

How to Distinguish Regression Through the Origin from Correlation

Authored by
stats writer

December 2, 2025

RECOMMENDED CITATION

stats writer (2025). *How to Distinguish Regression Through the Origin from Correlation*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=103498>

The definitive answer is no: regression through the origin is fundamentally distinct from Correlation. While both statistical tools are utilized to analyze relationships between variables, they serve different analytical purposes and operate under distinct mathematical assumptions. Regression through the origin (RTO) is a specialized form of linear regression where the fitted regression line is mathematically constrained to pass through the point (0, 0), or the origin. Correlation, conversely, is a descriptive metric, usually quantified by the correlation coefficient (r), which measures the strength and direction of the linear association between two variables, irrespective of whether the data passes through the origin or not.

The Fundamentals of Simple Linear Regression

Simple linear regression (SLR) is one of the most widely used methods in statistics for modeling the relationship between a single predictor variable (independent variable) and a single response variable (dependent variable). The primary goal of SLR is to establish a linear equation that best predicts the value of the response variable based on the value of the predictor variable. This predictive framework allows researchers to understand, quantify, and forecast associations within various data sets, ranging from economics to biology. SLR seeks the line that minimizes the sum of squared residuals, providing the best linear unbiased estimates (BLUE) for the coefficients.

The standard equation for a simple linear regression statistical model includes both a slope and an intercept term. This intercept term is crucial as it anchors the regression line, representing the baseline value of the response variable when the predictor variable is zero. Ignoring or misinterpreting the intercept can lead to significant modeling errors, particularly when extrapolating results beyond the observed data range. The standard model form is universally recognized and essential for nearly all introductory statistical applications.

A simple linear regression model takes on the following form, which dictates how the response variable is estimated based on the inputs:

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

where the terms are defined precisely to provide clear interpretability of the model parameters:

y: Represents the predicted value of the response variable.

β_0 : Known as the **intercept** term. This is the estimated value of y when the predictor x is exactly 0.

β_1 : Known as the **slope** coefficient. This measures the average change (increase or decrease) in the response variable y associated with a one-unit increase in the predictor variable x.

x: Represents the observed value of the predictor variable.

Defining the Correlation Coefficient

While regression focuses on prediction and quantifying causality (or association in a predictive sense), Correlation provides a measure of linear dependence between two variables, X and Y. This measure is descriptive rather than predictive. The most common measure is the Pearson product-moment correlation coefficient (r), which summarizes how closely the data points cluster around any potential straight line, whether that line passes through the origin or not. It is a normalized measurement, meaning its value is always constrained between -1 and +1.

A correlation coefficient of +1 indicates a perfect positive linear relationship, where as X increases, Y increases proportionally. Conversely, a value of -1 indicates a perfect negative linear relationship. A value near 0 suggests a weak or non-existent linear relationship. It is critical to understand that correlation only captures the strength of the linear association; it does not account for the slope of the relationship or the specific baseline value (the intercept) of the variables. Correlation simply tells us how well a straight line describes the data, not what that line specifically is.

Furthermore, correlation is symmetrical; the correlation between X and Y is identical to the correlation between Y and X. Regression, however, is asymmetrical: fitting Y predicted by X is different from fitting X predicted by Y. This fundamental difference underscores why regression is used for prediction and modeling conditional relationships, while correlation is used solely for characterizing mutual association. Correlation does not imply the existence of a specific regression line, nor does it necessitate that the relationship passes through the origin.

Understanding Regression Through the Origin (RTO)

A specialized modification of the standard SLR model is known as **regression through the origin** (RTO). This technique is implemented when there is compelling theoretical or physical evidence suggesting that if the predictor variable (x) is zero, the response variable (y) must also logically be zero. By imposing this constraint, we force the regression line to pass directly through the point (0, 0), thereby eliminating the intercept term (β_0) from the equation. This is a critical assumption that transforms the minimization problem, as the line must pivot exactly at the origin.

This type of model takes on the following simplified form, effectively setting the intercept (β_0) to zero:

$$y = \beta_1 x$$

Notice that the **intercept term** (β_0) has been completely dropped from the model. In standard regression, the intercept is estimated from the data, but in RTO, it is assumed, or fixed, at zero. This adjustment changes the estimation process (usually Ordinary Least Squares) because the

slope (β ?) is now calculated to minimize the sum of squared errors subject to the constraint that the line must pivot around the origin. This calculation typically results in a different slope estimate than the one derived from the full SLR model, often providing a steeper slope to account for the forced zero intercept.

Justification and Use Cases for RTO

RTO should only be employed when the non-negativity of both variables and the logical necessity of a zero-intercept are certain. Applying RTO inappropriately can severely bias the slope estimate and result in a poorly fitting model, particularly if the true underlying relationship possesses a non-zero intercept within the observed data range. A slight positive intercept, for instance, could skew the RTO slope upwards, leading to overestimation.

In practice, this constrained model is used most commonly in specific scientific or engineering fields where physical laws dictate the relationship, such as in certain areas of forestry, physics (e.g., Hooke's Law within certain limits), or stoichiometry in chemistry. For example, if measuring the volume of a substance based on its mass, zero mass must logically correspond to zero volume. Another example involves measuring the dosage response of a drug: if the dose (X) is zero, the response (Y) should be zero, regardless of the patient's baseline.

A classic illustration involves predicting tree height based on tree circumference. If a theoretical tree has a circumference of zero (meaning it does not exist or is just a seedling), its height must logically be zero. Thus, when fitting a regression model to this data, it would not make physical or logical sense for the estimated intercept term (the height when circumference is zero) to be a non-zero, positive value, as the standard SLR model might estimate based purely on minimizing squared error across the observed data points.

Practical Demonstration: Comparing SLR and RTO

To highlight the practical differences, we will compare the results of fitting an ordinary simple linear regression model against a model that implements regression through the origin using a real-world dataset. This example involves a biologist measuring tree circumference to predict tree height, reinforcing the previous discussion regarding the logical necessity of a zero intercept in this context.

Suppose the biologist collects the following measurements for a sample of 15 trees:

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 following R code demonstrates how both models are fitted and visualized on a scatterplot. Note the specific syntax used in R's `lm()` function to enforce the zero intercept using the `0 + .` term:

#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 (includes intercept)

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

#fit regression through the origin (syntax '0 + .' suppresses the intercept)

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

#create scatterplot

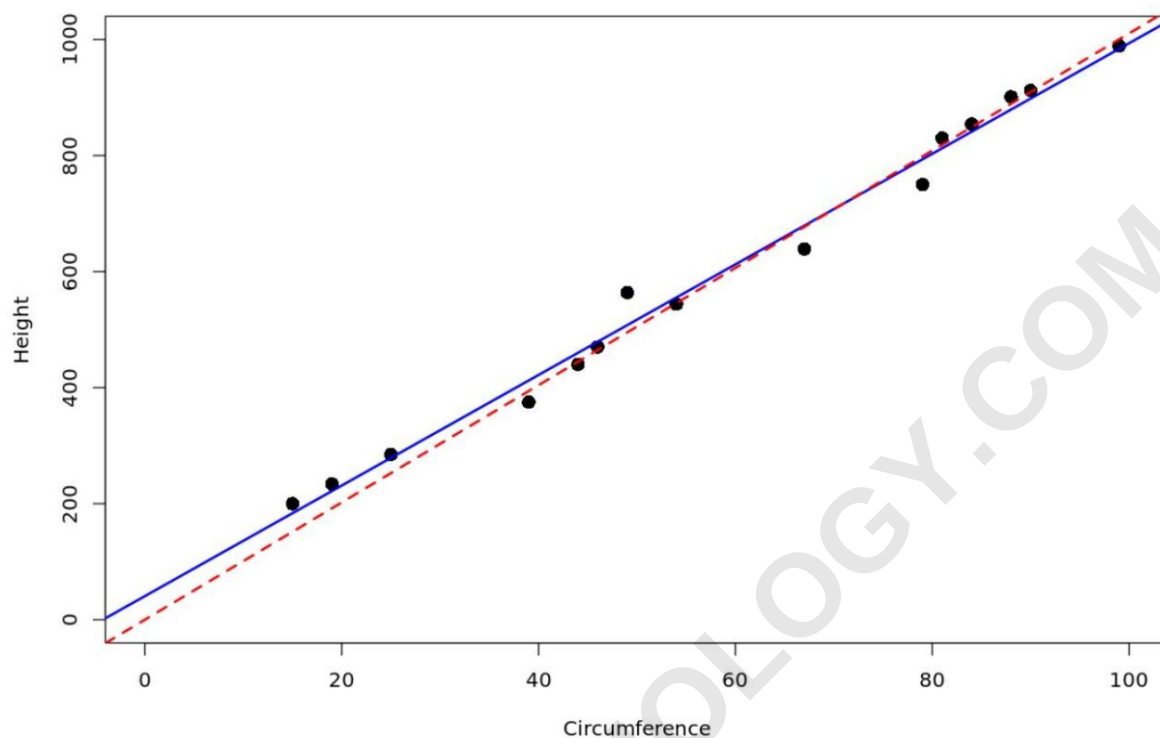
```
plot(df$circ, df$height, xlab='Circumference', ylab='Height',
cex=1.5, pch=16, ylim=c(0,1000), xlim=c(0,100))
```

