Stepwise Regression

You would like to able to predict what sales will be. You have a database that contains 40 different variables that might impact sales. How can you go through these 40 variables to see which ones really impact sales and could become part of a model to predict sales? Well, you could run a full regression analysis with all 40 variables and see which ones are "significant." But regression models change as variables are added or removed.

This is where stepwise regression can help. This is an automated process that builds a regression model for you by going through a series of steps of adding the most significant variable or removing the least significant variable.

This month's publication takes a look at stepwise regression. On the surface, this technique sounds great. Sit back and let the model be built for you. As with all techniques, there are some caveats about using stepwise regression. So, we will take a look at how stepwise regression can easily build a model for you as well as a few of the drawbacks of stepwise regression.

In this issue:

Regression Review

In regression, we are trying to build a model to predict Y based on certain predictor variables (x_1, x_2, etc.). For example, if we have two predictor variables, we would be building a regression equation of the form:

$$Y = b_o + b_1x_1 + b_2x_2$$

where b_0 is the y-intercept and b_1 and b_2 are the coefficients for the predictor variables x_1 and x_2.

If a predictor variable (e.g., x_1) does not impact Y, we would expect the coefficient (e.g., b_1) to be zero. However, there is variation in our processes and, when you run a regression, the coefficients that do not impact Y are not zero. We need a method of determining if a coefficient is sufficiently close to zero to be called zero or is far enough away from zero to be considered significant. We do this through the p value associated with a t-test.

For example, consider the dataset in Table 1. We have collected data on x_1, x_2 and Y. We want to use regression analysis to build a model for Y.

Table 1: Regression Dataset

Sample	x1	x2	Y		Sample	x1	x2	Y
1	88	52	186		11	95	57	196
2	82	57	174		12	117	29	242
3	95	60	197		13	95	21	199
4	101	25	210		14	104	22	218
5	102	41	214		15	102	42	209
6	97	33	204		16	87	24	183
7	101	47	209		17	105	60	219
8	99	24	208		18	101	20	208
9	92	37	190		19	106	60	220
10	93	31	196		20	110	42	228

Part of the output from the regression analysis using SPC for Excel is shown in Table 2. The table is described below.

Table 2: Partial Regression Output

	Coeff.	t Stat	p Value	95% Lower	95% Upper
Intercept	13.62	2.755	0.0135	3.189	24.05
x_1	1.954	41.056	0.0000	1.854	2.055
x_2	-0.0206	-0.774	0.4497	-0.0770	0.0357

The second column in Table 2 gives the coefficients (the b values in the regression equation). The "t Stat" column gives the t statistic associated with the coefficient. The "p Value" column gives the p values associate with the t statistic.

This p value is one key to interpreting the results. It is testing whether the coefficient is equal to zero (the null hypothesis in statistical jargon). A low value of p (usually < 0.05) suggests that the coefficient is not zero. This means that the coefficient most likely can be added to the model because it appears that changes in that predictor variable impacts the Y response. A high value of p suggests that the coefficient is zero. This means that the predictor variable does not impact the Y response and should not be included in the model.

How close to zero is the p value?

Table 2 also shows the 95% confidence interval. If that interval contains zero, then it is possible that the coefficient is zero (at that confidence interval) and should not be included in the model.

From Table 2, it can be seen that x_1 appears to be significant (p value < 0.05 and interval does not contain 0), while x_2 is not (p value > 0.05 and interval contains zero). The regression could be re-run with just x_1 in the model to create the final model.

Stepwise regression uses the p value to add or remove predictor variables from the model.

Introduction to Stepwise Regression

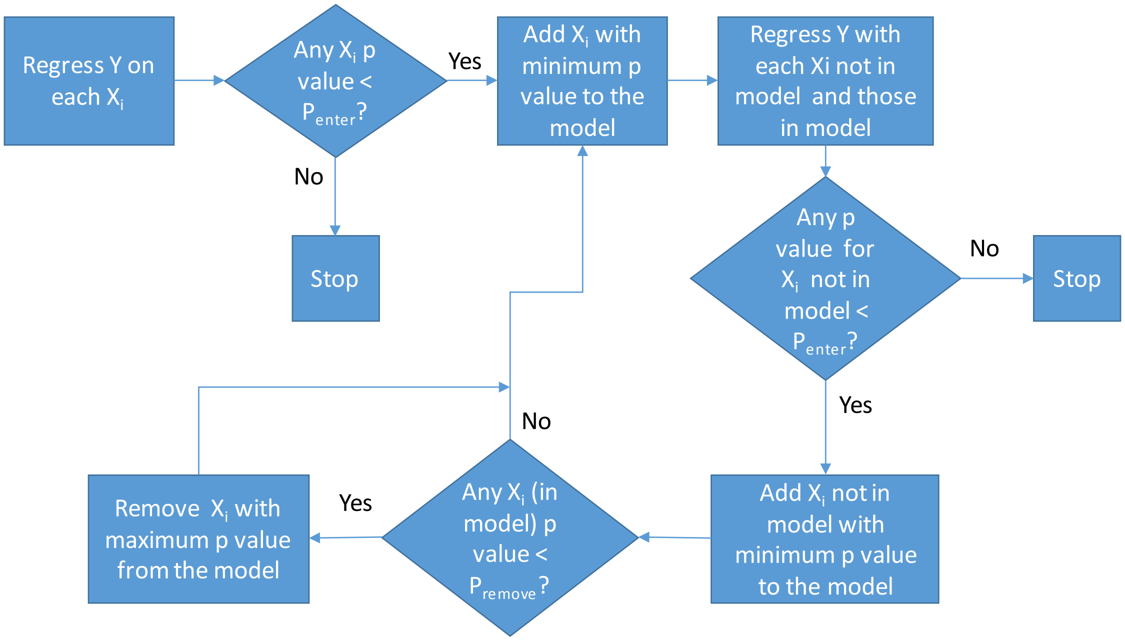
Stepwise regression adds or removes predictor variables based on their p values. The first step is to determine what p value you want to use to add a predictor variable to the model or to remove a predictor variable from the model. A common approach is to use the following:

$$\text{p value to enter} = P_{enter} = 0.15$$

$$\text{p value to remove} = P_{remove} = 0.15$$

A process flow diagram for stepwise regression is shown in Figure 1.

Figure 1: Stepwise Regression


Regress Y on each X_i
Any X_i p value < P_{enter}?
Yes
No
Stop
Add X_i with minimum p value to the model
Regress Y with each Xi not in model and those in model
Any p value for X_i not in model < P_{enter}?
No
Stop
Yes
Add X_i not in model with minimum p value to the model
Any X_i (in model) p value < P_{remove}?
No
Yes
Remove X_i with maximum p value from the model

You start with no predictor variables in the model. Regress each predictor variable, individually, with Y. Then you compare the p values for each predictor variable. If there are no p values less than P_{enter}, then there are no predictor variables that go into the model and the stepwise regression ends. If there are p values less than P_{enter}, then the predictor variable with the lowest p value is added to the model.

At this point, the model contains 1 predictor variable. The looping action begins now. Each individual predictor variable that is not in the model is added to the model and the regression run. So, if x_1 is in the model, you would run the regression for (x_1, x_2), (x_1, x_3) up to (x_1, x_k) where k is the number of predictor variables. If any p value for a x_i not in the model is less than P_{enter}, then the smallest p value for the x_i not in the mode is added to the model.

Adding additional predictor variables to the model can change the p values of the predictor variables already in the model. A check is made to see if any predictor variable in the model have a p value greater than P_{remove}. If so, the predictor variable with the greatest p value is removed.

This process continues until no more predictor variables can be added to or removed from the model. The example below shows the process of adding and removing predictor variables.

Stepwise Regression Example

You are a VP of sales and have responsibility for 44 stores. You have collected data from the stores on advertising costs, store size in square feet, % employee retention, customer satisfaction score, whether a promotion was run or not and sales. You want to build a model that can predict sales based on these five variables. The data are shown in Table 3.

Table 3: Store Data

Store	Advertising Costs	Size (Sq. Ft)	% Employee Retention	Customer Satisfaction Score	Promotion (1 = Ran)	Sales
1	124.4	22560	59	32	1	1581
2	154	31181	62	33	1	2139
3	123.5	16314	78	28	0	1043
4	163	24205	66	43	0	1702
5	107	17574	82	22	1	1339
6	143.9	19584	67	34	0	521
7	133.7	22682	57	32	1	1720
8	121.4	23398	64	23	1	1197
9	104.6	19507	88	25	0	950
10	99.2	11443	87	17	0	266
11	93.8	16832	82	29	1	1718
12	133.8	24326	70	37	1	1820
13	131.3	18541	86	27	1	1805
14	123.2	22099	67	30	0	1042
15	88.3	16928	80	21	0	655
16	154.3	16237	69	28	1	1480
17	112.1	15290	87	19	1	1057
18	114	12947	80	23	0	953
19	91.7	16326	77	24	0	364
20	113.7	13024	82	27	0	783
21	105.7	22054	86	28	1	792
22	161.3	22637	75	34	1	2185
23	143	18733	64	26	1	1051
24	113.7	21126	58	30	0	1456
25	69	16819	57	18	0	146
26	106.7	16992	74	22	0	899
27	125.8	18355	88	25	1	1243
28	52.4	13958	87	25	0	421
29	114.7	18298	71	24	0	318

Store	Advertising Costs	Size (Sq. Ft)	% Employee Retention	Customer Satisfaction Score	Promotion (1 = Ran)	Sales
30	142.5	18016	55	27	0	383
31	114.6	22317	80	27	1	993
32	150.5	26221	71	31	1	1766
33	74.5	19494	68	24	1	1123
34	111.4	18406	85	24	1	1523
35	117.7	18880	86	24	1	1281
36	108.4	21312	81	30	1	1899
37	171.7	26618	72	35	1	2508
38	156.5	15981	72	25	0	1042
39	156.3	20749	61	32	0	1416
40	140.6	22458	77	27	1	1659
41	155.1	18579	85	32	0	1370

A stepwise regression was done on these data using the SPC for Excel software. The p values to add and remove were both set at 0.15.

The first step was to regress Y on each predictor variable. This simply means run regression for each predictor variable alone versus Y. Then, the predictor variable with the lowest p value is added to the model (as long as is there is a predictor variable with a p value < 0.15. The store size had the lowest p value so it is added to the model in the first step. The output is shown below.

Step 1: Added Size (Sq. Ft)

Variable	Coefficient	t Stat	p Value
Intercept	-633.8	-1.893	0.066
Size (Sq. Ft)	0.0946	5.618	0.000

The second step is then to include each predictor variables (one at time) in the model that includes store size and run the regression. If any p value for a predictor variable that is not in the model is less than 0.15, that predictor variable is added to the model. Note that if two of the predictor variables have a p value less than 0.15, the predictor variable with the lowest p value is added to the model. You are only adding or removing one variable at a time. In this case, promotion had the lowest p value, so it is added to the model.

Step 2: Added Promotion

Variable	Coefficient	t Stat	p Value
Intercept	-355.3	-1.163	0.252
Size (Sq. Ft)	0.0675	4.036	0.000
Promotion	464.5	3.501	0.001

After a variable has been added, you check to see if any of the predictor variables in the model now have a p value greater than 0.15 (the p value to remove). This is not the case, so the model now has two terms: store size and promotion.

The process repeats. Each predictor variable, again by itself, not in the model is added to the model with two predictor variables and the regression is run. In this step, customer satisfaction had the lowest p value below 0.15 and is added to the model as shown below.

Step 3: Added Customer Satisfaction Score

Variable	Coefficient	t Stat	p Value
Intercept	-826.8	-2.941	0.006
Size (Sq. Ft)	0.0120	0.613	0.543
Customer Satisfaction Score	54.15	4.107	0.000
Promotion	594.5	5.132	0.000

Since p values change as additional predictor variables are added, you have to check to see if any of the predictor variables in the model now have a p value greater than 0.15 (the p value to remove). The store size has a p value greater than 0.15, so it is removed from the model.

Step 4: Removed Size (Sq. Ft)

Variable	Coefficient	t Stat	p Value
Intercept	-766.8	-2.933	0.006
Customer Satisfaction Score	59.76	6.343	0.000
Promotion	630.7	6.382	0.000

The process repeats again using this model that now contains customer satisfaction score and promotion. Advertising costs has the lowest p value under 0.15 and is added to the model.

Step 5: Added Advertising Costs

Variable	Coefficient	t Stat	p Value
		-	
Intercept	-899.8	3.423	0.002
Advertising Costs	4.456	1.880	0.068
Customer Satisfaction Score	45.13	3.764	0.001
Promotion	608.8	6.315	0.000

When the process is repeated at this point, there are no predictor variables not in the model that have a p value less than 0.15. So, there are no predictor variables to add or remove and the stepwise regression is completed. This analysis led to three of the predictor variables being included in the model: advertising costs, customer satisfaction score and whether the promotion was run or not. The model is:

Y = -899.8 + 4.456(Advertising Costs) + 45.13(Customer Satisfaction Score) + 608.8(Promotion)

You can easily run a full regression analysis that includes only these variables to complete the analysis.

Caveats about Stepwise Regression

This process seems very inviting. You have lots of possible predictor variables and this process adds or removes predictor variables until a single model is reached. Pretty neat.

However, if you search the internet about potential problems with stepwise regression, you will get quite a few hits. To me, the biggest problem (not unique to stepwise regression) is that encourages us not to think. Here is the model. You are done. The problem is that the process is automated. It can't contain the knowledge of the subject expert. You need to look at the model generated by stepwise regression from a practical point of view – does it make sense to the people who are the experts.

Another concern is, if there is excessive linear dependence (called multicollinearity), the procedure can end up throwing most of the variables in the model. This can also occur if there the number of variables to be tested in the model is large compared to the number of samples in your data. Stepwise regression may not give you the model with highest R^2 value (measure of how well the model explains the variation in the data). Some even say that stepwise regression usually doesn't pick the best model.

But, in reality, you have to use your knowledge of the process to decide if the model makes sense. And you can always run validation experiments to confirm the model.

Summary

The month's publication introduced stepwise regression. This is an automated technique for building a model from a larger number of predictor variables. The procedure is based on adding or removing predictor variables from a model based on p values. This procedure, like any automated procedure, cannot take into account your knowledge of the process. When you have your final model from stepwise regression, you should ensure that it makes sense to you, the subject expert. And, you should also run confirmation runs to ensure that the model is valid.

Quick Links

Visit our home page

SPC for Excel Software

SPC Training

SPC Consulting

SPC Knowledge Base

Ordering Information

Thanks so much for reading our publication. We hope you find it informative and useful. Happy charting and may the data always support your position.

Sincerely,

Dr. Bill McNeese
BPI Consulting, LLC