16 Oct 2025

In lec.2 we defined the three phases of the analysis:
1) model specification ✓
2) estimate ✓
3) model checking ← we will now focus on this

MODEL CHECKING / DIAGNOSTICS

The inference we outlined was performed under the assumption that the hypotheses were met. However, we have to make sure this is the case.
After we fit the model, we need to evaluate its validity.
We want to define a set of tools (plots, tests) to evaluate the validity of the assumptions at the basis of the model we have specified and fitted.

We should assess whether the model satisfies the underlying assumptions:
1. normality $Y_i = \mu_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$ $i = 1, \dots, n$
   → we are assuming that $(y_1, \dots, y_n)$ is a sample from some normal distribution
2. linearity $\mu_i = \beta_1 + \beta_2 x_i$
   → we are assuming that Y_i (actually, $\mathbb{E}[Y_i]$) depends on x_i linearly
3. homoscedasticity $\mathrm{var}(\varepsilon_i) = \sigma^2$ for all $i = 1, \dots, n$
4. independence $\mathrm{cov}(\varepsilon_i, \varepsilon_k) = 0$ for $i \neq k$ $i, k = 1, \dots, n$
   → we are assuming that $(y_1, \dots, y_n)$ are generated from independent r.v. with constant variance

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

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

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
⇒ if the properties of the observed residuals are not coherent with the theoretical properties, we conclude that the hypotheses on ε_i are not satisfied by the analyzed data.

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

DESCRIPTIVE PROPERTIES

a) zero mean $\bar{e} = \frac{1}{n}\sum_{i=1}^n e_i = 0$

b) orthogonality w.r.t. x : $\sum_{i=1}^n x_i e_i = 0$

indeed, $\sum_{i=1}^n x_i e_i = \sum_{i=1}^n x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)$ ← 2nd 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) $\mathrm{corr}(x, e) = 0$

indeed, $\mathrm{corr}(x, e) = 0 \Leftrightarrow \mathrm{cov}(x, e) = 0$

$\mathrm{cov}(x, e) = \sum_{i=1}^n (e_i - \cancel{\bar{e}})(x_i - \bar{x}) = \underbrace{\sum_{i=1}^n e_i x_i}_{(b)} - \bar{x} \underbrace{\sum_{i=1}^n e_i}_{(a)} = 0$

INFERENTIAL PROPERTIES

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

$E_i = Y_i - \hat{Y}_i = Y_i - \hat{B}_1 - \hat{B}_2 x_i = Y_i - \sum_{k=1}^n v_k Y_k - x_i \sum_{k=1}^n w_k Y_k = \sum_{k=1}^n c_k Y_k$

for some constants c_k.

Hence E_i is a linear combination of normal r.v.'s ⇒ $E_i \sim N(\cdot, \cdot)$ normal

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

⇒ $\mathbb{E}[E_i] = 0$

iii. $\mathrm{var}(E_i) = \sigma^2 (1 - h_i)$

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"

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

⇒ we have the theoretical distributive properties of the residuals

However, are the assumptions satisfied BY THE DATA ?

Now, we look at the realizations e_i and evaluate whether:
- it is reasonable that $(e_1, \dots, e_n)$ is a sample from a normal distribution with zero mean
- it is reasonable that $(\tilde{e}_1, \dots, \tilde{e}_n)$ is a sample from a normal distribution with zero mean and constant variance
- the residuals are uncorrelated with x and $\hat{y}$