How to Easily Perform Stepwise Regression in SAS

Authored by

stats writer

November 20, 2025

Stepwise regression within the SAS environment is a powerful and efficient method utilized for systematic variable selection in statistical modeling. This iterative technique employs a combination of forward and backward procedures to meticulously sift through a pool of potential predictor variables, ultimately identifying the optimal subset that provides the most robust and parsimonious explanation for the variation observed in the designated response variable. Unlike traditional methods that might require manual testing of numerous combinations, stepwise regression automates this process by evaluating models based on predefined statistical metrics.

Understanding the Stepwise Regression Procedure

The core objective of Stepwise regression is to construct a high-quality regression model by strategically entering and removing predictor variables in a sequential, iterative fashion. This automated process continues until statistical criteria are satisfied, indicating that no further variables can be added or removed without compromising the model's overall fit or efficiency. This careful balancing act ensures that the final model is both statistically sound and practically interpretable, avoiding issues like overfitting that can arise when too many predictors are included indiscriminately.

The procedure is designed to converge on a model that incorporates all predictor variables exhibiting a statistically significant relationship with the outcome. The primary difference between different forms of stepwise selection (forward, backward, or actual stepwise) lies in how variables are introduced and eliminated. The traditional stepwise approach cycles between adding the most significant variable (forward step) and checking if any previously included variable has become non-significant due to the presence of the new variable (backward step). This continuous refinement is why the method is so effective for exploratory data analysis.

In the SAS statistical software environment, executing stepwise regression is straightforward. Analysts leverage the powerful PROC REG procedure, combined with a specific **SELECTION** statement, to define the method of variable entry/removal and the metrics used for evaluation. This allows users to quickly implement complex model selection logic using concise SAS code, focusing on the interpretation of results rather than the mechanics of variable testing.

Implementing Stepwise Regression in SAS: An Example

To illustrate the practical application of stepwise regression, we will define a sample dataset containing several candidate predictor variables and a single response outcome. The objective here is to identify which combination of these predictors offers the optimal fit for explaining the variation in the response. This example demonstrates the necessary data preparation steps and the subsequent execution of the modeling procedure using SAS.

Assume we have collected data represented by four independent variables (x1, x2, x3, x4) and one dependent variable (y). This synthetic dataset, while small, provides a clear environment to demonstrate how the stepwise process identifies the contributing variables efficiently. Successful analysis requires that the data be loaded into a SAS dataset structure before any procedural modeling can commence.

The following code block demonstrates how to structure and input this data into a SAS dataset named `my_data`, followed by the verification step using `PROC PRINT` to ensure data integrity and visualization of the input variables. This foundational step is critical for any subsequent statistical analysis.

```
/*create dataset: defining predictor and response variables*/
data my_data;
input x1 x2 x3 x4 y;
datalines;
1 4 10 13 78
2 4 12 14 81
5 3 7 10 75
8 2 13 9 97
10 5 12 5 95
14 7 8 6 90
17 8 10 6 86
19 5 15 5 90
20 5 12 4 93
21 4 10 3 95
;
run;

/*view dataset to confirm successful loading*/
proc print data=my_data;
```

Obs	x1	x2	x3	x4	y
1	1	4	10	13	78
2	2	4	12	14	81
3	5	3	7	10	75
4	8	2	13	9	97
5	10	5	12	5	95
6	14	7	8	6	90
7	17	8	10	6	86
8	19	5	15	5	90
9	20	5	12	4	93
10	21	4	10	3	95

Defining Metrics for Optimal Model Selection

After the dataset is prepared, the next phase involves executing the regression and determining which specific combination of predictors yields the "best" model. When discussing the quality of a statistical model, "best" is quantified by metrics that assess both goodness-of-fit and model complexity. The goal is to maximize explanatory power while penalizing unnecessary parameters, ensuring the model generalizes well beyond the training data.

Two specific metrics are commonly employed in stepwise regression--and indeed across many model comparison tasks--to evaluate potential models generated during the selection process. These metrics provide quantitative standards by which competing models (e.g., a model using x1, x2, and x3 versus one using only x3 and x4) can be rigorously assessed and ranked.

SAS facilitates the automatic calculation of these metrics for all possible combinations, or for the combinations tested during the stepwise process, allowing the analyst to quickly identify the statistically preferred structure. Understanding the implications of each metric is crucial for making an informed final selection.

There are two highly important metrics we use to assess the comparative performance of regression models:

1. Adjusted R-squared: This metric provides an estimate of the proportion of variance in the response variable that is explained by the predictors, adjusted for the number of predictors included in the model. Crucially, the adjustment prevents the R-squared value from artificially inflating simply by adding more variables. The goal is always to find the model exhibiting the

highest adjusted R-squared value.

2. AIC: The Akaike Information Criterion (AIC) serves as a comparative measure that estimates the relative quality of statistical models for a given set of data. AIC balances the tradeoff between the complexity of the model (number of parameters) and how well the model fits the data. When comparing models, the model with the **lowest AIC value** is generally preferred, as it represents a better balance of fit and parsimony.

Executing the Stepwise Multiple Linear Regression in SAS

The power of PROC REG allows for the simultaneous calculation and comparison of models based on these vital metrics. By incorporating the `SELECTION` statement within `PROC REG`, we instruct SAS to perform a comprehensive search across the model space, evaluating each potential combination. Although the standard stepwise procedure typically relies on p-values (significance levels) for entry and removal, specifying metrics like ADJRSQ and AIC forces SAS to evaluate models based on overall fit statistics rather than just individual variable significance.

In this specific instance, we are using the `selection=adjrsq aic` option, which compels SAS to identify the model that maximizes Adjusted R-squared and minimizes AIC. This combined approach is highly robust for model selection. Furthermore, the `OUTEST` and `OUTPUT` statements are included to store the estimated parameters and prediction results for further post-modeling analysis, demonstrating best practices in statistical programming.

The following SAS code block executes the stepwise multiple linear regression, directing the procedure to compare all possible subsets of predictor variables (x1, x2, x3, x4) against the response variable (y), selecting the best based on our defined criteria.

```
/*perform stepwise multiple linear regression using ADJRSQ and AIC criteria*/
proc reg data=my_data outest=est;
model y=x1 x2 x3 x4 / selection=adjrsq aic ;
output out=out p=p r=r;
run;
quit;
```

The REG Procedure
Model: MODEL1
Dependent Variable: y

Adjusted R-Square Selection Method

Number of Observations Read	10
Number of Observations Used	10

Number in Model	Adjusted R-Square	R-Square	AIC	Variables in Model
2	0.5923	0.6829	34.2921	x3 x4
3	0.5854	0.7236	34.9191	x1 x3 x4
3	0.5648	0.7098	35.4051	x2 x3 x4
4	0.5205	0.7336	36.5509	x1 x2 x3 x4
2	0.4727	0.5899	36.8655	x2 x4
1	0.4639	0.5235	36.3653	x4
2	0.4081	0.5396	38.0206	x1 x3
2	0.4013	0.5344	38.1345	x1 x4
3	0.3867	0.5911	38.8348	x1 x2 x4
3	0.3503	0.5669	39.4109	x1 x2 x3
1	0.3285	0.4031	38.6186	x1
2	0.3271	0.4766	39.3031	x1 x2
1	0.1533	0.2474	40.9361	x3
2	0.0583	0.2675	42.6646	x2 x3
1	-.1213	0.0033	43.7454	x2

Interpreting the Stepwise Regression Output

The output generated by PROC REG, particularly when using specific selection criteria like ADJRSQ and AIC, provides a detailed summary table listing various subsets of variables and their corresponding statistical scores. Careful examination of this table is required to pinpoint the optimal model structure. Analysts must scan the column designated for Adjusted R-squared and identify the highest value, simultaneously checking the AIC column for the lowest corresponding value.

Based on the results displayed in the generated output image, we observe a clear consensus across both metrics for a specific subset of predictor variables. The combination that includes only **x3** and **x4** achieves the maximum adjusted R-squared and the minimum AIC value compared to all other potential models, including those using all four predictors or smaller subsets. This strong

agreement simplifies the selection process considerably.

Consequently, the model deemed "best" or most efficient out of all possible linear combinations is the one derived solely from predictors x3 and x4. This outcome suggests that the variables x1 and x2 do not contribute significantly enough to the predictive power of the model to justify their inclusion, particularly once the influence of x3 and x4 is accounted for.

Selecting the Optimal Regression Model Structure

The identified optimal structure leads directly to the formulation of the final mathematical model. This model expresses the response variable (y) as a linear function of the selected predictors (x3 and x4), plus an intercept term (b0). The resulting linear equation formally captures the statistical relationship uncovered by the stepwise procedure.

Thus, we formally declare the following linear model to be the most appropriate choice among the tested possibilities:

y = b0 + b1(x3) + b2(x4)

This parsimonious regression model is characterized by the following key statistical performance metrics derived directly from the SAS output:

Adjusted R-squared value: **0.5923**. This indicates that approximately 59.23% of the variance in the response variable y is explained by the combination of x3 and x4.
AIC: **34.2921**. This relatively low value suggests a high-quality, balanced model when compared to competing structures.

Practical Notes on Model Selection and Parsimony

It is crucial to note that while Adjusted R-squared and AIC often point to the same optimal model, this concordance is not guaranteed. In complex real-world datasets, the model maximizing Adjusted R-squared might not perfectly align with the model minimizing AIC. These metrics should therefore serve as powerful statistical guidelines, but they are rarely the sole determinants.

When divergence occurs between these metrics, selecting the superior regression model requires more than just statistical tabulation. It necessitates the application of domain expertise. Analysts must consider practical factors such as the cost of measuring certain predictors, the interpretability of the model, and the underlying theoretical knowledge relevant to the subject matter. Sometimes, a slightly statistically inferior model is chosen because it offers greater practical relevance or ease of explanation to stakeholders.

A best practice in model building is the pursuit of a **parsimonious model**. This concept dictates

choosing the simplest possible model--the one utilizing the fewest predictor variables--that still achieves a satisfactory and desired level of goodness of fit. This principle is deeply rooted in the philosophy of Occam's Razor, often referred to as the "Principle of Parsimony," which posits that, all else being equal, the simplest explanation is usually the correct one.

Applied rigorously to statistics, the philosophy of parsimony favors a model with fewer parameters (predictors) that achieves a robust level of fit over a much more complex model that only marginally improves the goodness of fit. This preference helps guard against overfitting, leading to models that are more stable, generalizable, and easier to deploy in operational environments.

Further SAS Tutorials and Resources

For those looking to expand their skills beyond stepwise regression, SAS offers a vast suite of procedures for various statistical tasks. The following links provide access to tutorials detailing how to execute other common analytical tasks within the SAS programming environment.

The following tutorials explain how to perform other common tasks in SAS: