# Regressional Diagnosis

- We have postulate several conditions:

i) $E(Y)$ is L.F of $\{X_i\}_1^p$ (linearity)

ii) $\varepsilon_i$ are indept. (independence)

iii) $E(\varepsilon_i)=0$, $\sigma^2(\varepsilon_i)=\sigma^2$. (homogeneity)

iv) $\varepsilon_i$ is normal dist. $1\le i\le n$. (normality)

$\Rightarrow$ Whether these hypothesis are reasonable or not. If not, how can we recorrect it?

(1) Residual and Plot:

① $\hat{e} = Y - X\hat{\beta} = (I-H)Y$. $H = X(X^TX)^{-1}X^T$.

So $\hat{e}$ is connected with $H =: (h_{ij})_{n\times n}$

prop. i) $h_{ii} \in [0,1]$. Besides, $h_{ii}=1 \Rightarrow h_{ij}=0, j\neq i$

ii) $\sum_1^n h_{ii} = p+1$ ($p$ is number of variables)

Pf: i) By $H^2=H \Rightarrow h_{ii} = \sum h_{ij}^2 \geq h_{ii}^2$.

ii) $\sum h_{ii} = tr(H) = tr((X^TX)^{-1}(X^TX))$

Def: $r_j = \hat{e}_j / \sqrt{1-h_{ii}}\,S$. is normalized residual, where

$S = \sqrt{SSE/(n-p-1)}$. $\hat{e} = (\hat{e}_1 \cdots \hat{e}_n)$

Rmk: i) $r_i^2/(n-p-1) \sim B(\frac{1}{2}, \frac{n-p-2}{2})$

ii) $E(r_i) = 0$. $Cov(r_i, r_j) = -\dfrac{h_{ij}}{\sqrt{(1-h_{ii})}\sqrt{(1-h_{jj})}}$

iii) Generally, $\hat{\varepsilon}_i$ isn't indept with $s^2$.

But if the hypothesis hold, then $\{r_i\}$ $\overset{iid}{\sim} N(0,1)$, approximately.

① Plot:

We choose $r_i$ as $y$-axis of plot

And $x_i$ as $x$-axis of plot (or $\hat{y}$)

Rmk: If the model is reasonable. Then the positions of data points is random.

since $r_i \sim N(0,1)$. i.i.d.

If there's some rule of dist. of points

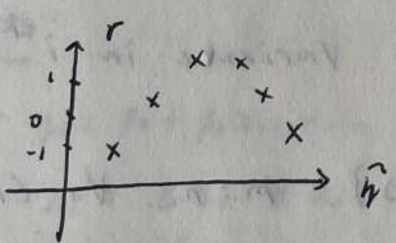
Then we can suspect the model fails.

(2) Model Diagnosis:

① linearity:

i) Criteria: $y$ is L.F. of $\{X_i\}_1^p$.

e.g.



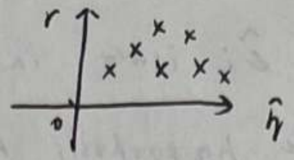

the dist. of points have a rule of quadratic function.

$\Rightarrow$ So we will suspect it's wrong.

ii) Recorrection:

Consider: $y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$

$\Rightarrow$ We have plot :




② Homogeneity :

If $\begin{cases} y_{ij} = \beta_0 + \beta_1 x_{ij1} \cdots + \beta_p x_{ijp} + \varepsilon_{ij} . & 1 \le i \le k . \ 1 \le j \le n . \\ \varepsilon_{ij} \sim N(0, \sigma_i^2) . \text{ indept.} \end{cases}$

Test : $H_0 : \sigma_1^2 = \sigma_2^2 \cdots = \sigma_k^2$ .

i) Diagnose by Plot :

If it's not homogeneius. Then plot $r - \hat{y}$ have a rule ( some trend )

ii) Diagnose by trivials :

If we can repeat the experiments. Then we can use the following three tests.

a) Hartley Test :

$$F = \frac{\max\limits_{1 \le i \le k} S_i^2}{\min\limits_{1 \le i \le k} S_i^2} . \quad \text{where } S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2$$

$\hat{y}_{ij} = x_{ij}^T \hat{\beta}$ . ( $S_i^2$ is variance in $i^{th}$ group )

$R = \{ F > F_{1-\alpha}(k, m-1) \}$ . $m = n_i$ . $\forall i$ . ( $m > 3$ ) .

b) Cochran Test :

$$G = \frac{\max\limits_{1 \le i \le k} S_i^2}{\sum_1^n S_i^2} \qquad R = \{ G > G_{(1-\alpha)}(k, m-1) \} .$$

$m = n_i$ . $\forall i$ . $m > 3$ .

c) Barlett Test:

$$\chi^2 = \frac{1}{c}\left[ f_e \ln s^2 - \sum_{i=1}^{k} (n_i-1) s_i^2 \right] \overset{H_0}{\sim} \chi^2(k-1).$$

where $f_e = \Sigma(n_i-1)$. $s^2 = \Sigma(n_i-1)s_i^2 / f_e$. $c = \frac{\Sigma \frac{1}{n_i-1} - \frac{1}{f_e}}{3(n-1)} + 1$

$R = \{ \chi^2 > \chi^2_{1-\alpha}(k-1) \}$.

iii) correction:

a) By generalised LSE:

Directly consider $\varepsilon \sim N(0, \sigma^2 \operatorname{diag}\{\sigma_1^2/\sigma^2 \cdots \sigma_k^2/\sigma^2\})$

b) By Transformation:

If $E(Y) = \mu = \mu(x_1 \cdots x_p)$. $Var(Y) = g(\mu(x_1 \cdots x_p))$

Set $Z = f(Y)$. Find $f$. st. $Var(Z) \simeq$ const.

