

Linear Statistical Models: The Method of Least Squares

In the sections that follow, we will discuss inferential procedures that are used when a response variable Y is a linear function of a single independent variable x . We will assume that the response variable is related to the independent variable by the simple linear model $Y_i = \mathbf{b}_0 + \mathbf{b}_1 x_i + \mathbf{e}_i$ where \mathbf{b}_0 and \mathbf{b}_1 are parameters and \mathbf{e}_i represents random error with $E(\mathbf{e}_i) = 0$. The notation Y_i represents some future observable value while y_i represents an observed value.

When we fit a model to a particular set of data, we estimate the parameters and develop a best fit line denoted by $\hat{y}_i = \hat{\mathbf{b}}_0 + \hat{\mathbf{b}}_1 x_i$. The *least squares procedure* for fitting a line through a set of n data points determines $\hat{\mathbf{b}}_0$ and $\hat{\mathbf{b}}_1$ so that the sum of squared errors

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left[y_i - (\hat{\mathbf{b}}_0 + \hat{\mathbf{b}}_1 x_i) \right]^2$$

is minimized with respect to $\hat{\mathbf{b}}_0$ and $\hat{\mathbf{b}}_1$. To find these values, we can take partial derivatives of SSE with respect to $\hat{\mathbf{b}}_0$ and then with respect to $\hat{\mathbf{b}}_1$. Then we can set the partial derivatives equal to zero and solve for the values of $\hat{\mathbf{b}}_0$ and $\hat{\mathbf{b}}_1$. This work is outlined below:

$$\begin{aligned} SSE &= \sum_{i=1}^n \left[y_i - (\hat{\mathbf{b}}_0 + \hat{\mathbf{b}}_1 x_i) \right]^2 \\ \frac{\partial SSE}{\partial \hat{\mathbf{b}}_0} &= 2 \sum_{i=1}^n \left[y_i - (\hat{\mathbf{b}}_0 + \hat{\mathbf{b}}_1 x_i) \right] (-1) \\ &= -2 \left(\sum_{i=1}^n y_i - n \hat{\mathbf{b}}_0 - \hat{\mathbf{b}}_1 \sum_{i=1}^n x_i \right) \\ &= -2n\bar{y} + 2n\hat{\mathbf{b}}_0 + 2n\hat{\mathbf{b}}_1 \bar{x}, \text{ since } \bar{y} = \frac{\sum y_i}{n} \text{ we can write } \sum y_i = n\bar{y}. \end{aligned}$$

Setting this partial derivative equal to zero yields:

$$\bar{y} = \hat{\mathbf{b}}_0 + \hat{\mathbf{b}}_1 \bar{x} \quad \text{or} \quad \hat{\mathbf{b}}_0 = \bar{y} - \hat{\mathbf{b}}_1 \bar{x}$$

$$\begin{aligned} \frac{\partial SSE}{\partial \hat{\mathbf{b}}_1} &= 2 \sum_{i=1}^n \left[y_i - (\hat{\mathbf{b}}_0 + \hat{\mathbf{b}}_1 x_i) \right] (-x_i) \\ &= -2 \left(\sum_{i=1}^n x_i y_i - \hat{\mathbf{b}}_0 \sum_{i=1}^n x_i - \hat{\mathbf{b}}_1 \sum_{i=1}^n x_i^2 \right) \end{aligned}$$

Setting this partial derivative equal to zero yields:

$$\begin{aligned} \hat{\mathbf{b}}_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i - \hat{\mathbf{b}}_0 \sum_{i=1}^n x_i \\ &= \sum_{i=1}^n x_i y_i - (\bar{y} - \hat{\mathbf{b}}_1 \bar{x}) \sum_{i=1}^n x_i \\ &= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i + \hat{\mathbf{b}}_1 \bar{x} \sum_{i=1}^n x_i \end{aligned}$$

$$\begin{aligned} \text{So } \hat{\mathbf{b}}_1 \sum_{i=1}^n x_i^2 - \hat{\mathbf{b}}_1 \bar{x} \sum_{i=1}^n x_i &= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i \\ \hat{\mathbf{b}}_1 &= \frac{\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i} = \frac{\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\ &= \frac{\sum_{i=1}^n x_i y_i - n\bar{x} \cdot \bar{y} - \bar{y} \sum_{i=1}^n x_i + n\bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2} \\ &= \frac{\sum_{i=1}^n x_i y_i - \bar{x} \sum_{i=1}^n y_i - \bar{y} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x}^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

When using the model $Y_i = \mathbf{b}_0 + \mathbf{b}_1 x_i + \mathbf{e}_i$, we assume that there is a linear relationship between x and $E(Y)$ with a true slope \mathbf{b}_1 and a true intercept \mathbf{b}_0 . Because of the error term \mathbf{e}_i , observations with equal x -values will not necessarily have equal y -values. We have stated previously that $E(\mathbf{e}_i) = 0$. We will now assume that $\mathbf{e}_i \sim N(0, \mathbf{s}^2)$.

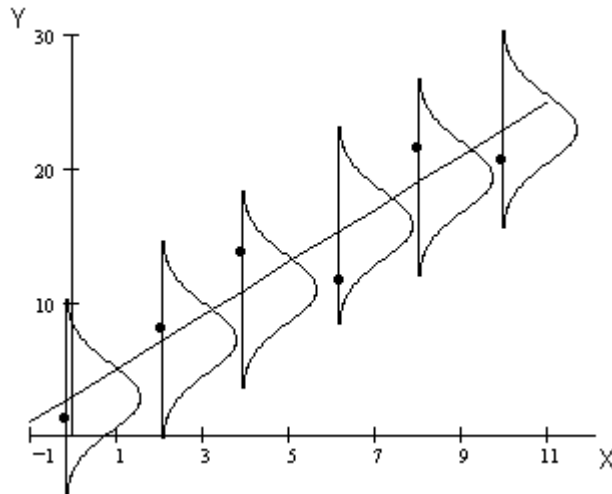


Figure 14: Graphical Representation of $Y_i \sim N(\mathbf{b}_0 + \mathbf{b}_1 x_i, \mathbf{s}^2)$

That is, for each value of x_i we assume that the errors are normally distributed with mean 0 and constant variance \mathbf{s}^2 . These assumptions about the distribution of the error terms underlie the distribution of Y_i for a fixed x_i . Specifically, for a fixed x_i , $Y_i \sim N(\mathbf{b}_0 + \mathbf{b}_1 x_i, \mathbf{s}^2)$.

The values Y_1, Y_2, \dots, Y_n are independent but are not identically distributed, since they have different values of \mathbf{m}_i . Here, $\mathbf{m}_i = \mathbf{b}_0 + \mathbf{b}_1 x_i$.

Note that:

$$\begin{aligned} E(Y_i) &= E(\mathbf{b}_0 + \mathbf{b}_1 x_i + \mathbf{e}_i) \\ &= E(\mathbf{b}_0) + E(\mathbf{b}_1 x_i) + E(\mathbf{e}_i) \\ &= \mathbf{b}_0 + \mathbf{b}_1 x_i + 0 \end{aligned}$$

$$\begin{aligned} V(Y_i) &= V(\mathbf{b}_0 + \mathbf{b}_1 x_i + \mathbf{e}_i) \\ &= V(\mathbf{b}_0) + V(\mathbf{b}_1 x_i) + V(\mathbf{e}_i) \\ &= 0 + 0 + \mathbf{s}^2 \end{aligned}$$

It can be shown that the least squares estimators of \mathbf{b}_0 and \mathbf{b}_1 are also maximum likelihood estimators:

$$L(\mathbf{b}_0, \mathbf{b}_1) = L(y_1, y_2, \dots, y_n | \mathbf{b}_0, \mathbf{b}_1)$$

Since Y_1, Y_2, \dots, Y_n are independent, we can multiply the density functions for each y_i .

$$\begin{aligned} L(\mathbf{b}_0, \mathbf{b}_1) &= \prod_{i=1}^n \frac{1}{\sqrt{2\mathbf{p}}} \frac{1}{\mathbf{s}} e^{\left(\frac{-1}{2\mathbf{s}^2}[y_i - \mathbf{m}_i]^2\right)} \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\mathbf{p}}} \frac{1}{\mathbf{s}} e^{\left(\frac{-1}{2\mathbf{s}^2}[y_i - (\mathbf{b}_0 + \mathbf{b}_1 x_i)]^2\right)} \\ &= \left(\frac{1}{\sqrt{2\mathbf{p}}} \frac{1}{\mathbf{s}}\right)^n e^{\left(\frac{-1}{2\mathbf{s}^2} \sum [y_i - (\mathbf{b}_0 + \mathbf{b}_1 x_i)]^2\right)} \end{aligned}$$

We want to maximize the likelihood with respect to \mathbf{b}_0 and \mathbf{b}_1 . Since the exponent on e is negative, large values of $L(\mathbf{b}_0, \mathbf{b}_1)$ are associated with small values of $\sum [y_i - (\mathbf{b}_0 + \mathbf{b}_1 x_i)]^2$. So to obtain the maximum likelihood estimator we need to minimize $\sum [y_i - (\mathbf{b}_0 + \mathbf{b}_1 x_i)]^2$, which is what we already did to obtain the least squares estimators $\hat{\mathbf{b}}_0$ and $\hat{\mathbf{b}}_1$! So, $\hat{\mathbf{b}}_0$ and $\hat{\mathbf{b}}_1$ are maximum likelihood estimators under the assumption $\mathbf{e} \sim N(0, \mathbf{s}^2)$.

Standardized Variables

Standardized variables offer another approach to the same problem. It is often advantageous to standardize the variables with the transformations

$$x_i^* = \frac{x_i - \bar{x}}{s_x} \quad \text{and} \quad y_i^* = \frac{y_i - \bar{y}}{s_y}.$$

Theorem: Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, $n > 1$ be a set of data, with $s_x, s_y > 0$. Then if $x_i^* = \frac{x_i - \bar{x}}{s_x}$, $y_i^* = \frac{y_i - \bar{y}}{s_y}$, the regression line of y on x becomes $\hat{\xi}_i^* = r\hat{\xi}_i^*$.

Proof: Recall that $x_i^* = \frac{x_i - \bar{x}}{s_x}$, $y_i^* = \frac{y_i - \bar{y}}{s_y}$ effectively transform the data into z -scores, and that, by definition,

$$\text{Pearson's } r = \frac{\sum_{i=1}^n \left[\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \right]}{n-1} = \frac{\sum_{i=1}^n x_i^* y_i^*}{n-1}.$$

We will now demonstrate two preliminary algebraic identities for x which will also hold for y . The first is an identity involving the sums of the transformed variables, the second involving the sums of squares of the transformed variables.

$$\begin{aligned} \sum_{i=1}^n x_i^* &= \sum_{i=1}^n \frac{x_i - \bar{x}}{s_x} & \sum_{i=1}^n (x_i^*)^2 &= \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right)^2 \\ &= \frac{1}{s_x} \sum_{i=1}^n (x_i - \bar{x}) & &= \frac{1}{s_x^2} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= 0. & &= \frac{1}{s_x^2} [(n-1)s_x^2] \\ & & &= n-1 \end{aligned}$$

We wish to find the least squares best fit line for our transformed variables. The line will have the form, $\hat{\xi}_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i^*$. We will show that for the line to be a least squares fit, $\hat{\beta}_0 = 0$, and $\hat{\beta}_1 = r$. That is, we will show that these are the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared errors SSE .

$$\begin{aligned}
SSE(\hat{\mathbf{b}}_0, \hat{\mathbf{b}}_1) &= g(\hat{\mathbf{b}}_0, \hat{\mathbf{b}}_1) = \sum_{i=1}^n \left[y_i^* - (\hat{\mathbf{b}}_0 + \hat{\mathbf{b}}_1 x_i^*) \right]^2 \\
&= \sum_{i=1}^n \left[(y_i^*)^2 - 2(y_i^*)(\hat{\mathbf{b}}_0 + \hat{\mathbf{b}}_1 x_i^*) + (\hat{\mathbf{b}}_0 + \hat{\mathbf{b}}_1 x_i^*)^2 \right] \\
&= \sum_{i=1}^n \left[(y_i^*)^2 - 2\hat{\mathbf{b}}_0 y_i^* - 2\hat{\mathbf{b}}_1 x_i^* y_i^* + \hat{\mathbf{b}}_0^2 + 2\hat{\mathbf{b}}_0 \hat{\mathbf{b}}_1 x_i^* + \hat{\mathbf{b}}_1^2 (x_i^*)^2 \right] \\
&= \sum_{i=1}^n (y_i^*)^2 - 2\hat{\mathbf{b}}_0 \sum_{i=1}^n y_i^* - 2\hat{\mathbf{b}}_1 \sum_{i=1}^n x_i^* y_i^* + n\hat{\mathbf{b}}_0^2 + 2\hat{\mathbf{b}}_0 \hat{\mathbf{b}}_1 \sum_{i=1}^n x_i^* + \hat{\mathbf{b}}_1^2 \sum_{i=1}^n (x_i^*)^2 \\
&= (n-1) - 2\hat{\mathbf{b}}_0(0) - 2\hat{\mathbf{b}}_1[(n-1)r] + n\hat{\mathbf{b}}_0^2 + 2\hat{\mathbf{b}}_0 \hat{\mathbf{b}}_1(0) + \hat{\mathbf{b}}_1^2(n-1) \\
&= (n-1) - 2\hat{\mathbf{b}}_1[(n-1)r] + n\hat{\mathbf{b}}_0^2 + \hat{\mathbf{b}}_1^2(n-1)
\end{aligned}$$

