9.4. Variable Selection

- Goal is to develop a model with the best set of independent variables
  - Easier to interpret if unimportant variables are removed
  - Lower probability of collinearity
- Stepwise regression procedure
  - Provide evaluation of alternative models as variables are added (or withdrawn)
- Best-subset approach
  - Try all combinations and select the best using various criteria, such as the highest adjusted R^2

```
> install.packages("olsrr")
> library(olsrr)
```

Data: mtcars

Dependent Variable: mpgMiles/(US) gallon

Independent Variables

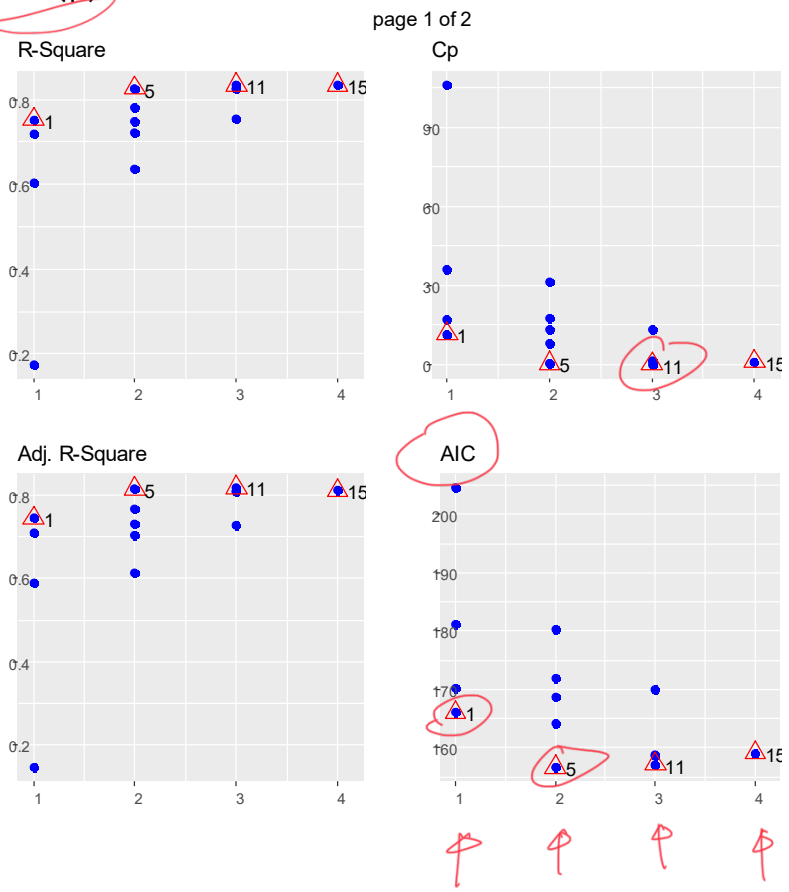
A data frame with 32 observations on 11 (numeric) variables.

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (1000 lbs)
[, 7]	qsec	1/4 mile time
[, 8]	vs	Engine (0 = V-shaped, 1 = straight)
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

All-possible subset selection

Assume we have independent variables: disp, hp, wt and qsec

```
reg <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
p <- ols_step_all_possible(reg)
p
# A table: 15 x 6
    Index N      Predictors  R-Square Adj. R-Square Mallow's Cp
3       1 1              wt 0.7528328     0.7445939   12.480939
1       2 1            disp 0.7183433     0.7089548   18.129607
2       3 1              hp 0.6024373     0.5891853   37.112642
4       4 1            qsec 0.1752963     0.1478062  107.069616
8       5 2           hp wt 0.8267855     0.8148396    2.369005
10      6 2         wt qsec 0.8264161     0.8144448    2.429492
6       7 2         disp wt 0.7809306     0.7658223    9.879096
5       8 2         disp hp 0.7482402     0.7308774   15.233115
7       9 2       disp qsec 0.7215598     0.7023571   19.602810
9      10 2         hp qsec 0.6368769     0.6118339   33.472150
14     11 3      hp wt qsec 0.8347678     0.8170643    3.061665
11     12 3     disp hp wt 0.8268361     0.8082829    4.360702
13     13 3    disp wt qsec 0.8264170     0.8078189    4.429343
12     14 3    disp hp qsec 0.7541953     0.7278591   16.257790
15     15 4 disp hp wt qsec 0.8351443     0.8107212    5.000000
> plot(p)
```



Stepwise regression (Use all of the available independent variables)

```
> reg <- lm(mpg ~ ., data=mtcars)
> summary(reg)
Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,   Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```
> start = lm(Life.Exp~1, data=statedata)
> fitALL = lm(Life.Exp~., data=statedata)

> stepwise <- step(start, direction="both",scope=formula(fitALL))
Start:  AIC=115.94
mpg ~ 1

       Df Sum of Sq     RSS     AIC
+ wt    1    847.73  278.32  73.217
+ cyl   1    817.71  308.33  76.494
+ disp  1    808.89  317.16  77.397
+ hp    1    678.37  447.67  88.427
+ drat  1    522.48  603.57  97.988
+ vs    1    496.53  629.52  99.335
+ am    1    405.15  720.90 103.672
+ carb  1    341.78  784.27 106.369
+ gear  1    259.75  866.30 109.552
+ qsec  1    197.39  928.66 111.776
<none>              1126.05 115.943

Step:  AIC=73.22
mpg ~ wt

       Df Sum of Sq     RSS     AIC
+ cyl   1     87.15  191.17  63.198
+ hp    1     83.27  195.05  63.840
+ qsec  1     82.86  195.46  63.908
+ vs    1     54.23  224.09  68.283
+ carb  1     44.60  233.72  69.628
+ disp  1     31.64  246.68  71.356
<none>              278.32  73.217
+ drat  1      9.08  269.24  74.156
+ gear  1      1.14  277.19  75.086
+ am    1      0.00  278.32  75.217
- wt    1    847.73 1126.05 115.943

Step:  AIC=63.2
mpg ~ wt + cyl

       Df Sum of Sq    RSS    AIC
+ hp    1    14.551 176.62 62.665
+ carb  1    13.772 177.40 62.805
<none>              191.17 63.198
+ qsec  1    10.567 180.60 63.378
+ gear  1     3.028 188.14 64.687
+ disp  1     2.680 188.49 64.746
+ vs    1     0.706 190.47 65.080
+ am    1     0.125 191.05 65.177
+ drat  1     0.001 191.17 65.198
- cyl   1    87.150 278.32 73.217
- wt    1   117.162 308.33 76.494

Step:  AIC=62.66
mpg ~ wt + cyl + hp
```

final model

	Df	Sum of Sq	RSS	AIC
<none>			176.62	62.665
- hp	1	14.551	191.17	63.198
+ am	1	6.623	170.00	63.442
+ disp	1	6.176	170.44	63.526
- cyl	1	18.427	195.05	63.840
+ carb	1	2.519	174.10	64.205
+ drat	1	2.245	174.38	64.255
+ qsec	1	1.401	175.22	64.410
+ gear	1	0.856	175.76	64.509
+ vs	1	0.060	176.56	64.654
- wt	1	115.354	291.98	76.750

Remark 9.4 Modern techniques for variable selection

Final model

```
>finalmodel <- lm(mpg ~ wt + cyl + hp,data=mtcars)
>summary(finalmodel)

Call:
lm(formula = mpg ~ wt + cyl + hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9290 -1.5598 -0.5311  1.1850  5.8986

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
cyl         -0.94162    0.55092  -1.709 0.098480 .
hp          -0.01804    0.01188  -1.519 0.140015
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,  Adjusted R-squared:  0.8263
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11

library(faraway)
m <- model.matrix(finalmodel) [,-1]

vif(m)
      wt      cyl       hp
2.580486 4.757456 3.258481
```