STAT 420

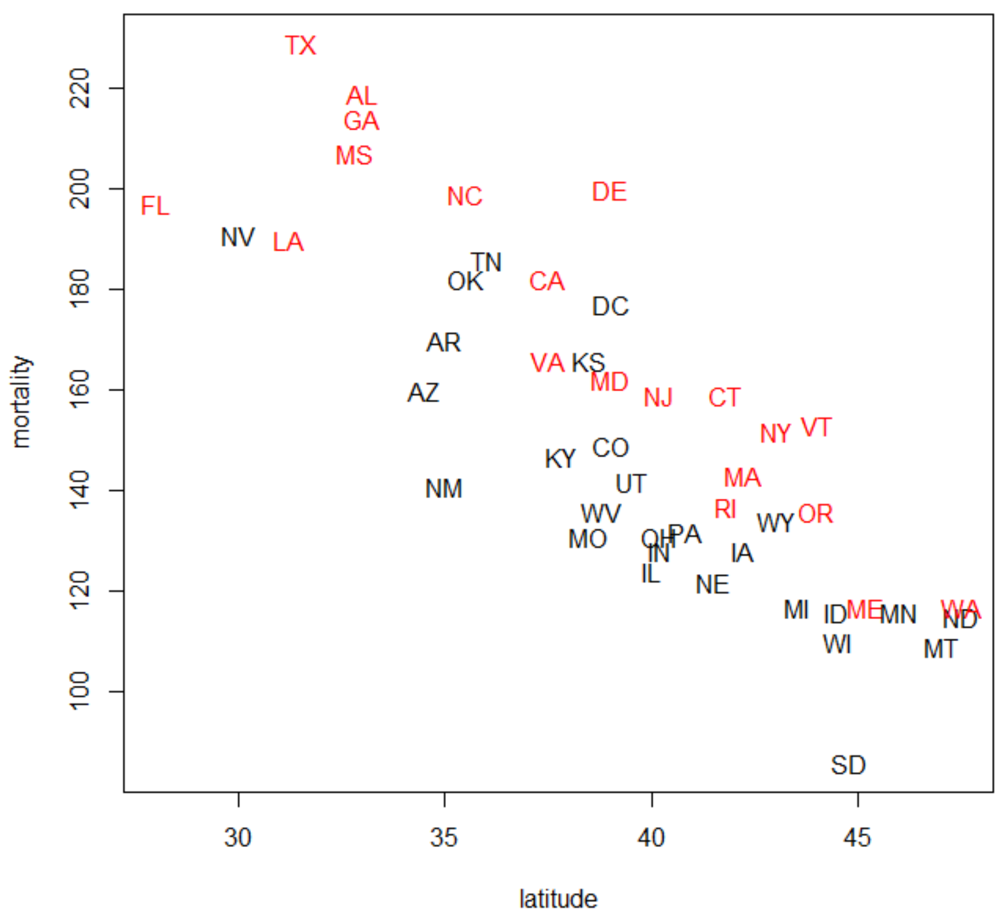
1. `skin.csv` contains the average annual mortality due to malignant melanoma for white males during 1950–1959 per 10 mil, for each state and the District of Columbia (Alaska and Hawaii [and New Hampshire, and South Carolina] are excluded), the latitude at the centroid of the state, and whether the state borders an ocean. (Fisher and Van Belle (1993). *Biostatistics: A methodology for the health sciences.*)

```
> plot(latitude, mortality, type="n")
```

to create an "empty" plot (`type="n"` for no plotting)

```
> text(latitude, mortality, as.character(state), col=ocean+1)
```

to add text from `state` to the plot at locations `(latitude, mortality)`.



Consider the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where x_1 is the latitude at the centroid of the state and x_2 is the dummy variable ($x_2 = 1$ for a state that borders an ocean, $x_2 = 0$ a state that does not).

