multiple regression model:
① Estimation
② Prediction
③ Model selection → need transformation
④ Diagnostics

9 Diagnostics and model building

9.1 Model validation and diagnostics

确认

The model is

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

$$\hat{\underset{\sim}{\epsilon}} = (\underset{\sim}{I} - \underset{\sim}{H})\underset{\sim}{y} = (\underset{\sim}{I} - \underset{\sim}{H})(\underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\epsilon})$$
$$= (\underset{\sim}{X}\underset{\sim}{\beta} - \underset{\sim}{H}\underset{\sim}{X}\underset{\sim}{\beta}) + (\underset{\sim}{I} - \underset{\sim}{H})\underset{\sim}{\epsilon}$$
$$(\underset{\sim}{H}\underset{\sim}{X} = \underset{\sim}{X}) \quad = (\underset{\sim}{X}\underset{\sim}{\beta} - \underset{\sim}{X}\underset{\sim}{\beta}) + (\underset{\sim}{I} - \underset{\sim}{H})\underset{\sim}{\epsilon}$$
$$= (\underset{\sim}{I} - \underset{\sim}{H})\underset{\sim}{\epsilon}$$

1. Residuals

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}} = \mathbf{y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}$$

where $\mathbf{x}$ is $n \times (k+1)$ and the hat matrix (projection matrix) is

$$\mathbf{H} = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'$$

$\underset{\sim}{X}$

In terms of the elements h_{ij} of $\mathbf{H}$, $n \times n$

$$\hat{\epsilon}_i = \epsilon_i - \sum_{j=1}^{n} h_{ij}\epsilon_j, \qquad i = 1, 2, ..., n.$$

Properties

proof by yourself

(a) $E(\hat{\boldsymbol{\epsilon}}) = \mathbf{0}$ (residual mean is the same as the error mean)

(b) $\text{Cov}(\hat{\boldsymbol{\epsilon}}) = \sigma^2(\mathbf{I} - \mathbf{H})$ (residuals are NOT independent)

(c) $\text{Cov}(\hat{\boldsymbol{\epsilon}}, \mathbf{y}) = \sigma^2(\mathbf{I} - \mathbf{H})$ (residuals correlated with the observations)

(d) $\text{Cov}(\hat{\boldsymbol{\epsilon}}, \hat{\mathbf{Y}}) = \mathbf{0}$ (residuals uncorrelated with the predicted values)

(e) $\bar{\hat{\epsilon}} = \sum_{i=1}^{n} \hat{\epsilon}_i / n = \hat{\boldsymbol{\epsilon}}'\mathbf{1}/n = 0$

(f) $\hat{\boldsymbol{\epsilon}}'\mathbf{y} = SSE$

(g) $\hat{\boldsymbol{\epsilon}}'\hat{\mathbf{Y}} = 0$

(h) $\hat{\boldsymbol{\epsilon}}'\mathbf{X} = \mathbf{0}'$ ($\hat{\boldsymbol{\epsilon}}'$ is orthogonal to each column of $\mathbf{X}$)

Properties:

(a) $E(\hat{\underset{\sim}{\varepsilon}}) = E[(\underset{\sim}{I} - \underset{\sim}{H})\underset{\sim}{\varepsilon}] = 0$

$(\underset{\sim}{I} - \underset{\sim}{H})$ is symmetric & idempotent.

(b) $Var(\hat{\underset{\sim}{\varepsilon}}) = Var[(\underset{\sim}{I} - \underset{\sim}{H})\underset{\sim}{\varepsilon}] = (\underset{\sim}{I} - \underset{\sim}{H})\sigma^2 I (\underset{\sim}{I} - \underset{\sim}{H})' = (\underset{\sim}{I} - \underset{\sim}{H})\sigma^2$

(c).

(d)

(e)

(f). $\hat{\underset{\sim}{\varepsilon}}\underset{\sim}{Y} = (\underset{\sim}{Y} - \underset{\sim}{X}\hat{\underset{\sim}{\beta}})'\underset{\sim}{Y} = \underset{\sim}{Y}'\underset{\sim}{Y} - \hat{\underset{\sim}{\beta}}'\underset{\sim}{X}'\underset{\sim}{Y} = SSE$

(g)

(h)

2. Model in centered form

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \\ &= \alpha + \beta_1 (x_{i1} - \bar{x}_1) + \beta_2 (x_{i2} - \bar{x}_2) + \cdots + \beta_k (x_{ik} - \bar{x}_k) + \epsilon_i \end{aligned}$$

In matrix form, the centered model is

$$\boldsymbol{y} = (\mathbf{1}, \boldsymbol{X}_c) \begin{pmatrix} \alpha \\ \boldsymbol{\beta}_1 \end{pmatrix} + \boldsymbol{\epsilon},$$

$$X_c = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots \\ \vdots & \vdots & \\ x_{n1} - \bar{x}_1 & x_{2n} - \bar{x}_2 & \cdots \end{pmatrix}$$

where

$$\boldsymbol{\beta}_1 = (\beta_1, \beta_2, ..., \beta_k)',$$

and

$$\boldsymbol{X}_c = \left(\boldsymbol{I} - \frac{1}{n} \boldsymbol{J} \right) \boldsymbol{X}_1$$

It can be shown from previous notes that the least squares estimators of the parameters are

$$\hat{\alpha} = \bar{y}$$

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_c' \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c' \boldsymbol{y}$$

Hence the predicted value is

$$\hat{\boldsymbol{y}} = \bar{y}\mathbf{1} + \boldsymbol{X}_c (\boldsymbol{X}_c' \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c' \boldsymbol{y} = \left(\frac{1}{n} \mathbf{1}' \boldsymbol{y} \right) \mathbf{1} + \boldsymbol{H}_c \boldsymbol{y} = \left(\frac{1}{n} \boldsymbol{J} + \boldsymbol{H}_c \right) \boldsymbol{y}$$

Since $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$, hence

nonnegative → elements in diagonal ≥ 0

$$\boldsymbol{H} = \frac{1}{n} \boldsymbol{J} + \boldsymbol{H}_c = \frac{1}{n} \boldsymbol{J} + \boldsymbol{X}_c (\underline{\boldsymbol{X}_c' \boldsymbol{X}_c})^{-1} \boldsymbol{X}_c' \quad (*)$$

positive definite

$\underset{\sim}{H} = (h_{ij}) \quad i, j = 1, \cdots, n.$

$h_{ii} \geq \frac{1}{n}$

$H = (\underset{\sim}{h_1}, \underset{\sim}{h_2}, \cdots \underset{\sim}{h_n})$

$\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$

$\hat{y} = \underset{\sim}{X} \hat{\beta} = \underbrace{X(X'X)^{-1}X'}_{\underset{\sim}{H}} y = \underset{\sim}{H} \underset{\sim}{y}$

$\underset{\sim}{H} = \underset{\sim}{H}^2 \Rightarrow h_{ii} = \underbrace{\underset{\sim}{h}_i^T}_{i\text{th row of } \underset{\sim}{H}} \underset{\sim}{h}_i = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$

$\Rightarrow 1 = h_{ii} + \frac{\sum_{j \neq i} h_{ij}^2}{h_{ii}} \Rightarrow h_{ii} \leq 1$

$\Rightarrow \frac{1}{n} \leq h_{ii} \leq 1$

3. Hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \{h_{ij}\}$ (Let $\mathbf{x}$ be a matrix with full column rank and with $\mathbf{1}$ as its first column)

Properties

leverage

(a) $1/n \leq h_{ii} \leq 1$ for $i = 1, 2, ..., n$.

(b) $-.5 \leq h_{ij} \leq .5$ for all $j \neq i$.

(c) $h_{ii} = 1/n + (\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)'(\mathbf{x}_c'\mathbf{x}_c)^{-1}(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)$, where $\mathbf{x}_{1i}' = (x_{i1}, x_{i2}, ..., x_{ik})$, $\bar{\mathbf{x}}_1' = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_k)$, and $(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)'$ is the ith row of the centered matrix $\mathbf{x}_c$.

