

How to Calculate AIC in SAS: A Simple Guide

Authored by
stats writer

November 19, 2025

RECOMMENDED CITATION

stats writer (2025). *How to Calculate AIC in SAS: A Simple Guide*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=97848>

The Akaike Information Criterion (AIC) is a fundamental concept in modern statistics, serving as a powerful tool for assessing the relative quality of multiple competing statistical models for a given set of data. Developed by statistician Hirotugu Akaike, AIC estimates the relative amount of information lost when a particular model is used to represent the complex process that generated the data. Essentially, it helps researchers navigate the trade-off between model complexity and goodness of fit, ensuring that the selected model is parsimonious yet highly explanatory. In the realm of SAS (Statistical Analysis System), the calculation and application of AIC are streamlined through specialized procedures like PROC REG and PROC MODEL, which automatically generate this metric alongside other critical model fit statistics.

The primary goal of using AIC is to select the model that provides the best balance. A model that perfectly fits the training data might be overly complex--a phenomenon known as overfitting--leading to poor performance on new, unseen data. Conversely, a model that is too simplistic might underfit the data, failing to capture essential relationships. AIC addresses this inherent conflict by introducing a penalty term for increasing the number of estimated parameters. This structural mechanism ensures that models with more predictors must demonstrate a substantially better fit to justify their added complexity, making AIC an indispensable resource for model validation and selection across various statistical disciplines.

While many statistical packages offer automated AIC calculations, understanding the underlying mechanism is crucial for accurate interpretation. When using SAS, researchers often rely on procedures designed for generalized linear models or regression analysis. For instance, the general PROC MODEL procedure is robust, capable of fitting models to data and outputting the AIC value. Furthermore, the **OUTMODEL** option within certain procedures allows users to save these fit statistics into a dataset. This capability is particularly useful for sophisticated analyses where AIC needs to be compared across dozens or hundreds of different model specifications, facilitating large-scale comparative model selection efficiently.

The Mathematical Foundation of AIC

The Akaike Information Criterion is derived from information theory and is formally defined by a concise mathematical equation that incorporates both the model's fit and its complexity. Understanding this formula is key to appreciating how AIC operates as a tool for statistical inference. The mathematical expression for AIC ensures that the criterion adequately balances the error minimization against the principle of parsimony, rewarding a tight fit but simultaneously discouraging the inclusion of superfluous variables.

The standard calculation for AIC is given by the formula:

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

This formulation comprises two distinct parts. The term $-2\ln(L)$ represents the measure of badness of fit, where L is the maximum value of the likelihood function for the estimated model. Minimizing this term means maximizing the likelihood that the observed data was generated by the specified model. The second term, $2K$, is the penalty imposed for model complexity. As the number of parameters (K) increases, the penalty increases linearly, thereby raising the overall AIC score and discouraging overly complicated structures.

The variables within the AIC formula are defined as follows:

K: This variable denotes the total number of estimated parameters in the statistical model. When calculating K , it is crucial to include all terms being estimated, including the intercept (β_0) and the error variance (σ^2) if they are part of the estimation process. For simple linear regression models, the default number of estimated parameters often includes the intercept and the variance, meaning even a model with just one predictor variable will typically have $K = 3$ (Intercept + Predictor Coefficient + Error Variance estimate). This penalty ensures that complexity is consistently accounted for across different model types.

$\ln(L)$: This is the natural logarithm of the maximum log-likelihood function value achieved by the estimated model. The likelihood function quantifies how likely the observed data is, given the estimated parameters of the model. Statistical software, including SAS, automatically calculates and maximizes this value during the model fitting process, making the raw AIC calculation straightforward once the model is executed. A higher log-likelihood (and thus a lower $-2\ln(L)$ term) signifies a better fit to the observed data.

Interpreting AIC for Model Selection

The utility of AIC lies entirely in its comparative nature. AIC is an absolute measure of quality only in reference to other candidate models fitted to the exact same dataset. It is not an absolute measure of goodness of fit, unlike R-squared, but rather an index of information loss relative to other specifications. Consequently, researchers must fit several competitive regression models and then compare their respective AIC values to determine the most suitable structure. This comparative framework ensures that the selected model is not just 'good,' but optimal among the tested candidates.

