    - link function (how the mean of the Response and a linear Combination of the predictors are linked)
- Exponential family. (continuous, discrete r.v.s)
  general form

$$f(y \mid \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right]$$

eg: Normal, Poisson, Bernoulli, binomial, exponential, Beta, Gamma, Weibull)

$$E(Y) = b'(\theta)$$

$$Var(Y) = b''(\theta)\, a(\phi)$$

Link function

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

$$\eta = g(\mu)$$

Poisson ~~h~~GLM: μ: positive

$$\mu = e^{\eta} \Rightarrow \eta = \log(\mu)$$

Bernoulli

$$\eta = \log \frac{\mu}{1-\mu} \quad \text{logit}$$

$$\eta = \Phi^{-1}(\mu) \quad \text{probit}$$

$$\eta = \log[-\log(1-\mu)] \quad \text{cloglog}$$

Canonical link: $\eta = \theta$ (the canonical parameter of the exp family)

Estimation of β_0, β_1

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i} , \qquad \pi_i = P(y=1 \mid x_i)$$

$$\log L(\beta) = \sum_{i=1}^{n} \left[y_i \log \pi_i + (1-y_i) \log (1-\pi_i) \right]$$

$$= \sum_{i=1}^{n} \left[y_i \log \frac{\pi_i}{1-\pi_i} + \log (1-\pi_i) \right]$$

$$\pi_i = \frac{1}{1+e^{-\beta_0-\beta_1 x_i}}$$

$$= \sum_{i=1}^{n} \left[y_i (\beta_0 + \beta_1 x_i) - \log (1+e^{\beta_0+\beta_1 x_i}) \right]$$

Maximize numerically in β_0, β_1.

no explicite solution.

Binary

$y_i = 0$ or 1

$y_i = \begin{cases} 1 & \text{if the treatment is effective} \\ 0 & \text{- - - is not - - -} \end{cases}$

Special feature:

(1) not Normal

(2) bounded $\in [0, 1]$

while $\beta_0 + \beta_1 x \in (-\infty, \infty)$

Regression function $E y_i = P(y_i = 1) = \pi_i = G(x_i)$

logical approach: relate y_i to a numerical variable V_i.

Say: $y_i = 1 \iff V_i \le \vartheta$ (blood pressure is below 110)

linear regression for V_i

$$V_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\varepsilon_i \sim \text{Normal}(0, \sigma^2)$$

$$G(x) = P(y=1) = P(V \le \vartheta)$$

$$= P\left(Z \le \frac{\vartheta - \beta_0 - \beta_1 x}{\sigma}\right)$$

$$= \Phi\left(\frac{\vartheta-\beta_0}{\sigma} - \frac{\beta_1 x}{\sigma}\right) = \Phi(\beta_0^* + \beta_1^* x) = h(x, \beta^*)$$

link func. $g(\mu) = x\beta^* = \beta_0^* + \beta_1^* x$

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

is a probit

- Logistic regression

(1) $y_i = 1 \iff V_i \le \vartheta$ (2). $V_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$\varepsilon_i \sim$ iid logistic

i.e. $F(\varepsilon) = \dfrac{e^{\varepsilon}}{1+e^{\varepsilon}}$

Then $\mu = G(x) = P(y=1)$

$$= F(\beta_0^* + \beta_1^* x) = \frac{e^{\beta_0^* + \beta_1^* x}}{1+e^{\beta_0^* + \beta_1^* x}} = \frac{1}{1+e^{-\beta_0^* - \beta_1^* x}}$$

link func. $g(\mu) = \beta_0^* + \beta_1^* x = -\log\left(\frac{1}{\mu} - 1\right) = \log\left(\frac{\mu}{1-\mu}\right)$

log-odds link func for logit response.

· exponential family distr.

$$f(y \mid \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right]$$

θ : canonical parameter, represents the location

ϕ : dispersion parameter, ... scale

$E(Y) = b'(\theta)$

·· Normal

$$f(y \mid \theta, \phi) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

$$= \exp\left[\log\{ \quad \} \right]$$

$$= \exp\left[\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}\left\{ \frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right\} \right]$$

$\theta = \mu$, $b(\theta) = \frac{1}{2}\mu^2 = \frac{1}{2}\theta^2$, $\phi = \sigma^2$, $a(\phi) = \phi$

· Poisson. $f(y \mid \theta, \phi) = e^{-\lambda} \dfrac{\lambda^y}{y!} = \exp\left[y \log\lambda - \lambda - \log y! \right]$

$\theta = \log\lambda$, $b(\theta) = \lambda = e^\theta$, $\phi = 1$, $a(\phi) = 1$

---

*

$$1 = \int_{-\infty}^{\infty} f(y \mid \theta, \phi)\, dy$$

. . . .

$E(Y) = b'(\theta)$

- Binomial

$$f(y \mid \theta, \phi) = \binom{n}{y} \mu^{y} (1-\mu)^{n-y}$$

$$= \exp\left[y \log \frac{\mu}{1-\mu} + n \log(1-\mu) + \log \binom{n}{y} \right]$$

$$\theta = \log \frac{\mu}{1-\mu}, \quad b(\theta) = -n \log(1-\mu)$$

$$= n \log(1+e^{\theta})$$

Estimation of β

$$\log L(\beta) = \sum_{i=1}^{n} \left\{ y_i(\beta_0 + \beta_1 x_i) - \log(1 + e^{\beta_0 + \beta_1 x_i}) \right\}$$

Maximize in β_0, β_1

interpretation of β_1

$$\pi_i = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_i}} \quad \text{and} \quad \frac{\pi_i}{1-\pi_i} = e^{\beta_0 + \beta_1 x_i}$$

when x_i increase by 1

$\frac{\pi_i}{1-\pi_i}$ multiplies by e^{β_1}

$\Rightarrow \beta_1$ is a change in log-odds ratio caused by a unit increase in x.

interpretation of β_0.

when $x_i = 0 \Rightarrow \dfrac{\pi_i}{1-\pi_i} = e^{\beta_0}$

β_0 : "initial" log-odds ratio when "$x=0$".

Ex: Suppose one has a contingency table.

Response	$x_1 =$ [illegible]	$x_2 = 1$	$x_3 = 1$	....	$x_{k+1} = 1$
$y=0$	n_{01}	n_{02}	n_{03}	---	$n_{0\,k+1}$
$y=1$	n_{11}	n_{12}	n_{13}	---	$n_{1\,k+1}$
	$n_{01}+n_{11}=n_1$	n_2			n_{k+1}

$x_{ij} = 1$ if the i-th observation is in class j.
$j = 1, \cdots, k+1$

The model is.

$$E(y) = \pi = \frac{\exp(X^T\beta)}{1+\exp(X^T\beta)} = \frac{1}{1+\exp(-X^T\beta)}$$

$$X = (1, x_1, \cdots x_K)$$

Then. $\log\frac{\pi}{1-\pi} = X^T\beta = \beta_0 + \beta_1 x_1 + \cdots \beta_K x_K$

$$\log\left(\frac{\pi}{1-\pi}\right)\Big|_{x=(1,0,\cdots 0)} = \beta_0$$

$$\log\left(\frac{\pi}{1-\pi}\right)\Big|_{x=(1,1,\cdots 0)} = \beta_0 + \beta_1$$

So $\beta_1 = \log\frac{\pi}{1-\pi}\Big|_{x=(1,1,\cdots 0)} - \log\frac{\pi}{1-\pi}\Big|_{x=(1,0\cdots 0)}$

$$= \log\left\{\frac{P(y=1 \mid x_1=1)}{P(y=0 \mid x_1=1)}\right\} - \log\left\{\frac{P(y=1 \mid x_{k+1}=1)}{P(y=0 \mid x_{k+1}=1)}\right\}$$

$$= \log\left\{\frac{P(y=1 \mid x_1=1)}{P(y=0 \mid x_1=1)} \cdot \frac{P(y=0 \mid x_{k+1}=1)}{P(y=1 \mid x_{k+1}=1)}\right\}$$

$$\hat{\beta}_1 = \log\left\{\frac{\hat{P}(y=1 \mid x_1=1)}{\hat{P}(y=0 \mid x_1=1)} \cdot \frac{\hat{P}(y=0 \mid x_{k+1}=1)}{\hat{P}(y=1 \mid x_{k+1}=1)}\right\}$$

$$= \log \left\{ \frac{n_{11}}{n_{01}} \cdot \frac{n_{0\,k+1}}{n_{1\,k+1}} \right\}$$

Likewise $\hat{\beta}_j = \log \left(\frac{n_{1j}}{n_{0j}} \cdot \frac{n_{0\,k+1}}{n_{1\,k+1}} \right)$, $j = 2, 3, \cdots k$

log of odds ratio.

Goodness of fit test.

Pearson's goodness of fit test

$$y_{\cdot j} = \sum_{i=1}^{n_j} y_{ij} \sim \text{Binomial}$$

$j = 1, \cdots J$

$n_1 \cdots n_J$

observed counts, $O_{1j} = y_{\cdot j}$ — success

$O_{0j} = n_j - y_{\cdot j}$ — failure

$j = 1, \cdots J$

expected counts if the model is good.

$$E_{1j} = \widehat{E\,y_{\cdot j}} = n_j \hat{\pi}_j = \frac{n_j}{1 + e^{-x_j\beta}}$$

$$E_{0j} = \widehat{E(n_j - y_{\cdot j})} = n_j(1-\hat{\pi}_j)$$

$$= \frac{n_j}{1+e^{x_j\beta}}$$

$$\chi^2 = \sum_{j=1}^{J} \sum_{k=0}^{1} \frac{(O_{kj} - E_{kj})^2}{E_{kj}} \approx \chi^2_{df = J-P}$$

J : # of classes

P : # of parameter

- Reject for large χ^2.