LECTURE - Regression

Multiple Linear Regression Model (MLR Model)

Lecture - 4 Part - 3

Multiple Linear Regression (MLR) Model:

$$Y_{n\times 1} = X_{n\times(k+1)}\beta_{(k+1)\times 1} + e_{n\times 1}$$

Estimated coefficients: $\hat{\beta} = (X'X)^{-1}X'Y$.

Estimated Regression Hyperplane: $\hat{Y} = X\hat{\beta}$

Leverage Points

In using SLR model we have seen that it is not wise to predict at any point x_0 which is outside the range of x-values. But in MLR model, it is sometime difficult to apply the same concept. Even if all the individual x_{0i} are within the limits of the corresponding x_i's there is no guarantee that the point is within the ellipsoid of observed x-values. So in predicting new responses and in estimating the mean response at a given point $(x_{01}, x_{02}, \ldots, x_{0k})$ one must be careful about extrapolating beyond the region containing the original observations. It is very possible that a model that fits well in the region of the original data will perform poorly outside that region. In multiple regression it is easy to inadvertently extrapolate, since the range of the regressors jointly defines the region.

Consider the simple two dimensional graph below where the new point $x_0 = (x_{01}, x_{02})$ is outside the region where each individual coordinates are within the range of their respective direction.

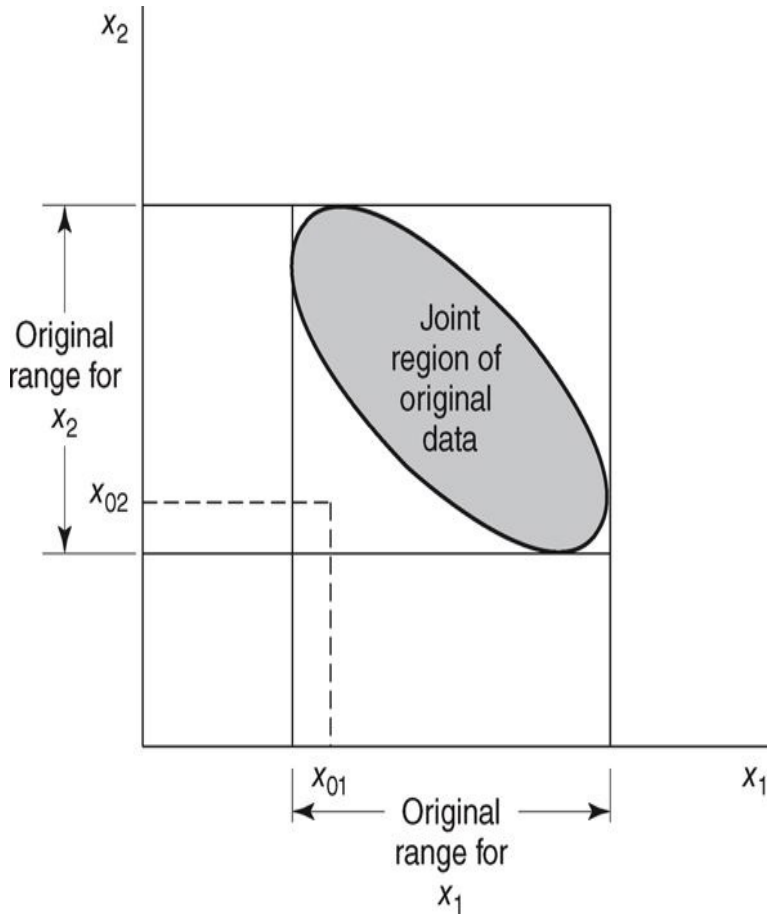

Figure 1: (x_{01}, x_{02}) is outside the ellipsoid

Since simply comparing the levels of the x's for a new data point with the ranges of the original x's will not always detect a hidden extrapolation, it would be helpful to have a formal procedure to do so. We will define the smallest convex set containing all of the original n data points $(x_{i1}, x_{i2}, \ldots, , x_{ik})$, $i = 1, 2, \cdots, n$. as the regressor variable hull (RVH). If a point $(x_{01}, x_{02}, ., x_{0k})$ lies inside or on the boundary of the RVH, then prediction or estimation involves interpolation, while if this point lies outside the RVH, extrapolation is required.

