Example: Body Fat

Yuqing Xu

Department of Statistics
University of Wisconsin-Madison

November 10, 2017

Outline

How to Select?

- Use criteria:
  - Mallow's Cp
  - Adjusted R^2
  - AIC and BIC
- Optimization method: Forward/Backward/Stepwise

Body Fat

- " A variety of popular health books suggest that readers assess their health, at least in part, by estimating their percentage of body fat. Bailey (1994, pp. 179-186), for instance, presents tables of estimates based on age, gender, and various skinfold measurements obtained using a caliper. Bailey (1991, p. 18) suggests that "15 percent fat for men and 22 percent fat for women are maximums for good health." "
- Fitting body fat to the other measurements using multiple regression provides a convenient way of estimating body fat for men using only a scale and a measuring tape.
- In the dataset provided by Dr. A. Garth Fisher (personal communication, October 5, 1994), age, weight, height, and 10 body circumference measurements are recorded for 252 men. Each man's percentage of body fat was accurately estimated by an underwater weighing technique discussed below.

Body Fat

Percentage of body fat for an individual can be estimated once body density has been determined. Folks (e.g. Siri (1956)) assume that the body consists of two components - lean body tissue and fat tissue. Letting

D = Body Density (gm/cm^3)
A = proportion of lean body tissue
B = proportion of fat tissue ($A + B = 1$)
a = density of lean body tissue (gm/cm^3)
b = density of fat tissue (gm/cm^3)

we have

$$D = 1/[(A/a) + (B/b)].$$

Solving for B we find

$$B = (1/D) * [ab/(a - b)] - [b/(a - b)].$$

Body Fat

```
> address <- "http://www.stat.wisc.edu/~cui/fall_2013/bodyfat.csv"
> str(bodyfat <- read.csv(address, header=TRUE))

'data.frame': 252 obs. of  15 variables:
 $ density: num  1.07 1.09 1.04 1.08 1.03 1.05 1.05 1.07 1.09 1.07 ...
 $ siri   : num  12.3 6.1 25.3 10.4 28.7 20.9 19.2 12.4 4.1 11.7 ...
 $ age    : int  23 22 22 26 24 24 26 25 25 23 ...
 $ weight : num  154 173 154 185 184 ...
 $ height : num  67.8 72.2 66.2 72.2 71.2 ...
 $ neck   : num  36.2 38.5 34 37.4 34.4 39 36.4 37.8 38.1 42.1 ...
 $ chest  : num  93.1 93.6 95.8 101.8 97.3 ...
 $ abdomen: num  85.2 83 87.9 86.4 100 94.4 90.7 88.5 82.5 88.6 ...
 $ hip    : num  94.5 98.7 99.2 101.2 101.9 ...
 $ thigh  : num  59 58.7 59.6 60.1 63.2 66 58.4 60 62.9 63.1 ...
 $ knee   : num  37.3 37.3 38.9 37.3 42.2 42 38.3 39.4 38.3 41.7 ...
 $ ankle  : num  21.9 23.4 24 22.8 24 25.6 22.9 23.2 23.8 25 ...
 $ biceps : num  32 30.5 28.8 32.4 32.2 35.7 31.9 30.5 35.9 35.6 ...
 $ forearm: num  27.4 28.9 25.2 29.4 27.7 30.6 27.8 29 31.1 30 ...
 $ wrist  : num  17.1 18.2 16.6 18.2 17.7 18.8 17.7 18.8 18.2 19.2 ...
```

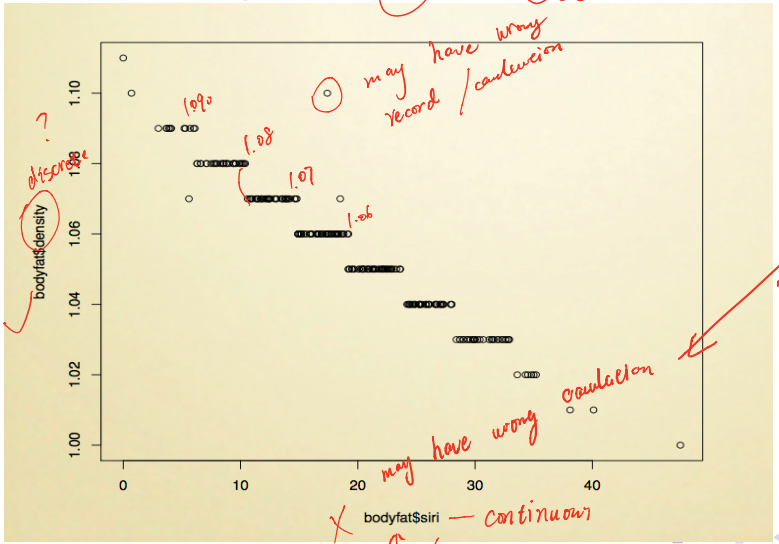
Clean up the data

We need to do some data cleaning first. Notice that there are some points outside of the cloud of data, which means they're outlier candidates, that is, either leverage points or outliers in Y.

First, looking at the data description, we learn that there's a linear relation between the percentage of body fat (in this case the response, density) and the so called Siri formula, given by (100*B) = 495/D - 450. In fact,

Clean up the data

```
> plot(bodyfat$density ~ 1/bodyfat$siri)
```



Clean up the data

I'll remove siri index from the dataset and just do the regression on density.

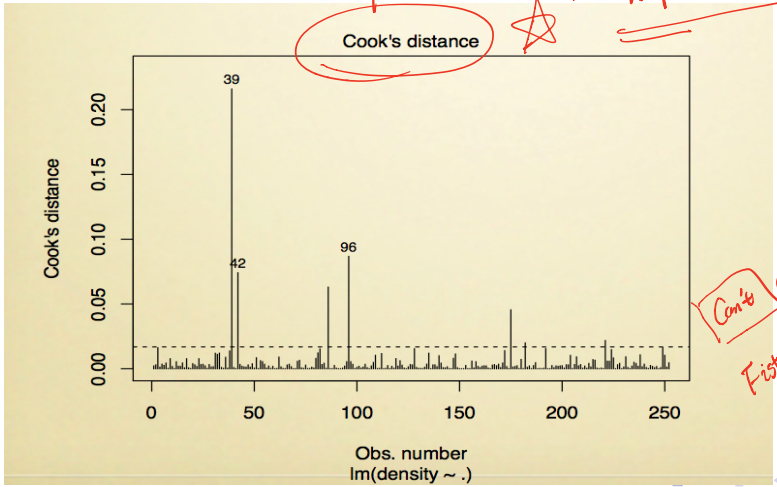
```
> summary(model <- lm(density ~ ., data=subset(bodyfat,select=-siri)))$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.1387217190	4.125611e-02	27.6012868	3.917932e-76
age	-0.0001394986	7.692789e-05	-1.8133689	7.103452e-02
weight	0.0002260866	1.272874e-04	1.7761896	7.697922e-02
height	0.0001245953	2.283098e-04	0.5457293	5.857631e-01
neck	0.0009546315	5.528236e-04	1.7268285	8.549599e-02
chest	0.0001466833	2.357782e-04	0.6221242	5.344558e-01
abdomen	-0.0022432951	2.055803e-04	-10.9120146	9.612408e-23
hip	0.0006535345	3.469850e-04	1.8834662	6.085654e-02
thigh	-0.0007432684	3.432929e-04	-2.1651145	3.137316e-02
knee	-0.0001192582	5.754391e-04	-0.2072473	8.359939e-01
ankle	-0.0005523390	5.266624e-04	-1.0487535	2.953554e-01
biceps	-0.0005191862	4.069478e-04	-1.2758054	2.032681e-01
forearm	-0.0009744601	4.735425e-04	-2.0578090	4.069786e-02
wrist	0.0039247716	1.272139e-03	3.0851740	2.274875e-03

Clean up the data

Rule of thumb: classify as leverages anything above 4/(n-p). (Fox, 1997)

```
> plot(model, which=4)
> abline( h = 4/(252-15),lty=2 )
```



Clean up the data

Who is the 39 guy?