For the states that do not border an ocean:

$$Y = \beta_0 + \beta_1 x_1 + e.$$

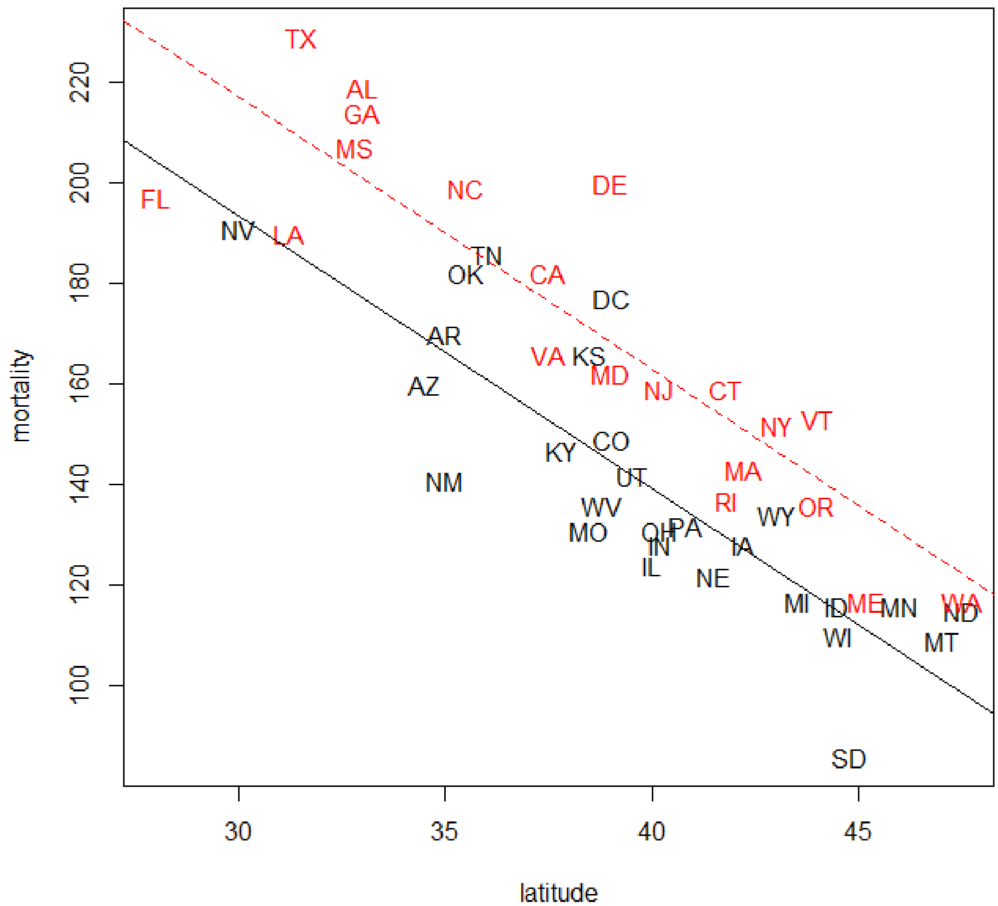
For the states that do border an ocean:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 + e = (\beta_0 + \beta_2) + \beta_1 x_1 + e.$$

The dummy variable splits the regression relationship into two parallel lines, one for each level (0 or 1) of the qualitative dummy variable. The distance between the two parallel lines (measured as the distance between the two y-intercepts) is equal to the estimated coefficient of the dummy variable x_2.

```
> fit = lm(mortality ~ latitude + ocean)
> summary(fit)

Call:
lm(formula = mortality ~ latitude + ocean)

Residuals:
    Min      1Q  Median      3Q     Max 
-31.065  -9.118  -2.384  10.036  32.290 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 355.6328    18.9405  18.776  < 2e-16 ***
latitude     -5.4083     0.4668 -11.586 5.90e-15 ***
ocean        23.8640     4.4813   5.325 3.27e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 14.94 on 44 degrees of freedom
Multiple R-squared: 0.8126,     Adjusted R-squared: 0.8041 
F-statistic:  95.4 on 2 and 44 DF,  p-value: < 2.2e-16
```

```
> abline(fit$coeff[1],fit$coeff[2],col=1,lty=1)
> abline(fit$coeff[1]+fit$coeff[3],fit$coeff[2],col=2,lty=2)
```



\- -

Consider the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$.

For the states that do not border an ocean:

$$Y = \beta_0 + \beta_1 x_1 + e.$$

For the states that do border an ocean:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 + \beta_3 x_1 + e = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + e.$$

Two lines, not necessarily parallel.

