

Calculate AIC in SAS (With Example)

Authored by
Mohammed looti

November 15, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Calculate AIC in SAS (With Example)*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=1941>

The Crucial Role of Model Selection and the Akaike Information Criterion

In the expansive field of statistical analysis, especially when working with [regression models](#), one of the most intellectually demanding tasks is selecting the optimal model. Analysts frequently develop several competing models, each incorporating a different set of [predictor variables](#), all aiming to explain the same outcome or dependent variable. The core challenge lies in objectively comparing these candidates to identify the one that achieves the ideal balance: maximizing explanatory power while maintaining simplicity.

To navigate this complexity, researchers rely on robust model selection criteria. Among these, the [Akaike Information Criterion \(AIC\)](#) stands out as a globally respected and widely applied metric. Conceived by Hirotugu Akaike in 1971, AIC provides a mathematically rigorous framework for evaluating models by simultaneously weighing two fundamental, often conflicting, characteristics: their [goodness of fit](#) to the observed data and their inherent [model complexity](#).

The philosophical foundation of AIC is rooted in the principle of parsimony. It seeks to identify the model that is most likely to represent the true data-generating process while actively penalizing the inclusion of extraneous [model parameters](#). An overly complex model, though it might achieve a perfect fit on the training data, often suffers from overfitting, rendering it useless for generalization to new, unseen data. AIC thus acts as a vital compass, guiding the selection process toward models that are both statistically accurate and practically interpretable.

Deconstructing the Akaike Information Criterion (AIC) Formula

The [Akaike Information Criterion](#) is fundamentally an estimator of the relative amount of information lost when a particular model is used to approximate the reality of the data generation process. It allows for a quantitative comparison of different statistical models derived from the same dataset. A model that performs well while remaining relatively simple is preferred. Its calculation integrates the model's maximized likelihood with a term that accounts for its complexity.

The standard formula used for calculating AIC is elegantly simple yet powerful:

$$AIC = 2K - 2\ln(L)$$

Let's meticulously define the components of this critical equation:

K (Number of Parameters): This term represents the number of [model parameters](#) that are estimated. In the context of standard linear [regression models](#), K typically includes the intercept term, the number of [predictor variables](#) (or coefficients), and often, the estimate of the error variance. For instance, if we fit a model with two predictors and an intercept, K would be calculated as 2 (predictors) + 1 (intercept) + 1 (error variance) = 4.

$\ln(L)$ (Log-Likelihood): This refers to the natural logarithm of the model's maximum [log-likelihood](#) function. The likelihood function itself is a measure of how well the model aligns with the observed data--a proxy for the [goodness of fit](#). A higher log-likelihood suggests a superior fit to the data. Crucially, modern [SAS](#) and other sophisticated statistical packages automatically calculate this value during the model fitting procedure, making the computation of AIC highly streamlined.

The structure of the AIC formula reveals its purpose: the $-2\ln(L)$ term decreases as the model fits the data better (L increases), thus rewarding [goodness of fit](#). Conversely, the $2K$ term increases as more parameters are added, imposing a penalty for increased complexity. When comparing multiple models, the model exhibiting the **lowest AIC value** is unequivocally the preferred choice, as it successfully minimizes the informational loss relative to the true process while avoiding unnecessary parameterization.

Setting Up the Practical Example: Calculating AIC in SAS

To transition from theory to practical application, we will demonstrate how to calculate and utilize the [AIC](#) within the [SAS](#) statistical environment. Consider a common scenario in quantitative research: predicting academic performance. Our objective is to predict students' final exam scores based on two potential factors, and we must determine which combination of these factors yields the most efficient prediction model.

For this illustration, we will compare three distinct [regression models](#), each built upon a different configuration of [predictor variables](#). By fitting these models in [SAS](#) and requesting the respective AIC values, we can establish an objective basis for model selection.

The three candidate models are designed as follows:

Model 1 (Simplicity Focus): This model uses "hours spent studying" as the singular [predictor variable](#). It tests the hypothesis that study time alone is the most dominant factor in determining scores.

Model 2 (Alternative Factor): This model relies solely on the "number of practice exams taken" as its [predictor variable](#). This assesses whether active preparation through practice is a better standalone predictor.

Model 3 (Combined Complexity): This model incorporates both "hours spent studying" and "number of practice exams taken" as [predictor variables](#). It investigates the cumulative influence of both factors, which might offer a more exhaustive explanation at the cost of increased complexity.

Step-by-Step SAS Implementation and Code Execution

The first prerequisite for any statistical analysis in [SAS](#) is the preparation of the raw data. We begin by creating a hypothetical dataset containing information for 20 students, defining variables for the hours studied (`hours`), the number of practice exams taken (`prep\_exams`), and the resulting final exam score (`score`).

The following [DATA step](#) code constructs our sample dataset, named `exam\_data`:

```
/*create dataset*/  
data exam_data;  
input hours prep_exams score;  
datalines;  
1 1 76  
2 3 78  
2 3 85  
4 5 88  
2 2 72  
1 2 69  
5 1 94  
4 1 94  
2 0 88  
4 3 92  
4 4 90  
3 3 75  
6 2 96  
5 4 90  
3 4 82  
4 4 85  
6 5 99  
2 1 83  
1 0 62  
2 1 76  
;  
run;
```

Once the dataset is ready, we proceed to fit our candidate models using [PROC REG](#), the standard procedure for performing linear [regression models](#) in [SAS](#). Crucially, to instruct SAS to evaluate and present model selection criteria for all possible subsets of the specified predictors, we include the `selection=adjrsq sse aic` option within the `MODEL` statement. This command ensures the

calculation and display of the [AIC](#), alongside other measures like [Adjusted R-squared \(ADJRSQ\)](#) and [Sum of Squared Errors \(SSE\)](#).

/*fit multiple linear regression models and calculate AIC for each model*/

proc reg data=exam_data;

model score = hours prep_exams / selection=adjrsq sse aic;

run;

The resulting output generated by [PROC REG](#) provides a comprehensive summary table detailing the performance metrics for the three candidate models. The image below presents the relevant section of the SAS output, specifically focusing on the calculated AIC values for each model configuration:

The REG Procedure					
Model: MODEL1					
Dependent Variable: score					
Adjusted R-Square Selection Method					
Number of Observations Read				20	
Number of Observations Used				20	

Number in Model	Adjusted R-Square	R-Square	AIC	SSE	Variables in Model
1	0.7121	0.7273	68.4537	501.90937	hours
2	0.7027	0.7340	69.9507	489.44312	hours prep_exams
1	0.0888	0.1368	91.4967	1588.53461	prep_exams

From this critical output table, we can extract the specific [AIC](#) values for our three comparison models:

AIC for Model 1 (using only "hours spent studying"): **68.4537**

AIC for Model 3 (using both "hours spent studying" and "practice exams taken"): **69.9507**

AIC for Model 2 (using only "number of practice exams taken"): **91.4967**

Interpretation of Results and Final Model Selection

The interpretation of the [SAS](#) output is the most critical stage of the model selection process. The

established rule for the [AIC](#) dictates that the model with the lowest value is the most favorable, as it represents the most judicious trade-off between minimizing error and reducing complexity. Analyzing our calculated AIC figures--68.4537 (Model 1), 69.9507 (Model 3), and 91.4967 (Model 2)--the choice becomes evident.

The model that incorporates only "hours spent studying" (Model 1) yields the lowest AIC of **68.4537**. This decisive result signifies that, relative to the other two options, Model 1 is the most efficient and robust statistical representation of the data. Although Model 3, which included two [predictor variables](#), showed a slightly better fit (a lower SSE, as seen in the full output), the penalty imposed by the AIC criterion for the extra complexity in Model 3 ($K=4$ vs $K=3$) was sufficient to elevate its AIC score above Model 1. This suggests that the small predictive gain from including "practice exams taken" did not justify the additional parameter.

Consequently, we formally select the model based solely on "hours spent studying." This decision implies that the relationship between the outcome and the predictors is most adequately captured by the simplest model, represented by the linear equation: $\text{Score} = \beta_0 + \beta_1(\text{Hours Studied})$. Once the optimal model is identified using AIC, subsequent analysis involves a deep dive into the properties of this selected model. Key steps include assessing the [R-squared](#) value to understand the variance explained, and meticulously examining the [beta coefficients](#), which quantify the magnitude and direction of the relationship between the chosen [predictor variable](#) and the final exam score.

Limitations and Comparative Considerations of AIC

While the [AIC](#) is an exceptionally valuable diagnostic tool, its effective use requires an understanding of its inherent limitations. Foremost among these is the fact that AIC is strictly a **relative measure**. It provides an ordinal ranking, telling us which model performs best among the specific set of candidates we provided; it does not offer any insight into whether the models themselves are statistically adequate or whether the chosen "best" model accurately reflects the true underlying mechanism of the population.

Furthermore, the integrity of AIC comparisons rests entirely on the consistency of the data used. All candidate [regression models](#) must be fitted to the exact same dataset, including the same number of observations and the same data transformations. Diverting from this consistency invalidates direct comparison of the AIC values. Researchers must strictly adhere to identical data subsets when evaluating models.

It is also important to recognize AIC as one member of a family of information criteria. Its close cousin, the [Bayesian Information Criterion \(BIC\)](#), serves a similar purpose but enforces a significantly heavier penalty for the number of [model parameters](#), especially as the sample size increases. This difference often results in BIC selecting a simpler, more parsimonious model than

AIC. The choice between AIC (which prioritizes asymptotic efficiency) and BIC (which prioritizes consistency) depends heavily on the specific modeling goals and the researcher's philosophy regarding model complexity.

Conclusion: The Efficiency of AIC in Statistical Modeling

The [Akaike Information Criterion](#) remains an indispensable cornerstone of the model selection process in statistics. It empowers researchers and analysts to move beyond subjective judgment, enabling the objective comparison of competing [regression models](#). By meticulously quantifying the trade-off between explanatory power and intrinsic complexity, AIC guides the selection toward models that are optimally configured for generalization and robustness.

Our comprehensive example utilizing the [SAS](#) programming language effectively demonstrated the systematic application of this criterion, ranging from initial data coding to the final interpretation of comparative AIC values. This methodology underscores the necessity of a data-driven and principled approach to model building, ensuring that the chosen analytical structure is maximally efficient.

To further refine your analytical skills, continued exploration of [SAS](#) programming and the broader landscape of model selection methodologies is highly recommended. Consulting official documentation and specialized tutorials will provide deeper insights into advanced statistical techniques and the functionality of various [statistical software](#), ultimately enhancing your capacity to perform rigorous and defensible quantitative analysis.