The core principle for interpretation is straightforward: **the model with the lowest AIC value is considered the best fit among the set of candidate models**. A lower AIC indicates a model that minimizes the information loss, achieving an optimal balance between maximizing the log-likelihood (goodness of fit) and minimizing the number of parameters (parsimony). When comparing two models, Model A and Model B, if $AIC(A)$ is significantly lower than $AIC(B)$, Model A is statistically preferred. However, it is also important to consider the magnitude of the difference. If the AIC values are very close, the models might be considered statistically equivalent, and simpler

models may still be preferred based on ease of interpretation or domain knowledge.

The interpretation often extends beyond simply identifying the minimum value. Some statisticians utilize AIC differences (Δ_i) where $\Delta_i = \text{AIC}_i - \text{AIC}_{\min}$. Models with $\Delta_i < 2$ are generally considered to have substantial support and are virtually indistinguishable from the best model. Models with $4 < \Delta_i \leq 10$ essentially have no empirical support and should be disregarded. This nuanced approach prevents analysts from arbitrarily choosing a model based on a marginal difference of AIC, encouraging a more robust selection process that acknowledges uncertainty.

Calculating AIC in SAS: The Role of Procedures

The SAS Statistical Analysis System provides several powerful procedures that automatically calculate the Akaike Information Criterion as part of their standard output, particularly those used for model fitting and estimation. While procedures like PROC GLM, PROC LOGISTIC, and PROC MIXED all generate AIC for their respective model types, one of the most common and versatile procedures for calculating AIC in the context of standard linear regression models is **PROC REG**. PROC REG allows for the fitting of ordinary least squares models and offers powerful options for model selection, including the calculation of AIC.

When executing a model fitting procedure in SAS, the software automatically performs the necessary steps: maximizing the log-likelihood function (L) based on the specified model structure and data, determining the number of estimated parameters (K), and applying the AIC formula. For complex models or generalized linear models, the procedure **PROC GENMOD** is often used, which also routinely outputs AIC, along with related criteria such as BIC (Bayesian Information Criterion) and AICC (Corrected AIC), providing a comprehensive set of metrics for assessment.

To explicitly request AIC calculation within a regression context, specific model selection options must often be included in the model statement. In **PROC REG**, parameters like **SELECTION=AIC** instruct SAS to perform an analysis of models based on this criterion. Furthermore, many procedures offer the capability to save model fit statistics using options like **OUTSTAT** (in PROC REG) or **OUTMODEL** (in other procedures), enabling subsequent analysis or visualization of AIC values across different specifications in a clean, tabular format. This ability to automate the comparison across multiple specifications is a key strength of the SAS environment.

Practical Example: Setting Up the Data in SAS

To demonstrate the calculation and comparison of AIC in SAS, let us consider a practical scenario where we aim to predict student exam scores based on different potential predictor variables. Suppose we want to compare three competing regression models designed to predict a student's final exam score in a course. The models will use varying combinations of predictors hypothesized

to influence the outcome.

The three candidate models we wish to compare are structured as follows:

Model 1: Predicts score using only the variable for hours spent studying (a measure of effort).

Model 2: Predicts score using only the variable for the number of practice exams taken (a measure of preparation).

Model 3: Predicts score using both hours spent studying and practice exams taken (a combined measure).

Before fitting these models, we must create a dataset in SAS containing the necessary variables for 20 students. We use the **DATA step** and **DATALINES** statement to input the variables: hours (hours spent studying), prep_exams (number of practice exams taken), and score (the final exam score). This structured input process is fundamental to all subsequent statistical analysis within the SAS environment.

The following code block executes the data creation process, establishing the exam_data dataset:

```
/*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;
```

Fitting Multiple Regression Models Using PROC REG

Once the data is successfully loaded into the `exam_data` dataset, we employ the powerful **PROC REG** procedure. While PROC REG is primarily designed for fitting standard least squares regression models, it also possesses advanced capabilities for comparing competing models through various selection methods. To simultaneously fit and compare all subsets of our defined predictor variables (hours and prep_exams) and automatically calculate the AIC for each resulting model, we utilize the SELECTION option within the MODEL statement.

The syntax for **PROC REG** is designed to be highly flexible. By specifying `selection=adjrsq sse aic`, we instruct SAS not only to calculate the Adjusted R-squared (adjrsq) and Sum of Squared Errors (sse) but critically, to calculate the AIC for every possible combination of predictors derived from the set specified in the model statement. Since we listed hours and prep_exams, SAS will evaluate three models: `score = hours`, `score = prep_exams`, and `score = hours + prep_exams`. This process provides a robust and automated way to screen potential model candidates without running three separate PROC REG commands.

The following SAS code executes the model fitting and comparison procedure:

```
/*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 output generated by this procedure will contain detailed summaries for the models, including parameter estimates, ANOVA tables, and, most importantly for our selection purposes, a table summarizing the fit statistics for all evaluated subsets. This table clearly presents the AIC value associated with each unique model specification, allowing for immediate comparison based on the fundamental criterion that a lower AIC signifies a superior, more parsimonious fit to the data.

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

Interpreting the SAS Output for Model Comparison

Upon execution of the **PROC REG** procedure with the **SELECTION=AIC** option, SAS generates a comprehensive output table listing the results for all possible subsets of predictor variables. This table is the critical juncture for model selection, as it clearly presents the calculated AIC value for each specification. The analyst's task is then to systematically review these values and identify the model exhibiting the minimum AIC score, as per the rules of information criterion analysis.

Based on the analysis performed on the `exam_data` dataset, the AIC values for the three candidate models are summarized below:

AIC for Model 1 (Predictor: hours spent studying): **68.4537**

AIC for Model 3 (Predictors: hours and practice exams): **69.9507**

AIC for Model 2 (Predictor: practice exams taken): **91.4967**

A direct comparison of these values immediately reveals that Model 1, which incorporates only 'hours spent studying' as the predictor, yields the lowest AIC value (68.4537). This result signifies that, relative to the other two tested models, the single-predictor Model 1 minimizes the estimated information loss. It achieves the best trade-off between maximizing the log-likelihood fit and maintaining model simplicity, effectively penalizing the slight improvement in fit (if any) offered by the more complex Model 3.

Conversely, Model 2, utilizing only 'practice exams taken,' resulted in a significantly higher AIC (91.4967), indicating that it is a poor fit for the data compared to the other specifications. Even

Model 3, which is the most complex (having two predictors), shows an AIC (69.9507) that is higher than the simplest Model 1. While the difference between Model 1 and Model 3 is small ($\Delta AIC \approx 1.5$), favoring Model 1 adheres to the principle of parsimony and the AIC selection rule. Thus, we conclude that the model that best fits the data, according to the Akaike Information Criterion, is the simple linear regression model predicting score based solely on hours studied, formally expressed as: $\text{Score} = \beta_0 + \beta_1 (\text{Hours Studied})$.

Advanced Considerations and Best Practices

While the Akaike Information Criterion provides a robust, objective measure for statistical model selection, its proper application requires awareness of several advanced considerations and best practices, particularly within the SAS environment. One key consideration is the use of AICc, the corrected AIC, which is particularly relevant when dealing with small sample sizes relative to the number of parameters (specifically when $n/K < 40$). In such cases, the standard AIC tends to favor overly complex models, and AICc provides a more reliable estimate. Many SAS procedures, such as **PROC GENMOD**, automatically calculate AICc alongside AIC, offering analysts a more conservative selection tool when data is limited.

Another important aspect is the comparison of non-nested models. AIC is highly effective for comparing models that are not hierarchically related (i.e., where one model is not simply a subset of the other), which is a weakness of traditional hypothesis testing methods like F-tests. Since AIC comparison relies only on the maximum log-likelihood and the number of parameters, it provides a universal metric applicable across vastly different model structures, provided they are fitted to the identical response variable and dataset. This flexibility makes AIC indispensable in fields like ecology and econometrics where model forms can vary dramatically.

Finally, analysts should always supplement AIC-based selection with diagnostic checks and domain knowledge. Although AIC identifies the statistically "best" model, a chosen model must still be interpretable and satisfy underlying assumptions (such as linearity, normality of residuals, and homoscedasticity for regression models). After using the AIC to identify Model 1 ($\text{Score} = \beta_0 + \beta_1 (\text{Hours Studied})$) as superior in our example, the next logical step would be to run a final, dedicated **PROC REG** on Model 1 to thoroughly analyze its residual plots, R-squared value, and the precise coefficient estimates (β_1). This integrated approach ensures that the model chosen is not only statistically optimal but also scientifically sound and reliable for prediction.