

Derivations of the LSE for Four Regression Models

1. Introduction

The least squares method goes back to 1795, when Carl Friedrich Gauss, the great German mathematician, discovered it when he was eighteen years old. It arose in the context of astronomy. Gauss' method was used on data collected by the Italian astronomer Giuseppe Piazzi to successfully calculate the position of the asteroid Ceres after it reemerged from behind the sun. In the next two centuries, several variations of the least squared method were successfully used to make predictions for many different applications. These applications were crucial to many more discoveries in astronomy, in particular navigation at sea, where accurate measurements of the positions of stars were used to determine the positions of ships navigating the world's oceans.

In 1805, Gauss published the second volume of his work on celestial mechanics. He showed there that his least squared method was optimal in the sense that it has the minimum variance out of all unbiased estimators. Today, this is called the Gauss-Markov Theorem. Although the least squares method was also developed independently by Adrien-Marie Legendre and Robert Adrian in the early 1800's, Gauss is the historical figure that comes to mind first as the discoverer of the least squares method.

In general, if the error for an observation is defined as the actual observation minus the predicted value, then the least squares estimator (LSE) is the one that minimizes the sum of squares of the errors. Four special cases will be covered in the following sections: (2) Horizontal Line Regression, (3) Simple Linear Regression, (4) Multiple Linear Regression, (5) Multiple Linear Regression.

Reference: Wikipedia, *Least Squares, History*,
<http://en.wikipedia.org/wiki/Least_squares#History>,
accessed on 9/2/2011.

2. Horizontal Line Regression

Horizontal line regression (also called the ideal measurement model) assumes that there are no independent variables. The regression model is defined as

$$y_i = \mu + \epsilon_i.$$

If the predicted value of y_i is denoted by \hat{y}_i , and the residuals are defined as $\epsilon_i = y_i - \hat{y}_i$, then the relationship between y_i , \hat{y}_i , and ϵ_i is shown in Figure 1:

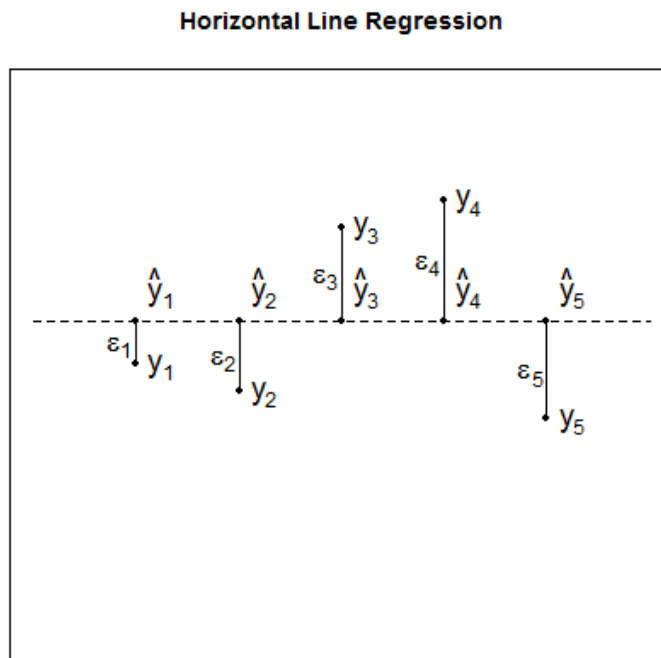


Figure 1: The LSE for horizontal line regression is $y = \bar{y}$. The slope of the regression line is constrained to be zero.

