LINEAR REGRESSION ANALYSIS

MODULE – III

Lecture - 14

Multiple Linear Regression Analysis

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Confidence interval estimation

The confidence intervals in multiple regression model can be constructed for individual regression coefficients as well as jointly . We consider both of them as follows:

Confidence interval on the individual regression coefficient

Assuming ε_i's are identically and independently distributed following $N(0,\sigma^2)$ in $y = X\beta + \varepsilon$, we have

$$y \sim N(X\beta, \sigma^2 I)$$

$$b \sim N(\beta, \sigma^2 (X'X)^{-1}).$$

Thus the marginal distribution of any regression coefficient estimate

$$b_j \sim N(\beta_j, \sigma^2 C_{jj})$$

where C_{jj} is the j^{th} diagonal element of $(X'X)^{-1}$.

Thus

$$t_j = \frac{b_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t(n-k) \text{ under } H_0,\ j = 1,2,...$$

where

$$\hat{\sigma}^2 = \frac{SS_{res}}{n-k} = \frac{y'y - b'X'y}{n-k}.$$

So the $100(1-\alpha)\%$ confidence interval for $\beta_j\,(j=1,2,...,k)$ is obtained as follows:

$$P\left[-t_{\frac{\alpha}{2},n-k} \le \frac{b_j-\beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \le t_{\frac{\alpha}{2},n-k}\right]=1-\alpha$$

$$P\left[b_j - t_{\frac{\alpha}{2},n-k}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_j \le b_j + t_{\frac{\alpha}{2},n-k}\sqrt{\hat{\sigma}^2 C_{jj}}\right]=1-\alpha.$$

So the confidence interval is

$$\left(b_j - t_{\frac{\alpha}{2},n-k}\sqrt{\hat{\sigma}^2 C_{jj}}, b_j + t_{\frac{\alpha}{2},n-k}\sqrt{\hat{\sigma}^2 C_{jj}}\right).$$

Simultaneous confidence intervals on regression coefficients

A set of confidence intervals that are true simultaneously with probability $(1-\alpha)$ are called simultaneous or joint confidence intervals.

It is relatively easy to define a joint confidence region for β in multiple regression model.

Since

$$\frac{(b-\beta)'X'X(b-\beta)}{k\,MS_{res}} \sim F_{k,n-k}$$

$$\Rightarrow P\left[\frac{(b-\beta)'X'X(b-\beta)}{k\,MS_{res}} \le F_{\alpha}(k,n-k)\right] = 1-\alpha.$$

So a 100 $(1-\alpha)\%$ joint confidence region for all of the parameters in β is

$$\frac{(b-\beta)'X'X(b-\beta)}{k\,MS_{res}} \sim F_{\alpha}(k,n-k)$$

which describes an elliptically shaped region.

Coefficient of determination (R^2) and adjusted R^2

Let R be the multiple correlation coefficient between y and $X_1, X_2, ..., X_k$. Then square of multiple correlation coefficient (R^2) is called as coefficient of determination. The value of R^2 commonly describes that how well the sample regression line fits to the observed data. This is also treated as a measure of **goodness of fit** of the model.

Assuming that the intercept term is present in the model as

$$y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_k X_{ik} + u_i,\ i = 1, 2, ..., n$$

then

$$R^2 = 1 - \frac{e'e}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= 1 - \frac{SS_{res}}{SS_T}$$

$$= \frac{SS_{reg}}{SS_T}$$

where

SS_{res} : sum of squares due to residuals,

SS_T : total sum of squares,

SS_{reg} : sum of squares due to regression.

R^2 measure the explanatory power of the model which in turn reflects the goodness of fit of the model.

It reflects the model adequacy in the sense that how much is the explanatory power of explanatory variable.