```
> bodyfat[39,]
```

```
     density siri age  weight  height neck chest abdomen
39      1.02  35.2  46  363.15 72.25 51.2 136.2   148.1
   hip  thigh  knee  ankle biceps forearm wrist
 147.7  87.3  49.1   29.6     45      29  21.4
```

He weights 363 pounds. Let's say we remove him from the model.

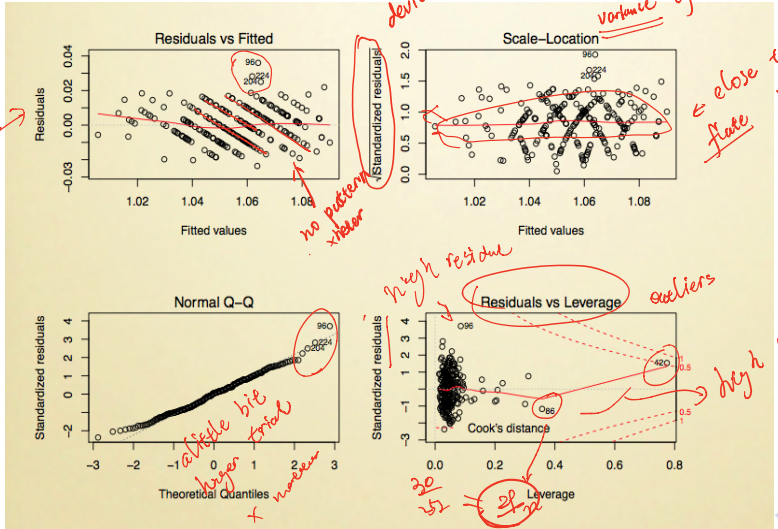
```
> summary(model <- lm(density ~ ., data=bodyfat[-39,-2]))

Call:
lm(formula = density ~ ., data = bodyfat[-39, -2])
Residuals:
      Min        1Q    Median        3Q       Max
-0.023655 -0.007565  0.000514  0.006824  0.036024

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.1249672  0.0413521  27.205  < 2e-16 ***
age         -0.0001579  0.0000767  -2.059 0.040628 *
weight       0.0001473  0.0001309   1.126 0.261498
height       0.0002031  0.0002290   0.887 0.376051
neck         0.0007359  0.0005565   1.322 0.187323
chest        0.0002745  0.0002405   1.142 0.254814
abdomen     -0.0021938  0.0002050 -10.702  < 2e-16 ***
hip          0.0005416  0.0003476   1.558 0.120486
thigh       -0.0006684  0.0003420  -1.955 0.051795 .
knee         0.0000790  0.0005772   0.137 0.891245
ankle       -0.0006057  0.0005227  -1.159 0.247675
biceps      -0.0005495  0.0004037  -1.361 0.174785
forearm     -0.0006431  0.0004918  -1.308 0.192208
wrist        0.0042365  0.0012687   3.339 0.000976 ***
```

Clean up the data

```
> layout(matrix(1:4, ncol=2))
> plot(model)
```



Clean up the data

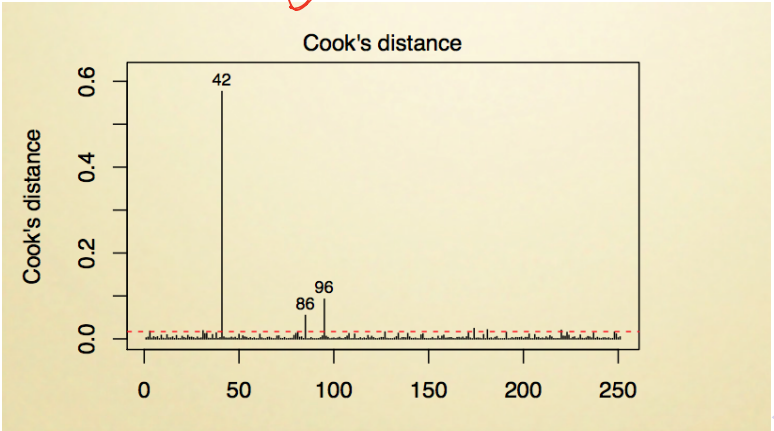
Let's find out what's wrong with those other guys.

```
> bodyfat[c(42,86),] # High Leverage
    density siri age weight height neck chest abdomen
42     1.03 32.9  44    205   29.5 36.6 106.0   104.3
86     1.04 26.6  67    167   67.5 36.5  98.9    89.7
   hip thigh knee ankle biceps forearm wrist
42 115.5  70.6 42.5  23.7   33.6    28.7  17.4
86  96.2  54.7 37.8  33.7   32.4    27.7  18.2
> bodyfat[c(96,204,224),] # Possible outliers
    density siri age weight height neck chest abdomen
96     1.10 17.4  53 224.50  77.75 41.1 113.2
204    1.09  6.0  44 184.00  74.00 37.9 100.8
224    1.09  5.2  55 142.25  67.25 35.2  92.7
    hip thigh knee ankle biceps forearm
96  99.2 107.5  61.7 42.3  23.2    32.9    30.8
204 89.1 102.6  60.6 39.0  24.0    32.9    29.2
224 82.8  91.9  54.4 35.2  22.5    29.4    26.8
    wrist
96   20.4
204  18.4
224  17.0
```

Clean up the data

Notice observation 42 is from someone only 30 inches tall (possibly an input error given the circumference measurements are large, and he weights 200 pounds). In fact

```
> plot(model, which=4)
> abline( h = 4/(251-15), col='red', lty=2)
```



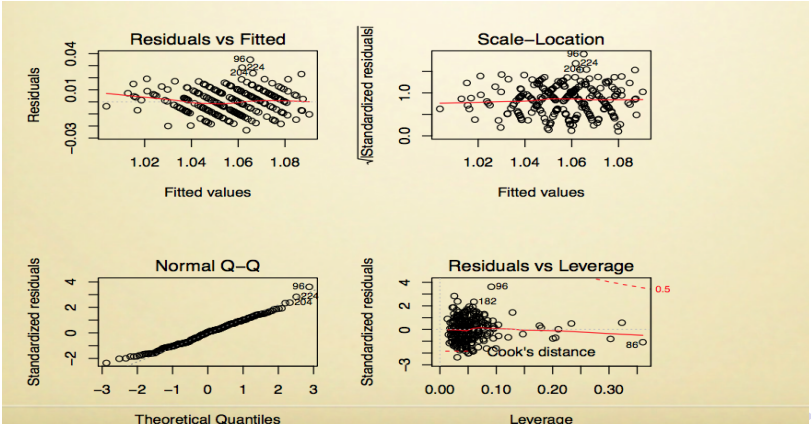
Diagonostics

```
> summary(model <- lm(density ~ ., data=bodyfat[-c(39,42),-2]))
Call:
lm(formula = density ~ ., data = bodyfat[-c(39, 42), -2])
Residuals:
      Min        1Q    Median        3Q       Max
-0.023523 -0.007726  0.000525  0.007110  0.034870
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.068e+00  5.559e-02  19.203  < 2e-16 ***
age         -1.616e-04  7.652e-05  -2.112 0.035706 *
weight       5.854e-06  1.597e-04   0.037 0.970788
height       8.078e-04  4.546e-04   1.777 0.076884 .
neck         8.342e-04  5.586e-04   1.493 0.136683
chest        4.187e-04  2.575e-04   1.626 0.105211
abdomen     -2.080e-03  2.174e-04  -9.568  < 2e-16 ***
hip          5.663e-04  3.469e-04   1.632 0.103926
thigh       -5.522e-04  3.492e-04  -1.581 0.115164
knee        -7.481e-05  5.842e-04  -0.128 0.898208
ankle       -5.407e-04  5.229e-04  -1.034 0.302156
biceps      -4.770e-04  4.053e-04  -1.177 0.240356
forearm     -5.786e-04  4.922e-04  -1.176 0.240880
wrist        4.364e-03  1.268e-03   3.442 0.000682 ***
---
Signif. codes:  0
```

Diagonostics

For the other observations, it is not so clear why they're influential, if they are. I will also not look at outlier tests until we remove observation 42 as well.

