HW - 3 : M (2-05) | Quiz - 2 : M (2-05) | HW - 4 : M (2-12) | Quiz - 3 : M (2-12) | Class-4

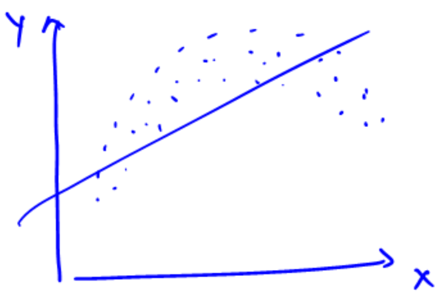
Residual plots (plots of residuals or Semi Studentized residuals) can be used to identify above departures from the model.

1. Non-linearity of regression function

There are two types of plots

a) Scatter plot (Y vs X)

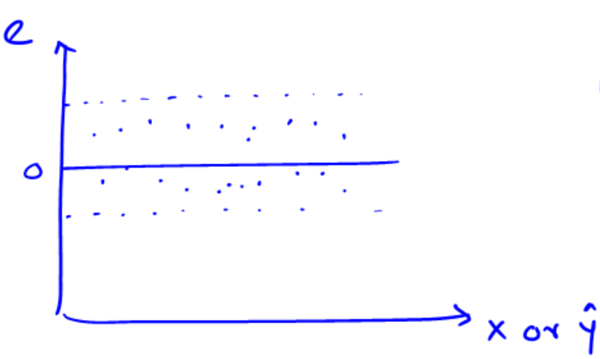
→ linear ⇒ good fit



→ not linear ⇒ not a good fit.

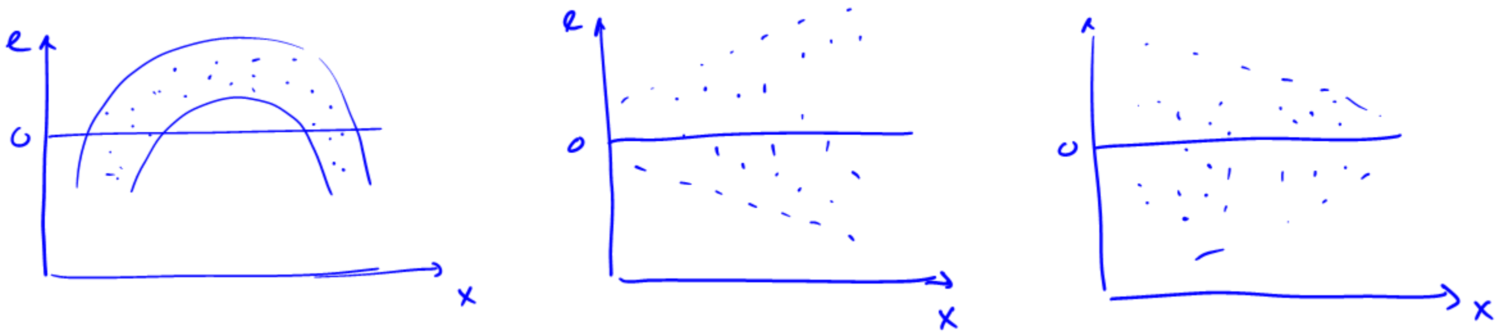
b) Residual plots (residuals vs X or residuals vs $\hat{Y}$)

For SLR model plots for residual vs X and residual vs $\hat{Y}$ show the same pattern.



All the residuals fall within a horizontal band around 0.
⇒ good fit.

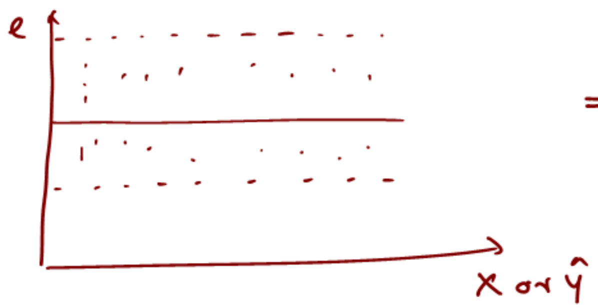
Departures:



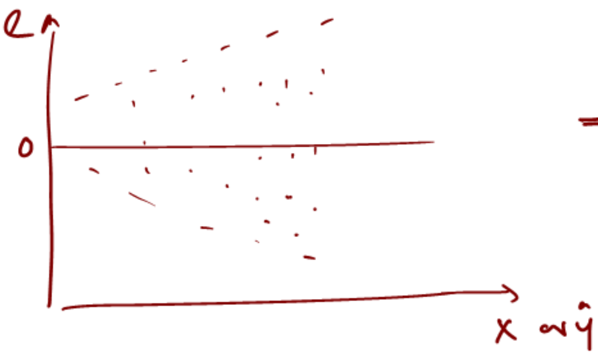
$\Rightarrow$ not a good fit.

(2) Non-constancy of error variance

a) Residual plots (vs X or $\hat{y}$)

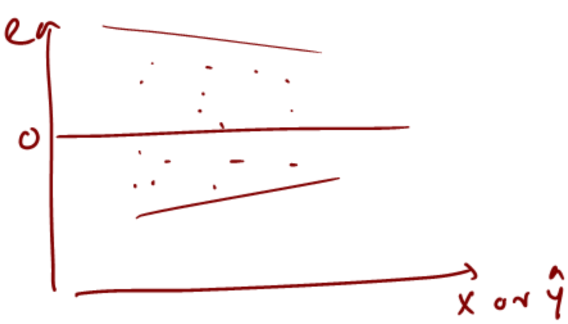


$\Rightarrow$ constant variance.



$\Rightarrow$ variation increases with X

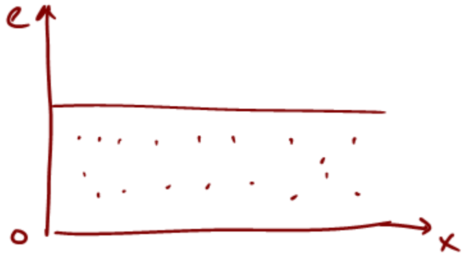
(not a good fit)



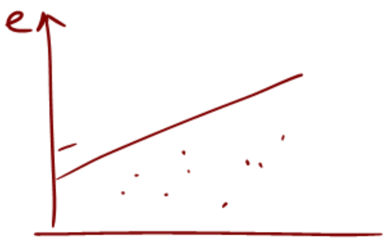
$\Rightarrow$ variation decreases with X

(not a good fit).

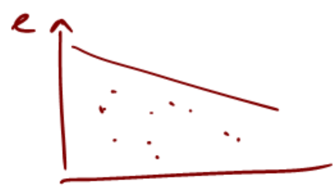
b) Absolute residuals vs X (or $\hat{y}$)



$\Rightarrow$ good fit



or



$\Rightarrow$ not a good fit.

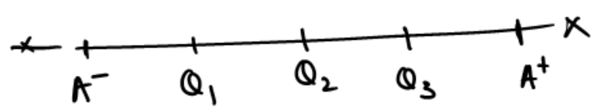
This is usfull when the sample size is small.

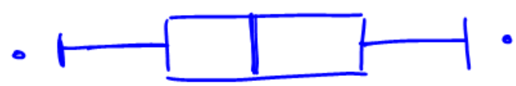
③ Presence of outliers

A^-, Q_1, Q_2, Q_3, A^+

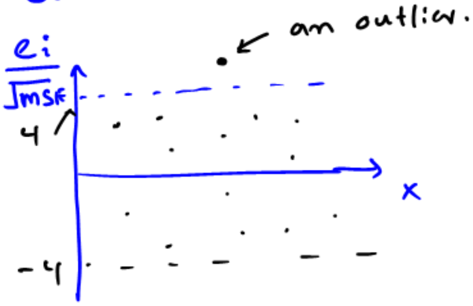
$$A^- = Q_1 - 1.5\,(Q_3 - Q_1)$$

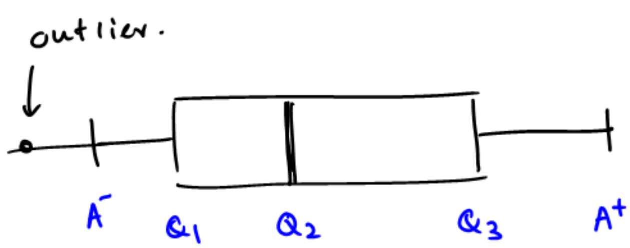
$$A^+ = Q_3 + 1.5\,(Q_3 - Q_1)$$





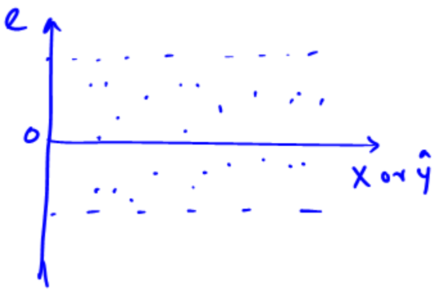
If $\left|\frac{e_i}{\sqrt{MSE}}\right| \geq 4$, then the i^{th} data point is an outlier. Residual plots or box plot of residuals can be used to identify outliers.





④ Non-independence of error terms

Residual plots (e vs X or e vs $\hat{y}$).



$\Rightarrow$ No pattern $\Rightarrow$ independent

$\Rightarrow$ Some pattern $\Rightarrow$ not independent.

⑤ Non-normality of error terms

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

Box-plot or histogram can be used to check the shape of the distribution.

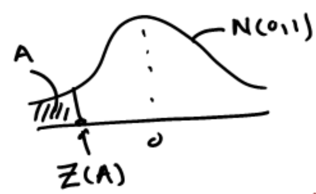
Normal probability plot can be used to check the normality.

Normal probability plot

plot of residuals vs their expected values under normality.

Caculating expected values:

Expected value for e_i $= \sqrt{MSE}\left(Z\left(\frac{K - 0.375}{n + 0.25} \right) \right)$,



where $Z(A)$ is the $A.100$ percentile for the standard normal distribution and K is the rank of the residual e_i. If there is a tie, take the average of the ranks.

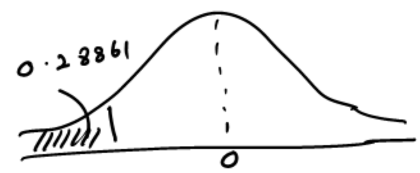
Eg:- -2.43, -2.31, 0.01, 0.01, 0.5, 0.67

	1	2	3	4	5	6
rank	1	2	3.5	3.5	5	6

Eg:- Plastic hardness example:

Residual (e_1) = -2.150 rank (e_1) = $K = 5$, $n = 16$

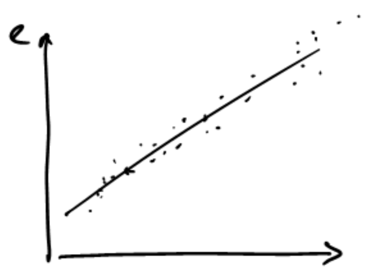
So $\dfrac{K - 0.375}{n + 0.25} = \dfrac{5 - 0.375}{16 + 0.25} = 0.28861$



$MSE = (3.23403)^2$, $Z(0.28861) = -0.55748$

Expected value $= (3.23403) * (-0.55748)$

$= -1.802903.$



linear $\Rightarrow$ normality.

non linear $\Rightarrow$ non-normality.

Test for Lack of fit (Numerical tools)

* Correlation test for normality

Here calculate the coefficient of correlation residuals and their expected values under normality. Then compare that with the corresponding value in table B6 (in the text). Values in the table B6 are the percentiles of the distribution of the correlation with normally distributed error terms.

If the observed coefficient > table value

$\Rightarrow$ error terms are normal.

Note:

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma(X)\cdot\sigma(Y)} = \frac{E(XY) - E(X)\cdot E(Y)}{\sigma(X)\cdot\sigma(Y)} \quad \text{— parameter.}$$

Sample term:

$$r = \frac{S^2_{xy}}{\sqrt{S^2_x \cdot S^2_y}} \quad \text{— Sample correlation ,}$$

where, $S^2_{xy} = \Sigma(XY) - \frac{\Sigma(X)\cdot\Sigma(Y)}{n}$,

$$S^2_x = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

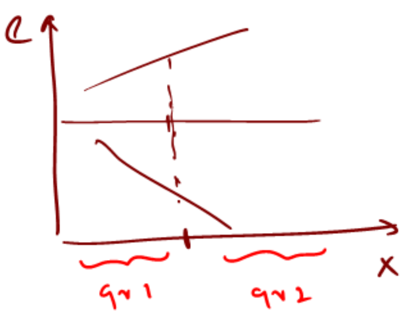
$$S^2_y = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

* Test for constant error variance

Brown – Forsythe test

Idea:

when error variance is not constant.



Steps:

1) devide residuals into two groups based on the level of X.

group 1 : n_1 values (lower X values)

group 2 : n_2 values (higher X values)

2) calculate

$$d_{i1} = | e_{i1} - \tilde{e}_1 |$$

$$d_{i2} = | e_{i2} - \tilde{e}_2 | ,$$

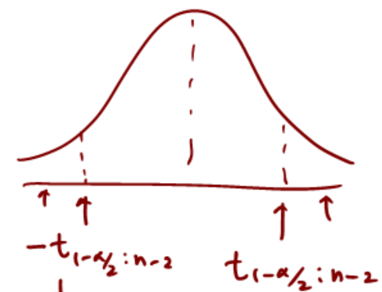
where e_{ik} – ith residual for the k^{th} group ($k=1,2$),

$\tilde{e}_k$ – median of the residuals for the k^{th} group ($k=1,2$).

