Regression Analysis lab 8

1 Variable selection and model building

1.1 Import data

```
cement<-read.csv(file="D:/chilo/Regression 8/cement.csv", header=T)
cement
```

```
       y x1 x2 x3 x4
1   78.5  7 26  6 60
2   74.3  1 29 15 52
3  104.3 11 56  8 20
4   87.6 11 31  8 47
5   95.9  7 52  6 33
6  109.2 11 55  9 22
7  102.7  3 71 17  6
8   72.5  1 31 22 44
9   93.1  2 54 18 22
10 115.9 21 47  4 26
11  83.8  1 40 23 34
12 113.3 11 66  9 12
13 109.4 10 68  8 12
```

2 All Subsets Regression

```
library(leaps)
attach(cement)
yy<-cement$y
yy
```

```
 [1]  78.5  74.3 104.3  87.6  95.9 109.2 102.7  72.5  93.1 115.9  83.8
[12] 113.3 109.4
```

```
xx<-cement[,2:5]
xx
```

```
   x1 x2 x3 x4
1   7 26  6 60
2   1 29 15 52
3  11 56  8 20
4  11 31  8 47
5   7 52  6 33
```

```
6  11 55  9 22
7   3 71 17  6
8   1 31 22 44
9   2 54 18 22
10 21 47  4 26
11  1 40 23 34
12 11 66  9 12
13 10 68  8 12

n<-length(yy)
leap1<-leaps(xx,yy, int=TRUE, names=names(cement)[2:5],
             method="Cp", nbest=16)
leap1

$which
     x1    x2    x3    x4
1 FALSE FALSE FALSE  TRUE
1 FALSE  TRUE FALSE FALSE
1  TRUE FALSE FALSE FALSE
1 FALSE FALSE  TRUE FALSE
2  TRUE  TRUE FALSE FALSE
2  TRUE FALSE FALSE  TRUE
2 FALSE FALSE  TRUE  TRUE
2 FALSE  TRUE  TRUE FALSE
2 FALSE  TRUE FALSE  TRUE
2  TRUE FALSE  TRUE FALSE
3  TRUE  TRUE FALSE  TRUE
3  TRUE  TRUE  TRUE FALSE
3  TRUE FALSE  TRUE  TRUE
3 FALSE  TRUE  TRUE  TRUE
4  TRUE  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "x1"          "x2"          "x3"          "x4"

$size
 [1] 2 2 2 2 3 3 3 3 3 4 4 4 4 5

$Cp
 [1] 138.731 142.486 202.549 315.154   2.678   5.496  22.373  62.438
 [9] 138.226 198.095   3.018   3.041   3.497   7.337   5.000

leap2<-leaps(xx,yy, int=TRUE, names=names(cement)[2:5],
             method="r2", nbest=16)
leap2
```

```
$which
     x1    x2    x3    x4
1 FALSE FALSE FALSE  TRUE
1 FALSE  TRUE FALSE FALSE
1  TRUE FALSE FALSE FALSE
1 FALSE FALSE  TRUE FALSE
2  TRUE  TRUE FALSE FALSE
2  TRUE FALSE FALSE  TRUE
2 FALSE FALSE  TRUE  TRUE
2 FALSE  TRUE  TRUE FALSE
2 FALSE  TRUE FALSE  TRUE
2  TRUE FALSE  TRUE FALSE
3  TRUE  TRUE FALSE  TRUE
3  TRUE  TRUE  TRUE FALSE
3  TRUE FALSE  TRUE  TRUE
3 FALSE  TRUE  TRUE  TRUE
4  TRUE  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "x1"          "x2"          "x3"          "x4"

$size
 [1] 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5

$r2
 [1] 0.6745 0.6663 0.5339 0.2859 0.9787 0.9725 0.9353 0.8470 0.6801 0.5482
[11] 0.9823 0.9823 0.9813 0.9728 0.9824
```

```
leap3<-leaps(xx,yy, int=TRUE, names=names(cement)[2:5],
             method="adjr2", nbest=16)
leap3
```

```
$which
     x1    x2    x3    x4
1 FALSE FALSE FALSE  TRUE
1 FALSE  TRUE FALSE FALSE
1  TRUE FALSE FALSE FALSE
1 FALSE FALSE  TRUE FALSE
2  TRUE  TRUE FALSE FALSE
2  TRUE FALSE FALSE  TRUE
2 FALSE FALSE  TRUE  TRUE
2 FALSE  TRUE  TRUE FALSE
2 FALSE  TRUE FALSE  TRUE
2  TRUE FALSE  TRUE FALSE
3  TRUE  TRUE FALSE  TRUE
3  TRUE  TRUE  TRUE FALSE
```

