Automated Variable Selection Procedures in R

Best subsets proceedure in R

To obtain a best subsets analysis for each possible number of predictor variables (nbest specifies how many of such top models you want to consider), we use a function in the `leaps` package called `regsubsets()`.

```
library(leaps)
all <- regsubsets(Response ~ ., nbest = 2, data=data_object)
```

Notice here we use a shortcut to specifying the model:`Response ~ .` tells R to regress the response variable on all of the predictors in the data object. R will automatically assume each column besides `Response` is a valid predictor if you use this shortcut.

We see the results of the `regsubsets()` command with

```
all %>% summary
```

It may be more informative to create a plot that shows which variables are in each model:

```
plot(all, scale = "adjr2")
```

If we wanted to rank the models by Mallow's Cp rather than by adjusted R2, then we use:

```
plot(all, scale = "Cp")
```

Backwards ellimination procedure in R

To use backward elimination, we use the `step()` function. We start by fitting the model with all predictors, and then we feed that model into the `step()` function.

```
full <- lm(GPA ~ ., data = FirstYearGPA)

step(full, direction = backward)
```

If we want to conduct backward elimination but with Cp rather than AIC as the criterion used to decide which predictor to drop at each step, then we need to calculate MSE and include it in the `step()` command.

```
MSE <- (summary(full)$sigma^2)

step(full, scale=MSE, direction="backward")
```

Stepwise selection procedure in R

To use stepwise regression, we fit the model with no predictors (just the constant, 1). Then we tell the `step()` command to start with that model (called `none` here) and to look for models up to and including the full model.

```
full <- lm(GPA ~ ., data=FirstYearGPA)
none <- lm(GPA ~ 1, data=FirstYearGPA)

MSE <- (summary(full)$sigma^2)

step(none, scope=formula(full), scale=MSE)
```