(d) $\boxed{\text{tr}(\mathbf{H}) = \Sigma_{i=1}^{n} h_{ii} = k + 1.}$

(d). $\text{tr}(\underset{\sim}{H}) = \text{tr}(\underset{\sim}{X}) = k+1$

$h_{ii} = \sum_{r=1}^{n} h_{ir}^2$

$h_{ii} = h_{ii}^2 + h_{ij}^2 + \sum_{r \neq i,j} h_{ir}^2$

$\Rightarrow h_{ii} - h_{ii}^2 = h_{ij}^2 + \sum_{r \neq i,j} h_{ir}^2 \geq h_{ij}^2$

$\Rightarrow h_{ij}^2 \leq h_{ii} - h_{ii}^2 = \frac{1}{4} - (h_{ii} - \frac{1}{2})^2 \leq \frac{1}{4}$

i.e. $-\frac{1}{2} \leq h_{ij} \leq \frac{1}{2}$

Prediction: $\hat{\underset{\sim}{Y}} = \underset{\sim}{H}\underset{\sim}{Y}$

$$\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} h_{11} & \cdots & h_{1n} \\ h_{21} & \cdots & h_{2n} \\ \vdots & \ddots & \vdots \\ h_{n1} & \cdots & h_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}. \quad i = 1, \cdots, n$$

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

or $\varepsilon_i \sim N(0, \sigma^2)$

(i). (ii) are two ways to scale the residuals $\Rightarrow$ same variance

4. Outliers

   (a) Variance of the residuals is not constant $Cov(\hat{\varepsilon}) = \sigma^2(I - H) \Rightarrow Var(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$

      i. Studentized residual

      ii. Studentized deleted residual (externally studentized residual)

   (b) Deleted residuals

   (c) Press (prediction sum of squares)

$\underset{\sim}{\varepsilon} = \underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta}$, residual $\hat{\underset{\sim}{\varepsilon}} = \underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\beta}} = (\underset{\sim}{I} - \underset{\sim}{H})\underset{\sim}{y} = (\underset{\sim}{I} - \underset{\sim}{H})\underset{\sim}{\varepsilon}$

$Var(\hat{\varepsilon}) = (I - H)\sigma^2$

$Var(\hat{\varepsilon}_i) = Var(y_i - \hat{y}_i) = \sigma^2(1 - h_{ii})$

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} = \frac{\hat{\varepsilon}_i}{S\sqrt{1-h_{ii}}}, \quad \text{where } S^2 = \hat{\sigma}^2 = \frac{SSE}{n-k-1}$$

$$t_i = \frac{\hat{\varepsilon}_i}{S_{(i)}\sqrt{1-h_{ii}}}, \quad S_{(i)}^2 = \frac{SSE_{(i)}}{(n-1)-k-1} = \frac{\sum_{j \neq i}(y_j - \hat{y}_j)^2}{n-k-2}$$

(当一个女孩被分到男孩组中时，会被分离的更远一些)

(b) deleted residuals:

$\hat{\varepsilon}_{(i)} = y_i - \hat{y}_{(i)} = y_i - \underset{\sim}{x}_i' \hat{\underset{\sim}{\beta}}_{(i)}$ estimate of β using data set $\{y_i, \underset{\sim}{x}_j, j \neq i, i = 1, \cdots, n\}$

(c) Press:

$$= \sum_{i=1}^{n} \hat{\varepsilon}_{(i)}^2 = \sum_{i=1}^{n} \left(\frac{\hat{\varepsilon}_i}{1-h_{ii}}\right)^2$$

Cross-Validation

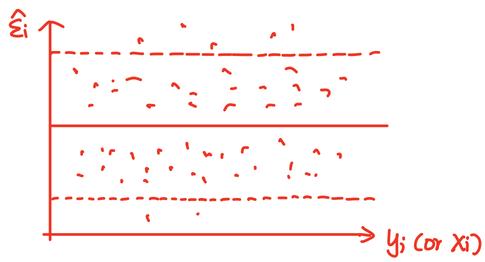
leave-one-out CV

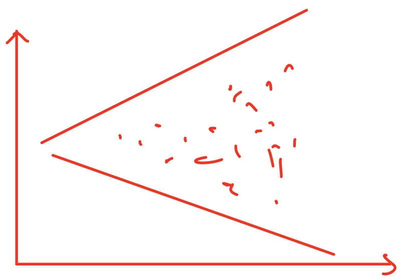
$$CV = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_{(i)}^2 \rightarrow MSE = E[(y_i - \hat{y}_i)^2]$$

$$CV = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_{(i)}^2 \longrightarrow MSE = E(y_i - \hat{y}_i)^2$$

