**Krista Miller**

**Part 1: Write a short description of the following three methods of automated model selection.**

Forward selection:

Forward selection is a procedure to build a model by starting with an empty model and then iterating over all features. Each feature is iteratively added by the most contributive predictors, and then stops when the improvement is no longer statistically significant.

Backward selection:

This process is an inverse of the forward selection process. It starts with all the predictors in the model, iteratively removes the least contributive predictors, and stops at a model where all the predictors are statistically significant.

\*Stepwise selection is also discussed in the asynch. You can go back and forth with predictors. A predictor can be added or dropped at later steps.

Best subsets selection:

This process consists of testing all possible combinations of the predictor variables, and then selecting the best model.

Testing models of each possible size: all 1 predictor models, all 2 predictor models, all 3 predictor models, etc…. doesn’t necessarily include interaction terms as predictors (but you could). You could also include squares and cubes in pool of potential predictors as well.

**Part 2: Conducting automated model selection in R (forward and backward)**

Look at the documentation for the step() function, which doesn’t require any special packages to run in RStudio, and answer the questions.

1) The first argument shown in the step() function documentation is called “object”. This is a required argument. What goes here when conducting a forward selection? What goes here when conducting a backward selection?

Object is the initial model used in the stepwise search.

For forward selection:

One method is by first creating a null model:

lm.null <- lm(y ~ 1, data= my\_dataframe)

and then the model can be called:

model.aic.forward<- step(lm.null, direction = ‘forward’, scope= ~x1 + x2 + x3 +…)

Another method is:

Fwd.model<- step(lm(y ~1, data=my.data\_frame), direction= ‘forward’, scope= ~ x1 + x2+x3+…)

For backward selection:

One method is by creating a full model:

lm.full<- lm(y ~ ., data= my\_dataframe)

and then calling the model:

model.aic.backward<- step(lm.full, direction = ‘backward’)

Another method is:

Bkwd.model<- step(lm(y ~ ., data= my\_dataframe), direction = ‘backward’)

2) The second argument shown in the step() function documentation is called “scope”. This is a required argument. What goes here when conducting a forward selection? What goes here when conducting a backward selection?

Scope is a formula that specifies which predictors we want to attempt to enter into the model. It can also be a list including two formulas (lower and upper), specifying the models considered in the step function. It defaults to NULL.

Here are several equivalent ways to select the scope for forward selection:

lm.null <- lm(y ~ 1, data= my\_dataframe)

model.aic.forward<- step(lm.null, direction = ‘forward’, scope= ~x1 + x2 + x3 +…)

or:

model.aic.forward<- step(lm.null, direction= ‘forward’, scope= formula(all))

or:

model.aic.foward<- step(lm.null, direction= ‘forward’, scope= list(lower=lm.null, upper=~.)

In terms of backward selection:

One could start with the full model and use the default null value for scope:

lm.full<- lm(y ~ ., data= my\_dataframe)

model.aic.backward<- step(lm.full, direction = ‘backward’)

or explicitly call it:

model.aic.backward<- step(lm.full, direction= ‘backward’, scope=formula(all))

3) The fourth argument shown in the step() function documentation is called “direction”. What are your three choices here? What is the default if you leave this argument out?

The R documentation describes the direction argument for the step() function as: a character string describing the direction of stepwise algorithm. Directions that can be selected are “both”, “backward”, or “forward”. This defaults to “both”. \*If no scope argument is provided, the default is “backward”.\*

4) The fifth argument shown in the step() function documentation is called “trace”. What does this do if trace=1? What does it do when trace=0?

If the argument trace = TRUE or trace = 1, R displays the single steps during the stepwise procedure. If trace = FALSE or trace = 0, it tells R not to display the full results of the stepwise selection. I read that trace can take up quite a bit of space if there are a large amount of predictor variables. This argument defaults to TRUE.

\*\* use trace=1 in this course \*\*

5) By default, what criterion is used to determine whether variables are added or dropped from the model? By default, the AIC is used to determine whether variables are added or dropped from the model.

Adding features:

First, we fit the intercept-only model and get an AIC value.

Next, we fit every possible one-predictor model. The model that procedud the lowest AIC AND also had a statistically significant reduction in AIC compared to the intercept-only model is then selected. ANOVA is used to determine statistical significance. Stops when no variable produces an F-ratio greater than the 90th or 95th percentile of the F1,n-k-2 distribution.

We continue this process until we cannot create a model that produces a significant reduction in AIC.

#view results of forward stepwise regression:

model.aic.forward$anova

#view final model:

model.aic.forward$coefficients

Dropping features:

First, we fit a model using all predictors.

Next, for k=p, p-1, …1, we fit all k models that contain all but one of the predictors in the original model, for a total of k-1 predictor variables.

Lastly, we select a single best model using AIC. Stops when each variable in the model produces an F-ratio greater than the 90th or 95th percentile of the F1,n-k-2.

#view results of backward stepwise regression:

model.aic.backward$anova

#view final model:

model.aic.backward$coefficients

**Part 3: Conducting automated model selection in R (best subsets)**

Look at the documentation for the regsubsets() function, which is a function from the leaps package, and answer the questions.

The regsubsets() function performs best subset selection by identifying the best model that contains a given number of predictors, where best is quantified using RSS. The model selection is by exhaustive search, forward or backward stepwise, or sequential replacement.

1) The first argument shown in the regsubset() function is “x”. What goes here?

X= the design matrix or model formula for full model, or ‘biglm’ object.

The data, in a specific form.

2) The third argument shown in the regsubset() function is “y”. What goes here?

Y= response vector

Column of outcomes

3) The sixth argument shown in the regsubset() function is “nvmax”. What does this do? What is the default value of nvmax if the argument isn’t included?

Nvmax= maximum size of subsets to examine. By default, regsubsets() only reports results up to the best eight-variable model. The nvmax option can be used in order to return as many variables as are desired.

4) The tenth argument shown in the regsubset() function is “method”. If you want to conduct best subsets selection, which option should you choose?

Method= use exhaustive search, forward selection, backward selection, or sequential replacement to search. method=c("exhaustive", "backward", "forward", "seqrep"). If you want to conduct the best subsets selection, you should choose ‘exhaustive’.

Method will always = exhaustive in this course.

Helpful resources used for this assignment:

<https://www.statology.org/stepwise-regression-r/>

<http://www.science.smith.edu/~jcrouser/SDS293/labs/lab8-r.html>