$H_0\colon\ \beta_3 = 0$

```
> fit2 = lm(mortality ~ latitude + ocean + I(latitude*ocean))
> summary(fit2)

Call:
lm(formula = mortality ~ latitude + ocean + I(latitude * ocean))

Residuals:
    Min       1Q   Median       3Q      Max
-32.243  -8.829  -0.994   9.431  32.442

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         350.1125    28.2987  12.372 9.26e-16 ***
latitude             -5.2706     0.7019  -7.509 2.38e-09 ***
ocean                33.7399    37.5582   0.898    0.374
I(latitude * ocean)  -0.2511     0.9481  -0.265    0.792
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.1 on 43 degrees of freedom
Multiple R-squared: 0.8129,     Adjusted R-squared: 0.7999
F-statistic: 62.28 on 3 and 43 DF,  p-value: 1.078e-15
```

OR

```
> anova(fit,fit2)
Analysis of Variance Table

Model 1: mortality ~ latitude + ocean
Model 2: mortality ~ latitude + ocean + I(latitude * ocean)
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     44 9826.5
2     43 9810.5  1    16.008 0.0702 0.7924
```

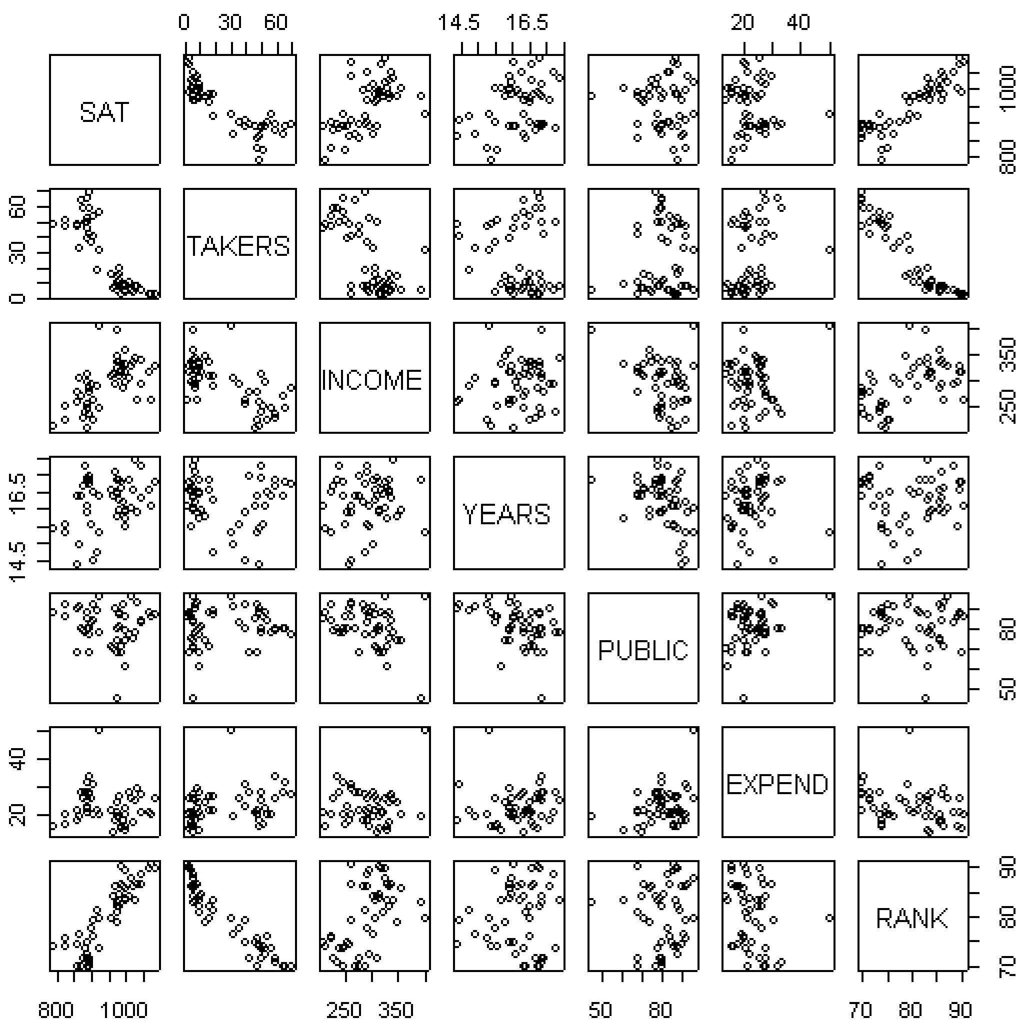
Do NOT Reject $H_0\colon\ \beta_3 = 0$ at any reasonable level of significance.

