LM, LNp.7-1~2

Residual Analysis: Theory

- Theory: define the **residual** for the i^{th} observation (x_i, y_i) as

  $\hat{\varepsilon}_i = r_i = y_i - \hat{y}_i, \quad \hat{y}_i = \mathbf{x}_i^T \hat{\beta}$, ($i$th row of model matrix X)

  $\hat{y}_i$ contains information given by the model; r_i is the "difference" between y_i (observed) and $\hat{y}_i$ (fitted) and contains information on possible *model inadequacy*.

  $\hat{\varepsilon}_i$ often used to check assumptions:
  ① ε_i's i.i.d. $N(0, \sigma^2)$
  ② $E(Y) = X\beta$ is a correct mean structure

- Vector of residuals $\mathbf{r} = (r_1, \ldots, r_N)^T = \mathbf{y} - \mathbf{X}\hat{\beta}$. $= (I-H)Y$ (H: hat matrix)

  overall pattern
  individual observation (outlier)

- Under the model assumption $E(\mathbf{y}) = \mathbf{X}\beta$, it can be shown that (i.e., $X\beta$ is correct model)

  (a) $E(\mathbf{r}) = 0$, ($\because r \perp \hat{Y}$)

  (b) $\mathbf{r}$ and $\hat{\mathbf{y}}$ are independent,

  (c) variances of r_i are nearly constant for "nearly balanced" designs.

  Note. variance of r_i $\propto 1-h_i$ (leverage) $= 1 - H_{ii}$
  $\propto$ Mahananobis dist. btwn design pts & $\bar{x}$

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

① $Y = X\beta + \varepsilon = \hat{Y} + \hat{\varepsilon}$

② $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon = (X_1\beta_1 + H_1X_2\beta_2) + ((I-H_1)X_2\beta_2 + \varepsilon) = \hat{Y}_{X_1} + \hat{\varepsilon}_{X_1}$

Under fitted model: $Y = X_1\beta_1 + \varepsilon^*$ ($H_1 = X_1(X_1^TX_1)^{-1}X_1^T$)

Residual Plots

LM, LNp.7-7~9

$\sigma_x \leftrightarrow \mu_x$

- Plot r_i vs. $\hat{y}_i$ (see Figure 1): It should appear as a parallel band around 0. Otherwise, it would suggest model violation. If spread of r_i increases as $\hat{y}_i$ increases, error variance of y increases with mean of y. May need a transformation of y. (Will be explained in future lecture.) (LNp.24)

- Plot r_i from replicates per treatment (see Figure 2): to see if error variance depends on treatment. (LNp.25) (Note. saturated model)

  ($\hat{\varepsilon}$ vs. predictor)

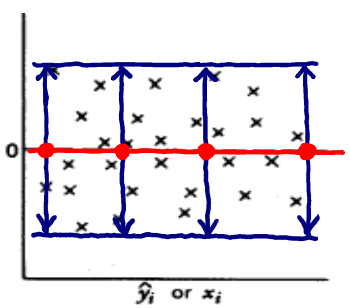
  quali-tative factor → Box plots

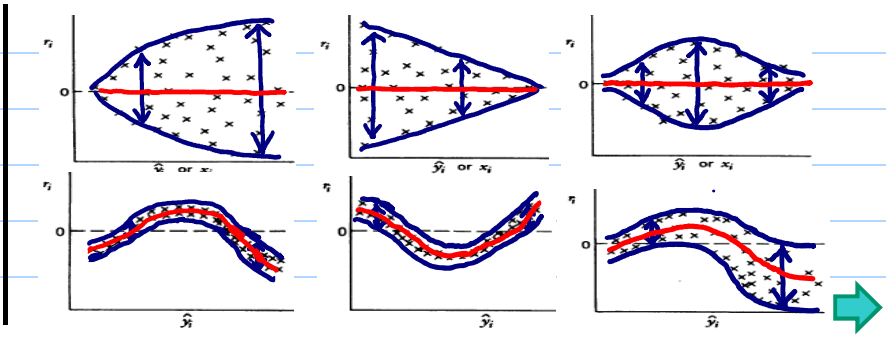
- Plot r_i vs. x_i: If not a parallel band around 0, relationship between y_i and x_i not fully captured, revise the $\mathbf{X}\beta$ part of the model.

