

Understanding Regression Through the Origin: A Comprehensive Guide

Authored by
Mohammed loot

November 1, 2025

RECOMMENDED CITATION

Mohammed loot (2025). *Understanding Regression Through the Origin: A Comprehensive Guide*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=7716>

The Foundation of Linear Modeling: Simple Linear Regression (SLR)

[Simple linear regression](#) (SLR) serves as a fundamental statistical methodology used extensively across science and industry. Its core purpose is to accurately quantify the linear relationship between two variables: a single [predictor variable](#) (often denoted as x) and a corresponding [response variable](#) (y). By modeling this dependency, SLR enables analysts to perform reliable prediction, detailed inference, and relationship assessment within a defined data space.

The standard mathematical expression for a simple linear regression model incorporates two estimated parameters and an error component. This robust structure accounts for both the starting point of the relationship and its rate of change. The canonical formulation is critical for understanding conventional statistical modeling:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

A precise understanding of the terms within this equation is absolutely essential for correctly interpreting the resulting statistical outcomes and ensuring model validity. Each component plays a specific role in defining the overall fitted line:

y : Represents the observed or predicted value of the [response variable](#), which is the outcome we are attempting to model.

β_0 : This is the critical [intercept term](#). Statistically, it predicts the mean value of the response variable (y) when the predictor variable (x) is exactly zero.

β_1 : Known as the slope or the [coefficient estimate](#), this parameter quantifies the average magnitude of change in the response variable associated with every one-unit increase in the predictor variable (x).

x : Represents the specific value of the [predictor variable](#) (the independent input driving the prediction).

ε : Denotes the random error term, which captures all unobserved factors and inherent variability not explained by the linear relationship between x and y .

In standard regression practice, both the intercept (β_0) and the slope (β_1) are simultaneously estimated using optimization techniques, most commonly [Ordinary Least Squares](#) (OLS). The inclusion of the intercept ensures that the fitted regression line is positioned optimally to minimize the sum of squared errors relative to all data points. This standard approach is used even if the theoretical relationship does not physically mandate the line to pass through the origin (0, 0).

Defining Regression Through the Origin (RTO)

While standard SLR provides maximum flexibility in fitting the data, certain real-world phenomena require a constrained modeling approach. This specialized modification is known as **regression**

through the origin (RTO). This model is invoked specifically when robust theoretical principles or established external [domain knowledge](#) strictly dictate that the response variable **must** be zero whenever the predictor variable is zero.

The defining characteristic of RTO is that it forces the fitted regression line to pass precisely through the coordinates (0, 0) on the plane. Mathematically, this constraint is enforced by explicitly removing the [intercept term](#) (β_0) from the standard model equation. The resulting RTO model is simplified into a proportional form:

$$y = \beta_1 x + \varepsilon$$

The primary statistical implication of RTO is that the estimation process focuses solely on calculating the single remaining parameter: the slope [coefficient estimate](#) (β_1). This coefficient then represents the constant factor of proportionality or the rate of change between the predictor (x) and the response (y). Furthermore, the definition of the model guarantees that if $x=0$, the predicted outcome y must also be zero.

It is crucial to understand that imposing this zero-intercept constraint fundamentally alters the underlying mathematical optimization used by [OLS](#). Since the line is restricted, the slope derived from an RTO model will almost certainly differ from the slope calculated using a standard SLR model on the identical dataset. The constraint influences the overall minimization of the [sum of squared errors](#), often resulting in a larger overall error compared to standard SLR, but yielding a theoretically sounder model.

Theoretical Justification: When RTO is Necessary

The decision to utilize **regression through the origin** should never be based merely on achieving a better statistical fit (e.g., a higher R-squared value) but must be rigorously grounded in strong, non-statistical reasoning or established scientific principles. RTO is appropriate only when the relationship represents an intrinsic [proportionality](#) where the absence of the independent variable necessarily implies the absence of the dependent variable.

