MTL108
Introduction to Linear Regression (Optional)

Rahul Singh

Introduction

At the heart of many scientific and engineering endeavors lies a fundamental question: *How exactly do different variables influence one another?*

Consider a chemical manufacturing plant. An engineer might need to determine the precise mathematical relationship between the final production yield (the response variable), the operating temperature, and the volume of catalyst used. Quantifying this relationship offers a tremendous operational advantage. It allows professionals to transition from passive observation to active forecasting, enabling them to *predict* manufacturing outcomes under a variety of hypothetical conditions. Establishing this reliable, predictive framework is the primary objective of linear regression analysis.

In this module, we focus on the mathematical modeling of a single dependent variable driven by one independent variable. This specific framework is formally known as a simple linear regression model (as opposed to a multiple linear regression model, which incorporates several explanatory variables).

The Linear Model

We define the standard simple linear regression model mathematically as:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Here, Y represents the dependent (or study) variable, while X serves as the independent (or explanatory) variable. The parameters β_0 (the intercept) and β_1 (the slope) dictate the true, underlying linear relationship and are collectively referred to as the regression coefficients.

Because empirical data almost never falls perfectly along a straight line, we introduce ϵ, the unobservable error (or disturbance) term. This term captures the deviation between the actual observed value of Y and the deterministic component of the model ($\beta_0 + \beta_1 X$). These deviations can arise from omitted variables, qualitative factors, or inherent natural randomness. For the purposes of this model, we assume that ϵ acts as an independent and identically distributed (i.i.d.) random variable with a mean of zero and a constant variance of σ^2.

Typically, the independent variables are treated as predetermined values tightly controlled by the researcher, rendering them non-stochastic (fixed). Consequently, the response Y becomes a random variable whose behavior is defined by its expected value:

$$E(Y) = \beta_0 + \beta_1 X$$

and its variance:

$$Var(Y) = \sigma^2.$$

Remark 1. In certain observational studies, X may also be treated as a random variable rather than a fixed constant. Under these circumstances, we shift our focus from marginal moments to the conditional mean of Y given a specific realization X:

$$E(Y|X) = \beta_0 + \beta_1 X$$

along with the conditional variance:

$$Var(Y|X) = \sigma^2.$$

If the true population values of β_0, β_1, and σ^2 were known, our statistical model would be completely defined. In practice, however, these parameters are unknown, and the true errors (ϵ) remain unseen. Therefore, successfully utilizing this model relies entirely on estimating β_0, β_1, and σ^2 from sample data. By collecting a sample of n paired observations (X_i, Y_i) for $i = 1, ..., n$, we can apply various estimation techniques to approximate these parameters. The two most prominent approaches are the Method of Least Squares and the Method of Maximum Likelihood.

Least Squares Estimation

Assume we possess a sample of n paired observations, (X_i, Y_i). If these observations adhere to our simple linear regression structure, we can express the model for each individual data point as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad (i = 1, 2, ..., n).$$

The core philosophy behind least squares estimation is to determine the coefficients β_0 and β_1 that minimize the total discrepancy between the actual data points and the proposed regression line. This discrepancy can be measured geometrically in several ways:

- **Direct Regression:** Minimizes the *vertical* distances (errors) between the observations and the fitted line.
- **Reverse Regression:** Minimizes the *horizontal* discrepancies.
- **Orthogonal Regression:** Minimizes the *perpendicular* distances from the points to the line (useful when both variables contain measurement error).
- **Reduced Major Axis Regression:** Minimizes the triangular or rectangular areas formed between the observed points and the line.
- **Least Absolute Deviation (LAD):** Minimizes the sum of absolute vertical errors rather than squared errors, offering greater robustness against extreme outliers.

Importantly, deriving least squares estimators requires no strict assumptions about the underlying probability distribution of the error terms (ϵ_i). We only require the baseline assumptions that $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, and $Cov(\epsilon_i, \epsilon_j) = 0$ to establish the basic properties (mean and variance) of the estimators. The stronger assumption—that the errors are normally distributed—is only necessary later when constructing confidence intervals and conducting formal hypothesis tests.

While the various geometric approaches yield different statistical behaviors, the direct regression (vertical error) approach is overwhelmingly the most prevalent. Thus, the term "Ordinary Least Squares" (OLS) universally refers to this direct vertical minimization.

Deriving the Ordinary Least Squares (OLS) Estimators

The direct regression framework achieves parameter estimation by minimizing the residual sum of squares function:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

To find the minimum, we compute the partial derivatives of $S(\beta_0, \beta_1)$ with respect to both β_0 and β_1:

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) X_i$$

Setting these partial derivatives to zero yields the normal equations:

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0, \quad \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

Solving this system provides the closed-form OLS estimators, denoted b_0 and b_1:

$$b_1 = \frac{S_{XY}}{S_{XX}}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

where the sum of squares and cross-products are defined as:

$$S_{XY} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}), \quad S_{XX} = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

and the sample means are $\bar{X} = \frac{1}{n} \sum X_i$ and $\bar{Y} = \frac{1}{n} \sum Y_i$.

To mathematically guarantee that these estimates represent a global minimum (rather than a maximum or saddle point), we evaluate the Hessian matrix of second-order partial derivatives:

$$H^* = \begin{pmatrix} \frac{\partial^2 S}{\partial \beta_0^2} & \frac{\partial^2 S}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 S}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 S}{\partial \beta_1^2} \end{pmatrix} = 2 \begin{pmatrix} n & n\bar{X} \\ n\bar{X} & \sum_{i=1}^{n} X_i^2 \end{pmatrix}$$

For the function to possess a minimum, H^* must be positive definite. We calculate its determinant as:

$$|H^*| = 4 \left(n \sum_{i=1}^{n} X_i^2 - n^2 \bar{X}^2 \right) = 4n \sum_{i=1}^{n} (X_i - \bar{X})^2 \geq 0$$

Provided there is actual variability in the independent variable ($\sum (X_i - \bar{X})^2 > 0$), it strictly follows that $|H^*| > 0$. Therefore, H^* is positive definite, confirming that $S(\beta_0, \beta_1)$ reaches a global minimum at (b_0, b_1).

The resulting fitted linear regression equation is:

$$Y = b_0 + b_1 X$$

Using this model, the predicted value for any data point i is $\hat{Y}_i = b_0 + b_1 X_i$. The residual ($e_i$) is the empirical deviation between the actual observed value and this prediction:

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$$

Properties of the OLS Estimators

1. Unbiasedness:
Both estimators are fundamentally linear combinations of the response variables Y_i. We can rewrite the slope estimator as $b_1 = \sum k_i Y_i$, where the weights are $k_i = (X_i - \bar{X})/S_{XX}$. Recognizing that $\sum k_i = 0$ and $\sum k_i X_i = 1$, we evaluate its expectation:

$$E(b_1) = \sum_{i=1}^{n} k_i E(Y_i) = \sum_{i=1}^{n} k_i(\beta_0 + \beta_1 X_i) = \beta_1$$

This proves that b_1 is an unbiased estimator for β_1. Similarly for the intercept:

$$E(b_0) = E[\bar{Y} - b_1\bar{X}] = (\beta_0 + \beta_1\bar{X}) - \beta_1\bar{X} = \beta_0$$

confirming that b_0 is an unbiased estimator for β_0.

2. Variances and Covariance:
Assuming the observations Y_i are independent, the variance for the slope estimator evaluates to:

$$Var(b_1) = \sum_{i=1}^{n} k_i^2 Var(Y_i) = \sigma^2 \sum_{i=1}^{n} k_i^2 = \frac{\sigma^2}{S_{XX}}$$

For the intercept estimator b_0, after noting that $Cov(\bar{Y}, b_1) = 0$, the variance simplifies to:

$$Var(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)$$

Furthermore, the covariance linking the two estimators is:

$$Cov(b_0, b_1) = -\frac{\bar{X}}{S_{XX}}\sigma^2$$

Under the Gauss-Markov Theorem, these OLS estimators achieve the lowest possible variance among all linear, unbiased estimators, making them the Best Linear Unbiased Estimators (BLUE).

3. Residual Sum of Squares (RSS) and Variance Estimation:
The total variation left unexplained by the regression is captured by the residual sum of squares:

$$SS_{res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 = S_{YY} - b_1 S_{XY}$$

where $S_{YY} = \sum (Y_i - \bar{Y})^2$.

To accurately estimate the underlying population variance σ^2, we utilize this residual sum of squares. Under the assumption of normally distributed errors, the quantity SS_{res}/σ^2 follows a Chi-square (χ^2) distribution with $(n-2)$ degrees of freedom. Consequently, $E(SS_{res}) = (n-2)\sigma^2$, which yields a strictly unbiased estimator for the variance:

$$s^2 = \frac{SS_{res}}{n-2}$$

Two degrees of freedom are lost because the calculation relies on the prior estimation of two parameters (b_0 and b_1).

With s^2 established, we can estimate the precision of our regression coefficients by replacing the unknown σ^2 with s^2:

$$\hat{Var}(b_1) = \frac{s^2}{S_{XX}} \quad \text{and} \quad \hat{Var}(b_0) = s^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)$$

Theorem 1 (Without proof)**.** *Within the OLS framework featuring a non-stochastic covariate, the following algebraic identities strictly hold: (i) Predictors and residuals are orthogonal:* $\sum X_i e_i = 0$*. (ii) Predictions and residuals are orthogonal:* $\sum \hat{Y}_i e_i = 0$*. (iii) The sum of observed values equals the sum of predicted values:* $\sum Y_i = \sum \hat{Y}_i$*. (iv) The regression line continuously traverses the center of mass of the data:* $(\bar{X}, \bar{Y})$*.*