  quanti-tative factor

- Plot r_i vs. time sequence: to see if there is a time trend or autocorrelation over time. (if available (or run order, measure order, ...))

null plot





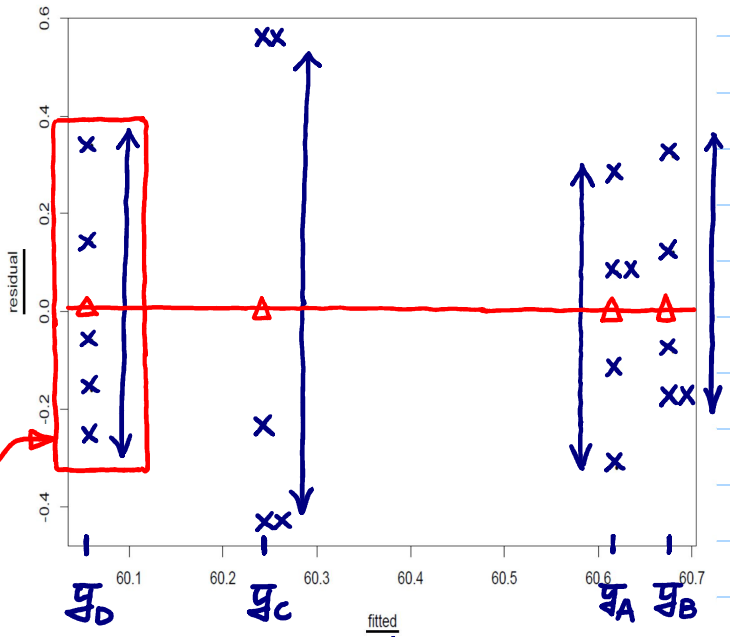
Plot of r_i vs. $\hat{y}_i$

no replicate data

simple model $\hat{\varepsilon}$

complex model $\hat{\varepsilon}$

saturated model $\hat{\varepsilon}$



replicates → allow "pure error" variance estimation

$\hat{y}_D$ $\hat{y}_C$ $\hat{y}_A$ $\hat{y}_B$

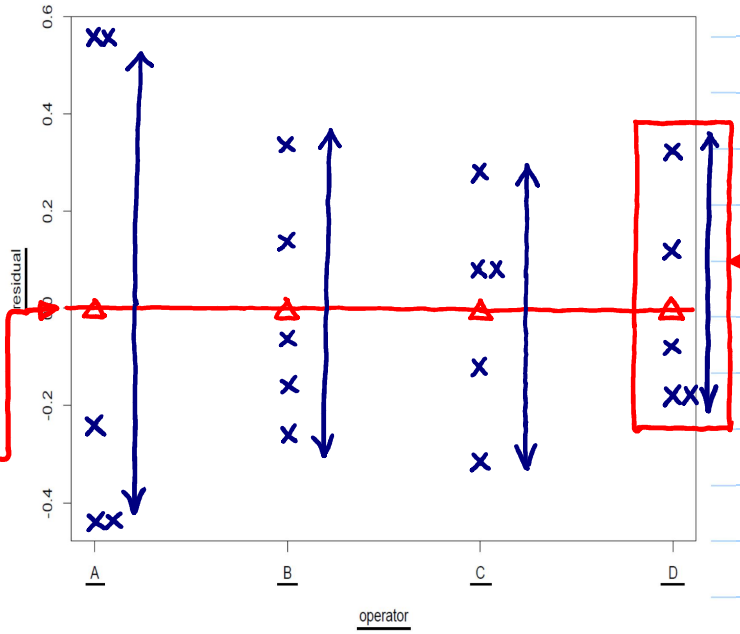
$\hat{y}_i$ ← a function of treatments A,B,C,D