```
> layout(matrix(1:4, ncol=2))
> plot(model)
```

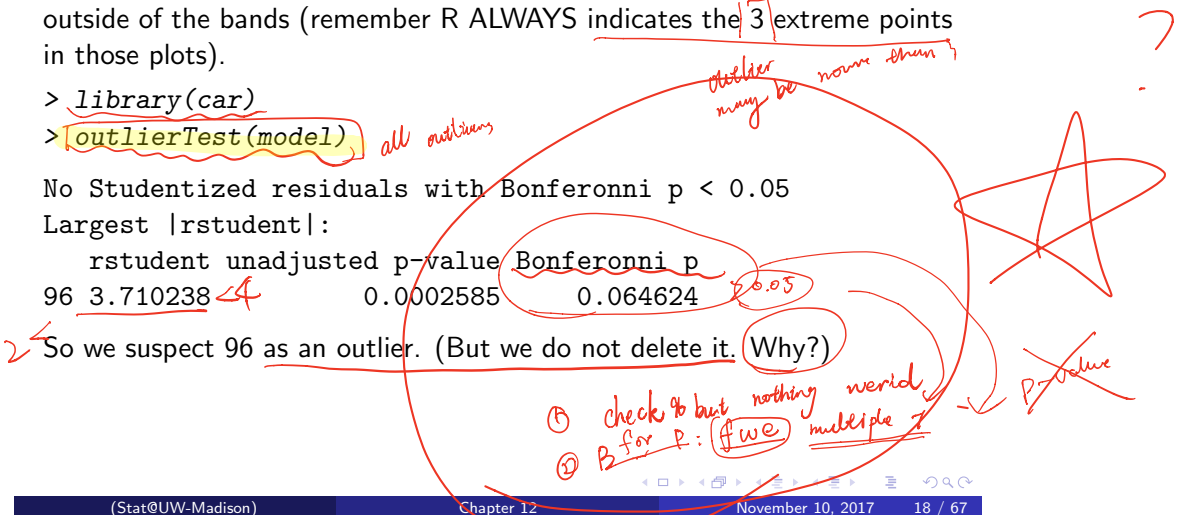


Outliers

The set c(96,204,224) still shows up as outliers. They also seem to be outside of the bands (remember R ALWAYS indicates the 3 extreme points in those plots).

```
> library(car)
> outlierTest(model)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferonni p
96 3.710238       0.0002585      0.064624
```

So we suspect 96 as an outlier. (But we do not delete it. Why?)



Eyeballing p-values

- So far, we have finished data cleaning and could get variable selection started.
- First idea: p-values
- Our principle: delete one variable (usually with p-value >0.1 or 0.2) each time(why?).

Eyeballing p-values

```
> summary(model)
Call:
lm(formula = density ~ ., data = bodyfat[-c(39, 42), -2])

Residuals:
      Min        1Q    Median        3Q       Max
-0.023523 -0.007726  0.000525  0.007110  0.034870

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.068e+00  5.559e-02  19.203  < 2e-16 ***
age         -1.616e-04  7.652e-05  -2.112 0.035706 *
weight       5.854e-06  1.597e-04   0.037 0.970788
height       8.078e-04  4.546e-04   1.777 0.076884 .
neck         8.342e-04  5.586e-04   1.493 0.136683
chest        4.187e-04  2.575e-04   1.626 0.105211
abdomen     -2.080e-03  2.174e-04  -9.568  < 2e-16 ***
hip          5.663e-04  3.469e-04   1.632 0.103926
thigh       -5.522e-04  3.492e-04  -1.581 0.115164
knee        -7.481e-05  5.842e-04  -0.128 0.898208
ankle       -5.407e-04  5.229e-04  -1.034 0.302156
biceps      -4.770e-04  4.053e-04  -1.177 0.240356
forearm     -5.786e-04  4.922e-04  -1.176 0.240880
wrist        4.364e-03  1.268e-03   3.442 0.000682 ***
```

$\beta_j = 0$?

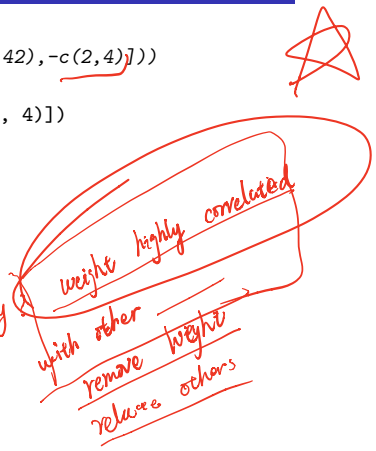
delete — p.value most large

Eyeballing p-values

Remove Weight:

```
> summary(model.eye <- lm(density ~ ., data=bodyfat[-c(39,42),-c(2,4)]))
Call:
lm(formula = density ~ ., data = bodyfat[-c(39, 42), -c(2, 4)])
Residuals:
      Min        1Q    Median        3Q       Max
-0.023527 -0.007719  0.000513  0.007119  0.034876

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.066e+00  2.053e-02  51.917  < 2e-16 ***
age         -1.622e-04  7.476e-05  -2.170 0.031015 *
height       8.193e-04  3.274e-04   2.502 0.013016 *
neck         8.394e-04  5.392e-04   1.557 0.120877
chest        4.239e-04  2.148e-04   1.973 0.049611 *
abdomen     -2.077e-03  2.018e-04 -10.293  < 2e-16 ***
hip          5.716e-04  3.146e-04   1.817 0.070503 .
thigh       -5.498e-04  3.423e-04  -1.606 0.109519
knee        -7.204e-05  5.780e-04  -0.125 0.900925
ankle       -5.373e-04  5.132e-04  -1.047 0.296230
biceps      -4.741e-04  3.964e-04  -1.196 0.232918
forearm     -5.765e-04  4.875e-04  -1.183 0.238172
wrist        4.372e-03  1.246e-03   3.509 0.000539 ***
---
```



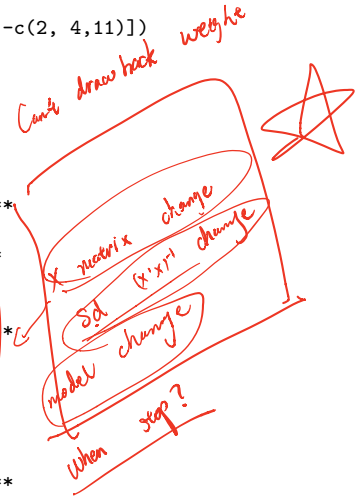
Eyeballing p-values

Remove knee:

```
> summary(model.eye <- lm(density ~ ., data=bodyfat[-c(39,42),-c(2,4,11)]))

Call:
lm(formula = density ~ ., data = bodyfat[-c(39, 42), -c(2, 4,11)])
Residuals:
      Min        1Q    Median        3Q       Max
-0.023480 -0.007728  0.000462  0.007191  0.034853