This proportional model finds frequent application in specialized quantitative fields like engineering, chemistry, and physics, where measurements are absolute and physical laws govern the variables. A classic example is modeling the relationship between the mass of a substance (response) and its volume (predictor). If the volume is zero, the mass must be zero. Attempting to use standard [simple linear regression](#) in such a case would force the model to estimate a non-zero intercept ($\beta_0 \neq 0$) simply to minimize residual error across the sampled data.

This resulting non-zero intercept creates a critical problem: it leads to nonsensical predictions when extrapolating back to $x=0$, severely undermining the model's physical or biological validity.

For instance, predicting a positive weight when the volume is zero is physically impossible. Therefore, RTO ensures theoretical consistency, even if it sacrifices a small degree of empirical fit within the sampled range.

Statistically, implementing RTO means that one fewer parameter (the intercept) requires estimation. This results in a gain of one [degree of freedom](#). While this efficiency gain can be relevant in scenarios involving extremely small sample sizes, analysts must be wary of using RTO merely for statistical convenience. The primary driver must remain the theoretical integrity dictated by the [domain knowledge](#) surrounding the variables.

Practical Application: Implementing RTO in R (Tree Growth Example)

To illustrate the tangible differences between the standard and constrained models, we will use the [R programming environment](#) to model a biologically constrained scenario: predicting tree height based on trunk circumference. In this context, it is biologically sound to assume that a circumference of zero must yield a height of zero, making RTO the theoretically superior choice. We utilize data collected from 15 hypothetical tree samples.

The dataset visualization below shows the relationship between the two variables before modeling:

Circumference (inches)	Height (inches)
15	200
19	234
25	285
39	375
44	440
46	470
49	564
54	544
67	639
79	750
81	830
84	854
88	901
90	912
99	989

The subsequent R code block demonstrates the essential steps: creating the data structure, fitting the standard SLR model (`model1`), and fitting the **regression through the origin** model

(`model_origin`). Note the key syntax difference in R's `lm` function: the addition of `+ 0` or `- 1` explicitly removes the intercept parameter, enforcing the constraint. The code then generates a visual comparison of the two fitted lines.

#create data frame

```
df <- data.frame(circ=c(15, 19, 25, 39, 44, 46, 49, 54, 67, 79, 81, 84, 88, 90, 99),  
height=c(200, 234, 285, 375, 440, 470, 564, 544, 639, 750, 830, 854,  
901, 912, 989))
```

#fit a simple linear regression model (standard SLR)

```
model <- lm(height ~ circ, data = df)
```

#fit regression through the origin (Note: using '0 +' explicitly removes the intercept)

```
model_origin <- lm(height ~ 0 + ., data = df)
```