3) calculate

$$T = \frac{\bar{d}_1 - \bar{d}_2}{S\sqrt{1/n_1 + 1/n_2}} \sim t_{n-2},$$

where $$S^2 = \frac{\sum (d_{i1} - \bar{d}_1)^2 + \sum (d_{i2} - \bar{d}_2)^2}{n-2}$$



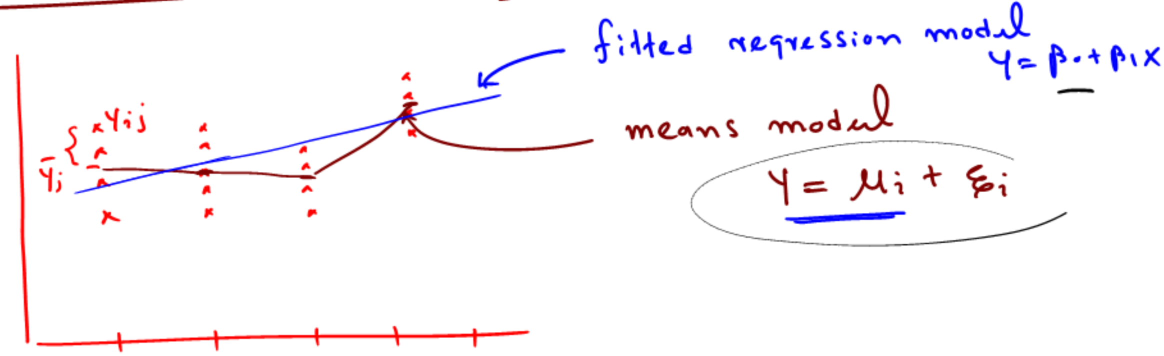
4) Suppose t^* be the observed value of T.

If $|t^*| \le t_{1-\alpha/2:n-2}$ $\Rightarrow$ error variance is constant

$|t^*| > t_{1-\alpha/2:n-2}$ $\Rightarrow$ error variance is not constant.

* Test for non-linearity (general linear test)

Idea:



Assumptions:

- $Y_i \overset{ind}{\sim} N(E(Y), \sigma^2)$ ← constant
- There are repeat observations at one or more X-levels.

Note:

Repeat observations for the same X-level are called replicates.

Notation:

Data should be arranged by level of X and replicate number.

level of X	$j=1$ (X_1)	$j=2$ (X_2)	$j=3$	....	$j=c$
replicate					
$i=1$	Y_{11}	Y_{12}	Y_{13}		Y_{1c}
$i=2$	Y_{21}	Y_{22}	Y_{23}		Y_{2c}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	$Y_{n_1 1}$	$Y_{n_2 2}$	$Y_{n_3 3}$		$Y_{n_c c}$
	$\bar{Y}_1$	$\bar{Y}_2$	$\bar{Y}_3$	- . . . -	$\bar{Y}_c$

Hypotheses:

H_0 : $E(Y) = \beta_0 + \beta_1 X$ (ie Regression function is linear)

H_1 : $E(Y) \neq \beta_0 + \beta_1 X$ (ie Regression function is not linear)

Test statistic:

* Full model:

If a regression model is not fitted (ie predicter variable is not used) the model is

$$Y_{ij} = \mu_j + \varepsilon_{ij}$$, — means model (ANOVA model)

$i = 1, 2, \dots n_j$

$j = 1, 2, \dots c.$

$\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.

$$E(Y_{ij}) = \mu_j$$

* least square estimate of $\mu_j = \hat{\mu}_j = \bar{Y}_j \leftarrow$ Sample mean of the j^{th} group.

* $SSE(F) = \underbrace{\sum_j \sum_i (\underbrace{Y_{ij} - \bar{Y}_j}) ^2}$

This is also called as pure error sum of squares (SSPE).

* df of $SSE(F) = \underbrace{\sum_{j=1}^{c} (n_j - 1)} = n - c$.

* <u>Reduced model:</u> (SLR model)

model : $Y_{ij} = \beta_0 + \beta_1 X + \varepsilon_{ij}$, $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

$E[Y_{ij}] = \beta_0 + \beta_1 X_j$ — does depend on X .

