$C_p = 2p' - n + \frac{\text{Res S.S.}/p'}{\hat{\sigma}^2_{\text{full model}}}$

1. Dec

3. (15 marks) An experiment was conducted to model *Y* with five explanatory variables X_1, X_2, X_3, X_4 and X_5. We desire an equation of relating *Y* to the other variables. The goal is to find variables that should be further studied with the eventual goal of developing a prediction equation. The following table gives RSS for all possible regressions. Total sum of squares is equal to 5.0634 and the number of observations is equal to 20.

No. of parameters in the model	*RSS*	Model
2	2.0338	X_1
2	5.0219	X_2
2	$\boxed{1.5370}$	X_3
2	2.5044	X_4
2	1.5563	X_5
3	1.5921	X_1, X_2
3	1.4397	X_1, X_3
3	1.7462	X_1, X_4
3	1.4963	X_1, X_5
3	1.4707	X_2, X_3
3	2.4381	X_2, X_4
3	1.4388	X_2, X_5
3	1.4590	X_3, X_4
3	$\boxed{1.0850}$	X_3, X_5
3	1.3287	X_4, X_5
4	1.2582	X_1, X_2, X_3
4	1.4257	X_1, X_2, X_4
4	1.2764	X_1, X_2, X_5
4	1.3894	X_1, X_3, X_4
4	1.0644	X_1, X_3, X_5
4	1.3204	X_1, X_4, X_5
4	1.3900	X_2, X_3, X_4
4	$\boxed{0.9871}$	X_2, X_3, X_5
4	1.2178	X_2, X_4, X_5
4	1.0634	X_3, X_4, X_5
5	1.2199	X_1, X_2, X_3, X_4
5	0.9871	X_1, X_2, X_3, X_5
5	1.1565	X_1, X_2, X_4, X_5
5	1.0388	X_1, X_3, X_4, X_5
5	$\boxed{0.9653}$	X_2, X_3, X_4, X_5
6	0.9652	X_1, X_2, X_3, X_4, X_5

$\Rightarrow C_p = 2*2 - 20 + \frac{1.5370}{0.9652/14} = 6.294$

$\Rightarrow C_p = 2*3 - 20 + \frac{1.085}{0.9652/14} = \boxed{1.738}$

Best model = X_3, X_5

$\Rightarrow C_p = 2*4 - 20 + \frac{0.9871}{0.9652/14} = 2.318$

$\Rightarrow C_p = 2*5 - 20 + \frac{0.9653}{0.9652/14} = 4.001$

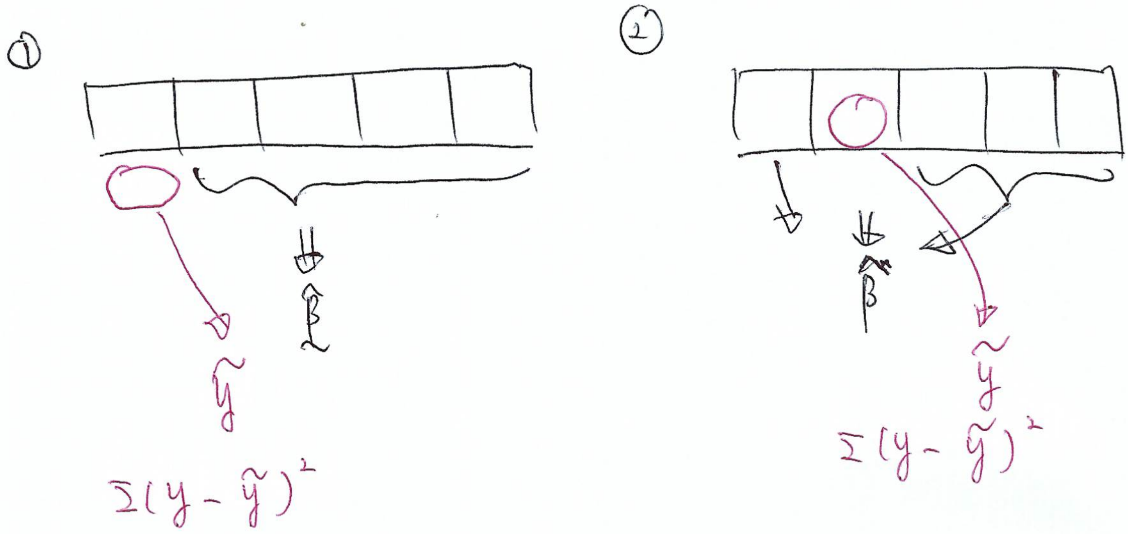
$\Rightarrow C_p = 2p' - n + \frac{0.9652}{0.9652/(n-p')} = p' = 6$

