Model Selection Using Stepwise Function

Pranav Gopalkrishna, 1940223

2021-09-02

# **Setting work directory**

setwd("~/Documents/Study/computerScience/programming/r/data/")

# Data set and purpose of using it

The data set used here does not have an apparent practical use, as its fields are simply one generic response variable with nine generic independent variables, i.e. y, x1, x2... x9.

The purpose of working with this data set is to figure out the best fitting model for the response y among the possible models that can be made using the available regressors. Hence, we need to decide which regressors must be selected, and how they must be fitted to the data i.e. how should y be modelled using these selected regressors. Note that ultimately, we are comparing the possible models, to find out which one is best.

head(myData)

## y x1 x2 x3 x4 x5 x6 x7 x8 x9  
## 1 25.9 4.9176 1 3.472 0.998 1 7 4 42 0  
## 2 29.5 5.0208 1 3.531 1.500 2 7 4 62 0  
## 3 27.9 4.5429 1 2.275 1.175 1 6 3 40 0  
## 4 25.9 4.5573 1 4.050 1.232 1 6 3 54 0  
## 5 29.9 5.0597 1 4.455 0.988 1 6 3 56 0  
## 6 30.9 5.8980 1 5.850 1.240 1 7 3 51 1

# Model with all regressors (raw model)

A full model with all the available regressors is to be created. The best fitting model will be created using this raw model as the source for drawing and studying the available regressors.

rawModel = lm(y~., data = myData)  
summary(rawModel)

##   
## Call:  
## lm(formula = y ~ ., data = myData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.8504 -1.4017 0.0929 1.7541 3.7206   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.11351 5.88549 2.908 0.0131 \*  
## x1 2.39009 1.05740 2.260 0.0432 \*  
## x2 5.74422 4.35113 1.320 0.2114   
## x3 0.12998 0.52530 0.247 0.8087   
## x4 2.63623 4.34493 0.607 0.5553   
## x5 2.32382 1.46160 1.590 0.1378   
## x6 -1.62471 2.40137 -0.677 0.5115   
## x7 -0.09723 3.38794 -0.029 0.9776   
## x8 -0.04445 0.06212 -0.716 0.4879   
## x9 2.03656 1.97372 1.032 0.3225   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.841 on 12 degrees of freedom  
## Multiple R-squared: 0.8774, Adjusted R-squared: 0.7854   
## F-statistic: 9.539 on 9 and 12 DF, p-value: 0.0003125

# Step function

## Purpose

This function chooses the best regression model using the AIC stepwise model selection algorithm. Best in this context means the model that is has the regressors and coefficients that best explain or match the responses, given the data. Hence, it is not only best fitting for a given set of regressors, it is also best fitting among all possible models using the available regressors.

*(AIC stands for* ***Akaike’s information criterion****. It is a stepwise model selection method compares the quality of a set of statistical models to each other. The lower the AIC value, the better fitting the model)*

## Usage

The following only presents the range of options we will be using for this function. There are more options, however.

step( object,  
 scope,  
 direction = c("both", "backward", "forward"))

**Argument “object”** is an object representing a model of an appropriate class (mainly “lm” and “glm”). This is used as the initial model in the stepwise search (variable selection) for the best regressors for modelling the given response. Initial model implies the model with the response and an initial set of regressors and coefficients on top of which more regressors will be added. Typically, it is a model with only the response, intercept and error term.

**Argument “scope”** defines the range of models examined in the stepwise search. It holds the model or models containing the different regressors that may be selected for the final model returned by the function. This option could contain a single model, or two models “lower” and “upper”, wherein the regressors in the lower model are a subset of the regressors in the upper model. In the case of “lower” and “upper” models, the step function performs a stepwise search for every model from the lower to the upper (and the models in between, with respect to rhe regressors used).

**Argument "direction"** the mode of stepwise search, can be one of "both", "backward", or "forward", with a default of "both". If the scope argument is missing the default for direction is "backward. Backward implies that we start with all regressors, and keep removing insignificant regressors (if removing them improves the fit of the model) to arrive at the most explanatory i.e. best fitting model.

# Backward selection and initial model

Since this is backward selection, the initial model is the full model. Backward selection, unlike forward selection, starts with all regressors and eliminates insignificant regressors (if removing them improves the fit of the model) to arrive at the most explanatory i.e. best fitting model.

# Performing stepwise search

step(rawModel, direction = "backward")

## Start: AIC=52.61  
## y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9  
##   
## Df Sum of Sq RSS AIC  
## - x7 1 0.007 96.868 50.611  
## - x3 1 0.494 97.355 50.721  
## - x4 1 2.971 99.832 51.274  
## - x6 1 3.695 100.556 51.433  
## - x8 1 4.133 100.994 51.528  
## - x9 1 8.594 105.455 52.479  
## <none> 96.861 52.609  
## - x2 1 14.068 110.929 53.593  
## - x5 1 20.404 117.265 54.815  
## - x1 1 41.240 138.101 58.413  
##   
## Step: AIC=50.61  
## y ~ x1 + x2 + x3 + x4 + x5 + x6 + x8 + x9  
##   
## Df Sum of Sq RSS AIC  
## - x3 1 0.488 97.356 48.721  
## - x4 1 2.983 99.850 49.278  
## - x8 1 4.613 101.480 49.634  
## <none> 96.868 50.611  
## - x9 1 12.927 109.795 51.367  
## - x6 1 14.410 111.277 51.662  
## - x2 1 15.494 112.362 51.875  
## - x5 1 21.721 118.588 53.062  
## - x1 1 55.163 152.031 58.527  
##   
## Step: AIC=48.72  
## y ~ x1 + x2 + x4 + x5 + x6 + x8 + x9  
##   
## Df Sum of Sq RSS AIC  
## - x4 1 4.954 102.310 47.813  
## - x8 1 5.101 102.457 47.845  
## <none> 97.356 48.721  
## - x6 1 14.955 112.311 49.865  
## - x2 1 15.330 112.686 49.938  
## - x9 1 19.445 116.800 50.727  
## - x5 1 21.698 119.054 51.148  
## - x1 1 77.178 174.534 59.564  
##   
## Step: AIC=47.81  
## y ~ x1 + x2 + x5 + x6 + x8 + x9  
##   
## Df Sum of Sq RSS AIC  
## - x8 1 5.452 107.76 46.955  
## <none> 102.31 47.813  
## - x6 1 10.963 113.27 48.053  
## - x9 1 17.202 119.51 49.232  
## - x5 1 21.654 123.96 50.037  
## - x2 1 31.807 134.12 51.769  
## - x1 1 88.167 190.48 59.487  
##   
## Step: AIC=46.96  
## y ~ x1 + x2 + x5 + x6 + x9  
##   
## Df Sum of Sq RSS AIC  
## <none> 107.76 46.955  
## - x9 1 13.719 121.48 47.592  
## - x5 1 21.365 129.13 48.935  
## - x6 1 25.110 132.87 49.564  
## - x2 1 28.366 136.13 50.096  
## - x1 1 202.251 310.01 68.203

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x5 + x6 + x9, data = myData)  
##   
## Coefficients:  
## (Intercept) x1 x2 x5 x6 x9   
## 16.182 3.055 6.364 2.195 -1.837 1.823

From the function’s results, we see that x1, x2, x5, x6 and x9 are the best regressors for our response y, with the given coefficients leading to the best fitting model possible for the data and available regressors.

# Final model (best model, based on this method)

cookedModel = lm(y~x1+x2+x5+x6+x9, data = myData)  
summary(cookedModel )

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x5 + x6 + x9, data = myData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.8842 -1.5551 0.1727 1.5600 3.5507   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.1817 4.5945 3.522 0.00283 \*\*   
## x1 3.0551 0.5575 5.480 5.04e-05 \*\*\*  
## x2 6.3639 3.1010 2.052 0.05688 .   
## x5 2.1949 1.2324 1.781 0.09389 .   
## x6 -1.8373 0.9516 -1.931 0.07141 .   
## x9 1.8231 1.2774 1.427 0.17274   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.595 on 16 degrees of freedom  
## Multiple R-squared: 0.8636, Adjusted R-squared: 0.8209   
## F-statistic: 20.25 on 5 and 16 DF, p-value: 2.091e-06

# Conclusions

Hence, we see that the best fitting model i.e. the most explanatory model (using backward elimination method) for the response variable is given by

*y = 16.182 + 3.055x1 + 6.364x2 + 2.195x5 - 1.837x6 + 1.823x9*

Based on the summary, only the intercept and x1 are significant given a 0.05 significance level, but not the rest. However, the other variables increase the fit of the model to the best level, hence they are retained. We also see that see that the adjusted R-squared value is 82.09, much greater than the full model's adjusted R-squared value 78.54. This means a greater majority (greater than the full model) of the variation in the response in the sample is explained by this model.