Figure 1: r_i vs. $\hat{y}_i$, Pulp Experiment

LNp.1

Plot of r_i vs. treatment

fitted model: saturated



Same 4 groups of residuals as in the residual plot of r_i vs. $\hat{y}_i$ (LNp.3-24)

no need to check the trend in the mean of residuals

$\sum_j \hat{\varepsilon}_{ij} = 0$ for any i

$X =$

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & & 0 \\ 1 & 1 & 1 & & 0 \\ 1 & 1 & 1 & & 0 \\ 1 & 1 & 1 & & 0 \\ 1 & 1 & 1 & & 0 \\ 0 & 1 & 1 & & 1 \end{bmatrix}$$

Figure 2: r_i vs. treatment, Pulp Experiment

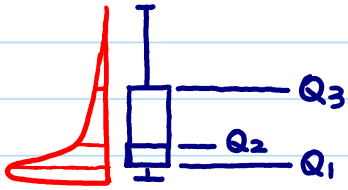
LNp.1

$Y_1, \ldots, Y_n$ iid $\sim$ cdf F — Box-(Whisker) Plot

- A powerful graphical display (due to Tukey) to capture the location, dispersion, skewness and extremity of a distribution. See Figure 3. ↑LNp.27
- Q_1 = lower quartile (25% quantile), Q_3 = upper quartile (75% quantile), Q_2 = median (50% quantile, estimate of *location* parameter) is the white line in the box. Q_1 and Q_3 are boundaries of the *black box*.
- IQR = interquartile range (length of box) = $Q_3 - Q_1$: measure of *dispersion*.
- Minimum and maximum of **observed** values within

$$[Q_1 - 1.5 \times IQR,\ \ Q_3 + 1.5 \times IQR]$$

are denoted by two *whiskers*. Any values outside the whiskers are regarded as extreme values and displayed (possible *outliers*).



- If Q_1 and Q_3 are not symmetric around the median, it indicates *skewness*.
- Side-by-side box plots (LNp. 3-2~3) are useful to compare the difference between the distributions of several groups of data.

Box-(Whisker) Plot

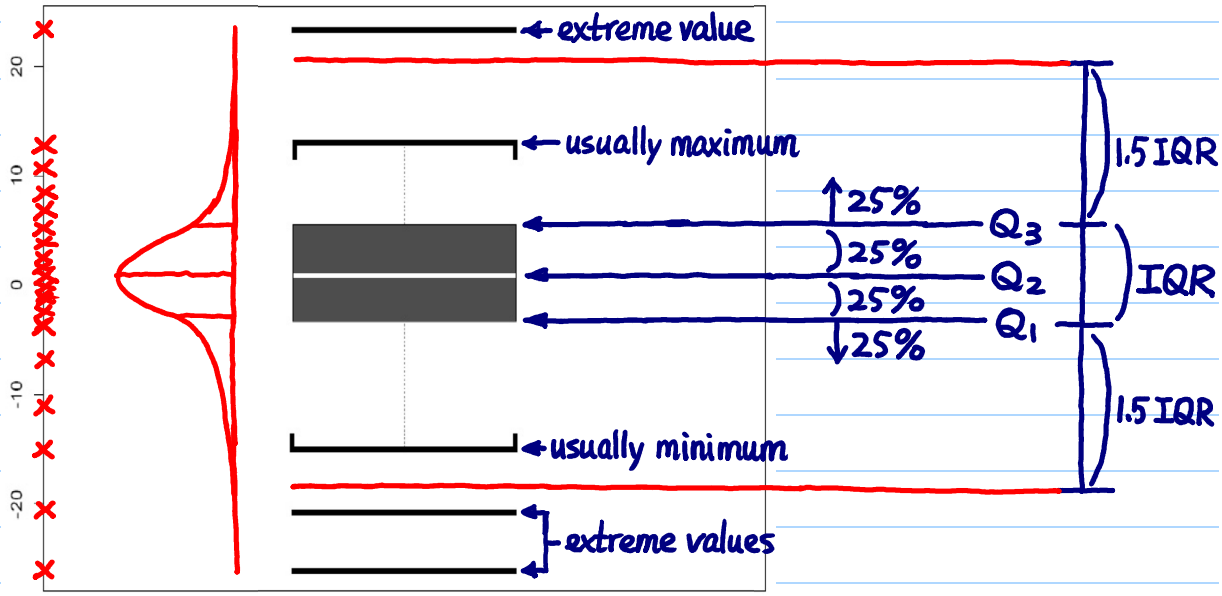
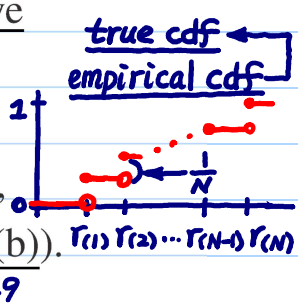


