

Ma 2b

Notes on Simple Linear Regression G. Lorden

These notes are a supplement to the handout on regression analysis, which contains the general information,

"Simple linear regression" refers to the problem of "fitting a line" to (x, y) data:

assume $(x_1, Y_1), \dots, (x_n, Y_n)$ are given, where

x_1, \dots, x_n are known constants
and Y_1, \dots, Y_n are independent $N(EY_i, \sigma^2)$
where $\sigma^2 > 0$ is unknown (the same for all Y_i 's)
and the true means, $\{EY_i\}$, lie on a line, i.e.,

$$EY_i = \alpha + \beta x_i,$$

where α, β are unknown parameters.

Note that this fits the setup in the general regression problem,

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

where $\varepsilon_1, \dots, \varepsilon_n$ are iid $\sim N(0, \sigma^2)$.

Using $X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ in the general formulas

we get

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (X'X)^{-1} X'Y.$$

It is a matter of straightforward calculation to derive

$$(X'X) = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & n\bar{x}^2 \end{pmatrix},$$

where $\bar{x} = (x_1 + \dots + x_n)/n$ and $\bar{x}^2 = (x_1^2 + \dots + x_n^2)/n$.

Using the usual formula for the inverse of a 2×2 matrix, it turns out that

$$(X'X)^{-1} = \frac{1}{n(\bar{x}^2 - \bar{x}^2)} \begin{pmatrix} \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}.$$

Since $X'Y = \begin{pmatrix} n\bar{y} \\ n\bar{xy} \end{pmatrix}$, where $\bar{xy} = \frac{\sum x_i y_i}{n}$,

plugging into the formula for $\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}$ yields

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \frac{1}{\bar{x}^2 - \bar{x}^2} \begin{pmatrix} \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} \bar{y} \\ \bar{xy} \end{pmatrix},$$

so $\hat{\alpha} = \frac{\bar{x}^2 \bar{y} - \bar{x} \bar{xy}}{\bar{x}^2 - \bar{x}^2}$ and $\hat{\beta} = \frac{\bar{xy} - \bar{x} \bar{y}}{\bar{x}^2 - \bar{x}^2}$.

It can be checked that

$$(*) \quad \hat{\alpha} + \hat{\beta} \bar{x} = \bar{y},$$

so the point (\bar{x}, \bar{y}) lies on the "fitted line",

whose equation is $\hat{\alpha} + \hat{\beta}x = y$.

Using the formula for $\hat{\beta}$ along with $(*)$ determines $\hat{\alpha}$, so the formula for $\hat{\alpha}$ isn't needed.

The diagonal entries of $(X'X)^{-1}$ are

$$d_1 = \frac{\bar{x}^2}{n(\bar{x}^2 - \bar{x}^2)} \quad \text{and} \quad d_2 = \frac{1}{n(\bar{x}^2 - \bar{x}^2)}.$$

So the Student's t confidence interval for α is (since the number of " β 's" is $r=2$)

$$\hat{\alpha} \pm t_{n-2, 1-\frac{\alpha}{2}} \sqrt{\frac{\bar{x}^2}{n(\bar{x}^2 - \bar{x}^2)}},$$

where

$$s^2 = \frac{SSR}{n-2} = \frac{\sum_i (y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n-2}.$$

Similarly, the Student's t confidence interval for β is

$$\hat{\beta} \pm t_{n-2, 1-\frac{\alpha}{2}} \sqrt{\frac{1}{n(\bar{x}^2 - \bar{x}^2)}}.$$

As always, testing

$H: \beta = b$ (given) vs. $K: \beta \neq b$
at the α ($= .05$ or whatever) level of significance
can be carried out by the rule

"Reject H if b is not in the $100(1-\alpha)\%$
confidence interval for β ."

Equivalently, we can reject H if $|t| \geq t_{n-2, 1-\frac{\alpha}{2}}$

where

$$t = \frac{\hat{\beta} - b}{\sqrt{\frac{1}{n(\bar{x}^2 - \bar{x}^2)}}}$$

Often in fitting a straight line, one is interested in giving a confidence interval for "the true mean at the point x ", where x is any point, and may or may not be a value of one or more of the given x_i 's. That means that we want a confidence interval for $\alpha + \beta x$. The general notes derive a confidence interval for any linear combination of the regression coefficients - i.e. in this case

$$c_1 \alpha + c_2 \beta, \text{ where } c_1 = 1 \text{ and } c_2 = x.$$

Applying that result gives the confidence interval (with confidence level $100(1-\alpha)\%$)

$$\alpha + \beta x \in \hat{\alpha} + \hat{\beta}x \pm t_{n-2, 1-\frac{\alpha}{2}} \Delta W,$$

$$\text{where } W = \sqrt{(c_1, c_2)(X'X)^{-1} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}}$$

$$= \sqrt{(1 \ x) \begin{pmatrix} \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix} \cdot \frac{1}{n(\bar{x}^2 - \bar{x}^2)}}$$

$$\text{(after algebra)} = \sqrt{\frac{1}{n} \left(1 + \frac{(x - \bar{x})^2}{\bar{x}^2 - \bar{x}^2} \right)}$$

Note

- 1) We don't really need to use these special formulas for "simple linear regression", since we can just "plug in" to the general notes' matrix and vector calculations.
- 2) But the formula for w (on this page) is informative, since it shows that the confidence interval is wider as x moves away from \bar{x} .