Examples for 10/16/2012 (2)

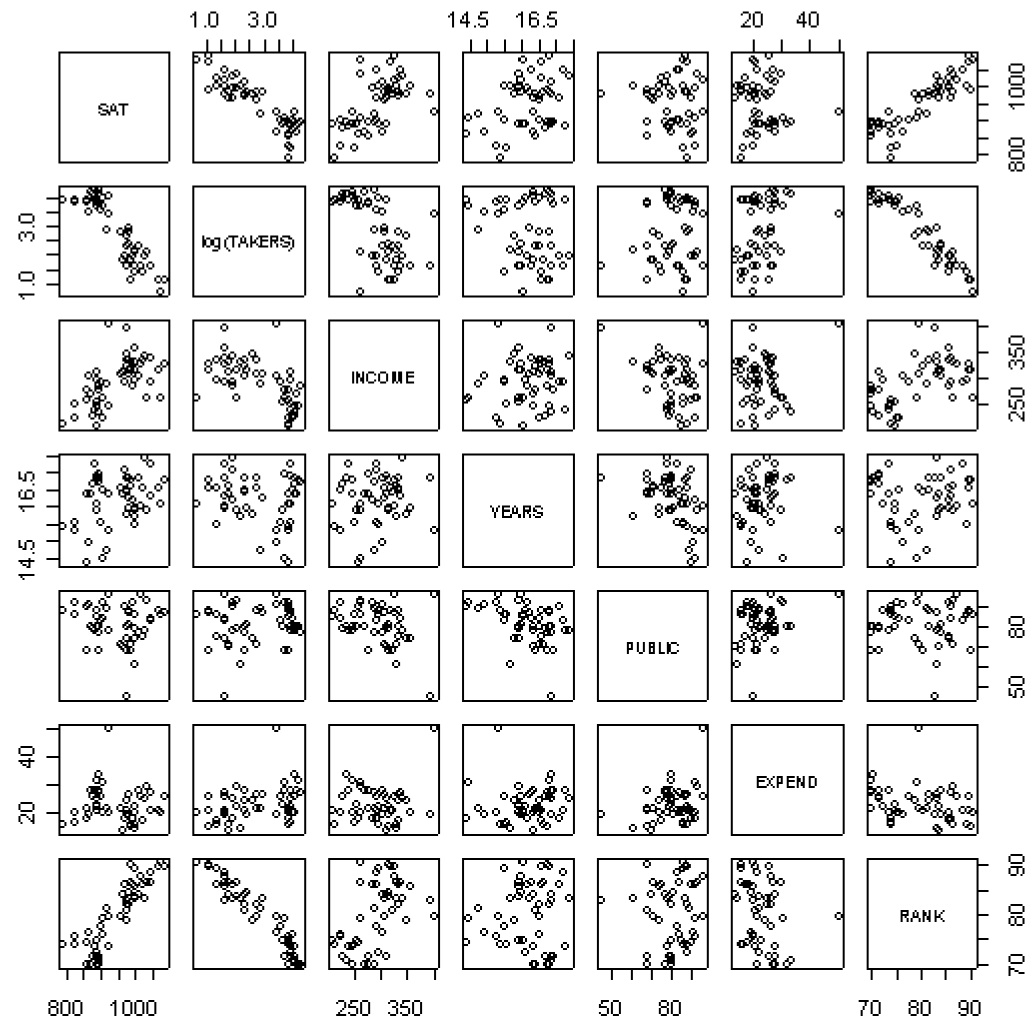
1. The worksheet `case1201.csv` contains data on the average SAT scores by state. The states have been ordered by how well their students did on the SAT on average. Researchers have tried to explain the state by state differences in scores. Column 2 is the average SAT scores, along with six variables that may be associated with the SAT differences among states: percentage of the total eligible students who took the exam, median income of families of test takers, average number of years that the test takers had formal studies in social studies, natural sciences, humanities, percentage of test takers who attended public secondary schools, total state expenditure on secondary schools (dollars per student), and median percentile ranking of the test takers within their secondary school classes.

```
> case1201.dat = read.table(" ... /case1201.csv", sep=",", header=T)
```

```
> pairs(SAT ~ TAKERS+INCOME+YEARS+PUBLIC+EXPEND+RANK, case1201.dat)
```



```
> pairs(SAT ~ log(TAKERS)+INCOME+YEARS+PUBLIC+EXPEND+RANK, case1201.dat)
```




```
> case1201.dat = subset(case1201.dat, STATE != "Alaska")

> case1201.fit = lm(SAT ~ log(TAKERS)+INCOME+YEARS+PUBLIC+EXPEND+RANK,
case1201.dat)

> summary(case1201.fit)

Call:
lm(formula = SAT ~ log(TAKERS) + INCOME + YEARS + PUBLIC + EXPEND +
    RANK, data = case1201.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-47.447 -10.361  -2.626  11.101  59.001
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 287.5242   259.4170   1.108   0.2740
log(TAKERS) -30.2149    14.7079  -2.054   0.0462 *
INCOME        0.1029     0.1259   0.817   0.4183
YEARS        13.1073     5.8798   2.229   0.0312 *
PUBLIC       -0.1011     0.5105  -0.198   0.8439
EXPEND        3.9367     0.8486   4.639 3.40e-05 ***
RANK          5.2738     2.2997   2.293   0.0269 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 22.57 on 42 degrees of freedom
Multiple R-Squared: 0.9128,     Adjusted R-squared: 0.9003
F-statistic: 73.28 on 6 and 42 DF,  p-value: < 2.2e-16
```

BACKWARD ELIMINATION

Set $\alpha_{crit} = 0.10$ or 0.05.

`PUBLIC` is the least significant variable, p-value = 0.8439.

```
> case1201.fit1 = update(case1201.fit, .~. - PUBLIC)

> summary(case1201.fit1)

Call:
lm(formula = SAT ~ log(TAKERS) + INCOME + YEARS + EXPEND + RANK,
    data = case1201.dat)

Residuals:
   Min     1Q Median     3Q    Max
-47.73 -10.27  -2.73  10.79  59.38

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 291.1605   255.8598   1.138   0.2614
log(TAKERS) -31.1553    13.7646  -2.263   0.0287 *
INCOME        0.1135     0.1126   1.007   0.3194
YEARS        13.4921     5.4875   2.459   0.0180 *
EXPEND        3.8718     0.7739   5.003 1.00e-05 ***
RANK          5.0601     2.0084   2.520   0.0155 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

```
Residual standard error: 22.32 on 43 degrees of freedom
Multiple R-Squared: 0.9127,     Adjusted R-squared: 0.9026
F-statistic: 89.93 on 5 and 43 DF,  p-value: < 2.2e-16
```

`INCOME` is the least significant variable, p-value = 0.3194.

```
> case1201.fit2 = update(case1201.fit1, .~. - INCOME)

> summary(case1201.fit2)

Call:
lm(formula = SAT ~ log(TAKERS) + YEARS + EXPEND + RANK, data =
case1201.dat)

Residuals:
     Min       1Q   Median       3Q      Max
-52.3043  -9.9170   0.5963  11.8798  59.2026

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 399.1147   232.3716   1.718  0.09291 .
log(TAKERS) -38.1005    11.9152  -3.198  0.00257 **
YEARS        13.1473     5.4778   2.400  0.02069 *
EXPEND        3.9957     0.7642   5.228 4.52e-06 ***
RANK          4.4003     1.8989   2.317  0.02520 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 22.32 on 44 degrees of freedom
Multiple R-Squared: 0.9107,     Adjusted R-squared: 0.9025
F-statistic: 112.1 on 4 and 44 DF,  p-value: < 2.2e-16
```

All variables are significant at α_{crit}.

```
> anova(case1201.fit2,case1201.fit)
Analysis of Variance Table

Model 1: SAT ~ log(TAKERS) + YEARS + EXPEND + RANK
Model 2: SAT ~ log(TAKERS) + INCOME + YEARS + PUBLIC + EXPEND + RANK
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     44 21922.1
2     42 21396.7  2     525.4 0.5156 0.6009
```

- -

AKAIKE'S INFORMATION CRITERION (AIC):

Akaike proposed to choose the model that minimises

$$\text{AIC} = -2 \times (\text{Maximized log-likelihood}) + 2 \times (\text{number of parameters in the model})$$
$$= n + n \ln(2\pi) + n \ln\left({}^{\text{RSS}}\!/_{n}\right) + 2p$$

R: $\text{AIC} = n \ln\left({}^{\text{RSS}}\!/_{n}\right) + 2p$

BAYESIAN INFORMATION CRITERION (BIC):

$$\text{BIC} = -2 \times (\text{Maximized log-likelihood}) + \ln(n) \times (\text{number of parameters in the model})$$

```
> RSS = sum(case1201.fit$residuals^2)
> RSS
[1] 21396.74

> 49*log(RSS/49)+2*7
         [,1]
[1,] 311.8795
> extractAIC(case1201.fit)
[1]   7.0000 311.8795
```

$$\text{AIC} = 49 + 49 \ln(2\pi) + 49 \ln\left({}^{21396.74}\!/_{49}\right) + 2 \times 7 = \mathbf{450.9355}.$$

BACKWARD ELIMINATION

```
> step(case1201.fit, direction = "backward")
Start:  AIC= 311.88
 SAT ~ log(TAKERS) + INCOME + YEARS + PUBLIC + EXPEND + RANK

              Df Sum of Sq   RSS   AIC
- PUBLIC       1        20 21417   310
- INCOME       1       340 21737   311
<none>                     21397   312
- log(TAKERS)  1      2150 23547   315
- YEARS        1      2532 23928   315
- RANK         1      2679 24076   316
- EXPEND       1     10964 32361   330
```

```
Step:  AIC= 309.93
 SAT ~ log(TAKERS) + INCOME + YEARS + EXPEND + RANK

              Df Sum of Sq   RSS   AIC
- INCOME       1       505 21922   309
<none>                     21417   310
- log(TAKERS)  1      2552 23968   313
- YEARS        1      3011 24428   314
- RANK         1      3162 24578   315
- EXPEND       1     12465 33882   330

Step:  AIC= 309.07
 SAT ~ log(TAKERS) + YEARS + EXPEND + RANK

              Df Sum of Sq   RSS   AIC
<none>                     21922   309
- RANK         1      2676 24598   313
- YEARS        1      2870 24792   313
- log(TAKERS)  1      5094 27016   317
- EXPEND       1     13620 35542   331

Call:
lm(formula = SAT ~ log(TAKERS) + YEARS + EXPEND + RANK, data =
case1201.dat)

Coefficients:
(Intercept)  log(TAKERS)        YEARS       EXPEND         RANK
    399.115      -38.100       13.147        3.996        4.400
```

Another way:

```
> drop1(case1201.fit)
Single term deletions

Model:
SAT ~ log(TAKERS) + INCOME + YEARS + PUBLIC + EXPEND + RANK
            Df Sum of Sq   RSS   AIC
<none>                   21397   312
log(TAKERS)  1      2150 23547   315
INCOME       1       340 21737   311
YEARS        1      2532 23928   315
PUBLIC       1        20 21417   310
EXPEND       1     10964 32361   330
RANK         1      2679 24076   316
```

AIC will be lowest, 310, if `PUBLIC` is dropped.

```
> case1201.fit1 = update(case1201.fit, .~. - PUBLIC)
```

```
> drop1(case1201.fit1)
Single term deletions

Model:
SAT ~ log(TAKERS) + INCOME + YEARS + EXPEND + RANK
            Df Sum of Sq   RSS   AIC
<none>                   21417   310
log(TAKERS)  1      2552 23968   313
INCOME       1       505 21922   309
YEARS        1      3011 24428   314
EXPEND       1     12465 33882   330
RANK         1      3162 24578   315
```

AIC will be lowest, 309, if `INCOME` is dropped.

```
> case1201.fit2 = update(case1201.fit1, .~. - INCOME)

> drop1(case1201.fit2)
Single term deletions

Model:
SAT ~ log(TAKERS) + YEARS + EXPEND + RANK
            Df Sum of Sq   RSS   AIC
<none>                   21922   309
log(TAKERS)  1      5094 27016   317
YEARS        1      2870 24792   313
EXPEND       1     13620 35542   331
RANK         1      2676 24598   313
```

Dropping any of the remaining variables will result in higher AIC.

FORWARD SELECTION

```
> attach(case1201.dat)
> step(lm(SAT ~ 1), SAT ~ log(TAKERS)+INCOME+YEARS+PUBLIC+EXPEND+RANK,
direction = "forward")
Start:  AIC= 419.42
 SAT ~ 1

              Df Sum of Sq    RSS    AIC
+ log(TAKERS)  1    199007  46369    340
+ RANK         1    190297  55079    348
+ INCOME       1    102026 143350    395
+ YEARS        1     26338 219038    416
<none>                     245376    419
+ PUBLIC       1      1232 244144    421
+ EXPEND       1       386 244991    421
```

```
Step:  AIC= 339.78
 SAT ~ log(TAKERS)

         Df Sum of Sq   RSS   AIC
+ EXPEND  1     20523 25846   313
+ YEARS   1      6364 40006   335
<none>                46369   340
+ RANK    1       871 45498   341
+ INCOME  1       785 45584   341
+ PUBLIC  1       449 45920   341


Step:  AIC= 313.14
 SAT ~ log(TAKERS) + EXPEND

         Df Sum of Sq     RSS     AIC
+ YEARS   1    1248.2 24597.6   312.7
+ RANK    1    1053.6 24792.2   313.1
<none>                25845.8   313.1
+ INCOME  1      53.3 25792.5   315.0
+ PUBLIC  1       1.3 25844.5   315.1


Step:  AIC= 312.71
 SAT ~ log(TAKERS) + EXPEND + YEARS

         Df Sum of Sq     RSS     AIC
+ RANK    1    2675.5 21922.1   309.1
<none>                24597.6   312.7
+ PUBLIC  1     287.8 24309.8   314.1
+ INCOME  1      19.2 24578.4   314.7


Step:  AIC= 309.07
 SAT ~ log(TAKERS) + EXPEND + YEARS + RANK

         Df Sum of Sq     RSS     AIC
<none>                21922.1   309.1
+ INCOME  1     505.4 21416.7   309.9
+ PUBLIC  1     185.0 21737.1   310.7


Call:
lm(formula = SAT ~ log(TAKERS) + EXPEND + YEARS + RANK)

Coefficients:
(Intercept)  log(TAKERS)       EXPEND        YEARS         RANK
    399.115      -38.100        3.996       13.147        4.400
```

STEPWISE REGRESSION

```
> step(case1201.fit, direction = "both")
Start:  AIC= 311.88
 SAT ~ log(TAKERS) + INCOME + YEARS + PUBLIC + EXPEND + RANK

              Df Sum of Sq   RSS   AIC
- PUBLIC       1        20 21417   310
- INCOME       1       340 21737   311
<none>                     21397   312
- log(TAKERS)  1      2150 23547   315
- YEARS        1      2532 23928   315
- RANK         1      2679 24076   316
- EXPEND       1     10964 32361   330


Step:  AIC= 309.93
 SAT ~ log(TAKERS) + INCOME + YEARS + EXPEND + RANK

              Df Sum of Sq   RSS   AIC
- INCOME       1       505 21922   309
<none>                     21417   310
+ PUBLIC       1        20 21397   312
- log(TAKERS)  1      2552 23968   313
- YEARS        1      3011 24428   314
- RANK         1      3162 24578   315
- EXPEND       1     12465 33882   330


Step:  AIC= 309.07
 SAT ~ log(TAKERS) + YEARS + EXPEND + RANK

              Df Sum of Sq   RSS   AIC
<none>                     21922   309
+ INCOME       1       505 21417   310
+ PUBLIC       1       185 21737   311
- RANK         1      2676 24598   313
- YEARS        1      2870 24792   313
- log(TAKERS)  1      5094 27016   317
- EXPEND       1     13620 35542   331


Call:
lm(formula = SAT ~ log(TAKERS) + YEARS + EXPEND + RANK, data =
case1201.dat)

Coefficients:
(Intercept)  log(TAKERS)        YEARS       EXPEND         RANK
    399.115      -38.100       13.147        3.996        4.400
```

Examples for 10/16/2012 (3)

Mallows' C_p :

$$C_p = \frac{\text{SSResid}_{\text{New}}}{\text{MSResid}_{\text{Full}}} - n + 2 \cdot (\#\text{ of parameters in New}).$$

For Full model,

$$C_p = \frac{\text{SSResid}_{\text{Full}}}{\text{MSResid}_{\text{Full}}} - n + 2 \cdot (\#\text{ of parameters in Full}).$$

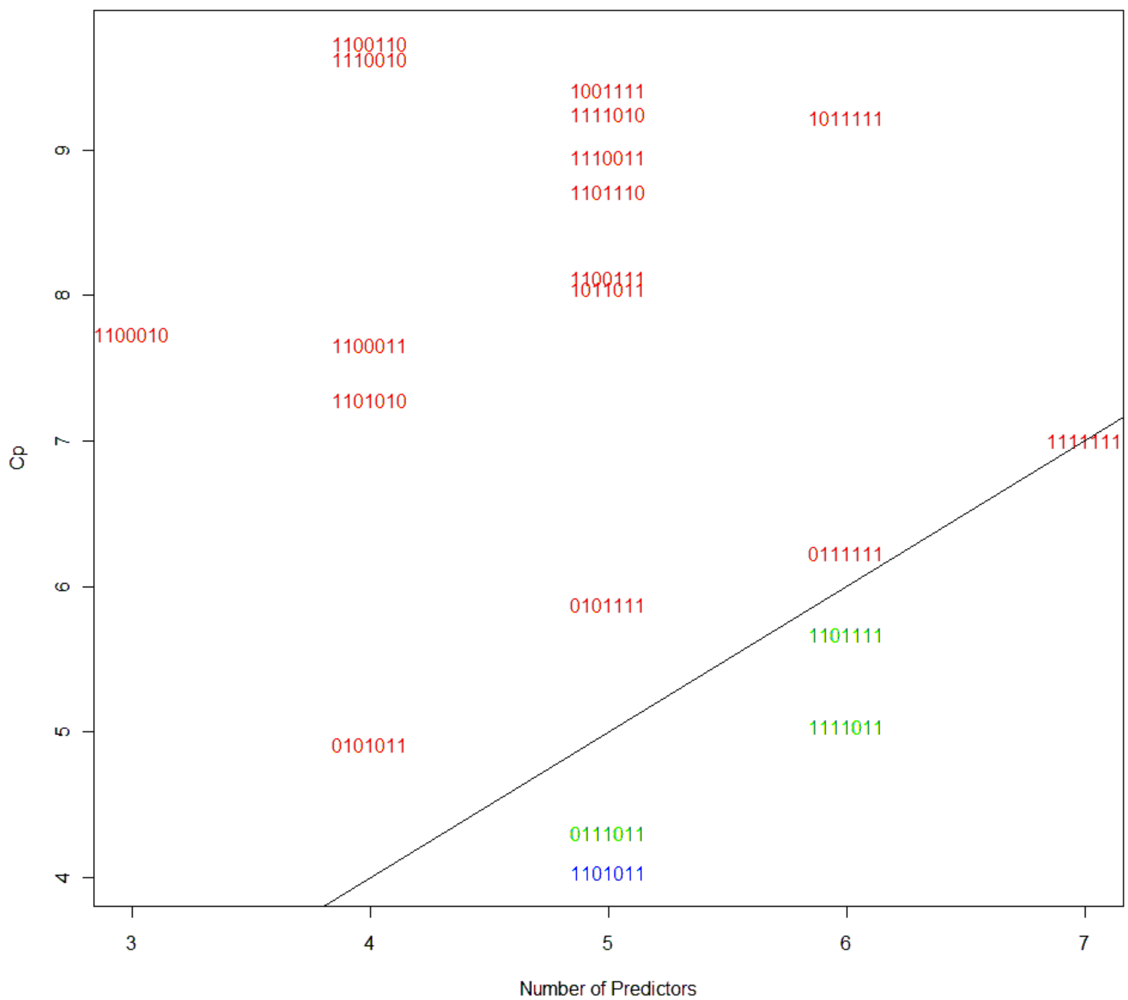
$$= [n - (\#\text{ of parameters in Full})] - n + 2 \cdot (\#\text{ of parameters in Full})$$

$$= (\#\text{ of parameters in Full}).$$

$\Rightarrow$ Want models with C_p close to or less than $(\#\text{ of parameters in New})$.

```
> case1201.dat = read.table(" . . . /case1201.csv", sep=",", header=T)
> case1201.dat = subset(case1201.dat, STATE != "Alaska")
>
> case1201.fit = lm(SAT ~ log(TAKERS)+INCOME+YEARS+PUBLIC+EXPEND+RANK,
+ data=case1201.dat)
>
> library(wle)
> mle.cp(case1201.fit)

Call:
mle.cp(formula = case1201.fit)


Mallows Cp:
     (Intercept) log(TAKERS) INCOME YEARS PUBLIC EXPEND RANK    cp
[1,]           1           1      0     1      0      1    1 4.031
[2,]           0           1      1     1      0      1    1 4.305
[3,]           1           1      1     1      0      1    1 5.039
[4,]           1           1      0     1      1      1    1 5.668
[5,]           1           1      1     1      1      1    1 7.000

Printed the first  5  best models
>
> case1201.fit2 = lm(SAT ~ log(TAKERS)+YEARS+EXPEND+RANK,
+ data=case1201.dat)
> SSResidNew = sum(case1201.fit2$residuals^2)
> MSResidFull = sum(case1201.fit$residuals^2)/(49-7)
>
> SSResidNew/MSResidFull - 49 + 2*5
[1] 4.031249
```

```
> mallows_cp = mle.cp(case1201.fit)
> plot(mallows_cp)
```



= =

Adjusted R-squared $= 1 - \dfrac{n-1}{n-p} \cdot \left(1 - R^2\right)$

```
Multiple R-Squared: 0.9128,     Adjusted R-squared: 0.9003
```

$$1 - \frac{n-1}{n-p} \cdot \left(1 - R^2\right) = 1 - \frac{49-1}{49-7} \cdot (1 - 0.9128) = 0.900343.$$