MODEL CHECKING / DIAGNOSTICS

The inference we outlined was performed under the assumption that the hypotheses were met. However, we have to make sure it is the case.

After we fit the model, we need to evaluate the validity of the model.
We should assess whether the model satisfies the underlying assumptions:

1. normality $Y_i \sim N(\mu_i, \sigma^2)$ $i=1,\dots,n$
2. linearity $\mu_i = \beta_1 + \beta_2 x_i$
3. homoscedasticity $var(Y_i) = \sigma^2$ for all $i=1,\dots,n$
4. independence $cov(Y_i, Y_k) = 0$ for $i \neq k$

or, equivalently, $Y_i = \mu_i + \varepsilon_i$ $i=1,\dots,n$; and we formulate the hypotheses on the errors ε_i ($\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$ $i=1,\dots,n$).

Other possible issues to evaluate are:
- is the functional form adequate? The model may be missing needed covariates, or nonlinear transformations of the variables
- is there any outlier? Unusual observations may have too much influence on the model fit.

(we will focus more on these issues with the exercises)

Classical tools to perform the model's diagnostics:
- visual inspection (plots)
- tests

ANALYSIS of RESIDUALS

We make assumptions on the model's error terms ε_i, which are not observable.
However, after we estimate the model, we can compute the RESIDUALS, which are the "analogous" sample quantity (NOT an estimate!).
The assumptions on ε_i have implications on the properties of e_i
$\Rightarrow$ if the properties of e_i do not hold for the estimated model, we conclude that the hypotheses on ε_i were not satisfied.

The residuals are $e_i = y_i - \hat{y}_i$ $i=1,\dots,n$.
We have already showed some properties of e_i:

a) zero mean: $\frac{1}{n}\sum_{i=1}^n e_i = 0$
(see Lecture 2)

b) orthogonality w.r.t. x: $\sum_{i=1}^n x_i e_i = 0$
indeed, $\sum_{i=1}^n x_i e_i = \sum_{i=1}^n x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)$ ← 2^{nd} likelihood equation

c) orthogonality w.r.t. $\hat{y}$: $\sum_{i=1}^n e_i \hat{y}_i = 0$
indeed, $\sum_{i=1}^n e_i \hat{y}_i = \sum_{i=1}^n e_i (\hat{\beta}_1 + \hat{\beta}_2 x_i) = \hat{\beta}_1 \underbrace{\sum_{i=1}^n e_i}_{(a)} + \hat{\beta}_2 \underbrace{\sum_{i=1}^n e_i x_i}_{(b)} = 0$

d) $corr(x, e) = 0$
indeed, $corr(x,e) = 0 \iff cov(x,e) = 0$
$cov(x,e) = \sum_{i=1}^n (e_i - \bar{e})(x_i - \bar{x}) = \underbrace{\sum_{i=1}^n e_i x_i}_{(b)} - \bar{x} \underbrace{\sum_{i=1}^n e_i}_{(a)} = 0$ (with $\bar{e}$ crossed out)

Before observing the data, we have the random variables $E_i = Y_i - \hat{Y}_i$ $i=1,\dots,n$.

DISTRIBUTION of E_i

i. they have normal distribution
$E_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1(Y) - \hat{\beta}_2(Y) x_i = Y_i - \sum_{k=1}^n v_k Y_k - x_i \sum_{k=1}^n w_k Y_k = \sum_{k=1}^n c_k Y_k$
for some constants c_k.
Hence E_i is a linear combination of normal r.v.'s $\Rightarrow E_i \sim N(\cdot,\cdot)$ normal

ii. $\mathbb{E}[E_i] = 0$
$\mathbb{E}[E_i] = \mathbb{E}[Y_i - \hat{Y}_i] = \mathbb{E}[Y_i] - \mathbb{E}[\hat{Y}_i] =$
$= \beta_1 + \beta_2 x_i - \mathbb{E}[\hat{\beta}_1 + \hat{\beta}_2 x_i] = \beta_1 + \beta_2 x_i - \beta_1 - \beta_2 x_i = 0$

iii. $var(E_i) = \sigma^2(1 - h_i)$
with $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2}$ h_i is called "LEVERAGE"
$\Rightarrow$ NOT homoscedastic! (they depend on the index i)
Moreover, they are NOT independent

- Distribution of the residuals: $E_i \sim N(0, \sigma^2(1-h_i))$ $i=1,\dots,n$

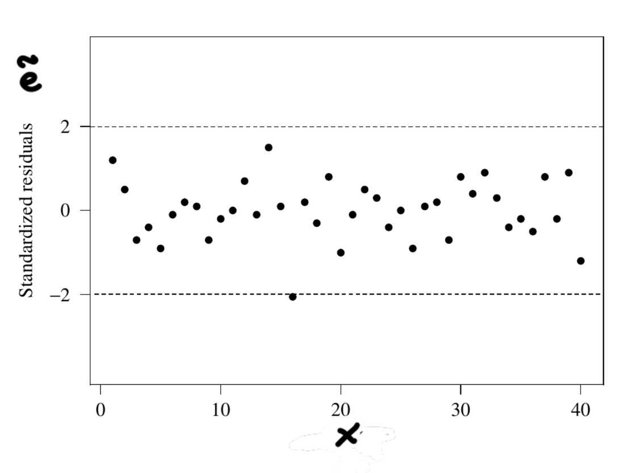
ALTERNATIVE DEFINITIONS:
- Standardized residuals $\tilde{E}_i = \frac{E_i}{\sqrt{1-h_i}}$ with $\mathbb{E}[\tilde{E}_i] = 0$, $var(\tilde{E}_i) = \sigma^2$
$\tilde{E}_i \sim N(0, \sigma^2)$
homoscedastic, but σ^2 is unknown
- Studentized residuals $R_i = \frac{E_i}{\sqrt{S^2(1-h_i)}}$ with $\mathbb{E}[R_i] = 0$, $var(R_i) = 1$
we don't have a nice exact distribution, but approximately $R_i \,\dot\sim\, N(0,1)$

How can we use these quantities to check the model's assumptions?

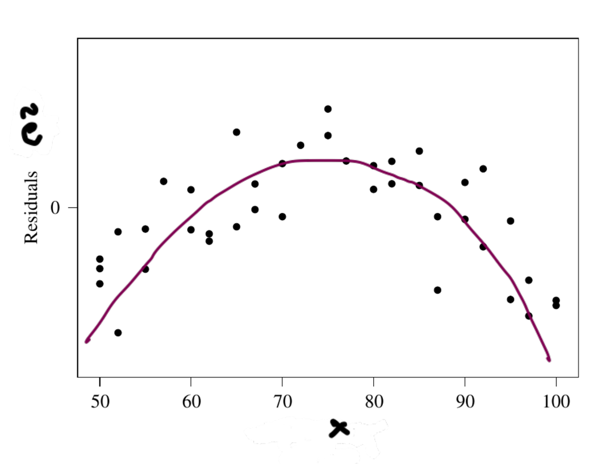
(1) ORTHOGONALITY wrt x and HOMOSCEDASTICITY

RESIDUALS vs COVARIATE → SCATTERPLOT of x_i vs $\tilde{e}_i$ (standardized residuals)
we can use it to check
- the covariate and residuals have correlation = 0
- the standardized residuals are homoscedastic

note: we use standardized or studentized residuals to have constant variance
if the assumption is satisfied, the plot should show a random pattern (no systematic behaviors) and homogeneous variability