* least square estimator $E[Y_{ij}] = b_0 + b_1 X_j$

* error sum of squares = $SSE(R) = \sum_i \sum_{j=} (Y_{ij} - b_0 - b_1 X_j)^2$

$= \sum \sum (Y_{ij} - \hat{Y}_{ij})^2 = SSE$

df or $SSE(R) = n-2$.

Test statistic

$$F = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

$$= \frac{SSE - SSPE}{n-2-(n-c)} \div \frac{SSPE}{n-c}$$

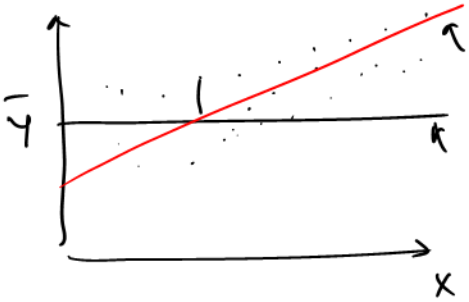
But $SSE - SSPE = SSLF$ — lack of fit sum of squares.

Now $F = \dfrac{SSLF}{c-2} \Big/ \dfrac{SSPE}{n-c} = \dfrac{MSLF}{MSPE}$ ← lack of fit mean square / ← pure error mean square

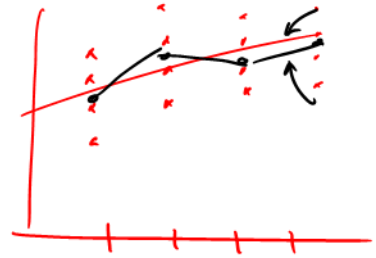
<u>Decision:</u> Let f^* be observed value of F.

* $f^* \le F(1-\alpha, c-2, n-c) \Rightarrow$ conclude H_0 (ie regression function is linear)

$f^* > F(1-\alpha, c-2, n-2) \Rightarrow$ conclude H_1 (ie regression function is not linear).

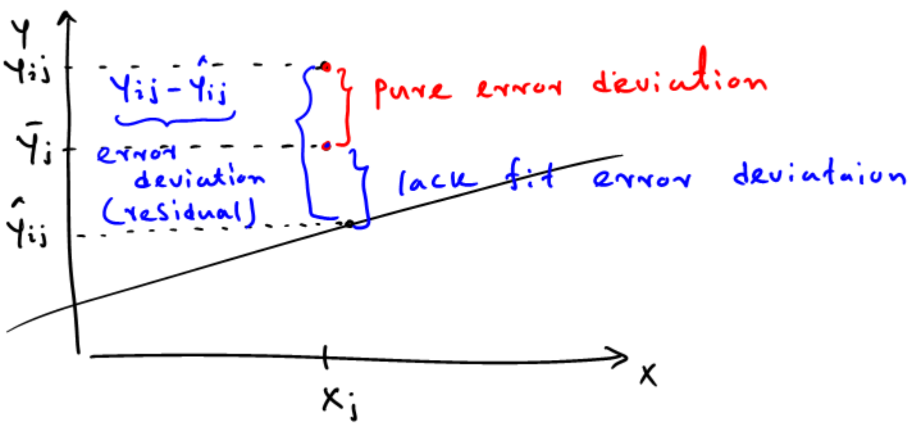


$$y_i = \underset{\uparrow \atop \bar{y}}{\mu} + \varepsilon_i$$



$$y_{ij} = \underset{\uparrow \atop \bar{y}_j}{\mu_j} + \varepsilon_{ij}$$

Note:



From the graph

$$y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_j + \bar{y}_j - \hat{y}_{ij}$$

we also can show.

$$\underbrace{\sum (y_{ij} - \hat{y}_{ij})^2}_{SSE} = \underbrace{\sum (y_{ij} - \bar{y}_j)^2}_{SSPE} + \underbrace{\sum (\bar{y}_j - \hat{y}_{ij})^2}_{SSLF}$$

Proof: HW.

All of above values are summarized in a table called general ANOVA table.

General ANOVA table:

Source of variation	SS	df	MS
Regression	$SSR = \sum\sum (Y_{ij} - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
error	$SSE = \sum\sum (Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Lack of fit	$SSLF = \sum\sum (\bar{Y}_j - \hat{Y}_{ij})^2$	$c-2$	$MSLF = \frac{SSLF}{c-2}$
Pure error	$SSPE = \sum\sum (Y_{ij} - \bar{Y}_j)^2$	$n-c$	$MSPE = \frac{SSPE}{n-c}$
Total	$SSTO = \sum\sum (Y_{ij} - \bar{Y})^2$	$n-1$	