Variable Selection

In our discussion of regression to date we have assumed that all the explanatory variables included in the model are chosen in advance. However, in many situations the set of explanatory variables to be included is not predetermined and selecting them becomes part of the analysis.

There are two main approaches towards variable selection: the all possible regressions approach and automatic methods.

The all possible regressions approach considers all possible subsets of the pool of explanatory variables and finds the model that best fits the data according to some criteria (e.g. Adjusted R^2, AIC and BIC). These criteria assign scores to each model and allow us to choose the model with the best score.

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().

Before using the function for the first time you will need to install the library using the R GUI. Alternatively, you can use the command install.packages("leaps") to install it.

Ex. Data was collected on 100 homes recently sold in a city. It consisted of the sales price (in $), house size (in square feet), the number of bedrooms, the number of bathrooms, the lot size (in square feet) and annual real estate tax (in $).

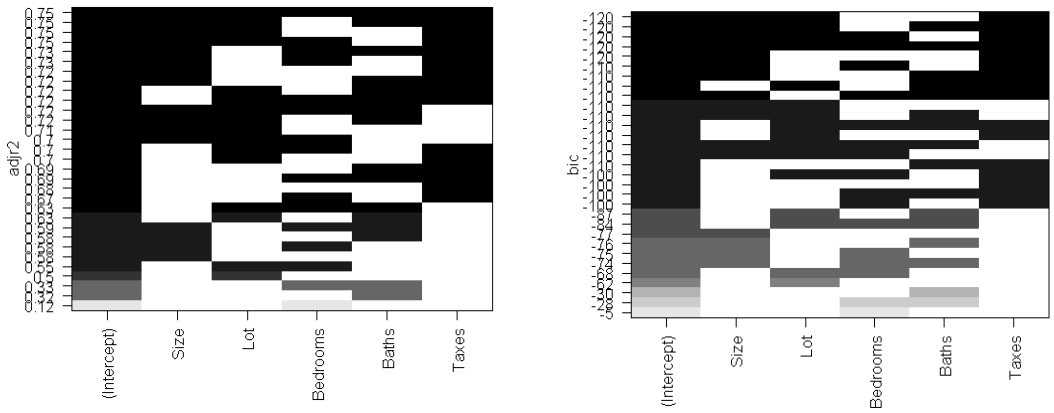
Use price as the response variable and determine which of the five explanatory variables should be included in the regression model using the all possible regressions approach.

```
> Housing = read.table("C:/W2024/housing.txt", header=TRUE)

> library(leaps)
> leaps=regsubsets(Price~Size+Lot+Bedrooms+Baths+Taxes,
                   data=Housing, nbest=10)
```

To view the ranked models according to the adjusted R-squared criteria and BIC, respectively, type:

```
> plot(leaps, scale="adjr2")
> plot(leaps, scale="bic")
```



Here black indicates that a variable is included in the model, while white indicates that they are not. The model containing all variables minimizes the adjusted R-square criteria (left), while the model including Size, Lot and Taxes minimizes the BIC (right). Looking at the values on the y-axis of the plot indicates that the top four models have roughly the same adjusted R-square and BIC values, thus possibly explaining the discrepancy in the results.

Automatic methods are useful when the number of explanatory variables is large and it is not feasible to fit all possible models. In this case, it is more efficient to use a search algorithm (e.g., Forward selection, Backward elimination and Stepwise regression) to find the best model.

The R function step() can be used to perform variable selection. To perform forward selection we need to begin by specifying a starting model and the range of models which we want to examine in the search.

```
> null=lm(Price~1, data=Housing)
> null
Call:
lm(formula = Price ~ 1, data = Housing)
Coefficients:
(Intercept)
     126698

> full=lm(Price~., data=Housing)
> full
Call:
lm(formula = Price ~ ., data = Housing)
Coefficients:
(Intercept)       Taxes     Bedrooms       Baths        Size         Lot
  6633.800       20.644    -6469.686   11824.488      33.571       1.616
```

We can perform forward selection using the command:

```
> 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:

```
Start:  AIC=2188.89
Price ~ 1
             Df  Sum of Sq       RSS     AIC
+ Taxes       1  2.1337e+11  1.0107e+11  2077.4
+ Size        1  1.8222e+11  1.3221e+11  2104.2
+ Lot         1  1.6020e+11  1.5424e+11  2119.7
+ Baths       1  1.0258e+11  2.1186e+11  2151.4
+ Bedrooms    1  4.1519e+10  2.7291e+11  2176.7
<none>                       3.1443e+11  2188.9

Step:  AIC=2077.39
Price ~ Taxes
             Df  Sum of Sq       RSS     AIC
+ Size        1  1.6245e+10  8.4820e+10  2061.9
+ Lot         1  7.9706e+09  9.3095e+10  2071.2
+ Baths       1  6.2487e+09  9.4817e+10  2073.0
<none>                       1.0107e+11  2077.4
+ Bedrooms    1  4.5450e+08  1.0061e+11  2078.9

Step:  AIC=2061.86
Price ~ Taxes + Size
             Df  Sum of Sq       RSS     AIC
+ Lot         1  8390274108  7.6430e+10  2053.4
<none>                       8.4820e+10  2061.9
+ Bedrooms    1  1639261644  8.3181e+10  2061.9
+ Baths       1   816031917  8.4004e+10  2062.9

Step:  AIC=2053.45
Price ~ Taxes + Size + Lot
             Df  Sum of Sq       RSS     AIC
+ Baths       1  1674694483  7.4755e+10  2053.2
<none>                       7.6430e+10  2053.4
+ Bedrooms    1   793123582  7.5636e+10  2054.4
```

```
Step:  AIC=2053.23
Price ~ Taxes + Size + Lot + Baths
                Df   Sum of Sq        RSS          AIC
<none>                            7.4755e+10    2053.2
+ Bedrooms  1   1160850856    7.3594e+10    2053.7

Call:
lm(formula = Price ~ Taxes + Size + Lot + Baths, data = Housing)
Coefficients:
(Intercept)       Taxes        Size           Lot       Baths
  -5363.254       20.517       29.484        1.689    10606.892
```

According to this procedure, the best model is the one that includes the variables Taxes, Size, Lot and Baths.

We can perform backward elimination on the same data set using the command:

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

and stepwise regression using the command:

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

Both algorithms give rise to results that are equivalent to the forward selection procedure in the Housing example.