#create scatterplot

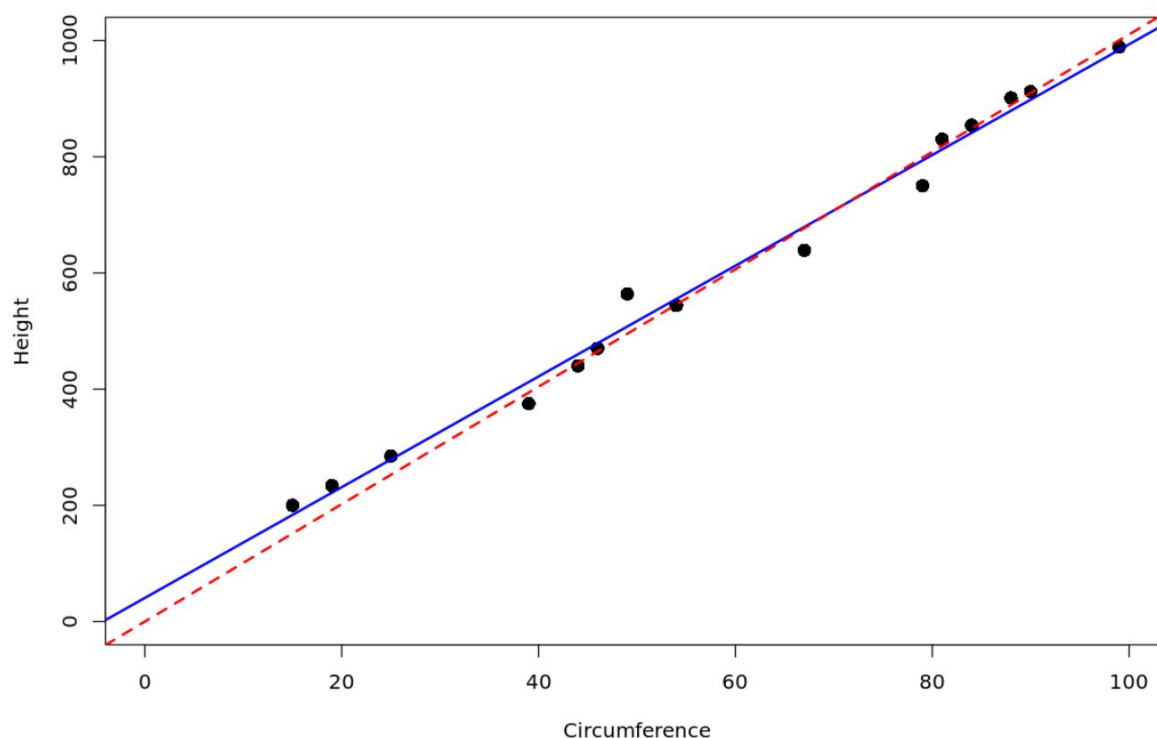
```
plot(df$circ, df$height, xlab='Circumference', ylab='Height',  
cex=1.5, pch=16, ylim=c(0,1000), xlim=c(0,100))
```

#add the fitted regression lines to the scatterplot

```
abline(model, col='blue', lwd=2)
```

```
abline(model_origin, lty='dashed', col='red', lwd=2)
```

The resulting visualization clearly contrasts the outcome of the two modeling approaches.



The **blue solid line** represents the standard [simple linear regression](#) fit, which clearly estimates a positive intercept (a starting point above zero). Conversely, the **red dashed line** represents the constrained **regression through the origin** model, which is visually and mathematically anchored at the point (0, 0), demonstrating the power of the imposed constraint in maintaining theoretical fidelity.

Analyzing Results and Crucial Statistical Cautions

To quantify the practical impact of forcing the line through the origin, we examine the extracted [coefficient estimates](#) for both the standard and constrained models. This comparison highlights how the removal of one parameter necessitates a significant adjustment in the remaining slope parameter to compensate for the restricted fitting space.

```
#display coefficients for simple linear regression model
coef(model)
```

```
(Intercept) circ
40.696971 9.529631
```

```
#display coefficients for regression model through the origin
coef(model_origin)
```

```
circ
```

10.10574

The standard simple linear regression model yields the following fitted equation: **Height = 40.6969 + 9.5296(Circumference)**. In this equation, the [intercept term](#) of 40.7 units predicts that a tree with zero circumference would still possess a height of approximately 40.7 units. As previously discussed, this interpretation is physically or biologically inconsistent in the context of tree growth.

In stark contrast, the **regression through the origin** model produces the equation: **Height = 10.1057(Circumference)**. The model satisfies the theoretical constraint that zero circumference results in zero height. Furthermore, notice the significant shift in the slope coefficient: 9.5296 for the standard model versus 10.1057 for the RTO model. This difference underscores how the slope must adjust upward when the line is forced to pivot through the origin, demonstrating a higher estimated rate of change between the variables.

While RTO provides theoretical fidelity in proportional relationships, analysts must exercise significant caution. If the true underlying relationship does **not** pass through the origin--perhaps due to measurement errors, calibration biases, or a non-proportional physical law--forcing the constraint will inevitably introduce statistical [bias](#) into the slope estimate (β_1). This bias can lead to severely inaccurate predictions, especially when extrapolating outside the range of observed data points. Therefore, the use of **regression through the origin** should always be meticulously justified by strong domain knowledge and clearly documented in any final analysis or statistical report.

Additional Resources for Linear Modeling

For those interested in delving deeper into linear modeling techniques, the assumptions underlying regression, and advanced constraint methods, the following resources provide valuable additional information: