Quantifying Health

Understand Forward and Backward Stepwise Regression

Regression Analysis

Running a regression model with many variables including irrelevant ones will lead to a needlessly complex model.

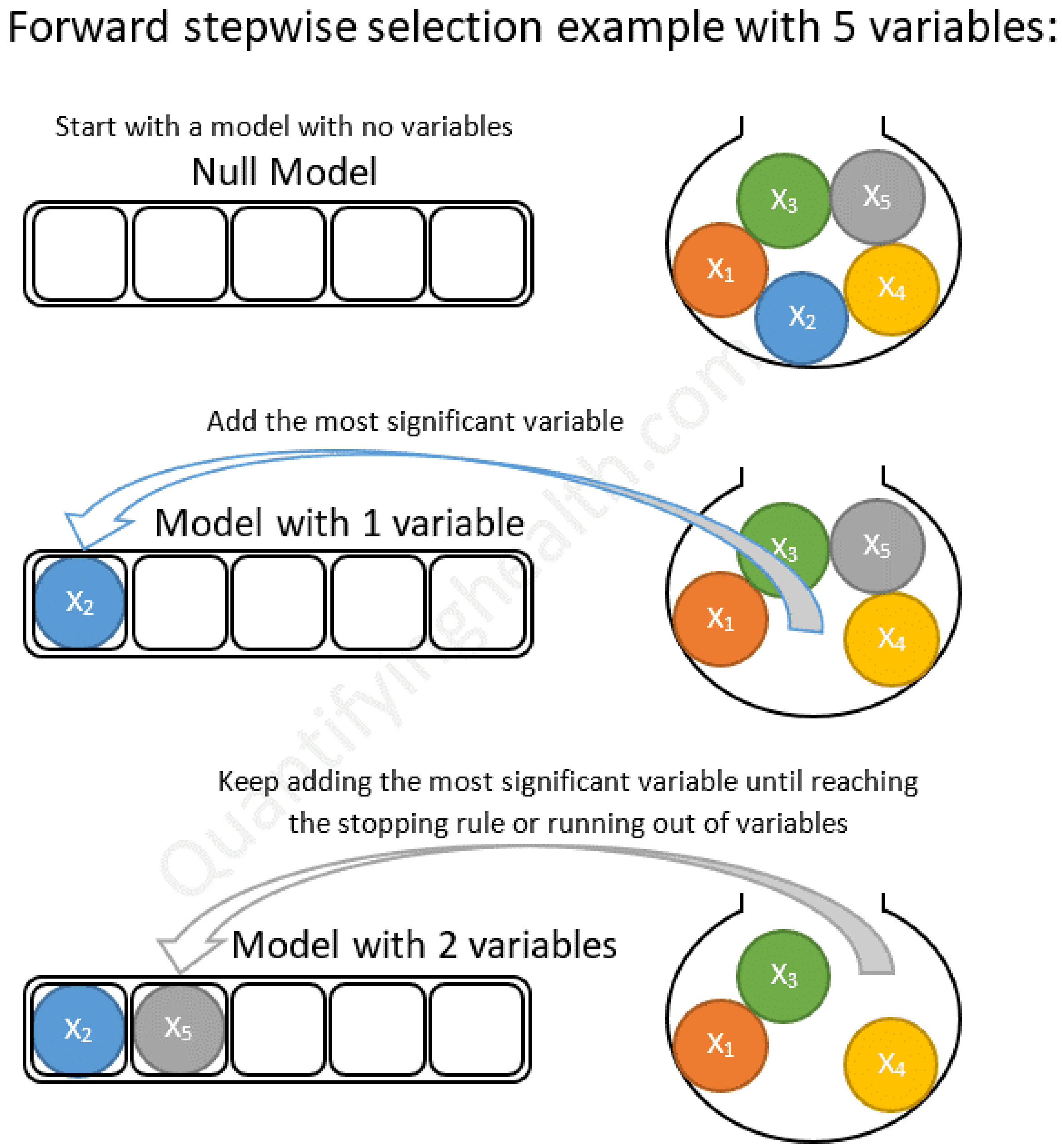
Stepwise regression is a way of selecting important variables to get a simple and easily interpretable model. Below we discuss Forward and Backward stepwise selection, their advantages, limitations and how to deal with them.

Forward stepwise

Forward stepwise selection (or **forward selection**) is a variable selection method which:

1. **Begins** with a model that contains no variables (called the *Null Model*)
2. **Then** starts adding the most significant variables one after the other
3. **Until** a pre-specified stopping rule is reached or until all the variables under consideration are included in the model

Here’s an example of forward selection with 5 variables:


Forward stepwise selection example with 5 variables:
Start with a model with no variables
Null Model
X3
X5
X1
X4
X2
Add the most significant variable
Model with 1 variable
X2
Keep adding the most significant variable until reaching the stopping rule or running out of variables
Model with 2 variables
X2
X5
QuantifyingHealth.com

In order to fully understand how forward selection works, we need to know:

1. How to determine the most significant variable at each step
2. How to choose a stopping rule

1. Determine the most significant variable to add at each step

The most significant variable can be chosen so that, when added to the model:

- Has the smallest p-value, or
- Provides the highest increase in R^2, or
- Provides the highest drop in model RSS (Residuals Sum of Squares) compared to other predictors under consideration.

2. Choose a stopping rule

The stopping rule is satisfied when all remaining variables to consider have a p-value larger than some threshold if added to the model.

When we reach this state, forward selection will terminate and return a model that only contains variables with p-values < threshold.

How to determine the threshold?

The threshold can be:

1. A fixed value (for instance: 0.05 or 0.2 or 0.5)
2. Determined by AIC (Akaike Information Criterion)
3. Determined by BIC (Bayesian information criterion)

If we choose a fixed value, the threshold will be the same for all variables.

However, if we let AIC or BIC automatically determine the threshold, it will be different for each variable.

Fortunately, computers nowadays calculate these thresholds automatically so we do not have to bother with the details. However, I think it is interesting to have at least some understanding of what is going on under the hood.

Ready for some geeky details?

How does AIC determine the threshold?

AIC chooses the threshold according to how many degrees of freedom the variable under consideration has.

Take for example the case of a binary variable (by definition it has 1 degree of freedom): According to AIC, if this variable is to be included in the model, it needs to have a p-value < 0.157.

The more degrees of freedom a variable has, the lower the threshold will be.

How does BIC determine the threshold?

BIC chooses the threshold according to the effective sample size n.

For instance, for n = 20, a variable will need a p-value < 0.083 in order to enter the model.

The larger n is, the lower the threshold will be.

BIC is a more restrictive criterion than AIC and so yields smaller models. Therefore it is only recommended when working with large sample sizes — where the sample size (or number of events in case of logistic regression) exceeds 100 per independent variable [Heinze et al.].

Note that both AIC (and BIC) can be applied to the pooled degrees of freedom of all unselected predictors. But applying it to individual variables (like we described above) is far more prevalent in practice.

Backward stepwise

Backward stepwise selection (or **backward elimination**) is a variable selection method which:

1. **Begins** with a model that contains all variables under consideration (called the *Full Model*)
2. **Then** starts removing the least significant variables one after the other
3. **Until** a pre-specified stopping rule is reached or until no variable is left in the model

Here's an example of backward elimination with 5 variables:

Like we did with forward selection, in order to understand how backward elimination works, we will need discuss how to determine:

1. The least significant variable at each step
2. The stopping rule

1. Determine the least significant variable to remove at each step

The least significant variable is a variable that:

- Has the highest p-value in the model, or
- Its elimination from the model causes the lowest drop in R^2, or
- Its elimination from the model causes the lowest increase in RSS (Residuals Sum of Squares) compared to other predictors

2. Choose a stopping rule

The stopping rule is satisfied when all remaining variables in the model have a p-value smaller than some pre-specified threshold.

When we reach this state, backward elimination will terminate and return the current step's model.

How to determine the threshold?

As with forward selection, the threshold can be:

1. A fixed value (for instance: 0.05 or 0.2 or 0.5)
2. Determined by AIC (Akaike Information Criterion)
3. Determined by BIC (Bayesian information criterion)

Should you use forward or backward stepwise selection?

Where forward stepwise is better

Unlike backward elimination, forward stepwise selection can be applied in settings where the number of variables under consideration is larger than the sample size!

This is because forward selection starts with a null model (with no predictors) and proceeds to add variables one at a time, and so unlike backward selection, it DOES NOT have to consider the full model (which includes all the predictors).

In fact, it will only consider models with number of variables less than:

- The sample size (for linear regression)
- The number of events (for logistic regression)

Where backward stepwise is better

Starting with the full model has the advantage of considering the effects of all variables simultaneously.

This is especially important in case of collinearity (when variables in a model are correlated which each other) because backward stepwise may be forced to keep them all in the model unlike forward selection where none of them might be entered [see Mantel].

BOTTOM LINE:

Unless the number of candidate variables > sample size (or number of events), use a backward stepwise approach.

Advantages and limitations of stepwise selection

Advantages of stepwise selection:

(Note that these advantages are shared by most automated methods that reduce the number of predictors).

Stepwise selection is easy to run in most statistical packages. For example, here’s how to run forward and backward selection in SPSS:

Note: Before you run stepwise regression, make sure to impute missing values, otherwise your sample size will be restricted to observations that do not have any missing values in any of the variables under consideration.

A regression model fitted in cases where the sample size is not much larger than the number of predictors will perform poorly in terms of out-of-sample accuracy. In these cases, reducing the number of predictors in the model by using stepwise regression will improve out-of-sample accuracy (generalizability).

When it comes to interpreting a statistical model's output, we can all agree that a smaller model is more desirable than a complicated one. By reducing the number of variables, stepwise selection will yield a simple and easily interpretable model.

It also provides a reproducible and objective way to reduce the number of predictors compared to manually choosing variables based on expert opinion which, more often than we would like to admit, is biased towards proving one's own hypothesis.

Note: Automated variable selection is not meant to replace expert opinion. In fact, important variables judged by background knowledge should still be entered in the model even if they are statistically non-significant.

Where automated variable selection is most helpful is in exploratory data analysis especially when working on new problems not already studied by other researchers (where background knowledge is not available).

Limitations of stepwise selection:

Stepwise selection does not consider all possible combination of potential predictors.

This is both a feature and a bug as:

- It will provide a computational advantage over methods that do consider all these combinations
- It is not guaranteed to select the best possible combination of variables

The regression coefficients, confidence intervals, p-values and R^2 outputted by stepwise selection are biased and cannot be trusted. The direction of the bias is as follows:

Stepwise regression output	**Bias direction**
Regression coefficients	Will appear larger
Confidence intervals	Will appear narrower
p-values	Will appear smaller; Also invalid
R^2	Will appear larger

The selection of variables using a stepwise regression will be highly unstable, especially when we have a small sample size compared to the number of variables we want to study.

This is because many variable combinations can fit the data in a similar way!

Note: You can test the instability of the stepwise selection by rerunning the stepwise regression on different subsets of your data. When there is instability, you will notice that you'll get a different selection of variables each time.

This instability is reduced when we have a sample size (or number of events) > 50 per candidate variable [Steyerberg et al.].

In case you didn't notice, 50 is a really HUGE number:

Imagine that for a stepwise regression with only 10 candidate variables you will need 500 events to reduce the instability of the stepwise selection algorithm!

Finally, stepwise regression, like all other automated methods, is easy to run without even thinking about the problem at hand.

For a more technical discussion of these limitations, I recommend the following books:

- *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* – by Frank Harrell

- *Clinical prediction models: A practical approach to development, validation and updating* – by Ewout Steyerberg.

How to deal with limitations of the stepwise approach

First of all you have to accept the fact that stepwise regression does not always select the best possible combination of variables.

There is no way around it!

Therefore, when reporting your results NEVER use the words: “the best predictors were…” or “the best model contains the following variables…”.

For obtaining an unbiased estimation of the regression coefficients, confidence intervals, p-values and R^2, you can divide the sample into training and testing sets:

1. Use the first set to run a stepwise selection (i.e. selecting important variables)
2. Use the second set to run a model with the selected variables to estimate the regression coefficients, p-values and R^2.

This approach certainly has the drawback of throwing half the sample you collected and therefore is very costly in certain cases.

For checking the stability of the selection, you can use the bootstrap method. It works as follows:

1. Take sub-samples from your original sample (with replacement) and perform stepwise selection on these sub-samples
2. The most important variables will be those that have a high frequency of inclusion in these sub-samples

For a proper discussion of how this method works, how to use it in practice and how to report its results see Heinze et al.

Finally, take a moment to consider other variable selection methods like:

- Selection based on prior research
- Shrinkage methods such as LASSO regression
- Dimensionality reduction methods like principle components analysis

However, this does not mean that you should never use stepwise regression, just remember that it comes with a very high cost.

To help you remember that last note, I want to leave you with the following 2 quotes:

The first is from IBM, the developers of SPSS themselves:

> *The significance values [a.k.a. p-values] are generally invalid when a stepwise method (stepwise, forward, or backward) is used.*
>
> ***IBM Knowledge Center***

And the other quote is from a statistics book:

> *Stepwise variable selection has been a very popular technique for many years, but if this procedure had just been proposed as a statistical method, it would most likely be rejected because it violates every principle of statistical estimation and hypothesis testing.*

> *Regression modeling strategies – Frank Harrell*

Further reading

- Which Variables to Include in a Regression Model
- Standardized vs Unstandardized Regression Coefficients
- Understand Best Subset Selection
- Understand Regularized Regression
- Why and When to Include Interactions in a Regression Model

← Previous Post Next Post →

About Me

I am George Choueiry, PharmD, MPH, my objective is to help you analyze data and interpret study results without assuming a formal background in either math or statistics.

Your Feedback

Use the textbox below to suggest a topic you would like me to write about, report an error, or criticize anything you see on this website. Honesty is welcome. I want this website to be helpful to you more than I want you to be kind to me!

Send