Test for the significance of the regressor variables in a multiple regression

Suppose we have a set of k independent variables, $X_1, \ldots, X_k$, and the dependent variable Y. We denote the observations as $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$ for $i = 1(1)n$, the sample size. Now, a multiple linear regression equation

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + e_i, \quad \ldots (1)$$

for $i = 1(1)n$, is considered. In this model, e_i is assumed to be normally distributed with mean 0 and variance σ_e^2 and e_i's are independent over $i = 1(1)n$, i.e. $e_i \overset{iid}{\sim} N(0, \sigma_e^2)$.

Now, one may be interested to test wheather all the k regressors, $X_1, \ldots, X_k$, have significant presence in the regression model (1). To address the question, we do a statistical test for

$$H_0: \quad \beta_j = 0 \; \forall \; j = 1(1)k$$

which means that there is no dependence of Y on $X_1, X_2, \ldots, X_k$. Commonly, H_A can be considered as $H_A: \beta_j \neq 0$ for atleast one j. For simplicity in the calculation of the least-square estimates of the parameters α and β_j's, we can re-write the model (1) as

$$y_i = \alpha' + \beta_1 (x_{1i} - \bar{x}_1) + \cdots + \beta_k (x_{ki} - \bar{x}_k) + e_i \quad \ldots (2)$$

, where $\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ji}$.

$$\text{or,} \quad y_i = \alpha' + \beta_1 x'_{1i} + \beta_2 x'_{2i} + \cdots + \beta_k x'_{ki} + e_i, \quad \ldots (2)$$

where $x'_{ji} = x_{ji} - \bar{x}_j \quad \forall \; j = 1(1)k$.

Using Least-square method, estimates of β_j's are

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} (x'_{ji} y_i}{\sum_{i=1}^{n} x'^2_{ji}} = b_j, \text{ say.} \quad \ldots (3)$$

Averaging the equation (2) over i, we have $\bar{y} = \alpha' + \sum_{j=1}^{k} \beta_j \cdot 0$.

$$\Rightarrow \hat{\alpha}' = \bar{y}.$$

Now, the unrestricted residual sum of squares (SS) is

$$S_1^2 = \min_{\alpha', \beta_j} \sum_{i=1}^{n} \left(y_i - \alpha' - \sum_{j=1}^{k} \beta_j \cdot (x'_{ji})\right)^2$$

$$= \sum_{i=1}^{n} \left(y_i - \hat{\alpha}' - \sum_{j=1}^{k} b_j x'_{ji}\right)$$ [replacing parameters by their least-square estimates]

$$= \sum_{i=1}^{n} \left[(y_i - \bar{y})^2 + \left\{\sum_{j=1}^{k} b_j \cdot x'_{ji}\right\}^2 - 2(y_i - \bar{y}) \sum_{j=1}^{k} b_j \cdot x'_{ji}\right]$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} \sum_{j=1}^{k} b_j^2 x'^2_{ji} - 2 \sum_{j=1}^{k} b_j \sum_{i=1}^{n} y_i \cdot x'_{ji}$$

$$\left[\because 2 \sum_j \sum_{j'} \sum_i b_j b_{j'} x'_{ji} x'_{j'i} = 0 \text{ as } \sum_i x'_{ji} = 0\right]$$

$$= \sum_i (y_i - \bar{y})^2 + \sum_j b_j^2 \sum_i x'^2_{ji} - 2 \sum_j b_j \sum_i x'_{ji} y_i$$

$$= \sum_i (y_i - \bar{y})^2 - \sum_j b_j \sum_i x'_{ji} y_i \quad \left[\text{from } (3), \; b_j \sum_i x'^2_{ji} = \sum_i x'_{ji} y_i\right]$$

$$= \sum_i (y_i - \bar{y})^2 - \sum_{j=1}^{k} b_j \cdot P_j \quad , \text{where } P_j = \sum_{i=1}^{k} x'_{ji} y_i$$

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

$$S_2^2 = \min_{\substack{\alpha', \beta_j \\ H_0}} \sum_{i=1}^{n} \left(y_i - \alpha' - \sum_{j=1}^{k} \beta_j \cdot x'_{ji}\right)^2$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad , \text{since under } H_0, \; \beta_j = 0 \; \forall j = 1(1)k$$

The d.f. of S_1^2 is $(n-k-1)$.

[$\because$ the first part of S_1^2 consists n observations on Y with one restriction of sample mean, $\bar{y}$, so it has $(n-1)$ d.f and second part consists k estimates $b_1, b_2, \ldots, b_k$, that means k more restrictions]

The d.f. of S_2^2 is $(n-1)$.

So, SS due to regression $= S_2^2 - S_1^2 = \sum_{j=1}^{k} b_j P_j = SSR$, say.

SS due to error $= S_1^2 = SSE$, say.

So, the d.f. of SSR $= d.f.(S_2^2) - d.f.(S_1^2) = \overline{n-1} - \overline{n-k-1} = k$

Under H_0, SSR and SSE both follow χ^2 distribution with respective d.f.s. and they are independent.

Hence,

$$F = \frac{SSR/k}{SSE/(n-k-1)} \sim F_{k, n-k-1}$$

Hence, if the observed F, say F_0, is greater than $F_{\alpha; k, n-k-1}$. we reject H_0 at level α, otherwise we do not reject H_0.