Find the best model by C_p, forward selection, backward selection and stepwise selection. Write down how to get the best model on details. Choose critical values for both ENTRY and STAY to be 2. Comment the results.

$\hat{\sigma}^2_{\text{full model}} = \frac{0.9652}{20-6} = \frac{0.9652}{14} = 0.06894$

1

(6) Cross - validation



Calculate $\Sigma\,\Sigma(y-\tilde{y})^2$ all obs. without ith obs.

CVPRESS

PRESS statistic

Predicted

$n_1 = n-1 \rightarrow$ fit a model

$\rightarrow \hat{\beta}$

$n_2 = 1 \rightarrow \tilde{y}_{i(-i)} \qquad i = 1, \ldots, n$

ith obs

$$\underbrace{\left(y_i - \tilde{y}_{i(-i)}\right)}_{\text{PRESS residual}}$$

$$\text{PRESS} = \sum_{i=1}^{n} \left(y_i - \tilde{y}_{i(-i)}\right)^2$$

$$= \sum_{i=1}^{n} \frac{\hat{e}_i^2}{(1-h_{ii})^2}$$

$\hat{e}_i = y_i - \hat{y}_i$ — residual

$h_{ii} = (i,i)^{th}$ element in $\underset{\sim}{H}$

where $\underset{\sim}{H} = \underset{\sim}{X}(\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{X}^T$

$h_{ii} = \underset{\sim}{x}_i^T (\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{x}_i$

leverage

$\underset{\sim}{x}_i^T$ — ith row in $\underset{\sim}{X}$

Simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad i = 1, \ldots, n$$

Centred model

$$\Rightarrow \underset{\sim}{X}^T \underset{\sim}{X} = \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix} \qquad (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{S_{xx}} \end{pmatrix}$$

$$h_{ii} = (1 \quad x_i - \bar{x}) \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{S_{xx}} \end{pmatrix} \begin{pmatrix} 1 \\ x_i - \bar{x} \end{pmatrix}$$

$$= \frac{1}{n} + \frac{\boxed{(x_i - \bar{x})^2}}{S_{xx}} \longleftarrow \text{diff. between } x_i \text{ \& } \bar{x}$$

$$\boxed{1} \geq h_{ii} \geq \boxed{\frac{1}{n}} \longleftarrow \underset{\sim}{H} - \frac{1}{n}\underset{\sim}{J} \text{ is positive semi definite}$$

$\underset{\sim}{I} - \underset{\sim}{H}$ is positive semi definite (Chapter 1)

$\Rightarrow$ all ~~diagon~~ diagonal elements are non-negative

$\Rightarrow 1 - h_{ii} \geq 0$

$\Rightarrow 1 \geq h_{ii}$

Define $\underset{\sim}{H_c} = \underset{\sim}{X_c} (\underset{\sim}{X_c}^T \underset{\sim}{X_c})^{-1} \underset{\sim}{X_c}^T$

Model $= \underset{\sim}{y} = \alpha \underset{\sim}{1} + \underset{\sim}{X_c} \underset{\sim}{\beta} + \underset{\sim}{e}$

$$\hat{\underset{\sim}{y}} = \hat{\alpha} \underset{\sim}{1} + \underset{\sim}{X_c} \hat{\underset{\sim}{\beta}}$$

$$= \hat{\alpha} \underset{\sim}{1} + \underset{\sim}{X_c} (\underset{\sim}{X_c}^T \underset{\sim}{X_c})^{-1} \underset{\sim}{X_c}^T \underset{\sim}{Y}$$

$$= [\tfrac{1}{n} \underset{\sim}{J} + \underset{\sim}{H_c}] \underset{\sim}{Y}$$

matrix with ~~eq~~ all elements equal to 1

$$= \underset{\sim}{H} \underset{\sim}{Y}$$

$$\Rightarrow \underset{\sim}{H} = \frac{1}{n}\underset{\sim}{J} + \underset{\sim}{H_c}$$

$$\Rightarrow \underset{\sim}{H} - \frac{1}{n}\underset{\sim}{J} = \underset{\sim}{H_c}$$

positive semi definite

$\Rightarrow$ all diagonal elements of $\underset{\sim}{H} - \frac{1}{n}\underset{\sim}{J}$ are non-negative

$$\Rightarrow h_{ii} - \frac{1}{n} \geq 0$$

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

PRESS $= \sum_{i=1}^{n} \frac{\hat{e}_i^2}{(1-h_{ii})^2}$ — smallest PRESS

(c) largest $R^2 = 1 - \frac{\text{Res S.S.}}{\text{total S.S.}}$

$\Leftrightarrow$ smallest Res S.S.

$R^2 \uparrow$ when # of indep. variables $\uparrow$

$\Rightarrow$ Best model = full model

(d) largest $R^2_{adj} = 1 - \frac{MSE}{\text{total M.S.}}$ $\leftarrow$ $\text{Res S.S.}/(n-p')$, $\leftarrow$ $\text{total S.S.}/(n-1)$

$$= 1 - \frac{\text{Res S.S.}}{\text{total S.S.}} * \frac{n-1}{n-p'}$$

R^2_{adj} may be negative

(f) AIC = Akaike Information Criteria — smallest

$$= -2\log L + \underbrace{2p'}_{\text{penality}}$$

(g) BIC = Bayesian Information Criteria — smallest

$$= -2\log L + \underbrace{\ln n * p'}_{\text{penality}}$$

~~ln n > 2~~ $\log_e(n) > 2$

Linear Regression

$$L(\hat{\beta}, \hat{\sigma}) = \frac{1}{(2\pi)^{n/2} (\hat{\sigma}^2)^{n/2}} \exp\left\{ - \frac{\sum_{i=1}^{n} (y_i - \underset{\sim}{x}_i^T \hat{\beta})^2}{2\hat{\sigma}^2} \right\}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \underset{\sim}{x}_i^T \hat{\beta})^2}{n} \quad \text{m.l.e}$$

$$\Rightarrow \log L = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\hat{\sigma}^2) - \frac{n}{2}$$

$$\Rightarrow -2 \log L = \underline{n \ln(2\pi)} + n \ln(\hat{\sigma}^2) + \underline{n}$$

$$= n \ln\left(\frac{\text{Res.S.S.}}{n}\right) + \text{Constant}$$

$$\hat{\sigma}^2 = \frac{\text{Res.S.S.}}{n - p'} \text{ to est. } \sigma^2 \text{ (unbiased est.)}$$

$$\Rightarrow AIC = n \ln\left(\frac{\text{Res.S.S.}}{n}\right) + 2p'$$

$$BIC = n \ln\left(\frac{\text{Res.S.S.}}{n}\right) + \ln(n) * p'$$

Chapter 5 Residual Analysis

Assume – $e_i \sim N(0, \sigma^2)$

– y & $\underset{\sim}{X}$ are linear related

Residual $\hat{e}_i = y_i - \hat{y}_i$

$E(\hat{e}_i) = 0$

$Var(\hat{e}_i) = \sigma^2(1 - h_{ii})$ ← Chapter 1

$\neq$ constant

Standardized residual

① $r_i = \dfrac{\hat{e}_i}{\sqrt{\hat{\sigma}^2 (1 - h_{ii})}}$ — (internally studentized residual)

② $t_i = \dfrac{\hat{e}_i}{\boxed{\hat{\sigma}_{(-i)}} \sqrt{1 - h_{ii}}}$ — (externally studentized residual)

$\underset{\sim}{x}_1, \;----\; \underset{\sim}{x}_{i-1}, \boxed{\underset{\sim}{x}_i}, \underset{\sim}{x}_{i+1} \;-----,\; \underset{\sim}{x}_n$

$y_1, \;-----\; y_{i-1}, \boxed{y_i}, y_{i+1} \;----,\; y_n$

Ignore the ith obs

$\Rightarrow$ $(n-1)$ obs $\Rightarrow$ Fit a model of y on $\underset{\sim}{x}$

$\hat\beta_{0(-i)}, \hat\beta_{1(-i)}, \;---,\; \hat\beta_{p(-i)}, \hat\sigma^2_{(-i)}$

Put $\underset{\sim}{x}_i$ into the fitted model

$\Rightarrow \hat{y}_{i(-i)}$

$$\hat\sigma^2_{(-i)} = \frac{\text{Res S.S.}_{(-i)}}{\underbrace{(n-1)}_{\#\text{ of obs.}} - \underbrace{p'}_{\#\text{ of unknown para. in the model}}}$$

$t_i \sim t_{(n-1-p')}$?

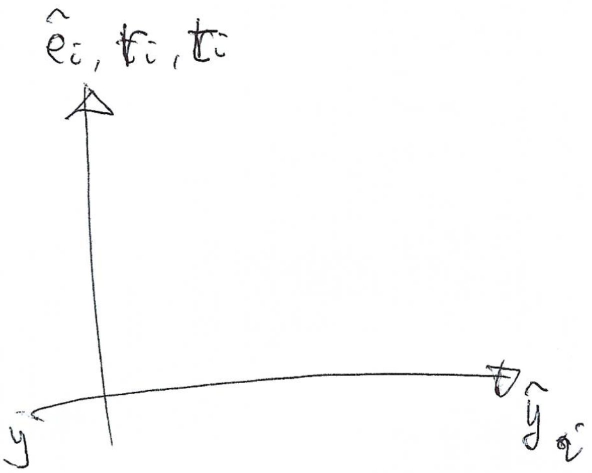
Use t_i to detect outlier

① Residual Plot

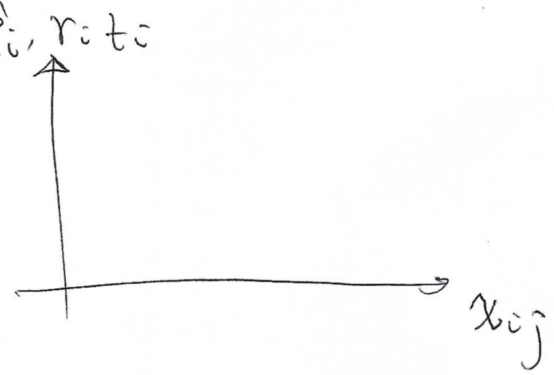
Check: linearity

constant variance

pattern $\Rightarrow$ transformation of y & $\underset{\sim}{x}$



* If there is pattern



$\cancel{\hat{e}_i}, r_i, t_i$

no pattern

$\Rightarrow$ Assumptions are valid

variance (variation of residuals) increases as $\hat{y}_i$ or x_i increases

non-linear $\longrightarrow$

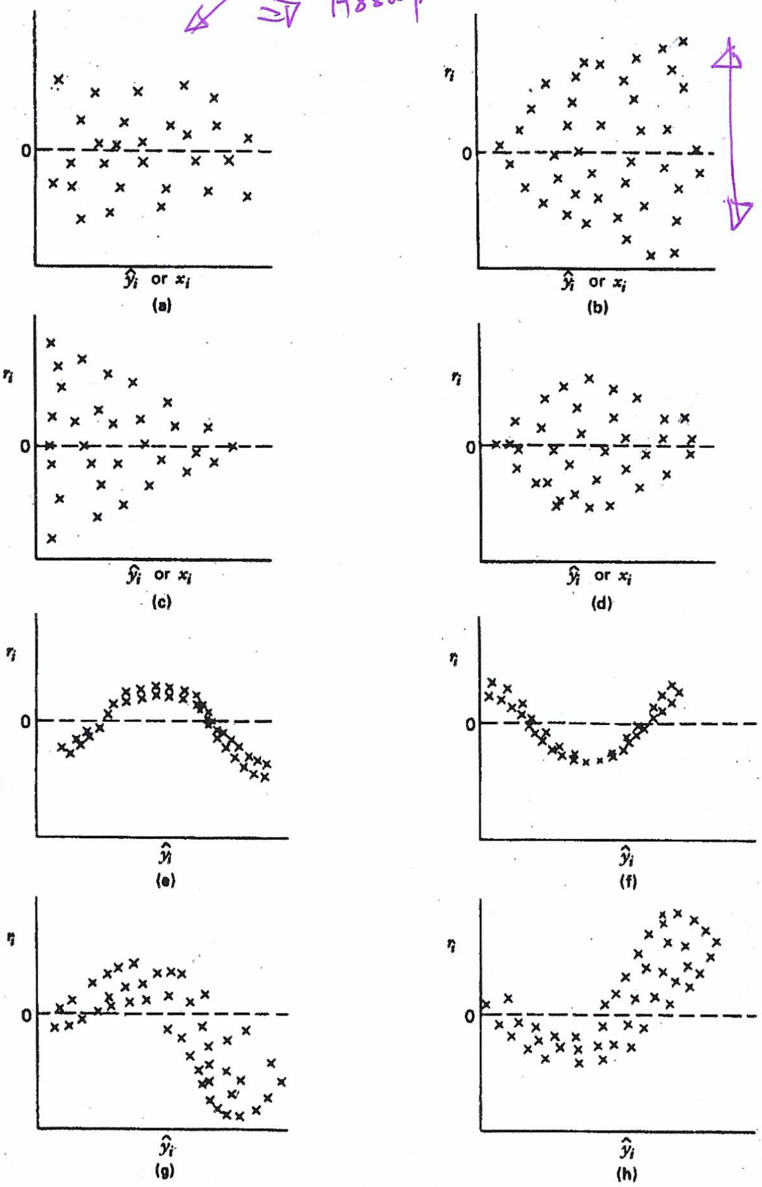


Figure 6.3 Residual plots: (*a*) null plot; (*b*) right-opening megaphone; (*c*) left-opening megaphone; (*d*) double outward bow; (*e*) nonlinearity; (*f*) nonlinearity; (*g*) nonlinearity and nonconstant variance; (*h*) nonlinearity and nonconstant variance.

$\Rightarrow$ transformation of y and/or x

e.g. Count data $\Rightarrow \sim$ Poisson(μ) $\quad E(Y) = \mu$

$Var(Y) = \mu$

$-\ \sqrt{y}\ ,\ \log(y)$

y − area $\quad m^2$

x − perimeter $\quad m$

$\sqrt{y}$	$\log(y)$	y
x	$\log(x)$	x^2

smallest $-2\log L$

largest likelihood

$\boxed{\text{smallest Res S.S}}$

↑

linear regression

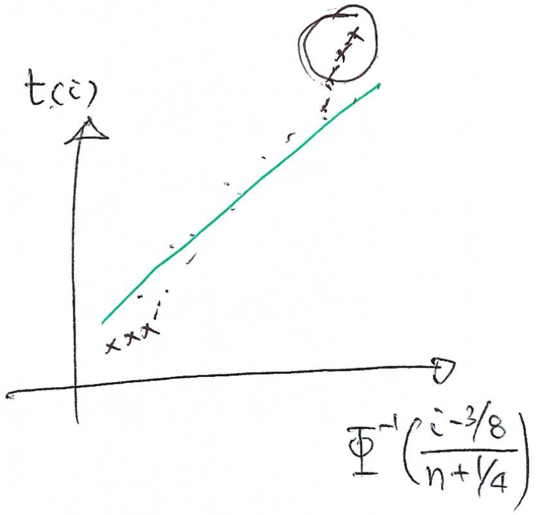
7

Normality assumption

Q-Q plot $\Rightarrow$ ~~normal~~ normality

Arrange $t_1, \ldots, t_n$

$\Rightarrow t_{(1)}, \ldots, t_{(n)}$



$t_{(i)}$

$\uparrow$

$$z_{\frac{i}{n}} = \Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right)$$

