DIAGNOSTICS

We analyze the empirical properties of the residuals to understand if they are coherent with the theoretical ones. We use plots:

(1) RESIDUALS vs FITTED (PREDICTED) → SCATTERPLOT of $\tilde{e}_i$ vs $\hat{y}_i$

if the model assumptions are satisfied, what should we observe?

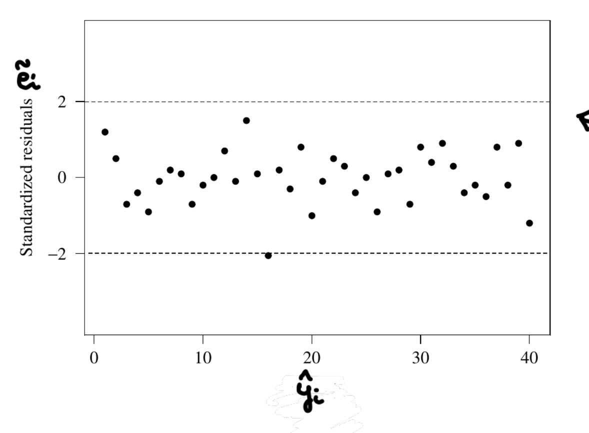
- SYMMETRIC DISTRIBUTION → balanced number of positive and negative residuals
- LINEARITY of the relationship between x and y → no systematic behavior left unexplained
- HOMOSCEDASTICITY → constant variability

In the case of the simple linear model, we can equivalently look at the plot of $\tilde{e}_i$ vs x_i (since $\hat{y}_i$ is a linear transformation of x_i).

note: we use standardized or studentized residuals to have constant variance

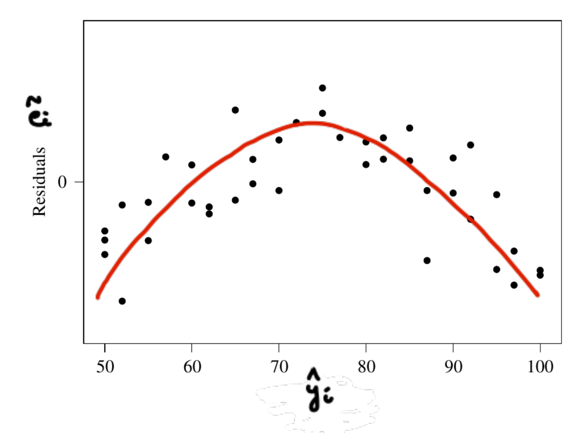
Examples:

if the assumption is satisfied, the plot should show a random pattern (no systematic behaviors) and homogeneous variability



ok!

- no patterns: positive and negative values, randomly spread
- constant dispersion: for all values of $\hat{y}_i$, the $\tilde{e}_i$'s lie approximately between $(-2, 2)$

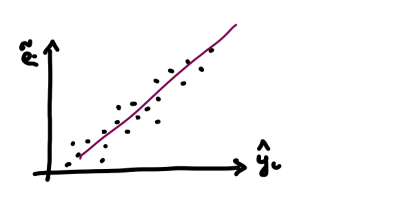


No!

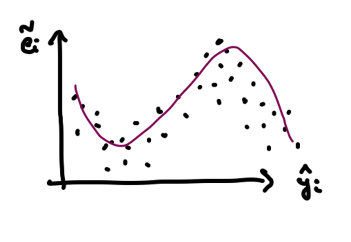
- presence of a SYSTEMATIC BEHAVIOR

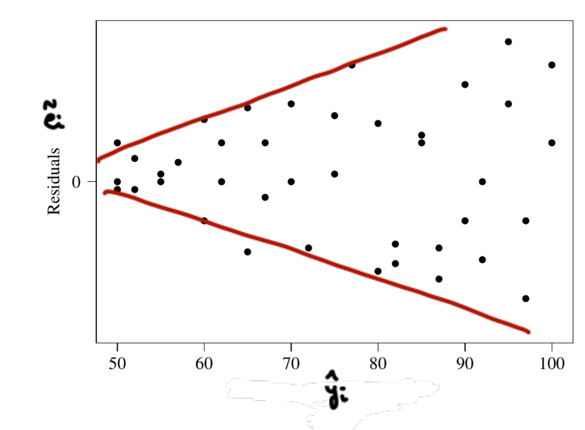
quadratic trend is suggesting that we should include x^2 in the model

other examples of systematic behaviors:



