

# Understanding the PRESS Statistic: A Guide to Evaluating Predictive Models

Authored by  
**Mohammed Iooti**

November 7, 2025

## RECOMMENDED CITATION

Mohammed Iooti (2025). *Understanding the PRESS Statistic: A Guide to Evaluating Predictive Models*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=12210>

## The Dual Purpose of Regression Analysis

In the field of [statistics](#), the construction and fitting of [regression models](#) serve two primary and distinct objectives. The first objective is often explanatory: seeking to understand and quantify the nature of the relationship between one or more potential causal factors, known as [explanatory variables](#) (or predictors), and the outcome of interest, referred to as the [response variable](#). This goal focuses on inference--determining which variables are significant and the magnitude and direction of their influence.

The second, equally critical purpose of model fitting is prediction. Here, the goal shifts from understanding the underlying mechanics to generating accurate estimates of the response variable for new, unseen data points, based solely on the observed values of the explanatory variables. When prediction is the paramount concern, the suitability of a model is not judged merely by how well it fits the existing data (a concept known as goodness-of-fit), but rather by its capacity to generalize effectively to novel observations. A model that performs exceptionally well during training but poorly on new data is considered overfitted and is ultimately insufficient for reliable forecasting.

To ensure that a selected model is truly the best candidate for future prediction, statisticians must employ metrics that rigorously test a model's external validity. Traditional metrics like the R-squared value primarily assess the fit to the sample data used for calibration and often fail to penalize model complexity adequately, thereby rewarding overfitting. This necessity for a robust measure of predictive accuracy leads us to specialized statistics designed specifically to estimate a model's performance on out-of-sample data, such as the crucial metric known as the PRESS Statistic.

## Understanding the PRESS Statistic

When the objective is strictly to predict the values of a response variable, we require a systematic way to evaluate and compare multiple competing regression models. The most effective model is the one that minimizes prediction error on data points not used during its initial construction. This is precisely the function of the **PRESS Statistic**, which is an acronym for the "**P**redicted **R**esidual **S**um of **S**quares." The PRESS Statistic provides an estimate of the sum of squared prediction errors that would occur if the model were applied to a new dataset.

Conceptually, the calculation of PRESS is fundamentally related to the methodology of cross-validation, specifically the [Leave-One-Out Cross-Validation \(LOOCV\)](#) procedure. In LOOCV, each observation in the dataset is sequentially removed, the model is refitted using the remaining  $n-1$  observations, and then the refitted model is used to predict the value of the removed observation. The difference between the actual observed value and this prediction is the prediction error for that observation. PRESS is simply the sum of the squares of these prediction errors across all

observations. This process ensures that every prediction is made using a model that did not rely on the actual observed data point for its parameter estimation, thus providing a much more honest assessment of the model's true predictive power compared to standard residual sum of squares (SSE).

By systematically evaluating the model's performance in this manner, the PRESS Statistic serves as an invaluable tool for model selection. Unlike metrics that focus on minimizing in-sample error, PRESS prioritizes models that exhibit strong generalization capabilities. Given a choice between several viable regression models, the model yielding the lowest PRESS value is statistically determined to be the one that is likely to make the most accurate predictions when deployed against a truly new, independent dataset. This makes the PRESS Statistic an indispensable diagnostic tool, especially in applications where predictive accuracy, rather than mere explanatory power, is the primary concern.

## The Mathematical Foundation of PRESS

The formal definition of the PRESS Statistic elegantly bypasses the need to iteratively refit the model  $n$  times (which would be computationally intensive for large datasets) by utilizing the properties of the hat matrix. The mathematical expression used to calculate PRESS is derived from the standardized residuals, effectively incorporating the influence of each observation directly into the error calculation. The formula is defined as follows:

$$\text{PRESS} = \sum (e_i / (1-h_{ii}))^2$$

This formula uses standard quantities derived from the initial model fit, allowing for quick calculation. Let us dissect the components of this critical formula to fully appreciate its implications for predictive accuracy.

The core components are defined as:

**$e_i$ :** This represents the  $i$ th [residual](#). The residual is the simple difference between the actual observed response value ( $y_i$ ) and the fitted or predicted value ( $\hat{y}_i$ ) derived from the full model. It quantifies the error of the model on the data used for fitting.

**$h_{ii}$ :** This value is the  $i$ th diagonal element of the [Hat Matrix](#). It is a measure of the influence, often termed "[leverage](#)," of the  $i$ th observation on the overall model fit. Values of  $h_{ii}$  close to 1 indicate that the observation heavily influences its own predicted value, suggesting that the model is fitting that specific point very closely, perhaps at the expense of overall generalizability.

The term  $(1-h_{ii})$  in the denominator is crucial because it accounts for the degree of self-influence an observation has on the model. When  $h_{ii}$  is large (meaning the observation has high leverage), the denominator  $(1-h_{ii})$  becomes small. Dividing the residual

$(e_i)$  by a small number effectively inflates the squared prediction error for that observation. This inflation is a mathematical representation of the prediction error that would occur if that observation were truly left out of the model fitting process. Consequently, the PRESS statistic heavily penalizes models that are overly reliant on high-leverage points, which is a common characteristic of overfitted or unstable models.

## Interpreting and Utilizing the PRESS Value

The primary utility of the PRESS Statistic lies in its straightforward application for comparative model selection. When a researcher has generated several competing [linear regression](#) models--perhaps differing in the number or combination of explanatory variables--the decision criterion is clear: the model associated with the lowest calculated PRESS value is deemed the superior predictor. This selection rule is rooted in the principle that minimizing the predicted residual sum of squares directly minimizes the expected prediction error on new data.

It is important to understand how the PRESS statistic relates to other popular model selection metrics, such as Akaike Information Criterion (AIC) or the Adjusted R-squared. While AIC and Adjusted R-squared provide ways to balance model complexity with goodness-of-fit, they are often derived from the in-sample residual error structure. PRESS, by contrast, is a purely predictive measure. It is a robust, data-driven approach that estimates the error variance when extrapolating the model beyond the training data. A model with a high R-squared but also a high PRESS statistic indicates a strong fit to the existing data but poor generalization potential--a classic sign of overfitting.

Furthermore, the PRESS statistic can be used to derive  $R^2_{\text{prediction}}$ , a related metric that normalizes the PRESS value against the total sum of squares (TSS). This allows analysts to express the predictive accuracy in a percentage format, similar to the standard  $R^2$ . However, even without this normalization, the raw PRESS score provides a sufficient basis for comparison. By favoring models that maintain small prediction errors even when individual data points are excluded, PRESS intrinsically promotes parsimony--the selection of the simplest model that adequately explains the data--which is a cornerstone of robust statistical modeling.

## Practical Example: Calculating PRESS in R

To illustrate the application of the PRESS Statistic in a real-world scenario, consider a small dataset designed to model a response variable,  $y$ , based on three potential [explanatory variables](#):  $x_1$ ,  $x_2$ , and  $x_3$ . We aim to determine which combination of predictors yields the best predictive model using the R statistical environment.

The sample data, consisting of ten observations, is initialized in R as follows:

```
data <- data.frame(x1 = c(2, 3, 3, 4, 4, 6, 8, 9, 9, 9),  
x2 = c(2, 2, 3, 3, 2, 3, 5, 6, 6, 7),  
x3 = c(12, 14, 14, 13, 8, 8, 9, 14, 11, 7),  
y = c(23, 24, 15, 9, 14, 17, 22, 26, 34, 35))
```

Next, we proceed to define and fit three distinct [linear regression](#) models using the standard `lm()` function. Model 1 is a simple regression using only  $x_1$ . Model 2 incorporates two predictors,  $x_1$  and  $x_2$ . Finally, Model 3 uses a different combination,  $x_2$  and  $x_3$ .

```
model1 <- lm(y~x1, data=data)
```

```
model2 <- lm(y~x1+x2, data=data)
```

```
model3 <- lm(y~x2+x3, data=data)
```

## Selecting the Optimal Predictive Model

Once the models are fitted, the PRESS Statistic must be calculated for each one to compare their external predictive performance. While standard statistical packages may not include a built-in function for PRESS, it is straightforward to implement a custom function in R using the mathematical definition, leveraging the model's residuals and the diagonal elements of the hat matrix (leverage values).

The following R code block demonstrates the definition of a custom function named `PRESS` and its subsequent application to the three fitted models:

```
#create custom function to calculate the PRESS statistic
```

```
PRESS <- function(model) {  
i <- residuals(model)/(1 - lm.influence(model)$hat)  
sum(i^2)  
}
```

```
#calculate PRESS for model 1
```

```
PRESS(model1)
```

```
590.2197
```

```
#calculate PRESS for model 2
```

```
PRESS(model2)
```

```
519.6435
```

```
#calculate PRESS for model 3
```

```
PRESS(model3)
```

```
537.7503
```

Upon reviewing the calculated PRESS statistics, a clear winner emerges. Model 1, relying solely on  $x_1$ , exhibits the highest PRESS value at **590.2197**, indicating the poorest predictive generalization among the three candidates. Model 3, incorporating  $x_2$  and  $x_3$ , shows a significant improvement with a PRESS value of **537.7503**. However, Model 2, which utilizes both  $x_1$  and  $x_2$ , yields the lowest calculated PRESS statistic, settling at **519.6435**.

According to the principles governing the use of the PRESS statistic, the model that minimizes this value is the one best suited for future prediction. Therefore, based on this rigorous assessment of out-of-sample error, we would confidently select Model 2 ( $Y \sim X_1 + X_2$ ) as the preferred predictive model for this dataset. This decision is robust because the PRESS statistic has specifically penalized any overfitting tendencies and confirmed Model 2's superior ability to generalize its relationship to new data points, solidifying its reliability for forecasting applications.

## Additional Resources

[Introduction to Simple Linear Regression](#)

[What is a Parsimonious Model?](#)

[What is a Good R-squared Value?](#)