The LSE for horizontal line regression is found by minimizing the sum of squares for error (SSE):

$$\min_{\mu} \text{SSE} = \min_{\mu} \sum_{i=1}^n \epsilon_i^2 = \min_{\mu} \sum_{i=1}^n (y_i - \mu)^2$$

To minimize the SSE, use the standard calculus procedure of setting the derivative of SSE to zero and solving for μ :

$$\frac{d}{d\mu} \text{SSE} = \frac{d}{d\mu} \sum_{i=1}^n (y_i - \mu)^2 = \sum_{i=1}^n 2(y_i - \mu)(-1) = 0$$

Divide by $-2n$ to obtain

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (y_i - \mu) &= 0 \\ \frac{1}{n} \left(\sum_{i=1}^n y_i - n\mu \right) &= 0 \\ \frac{1}{n} \sum_{i=1}^n y_i - \mu &= 0 \\ \bar{y} - \mu &= 0. \end{aligned}$$

Thus the least squares value for μ is the usual sample mean \bar{y} and the horizontal line regression equation is

$$\boxed{y = \bar{y}}$$

3. Regression through the Origin

For regression through the origin, the intercept of the regression line is constrained to be zero, so the regression line is of the form $y = ax$. We want to find the value of a that satisfies

$$\min_a \text{SSE} = \min_a \sum_{i=1}^n \epsilon_i^2 = \min_a \sum_{i=1}^n (y_i - ax_i)^2$$

This situation is shown in Figure 2. As in Section 1, set the derivative to zero and solve for the desired parameter, which is a in this case.

$$\frac{d}{da} \text{SSE} = \frac{d}{da} \sum_{i=1}^n (y_i - ax_i)^2 = \sum_{i=1}^n 2(y_i - ax_i)(-x_i) = 0$$

Regression through the Origin

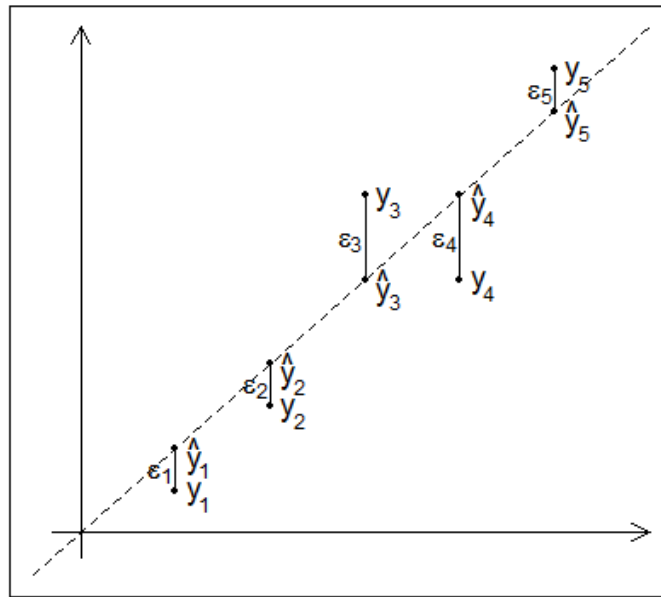


Figure 2: The LSE for regression through the origin is given by (1). The intercept of the regression line constrained to be zero.

Divide by -2 to obtain

$$\sum_{i=1}^n (y_i - ax_i)x_i = 0$$
$$\sum_{i=1}^n x_i y_i - a \sum_{i=1}^n x_i^2 = 0$$

Then solve for \hat{a} :

$$\hat{a} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

This gives the LSE for regression through the origin:

$$y = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \cdot x \quad (1)$$

4. Simple Straight Line Regression

The regression model for simple linear regression is

$$y = ax + b.$$

Finding the LSE is more difficult than for horizontal line regression or regression through the origin because there are two parameters a and b over which to optimize simultaneously. This involves two equations in two unknowns. The minimization problem is

$$\min_{a,b} \text{SSE} = \min_{a,b} \sum_{i=1}^n \epsilon_i^2 = \min_{a,b} \sum_{i=1}^n (y_i - (ax_i + b))^2 = \min_{a,b} \sum_{i=1}^n (y_i - ax_i - b)^2 \quad (2)$$

See Figure 3 for a picture of the relationship among y_i , \hat{y}_i , and ϵ_i .

We first find the optional value of b by setting the partial derivative of b to zero:

$$\frac{\partial}{\partial b} \text{SSE} = \frac{\partial}{\partial b} \sum_{i=1}^n (y_i - ax_i - b)^2 = \frac{\partial}{\partial b} \sum_{i=1}^n 2(y_i - ax_i - b)(-1) = 0$$

Divide by $-2n$ to obtain

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (y_i - ax_i - b) &= 0 \\ \frac{1}{n} \sum_{i=1}^n y_i - \frac{a}{n} \sum_{i=1}^n x_i - \frac{b}{n} \sum_{i=1}^n 1 &= 0 \\ \frac{1}{n} \sum_{i=1}^n y_i - a \frac{1}{n} \sum_{i=1}^n x_i - \frac{b}{n} n &= 0 \\ \bar{y} - a\bar{x} - b &= 0, \end{aligned}$$

Simple Straight Line Regression

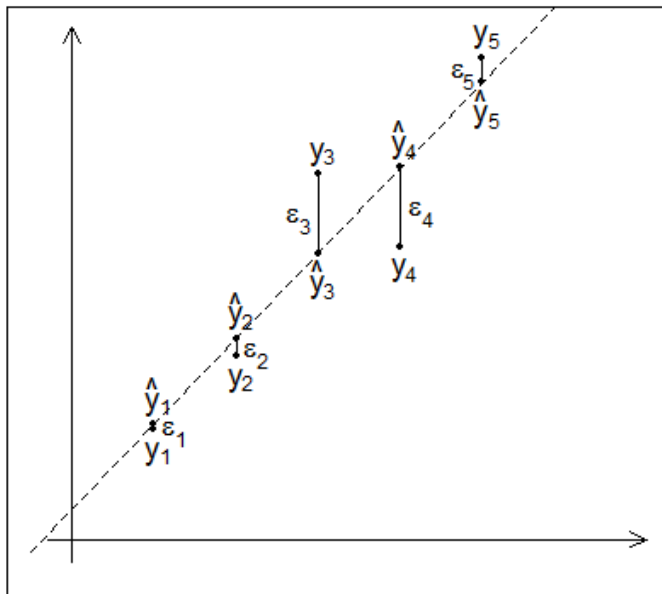


Figure 3: Least squared regression line is $\hat{y}_i = \hat{a}x_i + \hat{b} + \hat{\epsilon}_i$.

then solve for the optional b , which is \hat{b} :

$$\hat{b} = \bar{y} - a\bar{x} \quad (3)$$

Use the newly obtained optimal value of b from Equation (3) to substitute into Equation (2) to solve for the optimal a :

$$\min_{a,b} \text{SSE} = \min_a \sum_{i=1}^n (y_i - ax_i - (\bar{y} - a\bar{x}))^2 = \min_a \sum_{i=1}^n (y_i - a(x_i - \bar{x}) - \bar{y})^2$$

Set the derivative of SSE with respect to a to zero:

$$\begin{aligned} \frac{d}{da} \text{SSE} &= \frac{d}{da} \sum_{i=1}^n (y_i - a(x_i - \bar{x}) - \bar{y})^2 \\ &= \sum_{i=1}^n 2(y_i - a(x_i - \bar{x}) - \bar{y})(-(x_i - \bar{x})) = 0 \end{aligned}$$

Divide the preceding line by $-2/(n-1)$:

$$\begin{aligned} \frac{1}{n-1} \sum_{i=1}^n (y_i - a(x_i - \bar{x}) - \bar{y})(x_i - \bar{x}) &= 0 \\ \frac{1}{n-1} \sum_{i=1}^n ((y_i - \bar{y}) - a(x_i - \bar{x}))(x_i - \bar{x}) &= 0 \\ \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) - a \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) &= 0 \\ \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) - a \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 &= 0 \\ s_{xy} - a s_x^2 &= 0 \end{aligned}$$

