LINEAR REGRESSION ANALYSIS

MODULE – XIII

Lecture - 37

Variable Selection and Model Building

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

The complete regression analysis depends on the explanatory variables present in the model. It is understood in the regression analysis that only correct and important explanatory variables appear in the model. In practice, after ensuring the correct functional form of the model, the analyst usually has a pool of explanatory variables which possibly influence the process or experiment. Generally, all such candidate variables are not used in the regression modeling but a subset of explanatory variables is chosen from this pool. How to determine such an appropriate subset of explanatory variables to be used in regression is called the **problem of variable selection**.

While choosing a subset of explanatory variables, there are two possible options:

1. In order to make the model as realistic as possible, the analyst may include as many as possible explanatory variables.
2. In order to make the model as simple as possible, one may include only fewer number of explanatory variables.

Both the approaches have their own consequences. In fact, model building and subset selection have contradicting objectives. When large number of variables are included in the model, then these factors can influence the prediction of study variable *y*. On the other hand, when small number of variables are included then the predictive variance of $\hat{y}$ decreases. Also, when the observations on more number are to be collected, then it involves more cost, time, labour etc. A compromise between these consequences is striked to select the “best regression equation”.

The problem of variable selection is addressed assuming that the functional form of the explanatory variable, e.g., x^2, $\frac{1}{x}$, $\log x$ etc., is known and no outliers or influential observations are present in the data. Various statistical tools like residual analysis, identification of influential or high leverage observations, model adequacy etc. are linked to variable selection. In fact, all these processes should be solved simultaneously. Usually, these steps are iteratively employed. In the first step, a strategy for variable selection is opted and model is fitted with selected variables. The fitted model is then checked for the functional form, outliers, influential observations etc. Based on the outcome, the model is re-examined and selection of variable is reviewed again. Several iterations may be required before the final adequate model is decided.

There can be two types of incorrect model specifications.

1. Omission/exclusion of relevant variables.
2. Inclusion of irrelevant variables.

Now we discuss the statistical consequences arising from both the situations.

1. Exclusion of relevant variables

In order to keep the model simple, the analyst may delete some of the explanatory variables which may be of importance from the point of view of theoretical considerations. There can be several reasons behind such decision, e.g., it may be hard to quantity the variables like taste, intelligence etc. Sometimes it may be difficult to take correct observations on the variables like income etc.

Let there be *k* candidate explanatory variables out of which suppose *r* variables are included and (*k* – *r*) variables are to be deleted from the model. So partition the *X* and β as

$$X_{n\times k} = \left(X_{1\ n\times r} \quad X_{2\ n\times(k-r)} \right) \quad \text{and} \quad \beta = \left(\beta_{1\ r\times 1} \quad \beta_{2\ (k-r)\times 1} \right).$$

The model $y = X\beta + \varepsilon,\ E(\varepsilon) = 0,\ V(\varepsilon) = \sigma^2 I$ can be expressed as

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

which is called as **full model** or **true model.**

After dropping the *r* explanatory variable in the model, the new model is

$$y = X_1\beta_1 + \delta$$

which is called as **misspecified mode**l or **false model.**

Applying OLS to the false model, the OLSE of β_1 is

$$b_{1F} = (X_1'X_1)^{-1}X_1'y.$$

The estimation error is obtained as follows:

$$b_{1F} = (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon)$$

$$= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon$$

$$b_{1F} - \beta_1 = \theta + (X_1'X_1)^{-1}X_1'\varepsilon$$

where

$$\theta = (X_1'X_1)^{-1}X_1'X_2\beta_2.$$

Thus

$$E(b_{1F} - \beta) = \theta + (X_1'X_1)^{-1}E(\varepsilon)$$

$$= \theta$$

which is a linear function of β_2, i.e., the coefficients of excluded variables. So b_{1F} is biased, in general. The bias vanishes if $X_1'X_2 = 0$, i.e., X_1 and X_2 are orthogonal or uncorrelated.

The mean squared error matrix of b_{1F} is

$$MSE(b_{1F}) = E(b_{1F} - \beta)(b_{1F} - \beta)'$$

$$= E\left[\theta\theta' + \theta\varepsilon' X_1(X_1'X_1)^{-1} + (X_1'X_1)^{-1}X_1'\varepsilon\theta' + (X_1'X_1)^{-1}X_1'\varepsilon\varepsilon' X_1(X_1'X_1)^{-1}\right]$$

$$= \theta\theta' + 0 + 0 + \sigma^2(X_1'X_1)^{-1}X_1'IX_1(X_1'X_1)^{-1}$$

$$= \theta\theta' + \sigma^2(X_1'X_1)^{-1}.$$

So efficiency generally declines. Note that the second term is the conventional form of MSE.