Figure 3: Box-Whisker Plot

Normal Probability Plot ← Q-Q plot (LM, LNp.7-15~16)

- Original purpose : To test if a distribution is normal, e.g., if the residuals follow a normal distribution (see Figure 5). Q: Why need normality for error?
  → can be used to identify outlier ↑LNp.30
- More powerful use in factorial experiments (discussed in Units 5 and 6).
  → used to identify significant effects β_i's ↔ cf. replace t-tests
- Let $r_{(1)} \le \ldots \le r_{(N)}$ be the ordered residuals. The cumulative probability for $r_{(i)}$ is $p_i = (i - 0.5)/N$. Thus the plot of p_i vs. $r_{(i)}$ should be S-shaped as in Figure 4(a) ←LNp.29 if the errors are normal. By transforming the scale of the horizontal axis, the S-shaped curve is straightened to be a line (see Figure 4(b)). ↑LNp.29



- **Normal probability plot** of residuals :

$$\left(\Phi^{-1}\left(\frac{i-0.5}{N}\right), r_{(i)}\right), \quad i = 1, \ldots, N, \quad \Phi = \text{normal cdf}.$$

If the errors are normal, it should plot roughly as a straight line. See Figure 5. ←LNp.30

Regular and Normal Probability Plots of Normal

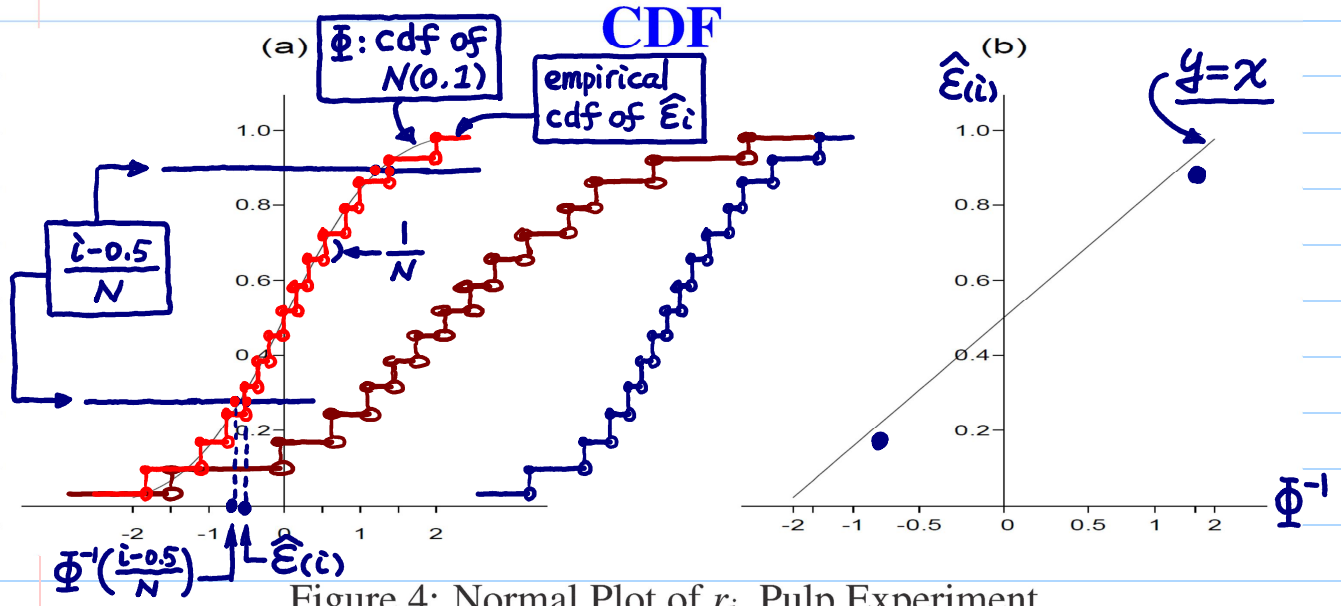


Figure 4: Normal Plot of r_i, Pulp Experiment

$z_1, \ldots, z_N$ iid $\sim N(\mu, \sigma^2)$

$z_i = \sigma w_i + \mu \Leftarrow w_i = (z_i - \mu)/\sigma$ iid $\sim N(0,1)$

(z_i: y', w_i: y)

Normal probability plot of w_i's $\Rightarrow w_{(i)}$ vs. Φ^{-1}: $y = x$

Normal probability plot of z_i's $\Rightarrow z_{(i)}$ vs. Φ^{-1}: $y' = \sigma x + \mu$

Normal Probability Plot : Pulp Experiment

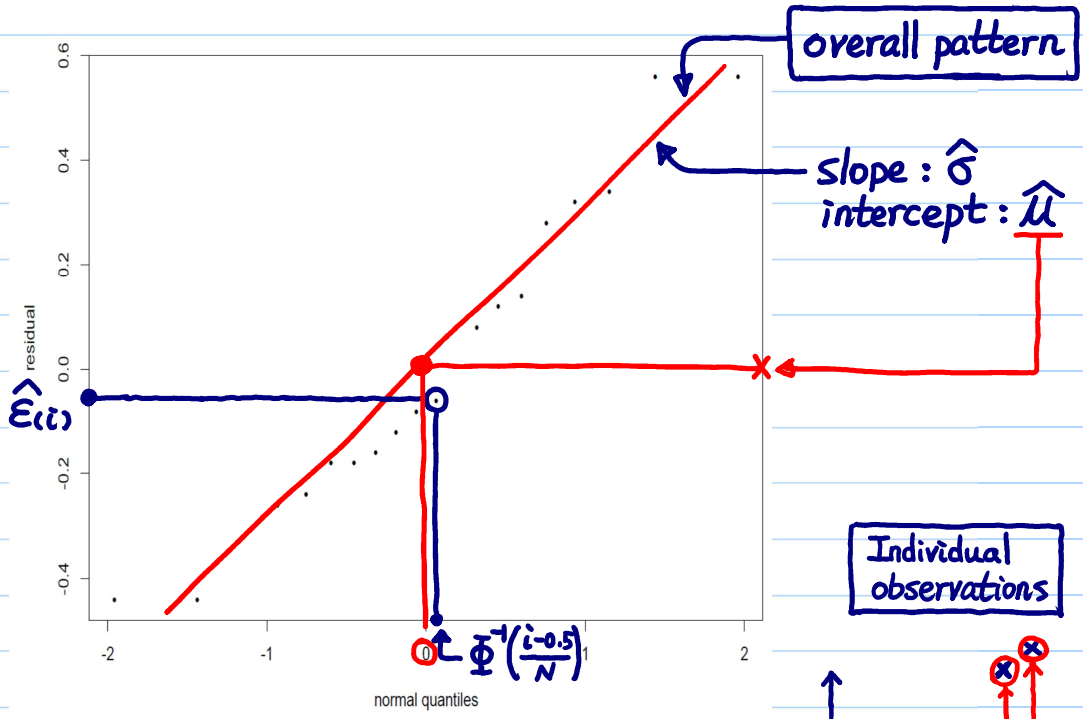
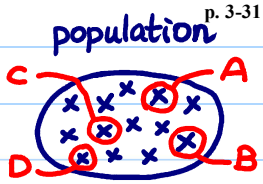


Figure 5: Normal Plot of r_i, Pulp Experiment

❖ **Reading**: textbook, 2.6

Pulp Experiment Revisited

Q: Are A.B.C.D a representitive sample of the population ?



LNp.1

- Compare the 2 scenarios

LNp.4

(S1) plant has only 4 operators (or only interested in these 4 operators)

effects of operators

after sampling & conditioning

- τ_i's: parameters (unknown fixed values)
  - interest: difference btwn the 4 specific τ_i's

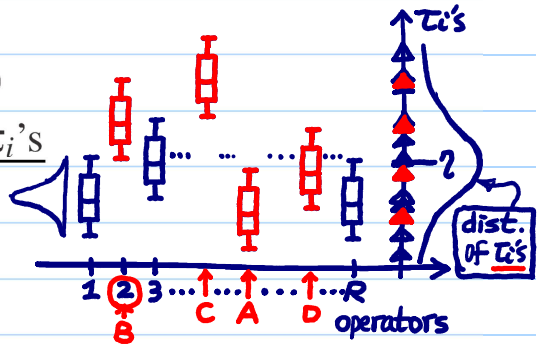
cf.

(S2) 4 operators randomly sampled from a large population of operators

before sampling

- τ_i's: random variables
  - interest: difference btwn all operators in this population

$\mu_i = E(y_{ij})$



- In the pulp experiment the effects τ_i are called *fixed* effects because the interest was in comparing the four *specific* operators in the study. If these four operators were chosen randomly from the population of operators in the plant, the interest would usually be in the variation among all operators in the population. Because the observed data are from operators randomly selected from the population, the variation among operators in the *population* is referred to as *random* effects.

conditioned on these 4 operators

- **One-way random effects model (REM $\overset{\text{cf.}}{\longleftrightarrow}$ FEM) :**

FEM in LNp.3-4&3-8 (cf.)

$$y_{ij} = \eta + \tau_i + \varepsilon_{ij},$$

intercept, parameter; τ_i: r.v.; cf. whole- & sub-plot errors in LNp.4-47~48

$y_{ij} \sim N(\eta, \sigma_\tau^2 + \sigma^2)$

Q: y_{ij}'s indep.?

where ε_{ij}'s: independent error terms with $N(0, \sigma^2)$,

τ_i's: independent $N(0, \sigma_\tau^2)$, (why?)

and τ_i and ε_{ij} are independent (Why? Give an example.);

σ^2 and σ_τ^2 are the two *variance components* of the model. (check Σ^*)

The variance among operators in the population is measured by σ_τ^2.

REM: parameters $\eta, \sigma_\tau^2, \sigma^2$

$Y = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{k1} \\ \vdots \\ y_{kn_k} \end{bmatrix}$

REM. (1) $E(Y) = \eta \mathbb{1}$

(2) $\text{cov}(y_{11}, y_{12}) = \text{cov}(\eta + \tau_1 + \varepsilon_{11}, \eta + \tau_1 + \varepsilon_{12}) = \sigma_\tau^2$

$\text{cov}(y_{11}, y_{21}) = \text{cov}(\eta + \tau_1 + \varepsilon_{11}, \eta + \tau_2 + \varepsilon_{22}) = 0$

$\text{cov}(Y) = \Sigma^* = \begin{bmatrix} \square_{n_1 \times n_1} & 0 & \cdots \\ 0 & \square_{n_2 \times n_2} & \\ \vdots & & \ddots \end{bmatrix}$, $\square = \begin{bmatrix} \sigma_\tau^2 + \sigma^2 & \sigma_\tau^2 & \cdots \\ \sigma_\tau^2 & \sigma_\tau^2 + \sigma^2 & \\ \vdots & & \ddots \end{bmatrix}$

$= \sigma_\tau^2 \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} + \sigma^2 \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$ (X)

FEM. $y_{ij} \overset{\text{indep.}}{\sim} N(\eta + \tau_i, \sigma^2)$

(1) $E(Y) = X\beta \leftarrow \eta + \tau_i$

(2) $\text{cov}(Y) = \sigma^2 I$ (X)

★ relationship btwn y_X & X

- In FEM, E(Y)
- In REM, cov(Y)