Solve for a and use the definition of the sample correlation $r_{xy} = s_{xy}/(s_x s_y)$ write the optimal value of a like this:

$$\hat{a} = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}}{s_x^2} \cdot \frac{s_y}{s_y} = \frac{s_{xy}}{s_x s_y} \cdot \frac{s_y}{s_x} = r \frac{s_y}{s_x}$$

Thus, the least squares regression equation for simple straight line regression is

$$y = \hat{a}x + \hat{b}, \tag{4}$$

where

$$\hat{a} = r \frac{s_y}{s_x} \tag{5}$$

and

$$\hat{b} = \hat{y} - a\hat{x} \tag{6}$$

We can substitute definitions (5) and (6) into (4) to write the LSE for simple linear regression equation as

$$\boxed{y - \bar{y} = r_{xy} \frac{s_y}{s_x} (x - \bar{x})}$$

5. Multiple Linear Regression

To efficiently solve for the least squares equation of the multiple linear regression model, we need an efficient method of representing the multiple linear regression model. A good way to do this is to use the matrix representation

$$y = X\beta + \epsilon$$

where

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \text{and} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

The parameter vector β and the matrix X depend on the model which is used. Here are the matrix formulations of the models in sections 2, 3, and 4:

1. **Horizontal Line Regression:** β is a 1×1 vector (simple variable) containing μ ; X is an $n \times 1$ matrix of ones.

$$\beta = \mu \quad X = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

We have

$$\begin{aligned} y &= X\beta + \epsilon \\ \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} &= \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \\ \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} &= \begin{pmatrix} \mu + \epsilon_1 \\ \vdots \\ \mu + \epsilon_n \end{pmatrix} \\ &= \begin{pmatrix} y_1 = \mu + \epsilon_1 \\ \vdots \\ y_n = \mu + \epsilon_n \end{pmatrix} \end{aligned}$$

2. **Regression through the Origin:** β is a 1×1 vector (simple variable) containing a ; X is an $n \times 1$ matrix containing the values of x_i .

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

Then we have

$$\begin{aligned}y &= X\beta + \epsilon \\ \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} &= \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} a + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \\ \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} &= \begin{pmatrix} ax_1 + \epsilon_1 \\ \vdots \\ ax_n + \epsilon_n \end{pmatrix} \\ &= \begin{pmatrix} y_1 = ax_1 + \epsilon_1 \\ \vdots \\ y_n = ax_n + \epsilon_n \end{pmatrix}\end{aligned}$$

3. **Simple Linear Regression:** β is a 1×2 vector containing a and b from the regression equation $y = ax + b$; X is an $n \times 2$ matrix:

$$\beta = \begin{pmatrix} a \\ b \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

Then

$$\begin{aligned}y &= X\beta + \epsilon \\ \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} &= \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \\ \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} &= \begin{pmatrix} ax_1 + b + \epsilon_1 \\ \vdots \\ ax_n + b + \epsilon_n \end{pmatrix} \\ &= \begin{pmatrix} y_1 = ax_1 + b + \epsilon_1 \\ \vdots \\ y_n = ax_n + b + \epsilon_n \end{pmatrix}\end{aligned}$$

To solve for the LSE in this general matrix form, solve the minimization problem

$$\min_{\beta} \sum_{i=1}^n \epsilon_i^2 = \min_{\beta} \epsilon^T \epsilon$$

To perform the minimization, generalize the standard calculus procedure by using vector calculus. Here are the formulas that are needed. Assume that u , v , w are variable vectors, c is a constant vector, and A is a symmetric matrix. The vector partial derivative is defined as

$$\frac{\partial}{\partial u} v = \begin{pmatrix} \frac{\partial v}{\partial u_1} \\ \vdots \\ \frac{\partial v}{\partial u_n} \end{pmatrix}$$

We use the following facts about vector derivatives. Also see the Section 6 of the Matrix Review document: Some Matrix Identities. A link is on the course Documents page.