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.066e+00  2.027e-02  52.599  < 2e-16 ***
age         -1.644e-04  7.253e-05  -2.266 0.024328 *
height       8.040e-04  3.030e-04   2.654 0.008496 **
neck         8.439e-04  5.369e-04   1.572 0.117280
chest        4.232e-04  2.143e-04   1.975 0.049446 *
abdomen     -2.078e-03  2.013e-04 -10.323  < 2e-16 ***
hip          5.662e-04  3.110e-04   1.821 0.069887 .
thigh       -5.634e-04  3.238e-04  -1.740 0.083147 .
ankle       -5.508e-04  5.005e-04  -1.100 0.272271
biceps      -4.729e-04  3.955e-04  -1.196 0.232983
forearm     -5.807e-04  4.853e-04  -1.197 0.232600
wrist        4.358e-03  1.238e-03   3.519 0.000518 ***
```



Mallow's Cp

- Mallow's Cp is a criteria based on the Model Error.

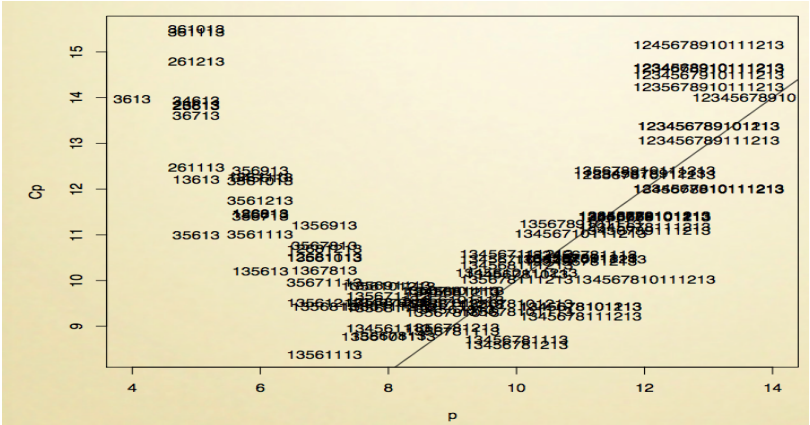
$$C_p(k) = \frac{RSS(k)}{s^2} + 2k - n.$$

- $\mathbb{E}[C_p(k)] \approx k + \frac{\beta' X'(I-P_1)X\beta}{\sigma^2}$. If the model fits well in the sense that $\mathbf{X}\beta$ is well approximated by vectors in $\mathcal{C}(\mathbf{X}_1)$, then the quantity $\frac{\beta' X'(I-P_1)X\beta}{\sigma^2}$ will be small.
- The estimation s^2 is usually based on full model.
- "leaps()" performs an exhaustive search for the best subsets of the variables for prediction y in linear regression.

```
> X <- model.matrix(model)[,-1]
> Y <- bodyfat[-c(39,42),1]
> library(leaps) # for leaps()
> library(faraway) # for Cpplot()
> g <- leaps(X, Y)
```

Mallow's Cp

```
> Cpplot(g)
```

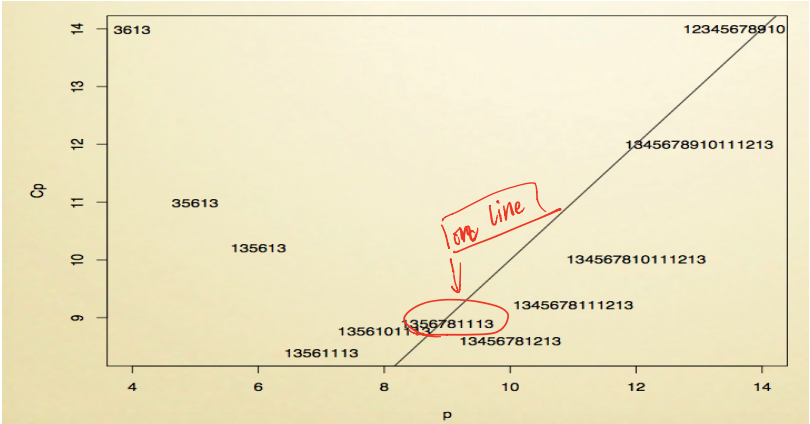


Mallow's Cp

```
> g <- leaps(X,Y, nbest=1)
> Cpplot(g)
```

PSs small Cp

smallest



Mallow's Cp

A good choice seems to be (1,3,5,6,7,8,11,13). Notice these are the covariates and the leaps() index does not include the intercept. Remember column 1 was the response and column 2 correspond to the siri index and it was removed, so we adjust the indexes.

Mallow's Cp

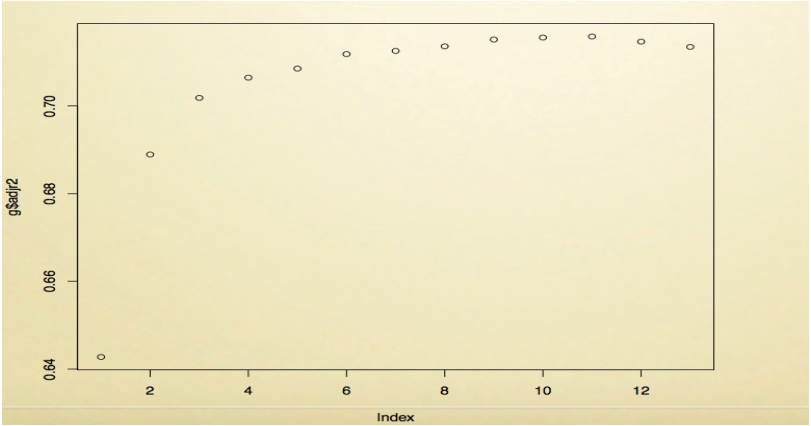
```
> cp.choice <- c(1,3,5,6,7,8,11,13)+2
> summary(model.cp <- lm(density ~ ., data=bodyfat[-c(39,42),c(1,cp.choice)]))

Call:
lm(formula = density ~ ., data = bodyfat[-c(39, 42), c(1, cp.choice)])
Residuals:
      Min        1Q    Median        3Q       Max
-0.024850 -0.007826  0.000398  0.006256  0.035963
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.068e+00  2.010e-02  53.130  < 2e-16 ***
age         -1.446e-04  7.172e-05  -2.016  0.04487 *
height       7.977e-04  3.005e-04   2.654  0.00848 **
chest        4.492e-04  2.085e-04   2.154  0.03222 *
abdomen     -2.023e-03  2.005e-04 -10.091  < 2e-16 ***
hip          4.907e-04  3.096e-04   1.585  0.11422
thigh       -5.355e-04  3.212e-04  -1.667  0.09675 .
biceps      -5.102e-04  3.706e-04  -1.377  0.16993
wrist        4.347e-03  1.077e-03   4.036 7.31e-05 ***
```

Adjusted R^2

The function leaps() uses Cp by default, but we can change a parameter and select covariates based on the

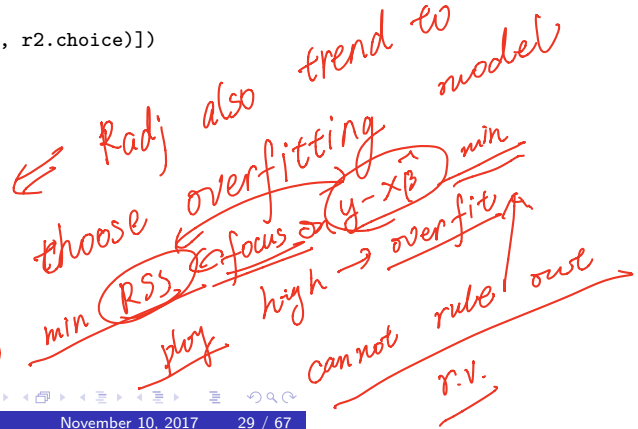
```
> g <- leaps(X,Y, nbest=1, method="adjr2")
> plot(g$adjr2)
```



Adjusted R^2

We have to look inside of the g object, since there's no Cpplot method for adjusted R^2.

