Unveiling the Statistical Foundations of Simple Linear Regression

An Aaalytical review of assumptions and estimators in the classical linear model.

Hernández Pérez Erick Fernando

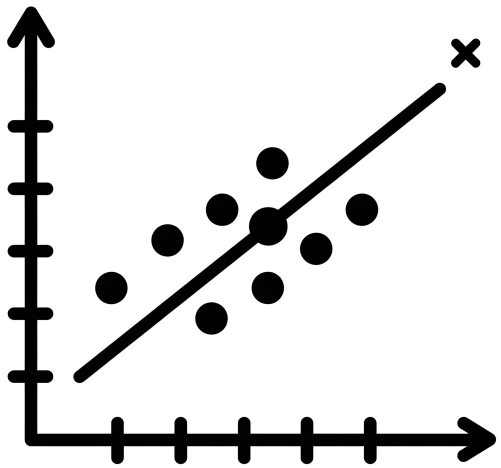

Figure 1

Linear regression stands as one of the most fundamental and widely utilized tools in the fields of statistics, econometrics, and data analysis. Its simplicity, interpretability, and broad applicability have made it a preferred starting point for modeling relationships between variables in countless disciplines. Specifically, simple linear regression addresses the relationship between a single independent variable and a dependent variable by fitting a straight line to the observed data, aiming to explain how changes in one variable are associated with changes in the other.

Beneath its practical and intuitive appeal, however, lies a well-defined statistical framework grounded in probability theory and inferential methods. The reliability of the results obtained from a linear regression model depends not only on the computational procedures involved but also on the validity of several underlying assumptions. These assumptions, concerning the behavior of the error terms and the properties of the explanatory variable, are essential for ensuring that the estimators derived from the model are unbiased, efficient, and suitable for hypothesis testing.

This article seeks to explore the statistical foundations that support simple linear regression, with a particular focus on the assumptions that sustain the model, the inferential techniques that accompany it, and the diagnostic tests that assess its adequacy. While the formal derivation of the ordinary least squares (OLS) method — the standard technique for estimating regression coefficients — will not be presented in this work, we will base our discussion on the estimators that result from the OLS procedure. These estimators serve as the core elements for analyzing the expected values, variances, and covariances within the model, and for constructing relevant hypothesis tests about the parameters involved.

What is a Simple Linear Regression?

> Simple linear regression provides a model of the relationship between the magnitude of one variable and that of a second - for example, as X increases, Y also increases. Or as X increases, Y decreases. (Bruce et al., 2020)[1]

Simple linear regression provides a framework for understanding and quantifying the relationship between two continuous variables. At its essence, it models how the expected value of one variable changes in response to variations in another. Typically denoted as Y (the dependent or response variable) and X (the independent or explanatory variable), the model posits that changes in X are associated with systematic changes in the average value of Y.

This relationship can take on different forms: as X increases, Y might also increase, suggesting a positive association between the variables; conversely, as X increases, Y could decrease, indicating a negative relationship. The strength and direction of this relationship are captured by the slope coefficient of the regression line, while the intercept represents the expected value of Y when X is zero.

The intuitive appeal of simple linear regression lies in its ability to summarize the general pattern of data through a straight line, making it possible to describe and predict outcomes with relative simplicity. By providing a quantitative measure of the association between two variables, this method serves as a foundational tool for statistical modeling, offering insights into how one variable behaves in relation to another in a wide range of applied contexts.

Ordinary Least Squares

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{1}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \tag{2}$$

And the matrix form

$$\beta = (x^T x)^{-1}(x^T y) \tag{3}$$

Regression functions

A regression function is a mathematical expression that describes the expected value of a dependent variable Y as a function of one or more independent variables $X_1, X_2, \ldots, X_k$. It expresses how the average value of Y changes in response to changes in the predictors.

$$E(Y|x_1, ..., x_k) = \beta_1 + \beta_2 x_1 + ... + \beta_{k+1} x_k \tag{4}$$

The regression coefficients β_i are parameters whose values will be estimated as $\hat{\beta}_i$.