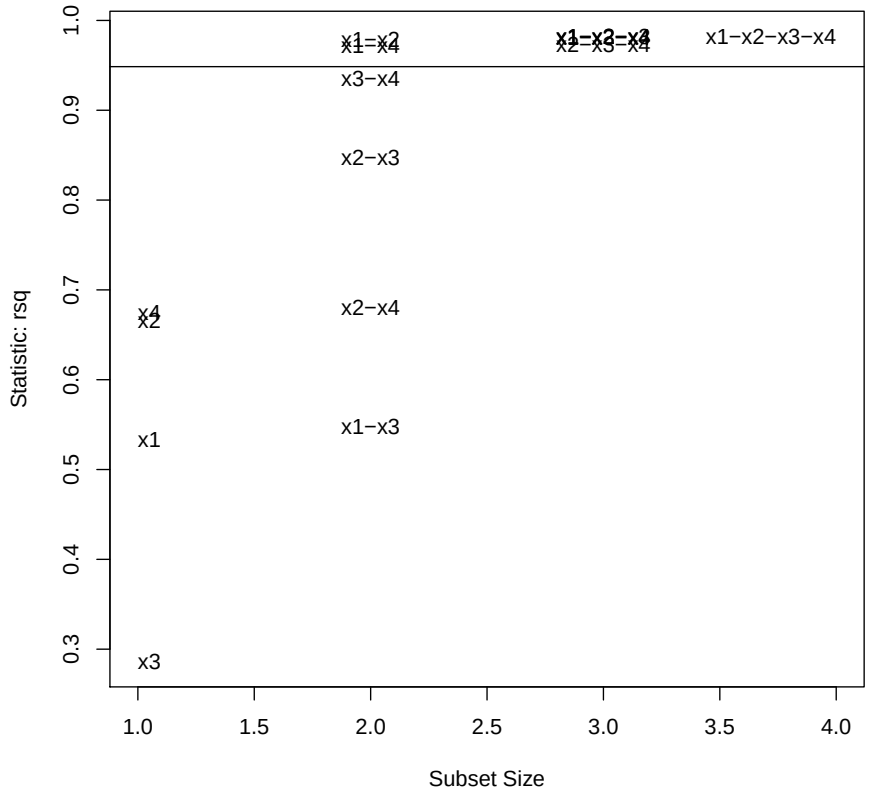
```
3  TRUE FALSE  TRUE  TRUE
3 FALSE  TRUE  TRUE  TRUE
4  TRUE  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "x1"          "x2"          "x3"          "x4"

$size
 [1] 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5

$adjr2
 [1] 0.6450 0.6359 0.4916 0.2210 0.9744 0.9670 0.9223 0.8164 0.6161 0.4578
[11] 0.9764 0.9764 0.9750 0.9638 0.9736

SST<-sum((yy-mean(yy))^2)
MSEp<-(1-leap3$adjr2)*SST/(n-1)
MSEp

 [1]  80.352  82.394 115.062 176.309   5.790   7.476  17.574  41.544
 [9]  86.888 122.707   5.330   5.346   5.648   8.202   5.983

# plot the results
leaps1<-regsubsets(y~x1+x2+x3+x4,data=cement,nbest=16)
summary(leaps1)

Subset selection object
Call: regsubsets.formula(y ~ x1 + x2 + x3 + x4, data = cement, nbest = 16)
4 Variables  (and intercept)
   Forced in Forced out
x1     FALSE      FALSE
x2     FALSE      FALSE
x3     FALSE      FALSE
x4     FALSE      FALSE
16 subsets of each size up to 4
Selection Algorithm: exhaustive
         x1  x2  x3  x4
1  ( 1 ) " " " " " " "*"
1  ( 2 ) " " "*" " " " "
1  ( 3 ) "*" " " " " " "
1  ( 4 ) " " " " "*" " "
2  ( 1 ) "*" "*" " " " "
2  ( 2 ) "*" " " " " "*"
2  ( 3 ) " " " " "*" "*"
2  ( 4 ) " " "*" "*" " "
2  ( 5 ) " " "*" " " "*"
2  ( 6 ) "*" " " "*" " "
```

```
3  ( 1 ) "*" "*" " " "*"
3  ( 2 ) "*" "*" "*" " "
3  ( 3 ) "*" " " "*" "*"
3  ( 4 ) " " "*" "*" "*"
4  ( 1 ) "*" "*" "*" "*"
```

```
library(car)
subsets(leaps1, statistic="rsq")
```

```
Error:  invalid coordinate lengths
```

```
abline(h=0.94855)
```




```
leaps2<-regsubsets(y~x1+x2+x3+x4,data=cement,nbest=16)
summary(leaps2)
```

