ANOVA (part II)

TEST ABOUT EQUALITY OF THE MEANS

Testing equality of the means is equivalent to testing

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

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

What are $\tilde{\Sigma}^2$ and $\hat{\Sigma}^2$ here?

- $\tilde{\Sigma}^2$ estimator under H_0: model $\underline{Y} = \beta_1 \cdot \underline{1} + \underline{\varepsilon} \Rightarrow \hat{\beta}_1 = \bar{y}$ overall mean

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

- $\hat{\Sigma}^2$ estimator 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

$$\begin{aligned} F &= \frac{\tilde{\Sigma}^2 - \hat{\Sigma}^2}{\hat{\Sigma}^2} \cdot \frac{N-G}{G-1} = \\ &= \frac{SST - SSE}{SSE} \cdot \frac{N-G}{G-1} = \\ &= \frac{SSR}{SSE} \cdot \frac{N-G}{G-1} = \\ &= \frac{R^2}{1-R^2} \cdot \frac{N-G}{G-1} \overset{H_0}{\sim} F_{G-1,\,N-G} \end{aligned}$$

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

$$\bar{y}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} y_{ig} \qquad g = 1, \dots, G$$

- 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

$$s_g^2 = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (y_{ig} - \bar{y}_g)^2 \qquad g = 1, \dots, G$$

The partition of the sum of squares in the linear model was $\overbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}^{SST} = \overbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}^{SSR} + \overbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)}^{SSE}$

We can specify it for this setting:

The total sum of squares here is $SST = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y})^2$

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

Hence we get

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

⇝ "ANOVA": Analysis of variance

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

$$\sum_{g=1}^{G} (n_g - 1) s_g^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y}_g)^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \hat{y}_{ig})^2 \qquad \text{ERROR SUM OF SQUARES}$$

$$\sum_{g=1}^{G} n_g (\bar{y}_g - \bar{y})^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\bar{y}_g - \bar{y})^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\hat{y}_{ig} - \bar{y})^2 \qquad \text{REGRESSION SUM OF SQUARES}$$

Hence, the F test can also be expressed as

$$F = \frac{\text{BETWEEN-GROUP S.S.}}{\text{WITHIN-GROUP S.S.}} \cdot \frac{N-G}{G-1}$$