Simple Linear Regression

Dr. Kiah Wah Ong

ILLINOIS TECH

Estimation of σ^2

Recall from the previous video, the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ is given by

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right)$$

and

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}.$$

Since both of these values involve the variance σ^2 of the error terms ϵ_i in the regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

an estimate of σ^2 is required.

Estimation of σ^2

Let us look at the quantities

$$y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, \cdots, n$$

which represent the differences between the actual responses (y_i) and their least squares estimators ($\hat{\beta}_0 + \hat{\beta}_1 x_i$).

We sometimes also write the above expression as $y_i - \hat{y}_i$ where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is called the fitted (or predicted) values.

The estimation of σ^2 is obtained from the residual sum of squares

$$SS_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Estimation of σ^2

It can be shown that $\frac{SS_R}{\sigma^2}$ follows a chi-square distribution with $n-2$ degree of freedom. That is,

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}.$$

Since a random variable with a chi-square distribution with k degree of freedom has its mean equal to k, we see that

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

From here, it implies

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

Estimation of σ^2

From

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

we conclude that $\dfrac{SS_R}{n-2}$ is an **unbiased estimator** of σ^2, and write

$$\hat{\sigma}^2 = \frac{SS_R}{n-2}.$$

Inferences Concerning β_0 and β_1

An important hypothesis to consider regarding the simple linear regression model

$$y = \beta_0 + \beta_1 x + \epsilon$$

is the hypothesis that $\beta_1 = 0$

This is the same as saying that the mean response $E(y)$ does not depend on the input x, or equivalently, that there is no regression on the input variable.

We also want to perform inference on β_0 as well.

How do we do all that?

Inferences Concerning β_0 and β_1

To do that, we need a distribution for the unknown β_0 and β_1.

This is the point where we need to make additional assumption that the model errors ϵ_i are normally distributed.

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Inferences Concerning β_0 and β_1

Since the errors

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

the observations y_i are now normal and independently distributed with mean $\beta_0 + \beta_1 x_i$ and variance σ^2, and we write

$$y_i \sim \mathrm{NID}(\beta_0 + \beta_1 x_i, \sigma^2)$$

Inferences Concerning β_0 and β_1

Recall that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{S_{xx}} = \sum_{i=1}^{n} c_i y_i$$

where

$$c_i = \frac{x_i - \overline{x}}{S_{xx}}, \quad i = 1, \cdots, n$$

hence from

$$y_i \sim \mathrm{NID}(\beta_0 + \beta_1 x_i, \sigma^2)$$

we know that $\hat{\beta}_1$ is also normally distributed with mean β_1 and variance σ^2/S_{xx} as shown before. That is

$$\hat{\beta}_1 \sim \mathrm{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Inferences Concerning β_0 and β_1

Since $\hat{\beta}_0$ can also be written as a linear combination of y_i, the distribution of $\hat{\beta}_0$ will also be normal with mean and variance as shown below

$$\hat{\beta}_0 \sim \mathrm{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right)$$

Inferences Concerning β_0 and β_1

Since

$$\hat{\beta}_1 \sim \mathrm{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

we have

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim \mathrm{N}(0,1)$$

hence

$$\sqrt{S_{xx}}\frac{\hat{\beta}_1 - \beta_1}{\sigma} \sim \mathrm{N}(0,1)$$

and it is actually independent of

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

Inferences Concerning β_0 and β_1

Recall from your previous statistics classes:

If Z and X_n^2 are independent r.v. with $Z \sim \mathrm{N}(0,1)$ and X_n^2 with chi-square with n degree of freedom , then the r.v.

$$T_n := \frac{Z}{\sqrt{X_n^2/n}} \sim t_n$$

that is T_n has a t-distribution with n degree of freedom.

Inferences Concerning β_0 and β_1

Applying this to

$$\sqrt{S_{xx}}\frac{\hat{\beta}_1 - \beta_1}{\sigma} \sim \mathrm{N}(0,1) \quad \text{and} \quad \frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

we obtain

$$\begin{aligned} T_{n-2} &:= \frac{\sqrt{S_{xx}}(\hat{\beta}_1 - \beta_1)/\sigma}{\sqrt{\frac{SS_R/\sigma^2}{n-2}}} \\ &= \sqrt{\frac{(n-2)S_{xx}}{SS_R}}(\hat{\beta}_1 - \beta_1) \sim t_{n-2} \end{aligned}$$

Inferences Concerning β_0 and β_1

Back to the question on whether $\beta_1 = 0$ for

$$y = \beta_0 + \beta_1 x + \epsilon$$

Let us perform hypothesis testing:

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0$$

A significance level α test for H_0 is to

$$\text{Reject} \quad H_0 \quad \text{if} \quad \sqrt{\frac{(n-2)S_{xx}}{SS_R}}|\hat{\beta}_1| > t_{\alpha/2,n-2}$$

$$\text{accept} \quad H_0 \quad \text{otherwise}$$

Inferences Concerning β_0 and β_1

That is, the test can be performed by computing the value of the test statistics

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}|\hat{\beta}_1|$$

and call the value ν, then rejecting H_0 if

$$P(|T_{n-2}| > \nu) = 2P(T_{n-2} > \nu) \leq \alpha$$

Inferences Concerning β_0 and β_1

It follows from

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}(\hat{\beta}_1 - \beta_1) \sim t_{n-2}$$

that, for any $\alpha, 0 < \alpha < 1$, a $100(1-\alpha)\%$ confidence interval for the estimator β_1 is

$$\left(\hat{\beta}_1 - \sqrt{\frac{SS_R}{(n-2)S_{xx}}}\ t_{\alpha/2,n-2},\ \ \hat{\beta}_1 + \sqrt{\frac{SS_R}{(n-2)S_{xx}}}\ t_{\alpha/2,n-2}\right)$$