The diagonal elements of the hat matrix $H = X(X'X)^{-1}X'$ are denoted as h_{ii} and those are useful in detecting hidden extrapolation. The values of h_{ii} depend both on the Euclidean distance of the point x_i from the centroid and on the density of the points in the RVH. In general, the point that has the largest value of h_{ii}, say h_{max}, will lie on the boundary of the RVH in a region of the x space where the density of the observations is relatively low. The set of points x (not necessarily data points used to fit the model) that satisfy

$$x_0(X'X)^{-1}x_0' \ \leq \ h_{max}$$

is an ellipsoid enclosing all points inside the RVH. Thus, if we are interested in prediction or estimation at the point $x_0 = (x_{01}, x_{02}), \ldots, , x_{0k})$, the location

of that point relative to the RVH is reflected by

$$h_{00} = x_0(X'X)^{-1}x_0'$$

Points for which $h_{00} > h_{max}$ are outside the ellipsoid enclosing the RVH and are extrapolation points. However, if $h_{00} < h_{max}$, then the point is inside the ellipsoid and possibly inside the RVH and would be considered an interpolation point because it is close to the cloud of points used to fit the the model.

If the estimated error for such a point is not very high but h_{ii} is very high and much higher than the other h_{ii} values then it is called leverage point (discussed later).

STANDARDIZED REGRESSION COEFFICIENTS

It is usually difficult to directly compare regression coefficients because the magnitude of $\hat{\beta}_j$ reflects the units of measurement of the regressor x_j. For this reason, it is sometimes helpful to work with scaled regressor and response variables that produce dimensionless regression coefficients. These dimensionless coefficients are usually called standardized regression coefficients.

Any variable can be standardized by subtracting its own mean and then dividing by its own standard deviation. (same concept of z-score).

For example, all the n values of the variable X_i (i.e x_{ij} for j=1,2,..., n) can be standardized as

$$z_{ij} = \frac{x_{ij} - \bar{x_j}}{s_j}, \quad \text{where } \bar{x_j} \,\&\, s_j$$

are the mean and sample standard deviation of x_j's.

Unit Normal Scaling: In this approach regressors and the response variable are all standardized by its own mean and standard deviation. Then all the new variables have mean 0 and standard deviation 1.

If $(Y^\star, Z_1, Z_2, \cdots, Z_k)$ are the new variable corresponding to the original

$(Y, X_1, X_2, \cdots, X_k)$ and Z is the corresponding matrix in the place of X then the estimator can be written as

$$\hat{b}^\star = (Z'Z)^{-1} Z' y^\star$$

Remark: Observe that under this scaling, means of all the variables are 0 and as a result estimate of β_0 is always zero (verify by the formula in SLR model).

Unit Length Scaling: There is another scaling which is popular in practice is called "Unit Length Scaling". It is not really very different from "Unit Normal Scaling". Instead of dividing by $s_j = \sqrt{\frac{\sum(x_{ij}-\bar{x_j})^2}{n-1}}$ this scaling divides by $s_{jj} = \sqrt{\sum(x_{ij} - \bar{x_j})^2}$. There is no technical advantage of it but the reason to do that if we set

$w_{ij} = \frac{x_{ij}-\bar{x_j}}{s_{jj}}$ then $\sqrt{\sum(w_{ij} - \bar{w_j})^2} = 1$

Under this scaling the corresponding estimates becomes

$\hat{b}^\star = (W'W)^{-1} W' y^\star$ where W is the corresponding design matrix (X)

MULTICOLLINEARITY

In MLR model, multicollinearity (also collinearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy. Multicollinearity implies near-linear dependence among the regressors. Multicollinearity, or near-linear dependence among the regression variables, is a serious problem that may dramatically impact the usefulness of a regression model. The regressors are the columns of the X matrix, so clearly an exact linear dependence would result in a singular X'X.

Suppose in a MLR model with 2 predictors, we use the unit length scaling. Then the $(X'X)$ matrix (previously called $(W'W)$) will be in the form of a correlation matrix.

In an ideal situation if the predictors are orthogonal, then

$$W'W = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{then} \quad (W'W)^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

But if they are correlated (r=0.8242) then

$$W'W = \begin{pmatrix} 1 & 0.8242 \\ 0.8242 & 1 \end{pmatrix} \quad \text{then} \quad (W'W)^{-1} = \begin{pmatrix} 3.12 & -2.57 \\ -2.57 & 3.12 \end{pmatrix}$$

