7 Nov - LEC 8

ONE-WAY ANOVA (Analysis of variance)

In the cuckoo exercise we had 2 groups of observations and we wanted to test whether the means of the two groups were equal (assuming normality and homoscedasticity). In particular, we showed the equivalence between the two-sample t-test and a test of significance on the regression parameter of a simple lm.

Let's generalize the setting and notation

Suppose we are testing the effectiveness of a treatment, and we measure the survival time Y on subjects divided into $G = 3$ groups
The question of interest is whether the mean survival times of the three groups are equal or different. If they are different, then the treatments have different effectiveness.

we can use 2 indices i, g $\quad i = 1, \dots, n_g$ (index of the unit in each group; n_g: number of units in group g)
$\quad g = 1, \dots, G$ (index of the group; G: number of groups)

$$\Rightarrow Y_{ig} \sim N(\mu_g, \sigma^2) \quad \text{independent}$$

Y_{ig}: survival time of individual i from group g
μ_g: group-specific mean of group g
σ^2: common variance for all groups

With 3 groups, for example, we get
- group 1: n_1 individuals $\rightarrow \underline{Y}_1 = [Y_{11}, \dots, Y_{n_1 1}]^T$
- group 2: n_2 individuals $\rightarrow \underline{Y}_2 = [Y_{12}, \dots, Y_{n_2 2}]^T$
- group 3: n_3 individuals $\rightarrow \underline{Y}_3 = [Y_{13}, \dots, Y_{n_3 3}]^T$

Let us denote with μ_g the mean survival time for group g $(g = 1, \dots, G)$
The estimates are
$$\hat{\mu}_g = \bar{y}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} y_{ig}$$

If we want to test equality of the treatments, we test:
$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_1: \text{at least one of them is different} \end{cases}$$

There are several ways to formulate a linear model for this problem.
Here, we only consider the case where we have the intercept $(\underline{x}_1 = \underline{1})$

First, we define the vector of the response by concatenating each group-specific vector $\underline{Y}_g$
$$\underline{Y} = [\underline{Y}_1^T \ \underline{Y}_2^T \ \underline{Y}_3^T]^T = [\underbrace{Y_{11}, Y_{21}, \dots, Y_{n_1 1}}_{\text{group 1}}, \underbrace{Y_{12}, \dots, Y_{n_2 2}}_{\text{group 2}}, \underbrace{Y_{13}, \dots, Y_{n_3 3}}_{\text{group 3}}]^T$$
vector with $N = n_1 + n_2 + n_3$ elements.

Then, we define the matrix X of the covariates
We use DUMMY VARIABLES where
$$x_{ig} = \begin{cases} 1 & \text{if individual } i \text{ belongs to group } g \\ 0 & \text{otherwise} \end{cases}$$
for $i = 1, \dots, n_g$ and $g = 1, \dots, G$.

Remark:
consider $G = 3$. If we define the matrix X as
$$X = [\underline{1} \ \underline{x}_1 \ \underline{x}_2 \ \underline{x}_3] = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix}$$
$\rightarrow \underline{1} = \underline{x}_1 + \underline{x}_2 + \underline{x}_3$
multicollinearity!
rank (X) = 3 < 4 (n. columns)

intercept; indicator of group 1; indicator of group 2; indicator of group 3

To encode G groups, if we keep the intercept, we only need G-1 dummy variables.

Consider removing $\underline{x}_1$. Then X becomes
$$X = [\underline{1} \ \underline{x}_2 \ \underline{x}_3] = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{array}{l} \left.\vphantom{\begin{matrix}1\\1\\1\\1\end{matrix}}\right\} n_1 \text{ obs.} \\ \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} n_2 \text{ obs.} \\ \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} n_3 \text{ obs.} \end{array} \quad (N \times G) \text{ matrix}$$

We can define a linear model with these quantities
$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \quad \text{with} \quad \underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$
$$\text{and} \quad \underline{\varepsilon} \sim N_N(\underline{0}, \sigma^2 I_N)$$

Let's study the expected value of observations in each group according to the model
$$\mathbb{E}[Y_{i1}] = \beta_1 \quad \text{for } i = 1, \dots, n_1$$
$$\mathbb{E}[Y_{i2}] = \beta_1 + \beta_2 \quad \text{for } i = 1, \dots, n_2$$
$$\mathbb{E}[Y_{i3}] = \beta_1 + \beta_3 \quad \text{for } i = 1, \dots, n_3$$

INTERPRETATION:
- INTERCEPT: β_1 is the mean of Y_{ig} when $g = 1$ (when all dummy variables are equal to zero)
 (mean of the group for which we removed the dummy variable)
 This group is said to be the REFERENCE GROUP: it is the BASELINE
 A classical example is the control group (i.e. the "no treatment") in clinical trials.
 $\Rightarrow \beta_1 = \mathbb{E}[Y_{i1}]$

The other groups are described in terms of DEVIATION FROM THE BASELINE.
- β_2 is the difference in the mean of Y_{i2} w.r.t. the mean of Y_{i1}
 Indeed from the model we have
 $\mathbb{E}[Y_{i2}] = \beta_1 + \beta_2$
 $\Rightarrow \beta_2 = \mathbb{E}[Y_{i2}] - \mathbb{E}[Y_{i1}]$
 $= \mu_2 - \mu_1$
- β_3 is the difference in the mean of Y_{i3} w.r.t. the mean of Y_{i1}
 $\mathbb{E}[Y_{i3}] = \beta_1 + \beta_3$
 $\Rightarrow \beta_3 = \mathbb{E}[Y_{i3}] - \mathbb{E}[Y_{i1}]$
 $= \mu_3 - \mu_1$

Remark: we automatically get the estimates of the regression parameters:
Reparametrization
$$\begin{cases} \mu_1 = \beta_1 \\ \mu_2 = \beta_1 + \beta_2 \\ \mu_3 = \beta_1 + \beta_3 \end{cases} \iff \begin{cases} \beta_1 = \mu_1 \\ \beta_2 = \mu_2 - \mu_1 \\ \beta_3 = \mu_3 - \mu_1 \end{cases}$$

Invariance of the MLE w.r.t. reparametrizations
$$\begin{cases} \hat{\beta}_1 = \hat{\mu}_1 \\ \hat{\beta}_2 = \hat{\mu}_2 - \hat{\mu}_1 \\ \hat{\beta}_3 = \hat{\mu}_3 - \hat{\mu}_1 \end{cases} \Rightarrow \begin{cases} \hat{\beta}_1 = \bar{y}_1 \\ \hat{\beta}_2 = \bar{y}_2 - \bar{y}_1 \\ \hat{\beta}_3 = \bar{y}_3 - \bar{y}_1 \end{cases}$$

We can easily compute the predicted values $\hat{y}_{ig}$
$$\hat{y}_{i1} = \hat{\beta}_1 = \bar{y}_1 \quad i = 1, \dots, n_1$$
$$\hat{y}_{i2} = \hat{\beta}_1 + \hat{\beta}_2 = \bar{y}_1 + \bar{y}_2 - \bar{y}_1 = \bar{y}_2 \quad i = 1, \dots, n_2$$
$$\hat{y}_{i3} = \hat{\beta}_1 + \hat{\beta}_3 = \bar{y}_1 + \bar{y}_3 - \bar{y}_1 = \bar{y}_3 \quad i = 1, \dots, n_3$$
$\Rightarrow$ The predicted values are the group-specific means.

Finally, the test about equality of the group-specific means becomes
$$\begin{cases} H_0: \beta_2 = \beta_3 = 0 \\ H_1: \text{at least one is } \neq 0 \end{cases}$$
$\rightarrow$ test about the significance of the model
$=$ test about the equality of the treatments on the survival time

SUM OF SQUARES DECOMPOSITION

Consider G groups, and n_g observations in each group.
$$Y_{ig} \sim N(\mu_g, \sigma^2) \text{ independent for } i = 1, \dots, n_g \text{ and } g = 1, \dots, G$$
Let $N = \sum_{g=1}^{G} n_g$ total sample size
- The group-specific means are
$$\bar{y}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} y_{ig} \quad g = 1, \dots, G$$
- The overall mean is
$$\bar{y} = \frac{1}{N} \sum_{g=1}^{G} \sum_{i=1}^{n_g} y_{ig} = \frac{1}{N} \sum_{g=1}^{G} n_g \bar{y}_g$$
- The group-specific estimates of the variance are
$$s_g^2 = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (y_{ig} - \bar{y}_g)^2 \quad g = 1, \dots, G$$

The partition of the sum of squares in the linear model was $\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)$
We can specify it for this setting:
The total sum of squares here is $SST = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y})^2$

$$\sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y})^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y}_g + \bar{y}_g - \bar{y})^2$$
$$= \sum_{g=1}^{G} \sum_{i=1}^{n_g} [(y_{ig} - \bar{y}_g)^2 + (\bar{y}_g - \bar{y})^2 + 2(y_{ig} - \bar{y}_g)(\bar{y}_g - \bar{y})]$$
$$= \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y}_g)^2 + \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\bar{y}_g - \bar{y})^2 + 2 \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\bar{y}_g - \bar{y})(y_{ig} - \bar{y}_g)$$
$$= \sum_{g=1}^{G} \underbrace{\sum_{i=1}^{n_g} (y_{ig} - \bar{y}_g)^2}_{= (n_g - 1) s_g^2} + \sum_{g=1}^{G} n_g (\bar{y}_g - \bar{y})^2 + \cancel{2 \sum_{g=1}^{G} (\bar{y}_g - \bar{y}) \underbrace{\sum_{i=1}^{n_g} (y_{ig} - \bar{y}_g)}_{= 0}}$$
$$= \sum_{g=1}^{G} (n_g - 1) s_g^2 + \sum_{g=1}^{G} n_g (\bar{y}_g - \bar{y})^2$$

Hence we get
$$\underbrace{\sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y})^2}_{\text{TOTAL SUM OF SQUARES}} = \underbrace{\sum_{g=1}^{G} (n_g - 1) s_g^2}_{\text{WITHIN-GROUP VARIABILITY}} + \underbrace{\sum_{g=1}^{G} n_g (\bar{y}_g - \bar{y})^2}_{\text{BETWEEN-GROUP VARIABILITY}}$$

Total sum of squares: deviations of each observation from the overall mean
Within-group sum of squares: deviations of each observation from the corresponding group-specific mean
Between-group sum of squares: deviations of each group-specific mean from the overall mean

Moreover, we have seen that $\bar{y}_g = \hat{y}_{ig}$ for $i = 1, \dots, n_g$
that is, the predicted values are the group-specific means
Thus
$$\sum_{g=1}^{G} (n_g - 1) s_g^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y}_g)^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \hat{y}_{ig})^2 \quad \text{ERROR SUM OF SQUARES}$$
$$\sum_{g=1}^{G} n_g (\bar{y}_g - \bar{y})^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\bar{y}_g - \bar{y})^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\hat{y}_{ig} - \bar{y})^2 \quad \text{REGRESSION SUM OF SQUARES}$$

TEST ABOUT EQUALITY OF THE MEANS

Testing equality of the means is equivalent to testing
$$\begin{cases} H_0: \beta_2 = \beta_3 = \dots = \beta_G = 0 \\ H_1: \overline{H_0} \end{cases} \quad \text{test about the overall significance}$$

We used $\quad F = \dfrac{\tilde{\Sigma}^2 - \hat{\Sigma}^2}{\hat{\Sigma}^2} \cdot \dfrac{N - G}{G - 1} \overset{H_0}{\sim} F_{G-1, N-G}$

What are $\tilde{\sigma}^2$ and $\hat{\sigma}^2$ here?
- $\tilde{\sigma}^2$ estimate under H_0: model $\underline{Y} = \beta_1 \cdot \underline{1} + \underline{\varepsilon} \Rightarrow \hat{\beta}_1 = \bar{y}$ overall mean
 Hence, $\tilde{\sigma}^2 = \dfrac{1}{N} \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y})^2 = \dfrac{SST}{N}$
- $\hat{\sigma}^2$ estimate under H_1: $\hat{\sigma}^2 = \dfrac{1}{N} \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y}_g)^2 = \dfrac{SSE}{N}$

Hence the test statistic becomes
$$F = \frac{\tilde{\Sigma}^2 - \hat{\Sigma}^2}{\hat{\Sigma}^2} \cdot \frac{N-G}{G-1} =$$
$$= \frac{SST - SSE}{SSE} \cdot \frac{N-G}{G-1} =$$
$$= \frac{SSR}{SSE} \cdot \frac{N-G}{G-1} =$$
$$= \frac{\text{BETWEEN-GROUP S.S.}}{\text{WITHIN-GROUP S.S.}} \cdot \frac{N-G}{G-1} =$$
$$= \frac{R^2}{1 - R^2} \cdot \frac{N-G}{G-1} \overset{H_0}{\sim} F_{G-1, N-G}$$