Stepwise AIC

May 8, 2020

```
In [22]: library(MASS)
```

```
In [2]: df <- read.table('oil.txt', col.names = c('spirit','gravity','pressure','distil','endpo
```

0.1 Fit a model with the three best explanatory variables

```
In [4]: lm.3 <- lm(spirit ~ gravity + distil + endpoint, data=df)

        summary(lm.3)
```

```
Call:
lm(formula = spirit ~ gravity + distil + endpoint, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5303 -1.3606 -0.2681  1.3911  4.7658

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.032034   7.223341   0.558   0.5811
gravity      0.221727   0.102061   2.173   0.0384 *
distil      -0.186571   0.015922 -11.718 2.61e-12 ***
endpoint     0.156527   0.006462  24.224  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.283 on 28 degrees of freedom
Multiple R-squared:  0.959,Adjusted R-squared:  0.9546
F-statistic: 218.5 on 3 and 28 DF,  p-value: < 2.2e-16
```

0.2 Using AIC (Akaike Information Criterion) to perform systematic, stepwise model fitting, starting with a base model with $p = 1$

A statistic that provides a SS_R Vs p trade off

$$AIC = n \ln \frac{SS_R}{n} + 2p + \text{constant}$$

```
In [17]: # essentially just fitting beta_zero paramater estimate, which is the sample mean of
         lm.base <- lm(spirit ~ 1, data=df)

In [20]: mean(df[,'spirit'])
```

19.659375

```
In [21]: summary(lm.base)
```

```
Call:
lm(formula = spirit ~ 1, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-16.859  -8.009  -1.859   7.391  26.041

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.659      1.895   10.37 1.33e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.72 on 31 degrees of freedom
```

So the stepwise AIC process will iterate through the remaining **scope** variables, and select the variable to add to the linear model regression that has the smallest AIC value

It will do this for all variables in the scope

The sum of squares is the **REGRESSION SUM OF SQUARES** as a result of adding the variable

The RSS is the **RESIDUAL SUM OF SQUARES** as a result of adding the variable

So your AIC will be based on the RSS and the number of parameters - which will just be a delta of 1 per step (as you're just looking at adding a single variable per step).

We actually see that the AIC score increases rather than decreases at the point when it adds **gravity** to the model.

Also note that the RSS - SS_R is always decreasing.

```
In [23]: stepAIC(lm.base, ~ gravity + pressure + distil + endpoint, data=df)
```

```
Start:  AIC=152.81
spirit ~ 1

           Df Sum of Sq    RSS    AIC
+ endpoint  1   1804.38 1759.7 132.23
+ pressure  1    525.74 3038.3 149.71
+ distil    1    353.70 3210.4 151.47
+ gravity   1    216.26 3347.8 152.81
<none>                  3564.1 152.81

Step:  AIC=132.23
spirit ~ endpoint

           Df Sum of Sq    RSS     AIC
+ distil    1   1589.08  170.6  59.557
+ pressure  1   1389.83  369.9  84.317
+ gravity   1    897.75  861.9 111.391
<none>                  1759.7 132.229
- endpoint  1   1804.38 3564.1 152.814

Step:  AIC=59.56
spirit ~ endpoint + distil

           Df Sum of Sq    RSS     AIC
+ gravity   1     24.61  146.0  56.572
<none>                   170.6  59.557
+ pressure  1      9.99  160.6  59.626
- distil    1   1589.08 1759.7 132.229
- endpoint  1   3039.77 3210.4 151.469

Step:  AIC=56.57
spirit ~ endpoint + distil + gravity

           Df Sum of Sq    RSS     AIC
+ pressure  1     11.20  134.8  56.019
<none>                   146.0  56.572
- gravity   1     24.61  170.6  59.557
- distil    1    715.95  861.9 111.391
- endpoint  1   3059.74 3205.7 153.423

Step:  AIC=56.02
spirit ~ endpoint + distil + gravity + pressure

           Df Sum of Sq     RSS     AIC
<none>                   134.80  56.019
- pressure  1     11.20  146.00  56.572
- gravity   1     25.82  160.62  59.626
- distil    1    130.68  265.48  75.706
```

```
- endpoint  1   2873.95 3008.76 153.393
```

```
Call:
lm(formula = spirit ~ endpoint + distil + gravity + pressure,
    data = df)

Coefficients:
(Intercept)     endpoint       distil      gravity     pressure
    -6.8208       0.1547      -0.1495       0.2272       0.5537
```

In [ ]: