LINEAR REGRESSION ANALYSIS

MODULE – II

Lecture - 4

Simple Linear Regression Analysis

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Maximum likelihood estimation

We assume that ε_i's $(i=1,2,...,n)$ are independent and identically distributed following a normal distribution $N(0,\sigma^2)$. Now we use the method of maximum likelihood to estimate the parameters of the linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i=1,2,...,n),$$

the observations y_i $(i=1,2,...,n)$ are independently distributed with $N(\beta_0+\beta_1 x_i, \sigma^2)$ for all $i=1,2,...,n$. The likelihood function of the given observations (x_i, y_i) and unknown parameters β_0, β_1 and σ^2 is

$$L(x_i, y_i; \beta_0, \beta_1, \sigma^2 = \prod_{i=1}^{n}\left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_1 x_i)^2\right].$$

The maximum likelihood estimates of β_0, β_1 and σ^2 can be obtained by maximizing $L(x_i, y_i; \beta_0, \beta_1, \sigma^2)$ or equivalently $\ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)$ where

$$\ln\ L(x_i, y_i; \beta_0, \beta_1, \sigma^2) = -\left(\frac{n}{2}\right)\ln 2\pi - \left(\frac{n}{2}\right)\ln\sigma^2 - \left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2.$$

The normal equations are obtained by partial differentiation of log-likelihood with respect to β_0, β_1 and σ^2 equating them to zero

$$\frac{\partial \ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)}{\partial \beta_0} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i) = 0$$

$$\frac{\partial \ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)}{\partial \beta_1} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)x_i = 0$$

and

$$\frac{\partial \ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2 = 0.$$

The solution of these normal equations give the maximum likelihood estimates of β_0, β_1 and σ^2 as

$$\tilde{b}_0 = \bar{y} - \tilde{b}_1 \bar{x}$$

$$\tilde{b}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}}$$

and

$$\tilde{s}^2 = \frac{\sum_{i=1}^{n}(y_i - \tilde{b}_0 - \tilde{b}_1 x_i)^2}{n},$$

respectively.

It can be verified that the Hessian matrix of second order partial derivation of ln L with respect to β_0, β_1, and σ^2 is negative definite at $\beta_0 = \tilde{b}_0$, $\beta_1 = \tilde{b}_1$, and $\sigma^2 = \tilde{s}^2$ which ensures that the likelihood function is maximized at these values.

Note that the least squares and maximum likelihood estimates of β_0 and β_1 are identical when disturbances are normally distributed. The least squares and maximum likelihood estimates of σ^2 are different. In fact, the least squares estimate of σ^2 is

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

so that it is related to maximum likelihood estimate as $\tilde{s}^2 = \frac{n-2}{n}s^2$.

Thus $\tilde{b}_0$ and $\tilde{b}_1$ are unbiased estimators of β_0 and β_1 whereas $\tilde{s}^2$ is a biased estimate of σ^2, but it is asymptotically unbiased. The variances of $\tilde{b}_0$ and $\tilde{b}_1$ are same as that of b_0 and b_1 respectively but the mean squared error

$$MSE(\tilde{s}^2) < Var(s^2).$$

Testing of hypotheses and confidence interval estimation for slope parameter

Now we consider the tests of hypothesis and confidence interval estimation for the slope parameter of the model under two cases, viz., when σ^2 is known and when σ^2 is unknown.

Case 1: When σ^2 is known

Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, 2, ..., n)$. It is assumed that ε_i's are independent and identically distributed and follow $N(0, \sigma^2)$.

First we develop a test for the null hypothesis related to the slope parameter

$$H_0 : \beta_1 = \beta_{10}$$

where β_{10} is some given constant.

Assuming σ^2 to be known, we know that

$$E(b_1) = \beta_1, \; Var(b_1) = \frac{\sigma^2}{s_{xx}}$$

and b_1 is a linear combination of normally distributed y_i's, so

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right)$$

and so the following statistic can be constructed

$$Z_1 = \frac{b_1 - \beta_{10}}{\sqrt{\frac{\sigma^2}{s_{xx}}}}$$

which is distributed as $N(0, 1)$ when H_0 is true.

A decision rule to test $H_1 : \beta_1 \neq \beta_{10}$ can be framed as follows:

Reject H_0 if $|Z_0| > z_{\alpha/2}$

where $z_{\alpha/2}$ is the $\alpha/2$ percentage points on normal distribution. Similarly, the decision rule for one sided alternative hypothesis can also be framed.

The $100(1-\alpha)\%$ confidence interval for β_1 can be obtained using the Z_1 statistic as follows:

$$P\left[-z_{\frac{\alpha}{2}} \leq Z_1 \leq z_{\frac{\alpha}{2}}\right] = 1-\alpha$$

$$P\left[-z_{\frac{\alpha}{2}} \leq \frac{b_1-\beta_1}{\sqrt{\frac{\sigma^2}{s_{xx}}}} \leq z_{\frac{\alpha}{2}}\right] = 1-\alpha$$

$$P\left[b_1 - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{s_{xx}}} \leq \beta_1 \leq b_1 + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{s_{xx}}}\right] = 1-\alpha.$$

So $100(1-\alpha)\%$ confidence interval for β_1 is

$$\left(b_1 - z_{\alpha/2}\sqrt{\frac{\alpha^2}{s_{xx}}}, b_1 + z_{\alpha/2}\sqrt{\frac{\alpha^2}{s_{xx}}}\right)$$

where $z_{\alpha/2}$ is the $\alpha/2$ percentage point of the $N(0,1)$ distribution.

Case 2: When σ^2 is unknown

When σ^2 is unknown, we proceed as follows. We know that

$$\frac{SS_{res}}{\sigma^2} \sim \chi^2(n-2).$$

and

$$E\left(\frac{SS_{res}}{n-2}\right) = \sigma^2.$$

Further, SS_{res} / σ^2 and b_1 are independently distributed. This result will be proved formally later in module on multiple linear regression. This result also follows from the result that under normal distribution, the maximum likelihood estimates, viz., sample mean (estimator of population mean) and sample variance (estimator of population variance) are independently distributed so b_1 and s^2 are also independently distributed.

Thus the following statistic can be constructed:

$$t_0 = \frac{b_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}}$$

$$= \frac{b_1 - \beta_1}{\sqrt{\frac{SS_{res}}{(n-2)s_{xx}}}} \sim t_{n-2}$$

which follows a t-distribution with $(n - 2)$ degrees of freedom, denoted as t_{n-2} , when H_0 is true.

A decision rule to test $H_1 : \beta_1 \neq \beta_{10}$ is to

reject H_0 if $|t_0| > t_{n-2,\alpha/2}$

where $t_{n-2,\alpha/2}$ is the $\alpha/2$ percentage point of the *t*-distribution with (*n* - 2) degrees of freedom.

Similarly, the decision rule for one sided alternative hypothesis can also be framed.

The $100(1-\alpha)\%$ confidence interval of β_1 can be obtained using the t_0 statistic as follows :

$$P\left[-t_{\frac{\alpha}{2}} \leq t_0 \leq t_{\frac{\alpha}{2}}\right] = 1-\alpha$$

$$P\left[-t_{\frac{\alpha}{2}} \leq \frac{b_1-\beta_1}{\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}} \leq t_{\frac{\alpha}{2}}\right] = 1-\alpha$$

$$P\left[b_1 - t_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}} \leq \beta_1 \leq b_1 + t_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}\right] = 1-\alpha.$$

So the $100(1-\alpha)\%$ confidence interval β_1 is

$$\left(b_1 - t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{(n-2)s_{xx}}}, b_1 + t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{(n-2)s_{xx}}}\right).$$

Testing of hypotheses and confidence interval estimation for intercept term

Now, we consider the tests of hypothesis and confidence interval estimation for intercept term under two cases, viz., when σ^2 is known and when σ^2 is unknown.

Case 1: When σ^2 is known

Suppose the null hypothesis under consideration is $H_0 : \beta_0 = \beta_{00}$,

where σ^2 is known, then using the result that $E(b_0) = \beta_0, Var(b_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_x}\right)$ and b_0 is a linear combination of normally distributed random variables, the following statistic

$$Z_0 = \frac{b_0 - \beta_{00}}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}} \sim N(0,1),$$

has a $N(0, 1)$ distribution when H_0 is true.

A decision rule to test $H_1 : \beta_0 \neq \beta_{00}$ can be framed as follows:

Reject H_0 if $|Z_0| > z_{\alpha/2}$

where $z_{\alpha/2}$ is the $\alpha/2$ percentage points on normal distribution.

Similarly, the decision rule for one sided alternative hypothesis can also be framed.

The $100(1-\alpha)\%$ confidence intervals for β_0 when σ^2 is known can be derived using the Z_0 statistic as follows::

$$P\left[-z_{\frac{\alpha}{2}} \le Z_0 \le z_{\frac{\alpha}{2}}\right] = 1-\alpha$$

$$P\left[-z_{\frac{\alpha}{2}} \le \frac{b_0-\beta_0}{\sqrt{\left(\frac{1}{n}+\frac{\overline{x}^2}{s_{xx}}\right)}} \le z_{\frac{\alpha}{2}}\right] = 1-\alpha$$

$$P\left[b_0 - z_{\frac{\alpha}{2}}\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{s_{xx}}\right)} \le \beta_0 \le b_0 + z_{\frac{\alpha}{2}}\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{s_{xx}}\right)}\right] = 1-\alpha.$$

So the $100(1-\alpha)\%$ of confidential interval of β_0 is

$$\left(b_0 - z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{s_{xx}}\right)}, b_0 + z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{s_{xx}}\right)}\right).$$

Case 2: When σ^2 is unknown

When σ^2 is unknown, then the statistic is constructed

$$t_0 = \frac{b_0 - \beta_{00}}{\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}}$$

which follows a *t*-distribution with (*n* - 2) degrees of freedom, i.e., t_{n-2} when H_0 is true.

A decision rule to test $H_1 : \beta_0 \neq \beta_{00}$ is as follows:

Reject H_0 whenever $|t_0| > t_{n-2,\alpha/2}$

where $t_{n-2,\alpha/2}$ is the $\alpha/2$ percentage point of the *t*-distribution with (*n* - 2) degrees of freedom.

Similarly, the decision rule for one sided alternative hypothesis can also be framed.

The $100(1-\alpha)\%$ of confidential interval of β_0 can be obtained as follows:

Consider

$$P\left[t_{n-2,\alpha/2} \le t_0 \le t_{n-2,\alpha/2}\right] = 1-\alpha$$

$$P\left[t_{n-2,\alpha/2} \le \frac{b_0-\beta_0}{\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}\right)}} \le t_{n-2,\alpha/2}\right] = 1-\alpha$$

$$P\left[b_0 - t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}\right)} \le \beta_0 \le b_0 + t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}\right)}\right] = 1-\alpha.$$

The $100(1-\alpha)\%$ confidential interval for β_0 is

$$\left[b_0 - t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}\right)}, b_0 + t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}\right)}\right].$$

Confidence interval for σ^2

A confidence interval for σ^2 can also be derived as follows. Since $SS_{res}/\sigma^2 \sim \chi^2_{n-2}$, thus consider

$$P\left[\chi^2_{n-2,\alpha/2} \le \frac{SS_{res}}{\sigma^2} \le \chi^2_{n-2,1-\alpha/2}\right] = 1-\alpha$$

$$P\left[\frac{SS_{res}}{\chi^2_{n-2,1-\alpha/2}} \le \sigma^2 \le \frac{SS_{res}}{\chi^2_{n-2,\alpha/2}}\right] = 1-\alpha.$$

The corresponding $100(1-\alpha)\%$ confidence interval for σ^2 is $\left(\frac{SS_{res}}{\chi^2_{n-2,1-\alpha/2}}, \frac{SS_{res}}{\chi^2_{n-2,\alpha/2}}\right)$.