#add the fitted regression lines to the scatterplot

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

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

Upon visualizing the results, the difference between the two models becomes immediately apparent:



The **red dashed line** clearly represents the regression through the origin model, which originates at (0, 0). The **blue solid line** represents the ordinary simple linear regression model, which is allowed to estimate its optimal intercept. Visually, the blue line appears to provide a slightly better fit to the observed data points near the center, but it estimates a positive height when the circumference is zero, which is biologically implausible and demonstrates the danger of extrapolation when the physical constraints are violated.

Interpreting Model Coefficients

The mathematical outcome of forcing the line through the origin is a change in the estimation of the slope coefficient (β). Since the intercept is fixed, the entire variance captured by the model must be channeled into the slope, resulting in a potentially altered slope estimate compared to the standard SLR model. We can examine the specific coefficient estimates generated by the R statistical software:

We use the following R code to extract and display the coefficient estimates for both models, providing a quantitative comparison:

#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 fitted equation derived from the standard simple linear regression model, including the estimated intercept term, is:

$$\text{Height} = 40.6969 + 9.5296(\text{Circumference})$$

This equation suggests that a tree with zero circumference would still be estimated to have a height of approximately 40.7 units, which is a major extrapolation error based on the physical constraints of the problem. However, the slope indicates that for every unit increase in circumference, the height increases by 9.53 units, representing the average relationship within the observed data range.

The fitted equation for the regression model through the origin is starkly different:

$$\text{Height} = 10.1057(\text{Circumference})$$

Notice that the coefficient estimates for the circumference variable are slightly different (9.53 vs. 10.11). By forcing the intercept to zero, the slope coefficient (10.11) increases to compensate for the lack of a baseline term (40.7). This ensures that the constrained line passes through the origin while still minimizing the overall prediction error relative to that fixed pivot point. For the specific application of tree growth, this RTO model provides a more logically sound prediction for extremely small or zero circumferences.

Critical Cautions When Employing RTO

The decision to use regression through the origin must be made with significant deliberation, as imposing this constraint can lead to a less optimal fit (in terms of maximizing R-squared) if the theoretical constraint is incorrect. Before using RTO, analysts must be absolutely certain--based on theoretical justification, physical laws, or prior knowledge--that a value of 0 for the predictor variable implies a value of 0 for the response variable. In many observational scenarios, it is almost impossible to know this causal relationship for sure, making the standard SLR model the safer and more robust choice.

One common, but often weak, justification for using RTO is to "save one degree of freedom" by avoiding the estimation of the intercept ($\beta?$). This rationale is rarely valid in modern statistical practice. If your sample size is large enough (N typically greater than 30 or 40), saving a single degree of freedom from estimating the intercept rarely makes a substantial difference to the overall model performance or inferential power. The potential bias introduced by forcing the line through the origin usually outweighs the marginal benefit of gaining one degree of freedom, particularly if the true intercept is, in fact, non-zero.

If you do choose to utilize regression through the origin, it is a professional requirement to state your precise reasoning and justification in your final analysis or report. Transparency is key, as RTO represents a significant deviation from the default assumptions of standard regression analysis.

Further Reading on Linear Modeling

The following tutorials provide additional information about linear regression and related modeling techniques that expand upon the foundational differences discussed here: