# Stepwise Regression (using R)

- SPSS can be very opaque in determining how particularly statistical routines are carried out. Conversely the statistical programming language R is usually quite clear, once a familiarity with the language has been developed.
- For variable selection procedures, R used the AIC criterion. When comparing multiple candidate models, the candidate model with the lowest AIC value is the best model. We will use R output to revise variable selection procedures. Recall that we used the ***mtcars*** data set.
- The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (197374 models).
- For this data, we tried to determine the optimal set of independent variables to predict the dependent variables ***mpg*** (miles per gallon).
- The possible predictor variables are

  **cyl** Number of cylinders

  **disp** Displacement (cu.in.)

  **hp** Gross horsepower

  **drat** Rear axle ratio

  **wt** Weight (lb/1000)

  **qsec** 1/4 mile time

  **vs** V/S engine type

  **am** Transmission (0 = automatic, 1 = manual)

  **gear** Number of forward gears

  **carb** Number of carburetors

## Backward Elimination

- The initial model contains all of the independent variables. Candidate models, whereby each of the independent variables are individually removed from the model are fitted.
- The AIC value for each reduced model is computed. The unreduced model is also used for comparison. The AIC values are tabulated to determine which removal results in the lowest AIC value.
- In this first case, the removal of cyl would reduced the AIC value from 70.898 (see bottom row) to 68.915. Thus the independent variable ***cyl*** is removed from the set of independent variables.

```
Start:  AIC=70.9
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq    RSS    AIC
- cyl   1    0.0799 147.57 68.915
- vs    1    0.1601 147.66 68.932
- carb  1    0.4067 147.90 68.986
- gear  1    1.3531 148.85 69.190
- drat  1    1.6270 149.12 69.249
- disp  1    3.9167 151.41 69.736
- hp    1    6.8399 154.33 70.348
- qsec  1    8.8641 156.36 70.765
<none>              147.49 70.898
- am    1   10.5467 158.04 71.108
- wt    1   27.0144 174.51 74.280
```

- In the second phase, the process is repeated. This time removing ***vs*** results in an AIC value of 66.973. It is then removed from the set of independent variables.
- For this phase, the unreduced model is the model fitted by all independent variables except ***cyl***, which was removed in the previous phase.

```
Step:  AIC=68.92
mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq    RSS    AIC
- vs    1    0.2685 147.84 66.973
- carb  1    0.5201 148.09 67.028
- gear  1    1.8211 149.40 67.308
- drat  1    1.9826 149.56 67.342
```

```
- disp  1    3.9009 151.47 67.750
- hp    1    7.3632 154.94 68.473
<none>              147.57 68.915
- qsec  1   10.0933 157.67 69.032
- am    1   11.8359 159.41 69.384
- wt    1   27.0280 174.60 72.297
```

- This process continues until the removal of an independent variable will not results in an improvement in AIC. This is indicated by having the $< none >$ ( i.e unreduced model) having the lowest AIC value.
- At the end of the output is the optimal model, according to the backward elimination procedure, using the independent variables : ***am*** , ***qsec*** and ***wt***.

```
Step:  AIC=61.31
mpg ~ wt + qsec + am

       Df Sum of Sq    RSS    AIC
<none>              169.29 61.307
- am    1    26.178 195.46 63.908
- qsec  1   109.034 278.32 75.217
- wt    1   183.347 352.63 82.790

Call:
lm(formula = mpg ~ wt + qsec + am)

Coefficients:
(Intercept)           wt         qsec           am
      9.618       -3.917        1.226        2.936
```

## Stepwise Regression

- Stepwise Regression differs from Backward Elimination, in that it allows independent variables to be re-introduced. Hence the + signs from the second phase onwards.

```
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq    RSS    AIC
- cyl   1    0.0799 147.57 68.915
- vs    1    0.1601 147.66 68.932
- carb  1    0.4067 147.90 68.986
- gear  1    1.3531 148.85 69.190
- drat  1    1.6270 149.12 69.249
- disp  1    3.9167 151.41 69.736
- hp    1    6.8399 154.33 70.348
- qsec  1    8.8641 156.36 70.765
<none>              147.49 70.898
- am    1   10.5467 158.04 71.108
- wt    1   27.0144 174.51 74.280

Step:  AIC=68.92
mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq    RSS    AIC
- vs    1    0.2685 147.84 66.973
- carb  1    0.5201 148.09 67.028
- gear  1    1.8211 149.40 67.308
- drat  1    1.9826 149.56 67.342
- disp  1    3.9009 151.47 67.750
- hp    1    7.3632 154.94 68.473
<none>              147.57 68.915
- qsec  1   10.0933 157.67 69.032
- am    1   11.8359 159.41 69.384
+ cyl   1    0.0799 147.49 70.898
- wt    1   27.0280 174.60 72.297
```

- Again, the procedure finishes when it is found that the unchanged model has the lowest of all possible AIC values.

```
Step:  AIC=61.31
mpg ~ wt + qsec + am

       Df Sum of Sq    RSS    AIC
<none>              169.29 61.307
+ hp    1     9.219 160.07 61.515
+ carb  1     8.036 161.25 61.751
+ disp  1     3.276 166.01 62.682
+ cyl   1     1.501 167.78 63.022
+ drat  1     1.400 167.89 63.042
+ gear  1     0.123 169.16 63.284
+ vs    1     0.000 169.29 63.307
- am    1    26.178 195.46 63.908
- qsec  1   109.034 278.32 75.217
- wt    1   183.347 352.63 82.790

Call:
lm(formula = mpg ~ wt + qsec + am)

Coefficients:
(Intercept)           wt         qsec           am
      9.618       -3.917        1.226        2.936
```