Since

$$e'e = y'\left[I - X(X'X)^{-1}X'\right]y = y'\bar{H}y,$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2,$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\ell' y$ with $\ell = (1,1,...,1)'$, $y = (y_1, y_2, ..., y_n)'$

Thus $$\sum_{i=1}^{n}(y_i - \bar{y})^2 = y'y - n\left(\frac{1}{n^2}\ell' yy'\ell\right)$$

$$= y'y - y'\ell\frac{1}{n}\ell' y$$

$$= y'y - y'\ell(\ell'\ell)^{-1}\ell' y$$

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

$$= y'Ay$$

where $A = I - \ell(\ell'\ell)^{-1}\ell'.$

So $$R^2 = 1 - \frac{y'\bar{H}y}{y'Ay}.$$

The limits of R^2 are 0 and 1, i.e., $0 \le R^2 \le 1.$

- R^2 = 0 indicates the poorest fit of the model.
- R^2 = 1 indicates the best fit of the model.
- R^2 = 0.95 indicates that 95% of the variation in *y* is explained by the explanatory variables. In simple words, the model is 95% good.
- Similarly any other value of R^2 between 0 and 1 indicates the adequacy of fitted model.

Adjusted R^2

If more explanatory variables are added to the model, then R^2 increases. In case the variables are irrelevant, then R^2 will still increase and gives an overly optimistic picture.

With a purpose of correction in overly optimistic picture, adjusted R^2, denoted as $\bar{R}^2$ or adj R^2 is used which is defined as

$$\bar{R}^2 = 1 - \frac{SS_{res}/(n-k)}{SS_T/(n-1)}$$

$$= 1 - \left(\frac{n-1}{n-k}\right)(1-R^2).$$

We will see later that $(n-k)$ and $(n-1)$ are the degrees of freedom associated with the distributions of SS_{res} and SS_T.

Moreover , the quantities $\frac{SS_{res}}{n-k}$ and $\frac{SS_T}{n-1}$ are based on the unbiased estimators of respective variances of e and y is the context of analysis of variance.

The adjusted R^2 will decline if the addition of an extra variable produces too small a reduction in $(1-R^2)$ to compensate for the increase is $\left(\frac{n-1}{n-k}\right)$.

Another limitation of adjusted R^2 is that it can be negative also. For example if $k=3, n=10, R^2=0.16,$ then

$$\bar{R}^2 = 1 - \frac{9}{7}\times 0.84 = -0.08 < 0$$

which has no interpretation.

Limitations

1. If constant term is absent in the model, then R^2 can not be defined. In such cases, R^2 can be negative. Some ad-hoc measures based on R^2 for regression line through origin have been proposed in the literature.
2. R^2 is sensitive to extreme values, so R^2 lacks robustness.
3. Consider a situation where we have following two models:

   The question is now which model is better?

$$y_i = \beta_1 + \beta_2 X_{i2} + ... + \beta_k X_{ik} + u_i, \quad i = 1, 2, .., n$$

$$\log y_i = \gamma_1 + \gamma_2 X_{i2} + ... + \gamma_k X_{ik} + v_i$$

For the first model, $R_1^2 = 1 - \dfrac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$

and for the second model, an option is to define R^2 as $R_2^2 = 1 - \dfrac{\sum_{i=1}^{n} (\log y_i - \log \hat{y}_i)^2}{\sum_{i=1}^{n} (\log y_i - \log \bar{y})^2}$.

As such R_1^2 and R_2^2 are not comparable.

If still, the two models are needed to be compared, a better proposition to define R^2 can be as follows:

$$R_3^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - anti\log\ \hat{y}_i^*)}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where $y_i^* = \widehat{\log y_i}$. Now R_1^2 and R_3^2 on comparison may give an idea about the adequacy of the two models.

Relationship of analysis of variance test and coefficient of determination

Assuming β_1 to be an intercept term, then for $H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$ the *F*-statistic in analysis of variance test is

$$F = \frac{MS_{reg}}{MS_{res}}$$

$$= \frac{(n-k)}{(k-1)} \frac{SS_{reg}}{SS_{res}} = \left(\frac{n-k}{k-1}\right) \frac{SS_{reg}}{SS_T - SS_{reg}} = \left(\frac{n-k}{k-1}\right) \frac{\frac{SS_{reg}}{SS_T}}{1 - \frac{SS_{reg}}{SS_T}}$$

$$= \left(\frac{n-k}{k-1}\right) \frac{R^2}{1-R^2}$$

where R^2 is the coefficient of determination.

So *F* and R^2 are closely related. When $R^2 = 0$, then $F = 0$.

In limit, when $R^2 = 1, F = \infty$ So both *F* and R^2 vary directly. Larger R^2 implies greater *F* value.

That is why the *F* test under analysis of variance is termed as the measure of overall significance of estimated regression.

It is also a test of significance of R^2. If *F* is highly significant, it implies that we can reject H_0, i.e. *y* is linearly related to *X's.*

Prediction of values of study variable

The prediction in multiple regression model has two aspects

1. Prediction of average value of study variable or mean response.
2. Prediction of actual value of study variable.

1. Prediction of average value of *y*

We need to predict $E(y)$ at a given $x_0 = (x_{01}, x_{02}, ..., x_{0k})'$.

The predictor as a point estimate is

$$p = x_0' b = x_0' (X'X)^{-1} X' y$$

$$E(p) = x_0' \beta.$$

So p is an unbiased predictor for $E(y)$.

Its variance is

$$Var(p) = E\left[p - E(y)\right]'\left[p - E(y)\right]$$

$$= \sigma^2 x_0' (X'X)^{-1} x_0.$$

The confidence interval on the mean response at a particular point, such as $x_{01}, x_{02}, ..., x_{0k}$ can be found as follows:

Define $x_0 = (x_{01}, x_{02}, ..., x_{0k})'$.

The fitted value at x_0 is $\hat{y}_0 = x_0' b$.

Then

$$E(\hat{y}_0) = x_0' \beta = E(y \mid x_0)$$

$$Var(\hat{y}_0) = \sigma^2 x_0' (X'X)^{-1} x_0$$

$$P\left[-t_{\frac{\alpha}{2}, n-k} \le \frac{\hat{y}_0 - E(y \mid x_0)}{\sqrt{\hat{\sigma}^2 x_0' (X'X)^{-1} x_0}} \le t_{\frac{\alpha}{2}, n-k} \right] = 1 - \alpha$$

$$P\left[\hat{y}_0 - t_{\frac{\alpha}{2}, n-k} \le \sqrt{\hat{\sigma}^2 x_0' (X'X)^{-1} x_0} \le E(y \mid x_0) \le \hat{y}_0 + t_{\frac{\alpha}{2}, n-k} \le \sqrt{\hat{\sigma}^2 x_0' (X'X)^{-1} x_0} \right] = 1 - \alpha.$$

The $100(1-\alpha)\%$ confidence interval on the mean response at the point $x_{01}, x_{02}, ..., x_{0k}$, i.e., $E(y \mid x_0)$ is

$$\left[\hat{y}_0 - t_{\frac{\alpha}{2}, n-k} \sqrt{\hat{\sigma}^2 x_0' (X'X)^{-1} x_0}, \hat{y}_0 + t_{\frac{\alpha}{2}, n-k} \le \sqrt{\hat{\sigma}^2 x_0' (X'X)^{-1} x_0} \right].$$

2. Prediction of actual value of *y*

We need to predict *y* at a given $x_0 = (x_{01}, x_{02}, ..., x_{0k})'$.

The predictor as a point estimate is

$$p_f = x_0' b$$

$$E(p_f) = x_0' \beta$$

So p_f is an unbiased predictor for y. It's variance is

$$Var(p_f) = E\left((p_f - y)(p_f - y)'\right)$$

$$= \sigma^2 \left[1 + x_0'(X'X)^{-1} x_0\right].$$

The $100(1-\alpha)\%$ confidence interval for this future observation is

$$\left(p_f - t_{\frac{\alpha}{2}, n-k} \sqrt{\hat{\sigma}^2 [1 + x_0'(X'X)^{-1} x_0]}, \quad p_f + t_{\frac{\alpha}{2}, n-k} \sqrt{\hat{\sigma}^2 [1 + x_0'(X'X)^{-1} x_0]} \right).$$