Maximum Likelihood Estimation (MLE)

To apply the Maximum Likelihood Estimation paradigm, we must formally assign a probability distribution to the random disturbances. We impose the strict assumption that $\epsilon_i \sim N(0, \sigma^2)$. Because X_i is treated as a fixed constant, the observed values Y_i become independent, normally distributed random variables:

$$Y_i \sim N(\beta_0 + \beta_1 X_i,\, \sigma^2) \quad \text{for all } i = 1, 2, \ldots, n \tag{1}$$

The likelihood function L represents the joint probability density function of our sample data. Since the observations are independent, this is the product of their marginal densities:

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp\left[-\frac{1}{2\sigma^2} (Y_i - \beta_0 - \beta_1 X_i)^2 \right] \tag{2}$$

Taking the natural logarithm simplifies differentiation while preserving the location of the maximum:

$$\ln L = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{3}$$

By taking the partial derivatives of $\ln L$ with respect to β_0 and β_1 and setting them to zero, we derive equations mathematically indistinguishable from the OLS normal equations. Therefore, the Maximum Likelihood Estimates for the regression coefficients perfectly coincide with the least-squares estimates: $\tilde{b}_1 = b_1$ and $\tilde{b}_0 = b_0$.

However, differentiating with respect to σ^2 yields a different variance estimator:

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = 0 \tag{4}$$

Solving this provides the MLE for the variance:

$$\tilde{s}^2 = \frac{\sum (Y_i - \tilde{b}_0 - \tilde{b}_1 X_i)^2}{n} = \frac{SS_{res}}{n} \tag{5}$$

Bias vs. Efficiency Trade-off: The deterministic ratio uniting the two variance estimators is $\tilde{s}^2 = \frac{n-2}{n} s^2$. Because it divides by n rather than $n-2$, the MLE $\tilde{s}^2$ consistently underestimates the true underlying variance in small finite samples (it is biased). However, as the sample size scales to infinity ($n \to \infty$), this bias decays to zero, making the MLE asymptotically unbiased and statistically efficient.

Hypothesis Testing and Confidence Intervals

With the assumption that $\epsilon_i \sim N(0, \sigma^2)$ firmly in place, the estimators b_0 and b_1 are normally distributed. This allows us to conduct formal hypothesis testing and construct confidence intervals.

Inference for the Slope Parameter (β_1)

We test the null hypothesis $H_0 : \beta_1 = \beta_{10}$ (most commonly $\beta_{10} = 0$, testing for a significant linear relationship).

Case 1: Known Variance (σ^2)

The test statistic follows a standard normal distribution:

$$Z_1 = \frac{b_1 - \beta_{10}}{\sqrt{\sigma^2/S_{XX}}} \sim N(0,1)$$

We reject H_0 if $|Z_1| > z_{\alpha/2}$. The corresponding $100(1-\alpha)\%$ Confidence Interval is:

$$\left[b_1 - z_{\alpha/2}\sqrt{\frac{\sigma^2}{S_{XX}}}, \quad b_1 + z_{\alpha/2}\sqrt{\frac{\sigma^2}{S_{XX}}} \right]$$

Case 2: Unknown Variance (Using s^2)

Replacing σ^2 with s^2 forces the test statistic to follow a Student's t-distribution with $n-2$ degrees of freedom:

$$t_1 = \frac{b_1 - \beta_{10}}{\sqrt{s^2/S_{XX}}} \sim t_{n-2}$$

We reject H_0 if $|t_1| > t_{n-2,\,\alpha/2}$. The Confidence Interval becomes:

$$\left[b_1 - t_{n-2,\,\alpha/2}\sqrt{\frac{s^2}{S_{XX}}}, \quad b_1 + t_{n-2,\,\alpha/2}\sqrt{\frac{s^2}{S_{XX}}} \right]$$

Inference for the Intercept Term (β_0)

Similarly, we test $H_0 : \beta_0 = \beta_{00}$.

Case 1: Known Variance (σ^2)

The standardized statistic is:

$$Z_0 = \frac{b_0 - \beta_{00}}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)}} \sim N(0,1)$$

We reject H_0 if $|Z_0| > z_{\alpha/2}$.

Case 2: Unknown Variance (Using s^2)

$$t_0 = \frac{b_0 - \beta_{00}}{\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)}} \sim t_{n-2}$$

We reject H_0 if $|t_0| > t_{n-2,\,\alpha/2}$. The corresponding Confidence Interval is:

$$\left[b_0 - t_{n-2,\,\alpha/2}\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)}, \quad b_0 + t_{n-2,\,\alpha/2}\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)} \right]$$

References

[1] Blitzstein, J. K., & Hwang, J. (2019). *Introduction to probability.* Chapman and Hall/CRC.

[2] Ross, Sheldon M. (2020). *Introduction to probability and statistics for engineers and scientists.* Academic press.

Disclaimer