# Variable selection with stepwise and best subset approaches

## Zhongheng Zhang

Department of Critical Care Medicine, Jinhua Municipal Central Hospital, Jinhua Hospital of Zhejiang University, Jinhua 321000, China
*Correspondence to:* Zhongheng Zhang, MMed. 351#, Mingyue Road, Jinhua 321000, China. Email: zh_zhang1984@hotmail.com.

*Author's introduction:* Zhongheng Zhang, MMed. Department of Critical Care Medicine, Jinhua Municipal Central Hospital, Jinhua Hospital of Zhejiang University. Dr. Zhongheng Zhang is a fellow physician of the Jinhua Municipal Central Hospital. He graduated from School of Medicine, Zhejiang University in 2009, receiving Master Degree. He has published more than 35 academic papers (science citation indexed) that have been cited for over 200 times. He has been appointed as reviewer for 10 journals, including *Journal of Cardiovascular Medicine*, *Hemodialysis International*, *Journal of Translational Medicine*, *Critical Care*, *International Journal of Clinical Practice*, *Journal of Critical Care*. His major research interests include hemodynamic monitoring in sepsis and septic shock, delirium, and outcome study for critically ill patients. He is experienced in data management and statistical analysis by using R and STATA, big data exploration, systematic review and meta-analysis.

Zhongheng Zhang, MMed.

**Abstract:** While purposeful selection is performed partly by software and partly by hand, the stepwise and best subset approaches are automatically performed by software. Two R functions stepAIC() and bestglm() are well designed for stepwise and best subset regression, respectively. The stepAIC() function begins with a full or null model, and methods for stepwise regression can be specified in the direction argument with character values "forward", "backward" and "both". The bestglm() function begins with a data frame containing explanatory variables and response variables. The response variable should be in the last column. Varieties of goodness-of-fit criteria can be specified in the IC argument. The Bayesian information criterion (BIC) usually results in more parsimonious model than the Akaike information criterion.

**Keywords:** Logistic regression; interaction; R; best subset; stepwise; Bayesian information criterion

Submitted Dec 25, 2015. Accepted for publication Jan 24, 2016.
doi: 10.21037/atm.2016.03.35
**View this article at:** http://dx.doi.org/10.21037/atm.2016.03.35

## Introduction

The previous article introduces purposeful selection for regression model, which allows incorporation of clinical experience and/or subject matter knowledge into statistical science. There are several other methods for variable selection, namely, the stepwise and best subsets regression. In stepwise regression, the selection procedure is automatically performed by statistical packages. The criteria for variable selection include adjusted R-square, Akaike information criterion (AIC), Bayesian information criterion (BIC), Mallows's Cp, PRESS, or false discovery rate (1,2). Main approaches of stepwise selection are the forward selection, backward elimination and a combination of the two (3). The procedure has advantages if there are numerous potential explanatory variables, but it is also criticized for being a paradigmatic example of data dredging that significant variables may be obtained from "noise" variables (4,5). Clinical experience and expertise are not allowed in model building process.

While stepwise regression select variables sequentially, the best subsets approach aims to find out the best fit model from all possible subset models (2). If there are p covariates, the number of all subsets is $2^p$. There are also varieties of statistical methods to compare the fit of subset models. In this article, I will introduce how to perform stepwise and best subset selection by using R.

## Working example

The working example used in the tutorial is from the package MASS. You can take a look at what each variable represents for.

```
> library(MASS)
> head(bwt)
```

| | low | age | lwt | race | smoke | ptd | ht | ui | ftv |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 19 | 182 | black | FALSE | FALSE | FALSE | TRUE | 0 |
| 2 | 0 | 33 | 155 | other | FALSE | FALSE | FALSE | FALSE | 2+ |
| 3 | 0 | 20 | 105 | white | TRUE | FALSE | FALSE | FALSE | 1 |
| 4 | 0 | 21 | 108 | white | TRUE | FALSE | FALSE | TRUE | 2+ |
| 5 | 0 | 18 | 107 | white | TRUE | FALSE | FALSE | TRUE | 0 |
| 6 | 0 | 21 | 124 | other | FALSE | FALSE | FALSE | FALSE | 0 |

The *bwt* data frame contains 9 columns and 189 rows. The variable *low* is an indicator variable with "0" indicates birth weight >2.5 kg and "1" indicates the presence of low birth weight. *Age* is mother's age in years. The variable *lwt* is mothers' weight in pounds. *Race* is mother's race and *smoke* is smoking status during pregnancy. The number of previous premature labor is *plt*. Other information includes history of hypertension (*ht*), presence of uterine irritability (*ui*), and the number of physician visits during the first trimester (*ftv*).

## Stepwise selection

We can begin with the full model. Full model can be denoted by using symbol "." on the right hand side of formula.

```
> full <- glm(low ~ ., family = binomial, data = bwt)
> summary(full)
Call:
glm(formula = low ~ ., family = binomial, data = bwt)
Deviance Residuals:
```

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -1.7038 | -0.8068 | -0.5008 | 0.8835 | 2.2152 |

Coefficients:

| Estimate | Std.Error | z | value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 0.82302 | 1.24471 | 0.661 | 0.50848 |
| age | -0.03723 | 0.03870 | -0.962 | 0.33602 |
| lwt | -0.01565 | 0.00708 | -2.211 | 0.02705* |
| raceblack | 1.19241 | 0.53597 | 2.225 | 0.02609* |
| raceother | 0.74069 | 0.46174 | 1.604 | 0.10869 |
| smokeTRUE | 0.75553 | 0.42502 | 1.778 | 0.07546. |
| ptdTRUE | 1.34376 | 0.48062 | 2.796 | 0.00518** |
| htTRUE | 1.91317 | 0.72074 | 2.654 | 0.00794** |
| uiTRUE | 0.68019 | 0.46434 | 1.465 | 0.14296 |
| ftv1 | -0.43638 | 0.47939 | -0.910 | 0.36268 |
| ftv2+ | 0.17901 | 0.45638 | 0.392 | 0.69488 |

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 195.48  on 178  degrees of freedom
AIC: 217.48
Number of Fisher Scoring iterations: 4
```

As you can see in the output, all variables except *low* are included in the logistic regression model. Variables *lwt*, *race*, *ptd* and *ht* are found to be statistically significant at conventional level. With the full model at hand, we can

 

begin our stepwise selection procedure.

```
> step <- stepAIC(full, trace = FALSE)
> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
low ~ age + lwt + race + smoke + ptd + ht + ui + ftv

Final Model:
low ~ lwt + race + smoke + ptd + ht + ui

   Step  Df   Deviance  Resid. Df Resid. Dev   AIC
1                       178       195.4755    217.4755
2 - ftv  2    1.358185  180       196.8337    214.8337
3 - age  1    1.017866  181       197.8516    213.8516
```

All arguments in the stepAIC() function are set to default. If you want to set direction of stepwise regression (e.g., backward, forward, both), the direction argument should be assigned. The default is both.

```
> forward<-stepAIC(full, direction="forward",trace = FALSE)
> forward$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
low ~ age + lwt + race + smoke + ptd + ht + ui + ftv

Final Model:
low ~ age + lwt + race + smoke + ptd + ht + ui + ftv

  Step  Df  Deviance  Resid.Df  Resid.Dev  AIC
1                     178       195.4755   217.4755
```

Because the forward stepwise regression begins with full model, there are no additional variables that can be added. The final model is the full model. Forward selection can begin with the null model (incept only model).

```
> backward<-stepAIC(full, direction="backward",trace =
FALSE)
> backward$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
low ~ age + lwt + race + smoke + ptd + ht + ui + ftv

Final Model:
low ~ lwt + race + smoke + ptd + ht + ui

   Step  Df   Deviance  Resid. Df Resid. Dev   AIC
1                       178       195.4755    217.4755
2 –ftv   2    1.358185  180       196.8337    214.8337
3 –age   1    1.017866  181       197.8516    213.8516
```

The backward elimination procedure eliminated variables *ftv* and *age*, which is exactly the same as the "both" procedure.

Different criteria can be assigned to the stepAIC() function for stepwise selection. The default is AIC, which is performed by assigning the argument k to 2 (the default option).

```
> BIC<-stepAIC(full,k=log(nrow(bwt))) #BIC method
```

The stepAIC() function also allows specification of the range of variables to be included in the model by using the *scope* argument. The lower model is the model with smallest number of variables and the upper model is the largest possible model. Both upper and lower components of scope can be explicitly specified. If scope is a single formula, it specifies the upper component, and the lower model is empty. If scope is missing, the initial model is the upper model.

```
> scope<-stepAIC(full,scope=list(lower=~smoke+age,upper=f
ull),trace=FALSE)
> scope$anova
Stepwise Model Path
Analysis of Deviance Table
Initial Model:
low ~ age + lwt + race + smoke + ptd + ht + ui + ftv

Final Model:
low ~ age + lwt + race + smoke + ptd + ht + ui

   Step  Df   Deviance  Resid. Df Resid. Dev   AIC
1                       178       195.4755    217.4755
```

```
2 - ftv     2     1.358185       180        196.8337       214.8337
```

When I specify the smallest model to include *age* variable, it will not be excluded by stepwise regression (e.g., otherwise, the age will be excluded as shown above). This function can help investigators to keep variables that are considered to be relevant by subject-matter knowledge. Next, we can have more complicated model for stepwise selection.

```
> step2 <- stepAIC(full, ~ .^2 + I(scale(age)^2)+ I(scale(lwt)^2),
trace = FALSE)
> step2$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
low ~ age + lwt + race + smoke + ptd + ht + ui + ftv

Final Model:
low ~ age + lwt + smoke + ptd + ht + ui + ftv + age:ftv + smoke:ui

  Step          Df   Deviance   Resid. Df   Resid. Dev   AIC
1                                 178        195.4755    217.4755
2 +age:ftv      2    12.474896    176        183.0006    209.0006
3 +smoke:ui     1    3.056805     175        179.9438    207.9438
4 -race         2    3.129586     177        183.0734    207.0734
```

Recall that "^" symbol denotes interactions up to a specified degree. In our case, we specified two-degree interactions among all possible combinations of variables. Elements within I() are interpreted arithmetically. The function scale(age) centers variable age at its mean and scales it by standard deviation. "~ .^2 + I(scale(age)^2)+ I(scale(lwt)^2)" is the scope argument and a single formula implies the upper component. The results show that the interaction between *age* and *ftv*, *smoke* and *ui* are remained in the final model. Other interactions and quadratic terms are removed.

## Best subset regression

Best subset regression selects the best model from all possible subsets according to some goodness-of-fit criteria. This approach has been in use for linear regression for several decades with the branch and bound algorithm (6). Later on, lawless and Singhal proposed an extension that can be used for non-normal error model (7). The application of best subsets for logistic regression model was described by Hosmer and coworkers (8). An R package called "bestglm" contains functions for performing best subsets selection. The bestglm() employs simple exhaustive searching algorithm as described by Morgan (9).

```
> library(leaps) # bestglm requires installation of leaps pack-
age
> library(bestglm)
> args(bestglm)
function (Xy, family = gaussian, IC = "BIC", t = "default", CVArgs
= "default",
    qLevel = 0.99, TopModels = 5, method = "exhaustive", inter-
cept = TRUE,
   weights = NULL, nvmax = "default", RequireFullEnumerationQ
= FALSE,
   ...)
```

Xy is a data frame containing independent variables and response variable. For logistic regression model when family is set to be binomial, the last column is the response variable. The sequence of Xy is important because a formula to specify response and independent variables are not allowed with bestglm() function. We can move the response variable low to the last column and assign a new name to the new data frame. The IC argument specifies the information criteria to use. Its values can be "AIC", "BIC", "BICg", "BICq", "LOOCV" and "CV" (10).

```
> bwt.move<-bwt[,-1]
> bwt.move$low<-bwt$low
```

Furthermore, factors with more than two levels should be converted to dummy variables. Otherwise, it returns an error message.

```
> library(dummies)
> race<-data.frame(dummy(bwt$race)[,c(1,2)])
> ftv<-data.frame(dummy(bwt$ftv)[,c(2,3)])
> bwt.dummy<-bwt[,-c(1,4,9)]
> low<-bwt$low
> bwt.dummy<-cbind(bwt.dummy,race,ftv, low)
```

To create dummy variables for factors with more than two levels, we use the *dummies* package. The dummy()

function passes a single variable and returns a matrix with the number of rows equal to that of given variable, and the number of columns equal to the number of levels of that variable. Because only n-1 dummy variables are needed to define a factor with n levels, I remove the base level by simple manipulation of vectors. Finally, a new data frame containing dummy variables is created, with the response variable in the last column.

```
> library(bestglm)
> bestglm(bwt.dummy,IC="BIC",family=binomial)
Morgan-Tatar search since family is non-gaussian.
BIC
BICq equivalent for q in (0.420012802643247,
0.585022045845753)
Best Model:
             Estimate     Std. Error   z value    Pr(>|z|)
(Intercept)  1.01736672   0.85333672   1.192222   0.233174251
lwt          -0.01728027  0.00678715   -2.546028  0.010895659
ptdTRUE      1.40676981   0.42850088   3.283003   0.001027075
htTRUE       1.89397147   0.72108967   2.626541   0.008625764
```

The model selection by AIC always keeps more variables in the model as follows.

```
> bestglm(bwt.dummy,IC="AIC",family=binomial)
Morgan-Tatar search since family is non-gaussian.
AIC
BICq equivalent for q in (0.797717473187408,
0.882261427360355)
Best Model:
             Estimate     Std.Error    zvalue     Pr(>|z|)
(Intercept)  0.64059802   0.859552154  0.7452695  0.456108807
lwt          -0.01424899  0.006583754  -2.1642657 0.030443965
smokeTRUE    0.92821842   0.398653203  2.3283857  0.019891632
ptdTRUE      1.12007778   0.450882008  2.4841927  0.012984553
htTRUE       1.85222596   0.705829936  2.6241816  0.008685745
uiTRUE       0.73543662   0.461731565  1.5927796  0.111209644
race.white   -1.01303877  0.396054355  -2.5578276 0.010532828
```

Readers can try other options available in bestglm() function. Different options may result in different models.

## Summary

The article introduces variable selection with stepwise and best subset approaches. Two R functions stepAIC() and bestglm() are well designed for these purposes. The stepAIC() function begins with a full or null model, and methods for stepwise regression can be specified in the direction argument with character values “forward”, “backward” and “both”. The bestglm() function begins with a data frame containing explanatory variables and response variables. The response variable should be in the last column. Factor variables with more than two levels should be converted before running bestglm(). The dummies package contains good function to convert factor variable to dummy variables. There are varieties of information criteria for selection of the best subset model.

## Acknowledgements

None.

## Footnote

*Conflicts of Interest:* The author has no conflicts of interest to declare.

## References

1. Hocking RR. A biometrics invited paper. The analysis and selection of variables in linear regression. Biometrics 1976;32:1-49.
2. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied Logistic Regression. Hoboken: John Wiley & Sons, Inc, 2013.
3. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag New York, 2013.
4. Freedman DA. A note on screening regression equations. The American Statistician 1983;37:152-5.
5. Flack VF, Chang PC. Frequency of selecting noise variables in subset regression analysis: a simulation study. The American Statistician 1987;41:84-6.
6. Furnival GM, Wilson RW. Regressions by leaps and bounds. Technometrics 1974;16:499-511.
7. Lawless JF, Singhal K. ISMOD: an all-subsets regression program for generalized linear models. II. Program guide and examples. Comput Methods Programs Biomed 1987;24:125-34.
8. Hosmer DW, Jovanovic B, Lemeshow S. Best subsets logistic regression. Biometrics 1989;45:1265-70.
9. Morgan JA, Tatar JF. Calculation of the residual sum

 

of squares for all possible regressions. Technometrics 1972;14:317-25.
10. McLeod AI, Changjiang X. bestglm: best subset GLM.– R package ver. 0.34. 2014. Available online: https://cran.r-project.org/web/packages/bestglm/bestglm.pdf

**Cite this article as:** Zhang Z. Variable selection with stepwise and best subset approaches. Ann Transl Med 2016;4(7):136. doi: 10.21037/atm.2016.03.35

 