LINEAR REGRESSION ANALYSIS

MODULE – II

Lecture - 6

Simple Linear Regression Analysis

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Prediction of values of study variable

An important use of linear regression modeling is to predict the average and actual values of study variable. The term prediction of value of study variable corresponds to knowing the value of $E(y)$ (in case of average value) and value of y (in case of actual value) for a given value of explanatory variable. We consider both the cases.

Case 1: Prediction of average value

Under the linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$, the fitted model is $y = b_0 + b_1 x$ where b_0 and b_1 are the OLS estimators of β_0 and β_1 respectively.

Suppose we want to predict the value of $E(y)$ for a given value of $x = x_0$. Then the predictor is given by

$$\widehat{E(y \mid x_0)} = \hat{\mu}_{y|x_0} = b_0 + b_1 x_0.$$

Predictive bias

The prediction error is given as

$$\begin{aligned}\hat{\mu}_{y|x_0} - E(y) &= b_0 + b_1 x_0 - E(\beta_0 + \beta_1 x_0 + \varepsilon) \\ &= b_0 + b_1 x_0 - (\beta_0 + \beta_1 x_0) \\ &= (b_0 - \beta_0) + (b_1 - \beta_1) x_0.\end{aligned}$$

Then

$$\begin{aligned}E\left[\hat{\mu}_{y|x_0} - E(y)\right] &= E(b_0 - \beta_0) + E(b_1 - \beta_1) x_0 \\ &= 0 + 0 = 0.\end{aligned}$$

Thus the predictor $\mu_{y|x_0}$ is an unbiased predictor of $E(y)$.

Predictive variance

The predictive variance of $\hat{\mu}_{y|x_0}$ is

$$PV(\hat{\mu}_{y|x_0}) = Var(b_0 + b_1 x_0)$$

$$= Var\left[\bar{y} + b_1(x_0 - \bar{x})\right]$$

$$= Var(\bar{y}) + (x_0 - \bar{x})^2 Var(b_1) + 2(x_0 - \bar{x})Cov(\bar{y}, b_1)$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{s_{xx}} + 0$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right].$$

Estimate of predictive variance

The predictive variance can be estimated by substituting σ^2 by $\hat{\sigma}^2 = MSE$ as

$$\widehat{PV}(\hat{\mu}_{y|x_0}) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right]$$

$$= MSE\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right].$$

Prediction interval estimation

The $100(1-\alpha)\%$ prediction interval for $E(y \mid x_0)$ is obtained as follows:

The predictor $\hat{\mu}_{y|x_0}$ is a linear combination of normally distributed random variables, so it is also normally distributed as

$$\hat{\mu}_{y|x_0} \sim N\left(\beta_0 + \beta_1 x_0, PV\left(\hat{\mu}_{y|x_0}\right)\right).$$

So if σ^2 is known, then the distribution of

$$\frac{\hat{\mu}_{y|x_0} - E(y \mid x_0)}{\sqrt{PV(\hat{\mu}_{y|x_0})}}$$

is *N*(0,1), so the $100(1-\alpha)\%$ prediction interval is obtained as

$$P\left[-z_{\frac{\alpha}{2}} \le \frac{\hat{\mu}_{y|x_0} - E(y \mid x_0)}{\sqrt{PV(\hat{\mu}_{y|x_0})}} \le z_{\frac{\alpha}{2}}\right] = 1-\alpha$$

which gives the prediction interval for $E(y \mid x_0)$ as

$$\left[\hat{\mu}_{y|x_0} - z_{\frac{\alpha}{2}}\sqrt{\sigma^2\left[\frac{1}{n} + \frac{(x_0-\overline{x})^2}{s_{xx}}\right]},\ \hat{\mu}_{y|x_0} + z_{\frac{\alpha}{2}}\sqrt{\sigma^2\left[\frac{1}{n} + \frac{(x_0-\overline{x})^2}{s_{xx}}\right]}\right].$$

When σ^2 is unknown, it is replaced by $\hat{\sigma}^2 = MSE$ and in this case the sampling distribution of

$$\frac{\hat{\mu}_{y|x_0} - E(y \mid x_0)}{\sqrt{MSE\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{s_{xx}}\right]}}$$

is *t*-distribution with (*n* - 2) degrees of freedom, i.e., t_{n-2}.

The 100(1- α)% prediction interval in this case is

$$P\left[-t_{\frac{\alpha}{2},n-2} \le \frac{\hat{\mu}_{y|x_0} - E(y \mid x_0)}{\sqrt{MSE\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{s_{xx}}\right]}} \le t_{\frac{\alpha}{2},n-2}\right] = 1 - \alpha$$

which gives the prediction interval as

$$\left(\hat{\mu}_{y|x_0} - t_{\frac{\alpha}{2},n-2}\sqrt{MSE\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{s_{xx}}\right)},\ \ \hat{\mu}_{y|x_0} + t_{\frac{\alpha}{2},n-2}\sqrt{MSE\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{s_{xx}}\right)}\right).$$

Note that the width of prediction interval $E(y|x_0)$ is a function of x_0. The interval width is minimum for $x_0 = \overline{x}$ and widens as $|x_0 - \overline{x}|$ increases. This is expected also as the best estimates of *y* to be made at *x*-values lie near the center of the data and the precision of estimation to deteriorate as we move to the boundary of the *x*-space.

Case 2: Prediction of actual value

If x_0 is the value of the explanatory variable, then the actual value predictor for *y* is .

$$\hat{y}_0 = b_0 + b_1 x_0.$$

Note that the form of predictor is same as of average value predictor but its predictive error and other properties are different. This is the **dual nature of predictor.**

Predictive bias

Then the prediction error of $\hat{y}_0$ is given as

$$\hat{y}_0 - y_0 = b_0 + b_1 x_0 - (\beta_0 + \beta_1 x_0 + \varepsilon)$$

$$= (b_0 - \beta_0) + (b_1 - \beta_1) x_0 - \varepsilon.$$

Thus, we find that

$$E(\hat{y}_0 - y_0) = E(b_0 - \beta_0) + E(b_1 - \beta_1) x_0 - E(\varepsilon)$$

$$= 0 + 0 + 0 = 0$$

which implies that $\hat{y}_0$ is an unbiased predictor of *y*.

Predictive variance

Because the future observation y_0 is independent of $\hat{y}_0$, the predictive variance of $\hat{y}_0$ is

$$PV(\hat{y}_0) = E(\hat{y}_0 - y_0)^2$$

$$= E[(b_0 - \beta_0) + (x_0 - \bar{x})(b_1 - \beta_1) + (b_1 - \beta_1)\bar{x} - \varepsilon_0]^2$$

$$= Var(b_0) + (x_0 - \bar{x})^2 Var(b_1) + \bar{x}^2 Var(b_1) + Var(\varepsilon) + 2(x_0 - \bar{x})Cov(b_0, b_1) + 2\bar{x}Cov(b_0, b_1) + 2(x_0 - \bar{x})Var(b_1)$$

[rest of the terms are 0 assuming the independence of ε_0 with $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$]

$$= Var(b_0) + [(x_0 - \bar{x})^2 + \bar{x}^2 + 2(x_0 - \bar{x})]Var(b_1) + Var(\varepsilon) + 2[(x_0 - \bar{x}) + 2\bar{x}]Cov(b_0, b_1)$$

$$= Var(b_0) + x_0^2 Var(b_1) + Var(\varepsilon) + 2x_0 Cov(b_0, b_1)$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right] + x_0^2 \frac{\sigma^2}{s_{xx}} + \sigma^2 - 2x_0 \frac{\bar{x}\sigma^2}{s_{xx}}$$

$$= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right].$$

Estimate of predictive variance

The estimate of predictive variance can be obtained by replacing σ^2 by its estimate $\hat{\sigma}^2 = MSE$ as

$$\widehat{PV}(\hat{y}_0) = \hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right]$$

$$= MSE \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right].$$

Prediction interval

If σ^2 is known, then the distribution of

$$\frac{\hat{y}_0 - E(\hat{y}_0)}{\sqrt{PV(\hat{y}_0)}}$$

is $N(0,1)$. So the $100(1-\alpha)\%$ prediction interval is obtained as

$$P\left[-z_{\frac{\alpha}{2}} \le \frac{\hat{y}_0 - E(\hat{y}_0)}{\sqrt{PV(\hat{y}_0)}} \le z_{\frac{\alpha}{2}}\right] = 1-\alpha$$

which gives the prediction interval for $\hat{y}_0$ as

$$\left(\hat{y}_0 - z_{\frac{\alpha}{2}}\sqrt{\sigma^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right)},\ \hat{y}_0 + z_{\frac{\alpha}{2}}\sqrt{\sigma^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right)}\right).$$

When σ^2 is unknown, then

$$\frac{\hat{y}_0 - E(\hat{y}_0)}{\sqrt{\widehat{PV}(\hat{y}_0)}}$$

follows a *t*-distribution with (*n* - 2) degrees of freedom.

The $100(1-\alpha)\%$ prediction interval for $\hat{y}_0$ in this case is obtained as

$$P\left[-t_{\frac{\alpha}{2},n-2} \le \frac{\hat{y}_0 - E(\hat{y}_0)}{\sqrt{\widehat{PV}(\hat{y}_0)}} \le t_{\frac{\alpha}{2},n-2}\right] = 1-\alpha$$

which gives the prediction interval

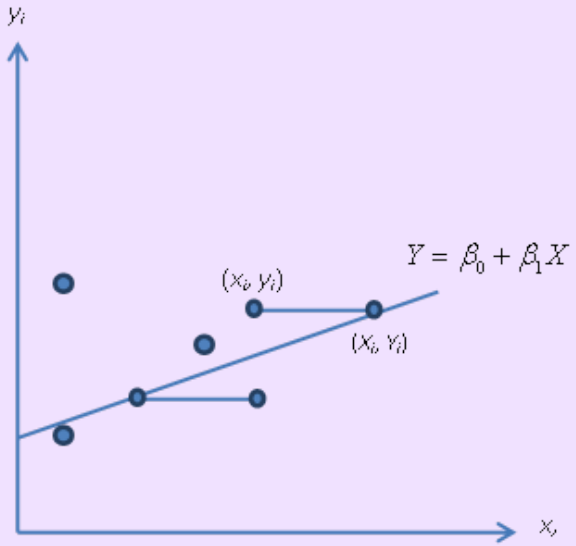
$$\left[\hat{y}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{MSE\left(1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{s_{xx}}\right)},\ \hat{y}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{MSE\left(1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{s_{xx}}\right)}\right].$$

The prediction interval is of minimum width at $x_0 = \overline{x}$ and widens as $\left|x_0 - \overline{x}\right|$ increases.

The prediction interval for $\hat{y}_0$ is wider than the prediction interval for $\hat{\mu}_{y|x_0}$ because the prediction interval for $\hat{y}_0$ depends on both the error from the fitted model as well as the error associated with the future observations.

Reverse regression method

The reverse (or inverse) regression approach minimizes the sum of squares of horizontal distances between the observed data points and the line in the following scatter diagram to obtain the estimates of regression parameters.



Reverse regression method

The reverse regression has been advocated in the analysis of sex (or race) discrimination in salaries. For example, if *y* denotes salary and *x* denotes qualifications and we are interested in determining if there is a sex discrimination in salaries, we can ask:

> **"**Whether men and women with the same qualifications (value of *x*) are getting the same salaries (value of *y*). This question is answered by the **direct regression**."

Alternatively, we can ask:

> "Whether men and women with the same salaries (value of *y*) have the same qualifications (value of *x*). This question is answered by the **reverse regression**, i.e., regression of *x* on *y*."

The regression equation in case of reverse regression can be written as $x_i = \beta_0^* + \beta_1^* y_i + \delta_i \quad (i = 1, 2, ..., n)$

where δ_i 's are the associated random error components and satisfy the assumptions as in the case of usual simple linear regression model.

The reverse regression estimates $\hat{\beta}_{OR}$ of β_0^* and $\hat{\beta}_{1R}$ of β_1^* for the model are obtained by interchanging the *x* and *y* in the direct regression estimators of β_0 and β_1. The estimates are obtained as

$$\hat{\beta}_{OR} = \overline{x} - \hat{\beta}_{1R}\overline{y}$$

and

$$\hat{\beta}_{1R} = \frac{s_{xy}}{s_{yy}}$$

for β_0^* and β_1^* respectively.

The residual sum of squares in this case is

$$SS_{res}^* = s_{xx} - \frac{s_{xy}^2}{s_{yy}}.$$

Note that

$$\hat{\beta}_{1R} b_1 = \frac{s_{xy}^2}{s_{xx} s_{yy}} = r_{xy}^2$$

where b_1 is the direct regression estimator of slope parameter and r_{xy} is the correlation coefficient between *x* and *y.* Hence if r_{xy}^2 is close to 1, the two regression lines will be close to each other.

An important application of reverse regression method is in solving the calibration problem.