Observe that $n\hat{\mathbf{b}}_0^2 \geq 0$. Then it must be true that for every value of $\hat{\mathbf{b}}_0$,

$$\begin{aligned}
g(\hat{\mathbf{b}}_0, \hat{\mathbf{b}}_1) &\geq g(0, \hat{\mathbf{b}}_1) = (n-1) - 2\hat{\mathbf{b}}_1(n-1)r + \hat{\mathbf{b}}_1^2(n-1) \\
&= (n-1)(1 - 2\hat{\mathbf{b}}_1 r + \hat{\mathbf{b}}_1^2)
\end{aligned}$$

$\therefore \hat{\mathbf{b}}_0 = 0$ will produce minimum SSE for each value of $\hat{\mathbf{b}}_1$.

Now observe that $\hat{\mathbf{b}}_1^2 - 2\hat{\mathbf{b}}_1 r + 1$ is quadratic in $\hat{\mathbf{b}}_1$ with a positive lead coefficient. Thus $g(0, \hat{\mathbf{b}}_1)$ must reach a minimum when $\hat{\mathbf{b}}_1 = \frac{-(-2r)}{2(1)} = r$. Therefore $SSE = g(\hat{\mathbf{b}}_0, \hat{\mathbf{b}}_1)$ reaches a minimum when $\hat{\mathbf{b}}_0 = 0$ and $\hat{\mathbf{b}}_1 = r$.

We note in passing that a different path in the proof above may be used to prove that Pearson's correlation (r) must assume values between -1 and 1 :

$$\begin{aligned}
g(0, r) &= (n-1)(1 - 2rr + r^2) \\
&= (n-1)(1 - r^2) \geq 0, \text{ since } g(0, r) \text{ is a sum of squares.}
\end{aligned}$$

$$\begin{aligned}
\text{Then, } (n-1)(1) - (n-1)r^2 &\geq 0 \\
(n-1) &\geq (n-1)r^2 \\
1 &\geq r^2 \\
-1 &\leq r \leq 1.
\end{aligned}$$

Properties of the least squares estimator for slope

The following properties concerning the least squares estimator for slope $\hat{\mathbf{b}}_1$ in the general linear model with normally distributed errors are important for performing inference procedures concerning the true slope \mathbf{b}_1 .

$$E(\hat{\mathbf{b}}_1) = \mathbf{b}_1 \qquad V(\hat{\mathbf{b}}_1) = \frac{\mathbf{s}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\mathbf{s}^2}{s_{xx}}$$

$$\hat{\mathbf{b}}_1 \sim N\left(\mathbf{b}_1, \frac{\mathbf{s}^2}{s_{xx}}\right) \qquad \frac{\hat{\mathbf{b}}_1 - \mathbf{b}_1}{\mathbf{s}/\sqrt{s_{xx}}} \sim N(0,1)$$

The residual variance is defined to be $S^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}$. The denominator of S^2 is $n-2$ indicating $n-2$ degrees of freedom, since there are two constraints on the residuals. The two constraints are

$$\sum (y_i - \hat{y}_i) = 0 \text{ and } \sum (y_i - \hat{y}_i)x_i = 0.$$

If

$$S^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2},$$

then, it can be shown that

$$\frac{(n-2)S^2}{\mathbf{s}^2} \sim \mathbf{c}_{n-2}^2.$$

It can also be shown that $\hat{\mathbf{b}}_1$ and S^2 are independent.

Now note $\frac{\frac{\hat{\mathbf{b}}_1 - \mathbf{b}_1}{\mathbf{s}/\sqrt{s_{xx}}}}{\sqrt{\frac{(n-2)S^2/\mathbf{s}^2}{n-2}}} = \frac{\hat{\mathbf{b}}_1 - \mathbf{b}_1}{S/\sqrt{s_{xx}}}$ is the ratio of a standard normal random variable to the

square root of an independent chi-square random variable divided by its degrees of freedom. Therefore, by definition, this result has a t -distribution with $n-2$ degrees of freedom and can be used for inferences regarding \mathbf{b}_1 . That is,

$$\frac{\hat{\mathbf{b}}_1 - \mathbf{b}_1}{S/\sqrt{s_{xx}}} \sim t_{n-2}.$$

The standard error of $\hat{\mathbf{b}}_1$ is $\frac{S}{\sqrt{s_{xx}}}$.

(For more details, see Sections 11.4 and 11.5 in *Mathematical Statistics with Applications*, Wackerly, Mendenhall, and Scheaffer, Duxbury Press, 1996.)