Variable selection: http://www.stat.columbia.edu/~martin/W2024/R10.pdf

There are two main approaches towards variable selection: the all possible

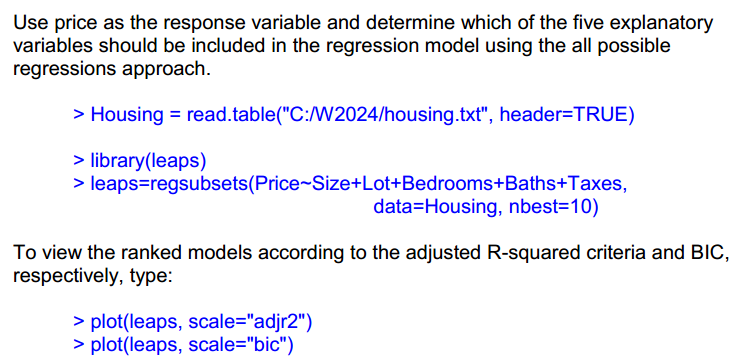
regressions approach and automatic methods.

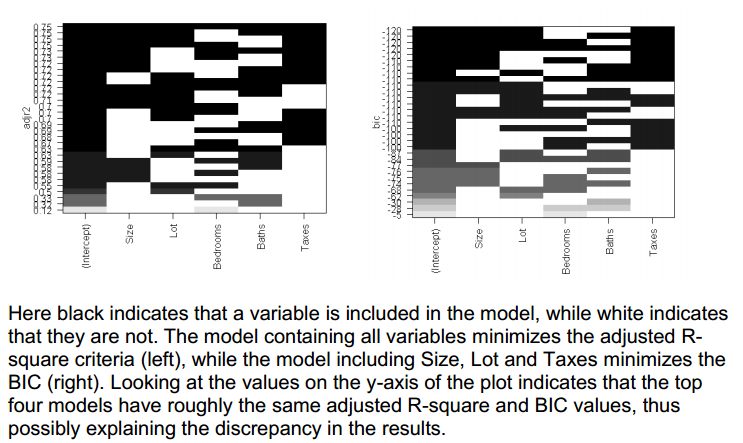
According to Adjusted R2 , AIC and BIC

The function regsubsets() in the library “leaps” can be used for regression subset

selection. Thereafter, one can view the ranked models according to different

scoring criteria by plotting the results of regsubsets().





About Stepwise:

Akaike information criterion: AIC = 2k - 2 log L = 2k + Deviance, where k = number of parameters

Small numbers are better

Penalizes models with lots of parameters

Penalizes models with poor fit

This example uses linear models, in logistic models, we should change lm into glm and add family=binomial (link=’logit’) like:

model <- glm(Survived ~.,family=binomial(link='logit'),data=train)

> null=lm(Price~1, data=Housing)

> null

> full=lm(Price~., data=Housing)

> full

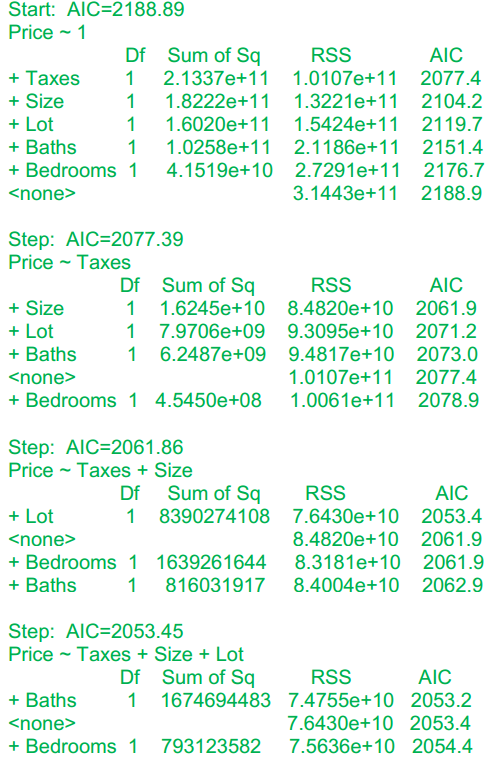
**forward selection**

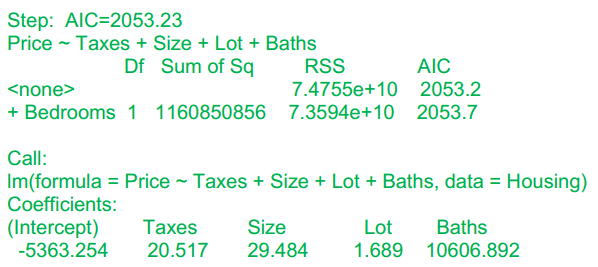
> step(null, scope=list(lower=null, upper=full), direction="forward")

This tells R to start with the null model and search through models lying in the

range between the null and full model using the forward selection algorithm. It

gives rise to the following output:





According to this procedure, the best model is the one that includes the variables

Taxes, Size, Lot and Baths.

**Backward selection:**

> step(full, data=Housing, direction="backward")

**stepwise selection:**

> step(null, scope = list(upper=full), data=Housing, direction="both")

There is another example in detail:

http://www.utstat.toronto.edu/~brunner/oldclass/appliedf11/handouts/2101f11StepwiseLogisticR.pdf