By Talyor Expansion: $Z \simeq f(\mu) + f'(\mu)(Y-\mu)$

$\Rightarrow Var(Z) = f'(\mu)^2 g(\mu) = c$.

Solve $f(y) = \int \sqrt{c/g(\mu)}\, d\mu$. Consider $Z$.

③ Independence:

e.g. $\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_i x_{it} + \varepsilon_i. & 1 \le i \le n. \\ \varepsilon_i = \rho \varepsilon_{i-1} + u_i \\ u_i \overset{i.i.d}{\sim} N(0,1) \end{cases}$ (one-order case)

$\Rightarrow$ Test: $H_0: \rho = 0$ v.s. $H_1: \rho \neq 0$.

(By Time-Series analysis)

④ Normality:

i) Diagnosis:

a) Since if $Y \sim N(X\beta, \sigma^2 I_n)$. Then $r_i \sim N(0,1)$.

Calculate the ratio of datas falled into $(-1,1)$, $(-2,2)$, $(-3,3)$.

b) Pearson - $\chi^2$ test.

ii) Correction:

We want to find a transformation on $Y$.

s.t. $\begin{pmatrix} g(y_1) \\ \vdots \\ g(y_n) \end{pmatrix} \sim N_n(X\beta, \sigma^2 I_n)$

Def: Box - Cox Transform: $y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \ln y, & \lambda = 0. \end{cases}$

$\Rightarrow$ Find $\lambda$, s.t. It's most likely normal dist.

$\Rightarrow$ Consider MLE $L(\beta, \sigma^2, \lambda) = |J| \cdot (2\pi\sigma^2)^{-\frac{n}{2}} \operatorname{Exp}(-\frac{1}{2\sigma^2}(Y^{(\lambda)} - X\beta)^T(Y^{(\lambda)} - X\beta))$

where $|J| = \prod |y_i|^{\lambda - 1}$.

$\Rightarrow \begin{cases} \hat{\beta}_\lambda = (X^T X)^{-1} X^T Y^{(\lambda)} \\ \hat{\sigma}_\lambda^2 = \frac{1}{n} Y^{(\lambda)T}(I - M) Y^{(\lambda)} \end{cases}$

$\Rightarrow$ Solve $\max\limits_{\lambda} L(\lambda, \hat{\beta}_\lambda, \hat{\sigma}_\lambda^2) = \max |J| (2\pi e \hat{\sigma}_\lambda^2)^{-\frac{n}{2}}$

By numerical method, find $\hat{\lambda}$.

(3) Data Diagnosis:

① Outlier:

It means that a data point departuring the model a lot

Consider $\begin{cases} y_i = x_i^T\beta + \varepsilon_i & i \neq j. \\ y_j = x_j^T\beta + \eta + \varepsilon_j. \end{cases}$ $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$

$\Rightarrow$ i.e. $Y = (X \;\; e_j(n)) \begin{pmatrix} \beta \\ \eta \end{pmatrix} + \varepsilon.$

$H_0: \eta = 0$ v.s. $H_1: \eta \neq 0.$

$$F = \frac{Y^T(M - M_0)Y/1}{Y^T(I - M)Y/(n-p-1)} = \frac{(n-p-1)r_j^2}{n-p-r_j^2} \sim F(1, n-p-1)$$

$R = \{ F > F_{1-\alpha}(1, n-p-1) \}.$

② Influential Observation:

It means the data point influences the estimate statistics a lot.

Rmk: i) It can provide more information than other points.

ii) It may be an outlier or may not be.

③ Analysis:

i) Cook Distance:

Denote: $IF_i = \hat{\beta}_{(-i)} - \hat{\beta}.$ $\hat{\beta}_{(-i)} = (X_{(-i)}^T X_{(-i)})^{-1} X_{(-i)}^T Y.$

Def: $D_i(M,c)$ is cook distance between $\hat{\beta}(-i)$ and $\hat{\beta}$ : $D_i(M,c) = (\hat{\beta}(-i) - \hat{\beta})^T M (\hat{\beta}(-i) - \hat{\beta}) / c$

Rmk: If $D_i(M,c) \uparrow$. Then it means $i^{th}$ group of data is influential.

prop. $D_i(M,c) = r_i^2 \cdot \frac{S^2}{c} \cdot P_i(M)$. $P_i(M) = \frac{x_i^T (X^T X)^{-1} M (X^T X)^{-1} x_i}{1 - h_{ii}}$

Rmk: i) $r_i^2$ is the evaluation of fitting degree.

ii) $P_i(M)$ describes the position of data point $\vec{x_i}$.

iii) When $D_i(M,c)$ is large $\Rightarrow$ $x_i$ is influential observation (Departure a lot)

When $h_{ii} \approx 1$. $x_i$ is high leverage point.

When $r_i$ is large. $x_i$ is outlier.

ii) AP - Statistics:

Denote: $Z = (X\ y)$. $Z(-I)$ is matrix depleted the rows with index in $I$.

Def: AP - statistics: $R_I = |Z^T(-I) Z(-I)| / |Z^T Z|$.

Rmk: $R_I$ is smaller $\Rightarrow$ It's more likely the data depleted is influential.