Derivations of the LSE for Four Regression Models

Nayel Bettache

September 10, 2024

1 Introduction

Least Squares Regression (LSR) is a fundamental statistical technique used to estimate the parameters of a model by minimizing the sum of the squared differences between observed values and the corresponding predicted values. The method provides a simple and computationally efficient way to fit linear models to data.

The key properties of LSR are rooted in its optimality and simplicity:

- **Unbiased Estimation:** The Least Squares Estimator (LSE) provides unbiased estimates of the model parameters under the assumption of homoscedasticity, meaning the variance of the errors is constant across all observations.

- **Minimum Variance:** According to the Gauss-Markov Theorem, among all linear and unbiased estimators, LSE has the smallest possible variance. This makes it an efficient estimator, particularly in the presence of normally distributed errors.

- **Linear Relationship:** LSR assumes a linear relationship between the dependent variable and one or more independent variables. This relationship can be expressed as a linear combination of the predictors, with the goal of finding the best-fitting line or hyperplane.

- **Error Minimization:** The method minimizes the sum of squared errors (SSE), which is the sum of the squared differences between the observed values and the predicted values. By focusing on squared errors, the method heavily penalizes large deviations, which ensures that outliers have a stronger influence on the final model compared to smaller deviations.

- **Interpretability:** The parameters estimated by LSR, such as the intercept and slope(s), have clear interpretations in terms of the relationship between the predictor(s) and the response variable. This makes LSR an easily interpretable method for understanding data patterns.
- **Sensitivity to Outliers:** While LSR is computationally efficient and easy to implement, it is sensitive to outliers. Large deviations from the trendline disproportionately affect the sum of squared errors, potentially leading to biased or misleading estimates.

LSR is widely used in a variety of fields, including economics, biology, engineering, and social sciences, to model relationships between variables and make predictions. This document explores four specific cases of regression: (1) Horizontal Line Regression, (2) Regression Through the Origin, (3) Simple Linear Regression, and (4) Multiple Linear Regression, detailing the derivation of the LSE for each model.

2 Horizontal Line Regression

Horizontal line regression, also called the ideal measurement model, assumes no independent variables. The model is:

$$y_i = \mu + \epsilon_i$$

The LSE minimizes the sum of squared errors (SSE):

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

To find the optimal value of μ, we take the derivative of the SSE with respect to μ and set it to zero:

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

Solving this, we get the LSE as the sample mean:

$$\mu = \bar{y}$$

Thus, the regression equation is:

$$y = \bar{y}$$

3 Regression Through the Origin

In regression through the origin, the intercept is constrained to zero, and the model is:

$$y_i = ax_i$$

The objective is to minimize the SSE:

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

Taking the derivative with respect to a and setting it to zero:

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

Solving for a, we obtain:

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

Thus, the LSE for regression through the origin is:

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

4 Simple Linear Regression

In simple linear regression, the model includes both an intercept and a slope:

$$y_i = ax_i + b$$

The SSE to be minimized is:

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

We first find the optimal b by taking the partial derivative of the SSE with respect to b and setting it to zero:

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

Solving, we find:

$$b = \bar{y} - a\bar{x}$$

Next, we substitute b into the SSE and solve for a, yielding:

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Thus, the regression equation is:

$$y = ax + b$$

5 Multiple Linear Regression

For multiple linear regression, the model is represented in matrix form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of parameters, and $\boldsymbol{\epsilon}$ is the error vector.

The LSE is found by minimizing:

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Taking the derivative and setting it to zero gives:

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$$

Solving for $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$