**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

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\beta_0$ and $\beta_1$ are parameters and $\epsilon_i$ represents random error with $E(\epsilon_i) = 0$. 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 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The least squares procedure for fitting a line through a set of $n$ data points determines $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the sum of squared errors

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

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

$$SSE = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

$$\frac{\partial SSE}{\partial \hat{\beta}_0} = 2 \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)](-1)$$

$$= -2 \left( \sum_{i=1}^{n} y_i - n \hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i \right)$$

$$= -2n\bar{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1 \bar{x}$$

Since $\sum y_i = n\bar{y}$, we can write $\sum y_i = n\bar{y}$.

Setting this partial derivative equal to zero yields:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

or

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\frac{\partial SSE}{\partial \hat{\beta}_1} = 2 \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)](-x_i)$$

$$= -2 \left( \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \right)$$

Setting this partial derivative equal to zero yields:
\begin{align*}
\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i \\
&= \sum_{i=1}^{n} x_i y_i - (\bar{y} - \hat{\beta}_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{\beta}_1 \sum_{i=1}^{n} x_i \\
\text{So } \beta_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 - \bar{y} \sum_{i=1}^{n} x_i + n \bar{x} \cdot \bar{y}}{\sum_{i=1}^{n} x_i^2 - 2 n \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 + \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 (x_i - \bar{x})^2}
\end{align*}

When using the model \( Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \), we assume that there is a linear relationship between \( x \) and \( E(Y) \) with a true slope \( \beta_1 \) and a true intercept \( \beta_0 \). Because of the error term \( \varepsilon_i \), observations with equal \( x \)-values will not necessarily have equal \( y \)-values. We have stated previously that \( E(\varepsilon_i) = 0 \). We will now assume that \( \varepsilon_i \sim N(0, \sigma^2) \).

Figure 14: Graphical Representation of \( Y_i \sim N\left( \beta_0 + \beta_1 x_i, \sigma^2 \right) \)
That is, for each value of \( x_i \) we assume that the errors are normally distributed with mean 0 and constant variance \( \sigma^2 \). These assumptions about the distribution of the error terms underlie the distribution of \( Y_i \) for a fixed \( x_i \). Specifically, for a fixed \( x_i \), \( Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \).

The values \( Y_1, Y_2, \ldots, Y_n \) are independent but are not identically distributed, since they have different values of \( \mu_i \). Here, \( \mu_i = \beta_0 + \beta_1 x_i \).

Note that:

\[
E(Y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) = E(\beta_0) + E(\beta_1 x_i) + E(\varepsilon_i) = \beta_0 + \beta_1 x_i + 0
\]

\[
V(Y_i) = V(\beta_0 + \beta_1 x_i + \varepsilon_i) = V(\beta_0) + V(\beta_1 x_i) + V(\varepsilon_i) = 0 + 0 + \sigma^2
\]

It can be shown that the least squares estimators of \( \beta_0 \) and \( \beta_1 \) are also maximum likelihood estimators:

\[
L(\beta_0, \beta_1) = L(y_1, y_2, \ldots, y_n | \beta_0, \beta_1)
\]

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

\[
L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2\sigma^2} \left[ y_i - (\beta_0 + \beta_1 x_i) \right]^2}
\]

\[
= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2\sigma^2} \left[ y_i - (\beta_0 + \beta_1 x_i) \right]^2}
\]

\[
= \left( \frac{1}{\sqrt{2\pi} \sigma} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2}
\]

We want to maximize the likelihood with respect to \( \beta_0 \) and \( \beta_1 \). Since the exponent on \( e \) is negative, large values of \( L(\beta_0, \beta_1) \) are associated with small values of \( \sum [y_i - (\beta_0 + \beta_1 x_i)]^2 \). So to obtain the maximum likelihood estimator we need to minimize \( \sum [y_i - (\beta_0 + \beta_1 x_i)]^2 \), which is what we already did to obtain the least squares estimators \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \). So, \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are maximum likelihood estimators under the assumption \( \varepsilon \sim N(0, \sigma^2) \).
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 \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), n > 1\) be a set of data, with \(s_x, s_y > 0\). Then if \(x^*_i = \frac{x_i - \bar{x}}{s_x}, \quad y^*_i = \frac{y_i - \bar{y}}{s_y}\), the regression line of \(y\) on \(x\) becomes \(\hat{y}^* = r \hat{y}^*\).