Associated ANOVA table

Source of variation	d.f.	SS	MS	F_0
Due to multiple linear regression	k	$SSR = \sum_{j=1}^{k} b_j P_j$	$MSR = \frac{SSR}{k}$	$F_0 = \frac{MSR}{MSE}$
Error	$n-k-1$	$SSE = S_1^2$	$MSE = \frac{SSE}{n-k-1}$	
Total	$n-1$	$\sum_i (y_i - \bar{y})^2 = S_2^2$	—	—

■ **Test for the significance of a subset of regressor variables in a given multiple linear regression model with k no. of regressor variables:**

Suppose we have a multiple linear regression model of the variable Y on a set of k independent regressor variables $X_1, X_2, \ldots, X_k$ and the regression equation has the following form

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + e_i, \quad \cdots\cdots (1)$$

for $i = 1(1)n$, where $e_i \overset{iid}{\sim} N(0, \sigma_e^2)$ $\forall i = 1(1)n$, by assumption.

Now, one may be interested to test whether a subset of s no. of regressor variables has significance in the multiple linear regression model (1), where $s < k$.

To address the question, we do a statistical hypothesis test for

$$H_0: \beta_j = 0 \quad \forall j = 1(1)s.$$

against H_A: not all β_j equal to 0 for $j = 1(1)s$,

i.e. $\beta_j \neq 0$ for atleast one j.

For simplicity in the calculation of the least-square estimates of the parameters, we can re-write the model (1) as

$$y_i = \alpha' + \beta_1(x_{1i} - \bar{x}_1) + \cdots\cdots + \beta_K(x_{Ki} - \bar{x}_K) + e_i$$

$$= \alpha' + \beta_1 x'_{1i} + \cdots\cdots + \beta_K x'_{Ki} + e_i \quad \cdots\cdots (2)$$

, where $x'_{ji} = x_{ji} - \bar{x}_j$, $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ji}$ $\forall\ j = 1(1)k$.

Therefore, under H_0, equation (2) reduces to

$$y_i = \alpha' + \beta_{s+1} x'_{s+1,i} + \cdots + \beta_K x'_{Ki} + e_i \quad \cdots\cdots (3)$$

Now the unrestricted residual SS from model (2) is

$$S_1^2 = \min_{\alpha', \beta_j\text{'s}} \sum_{i=1}^{n}\left(y_i - \alpha' - \sum_{j=1}^{K} \beta_j . x'_{ji}\right)^2$$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{j=1}^{k} b_j . P_j \quad , \text{ where } P_j = \sum_{i=1}^{n} x'_{ji} y_i$$

[Derivation of S_1^2 is already stated in earlier test]

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

$$S_2^2 = \min_{\substack{\alpha', \beta_j\text{'s} \\ H_0}} \sum_{i=1}^{n}\left(y_i - \alpha' - \sum_{j=1}^{K} \beta_j . x'_{ji}\right)^2$$

$$= \min_{\alpha', \beta_j\text{'s}} \sum_{i=1}^{n}\left(y_i - \alpha' - \sum_{j=s+1}^{k} \beta_j . x'_{ji}\right)^2$$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{j=s+1}^{k} b_j^* P_j \,,$$

[following the derivation similar to the derivation for S_1^2]

where b_j^*'s are the least square estimates of β_j's, respectively, for the restricted model (3).

The degrees of freedom of S_1^2 and S_2^2 are $(n-k-1)$ and $(n-(k-s)-1)$, respectively.

Thus, SS due to regression involving s regressors variables

$$= S_2^2 - S_1^2$$

$$= \sum_{j=1}^{K} b_j . P_j - \sum_{j=s+1}^{k} b_j^* . P_j \quad \text{with d.f.} = (n-k-1) - (n-\overline{k-s}-1) = s$$

$$= SSR_{s/k}, \text{say.}$$

SS due to error $= S_1^2 = SSE$, say.

Under H_0, $SSR_{s/k}$ and SSE both follow χ^2- distribution with respective d.f.s. and they are independent.

So, $$F = \frac{(SSR_{s/k})/s}{SSE/n-k-1} \sim F_{s,\, n-k-1}.$$

If the observed F, say F_0, is greater than $F_{\alpha;\, s,\, n-k-1}$, we reject H_0 at level α, otherwise we do not reject H_0.

Associated ANOVA table.

Source of Variation	d.f	SS		F_0
Due to multiple linear regression of Y on $X_1, \dots, X_s$, after fitting $X_1, \dots, X_k$	s	$SSR_{s/k} = \sum_{j=1}^{k} b_j P_j - \sum_{j=s+1}^{k} b_j^* P_j$	$MSR_{s/k}$	$F_0 = \frac{MSR_{s/k}}{MSE}$
Due to multiple linear regression of Y on $X_{s+1}, \dots, X_k$	$k-s$	$\sum_{j=s+1}^{k} b_j^* P_j$.	MSR_{k-s}	
Due to multiple linear regression of Y on $X_1, \dots, X_k$	k	$\sum_{j=1}^{k} b_j . P_j$	MSR_k	
Error	$n-k-1$	$\sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{j=1}^{k} b_j . P_j = S_1^2$	MSE	
Total	$n-1$	$\sum_{i=1}^{n} (y_i - \bar{y})^2$	—	—

Remark: If anyone wants to test whether the inclusion of variable X_1 is at all necessary in the multiple linear regression of Y on $X_1, X_2, \dots, X_k$, then the above test procedure can be carried out with $s=1$.