The residual sum of squares is

$$\hat{\sigma}^2 = \frac{SS_{res}}{n-r} = \frac{e'e}{n-r}$$

where

$$e = y - X_1 b_{1F} = \bar{H}_1 y,$$

$$\bar{H}_1 = I - X_1 (X_1' X_1)^{-1} X_1'.$$

Thus

$$\bar{H}_1 y = \bar{H}_1 (X_1\beta_1 + X_2\beta_2 + \varepsilon)$$

$$= 0 + \bar{H}_1 (X_2\beta_2 + \varepsilon)$$

$$= \bar{H}_1 (X_2\beta_2 + \varepsilon)$$

$$y'\bar{H}_1 y = (X_1\beta_1 + X_2\beta_2 + \varepsilon)\bar{H}_1 (X_2\beta_2 + \varepsilon)$$

$$= (\beta_2' X_2 \bar{H}_1' \bar{H}_1 X_2\beta_2 + \beta_2' X_2' \bar{H}_1 \varepsilon + \beta_2' X_2 \bar{H}_1' X_2\beta_2 + \beta_1' X_1' \bar{H}_1 \varepsilon + \varepsilon' \bar{H}_1' X_2\beta_2 + \varepsilon' \bar{H}_1 \varepsilon).$$

$$E(s^2) = \frac{1}{n-r}\left[E(\beta_2' X_2' \bar{H}_1 X_2\beta_2) + 0 + 0 + E(\varepsilon' \bar{H} \varepsilon)\right]$$

$$= \frac{1}{n-r}\left[\beta_2' X_2' \bar{H}_1 X_2\beta_2) + (n-r)\sigma^2\right]$$

$$= \sigma^2 + \frac{1}{n-r}\beta_2' X_2' \bar{H}_1 X_2\beta_2.$$

Thus s^2 is a biased estimator of σ^2 and s^2 provides an over estimate of σ^2. Note that even if $X_1' X_2 = 0$, then also s^2 gives an overestimate of σ^2. So the statistical inferences based on this will be faulty. The *t*-test and confidence region will be invalid in this case.

If the response is to be predicted at $x' = (x_1', x_2')$, then using the full model, the predicted value is

$$\hat{y} = x'b = x'(X'X)^{-1}X'y$$

with

$$E(\hat{y}) = x'\beta$$

$$Var(\hat{y}) = \sigma^2\left[1 + x'(X'X)^{-1}x\right].$$

When subset model is used then the predictor is

$$\hat{y}_1 = x_1' b_{1F}$$

and then

$$\begin{aligned} E(\hat{y}_1) &= x_1'(X_1'X_1)^{-1}X_1'E(y) \\ &= x_1'(X_1'X_1)^{-1}X_1'E(X_1\beta_1 + X_2\beta_2 + \varepsilon) \\ &= x_1'(X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2) \\ &= x_1'\beta_1 + x_1'(X_1'X_1)^{-1}X_1'X_2\beta_2 \\ &= x_1'\beta_1 + x_i'\theta. \end{aligned}$$

Thus $\hat{y}_1$ is a biased predictor of y. It is unbiased when $X_1'X_2 = 0$. The MSE of predictor is

$$MSE(\hat{y}_1) = \sigma^2\left[1 + x_1'(X_1'X_1)^{-1}x_1\right] + \left(x_1'\theta - x_2'\beta_2\right)^2.$$

Also

$$Var(\hat{y}) \geq MSE(\hat{y}_1)$$

provided $V(\hat{\beta}_2) - \beta_2\beta_2'$ is positive semidefinite.

2. Inclusion of irrelevant variables

Sometimes due to enthusiasm and to make the model more realistic, the analyst may include some explanatory variables that are not very relevant to the model. Such variables may contribute very little to the explanatory power of the model. This may tend to reduce the degrees of freedom (*n* - *k*) and consequently the validity of inference drawn may be questionable. For example, the value of coefficient of determination will increase indicating that the model is getting better which may not really be true.

Let the true model be

$$y = X\beta + \varepsilon,\ E(\varepsilon) = 0, V(\varepsilon) = \sigma^2 I$$

which comprise *k* explanatory variable. Suppose now *r* additional explanatory variables are added to the model and resulting model becomes

$$y = X\beta + Z\gamma + \delta$$

where Z is a $n \times r$ matrix of n observations on each of the *r* explanatory variables and γ is $r \times 1$ vector of regression coefficient associated with *Z* and δ is disturbance term. This model is termed as **false model.**

Applying OLS to false model, we get

$$\begin{pmatrix} b_F \\ c_F \end{pmatrix} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1} \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix} \begin{pmatrix} b_F \\ c_F \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

$$\Rightarrow X'Xb_F + X'ZC_F = X'y \qquad (1)$$

$$Z'Xb_F + Z'ZC_F = Z'y \qquad (2)$$

where b_F and C_F are the OLSEs of β and γ respectively.

Premultiply equation (2) by $X'Z(Z'Z)^{-1}$, we get

$$X'Z(Z'Z)^{-1}Z'Xb_F + X'Z(Z'Z)^{-1}Z'ZC_F = X'Z(Z'Z)^{-1}Z'y. \qquad (3)$$

Subtracting equation (1) from (3), we get

$$\left[X'X - X'Z(Z'Z)^{-1}Z'X\right]b_F = X'y - X'Z(Z'Z)^{-1}Z'y$$

$$X'\left[I - Z(Z'Z)^{-1}Z'\right]Xb_F = X'\left[I - Z(Z'Z)^{-1}Z'\right]y$$

$$\Rightarrow b_F = (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z y$$

where $\bar{H}_Z = I - Z(Z'Z)^{-1}Z'$.

