Variable Selection using leaps package by Amelia M

Ismail Khalil

6/28/2021

```
library(tidyverse)
library(Stat2Data)
library(skimr)
```

Loading the data:

```
data(FirstYearGPA)
skim(FirstYearGPA)
```

It is important to know that there are no "best" multiple regression model. Each model has its positive and negative features, and it is our role as data analyst to construct a model that is effective for our objective. This could mean passing by a model with a higher R^2 versus one that has a lower R^2 but has a much easier concept. Or on the contrary, adding polynomial terms that explain only a little bit of extra variation in the response.

There are several algorithms for optimizing a given criterion, we will work with the following procedures:

- Best subset Selection
- Backwards Eliminaion

- Forward Selection
- StepWise Regression

Each of these techniques can optimize a different criterion. There is no universal agreed-upon best method or criterion so far, however the following are the most used:

- [Adjusted R^2]
- [Residual Sum of Squares (RSS)]
- [Akaike's Information Criterion (AIC)]
- [Bayes Information Criterion (BIC)]

```
library(leaps)
# Report the two best models for each number of predictors

best <- regsubsets(GPA ~., data = FirstYearGPA, nbest = 2 , method = "exhaustive")
with(summary(best), data.frame(rsq, adjr2, cp, rss, outmat))
```

Best Subset Selection Based on R^2, it is not a very good criterion because the value keep increasing as we add on more variables(so of course we are going to choose the model with all the variables), so let's see the AdjustedR^2. Based on adjustedR2, the first model of size 6 seems to be the optimal

Based on C_p, we should pick the first model of size 5.

I'm going to investigate model 5 just to test a few things. Which has the following significant variables:

```r
model1 <- lm(GPA ~ HSGPA+SATV+HU+SS+White, data = FirstYearGPA)
summary(model1)
AIC(model1)
```

Note that the var. SS is not significant but will be selected according to the following models. #### Backward Elimination

Backward Elimination is a simple algorithn that begins by throwing all the terms into the model, and then greedily removing the ones that are least statistically significant

```r
BWD <- regsubsets(GPA~., data = FirstYearGPA, nbest = 1 , nvmax = 6, method = "backward")
summary(BWD)
with(summary(BWD), data.frame(cp, outmat))
```

Based on C_p We will choose the model of size 5, which drops the variable Male.

```r
model1 <- lm(GPA ~ HSGPA + SATV + HU + SS + White, data = FirstYearGPA)
AIC(model1)
```

Forward Selection Opposite idea of Backward selection. Here we begin on an empty model and greedily adding terms that have a statistical significant effect.

```r
FWD <- regsubsets(GPA~., data = FirstYearGPA, nbest = 1 , nvmax = 6 , method = "forward")

summary(FWD)
with(summary(FWD), data.frame(cp, outmat))
```

Model of size 5 again seems to be the best. Same variable we found using backward.

```r
model2 <- lm(GPA ~ HSGPA+SATV+HU+SS+White, data = FirstYearGPA)
AIC(model2)
```

Stepwise regression Combines both the concepts of backward elimination and forward selection to move in both directions.

```r
stepwise <- regsubsets(GPA~. ,data = FirstYearGPA, nbest=1, nvmax=6, method = "seqrep")
with(summary(stepwise),data.frame(cp, outmat))
```

```r
model3 <- lm(GPA ~ HSGPA+SATV+Male+HU+SS, data = FirstYearGPA)
AIC(model3)
```

Conclusion There is no best model, but if we have to pick we would avoid model which has 2 non significant variables but again this can change greatly if we use other models or just use another sample from the same data.