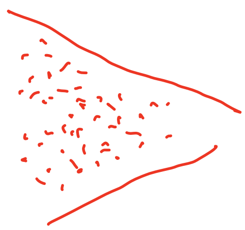
Remark 9.1:

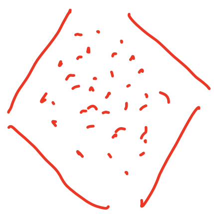
(i) check constant variance:



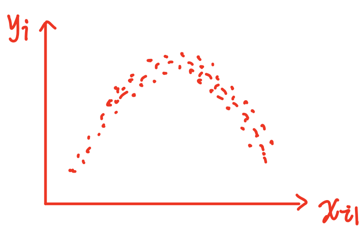


$Var(r_i) \propto \hat{y}_i$





(ii)



(iii) Q-Q plot p-value.

Residual analysis.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$\varepsilon_i = y_i - f(x_{i1}, \cdots, x_{ik}, \theta)$$

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - f(x_{i1}, \cdots, x_{ik}, \hat{\theta}) \qquad \theta = (\beta_0, \cdots, \beta_k, \sigma^2)$$

Linear model:

(i) constant variance $Var(y_i) = Var(\varepsilon_i) = \sigma^2$

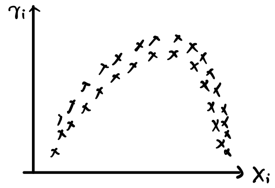
(ii) linearity

(iii) Normality

Residual plots.

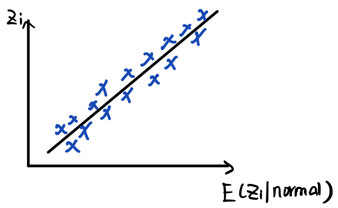
(i).

(ii)



not linear → maybe need x_i^2 or ...

(iii) Q-Q plot. $|\varepsilon_i|$ → ordered z_i



Outliers: make a large difference

5. Influential Observations

Guideline

it is an influential observation, if $D_i \geq F_{0.05,\ k+1,\ n-(k+1)}$

(a) Leverage h_{ii}

(b) Cook's distance

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(k+1)\hat{\sigma}^2}$$

(a) $\hat{\underset{\sim}{y}} = \underset{\sim}{H}\,\underset{\sim}{y}$ $\quad \hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j\neq i} h_{ij} y_j$

$1 = h_{ii} + \frac{\sum_{j\neq i} h_{ij}^2}{h_{ii}} \Rightarrow$ if $h_{ii} \to 1$, other h_{ij}^2 will be very small.

if h_{ii} is large $\Rightarrow$ the ith observation is a high leverage point.

Q: How to judge whether h_{ii} is large

$\frac{1}{n}\sum h_{ii} = \frac{k+1}{n}$

if $h_{ii} \geq 2\cdot\frac{k+1}{n} \to$ high leverage point guild line

Diffits $= \frac{\hat{Y}_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}\cdot\sqrt{h_{ii}}}$ where $\hat{y}_{(i)} = \underset{\sim}{x}_i'\,\hat{\underset{\sim}{\beta}}_{(i)}$

standardized deleted residual

(standardized difference of fitted value with or without using the ith observation

Guildline: the ith observation is influential if $|\text{diffits}_i| > 2\sqrt{\frac{k+1}{n}}$

$S_{(i)}^2 = \frac{SSE(i)}{n-k-2}$, $SSE_{(i)} = \underset{\sim}{y}_{(i)}'\underset{\sim}{y}_{(i)} - \hat{\underset{\sim}{\beta}}_{(i)}'\underset{\sim}{X}_{(i)}'y_{(i)}$

or $= SSE - \hat{\varepsilon}_i^2/1-h_{ii}$