The estimation error of b_F is

$$\begin{aligned} b_F - \beta &= (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z y - \beta \\ &= (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z(X\beta + \varepsilon) - \beta \\ &= (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z\varepsilon. \end{aligned}$$

Thus

$$E(b_F - \beta) = (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z E(\varepsilon) = 0,$$

so b_F is unbiased even when some irrelevant variables are added to the model.

The covariance matrix is

$$\begin{aligned} V(b_F) &= E\left(b_F - \beta\right)\left(b_F - \beta\right)^1 \\ &= E\left[\left(X'\bar{H}_Z X\right)^{-1}X'\bar{H}_Z\varepsilon\varepsilon'\bar{H}_Z X(X'\bar{H}_Z X)^{-1}\right] \\ &= \sigma^2\left(X'\bar{H}_Z X\right)^{-1}X'\bar{H}_Z I\bar{H}_Z X\left(X'\bar{H}_Z X\right)^{-1} \\ &= \sigma^2\left(X'\bar{H}_Z X\right)^{-1}. \end{aligned}$$

If OLS is applied to true model, then

$$b_T = (X'X)^{-1}X'y$$

with

$$E(b_T) = \beta$$

$$V(b_T) = \sigma^2(X'X)^{-1}.$$

To compare b_F and b_T we use the following result.

Result: If A and B are two positive definite matrices then $A - B$ is atleast positive semi definite if $B^{-1} - A^{-1}$ is also atleast positive semi definite.

Let

$$A = (X'\bar{H}_Z X)^{-1}$$

$$B = (X'X)^{-1}$$

$$B^{-1} - A^{-1} = X'X - X'\bar{H}_Z X$$

$$= X'X - X'X + X'Z(Z'Z)^{-1}Z'X$$

$$= X'Z(Z'Z)^{-1}Z'X$$

which is atleast positive semi definite matrix. This implies that the efficiency declines unless $X'Z = 0$. If $X'Z = 0$, i.e., X and Z are orthogonal, then both are equally efficient.

The residual sum of squares under false model is

$$SS_{res} = e_F'e_F$$

where

$$e_F = y - Xb_F - Zc_F$$

$$b_F = (X'\bar{H}_Z X)^{-1} X'\bar{H}_Z y$$

$$\begin{aligned} c_F &= (Z'Z)^{-1}Z'y - (Z'Z)^{-1}Z'Xb_F \\ &= (Z'Z)^{-1}Z'(y - Xb_F) \\ &= (Z'Z)^{-1}Z'\left[I - X(X'\bar{H}_Z X)^{-1}X'\bar{H}_z\right]y \\ &= (Z'Z)^{-1}Z'\bar{H}_{XZ}y \end{aligned}$$

$$\bar{H}_Z = I - Z(Z'Z)^{-1}Z'$$

$$\bar{H}_{Zx} = I - X(X'\bar{H}_Z X)^{-1}X'\bar{H}_Z$$

$\bar{H}_{ZX}^2 = \bar{H}_{ZX}$: idempotent.

So

$$\begin{aligned} e_F &= y - X(X'\bar{H}_Z X)^{-1}X'\bar{H}_Z y - Z(Z'Z)^{-1}Z'\bar{H}_{ZX}y \\ &= \left[I - X(X'\bar{H}_Z X)^{-1}X'\bar{H}_Z - Z(Z'Z)^{-1}Z'H_{ZX}\right]y \\ &= \left[\bar{H}_{ZX} - (I - \bar{H}_Z)\bar{H}_{ZX}\right]y \\ &= \bar{H}_Z\bar{H}_{ZX}y \\ &= \bar{H}_{ZX}^{*}y \quad \text{where} \quad \bar{H}_{ZX}^{*} = \bar{H}_Z\bar{H}_{ZX}. \end{aligned}$$

Thus

$$SS_{res} = e_F' e_F$$
$$= y' \bar{H}_Z \bar{H}_{ZX} \bar{H}_{ZX} \bar{H}_Z y$$
$$= y' \bar{H}_Z \bar{H}_{ZX} y$$
$$= y' \bar{H}_{ZX}^* y$$
$$E(SS_{res}) = \sigma^2 tr(\bar{H}_{ZX}^*)$$
$$= \sigma^2 (n-k-r)$$
$$E\left(\frac{SS_{res}}{n-k-r}\right) = \sigma^2.$$

So $\frac{SS_{res}}{n-k-r}$ is an unbiased estimator of σ^2 .

A comparison of exclusion and inclusion of variables is as follows:

	Exclusion type	**Inclusion type**
Estimation of coefficients	Biased	Unbiased
Efficiency	Generally declines	Declines
Estimation of disturbance term	Over-estimate	Unbiased
Conventional test of hypothesis and confidence region	Invalid and faulty inferences	Valid though erroneous