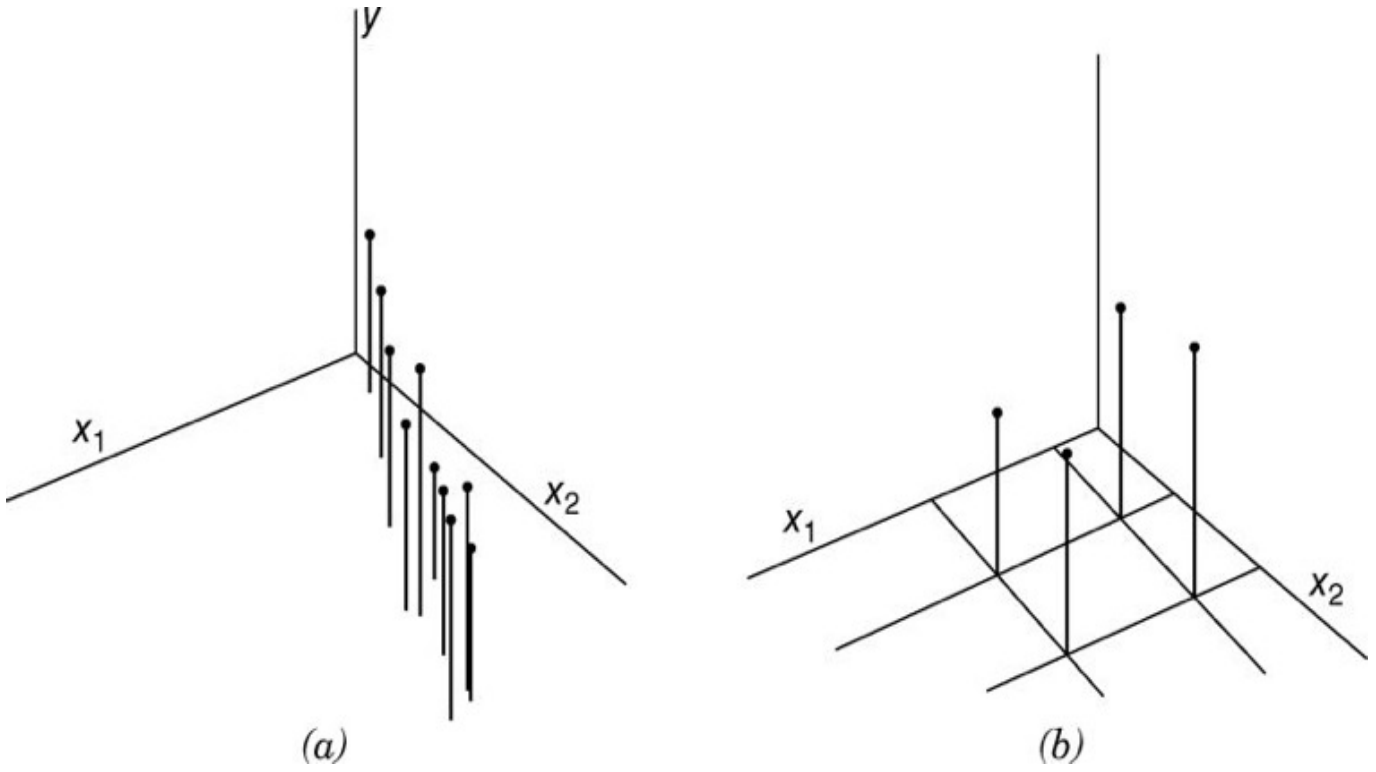

Figure 2: (a) Correlated Predictors (b) Orthogonal Predictors

The main diagonal elements of the inverse of the $(X'X)^{-1}$ matrix in correlation form $(W'W)^{-1}$ above are called variance inflation factors (VIFs), and they are an important multicollinearity diagnostic.

In general, the VIF for the jth regression coefficient can be written as

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of multiple determination obtained from regressing x_j on the other regressor variables. Clearly, if x_j is nearly linearly dependent on some of the other regressors, then R_j^2 will be near unity and VIF will be large.

Observe that in the above example

$$VIF_1 = VIF_2 = \frac{1}{1 - R_j^2} = \frac{1}{1 - r^2} = \frac{1}{1 - (0.8242)^2} = 3.12$$

Indicators that multicollinearity may be present in a model include the following:

(1) Large changes in the estimated regression coefficients when a predictor variable is added or deleted

(2) Insignificant regression coefficients for the affected variables in the multiple regression, but a rejection of the joint hypothesis that those coefficients are all zero (using an F-test)

(3) If a multivariable regression finds an insignificant coefficient of a particular explanator, yet a simple linear regression of the explained variable on this explanatory variable shows its coefficient to be significantly different from zero, this situation indicates multicollinearity in the multivariable regression.