NOTABLE EXAMPLES

ANOVA (Analysis of Variance)

- TWO-SAMPLE PROBLEM

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. Now we are going to generalize the procedure of comparing the means of several groups using the linear model.

Suppose we are testing the effectiveness of a treatment, and we measure the survival time Y on subjects divided into 2 groups:

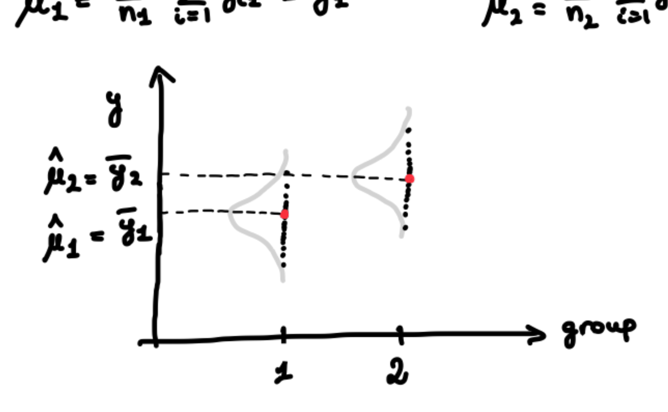
- group 1: n_1 individuals
- group 2: n_2 individuals

The question of interest is whether the mean survival time of the two groups are equal or different. If they are different, then the two treatments have different effectiveness.

we can use 2 indices i, g $\quad i = 1, \dots, n_g$ # unit $\quad g = 1, 2$ # group $\Rightarrow Y_{ig} \sim N(\mu_g, \sigma^2)$ independent

Let us denote with μ_1 the mean survival time for group 1, and with μ_2 the mean survival for group 2. The estimates are simply

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} y_{i1} = \bar{y}_1 \qquad \hat{\mu}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} y_{i2} = \bar{y}_2$$



If we want to test the effectiveness of the treatment, we test

$H_0: \mu_1 = \mu_2$ (no effect) vs $H_1: \mu_1 \neq \mu_2$

We can write this model as a linear model in several ways:

$$Y_{ig} = \begin{cases} \mu_1 + \varepsilon_{ig} & \text{if } g = 1 \\ \mu_2 + \varepsilon_{ig} & \text{if } g = 2 \end{cases} \qquad \varepsilon_{ig} \overset{iid}{\sim} N(0, \sigma^2) \qquad \begin{matrix} g = 1, 2 \\ i = 1, \dots, n_g \end{matrix}$$

But we can also write it in a more compact way as

$\underline{Y} = X\underline{\mu} + \underline{\varepsilon}$ where $\underline{Y} = (Y_{11}, \dots, Y_{n_1 1}, Y_{12}, \dots, Y_{n_2 2})^T$

$\underline{\mu} = (\mu_1, \mu_2)^T$

$\underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 I_n)$

X $(N \times 2)$ matrix $N = n_1 + n_2$

$$X = \begin{array}{c} \begin{matrix} \underline{x}_1 & \underline{x}_2 \end{matrix} \\ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \hline 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \end{array} \begin{matrix} \left.\vphantom{\begin{matrix}1\\1\\1\\1\end{matrix}}\right\} n_1 \text{ rows} \\ \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} n_2 \text{ rows} \end{matrix}$$

indeed, $\mathbb{E}[\underline{Y}] = X\underline{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_1 \\ \mu_2 \\ \vdots \\ \mu_2 \end{bmatrix} \begin{matrix} \}\, n_1 \\ \\ \}\, n_2 \end{matrix}$ and the estimates are $\hat{\mu}_1 = \bar{y}_1 \quad \hat{\mu}_2 = \bar{y}_2$

- an equivalent formulation

what if we wanted to include the intercept?

If we consider $X = [\underline{1}_n, \underline{x}_1, \underline{x}_2]$: not a good choice: $\underline{1}_n = \underline{x}_1 + \underline{x}_2$ (collinearity)

→ if we want the intercept, we have to remove either $\underline{x}_1$ or $\underline{x}_2$

if we write $\mu_2 = \mu_1 + \delta \Rightarrow \delta = \mu_2 - \mu_1$ difference of the means.

$\mathbb{E}[Y_{i1}] = \mu_1$

$\mathbb{E}[Y_{i2}] = \mu_1 + \delta$

How do we define a linear model with this parametrization?

$\underline{Y} = (Y_{11}, \dots, Y_{n_1 1}, Y_{12}, \dots, Y_{n_2 2})^T$

$$X = [\underline{1}_n, \underline{x}_2] = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \hline 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \qquad \underline{\beta} = \begin{bmatrix} \beta_1 \\ \delta \end{bmatrix}$$

$\Rightarrow \underline{Y} = X\underline{\beta} + \underline{\varepsilon}$

INTERPRETATION

- INTERCEPT: β_1 is the mean of y when $x_2 = 0$ $\Rightarrow$ mean of group 1
  Group 1 is the BASELINE since the mean of group 2 is defined in terms of deviation from $\beta_1 = \mu_1$.
- PARAMETER δ: is the difference of the mean of group 2 with respect to group 1.
  $\mathbb{E}[Y_{i2}] = \beta_1 + \delta$

Hence the interpretation of the regression parameters is different when the covariates are not quantitative variables.

suppose we have now $G = 4$ groups

The interest is again testing the efficacy of different treatments using the LM:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1: \overline{H_0} \end{cases}$$

How do we code X?

example for $G = 4$, and group 1 as baseline:

$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \qquad \underline{\beta} = [\beta_1\ \beta_2\ \beta_3\ \beta_4]^T$

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} \}\, n_1 \\ \\ \\ \}\, n_2 \\ \\ \\ \}\, n_3 \\ \\ \\ \}\, n_4 \end{matrix}$$

group 1 $\quad \mathbb{E}[Y_i] = \beta_1 = \mu_1$

group 2 $\quad \mathbb{E}[Y_i] = \beta_1 + \beta_2 = \mu_2$

group 3 $\quad \mathbb{E}[Y_i] = \beta_1 + \beta_3 = \mu_3$

group 4 $\quad \mathbb{E}[Y_i] = \beta_1 + \beta_4 = \mu_4$

β_1: mean of the baseline group

β_j $(j = 2, 3, 4)$: difference of the mean of y in group j compared to the baseline

Hence the hypotesis becomes

$$\begin{cases} H_0: \beta_2 = \beta_3 = \beta_4 = 0 \\ H_1: \overline{H_0} \end{cases}$$

we can automatically obtain the estimates of $\underline{\beta}$ (reparameterization)

$\hat{\mu}_1 = \bar{y}_1 \quad \Rightarrow \quad \hat{\beta}_1 = \hat{\mu}_1 = \bar{y}_1$

$\hat{\mu}_2 = \bar{y}_2 \quad \Rightarrow \quad \hat{\beta}_2 = \hat{\mu}_2 - \hat{\beta}_1 = \bar{y}_2 - \bar{y}_1$

$\hat{\mu}_3 = \bar{y}_3 \quad \Rightarrow \quad \hat{\beta}_3 = \hat{\mu}_3 - \hat{\beta}_1 = \bar{y}_3 - \bar{y}_1$

$\hat{\mu}_4 = \bar{y}_4 \quad \Rightarrow \quad \hat{\beta}_4 = \hat{\mu}_4 - \hat{\beta}_1 = \bar{y}_4 - \bar{y}_1$

The predicted values are $\hat{y}_{ig} = \begin{cases} \hat{\beta}_1 = \bar{y}_1 & \text{for } g = 1 \\ \hat{\beta}_1 + \hat{\beta}_g = \bar{y}_1 + \bar{y}_g - \bar{y}_1 = \bar{y}_g & \text{for } g = 2, 3, 4 \end{cases}$

ANOVA with G GROUPS

consider G groups and n_g observations for each group $(g = 1, \dots, G)$.

$Y_{ig} \sim N(\mu_g, \sigma^2)$ independent $\quad i = 1, \dots, n_g \quad g = 1, \dots, G$

and the test $\begin{cases} H_0: \mu_1 = \mu_2 = \dots = \mu_G \\ H_1: \overline{H_0} \end{cases}$

The group-specific means are $\bar{y}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} y_{ig}$

The overall mean is $\bar{y} = \frac{1}{N} \sum_{g=1}^{G} \sum_{i=1}^{n_g} y_{ig} = \frac{1}{N} \sum_{g=1}^{G} n_g \bar{y}_g \qquad N = \sum_{g=1}^{G} n_g$

The group-specific estimate of the variance is $s_g^2 = \frac{1}{(n_g - 1)} \sum_{i=1}^{n_g} (y_{ig} - \bar{y}_g)^2$

The total sum of squares can be partitioned into two parts

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

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

Indeed, $\sum_{g=1}^{G} (n_g - 1) s_g^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y}_g)^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \hat{y}_{ig})^2$ ERROR sum of squares

$\sum_{g=1}^{G} n_g (\bar{y}_g - \bar{y})^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\bar{y}_g - \bar{y})^2 = \sum_{g=1}^{G} \sum_{i=1}^{n} (\hat{y}_{ig} - \bar{y})^2$ REGRESSION sum of squares

Similarly to the previous example we can express this problem as a lm with

$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \qquad \varepsilon \sim N_n(\underline{0}, \sigma^2 I_n)$

$X = [\underline{1}_n\ \underline{x}_2\ \underline{x}_3\ \dots\ \underline{x}_G]$ where $x_{ig} = \begin{cases} 1 & \text{if } y_{ig} \text{ belongs to group } g \quad (g = 2, \dots, G) \\ 0 & \text{otherwise} \end{cases}$

Then $\hat{\beta}_1 = \bar{y}_1$

and $\hat{\beta}_g = \bar{y}_g - \bar{y}_1$

Testing equality of the means is equivalent to testing

$\begin{cases} H_0: \beta_2 = \beta_3 = \dots = \beta_G = 0 \\ H_1: \overline{H_0} \end{cases}$ test about the overall significance

We used $F = \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \cdot \frac{n - G}{G - 1} \overset{H_0}{\sim} F_{G-1, n-G}$

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

$\tilde{\sigma}^2 = \frac{1}{N} \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y})^2 = \frac{SST}{N}$

$\hat{\sigma}^2$ estimate under H_1: $\hat{\sigma}^2 = \frac{1}{N} \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y}_g)^2 = \frac{SSE}{N}$

$$\begin{aligned} \Rightarrow F = \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \cdot \frac{n - G}{G - 1} &= \frac{SST - SSE}{SSE} \cdot \frac{n - G}{G - 1} = \\ &= \frac{SSR}{SSE} \cdot \frac{n - G}{G - 1} = \\ &= \frac{\text{BETWEEN group variability}}{\text{WITHIN group variability}} \cdot \frac{n - G}{G - 1} \end{aligned}$$