How to Make Predictions Using Linear Regression

Authored by

stats writer

December 4, 2025

Linear regression is a fundamental statistical method used extensively across various scientific and professional domains, ranging from finance to healthcare. At its core, it is designed to model the relationship between a scalar dependent variable (or response variable) and one or more independent variables (or predictor variables). The primary goal of this technique is to create a linear model that can accurately quantify the expected value of the dependent variable based on the corresponding values of the independent variables.

The immense utility of linear regression stems from its ability to facilitate informed prediction and forecasting. By establishing a formalized linear relationship--often represented as a straight line in simple cases--researchers can project future outcomes or estimate values for observations not yet measured. This process involves fitting a mathematical equation to observed data points, thereby creating a robust framework for understanding and utilizing correlation between variables.

While the underlying mathematics can be complex, the concept is straightforward: if two or more variables exhibit a consistent, quantifiable relationship, we can use the known values of the predictors to estimate the unknown value of the response. This predictive power makes linear regression an indispensable tool for analysts seeking to understand trends, assess risks, and make data-driven decisions. The quality of these predictions, however, hinges entirely on the quality of the data and the appropriate application of the underlying statistical assumptions.

The Predictive Power of Linear Regression Models

Linear regression serves as the bedrock for predictive analytics in many disciplines. The ability to express complex relationships using a simple, interpretable equation allows practitioners to move beyond mere descriptive statistics. Instead of simply describing historical data, we use the model to generate quantitative forecasts. This predictive capability is perhaps the single most common reason for developing and fitting a regression model in practical applications.

The model precisely quantifies the contribution of each predictor variable to the overall variation in the response variable. For instance, in finance, a linear model might quantify how changes in interest rates and GDP affect stock prices. In healthcare, it might measure how age and diet contribute to a specific biometric outcome. The fitted model equation acts as a precise mechanism for estimating new data points, provided these new points fall within the scope and range of the original sample data used for training.

The foundation of making predictions rests on the assumption that the observed historical relationship will persist for future or unobserved instances. Therefore, ensuring the model is correctly specified--meaning that the relationship is truly linear and meets all necessary statistical assumptions--is paramount. A poorly fitted model that violates core assumptions will produce inaccurate and potentially misleading predictions, regardless of how precise the calculated equation appears.

Four Essential Steps for Prediction Modeling

To successfully harness a regression model for forecasting and prediction, analysts must follow a rigorous, methodical process. Skipping any of these essential steps can lead to invalid conclusions or forecasts that fail spectacularly when applied to real-world scenarios.

The general methodology for leveraging a fitted regression model to predict the values of new observations involves four core stages, ensuring statistical robustness and maximizing the confidence in the resulting predictions.

Step 1: Data Collection and Preparation. The process begins by gathering comprehensive, high-quality data relevant to the variables of interest. This includes measurements for both the predictor variables and the response variable. Data must be cleaned, checked for outliers, and prepared for modeling. Ensuring adequate sample size and representative sampling is crucial for the generalizability of the final model.

Step 2: Model Fitting. Once the data is prepared, a regression model (simple or multiple linear regression) is fitted to the dataset. This step involves calculating the coefficients (the intercept and the slopes for each predictor) that minimize the sum of the squared residuals--a process commonly known as the Ordinary Least Squares (OLS) method. The result is the final regression equation.

Step 3: Model Verification and Validation. Before using the equation for prediction, it is mandatory to verify that the model is statistically sound and fits the data well. This involves checking statistical significance of the coefficients, assessing the overall fit using measures like R-squared, and critically, confirming that the underlying assumptions of the linear regression model (e.g., linearity, independence of errors, normality of residuals, and homoscedasticity) are met. A model that violates these assumptions is inherently unreliable for prediction.

Step 4: Prediction Application. Finally, the verified and validated regression equation is used to predict the values of new, unobserved data points. By plugging the values of the predictor variables for a new observation into the fitted equation, a point estimate for the response variable is generated. This estimate forms the basis of the forecast.

Case Study 1: Making Predictions with a Simple Linear Regression Model

A simple linear regression model involves only one predictor variable used to estimate the response variable. This is the simplest form of regression, making it highly intuitive for illustrating predictive application. Consider a scenario in clinical research where a doctor aims to model the relationship between a patient's weight and their height.

Suppose a physician diligently collects paired data--height (in inches) and weight (in pounds)--from

a sample of 50 patients. The objective is to determine if weight can reliably predict height. The physician designates "weight" as the predictor variable and "height" as the response variable. After fitting the data using statistical software, the resulting fitted regression equation is derived:

The calculated regression equation is determined to be:

Height = 32.7830 + 0.2001 * (Weight)

Following the initial fitting, the physician meticulously checks the residuals, confirming that the underlying assumptions of the simple linear model are satisfied. This verification step confirms that the linear relationship holds true for the sample population and that the model possesses sufficient predictive validity. Armed with this verified model, the doctor can confidently use it to predict the height of new patients based solely on their measured weight, provided their weight falls within the range of the original 50 patients sampled.

For example, imagine a new patient arrives whose weight is 170 pounds. To calculate the predicted height (the point estimate), we substitute 170 into the established equation:

Height = 32.7830 + 0.2001 * (170) = **66.8 inches**

Thus, based on the statistical model derived from the sample, the best prediction for the height of a patient weighing 170 pounds is 66.8 inches.

Case Study 2: Making Predictions with a Multiple Linear Regression Model

When a single variable is insufficient to capture the complexity of a phenomenon, we turn to Multiple linear regression. This powerful extension allows for the incorporation of two or more independent variables to predict the outcome of a single dependent variable. This scenario is particularly common in fields like economics, where many factors simultaneously influence a key metric.

Consider an economist studying factors that influence personal income. The economist hypothesizes that both educational attainment and work effort contribute significantly to yearly income. Data is collected from 30 individuals, recording their total years of schooling, their average weekly hours worked, and their corresponding yearly income. Here, "yearly income" is the response variable, while "years of schooling" and "weekly hours worked" are the multiple predictor variables.

After running the regression analysis, the economist obtains the following fitted equation, which includes coefficients for the intercept and both predictor variables:

Income = 1,342.29 + 3,324.33 * (Years of Schooling) + 765.88 * (Weekly Hours Worked)

The critical step of model validation follows, where the economist confirms that the statistical assumptions inherent to multiple linear regression are met--including checks for multicollinearity, which is vital when using multiple predictors. With a validated model, the economist gains a robust tool for forecasting income based on specific educational and work profiles.

Suppose the economist wishes to predict the yearly income for a new job candidate who possesses 16 years of total schooling and works an average of 45 hours per week. Using the model, we substitute these values into the multiple regression equation:

Income = 1,342.29 + 3,324.33 * (16) + 765.88 * (45) = **$88,996.17**

This predicted value represents the best point estimate for the yearly income of an individual with those specific characteristics, based on the patterns observed in the economist's sample data.

Understanding Point Estimates and Confidence Intervals

When a regression model is used to make a forecast for a new observation, the single value calculated is known as a point estimate. While this point estimate represents the model's single best prediction, it is highly improbable that this value will exactly match the true outcome of the new observation due to inherent model error and sampling variability.

To effectively communicate the reliability and precision of a forecast, analysts must move beyond the single point estimate and incorporate measures of uncertainty using intervals. The Confidence Interval provides a range of values that is likely to contain a population parameter with a certain level of confidence.

For example, instead of predicting that a new patient will be exactly 66.8 inches tall, we may create a 95% Confidence Interval for the mean prediction of all 170-pound patients:

95% Confidence Interval =

We interpret this interval to mean that we are 95% confident that the true average height for all individuals weighing 170 pounds falls somewhere between 64.8 inches and 68.8 inches. It is crucial to note that the confidence interval addresses the uncertainty regarding the mean response of the population.

Distinguishing Confidence and Prediction Intervals

Although often confused, the Prediction Interval is arguably more relevant when forecasting a single, specific new observation, rather than the population mean. While the confidence interval estimates the mean response, the prediction interval estimates the actual value for a specific individual outcome.

Because the prediction interval must account for both the uncertainty in estimating the regression line (captured by the confidence interval) and the inherent random error (residual variability) of individual data points around that line, the prediction interval is always wider than the confidence interval for the same observation. When providing a forecast for a new individual, the prediction interval offers a more conservative and realistic estimate of the total expected uncertainty.

Crucial Cautions: Avoiding Extrapolation and Sampling Bias

While regression models offer significant predictive power, their application is not without serious constraints. Two primary pitfalls can invalidate predictions: extrapolation outside the data range and making predictions for a population different from the sampled group. Adhering to these cautions is fundamental for ethical and accurate statistical practice.

Caution 1: Only Predict Within the Range of Observed Data

The mathematical relationships defined by a fitted regression equation are only statistically validated across the range of the independent variables used to train the model. Using the model to predict outcomes for predictor values that fall outside this observed range is known as **extrapolation**, and it is highly discouraged.

For example, suppose we fit the height and weight regression model using a sample where patient weights ranged only between 120 pounds and 180 pounds. It would be invalid to use this model to estimate the height of an individual who weighed 250 pounds because this falls significantly outside the range of the predictor variable used to estimate the model.

The true relationship between weight and height might stabilize, curve, or change dramatically at weight values beyond 180 pounds. Since the model is unaware of data points outside its training range, using it to forecast an outcome via extrapolation would produce an invalid and potentially nonsensical point estimate. Predictions must be strictly constrained to the domain of the data used for the estimation of the regression model.

Caution 2: Only Predict for the Population Sampled

The validity and generalizability of the predictions are inherently tied to the statistical population from which the sample data was drawn. A regression model should only be used to make predictions for individuals or entities belonging to the same statistical population as the original sample. This caution addresses the critical issue of **sampling bias**.

In the economic example, if the economist collected their sample of 30 individuals exclusively from residents of a particular city, the resulting regression model accurately reflects the relationship between schooling, hours, and income for individuals *in that city*. It would be statistically

inappropriate to use this exact model to predict the income of an individual living in a vastly different economic or geographical region, where tax structures, cost of living, and wage standards are fundamentally different.

The fitted coefficients are specific to the population used for sampling. If a prediction is needed for a distinct population, a new, representative sample must be collected from that target population, and a new regression model must be fitted and validated. Failure to respect the boundaries of the sampled population compromises the reliability and applicability of the forecast.