PhUSE 2017

Paper AS06

"The Model Statistician", Insights into model selection techniques in Clinical Trials

Peter Williams, Veramed, Biohub (Alderley Park), UK

ABSTRACT

Maintaining the integrity of clinical trials and the frequent requirement for statisticians and programmers to remain blinded means that we are often constrained by pre-specified analysis plans where the statistical analysis is finalised and signed off prior to unblinding. This does not lend itself to using model selection techniques as a part of the statistical analysis, but does model selection have a place within a clinical trials setting? Model selection can be utilised as an exploratory tool on both early and late phase study data to aid future study design and analysis decisions and identify areas where further investigation may be worthwhile. Functionality within SAS® software facilitates efficient implementation of various model selection techniques. Finding the most parsimonious model is never an exact science, so how much influence should data driven conclusions have?

INTRODUCTION

The general term "statistical model" is sufficiently broad that statistical models underpin most efficacy and safety analyses performed in clinical trials. The type of model selected for an endpoint of interest is almost exclusively driven by the type of data collected and its expected distribution (e.g. for time-to-event data we would consider using a Cox proportional hazards model, but for a binary endpoint we would be likely to employ a logistic regression model). The extent of modern statistical techniques readily available within standard statistical software such as SAS means that there are occasions where more than one type of statistical model may be appropriate. However, discussion surrounding the type of model to choose is outside of the scope of this paper. The paper instead focuses on the variable selection techniques that can be efficiently implemented in SAS procedures and where/why we might consider utilising these techniques in a clinical trials setting.

Expert clinical knowledge of the therapeutic area is utilised at the planning and design stage of clinical trials to control for variables considered to be likely to have a confounding effect. Stratification is a common technique employed at randomisation to achieve balance across the treatment groups with respect to known confounders (if the randomisation is stratified then the primary analysis must be adjusted for these variables). Similarly, the planned analysis can be stratified to fix or restrict the level of the confounding variable in the performed analysis. However, particularly in cases where there are a lot of strata, stratified analysis can drastically reduce the sample size and this can lead to issues regarding the statistical assumptions underpinning the model and statistical power.

Multivariable statistical models are a commonly used alternative solution for controlling for confounders and provide a platform for multiple variables to be controlled for simultaneously, but which variables should we include? More often than not expert knowledge of the therapeutic area will point to these variables, but can a data driven approach help to inform and supplement these decisions? This paper outlines the overall concept of statistical model selection and explores the variable selection techniques available in SAS statistical procedures.

AN OVERVIEW OF STATISTICAL MODELS AND MODEL SELECTION

STATISTICAL MODELS

One of the main concepts of statistical analysis is using the information gained from a sample of individuals to make inferences about the relevant population [1]. This general principle applies to statistical models in that we use data taken from a sample of individuals to estimate a dependent response variable based on a selection of independent explanatory variables. Providing that the sample is representative of the relevant population and none of the model's statistical assumptions are violated, the model results can be used to make inferences about the population of interest.

MODEL SELECTION

Statistical model selection can be defined as a mechanism used to select the 'best' model from a set of candidate models. The definition of 'best' is less clear and the source of some controversy between statisticians. Generally speaking, good model selection techniques seek a balance between a well-fitting model based on the available data, and simplicity. That is, we ideally want the model selected to identify and describe trends in the data without being too complicated. Not only can the interpretation of complicated models be difficult, it can also become meaningless due to

overfitting if the model is too specific to the peculiarities of the sampled data to be applicable to the population as a whole.

Candidate models contain multiple potential explanatory variables and model selection procedures aim to systematically determine the set of explanatory variables (and interactions) that 'best' describe the response variable. There are various model selection techniques including forward, backward and stepwise variable selection available within SAS and these will be discussed further in a later section of this paper. Implementing these techniques appropriately using statistical software such as SAS can be a quick and efficient way of identifying correlations and trends of interest within data.

LINEAR REGRESSION EXAMPLE

To demonstrate the importance of finding a balance between goodness-of-fit and simplicity, we will use a simple example: We wish to create a model to estimate the weight of children aged between 11 and 16 using weight and height data for 19 subjects (10 males and 9 females) within this age range. The response variable is weight; height, gender and age are the potential explanatory variables available. Some simple exploratory plots show that, as we might expect, the variable with the clearest correlation with weight was height, but that there was also a positive correlation with age, and male subjects tended to weigh more than female subjects. Note: while this approach to exploring data is useful, it is rather crude and doesn't account for potential correlations present between the explanatory variables.

Figure 1 shows plots of weight against height including the model weight estimate produced by the following two linear regression models:

Model A: $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$

Model B: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$

where y_i is the response variable weight, x_{i1} and x_{i3} are the continuous covariates height and age respectively, x_{i2} is the factor gender, β_0 is the intercept, β_1, β_2 and β_3 are the parameter estimates associated with each explanatory variable and ε_i is the residual error. The points plotted are the actual observed weights for each of the 19 subjects and the solid blue line gives the weight estimates obtained from the linear regression model for each subject.

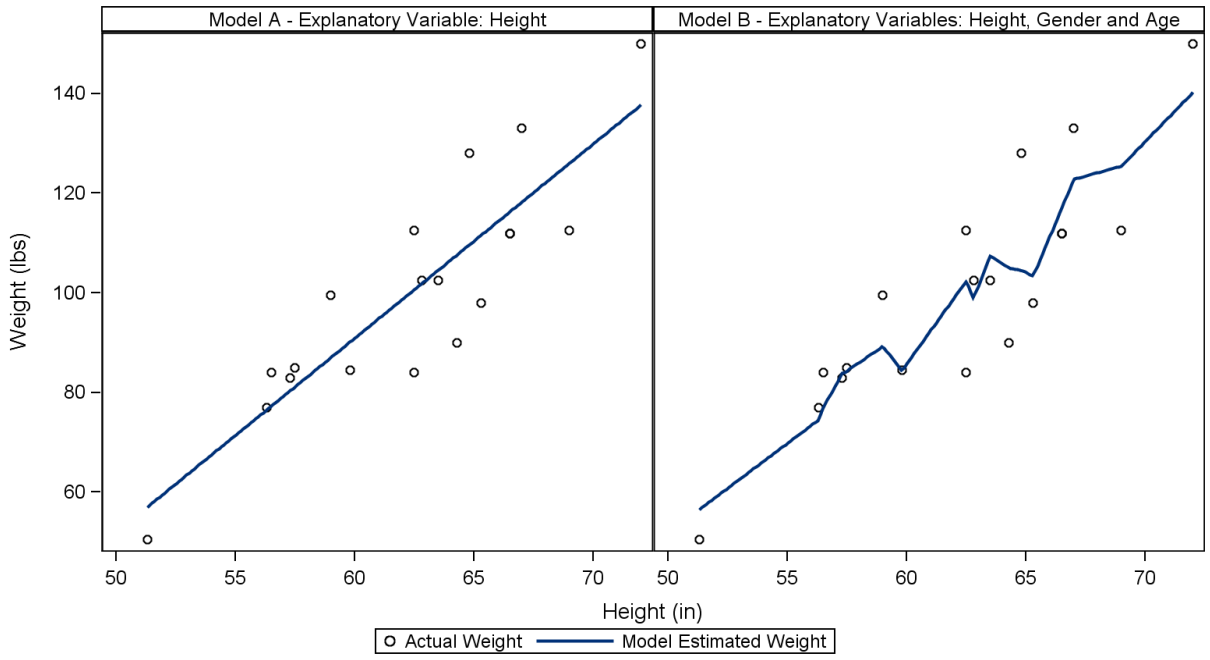

Figure 1: Model fit example.

Figure 1 reveals that Model A is relatively simple to interpret and even based on this small sample size we could be fairly confident that this linear relationship between weight and height would be true of the overall population this sample was taken from. Model B in Figure 1 shows that by additionally including gender and age in the model we've introduced complexity to the model estimated weights. Comparing against Model A and given the strong correlation present between weight and height it is questionable as to whether including the additional complexity has improved the model fit to the data. Furthermore, the interpretation in relation to the population of interest becomes more difficult.

Just looking at the two plots presented in Figure 1, my personal preference is the model represented by the left hand panel, but how would we systematically show whether this model is 'best'. The answer lies in variable selection techniques.

MODEL SELECTION TECHNIQUES AVAILABLE IN SAS

The iterative nature of most model selection techniques means that manual implementation can prove incredibly time consuming and ultimately inefficient. Fortunately, the most recent versions of certain SAS procedures contain functionality to perform the variable selection processes based on user defined parameters. Table 1 gives a summary of the model selection techniques available in SAS version 9.3/9.4 by SAS procedure.

Table 1: Model selection techniques available by SAS procedure

	PROC LOGISTIC	PROC REG	PROC PHREG	PROC GLMSELECT
Forward Selection	Y	Y	Y	Y
Backward Selection	Y	Y	Y	Y
Stepwise Selection	Y	Y	Y	Y
Branch-and-Bound Algorithm	Y		Y	
R^2 Selection (various)*		Y		
Mallows C_p Selection		Y		
Least Angle Regression				Y
LASSO				Y
Group LASSO				Y^
Elastic Net				Y^

*Including Maximum R^2 Improvement, Minimum R^2 Improvement, R^2 Selection and Adjusted R^2 Selection.
^Technique only available in SAS v9.4.

The following is a brief summary of what each of the above SAS procedures can be used for:

- **PROC LOGISTIC:** Fits linear logistic regression models for discrete response data by the method of maximum likelihood. It can also perform conditional logistic regression.
- **PROC REG:** A general purpose procedure for fitting regression models.
- **PROC PHREG:** Performs regression analysis of survival (or 'time to event') data based on the Cox proportional hazards model.
- **PROC GLMSELECT:** Performs effect selection in the framework of general linear models. The procedure bridges the gap between PROC GLM, which doesn't include effect selection functionality, and PROC REG, which doesn't include a CLASS statement.

STEPWISE TECHNIQUES

Forward, Backward and Stepwise selection are amongst the most widely used model selection techniques. They all follow a relatively intuitive process where candidate models are explored by either adding or deleting an explanatory variable/interaction term from the model before comparing against the model from the previous stage.

Forward selection starts with the null model (which includes no explanatory variables) and each explanatory variable is considered individually for inclusion in the model. The variable which makes the greatest contribution (explaining the greatest amount of variability in the response variable) to the model is included providing that the associated p-value is less than the pre-specified value (often set at 0.05 or 0.1). This model is then carried forward to the next step where the remaining explanatory variables are again considered one at a time for inclusion in the model. Once a variable is added to the model it cannot be removed even if the p-value increases to being above the pre-specified level when further variables are added to the model. The process is repeated until there are no further explanatory variables that meet the pre-specified criteria for inclusion in the model [2].

Backward selection is a similar iterative process to forward selection, but instead starts with the full model containing all explanatory variables. At each step the variable that contributes the least to the model is removed until all remaining explanatory variables have p-values less than the pre-specified value [2]. Similarly to forward selection, once a variable has been removed from the model it cannot be added back in. Note: Backward selection is sufficiently simple that 'manual' implementation is not overly time consuming and can allow a more pragmatic approach to the removal of explanatory variables from the model.

Stepwise selection is similar to forward selection, but less restrictive in that once an explanatory variable is added to the model it does not necessarily stay there. Thus, a significance level for entry to the model and a significance level for variables to stay within the model need to be specified for stepwise selection. These significance levels need not be equal. The following example SAS code uses a stepwise approach with entry and stay criteria of 0.05 and 0.1 respectively (the `@2` and `|` separators request that all two-way interactions are considered):

```
proc glmselect data = dsin;
  class sex;
  model weight = height|age|sex @2 / selection = stepwise select = sl sle = 0.05
                                     sls = 0.1;
  ods output parameterestimates = step_est;
run;
```

Despite this relatively simple implementation in SAS, stepwise variable selection techniques should be used with care. Such techniques require application of a hypothesis test intended for one test to many tests. It can be shown that this assumption violation biases p-values and standard errors towards zero, parameter estimates away from zero and can potentially lead the model selected being too complex [3]. There is no guarantee that the final model selected truly is the 'best' model! In fact, applying the forward and backward selection methods to the same dataset will not necessarily lead to the same final model being selected by both techniques. A final perceived weakness of stepwise approaches is that due to nature of the algorithms not all potential candidate models will be considered.

ALTERNATIVE TECHNIQUES

Various R^2 selection methods and Mallows C_p Selection fall under a set of model selection techniques often referred to as **Best Subset Selection**. These techniques consider all possible combinations of explanatory variables to find the 'best' model based on the criteria for the specific method selected. So for example, Adjusted R^2 selection uses an adjusted version of the R^2 statistic to select the 'best' model. One disadvantage to a method that considers all possible models is that it can become very computationally intensive and therefore time consuming where there are a large number of potential predictors to consider [2]. A further limitation to these techniques is that that they are only appropriate in cases where all potential explanatory variables are continuous.

Least Angle Regression (LARS) is similar to Forward Selection in that it starts with the null model and a parameter is added at each step. The algorithm starts by centering all variables and scaling the covariates so they have the same corrected sum of squares. After initially setting all parameters to zero, the predictor which is most correlated with the current residual is added to the model and the associated parameter estimate is determined by the distance moved in the direction of this predictor before a second predictor has as much correlation with the current residual. LARS then proceeds in a direction equiangular between the two predictors until a third variable has sufficient correlation with the current residual to be added to the 'most correlated' set. This process is continued until all predictors are in the model [4]. Either a number of steps or specific stopping criteria (e.g. Mallows C_p Selection) can be specified for LARS using the STOP= option in PROC GLMSELECT [5]. The stepwise nature of LARS means that it is prone to being biased towards more complex models [4].

LASSO (least absolute shrinkage and selection operator) selection arises from a constrained form of ordinary least squares regression where the sum of the absolute values of the parameter estimates is constrained to be smaller than a specified LASSO parameter. Providing that this specified LASSO parameter is small enough, some of the parameter estimates will be exactly zero and therefore the LASSO effectively selects a subset of explanatory variables for a given LASSO parameter. Incrementing the LASSO parameter in discrete steps yields a sequence of explanatory variables with non-zero parameter estimates which can be thought of as being selected for inclusion in the model. Stopping criteria in PROC GLMSELECT can be specified similarly to LARS [5].

MODEL SELECTION IN CLINICAL TRIALS

The very nature of model selection means that more often than not there will be an exploratory element to situations where model selection is appropriate. It is therefore unlikely that model selection techniques will be appropriate as part of the analysis in Phase II and Phase III studies. Nevertheless there are situations surrounding early phase and late phase/post marketing studies where model selection can prove a useful tool.

EARLY PHASE

Data collected during early phase studies can be explored using model selection techniques to investigate any correlations present between the endpoints of interest and other factors/covariates. It is a requirement that factors/covariates to be included in primary analysis must be pre-specified in the protocol. Therefore, any correlations detected could indicate variables that need careful consideration during both the study design and the planning of statistical analysis for the later phase studies. That said, these insights into the data should merely help inform decisions and never supersede expert knowledge in relation to the target indication. The nature of early phase studies means that sample sizes are likely to be small and hence this may limit the conclusions that you are able to draw from the available data.

LATE PHASE

Model selection techniques can be employed to explore data from late phase studies and post-marketing data in a number of ways. Some examples of situations where this may be appropriate and questions you may want to answer include:

- Identifying prognostic factors/covariates which have an association with disease outcome. If differences in disease outcome between subgroups were to be investigated then these factors and covariates should be adjusted for in any analysis performed.
- Exploring any additional/unexpected treatment effects present in the data. You may want to see whether the treatment had any unexpected additional benefits and if this seems to be the case, whether or not this was truly due to the treatment effect and not exaggerated or caused by confounding factors.
- Seeking additional details about the efficacy profile of a treatment successfully taken to market. For example: Is the treatment particularly efficacious for certain cohorts? Does the data indicate that the treatment may be efficacious for a broader age range than originally planned?
- Finding an appropriate set of factors/covariates to include in an imputation model (for example MCMC (Markov Chain Monte Carlo)) for an estimand that requires missing data points to be imputed.
- Performing exploratory analysis on data (could be pooled across multiple studies) with the view of confirming expected correlations and identifying any unexpected trends present in the data. Unexpected trends may point to areas where further investigation could be worthwhile. Further to this, any instances where statistical indications from the data are at all contrary to expert opinion may well be of interest.

One of the main considerations to make when using model selection techniques on pooled data across multiple studies is whether the data is complete enough to draw meaningful conclusions. Due to the sheer complexity of clinical trials, it is unlikely that data will have been collected in an identical fashion across all studies being pooled. If parameters of interest have not been collected for a subset of subjects then these subjects would, by default, be excluded from the model selection process by SAS. Techniques for dealing with missing data are available within SAS, but are outside of the scope of this paper.

A potential pitfall when performing exploratory analysis on a non-primary study endpoint is that data for known predictors may not have been collected because they were not required for the original planned analysis. Unfortunately there is little you can do in cases like this, but it's important to understand that the possible analysis may be limited by the data collected.

Care should always be taken not to confuse statistical significance with clinical significance. For example, a statistically significant difference detected between treatment groups for a parameter of interest will not necessarily be a large enough absolute difference for it to be considered clinically significant. This is a particularly relevant consideration to make when dealing with larger samples associated with late phase and post marketing data since the absolute difference required to constitute a statistically significant difference will decrease as the sample size increases. In a similar vein, a difference not considered statistically significant could still be of clinical interest and therefore selection criteria in exploratory model selection are often less stringent than those typically used in hypothesis testing. These points link back to the importance of data driven conclusions not overriding expert therapeutic knowledge.

CONCLUSION

There are areas in clinical trial analysis where model selection is warranted. Developments in recent versions of SAS mean that numerous model selection techniques can be implemented efficiently to explore clinical data. However, care should always be taken and the following main considerations adhered to:

- The statistical assumptions of the underlying statistical model should be satisfied.
- Data driven conclusions should always be considered in their clinical context and never trump expert knowledge in the therapeutic area.
- Statistical significance and clinical significance should never be confused when drawing conclusions.

There is no particular model selection technique that stands out above the rest as a preferred technique. My personal preference sways slightly towards the more intuitive stepwise techniques and I particularly like how the relative simplicity of backward variable selection can allow a more pragmatic approach through manual implementation if time permits. That said, my recommendation would be to try using more than one technique and see whether or not the same subset of variables are selected for inclusion in the final model. If the techniques agree then this adds to our confidence that the model selected is the most parsimonious, however if they do not then further investigation may be required for the variables that differ. Regardless of the results, the most important consideration when using model selection techniques is to not be lead astray by the data driven conclusions and remember that expert knowledge should primarily guide the research, not the data.

REFERENCES

[1] D. G. Altman, Practical Statistics for Medical Research, New York: Chapman & Hall/CRC, 1991.

[2] J. B. Kadane and N. A. Lazar, "Methods and Criteria for Model Selection," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 279-290, 2004.

[3] P. L. Flom and D. L. Cassell, "Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use," 29 September 2008. [Online]. Available: http://www.lexjansen.com/pnwsug/2008/DavidCassell-StoppingStepwise.pdf.

[4] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, "Least Angle Regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407-499, 2004.

[5] SAS Institute Inc. 2013, "The GLMSELECT Procedure," in *SAS/STAT® 13.1 User's Guide*, Cary, NC, SAS Institute Inc., 2013, pp. 3706-3858.

ACKNOWLEDGMENTS

I would like to acknowledge all of my Veramed colleagues who have contributed their thoughts and experiences to this paper. I wish to particularly thank Sally Garnett, Katherine Hutchinson and Andrew Holmes for their thorough review.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Peter Williams
Veramed Ltd
Biohub, Alderley Park
Alderley Edge
Cheshire, SK10 4TG
Work Phone: 01625 238705
Email: peter.williams@veramed.co.uk
Web: http://veramed.co.uk