10 oct - LEC 4

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 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, $Y_i = \mu_i + \varepsilon_i$ with $\mu_i = \beta_1 + \beta_2 x_i$ for $i = 1, \dots, n$
and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, \dots, n$.

Other possible issues to evaluate are:
- is the functional form adequate? The model may be missing needed covariates, or nonlinear transformations of the variables
- 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)

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 shown some properties of e_i:

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, $\sum_{i=1}^{n} x_i e_i = \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)$ ← 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, $corr(x, e) = 0 \Leftrightarrow cov(x, e) = 0$
$cov(x, e) = \sum_{i=1}^{n} (e_i - \cancel{\bar{e}})(x_i - \bar{x}) = \underbrace{\sum_{i=1}^{n} e_i x_i}_{(b)} - \bar{x} \underbrace{\sum_{i=1}^{n} e_i}_{(a)} = 0$

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 $\Rightarrow E_i \sim N(\cdot, \cdot)$ normal

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

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 $R_i \,\dot{\sim}\, N(0, 1)$

$\Rightarrow$ we have the theoretical distributive properties of the residuals $E_i = Y_i - \hat{Y}_i$.
Now, we look at the realizations e_i and study their empirical properties.