9 Diagnostics and model building

9.1 Model validation and diagnostics

The model is

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

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

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

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

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

Properties

(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 $\mathbf{X}$)

prove by yourself !

2. Model in centered form

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

In matrix form, the centered model is

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

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

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

positive definite

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

$h_{ii} \geq 1/n$

$H = H' \Rightarrow h_{ii} = h_i' h_i = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$

h_i: ith row of H

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

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

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

$\hat{Y} = X\hat{\beta}$

$= X(X'X)^{-1}X'Y$

$= HY$

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

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

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

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

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

Residuals

4. Outliers

(a) Variance of the residuals is not constant

i. Studentized residual

ii. ~~Studentized deleted residual~~ (externally studentized residual)

(b) Deleted residuals

(c) Press (prediction sum of squares)

$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$

$i = 1, \cdots, n.$

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

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

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

$= (I - H)\varepsilon$

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

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

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

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

(b) deleted residuals

$\hat{\varepsilon}_{(i)} = Y_i - \hat{Y}_{(i)} = y_i - x_i' \hat{\beta}_{(i)}$

$\hat{\beta}_{(i)}$ ← estimate of β using data set $\{y_j, x_j, j \neq i, i = 1, \cdots n\}$

(c) PRESS (prediction sum of squares)

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

Remark 9.1

CROSS - Validation

leave-one-out CV

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

if it is an influential observation, if $D_i > F_{.5, k+1, n-(k+1)}$ ← This is correct!

5. Influential Observations

(a) Leverage h_{ii}

(b) Cook's distance

outliers → make a large difference of $\hat{\beta}$ with or without using the observation / $\hat{\beta}_{(i)}$

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

(a) $\hat{y} = H y$, $\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$

$= h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$

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

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

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

if $h_{ii} \geq 2 \cdot \frac{k+1}{n}$ → high leverage point guideline!

Dffits $= \frac{\hat{y}_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)} \sqrt{h_{ii}}}$

$\hat{y}_{(i)} = x_i' \hat{\beta}_{(i)}$

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

guideline

the i-th observation is influential if

$|\text{dffits}_i| \geq 2 \sqrt{\frac{k+1}{68 n}}$

9.2 More complex models

9.3 Multicollinearity

9.4 Variable selection