$\uparrow$

$\Phi(z)$ = c.d.f. of $Z = N(0,1)$

Transformation on y

Box-Cox transformation (Power ~~t~~ transformation)

$$y_i^* = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log_e(y_i) & \lambda = 0 \end{cases}$$

λ - unknown

$\uparrow$

Find m.l.e. of λ

Assume $\underset{\sim}{Y}^* \sim N(\underset{\sim}{X}\underset{\sim}{\beta}, \sigma^2 \underset{\sim}{I})$

- Data: $\underset{\sim}{Y}, \underset{\sim}{X}$

— dist. of Y is unknown

- dist. of $\underset{\sim}{Y}^* \sim N$

$\Rightarrow$ dist. of Y

Likelihood $(\underset{\sim}{\beta}, \sigma^2, \lambda \mid \underset{\sim}{Y}, \underset{\sim}{X})$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{(\underset{\sim}{Y}^* - \underset{\sim}{X}\underset{\sim}{\beta})^2}{2\sigma^2}\right\} * J(\lambda, \underset{\sim}{Y})$$

where $J(\lambda, \underset{\sim}{Y}) = \boxed{\prod_{i=1}^{n} y_i}^{\lambda-1}$

$= GM(y)^{\lambda - 1}$

$\uparrow$ geometric mean

Linear regression

$$\log_e L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2) - \frac{n}{2}$$

$$= -\frac{n}{2}\log\left(\frac{Res.S.S}{n}\right) + \text{constant}$$

log likelihood ~~likelihood~~ $(\lambda) = -\frac{n}{2}\log(\tilde{\sigma}^2(\lambda))$

$$\tilde{\sigma}^2(\lambda) = \frac{\underset{\sim}{Y}^{*T}\underset{\sim}{Y}^{*} - \underset{\sim}{Y}^{*T}\underset{\sim}{X}^{*}(\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{X}^T\underset{\sim}{Y}^{*}}{n}$$

profile likelihood

Define $\underset{\sim}{Z}^\lambda = \underset{\sim}{Y}^* / J^{1/n}$

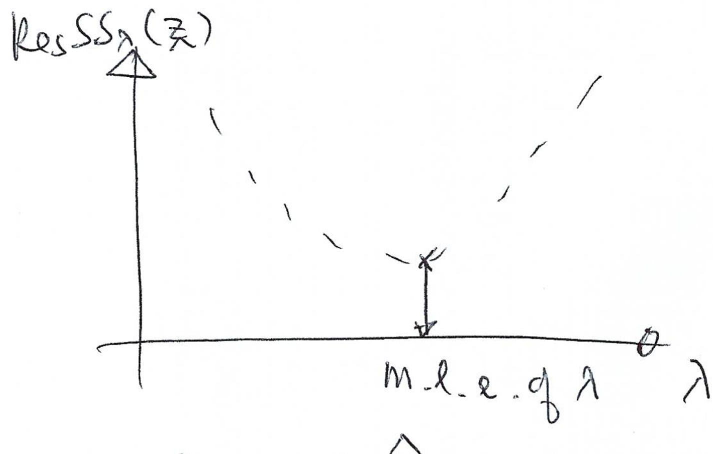
$\Rightarrow$ log likelihood $(\lambda) = -\frac{n}{2}\log(Res.S.S._\lambda(\underset{\sim}{Z}^\lambda))$

where $\underset{\sim}{Z}^\lambda$ is $n \times 1$ vector with $Z_i^\lambda = \begin{cases} \dfrac{y^\lambda - 1}{\lambda[GM(y)]^{\lambda-1}} & \lambda \neq 0 \\ GM(y)\ln(y_i) & \lambda = 0 \end{cases}$

For each λ, $\Rightarrow$ calculate Z_i^λ.

fit Z_i^λ on $\underset{\sim}{X}$

$\Rightarrow Res.S.S._\lambda(\underset{\sim}{Z})$



e.g. $\hat{\lambda} = 0.423$ ∴ 95% of $\hat{\lambda}$?

Does it over 0.5 ?