How do you perform stepwise regression in SAS, and can you provide an example?

Authored by

stats writer

June 25, 2024

Stepwise regression is a statistical technique used to select the most significant independent variables to include in a regression model, while also removing any redundant or insignificant variables. In SAS, this can be done using the "stepwise" option in the PROC REG procedure. This option automatically performs forward, backward, or bidirectional selection methods to determine the best model fit.

An example of performing stepwise regression in SAS would involve using the "stepwise" option in the PROC REG procedure, followed by specifying the dependent variable and a list of potential independent variables. SAS will then automatically perform the stepwise selection process and provide output displaying the selected variables and their corresponding coefficients, as well as other important statistical information such as the R-squared value and p-values. This procedure can help researchers and analysts to efficiently build regression models that accurately predict the relationship between variables.

Perform Stepwise Regression in SAS (With Example)

Stepwise regression is a procedure we can use to build a regression model from a set of predictor variables by entering and removing predictors in a stepwise manner into the model until there is no statistically valid reason to enter or remove any more.

The goal of stepwise regression is to build a regression model that includes all of the predictor variables that are statistically significantly related to the response variable.

To perform stepwise regression in SAS, you can use PROC REG with the SELECTION statement.

The following example shows how to perform stepwise regression in SAS in practice.

Example: Perform Stepwise Regression in SAS

Suppose we have the following dataset in SAS that contains four predictor variables (x1, x2, x3, x4) and one response variable (y):

```
/*create dataset*/
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*/

proc printdata=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

Now suppose that we would like to find which combination of predictor variables will produce the best .

When we say "best" regression model, we mean the model that maximizes or minimizes some metric.

There are two metrics we commonly use to assess which regression model is best among a group of potential models:

1. Adjusted R-squared: The tells us how useful a model

is, adjusted for the number of predictors in a model. The model with the highest adjusted R-squared value is considered the best.

2. AIC: The (AIC) is a metric that is used to compare the fit of different regression models. The model with the lowest AIC value is considered the best.

Fortunately, we can calculate both the adjusted R-squared and AIC values for regression models in SAS by using PROC REG with the SELECTION statement.

The following code shows how to do so:

```
/*perform stepwise multiple linear regression*/
proc regdata=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

From the output we can see that the value with the highest adjusted R-squared value *and* the lowest AIC value is the regression model that uses only x3 and x4 as the predictor variables.

Thus, we would declare the following model to be "best" out of all possible models:

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

This particular regression model has the following metrics:

Adjusted R-squared value: 0.5923AIC: 34.2921

Notes on Selecting the "Best" Regression Model

Note that sometimes the model with the highest adjusted R-squared value does not always have the lowest AIC value as well.

When it comes to deciding which regression model is best, adjusted R-squared and AIC serve as suggestions but in the real world you may have to use domain expertise to determine which model is best.

It can also be a good idea to choose a , which is a model that achieves a desired level of goodness of fit using as few predictor variables as possible.

The reasoning for this type of model stems from the idea of Occam's Razor (sometimes called the "Principle of Parsimony") which says that the simplest explanation is most likely the right one.

Applied to statistics, a model that has few parameters but achieves a satisfactory level of goodness of fit should be preferred over a model that has a ton of parameters and achieves only a slightly higher level of goodness of fit.

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