```
> (g$which)[which(g$adjr2 == max(g$adjr2)),]
    1     2     3     4     5     6     7     8     9     A     B     C     D
 TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
> r2.choice <- c(3,5:10,12:15)
> summary(model.r2 <- lm(density ~ ., data=bodyfat[-c(39,42),c(1,r2.choice)]))
Call:
lm(formula = density ~ ., data = bodyfat[-c(39, 42), c(1, r2.choice)])
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.066e+00  2.027e-02  52.599  < 2e-16 ***
age         -1.644e-04  7.253e-05  -2.266 0.024328 *
height       8.040e-04  3.030e-04   2.654 0.008496 **
neck         8.439e-04  5.369e-04   1.572 0.117280
chest        4.232e-04  2.143e-04   1.975 0.049446 *
abdomen     -2.078e-03  2.013e-04 -10.323  < 2e-16 ***
hip          5.662e-04  3.110e-04   1.821 0.069887 .
thigh       -5.634e-04  3.238e-04  -1.740 0.083147 .
ankle       -5.508e-04  5.005e-04  -1.100 0.272271
biceps      -4.729e-04  3.955e-04  -1.196 0.232983
forearm     -5.807e-04  4.853e-04  -1.197 0.232600
wrist        4.358e-03  1.238e-03   3.519 0.000518 ***
```



AIC, BIC

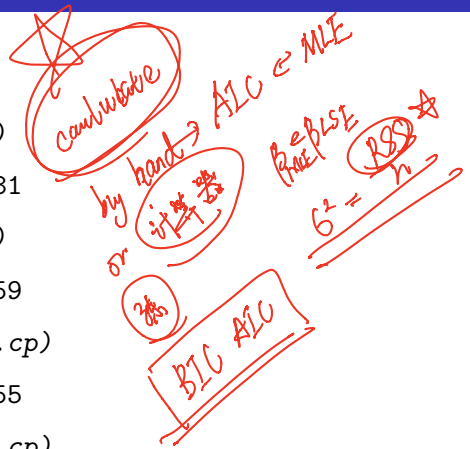
There are many information-based criterias. Arguably the most famous ones are the AIC and BIC.

$$AIC = -2LogLikelihood + 2p$$
$$BIC = -2LogLikelihood + log(n)p$$

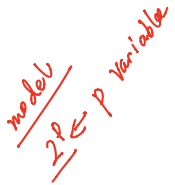
Since they're based on the negative log-likelihood, the best models have the smallest AIC. Also the number of parameters is included as a positive value, so it means large number of parameters are penalized. Notice both have built-in functions in R.

AIC, BIC

```
> AIC(model)
[1] -1571.481
> BIC(model)
[1] -1518.659
> AIC(model.cp)
[1] -1576.355
> BIC(model.cp)
[1] -1541.141
```



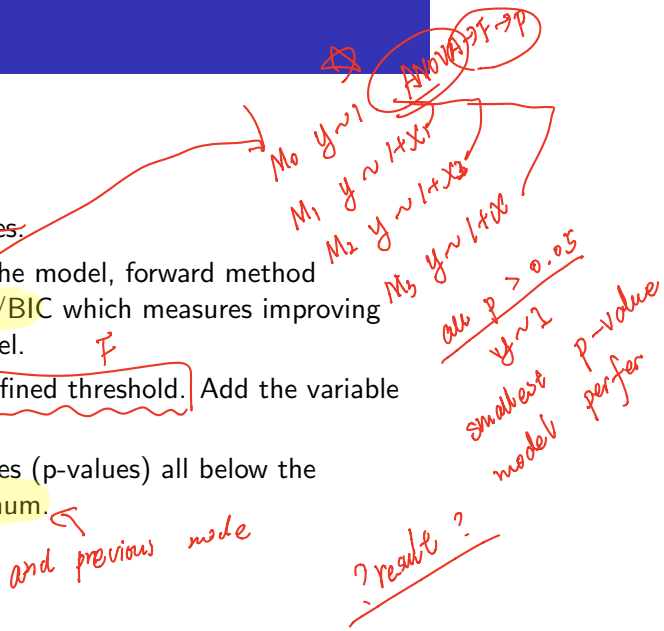
Optimization methods



- Leaps() does an exhaustive search and get the best subset. When p gets bigger, the computational burden is heavy.
- Use approximate optimization: Forward Selection/Backward Elimination/Stepwise Regression

forward selection

- Start with the model with no variables.
- For each of the variables outside of the model, forward method calculates the F-value (p-value)/AIC/BIC which measures improving if this variable is added into the model.
- Compare this value with some pre-defined threshold. Add the variable if the value is satisfactory.
- Repeat the procedure until all F-values (p-values) all below the threshold or AIC/BIC achieves minimum.



backward selection

- Start with the model with all variables
- Each time, remove one variable out and compare the changed model with original one. One F-value (p-value)/AIC/BIC which measures improving if this variable is removed from the model.
- Repeat the procedure until all F-values (p-values) all above the threshold or AIC/BIC achieves minimum.

stepwise selection

- At each step do forward addition and backward elimination.
- Need some condition to stop.

Drawbacks

- Because of the "one-at-a-time" nature of adding/dropping variables, it's possible to miss the "optimal" model.
- The p-values used should not be treated too literally.
- The procedures are not directly linked to final objectives of prediction or explanation and so may not really help solve the problem of interest.
- Stepwise variable selection tends to pick models that are smaller than desirable for prediction purposes.

AIC, BIC + optimization method

To use them as a criteria to select the covariates, use the step() function

```
> model.AIC <- step(model, k=2)

Start:  AIC=-2282.95
density ~ age + weight + height + neck + chest + abdomen + hip +
    thigh + knee + ankle + biceps + forearm + wrist

           Df Sum of Sq      RSS     AIC
- weight    1 0.0000001 0.024177 -2284.9
- knee      1 0.0000017 0.024179 -2284.9
- ankle     1 0.0001095 0.024287 -2283.8
- forearm   1 0.0001416 0.024319 -2283.5
- biceps    1 0.0001419 0.024319 -2283.5
<none>                 0.024177 -2282.9
- neck      1 0.0002285 0.024406 -2282.6
- thigh     1 0.0002561 0.024433 -2282.3
- chest     1 0.0002710 0.024448 -2282.2
- hip       1 0.0002730 0.024450 -2282.1
- height    1 0.0003234 0.024501 -2281.6
- age       1 0.0004571 0.024634 -2280.3
- wrist     1 0.0012139 0.025391 -2272.7
- abdomen   1 0.0093784 0.033556 -2203.0
```

AIC, BIC+ optimization method

```
Step:  AIC=-2284.95
density ~ age + height + neck + chest + abdomen + hip + thigh +
    knee + ankle + biceps + forearm + wrist

          Df Sum of Sq      RSS     AIC
- knee     1 0.0000016 0.024179 -2286.9
- ankle    1 0.0001118 0.024289 -2285.8
- forearm  1 0.0001427 0.024320 -2285.5
- biceps   1 0.0001459 0.024323 -2285.4
<none>                 0.024177 -2284.9
- neck     1 0.0002472 0.024425 -2284.4
- thigh    1 0.0002632 0.024441 -2284.2
- hip      1 0.0003367 0.024514 -2283.5
- chest    1 0.0003973 0.024575 -2282.9
- age      1 0.0004803 0.024658 -2282.0
- height   1 0.0006387 0.024816 -2280.4
- wrist    1 0.0012559 0.025433 -2274.3
- abdomen  1 0.0108071 0.034985 -2194.6
```

after delete weight

AIC, BIC+ optimization method

```
Step:  AIC=-2286.93
density ~ age + height + neck + chest + abdomen + hip + thigh +
    ankle + biceps + forearm + wrist

          Df Sum of Sq      RSS     AIC
- ankle    1 0.0001230 0.024302 -2287.7
- biceps   1 0.0001453 0.024324 -2287.4
- forearm  1 0.0001455 0.024325 -2287.4
<none>                 0.024179 -2286.9
- neck     1 0.0002511 0.024430 -2286.3
- thigh    1 0.0003076 0.024487 -2285.8
- hip      1 0.0003368 0.024516 -2285.5
- chest    1 0.0003962 0.024575 -2284.9
- age      1 0.0005218 0.024701 -2283.6
- height   1 0.0007155 0.024894 -2281.6
- wrist    1 0.0012584 0.025437 -2276.2
- abdomen  1 0.0108253 0.035004 -2196.4
```

AIC, BIC+ optimization method

```
Step:  AIC=-2287.66
density ~ age + height + neck + chest + abdomen + hip + thigh +
    biceps + forearm + wrist

          Df Sum of Sq      RSS     AIC
- biceps   1 0.0001408 0.024443 -2288.2
- forearm  1 0.0001521 0.024454 -2288.1
<none>                 0.024302 -2287.7
- neck     1 0.0002873 0.024589 -2286.7
- hip      1 0.0003192 0.024621 -2286.4
- thigh    1 0.0003440 0.024646 -2286.2
- chest    1 0.0003649 0.024667 -2285.9
- age      1 0.0004759 0.024778 -2284.8
- height   1 0.0006510 0.024953 -2283.1
- wrist    1 0.0011355 0.025438 -2278.2
- abdomen  1 0.0107485 0.035051 -2198.1
```

AIC, BIC+ optimization method

```
Step:  AIC=-2288.22
density ~ age + height + neck + chest + abdomen + hip + thigh +
    forearm + wrist

           Df Sum of Sq      RSS     AIC
<none>                  0.024443 -2288.2
- neck      1 0.0002443 0.024687 -2287.7
- forearm   1 0.0002838 0.024727 -2287.3
- chest     1 0.0002987 0.024742 -2287.2
- hip       1 0.0003235 0.024766 -2286.9
- age       1 0.0004919 0.024935 -2285.2
- thigh     1 0.0005299 0.024973 -2284.9
- height    1 0.0006447 0.025088 -2283.7
- wrist     1 0.0010874 0.025530 -2279.3
- abdomen   1 0.0106199 0.035063 -2200.0
```

AIC, BIC+ optimization method

```
> model.BIC <- step(model, k=log(250))
Start:  AIC=-2233.65
density ~ age + weight + height + neck + chest + abdomen + hip +
    thigh + knee + ankle + biceps + forearm + wrist

          Df Sum of Sq      RSS     AIC
- weight   1 0.0000001 0.024177 -2239.2
- knee     1 0.0000017 0.024179 -2239.2
- ankle    1 0.0001095 0.024287 -2238.0
- forearm  1 0.0001416 0.024319 -2237.7
- biceps   1 0.0001419 0.024319 -2237.7
- neck     1 0.0002285 0.024406 -2236.8
- thigh    1 0.0002561 0.024433 -2236.5
- chest    1 0.0002710 0.024448 -2236.4
- hip      1 0.0002730 0.024450 -2236.4
- height   1 0.0003234 0.024501 -2235.8
- age      1 0.0004571 0.024634 -2234.5
<none>                 0.024177 -2233.7
- wrist    1 0.0012139 0.025391 -2226.9
- abdomen  1 0.0093784 0.033556 -2157.2
```

AIC, BIC+ optimization method

```
Step:  AIC=-2239.17
density ~ age + height + neck + chest + abdomen + hip + thigh +
    knee + ankle + biceps + forearm + wrist

          Df Sum of Sq      RSS     AIC
- knee     1 0.0000016 0.024179 -2244.7
- ankle    1 0.0001118 0.024289 -2243.5
- forearm  1 0.0001427 0.024320 -2243.2
- biceps   1 0.0001459 0.024323 -2243.2
- neck     1 0.0002472 0.024425 -2242.2
- thigh    1 0.0002632 0.024441 -2242.0
- hip      1 0.0003367 0.024514 -2241.2
- chest    1 0.0003973 0.024575 -2240.6
- age      1 0.0004803 0.024658 -2239.8
<none>                 0.024177 -2239.2
- height   1 0.0006387 0.024816 -2238.2
- wrist    1 0.0012559 0.025433 -2232.0
- abdomen  1 0.0108071 0.034985 -2152.3
```

AIC, BIC+ optimization method

```
Step:  AIC=-2244.68
density ~ age + height + neck + chest + abdomen + hip + thigh +
    ankle + biceps + forearm + wrist

          Df Sum of Sq      RSS     AIC
- ankle    1 0.0001230 0.024302 -2248.9
- biceps   1 0.0001453 0.024324 -2248.7
- forearm  1 0.0001455 0.024325 -2248.7
- neck     1 0.0002511 0.024430 -2247.6
- thigh    1 0.0003076 0.024487 -2247.0
- hip      1 0.0003368 0.024516 -2246.7
- chest    1 0.0003962 0.024575 -2246.1
- age      1 0.0005218 0.024701 -2244.9
<none>                 0.024179 -2244.7
- height   1 0.0007155 0.024894 -2242.9
- wrist    1 0.0012584 0.025437 -2237.5
- abdomen  1 0.0108253 0.035004 -2157.7
```

AIC, BIC+ optimization method

```
Step:  AIC=-2248.93
density ~ age + height + neck + chest + abdomen + hip + thigh +
    biceps + forearm + wrist

          Df Sum of Sq      RSS     AIC
- biceps   1 0.0001408 0.024443 -2253.0
- forearm  1 0.0001521 0.024454 -2252.9
- neck     1 0.0002873 0.024589 -2251.5
- hip      1 0.0003192 0.024621 -2251.2
- thigh    1 0.0003440 0.024646 -2250.9
- chest    1 0.0003649 0.024667 -2250.7
- age      1 0.0004759 0.024778 -2249.6
<none>                 0.024302 -2248.9
- height   1 0.0006510 0.024953 -2247.8
- wrist    1 0.0011355 0.025438 -2243.0
- abdomen  1 0.0107485 0.035051 -2162.9
```

AIC, BIC+ optimization method

```
Step:  AIC=-2253
density ~ age + height + neck + chest + abdomen + hip + thigh +
    forearm + wrist

          Df Sum of Sq      RSS     AIC
- neck     1 0.0002443 0.024687 -2256.0
- forearm  1 0.0002838 0.024727 -2255.6
- chest    1 0.0002987 0.024742 -2255.5
- hip      1 0.0003235 0.024766 -2255.2
- age      1 0.0004919 0.024935 -2253.6
- thigh    1 0.0005299 0.024973 -2253.2
<none>                 0.024443 -2253.0
- height   1 0.0006447 0.025088 -2252.0
- wrist    1 0.0010874 0.025530 -2247.6
- abdomen  1 0.0106199 0.035063 -2168.3
```

AIC, BIC+ optimization method

```
Step:  AIC=-2256.04
density ~ age + height + chest + abdomen + hip + thigh + forearm +
    wrist

          Df Sum of Sq      RSS     AIC
- forearm  1 0.0001850 0.024872 -2259.7
- hip      1 0.0002620 0.024949 -2258.9
- thigh    1 0.0004271 0.025114 -2257.3
- chest    1 0.0004525 0.025140 -2257.0
- age      1 0.0004773 0.025164 -2256.8
<none>                 0.024687 -2256.0
- height   1 0.0007175 0.025405 -2254.4
- wrist    1 0.0016827 0.026370 -2245.1
- abdomen  1 0.0104043 0.035091 -2173.6
```

AIC, BIC+ optimization method

```
Step:  AIC=-2259.69
density ~ age + height + chest + abdomen + hip + thigh + wrist

          Df Sum of Sq      RSS     AIC
- hip      1 0.0002715 0.025144 -2262.5
- chest    1 0.0003515 0.025224 -2261.7
- age      1 0.0004072 0.025279 -2261.2
- thigh    1 0.0005324 0.025405 -2259.9
<none>                 0.024872 -2259.7
- height   1 0.0007020 0.025574 -2258.3
- wrist    1 0.0014981 0.026370 -2250.6
- abdomen  1 0.0102551 0.035127 -2178.9
```

AIC, BIC+ optimization method

```
Step:  AIC=-2262.5
density ~ age + height + chest + abdomen + thigh + wrist

          Df Sum of Sq      RSS     AIC
- thigh    1 0.0002844 0.025428 -2265.2
- chest    1 0.0004114 0.025555 -2264.0
- age      1 0.0005560 0.025700 -2262.6
<none>                 0.025144 -2262.5
- height   1 0.0010548 0.026198 -2257.8
- wrist    1 0.0016718 0.026815 -2251.9
- abdomen  1 0.0110972 0.036241 -2176.6
```

AIC, BIC+ optimization method

```
Step:  AIC=-2265.21
density ~ age + height + chest + abdomen + wrist

          Df Sum of Sq      RSS     AIC
- age      1 0.0002847 0.025713 -2267.9
- chest    1 0.0004104 0.025838 -2266.7
<none>                 0.025428 -2265.2
- height   1 0.0009925 0.026420 -2261.2
- wrist    1 0.0014372 0.026865 -2257.0
- abdomen  1 0.0171134 0.042541 -2142.1
```

AIC, BIC+ optimization method

```
Step:  AIC=-2267.95
density ~ height + chest + abdomen + wrist

          Df Sum of Sq      RSS     AIC
- chest    1 0.0005114 0.026224 -2268.6
<none>                 0.025713 -2267.9
- wrist    1 0.0012086 0.026921 -2262.0
- height   1 0.0016357 0.027348 -2258.1
- abdomen  1 0.0184554 0.044168 -2138.2
```

AIC, BIC+ optimization method

```
Step:  AIC=-2268.55
density ~ height + abdomen + wrist

          Df Sum of Sq      RSS     AIC
<none>                  0.026224 -2268.6
- height   1  0.001691 0.027915 -2258.4
- wrist    1  0.001782 0.028006 -2257.6
- abdomen  1  0.050426 0.076650 -2005.9
```

BIC

n large

penalty

stronger

AIC, BIC+ optimization method

```
> AIC.choice <- c(3,5:10,14,15)
> BIC.choice <- c(5,8,15)
```

Notice how BIC selects a very parsimonious model. However, it is consistent, as mentioned in the lecture.

AIC, BIC+ optimization method

To start form a base model,

```
> base <- lm(density~1,data=bodyfat)
> AIC.base <- step(base,direction="both",
+ scope=list(lower=~1,upper=model),trace=T)

Start:  AIC=-1995.51
density ~ 1

           Df Sum of Sq      RSS     AIC
+ abdomen   1  0.057243 0.033728 -2243.6
+ chest     1  0.041716 0.049255 -2148.1
+ hip       1  0.033159 0.057813 -2107.8
+ weight    1  0.031859 0.059112 -2102.2
+ thigh     1  0.027882 0.063089 -2085.7
+ knee      1  0.022467 0.068505 -2065.0
+ biceps    1  0.021403 0.069568 -2061.1
+ neck      1  0.020667 0.070305 -2058.4
+ forearm   1  0.011160 0.079811 -2026.5
+ wrist     1  0.009835 0.081136 -2022.3
+ age       1  0.007211 0.083760 -2014.3
+ ankle     1  0.006248 0.084724 -2011.4
+ height    1  0.000751 0.090220 -1995.6
<none>                 0.090971 -1995.5
```

each step
=T
final result

AIC, BIC+ optimization method

```
Step:  AIC=-2243.55
density ~ abdomen

          Df Sum of Sq      RSS     AIC
+ weight   1  0.005456 0.028272 -2286.0
+ wrist    1  0.003919 0.029809 -2272.7
+ neck     1  0.003115 0.030613 -2266.0
+ hip      1  0.003096 0.030632 -2265.8
+ height   1  0.002363 0.031365 -2259.9
+ knee     1  0.001536 0.032192 -2253.3
+ chest    1  0.001371 0.032356 -2252.0
+ ankle    1  0.001087 0.032641 -2249.8
+ age      1  0.000937 0.032791 -2248.7
+ thigh    1  0.000655 0.033072 -2246.5
+ biceps   1  0.000583 0.033145 -2245.9
+ forearm  1  0.000293 0.033435 -2243.8
<none>                0.033728 -2243.6
- abdomen  1  0.057243 0.090971 -1995.5
```

AIC, BIC+ optimization method

```
Step:  AIC=-2286.02
density ~ abdomen + weight

          Df Sum of Sq      RSS     AIC
+ wrist    1 0.0008900 0.027382 -2292.1
+ thigh    1 0.0007707 0.027501 -2291.0
+ biceps   1 0.0004949 0.027777 -2288.5
+ neck     1 0.0003767 0.027895 -2287.4
+ forearm  1 0.0003663 0.027906 -2287.3
<none>                 0.028272 -2286.0
+ height   1 0.0001762 0.028096 -2285.6
+ knee     1 0.0001083 0.028164 -2285.0
+ age      1 0.0000362 0.028236 -2284.3
+ ankle    1 0.0000352 0.028237 -2284.3
+ chest    1 0.0000269 0.028245 -2284.3
+ hip      1 0.0000030 0.028269 -2284.1
- weight   1 0.0054560 0.033728 -2243.6
- abdomen  1 0.0308404 0.059112 -2102.2
```

AIC, BIC+ optimization method

```
Step:  AIC=-2292.08
density ~ abdomen + weight + wrist

          Df Sum of Sq      RSS     AIC
+ forearm  1 0.0007064 0.026676 -2296.7
+ biceps   1 0.0006591 0.026723 -2296.2
+ thigh    1 0.0004666 0.026915 -2294.4
<none>                 0.027382 -2292.1
+ knee     1 0.0001902 0.027192 -2291.8
+ ankle    1 0.0001507 0.027231 -2291.5
+ height   1 0.0000935 0.027288 -2290.9
+ neck     1 0.0000801 0.027302 -2290.8
+ hip      1 0.0000782 0.027304 -2290.8
+ age      1 0.0000629 0.027319 -2290.7
+ chest    1 0.0000053 0.027377 -2290.1
- wrist    1 0.0008900 0.028272 -2286.0
- weight   1 0.0024271 0.029809 -2272.7
- abdomen  1 0.0296632 0.057045 -2109.1
```

AIC, BIC+ optimization method

```
Step:  AIC=-2296.67
density ~ abdomen + weight + wrist + forearm

          Df Sum of Sq      RSS     AIC
+ thigh    1 0.0003590 0.026317 -2298.1
+ biceps   1 0.0003046 0.026371 -2297.6
<none>                 0.026676 -2296.7
+ neck     1 0.0001933 0.026482 -2296.5
+ knee     1 0.0001851 0.026491 -2296.4
+ ankle    1 0.0001743 0.026501 -2296.3
+ age      1 0.0001357 0.026540 -2295.9
+ height   1 0.0000685 0.026607 -2295.3
+ chest    1 0.0000441 0.026631 -2295.1
+ hip      1 0.0000375 0.026638 -2295.0
- forearm  1 0.0007064 0.027382 -2292.1
- wrist    1 0.0012300 0.027906 -2287.3
- weight   1 0.0030541 0.029730 -2271.3
- abdomen  1 0.0303656 0.057041 -2107.1
```

AIC, BIC+ optimization method

```
Step:  AIC=-2298.08
density ~ abdomen + weight + wrist + forearm + thigh

          Df Sum of Sq      RSS     AIC
+ age      1 0.0004108 0.025906 -2300.0
+ hip      1 0.0003010 0.026016 -2299.0
<none>                 0.026317 -2298.1
+ biceps   1 0.0001792 0.026137 -2297.8
+ neck     1 0.0001784 0.026138 -2297.8
+ ankle    1 0.0001439 0.026173 -2297.5
+ knee     1 0.0000816 0.026235 -2296.9
- thigh    1 0.0003590 0.026676 -2296.7
+ height   1 0.0000111 0.026305 -2296.2
+ chest    1 0.0000036 0.026313 -2296.1
- forearm  1 0.0005988 0.026915 -2294.4
- wrist    1 0.0008663 0.027183 -2291.9
- weight   1 0.0030598 0.029376 -2272.4
- abdomen  1 0.0305230 0.056840 -2106.0
```

AIC, BIC+ optimization method