1. $\frac{\partial}{\partial u}(v + w) = \frac{\partial}{\partial u}v + \frac{\partial}{\partial u}w$
2. $\frac{\partial}{\partial u}c = 0$
3. $\frac{\partial}{\partial u}(a^T u) = a$
4. $\frac{\partial}{\partial u}(u^T A u) = 2A u$

We also use the fact that $A^T = A$ for a 1×1 matrix A .

Proceed with the minimization:

$$\begin{aligned}
 \min_{\beta}(\epsilon^T \epsilon) &= \min_{\beta}(y - X\beta)^T(y - X\beta) \\
 &= \min_{\beta}(y^T y - y^T X\beta - (X\beta)^T y + (X\beta)^T(X\beta)) \\
 &= \min_{\beta}(y^T y - y^T X\beta - \beta^T X^T y + (X\beta)^T(X\beta)) \\
 &= \min_{\beta}(y^T y - y^T X\beta - (\beta^T X^T y)^T + (X\beta)^T(X\beta)) \\
 &= \min_{\beta}(y^T y - y^T X\beta - y^T X\beta + \beta^T X^T X\beta) \\
 &= \min_{\beta}(y^T y - 2y^T X\beta - \beta^T X^T X\beta)
 \end{aligned}$$

Now compute the vector partial derivative and set it to the zero vector:

$$\begin{aligned}
 \frac{\partial}{\partial \beta}(\epsilon^T \epsilon) &= \frac{\partial}{\partial \beta}(y^T y - 2y^T X\beta + \beta^T X^T X\beta) = 0 \\
 -2(y^T X)^T + 2X^T X\beta &= 0 \\
 -2X^T y + 2X^T X\beta &= 0
 \end{aligned}$$

Divide by 2:

$$X^T X\hat{\beta} = X^T Y,$$

then solve for β to obtain

$$\boxed{\hat{\beta} = (X^T X)^{-1} X^T Y} \tag{7}$$

Now use Equation (7) to find each of the least squared regression equations for the models in Sections 2, 3, and 4:

1. Horizontal Line Regression:

$$\begin{aligned}
\hat{\beta} &= (X^T X)^{-1} X^T Y \\
&= \left[\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}^T \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right]^{-1} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\
&= \left[(1 \ \cdots \ 1) \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right]^{-1} (1 \ \cdots \ 1) \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\
&= \left(\sum_{i=1}^n 1 \cdot 1 \right)^{-1} \sum_{i=1}^n 1 \cdot y_i = n^{-1} \sum_{i=1}^n y_i = \bar{y}
\end{aligned}$$

Verify the this LSE

$$\boxed{y = \bar{y}}$$

is the same one obtained in Section 2.

2. Regression through the Origin:

$$\begin{aligned}
\hat{\beta} &= (X^T X)^{-1} X^T Y \\
&= \left[\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}^T \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \right]^{-1} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\
&= \left[(x_n \ \cdots \ x_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \right]^{-1} (x_1 \ \cdots \ x_n) \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\
&= \left(\sum_{i=1}^n x_i \cdot x_i \right)^{-1} \sum_{i=1}^n x_i \cdot y_i = \left(\sum_{i=1}^n x_i^2 \right)^{-1} \sum_{i=1}^n x_i y_i \\
&= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}
\end{aligned}$$

We obtain the same LSE

$$y = \begin{pmatrix} \frac{\sum_{i=1}^n x_i y_i}{\frac{1}{n} \sum_{i=1}^n x_i^2} \end{pmatrix} x$$

as we did in Section 3.

3. Simple Linear Regression:

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y \\ &= \left[\begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}^T \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \right]^{-1} \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\ &= \left[\begin{pmatrix} x_n & \cdots & x_1 \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \right]^{-1} \begin{pmatrix} x_1 & \cdots & x_n \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n 1 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^n x_i^2 & n\bar{x} \\ n\bar{x} & n \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ n\bar{y} \end{pmatrix} \end{aligned}$$

One can continue to simplify this expression along the line of Section 4. An alternative is to substitute the x_i and y_i values directly and solve numerically, as is done with statistical software.