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

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

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.

\[
\sum_{i=1}^{n} x^*_i = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} = \frac{1}{s_x} \sum_{i=1}^{n} (x_i - \bar{x}) = 0.
\]

\[
\sum_{i=1}^{n} (x^*_i)^2 = \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right)^2 = \frac{1}{s_x^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{s_x^2} [(n-1)s_x^2] = n-1.
\]

We wish to find the least squares best fit line for our transformed variables. The line will have the form, \(\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*_i\). 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 \(SSE\).
\[
SSE (\hat{\beta}_0, \hat{\beta}_1) = g (\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \left[ y_i^* - (\hat{\beta}_0 + \hat{\beta}_1 x_i^*) \right]^2
\]

\[
= \sum_{i=1}^{n} \left[ (y_i^*)^2 - 2 (y_i^*) (\hat{\beta}_0 + \hat{\beta}_1 x_i^*) + (\hat{\beta}_0 + \hat{\beta}_1 x_i^*)^2 \right]
\]

\[
= \sum_{i=1}^{n} \left[ (y_i^*)^2 - 2\hat{\beta}_0 y_i^* - 2\hat{\beta}_1 x_i^* y_i^* + \hat{\beta}_0^2 + 2\hat{\beta}_0 \hat{\beta}_1 x_i^* + \hat{\beta}_1^2 (x_i^*)^2 \right]
\]

\[
= \sum_{i=1}^{n} \left( y_i^* \right)^2 - 2\hat{\beta}_0 \sum_{i=1}^{n} y_i^* - 2\hat{\beta}_1 \sum_{i=1}^{n} x_i^* y_i^* + n\hat{\beta}_0^2 + 2\hat{\beta}_0 \hat{\beta}_1 \sum_{i=1}^{n} x_i^* + \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i^*)^2
\]

\[
= (n-1) - 2\hat{\beta}_0 (0) - 2\hat{\beta}_1 [(n-1)r] + n\hat{\beta}_0^2 + 2\hat{\beta}_0 \hat{\beta}_1 (0) + \hat{\beta}_1^2 (n-1)
\]

\[
= (n-1) - 2\hat{\beta}_1 [(n-1)r] + n\hat{\beta}_0^2 + \hat{\beta}_1^2 (n-1)
\]

Observe that \( n\hat{\beta}_0^2 \geq 0 \). Then it must be true that for every value of \( \hat{\beta}_0 \),

\[
g (\hat{\beta}_0, \hat{\beta}_1) \geq g (0, \hat{\beta}_1) = (n-1) - 2\hat{\beta}_1 (n-1)r + \hat{\beta}_1^2 (n-1)
\]

\[
= (n-1)(1 - 2\hat{\beta}_1 r + \hat{\beta}_1^2)
\]

\[
\therefore \hat{\beta}_0 = 0 \text{ will produce minimum } SSE \text{ for each value of } \hat{\beta}_1.
\]

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

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 \( 1 \):

\[
g(0,r) = (n-1)(1-2rr + r^2)
\]

\[
= (n-1)(1- r^2) \geq 0, \text{ since } g(0,r) \text{ is a sum of squares.}
\]

Then, \( (n-1)(1- (n-1)r^2) \geq 0 \)

\[
(n-1) \geq (n-1)r^2
\]

\[
1 \geq r^2
\]

\[
-1 \leq r \leq 1.
\]
Properties of the least squares estimator for slope

The following properties concerning the least squares estimator for slope $\hat{\beta}_1$ in the general linear model with normally distributed errors are important for performing inference procedures concerning the true slope $\beta_1$.

$$E(\hat{\beta}_1) = \beta_1$$

$$V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sigma^2}{s_{xx}}$$

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

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

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

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}.$$ 

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

Now note

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{s_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{(n-2)S^2/\sigma^2}}$$

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 $\beta_1$. That is,

$$\frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{s_{xx}}} \sim t_{n-2}.$$ 

The standard error of $\hat{\beta}_1$ 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.)