Two-sample t-test and MLR

- In this and the next few lectures, we will look at the relationships between MLR and some common statistical procedures.
- We will start with the pooled t-test
- Two sample t-test is used for comparing the mean of normal populations
- The setup of two-sample pooled t-test can be formulated as a MLR model

Two-sample pooled t-test

Consider independent observations

from:

Sample 1: $y_{11}, y_{12}, \dots, y_{1n_1}$ $N(\mu_1, \sigma^2)$

Sample 2: $y_{21}, y_{22}, \dots, y_{2n_2}$ $N(\mu_2, \sigma^2)$

with

$$Y_{ij} \sim N(\mu_i, \sigma^2) \text{ for } j = 1, 2, \dots, n_i; \mathrm{i} = 1, 2$$

$Y_{ij} = \mu_i + \varepsilon_i$ where $\varepsilon_i \sim$ iid $N(0, \sigma^2)$

We want to make inference about the difference between μ_2 and μ_1.

Let $\delta = \mu_2 - \mu_1$.

$H_0: \delta = 0$ vs $H_a: \delta \neq 0$

Set as a MLR model

$$Y = X\beta + \varepsilon$$

Sample 1

Sample 2

$$\boldsymbol{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \mu_1 \\ \mu_2 - \mu_1 \end{bmatrix}$$

$$X\beta = \begin{bmatrix} \mu_1 \\ \mu_1 \\ \vdots \\ \mu_1 \\ \cancel{\mu_1} + \mu_2 - \cancel{\mu_1} \\ \mu_2 \\ \vdots \\ \mu_2 \end{bmatrix}$$

Remarks:

1. It can be proved that the estimate ($\widehat{\boldsymbol{\beta}}$), standard error, hypothesis test, and confidence interval obtained from the multiple linear regression (MLR) setup are identical to the expressions we previously derived for the pooled t-test.

2. This also confirms that $\bar{Y}_2 - \bar{Y}_1$ is the BLUE for $\mu_2 - \mu_1$.

3. The two-sample t-test is a special case of MLR.

The estimate of $\boldsymbol{\beta}$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} = \boxed{\begin{bmatrix} \bar{Y}_{1\cdot} \\ \bar{Y}_{2\cdot} - \bar{Y}_{1\cdot} \end{bmatrix}}$$

where

$$\bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

Sample mean of Sample i, i=1,2

① $X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} n_1+n_2 & n_2 \\ n_2 & n_2 \end{bmatrix}$

n_1 n_2

$|X^T X| = (n_1+n_2)n_2 - n_2^2 = n_2(n_1 + \cancel{n_2} - \cancel{n_2}) = n_1 n_2$

$(X^T X)^{-1} = \frac{1}{n_1 n_2}\begin{bmatrix} n_2 & -n_2 \\ -n_2 & n_1+n_2 \end{bmatrix}$

② $$X^TY = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{2} \sum_{j=1}^{n_i} y_{ij} \\ \sum_{j=1}^{n_2} y_{2j} \end{bmatrix}$$

③ $$\hat{\beta} = (X^TX)^{-1} X^TY = \frac{1}{n_1 n_2} \begin{bmatrix} n_2 & -n_2 \\ -n_2 & n_1+n_2 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{2} \sum_{j=1}^{n_i} y_{ij} \\ \sum_{j=1}^{n_2} y_{2j} \end{bmatrix}$$

$$= \frac{1}{n_1 n_2} \begin{bmatrix} n_2 \sum_{i=1}^{2} \boxed{\sum_{j=1}^{n_i} y_{ij}} - \boxed{n_1 \sum_{j=1}^{n_2} y_{2j}} \\ -n_2 \sum_{i=1}^{2} \boxed{\sum_{j=1}^{n_i} y_{ij}} + (n_1 + \boxed{n_2}) \boxed{\sum_{j=1}^{n_2} y_{2j}} \end{bmatrix}$$

$$= \frac{1}{n_1 n_2} \begin{bmatrix} n_2 \sum_{j=1}^{n_1} y_{1j} \\ n_1 \sum_{j=1}^{n_2} y_{2j} - n_2 \sum_{j=1}^{n_1} y_{1j} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} \\ \boxed{\frac{1}{n_2} \sum_{j=1}^{n_2} y_{2j}} - \boxed{\frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j}} \end{bmatrix}$$

$$= \begin{bmatrix} \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} - \bar{y}_{1\cdot} \end{bmatrix}$$