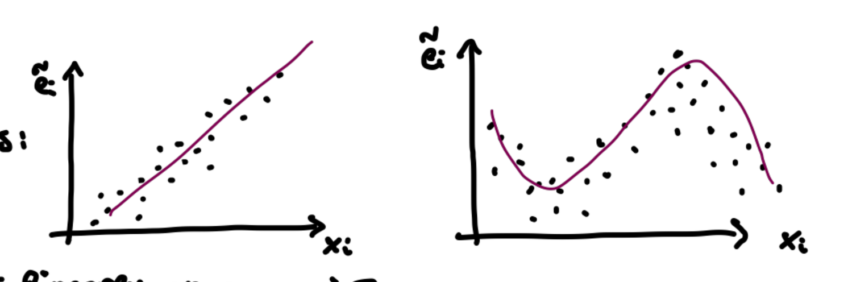


ok! ←
- no patterns
- constant dispersion

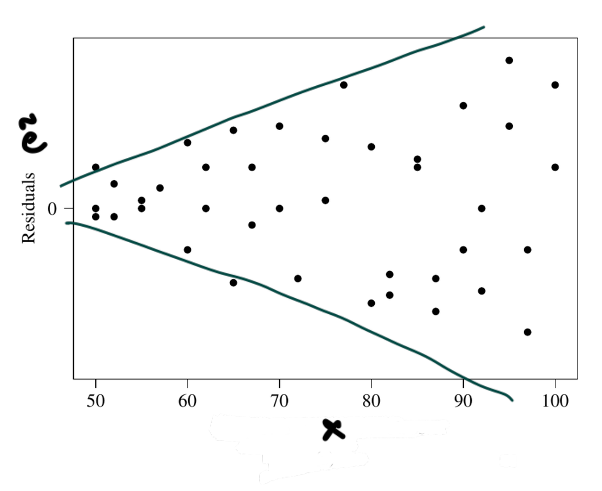


NO! systematic behaviors
quadratic trend is suggesting that we should include x^2 in the model

other examples:

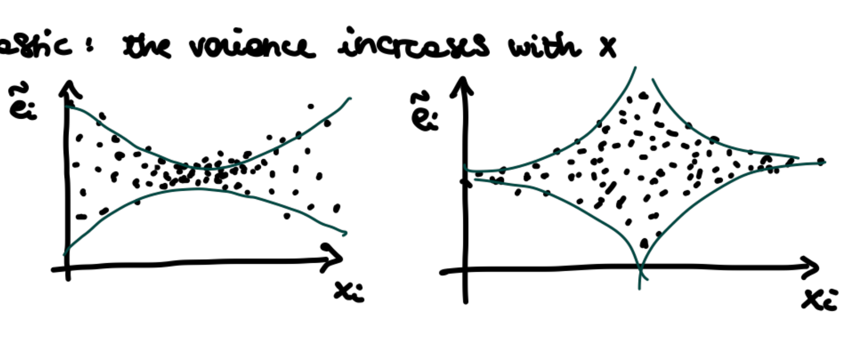


y depends linearly on a covariate not included in the model



NO! heteroscedastic: the variance increases with x

other examples



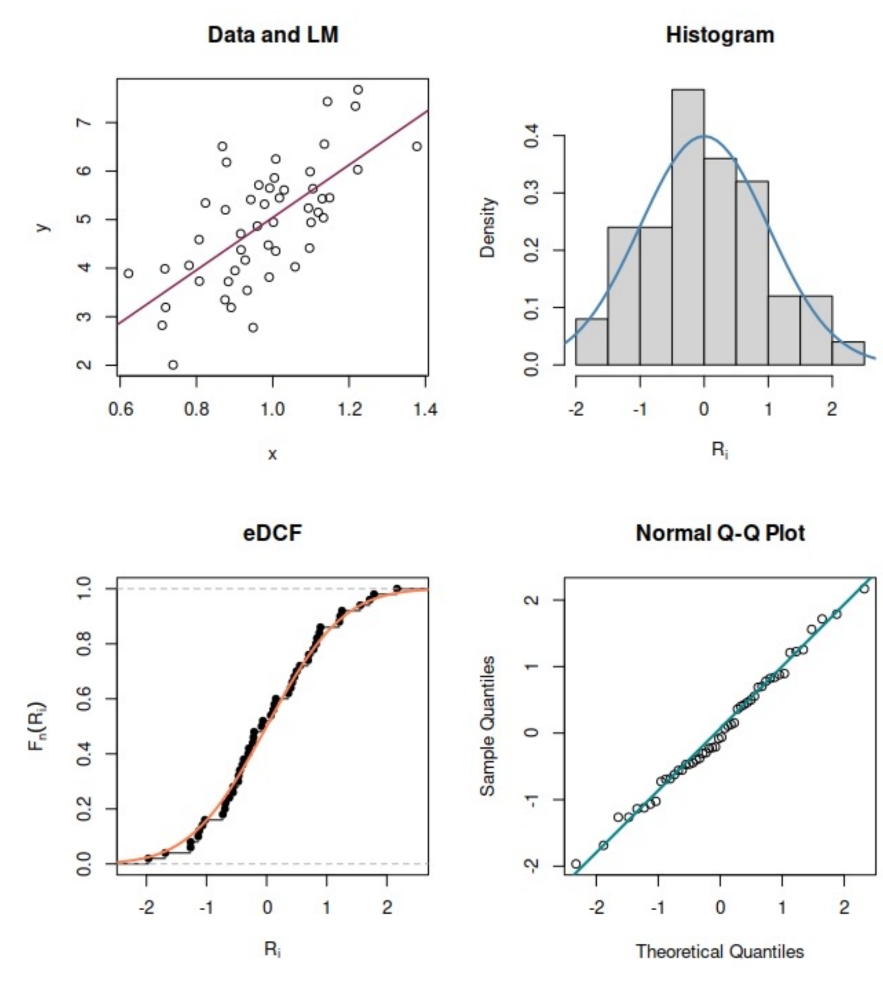
(2) NORMALITY ASSUMPTION

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

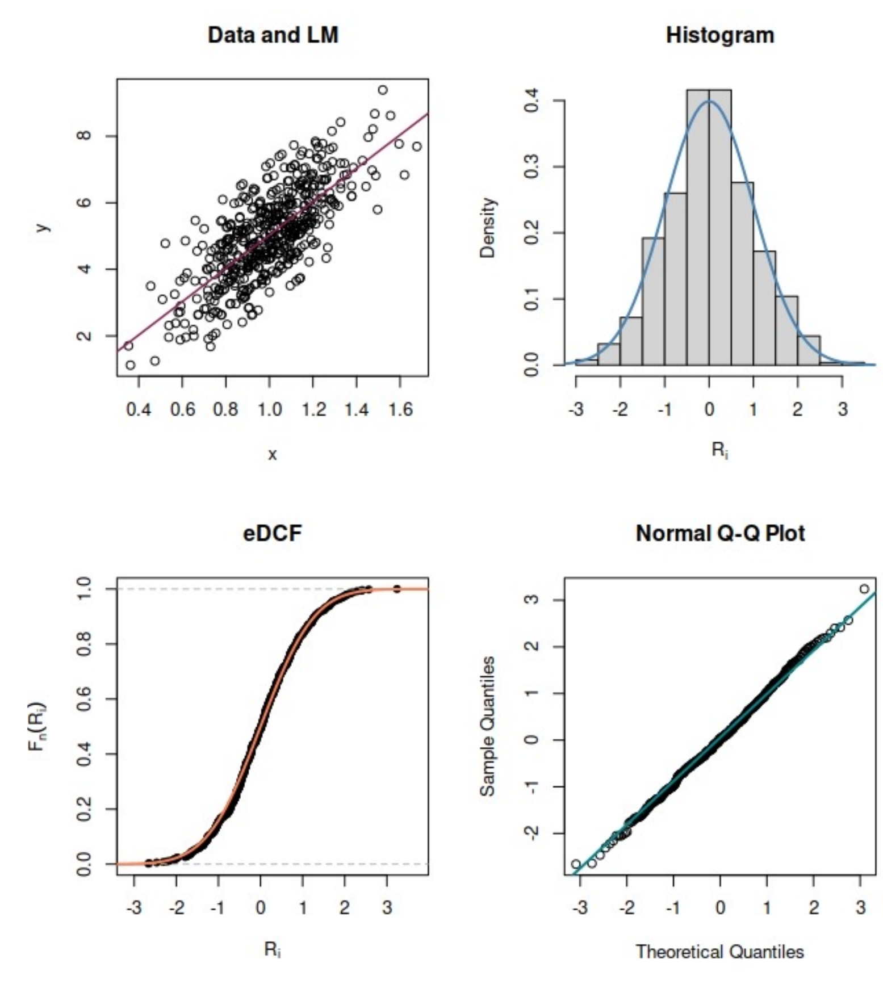
Recall: ECDF. If I observe a sample $u_1,\dots,u_n$ from U with $F_U(t) = \mathbb{P}(U \leq t)$
the empirical CDF is $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(u_i \leq t) = \frac{\text{number of observations} \leq t}{n}$
$\hat{F}_n(t)$ is an unbiased estimator of $F_U(t)$.

Recall: normal Q-Q plot. It is a plot of the empirical quantiles vs. the theoretical quantiles of a $N(0,1)$. If the observed sample comes from a $N(0,1)$ distribution, the points should lie on the straight line $y=x$.
(empirical quantiles = ordered sample $u_{(1)} \leq u_{(2)} \leq \dots \leq u_{(n)}$
theoretical quantiles = Φ^{-1} inverse of the CDF of a $N(0,1)$)

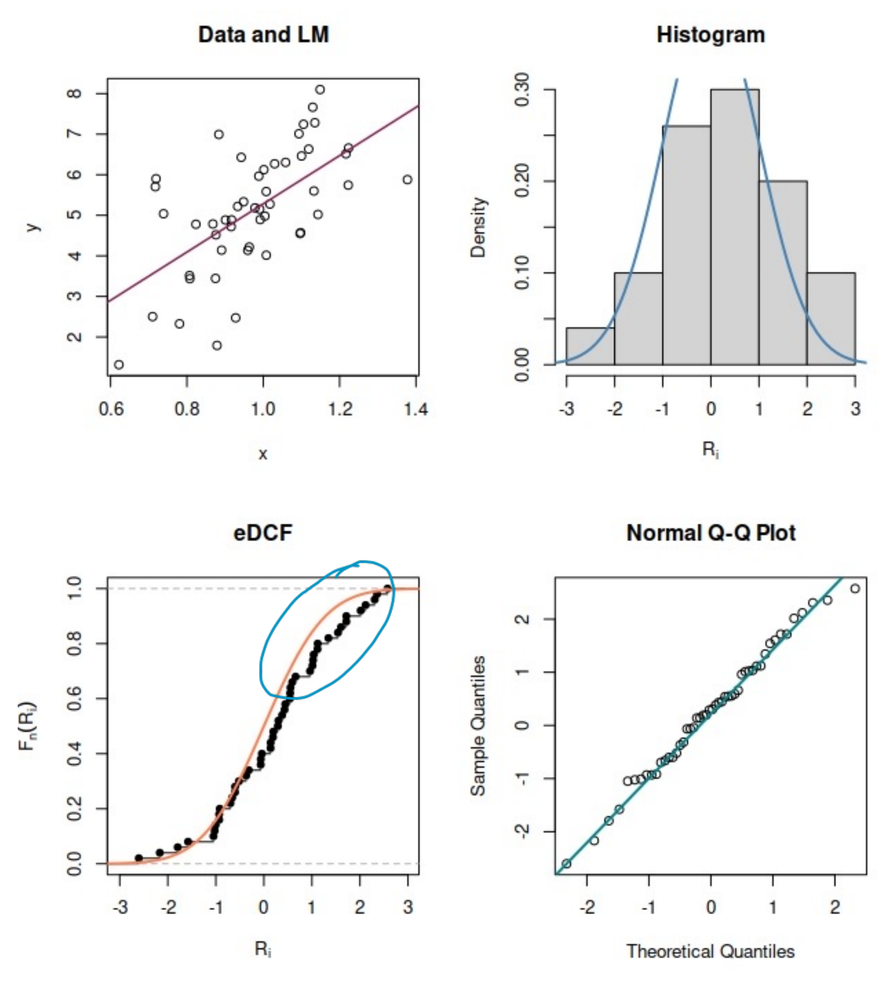
EXAMPLES: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i=1,\dots,n$
the normality assumption is satisfied $\Rightarrow R_i \,\dot\sim\, N(0,1)$
$\varepsilon_i \sim N(0,\sigma^2)$ and $n=50$



$\varepsilon_i \sim N(0,1)$, $n=500$



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 $n=500$