Highlights / Key Concepts in NKNW4 Chapter 3

- **§3.1 Dx for Predictor Variable**

Spread of x's ... range of validity ; efficiency (if have choice at design stage)

Time sequence of measurements ... "drift", learning effect etc

- **§3.2 Residuals**

Definition: $e = Y - \hat{Y}$ [e's observed; there is not a 1:1 correspondence with unobservable $\varepsilon = Y - E(Y)$]

[if model correct and ε 's follow $N(0,\sigma)$ distribution, can expect certain patterns in e's]

Even without any distributional assumptions, i.e. simply from their construction/calculation...

- ave of n e's = 0

- Variance of n e's = $\frac{1}{n-2}\Sigma e^2$ = SSE / (n-2) = Mean Square Error (average squared residual)

  n-2 "degrees of freedom" = n-2 indpendent assessments of σ^2 [if model correct]

  (no internal estimate of σ^2 if n=2 , since 2 y's define the fitted line)

- e's (almost) independent of each other (if n-2 reasonably large)

  2 constraints

  -- e's sum to zero

  -- x-weighted average of e's = 0 $\Sigma xe = 0$ ==> e's "orthogonal" to x's

Semi-studentized Residuals -- making the residuals independent of Y scale (akin to Z scores)

$$\frac{e\ -\ 0}{\text{estimate of }\sigma} = \frac{e}{MSE^{1/2}} = \frac{e}{\text{Root Mean Square Error}} = \frac{e}{RMSE}$$

Why "Semi-studentized" ... Even ifvar(ε) = σ^2 for all X, that does not mean that the e's have equal variability

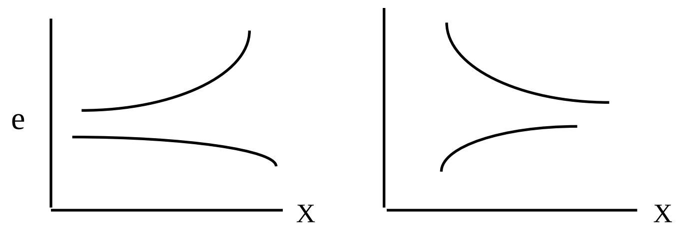
e's at X extremes vary less that e's at X centre [see the EMS spreadsheet] so e/RMSE somewhat crude.

- Departures from model and **§3.3 Diagnostics for Residuals**

	Plots					
	e's vs X	e's vs $\hat{Y}$	e's vs Time	e distribution	QQplot	(e's vs other X)
$\mu_{Y\|X}$ not linear in X	×	×				
$\sigma^2{}_{Y\|X}$ not constant over X	×					
ε's not Gaussian distribution				×	×	
ε's (serially/otherwise) correlated			×			
outlier observations	×	×				
omitted predictors						× (or symbols)

NON-CONSTANCY OF ERROR VARIANCE (common ; seldom critical*)

Visually



1) Compare absolute deviations of e's from their medians
categorized according to high/low x -- 2 sample t test (modified Levine test)

2) More quantitaive -- regress e-squared on X -- test for non-zero slope.

TX: IF NON-CONSTANCY OF VARIANCE IS ONLY "PROBLEM" ...

use Weighted Least Squares for more efficient estimates of regression coefficients

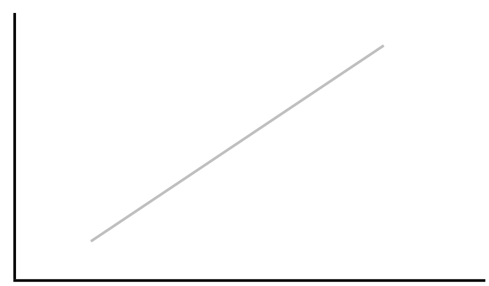
* if using for individual variation (e.g growth charts) model $\sigma(\varepsilon|X)$ as function of X

NON-NORMALITY CHECKS

Informallly (Visually)   Histogram   QQ Plot (Normal Probability Plot)

e e
e e e e
e e e e e e
e e e e e e e e
e e e e e e e e e e e
e e e e e e e e e e e e e

Observed Residual



Expected Value of residual under Normality

MSE × Z_i

use Z's to approximate the expected values of the n order statistics from N(0,1) distrn.

compute Z's corresponding to p = $\frac{5/8}{n + 1/4}$ $\frac{1\ 5/8}{n + 1/4}$ . . . $\frac{(n-1) + 5/8}{n + 1/4}$

if n = 40, the 40 order statistics are at approx. 2.5%ile, 5%ile, 7.5%ile, ... 97.5%ile.
(Could omit MSE from both and use studentized residual on vertical axis).