The residual variance S_e^2

$$S_e^2 = \frac{1}{n-p}\left\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right\|^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}. = S_p^2$$

$$X\hat{\beta} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} - \bar{y}_{1\cdot} \end{bmatrix} = \begin{bmatrix} \bar{y}_{1\cdot} \\ \bar{y}_{1\cdot} \\ \vdots \\ \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \bar{y}_{2\cdot} \\ \vdots \\ \bar{y}_{2\cdot} \end{bmatrix}$$

$$Y - X\hat{\beta} = \begin{bmatrix} y_{11} - \bar{y}_{1\cdot} \\ y_{12} - \bar{y}_{1\cdot} \\ \vdots \\ y_{1n_1} - \bar{y}_{1\cdot} \\ y_{21} - \bar{y}_{2\cdot} \\ \vdots \\ y_{2n_2} - \bar{y}_{2\cdot} \end{bmatrix}$$

$$\|Y - X\hat{\beta}\|^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 = \sum_{j=1}^{n_1}(y_{1j} - \bar{y}_{1\cdot})^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_{2\cdot})^2 = (n_1-1)S_1^2 + (n_2-1)S_2^2$$

$n = n_1 + n_2$

$p = 2$

$n - p = n_1 + n_2 - 2$

Hypothesis test

In the MLR setup:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 - \mu_1 \end{bmatrix}$$

$$\lambda = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\lambda^T\beta = \mu_2 - \mu_1$$

$$\overbrace{\begin{aligned} H_0&: \boxed{\mu_2 - \mu_1 = 0} \\ H_a&: \mu_2 - \mu_1 \neq 0 \end{aligned}}^{\lambda^T\beta = 0}$$

$H_0: \lambda^T\beta = 0$

vs $H_a: \lambda^T\beta \neq 0$

The appropriate test statistic is

$$T = \frac{\bar{Y}_{2\cdot} - \bar{Y}_{1\cdot}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

In the MLR setup:

$$T = \frac{\lambda^T\beta}{Se\sqrt{\lambda^T(X^TX)^{-1}\lambda}}$$

(Exercise: show $\lambda^T(X^TX)^{-1}\lambda = \frac{1}{n_1} + \frac{1}{n_2}$)

Coding binary predictors

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}$$

The second column of $\boldsymbol{X}$ is like indicators of which group the observation belongs to.

$$x_{i2} = \begin{cases} 1 & \text{if observation } i \text{ belongs to group 2} \\ 0 & \text{if observation } i \text{ belongs to group 1} \end{cases}$$

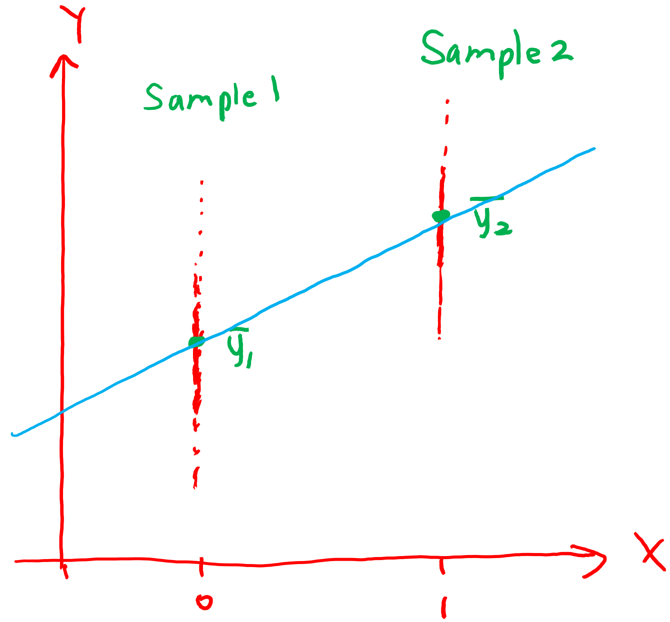
$Y_i = \beta_0 + \beta_1 x_{i2} + \varepsilon_i$ where $\varepsilon_i \sim \text{iid } N(0, \sigma^2)$

For Sample 1 $(x_{i2} = 0)$

$Y_i = \beta_0 + \varepsilon_i \quad = \mu_1 + \varepsilon_i$

For Sample 2 $(x_{i2} = 1)$

$Y_i = \beta_0 + \underbrace{\beta_1}_{\beta_1 = \mu_2 - \mu_1} + \varepsilon_i \quad = \mu_1 + (\mu_2 - \mu_1) + \varepsilon_i \quad = \mu_2 + \varepsilon_i$



$H_0: \mu_2 - \mu_1 = \beta_1 = 0$

We are essentially testing if the slope of the fitted line is zero.