Test for the homogeneity of a group of regression co-effecients

Suppose we have bivariate data sets on (X, Y) for p different groups. We denote the observations as (x_{ij}, y_{ij}) for $j = 1(1)n_i$ and $i = 1(1)p$ such that $\sum_{i=1}^{p} n_i = n$, total sample size. Now, for each group, a linear regression of Y on x is considered, then associated regression equation for i^{th} group is written as

$$E(y_{ij}) = \alpha_i + \beta_i (x_{ij} - \bar{x}_{i0}),$$

where $\bar{x}_{i0} = \sum_i \sum_j x_{ij}/n_i \quad \forall\, i = 1(1)p$. From the usual assumption of normality on the error term in regression, we have

$$y_{ij} \overset{iid}{\sim} N\left(E(y_{ij}), \sigma_e^2 \right)$$

Now, a natural question may arise that whether the all p regression equations are homogeneous or equivalently, all the regression lines are parallel to one another. To address the question, we do a statistical test for

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = \beta_0 \text{ (say)}$$

against H_1: atleast one inequality holds in H_0.

Let us first estimate the α_i's and β_i's using least-square method. The usual least-square estimates are

$$\hat{\beta}_i = \frac{\sum_j (x_{ij} - \bar{x}_{i0})(y_{ij} - \bar{y}_{i0})}{\sum_j (x_{ij} - \bar{x}_{i0})^2} = \frac{B_i}{A_i} = b_i, \text{ say}$$

Averaging the regression equation over j, we have

$$\bar{y}_{i0} = \alpha_i + \beta_i \cdot 0$$

$$\Rightarrow \hat{\alpha}_i = \bar{y}_{i0}.$$

Now the unrestricted residual SS is

$$s_1^2 = \min \sum_i \sum_j \left[y_{ij} - \alpha_i - \beta_i (x_{ij} - \bar{x}_{i0}) \right]^2$$

$$= \sum_i \sum_j \left[y_{ij} - \hat{\alpha}_i - \hat{\beta}_i (x_{ij} - \bar{x}_{i0}) \right]^2$$

$$= \sum_i \sum_j (y_{ij} - \bar{y}_{i0})^2 + \sum_i \sum_j b_i^2 (x_{ij} - \bar{x}_{i0})^2 - 2 \sum_i \sum_j (y_{ij} - \bar{y}_{i0}) b_i (x_{ij} - \bar{x}_{i0})$$

$$= \sum_i \sum_j (y_{ij} - \bar{y}_{i0})^2 + \sum_i b_i^2 \left[\sum_j (x_{ij} - \bar{x}_{i0})^2\right] - 2 \sum_i b_i \left[\sum_j (x_{ij} - \bar{x}_{i0})(y_{ij} - \bar{y}_{i0})\right]$$

$$= \sum_i \sum_j (y_{ij} - \bar{y}_{i0})^2 + \sum_i b_i^2 . A_i - 2 \sum_i b_i B_i$$

$$= \sum_i \sum_j (y_{ij} - \bar{y}_{i0})^2 - \sum_i b_i B_i \qquad \left(\because\ b_i^2 . A_i = \frac{B_i^2}{A_i^2} . A_i = \frac{B_i^2}{A_i} = b_i . B_i\right)$$

$$= \sum_i (C_i - b_i B_i), \text{ where } C_i = \sum_j (y_{ij} - \bar{y}_{i0})^2 = \sum_j (y_{ij} - \hat{\alpha}_i)^2$$

$$\boxed{\begin{aligned} &\cancel{\text{Now}},\ E\cancel{(y_{ij} - \bar{y}_{i0})} = \cancel{E(y_{ij})} - \cancel{E(\bar{y}_{i0})} \\ &\qquad = \cancel{\alpha_i + \beta_i (x_{ij} - \bar{x}_{i0})} - \cancel{[\alpha_i + \beta_i 0]} \\ &\qquad = \cancel{\beta_i (x_{ij} - \bar{x}_{i0})} \end{aligned}}$$

For each i, $C_i - b_i B_i = C_i - \hat{\beta}_i B_i$ consists n_i terms of Y subject to 2 restrictions for the estimates of α_i and β_i.

$\therefore$ $(C_i - b_i B_i)$ has d.f. $(n_i - 2)$.

$\Rightarrow$ S_1^2 has d.f. $\sum_{i=1}^{p} (n_i - 2) = n - 2p$

Next, the restricted (i.e. under H_0) residual SS is

$$S_2^2 = \min_{\substack{\alpha_i, \beta_i \\ H_0}} \sum_i \sum_j \left[y_{ij} - \alpha_i - \beta_i (x_{ij} - \bar{x}_{i0})\right]^2$$

$$= \sum_i \sum_j \left[y_{ij} - \hat{\alpha}_i - \hat{\beta}_0 (x_{ij} - \bar{x}_{i0})\right]^2,$$

where the least square estimates for α_i and for the common value of β_i under H_0 are

$$\hat{\alpha}_i = \bar{y}_{i0}, \quad \hat{\beta}_0 = \frac{\sum_i \sum_j (x_{ij} - \bar{x}_{i0})(y_{ij} - \bar{y}_{i0})}{\sum_i \sum_j (x_{ij} - \bar{x}_{i0})^2} = \frac{\sum_i B_i}{\sum_i A_i} = \frac{B_0}{A_0} = b \text{ (say)}.$$

$$= \sum_i \sum_j (y_{ij} - \bar{y}_{i0})^2 + b^2 A_0 - 2 b B_0$$

$$= \sum_i C_i - b B_0 \qquad \left(\because\ b^2 A = \frac{B_0^2}{A_0^2} . A_0 = b . B_0\right)$$

with d.f. $\left(\sum_i n_i - 1\right) - 1 = n - p - 1$.

Under H_0, S_1^2 and S_2^2 both follow χ^2 distribution with respective d.f.s.

Hence, we define

$$F = \frac{(S_2^2 - S_1^2)/(p-1)}{S_1^2/(n-2p)} \text{ which follows } F_{p-1, n-2p}$$

since $S_2^2 - S_1^2 \sim \chi^2_{n-p-1-\overline{n-2p}} \equiv \chi^2_{p-1}$

If the observed F, say F_0,

$$F_0 = \frac{MSR}{MSE} > F_{\alpha;\, \overline{p-1},\, \overline{n-2p}}$$

we reject H_0, otherwise we accept it.

Here, MSR = Mean Square due to regressions $= \dfrac{S_2^2 - S_1^2}{p-1}$

MSE = Mean Square Error $= \dfrac{S_1^2}{n-2p}$

Associated ANOVA table

Sources of Variation	d.f.	SS	MS	F_0
Heterogeneity between regression Lines for different groups	$p-1$	$SSR = S_2^2 - S_1^2 = \sum_i b_i B_i - bB_0$	$MSR = \dfrac{SSR}{p-1}$	$F_0 = \dfrac{MSR}{MSE}$
within groups	$n-2p$	$S_1^2 = \sum_i C_i - \sum_i b_i B_i$	$MSE = \dfrac{SSE}{n-2p} = \dfrac{S_1^2}{n-2p}$	

Scanned by CamScanner