FORMALLY

Correlation of e's and Expected e's -- Looney & Gulledge Table -- want high correlation

Kolmogorov - Smirnoff ...
...

- **§3.4 - §3.6 Formal tests involving residuals** [I don't use them]

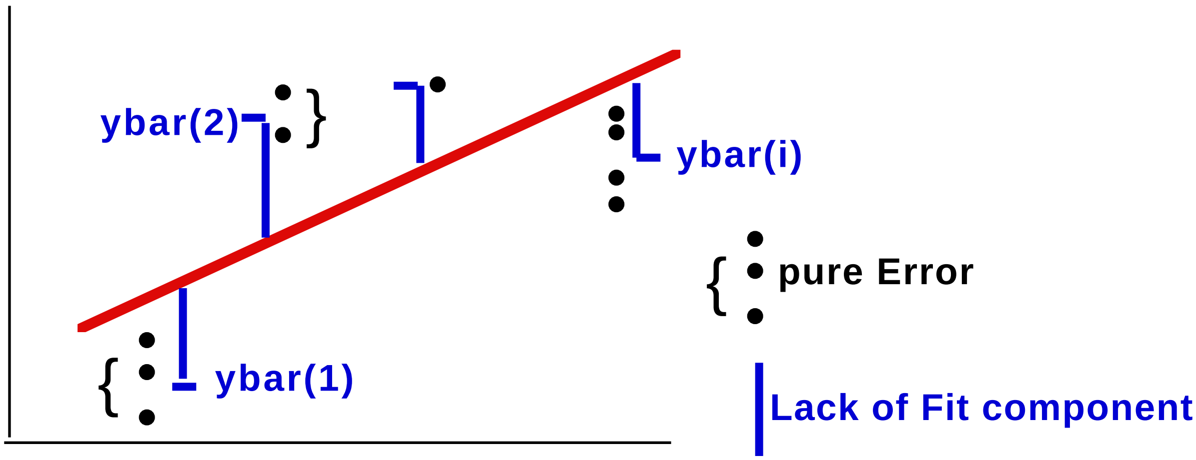
- **§3.7** TEST OF LACK OF FIT (WHEN HAVE REPLICATE Y'S AT SOME X'S) (say at c unique X values)

Saturated Model	Linear Regression Model
$Y_{ij} = \mu_i + \varepsilon_{ij}$	$Y_{ij} = \beta_0 + \beta_1 X_i + \varepsilon_{ij}$
estimate c separate means	estimate c means via single line (2 df)
use n-c df to estimate "pure" (within X) variation of Y	have n-2 df which we can split up into

SS pure error n - c df (*)

$SS_{\text{Lack of fit}}$	**c-2 df**
SSpure error	**n-c df [from *]**
SS residual	n-2 df

$$\textbf{Large F} = \frac{SS_{\text{Lack of fit}}/\{c-2\}}{SS_{\text{pure error}}/\{n-c\}} \implies \textbf{Lack of Fit}$$



To compute $SS_{\text{lack of fit}}$...

- obtain ss residual from lin. reg. model (includes 2 components)

- obtain ss "pure-error" fron ANOVA program

eg. in SAS
PROC GLM; CLASS X; (X as categoriacl variable)
MODEL Y =X; (1-way ANOVA)

- subtract the two

Source	SS
Between	{Model}
Within	{Error } <- *pure error.*

Idea in <u>Lack of Fit</u> Test i.e. <u>Saturated</u> Model vs <u>Reduced</u> Model is a prototype for similar tests in multiple regression

Model with p parameters	**n-p**	**df for error**
Model with p+ k parameters	**n-p-k**	**df ...**

$$F = \frac{[SS_{\text{residual}}\ \text{smaller model} - SS_{\text{residual}}\ \text{larger model}]\ /\ k}{SS_{\text{residual}}\ \text{larger model}\ /\ (n-p-k)}$$