Simple Linear Regression

Let's assume that for any given $X = x$, the random variable Y can be expressed as:

$$Y = \beta_1 + \beta_2 x + \epsilon$$

where ϵ is a random variable with a normal distribution $\mathcal{N}(0, \sigma^2)$ (mean equal to zero and variance equal to σ^2). From this assumption, we can deduce that the conditional distribution of Y given $X = x$ is a normal distribution $\mathcal{N}(\beta_1 + \beta_2 x, \sigma^2)$.

Additionally, it is usually assumed that the errors ϵ are uncorrelated, meaning that one error does not influence another, as is typically the case in time series models.

Now, suppose we obtain n pairs of observations $(x_1, Y_1), \ldots, (x_n, Y_n)$, and make the following assumptions:

- For any x_i, the random variables Y_i are independent.
- The distribution of Y_i is $\mathcal{N}(\beta_1 + \beta_2 x_i, \sigma^2)$.

- The errors are uncorrelated.
- Therefore, for given values of the vector $x = (x_1, \ldots, x_n)$ and the parameters β_1, β_2, and σ^2, the joint probability density function of $Y_1, \ldots, Y_n$ is:

$$f_n(y \mid x, \beta_1, \beta_2, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i)^2\right] \tag{5}$$

For any observed vector $y = (y_1, \ldots, y_n)$, this function will be the likelihood function of the parameters β_1, β_2, and σ^2.

From (5), it can be observed that, regardless of the value of σ^2, the values of β_1 and β_2 that maximize the likelihood function will be those that minimize the sum of squares (this analysis can be made more accessible by considering the function $f(x) = c_1 e^{-c_2 x}$, where c_1 and c_2 are arbitrary positive constants):

$$\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i)^2 \tag{6}$$

However, this sum of squares is precisely the one minimized by the ordinary least squares method. Therefore, the maximum likelihood estimators [MLE] of the regression coefficients are the same as the ordinary least squares estimators, which were previously reviewed in their corresponding section (Equations 1 and 2).

Finally, MLE of σ^2 can be obtained by first replacing β_1 and β_2 with their respective estimators, $\hat{\beta}_1$ and $\hat{\beta}_2$, and then maximizing the resulting expression with respect to σ^2:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 \tag{7}$$

A way of summarizing the assumptions made in the classical simple linear regression model is by explicitly listing them along with their corresponding statistical terms. These assumptions ensure the desirable properties of the ordinary least squares (OLS) estimators and the validity of statistical inference procedures. The assumptions are as follows:

- **Linearity in parameters:** The relationship between the dependent and independent variable is linear in the regression coefficients:

$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i \tag{8}$$

- **Exogeneity:** The expected value of the error term, conditional on the independent variable, is zero:

$$E(\epsilon_i | X_i) = 0 \tag{9}$$

- **Homoscedasticity:** The variance of the error term is constant for all values of the independent variable:

$$Var(\epsilon_i | X_i) = \sigma^2 \tag{10}$$

- **No autocorrelation of errors:** The error terms are uncorrelated across observations:

$$Cov(\epsilon_i, \epsilon_j) = 0, \text{ for all } i \neq j \tag{11}$$

- **Normality of errors:** The error terms follow a normal distribution:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{12}$$

  This assumption is mainly required for valid hypothesis testing and the construction of confidence intervals.
- **Independence of observations:** Each observation in the sample is independent of the others, meaning the pairs (X_i, Y_i) are independently drawn.

These assumptions are fundamental to the classical regression framework and ensure that the OLS estimators are unbiased, consistent, and efficient under the Gauss-Markov theorem, and that inference procedures based on normal theory are valid.

As a side note, one might wonder why, after carefully fitting a seemingly elegant regression model, we are then forced to run an arsenal of diagnostic tests like Durbin-Watson, Breusch-Pagan, Shapiro-Wilk, and others. The answer, as with many things in statistics, is that models love to lie beautifully on paper and fail spectacularly in practice. Hence, these tests act as vigilant sentinels, catching the misbehaviors our models would rather keep hidden.

Distribution of the Ordinary Least Squares Estimators

Let us examine the joint distribution of the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ when considered as functions of the random variables Y_i, for given values of x_i.

First, let us rewrite $\hat{\beta}_2$ as:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{13}$$

From (13) we can see that $\hat{\beta}_2$ is a linear combination of $Y_1, \ldots, Y_n$. Since the Y_i are independent and each follows a normal distribution, it follows that $\hat{\beta}_2$ also has a normal distribution. Moreover:

$$E(\hat{\beta}_2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(Y_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{14}$$

Since $E(Y_i) = \beta_1 + \beta_2 x_i$ for $i = 1, \ldots, n$, we obtain:

$$\begin{aligned}
E(\hat{\beta}_2) &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_1 + \beta_2 x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum \beta_1 x_i - \beta_1 \bar{x} + \beta_2 x_i^2 - \beta_2 x_i \bar{x}}{\sum (x_i - \bar{x})^2} \\
&= \frac{n\beta_1 \bar{x} - n\beta_1 \bar{x} + \sum \beta_2 (x_i^2 - x_i \bar{x})}{\sum (x_i - \bar{x})^2} \\
&= \beta_2 \frac{\sum x_i^2 - \bar{x} \sum x_i)}{\sum (x_i - \bar{x})^2} \\
&= \beta_2 \frac{\sum x_i^2 - n\bar{x}^2}{\sum (x_i - \bar{x})^2} \\
&= \beta_2 \frac{\sum x_i^2 - n\bar{x}^2}{\sum x_i^2 - n\bar{x}^2} \\
&= \beta_2
\end{aligned}$$

Therefore:

$$E(\hat{\beta}_2) = \beta_2 \tag{15}$$

And this means that it is an unbiased estimator. Moreover, since the Y_i are independent and each has variance σ^2, from (13) it follows that:

$$Var(\hat{\beta}_2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 Var(Y_i)}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{16}$$

For the case of β_1, it is necessary to recall that $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{x}$, and this makes it a linear function of $Y_1, \ldots, Y_n$. Therefore, $\hat{\beta}_1$ will have a normal distribution.

$$\begin{aligned}
E(\hat{\beta}_1) &= E(\bar{Y}) - \bar{x}E(\hat{\beta}_2) \\
&= E(\frac{1}{n}\sum Y_i) - \bar{x}\beta_2 \\
&= \frac{1}{n}\sum E(Y_i) - \bar{x}\beta_2 \\
&= \frac{1}{n}\sum(\beta_1 + \beta_2 x_i) - \bar{x}\beta_2 \\
&= \frac{1}{n}n\beta_1 + \frac{1}{n}\beta_2\sum x_i - \bar{x}\beta_2 \\
&= \beta_1 + \frac{1}{n}n\beta_2\bar{x} - \bar{x}\beta_2 \\
&= \beta_1 + \beta_2\bar{x} - \bar{x}\beta_2 \\
&= \beta_1
\end{aligned}$$

Therefore

$$E(\hat{\beta}_1) = \beta_1 \tag{17}$$

For the variance of $\hat{\beta}_1$, we have:

$$\begin{aligned}
Var(\hat{\beta}_1) &= Var(\bar{Y} - \bar{x}\hat{\beta}_2) \\
&= Var(\bar{Y}) + \bar{x}^2 Var(\hat{\beta}_2) - 2\bar{x}\,Cov(\bar{Y}, \hat{\beta}_2)
\end{aligned}$$

But first, let us analyze the covariance. Recall that:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{S_{xx}} \tag{18}$$

and

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i \tag{19}$$

Both are linear combinations of the Y_i, so their covariance is computed as:

$$\mathrm{Cov}(\bar{Y}, \hat{\beta}_2) = \sum_{i=1}^n a_i b_i \,\mathrm{Var}(Y_i) \tag{20}$$

where $a_i = \frac{1}{n}$ and $b_i = \frac{x_i - \bar{x}}{S_{xx}}$. Since $\mathrm{Var}(Y_i) = \sigma^2$ for all i, we have:

$$\mathrm{Cov}(\bar{Y}, \hat{\beta}_2) = \sigma^2 \sum_{i=1}^n \frac{1}{n} \cdot \frac{(x_i - \bar{x})}{S_{xx}} = \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^n (x_i - \bar{x}) \tag{21}$$

But since $\sum_{i=1}^n (x_i - \bar{x}) = 0$, then:

$$\mathrm{Cov}(\bar{Y}, \hat{\beta}_2) = 0 \tag{22}$$

In this way, returning to our variance calculation:

$$
\begin{aligned}
Var(\hat{\beta}_1) &= Var(\bar{Y}) + \bar{x}^2 Var(\hat{\beta}_2) - 2\bar{x}\, Cov(\bar{Y}, \hat{\beta}_2) \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - 0 \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + n\bar{x}^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(\frac{\sum x_i^2 - n\bar{x}^2 + n\bar{x}^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(\frac{\sum x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)
\end{aligned}
$$

Therefore:

$$
Var(\hat{\beta}_1) = \sigma^2 \left(\frac{\sum x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \tag{23}
$$

Finally, the covariance is

$$
\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\bar{x}\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{24}
$$

The Gauss-Markov Theorem in the Context of Simple Linear Regression

In the context of simple linear regression, one of the most important results in classical linear model theory is the **Gauss-Markov theorem**. This theorem provides a formal justification for the use of the ordinary least squares (OLS) estimators.

Under the classical assumptions of the linear regression model the Gauss-Markov theorem states that the OLS estimators of the regression coefficients are the *Best Linear Unbiased Estimators* (BLUE).

This means that, among all possible linear and unbiased estimators of the coefficients, the OLS estimators have the minimum variance. In other words, no other linear and unbiased estimator can be more precise than the one obtained through the least squares method.

The theorem highlights the optimality of OLS within a well-defined class of estimators and establishes a theoretical foundation for its widespread use in regression analysis.

Conclusion

As we have seen throughout this article, the mathematical and statistical foundations behind a so-called *simple* linear regression model are far from simple. Numerous derivations and properties were presented, and while every effort was made to explain these results as clearly and fluently as possible without omitting important steps, it ultimately all rests upon fundamental statistical concepts such as expectation, variance, normal distributions, and covariance, among others.

It is also important to emphasize that everything we have discussed so far has been purely theoretical. And while this theory lays out the path for estimating regression coefficients and making inferences, it does not immediately validate the use of a simple linear regression model with any given dataset. The theory essentially says, *"Yes, you may proceed with a simple linear regression, but..."* — and that "but" refers to the necessary verification of the model's underlying assumptions. Normality, homoscedasticity, independence of errors, and lack of autocorrelation are crucial to ensuring the reliability of the estimates and the validity of any subsequent inferences.

Lastly, it is worth noting that since this article focused on a simple linear regression model — involving only one explanatory variable — it spares us from having to deal with problems such as multicollinearity, which are

exclusive to multiple regression models. For this reason, issues of multicollinearity were deliberately left out of this discussion.

References

[1] Bruce, P., Bruce, A. & Gedeck, P. (2020). *Practical Statistics for Data Scientists*. O'Reilly.

[2] DeGroot, M. (1988). *Probabilidad y estadística*. Addison-Wesley Iberoamericana.

[3] Montgomery, D., Peck, E. & Vining, G. (2006). *Introducción al análisis de regresión lineal* [Introduction to linear regression analysis]. CECSA.