```
Step:  AIC=-2300.04
density ~ abdomen + weight + wrist + forearm + thigh + age

          Df Sum of Sq      RSS     AIC
+ hip      1 0.0002535 0.025652 -2300.5
+ neck     1 0.0002346 0.025671 -2300.3
<none>                 0.025906 -2300.0
+ ankle    1 0.0001842 0.025722 -2299.8
+ biceps   1 0.0001364 0.025769 -2299.4
+ knee     1 0.0000267 0.025879 -2298.3
+ chest    1 0.0000105 0.025895 -2298.2
- age      1 0.0004108 0.026317 -2298.1
+ height   1 0.0000011 0.025905 -2298.1
- thigh    1 0.0006340 0.026540 -2295.9
- forearm  1 0.0006999 0.026606 -2295.3
- wrist    1 0.0012319 0.027138 -2290.3
- weight   1 0.0020746 0.027980 -2282.6
- abdomen  1 0.0161930 0.042099 -2179.7
```

AIC, BIC+ optimization method

```
Step:  AIC=-2300.52
density ~ abdomen + weight + wrist + forearm + thigh + age +
    hip

          Df Sum of Sq      RSS     AIC
+ neck     1 0.0003550 0.025297 -2302.0
<none>                 0.025652 -2300.5
+ ankle    1 0.0001844 0.025468 -2300.3
- hip      1 0.0002535 0.025906 -2300.0
+ biceps   1 0.0001003 0.025552 -2299.5
- age      1 0.0003633 0.026016 -2299.0
+ knee     1 0.0000414 0.025611 -2298.9
+ chest    1 0.0000348 0.025617 -2298.9
+ height   1 0.0000195 0.025633 -2298.7
- forearm  1 0.0005375 0.026190 -2297.3
- thigh    1 0.0008757 0.026528 -2294.1
- weight   1 0.0009991 0.026651 -2292.9
- wrist    1 0.0012953 0.026948 -2290.1
- abdomen  1 0.0161687 0.041821 -2179.3
```

AIC, BIC+ optimization method

```
Step:  AIC=-2302.03
density ~ abdomen + weight + wrist + forearm + thigh + age +
    hip + neck

          Df Sum of Sq      RSS     AIC
<none>                 0.025297 -2302.0
+ biceps   1 0.0001485 0.025149 -2301.5
+ ankle    1 0.0001224 0.025175 -2301.3
- neck     1 0.0003550 0.025652 -2300.5
- hip      1 0.0003739 0.025671 -2300.3
+ height   1 0.0000296 0.025268 -2300.3
+ chest    1 0.0000207 0.025277 -2300.2
+ knee     1 0.0000085 0.025289 -2300.1
- age      1 0.0004209 0.025718 -2299.9
- weight   1 0.0005549 0.025852 -2298.6
- forearm  1 0.0006673 0.025965 -2297.5
- wrist    1 0.0008957 0.026193 -2295.3
- thigh    1 0.0009760 0.026273 -2294.5
- abdomen  1 0.0164838 0.041781 -2177.6
```

AIC, BIC+ optimization method

```
> BIC.base <- step(base,direction="both",
+ scope=list(lower=~1,upper=model),trace=T,k=log(250))

Start:  AIC=-1991.99
density ~ 1

           Df Sum of Sq      RSS     AIC
+ abdomen   1  0.057243 0.033728 -2236.5
+ chest     1  0.041716 0.049255 -2141.1
+ hip       1  0.033159 0.057813 -2100.7
+ weight    1  0.031859 0.059112 -2095.1
+ thigh     1  0.027882 0.063089 -2078.7
+ knee      1  0.022467 0.068505 -2057.9
+ biceps    1  0.021403 0.069568 -2054.1
+ neck      1  0.020667 0.070305 -2051.4
+ forearm   1  0.011160 0.079811 -2019.5
+ wrist     1  0.009835 0.081136 -2015.3
+ age       1  0.007211 0.083760 -2007.3
+ ankle     1  0.006248 0.084724 -2004.4
<none>                  0.090971 -1992.0
+ height    1  0.000751 0.090220 -1988.6
```

AIC, BIC+ optimization method

```
Step:  AIC=-2236.51
density ~ abdomen

          Df Sum of Sq      RSS     AIC
+ weight   1  0.005456 0.028272 -2275.4
+ wrist    1  0.003919 0.029809 -2262.1
+ neck     1  0.003115 0.030613 -2255.4
+ hip      1  0.003096 0.030632 -2255.2
+ height   1  0.002363 0.031365 -2249.3
+ knee     1  0.001536 0.032192 -2242.7
+ chest    1  0.001371 0.032356 -2241.4
+ ankle    1  0.001087 0.032641 -2239.2
+ age      1  0.000937 0.032791 -2238.1
<none>                0.033728 -2236.5
+ thigh    1  0.000655 0.033072 -2235.9
+ biceps   1  0.000583 0.033145 -2235.4
+ forearm  1  0.000293 0.033435 -2233.2
- abdomen  1  0.057243 0.090971 -1992.0
```

AIC, BIC+ optimization method

```
Step:  AIC=-2275.45
density ~ abdomen + weight

          Df Sum of Sq      RSS     AIC
+ wrist    1 0.0008900 0.027382 -2278.0
+ thigh    1 0.0007707 0.027501 -2276.9
<none>                 0.028272 -2275.4
+ biceps   1 0.0004949 0.027777 -2274.4
+ neck     1 0.0003767 0.027895 -2273.3
+ forearm  1 0.0003663 0.027906 -2273.2
+ height   1 0.0001762 0.028096 -2271.5
+ knee     1 0.0001083 0.028164 -2270.9
+ age      1 0.0000362 0.028236 -2270.3
+ ankle    1 0.0000352 0.028237 -2270.2
+ chest    1 0.0000269 0.028245 -2270.2
+ hip      1 0.0000030 0.028269 -2270.0
- weight   1 0.0054560 0.033728 -2236.5
- abdomen  1 0.0308404 0.059112 -2095.1
```

AIC, BIC+ optimization method

```
Step:  AIC=-2277.99
density ~ abdomen + weight + wrist

          Df Sum of Sq      RSS     AIC
+ forearm  1 0.0007064 0.026676 -2279.1
+ biceps   1 0.0006591 0.026723 -2278.6
<none>                 0.027382 -2278.0
+ thigh    1 0.0004666 0.026915 -2276.8
- wrist    1 0.0008900 0.028272 -2275.4
+ knee     1 0.0001902 0.027192 -2274.2
+ ankle    1 0.0001507 0.027231 -2273.9
+ height   1 0.0000935 0.027288 -2273.3
+ neck     1 0.0000801 0.027302 -2273.2
+ hip      1 0.0000782 0.027304 -2273.2
+ age      1 0.0000629 0.027319 -2273.1
+ chest    1 0.0000053 0.027377 -2272.5
- weight   1 0.0024271 0.029809 -2262.1
- abdomen  1 0.0296632 0.057045 -2098.6
```

AIC, BIC+ optimization method

```
Step:  AIC=-2279.06
density ~ abdomen + weight + wrist + forearm

          Df Sum of Sq      RSS     AIC
<none>                 0.026676 -2279.1
- forearm  1 0.0007064 0.027382 -2278.0
+ thigh    1 0.0003590 0.026317 -2276.9
+ biceps   1 0.0003046 0.026371 -2276.4
+ neck     1 0.0001933 0.026482 -2275.4
+ knee     1 0.0001851 0.026491 -2275.3
+ ankle    1 0.0001743 0.026501 -2275.2
+ age      1 0.0001357 0.026540 -2274.8
+ height   1 0.0000685 0.026607 -2274.2
+ chest    1 0.0000441 0.026631 -2273.9
+ hip      1 0.0000375 0.026638 -2273.9
- wrist    1 0.0012300 0.027906 -2273.2
- weight   1 0.0030541 0.029730 -2257.3
- abdomen  1 0.0303656 0.057041 -2093.1
```