y depends linearly on a covariate not included in the model



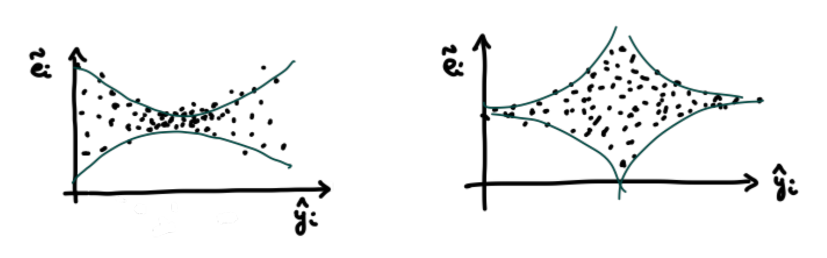


No!

- presence of HETEROSCEDASTICITY

the variance increases with x (or with $\hat{y}$)

other examples of heteroscedasticity



(2) NORMALITY ASSUMPTION

we can use the studentized residuals $R_i \sim N(0,1)$

- histogram of R_i vs. normal density (but it is not so simple to identify deviations)
- empirical cumulative distribution function (ECDF) vs CDF Φ of a $N(0,1)$
- normal Q-Q plot (quantile-quantile plot)

The EMPIRICAL CUMULATIVE DISTRIBUTION FUNCTION

Consider a random variable V with CDF $F_V(t) = \mathbb{P}(V \leq t)$.

Consider a sample $V_1, \dots, V_n$ from V.

We want to estimate the CDF of V based on $(v_1, \dots, v_n)$.

It is reasonable to estimate $F_V(t)$ with the number of observations smaller or equal to t:

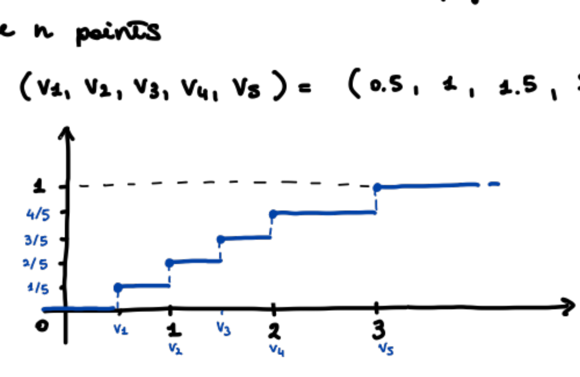
the EMPIRICAL CDF is $\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(V_i \leq t)$

where $\mathbb{1}(V_i \leq t) = \begin{cases} 1 & \text{if } V_i \leq t \\ 0 & \text{if } V_i > t \end{cases}$

$\hat{F}(t)$ is an unbiased estimator of $F_V(t)$

With the sample $(v_1, \dots, v_n)$, we obtain a step function that jumps up by $\frac{1}{n}$ at each of the n points

example: $(v_1, v_2, v_3, v_4, v_5) = (0.5, 1, 1.5, 2, 3)$



With the linear model, we can plot the ECDF of the studentized residuals against the theoretical CDF Φ of a $N(0,1)$

The NORMAL Q-Q PLOT

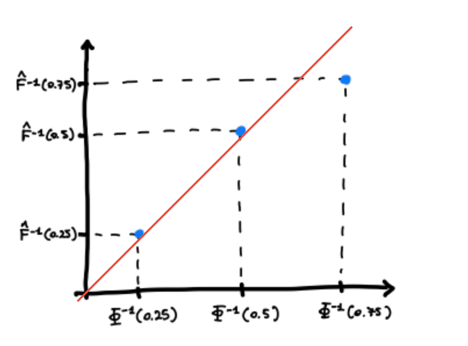
Instead of comparing the empirical and theoretical CDFs, we can compare the EMPIRICAL QUANTILES $\hat{F}^{-1}(\gamma)$ and theoretical quantiles $\Phi^{-1}(\gamma)$ for different values of γ.

These couples of points are then represented in a Q-Q (quantile-quantile) plot.

e.g. $\gamma = (0.25, 0.5, 0.75)$

I compute $\hat{F}^{-1}(0.25)$, $\hat{F}^{-1}(0.5)$, $\hat{F}^{-1}(0.75)$

$\Phi^{-1}(0.25)$, $\Phi^{-1}(0.5)$, $\Phi^{-1}(0.75)$



I plot these couples of points

If $\hat{F}^{-1} = \Phi^{-1}$ the points will be equal

⇒ they will lie on the line $y = x$

(bisector of the first quadrant)

With the linear model:

consider the ORDERED STUDENTIZED RESIDUALS $(r_{(1)}, r_{(2)}, \dots, r_{(n-1)}, r_{(n)})$ (in increasing order)

These are the EMPIRICAL QUANTILES of order $\left(\frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, 1\right)$.

If the assumptions are satisfied, $R_i \sim N(0,1)$.

We compute the theoretical quantiles at the same levels $\left(\frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, 1\right)$: $\Phi^{-1}\left(\frac{i - 1/2}{n}\right)$

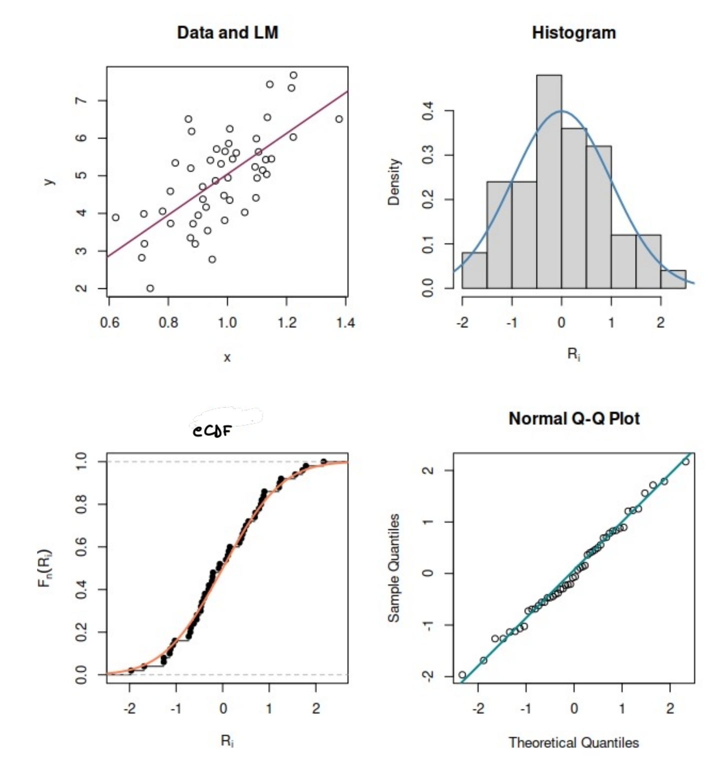
We plot the couples of points $\left(\Phi^{-1}\left(\frac{i-1/2}{n}\right) ; \ r_{(i)} \right)$ for $i = 1, \dots, n$.

we consider $\frac{i - 1/2}{n}$ instead of $\frac{i}{n}$ to avoid $\Phi(1)$ (which is not finite)

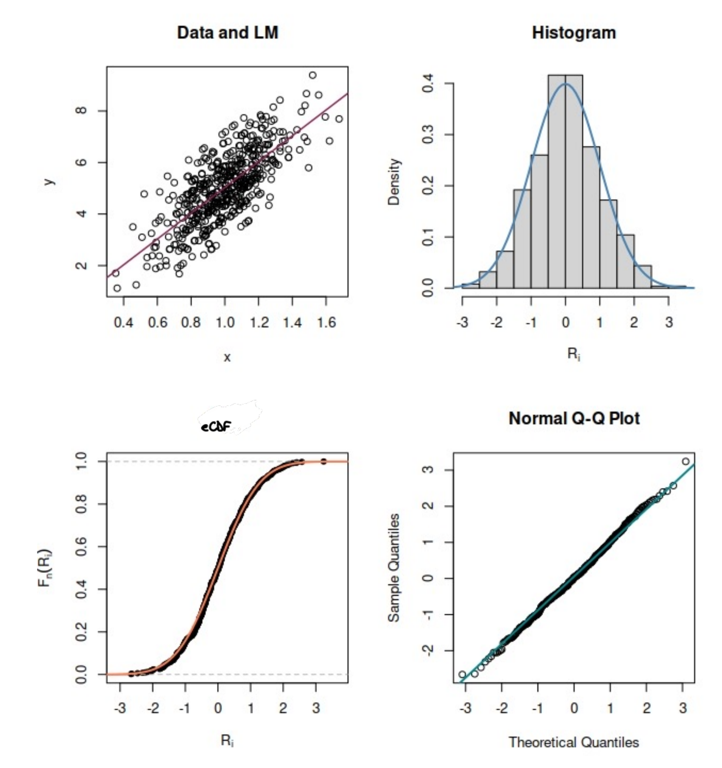
EXAMPLES: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \dots, n$

the normality assumption is satisfied ⇒ $R_i \sim N(0,1)$

$\varepsilon_i \sim N(0, \sigma^2)$, $n = 50$

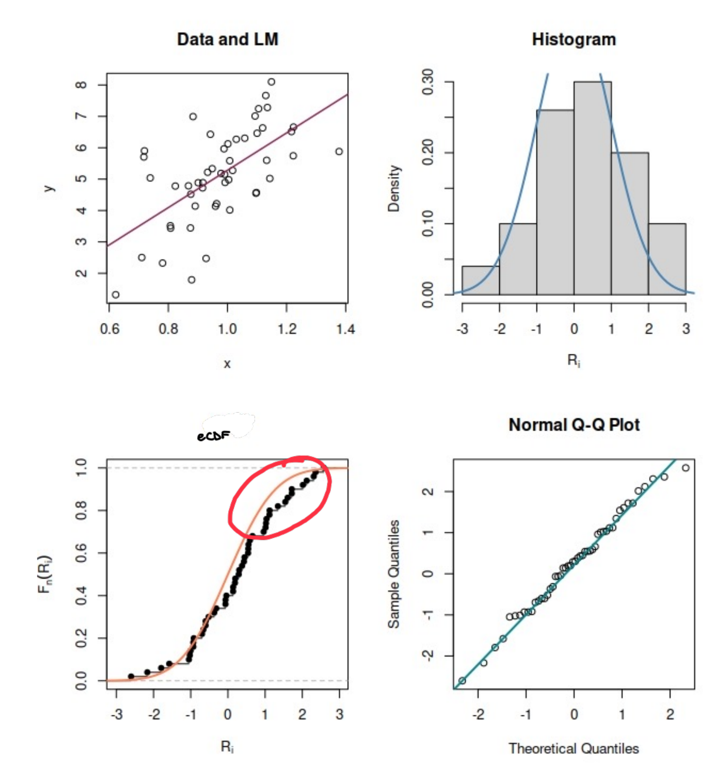


$\varepsilon_i \sim N(0, \sigma^2)$, $n = 500$

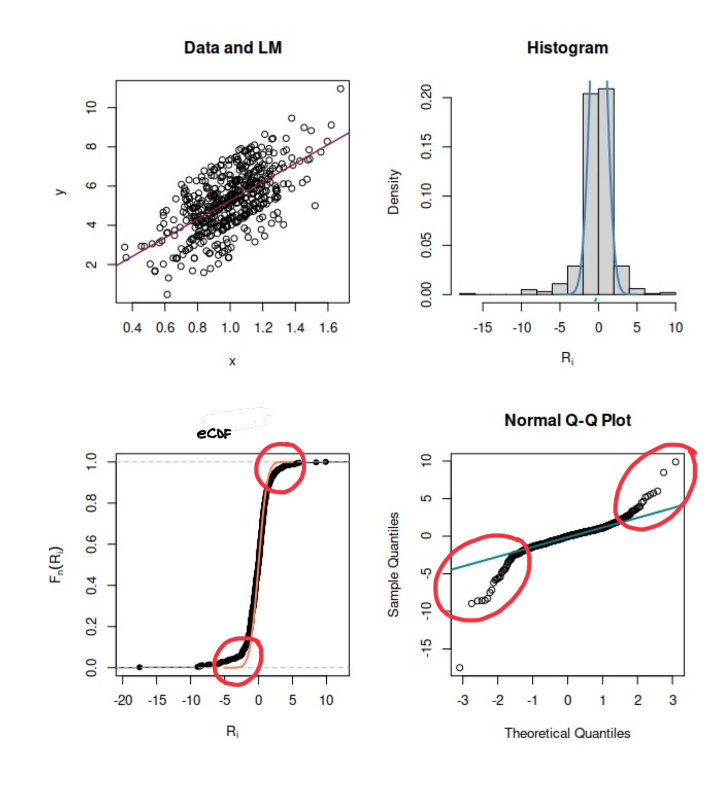


EXAMPLES: the normality assumption is not satisfied

$\varepsilon_i \sim t_2$ and $n = 50$ (t distribution has heavier tails)



$\varepsilon_i \sim t_2$ and $n = 500$



heavier tails