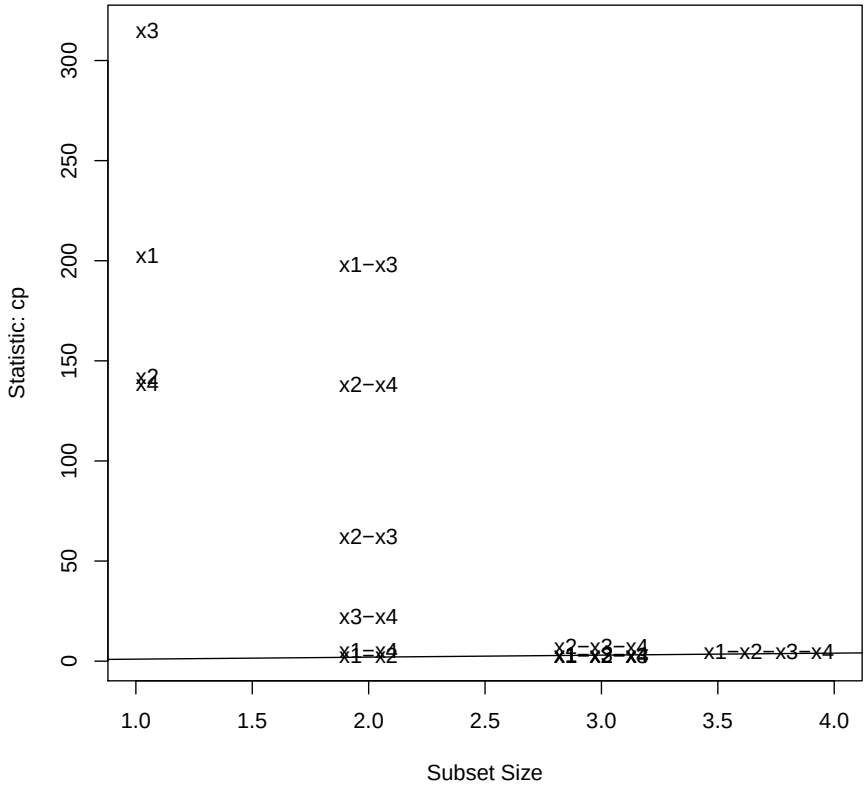
```
Subset selection object
```

```
Call: regsubsets.formula(y ~ x1 + x2 + x3 + x4, data = cement, nbest = 16)
4 Variables  (and intercept)
   Forced in Forced out
x1     FALSE      FALSE
x2     FALSE      FALSE
x3     FALSE      FALSE
x4     FALSE      FALSE
16 subsets of each size up to 4
Selection Algorithm: exhaustive
         x1  x2  x3  x4
1  ( 1 ) " " " " " " "*"
1  ( 2 ) " " "*" " " " "
1  ( 3 ) "*" " " " " " "
1  ( 4 ) " " " " "*" " "
2  ( 1 ) "*" "*" " " " "
2  ( 2 ) "*" " " " " "*"
2  ( 3 ) " " " " "*" "*"
2  ( 4 ) " " "*" "*" " "
2  ( 5 ) " " "*" " " "*"
2  ( 6 ) "*" " " "*" " "
3  ( 1 ) "*" "*" " " "*"
3  ( 2 ) "*" "*" "*" " "
3  ( 3 ) "*" " " "*" "*"
3  ( 4 ) " " "*" "*" "*"
4  ( 1 ) "*" "*" "*" "*"
```

```
library(car)
subsets(leaps2, statistic="cp")
```

```
Error:  invalid coordinate lengths
```

```
abline(0,1)
```

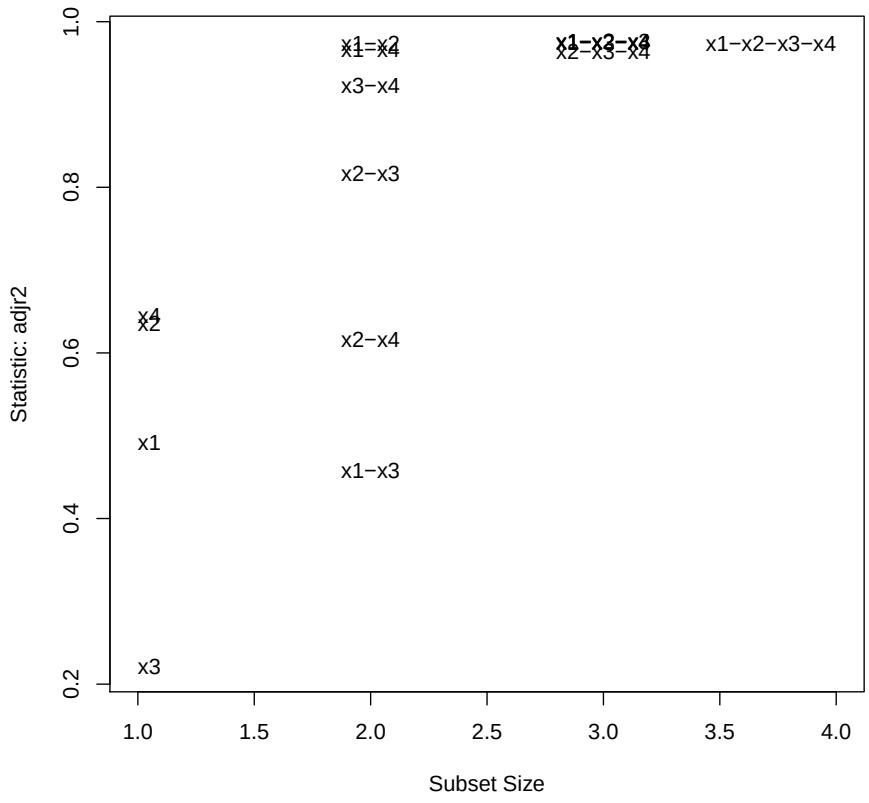



```
leaps3<-regsubsets(y~x1+x2+x3+x4,data=cement,nbest=16)
summary(leaps3)

Subset selection object
Call: regsubsets.formula(y ~ x1 + x2 + x3 + x4, data = cement, nbest = 16)
4 Variables  (and intercept)
   Forced in Forced out
x1     FALSE      FALSE
x2     FALSE      FALSE
x3     FALSE      FALSE
x4     FALSE      FALSE
16 subsets of each size up to 4
Selection Algorithm: exhaustive
         x1  x2  x3  x4
1  ( 1 ) " " " " " " "*"
1  ( 2 ) " " "*" " " " "
```

```
1  ( 3 ) "*" " " " " " " "
1  ( 4 ) " " " " " "*" " "
2  ( 1 ) "*" "*" " " " " "
2  ( 2 ) "*" " " " " " "*"
2  ( 3 ) " " " " " "*" "*"
2  ( 4 ) " " "*" "*" " " "
2  ( 5 ) " " "*" " " " "*"
2  ( 6 ) "*" " " "*" " " "
3  ( 1 ) "*" "*" " " " "*"
3  ( 2 ) "*" "*" "*" " " "
3  ( 3 ) "*" " " "*" "*" "
3  ( 4 ) " " "*" "*" "*" "
4  ( 1 ) "*" "*" "*" "*" "
```

```
library(car)
subsets(leaps3, statistic="adjr2")
```

```
Error:  invalid coordinate lengths
```



3 Forward selection

```
attach(cement)
```

```
The following objects are masked from cement (position 4):

    x1, x2, x3, x4, y
```

```
cor(cement)
```

```
         y      x1      x2       x3       x4
y   1.0000  0.7307  0.8163 -0.53467 -0.82131
x1  0.7307  1.0000  0.2286 -0.82413 -0.24545
x2  0.8163  0.2286  1.0000 -0.13924 -0.97295
x3 -0.5347 -0.8241 -0.1392  1.00000  0.02954
x4 -0.8213 -0.2454 -0.9730  0.02954  1.00000
```

```r
# step 1
# add x4
g <-lm(y ~ x4, data=cement)
summary(g)
```

```
Call:
lm(formula = y ~ x4, data = cement)

Residuals:
   Min     1Q Median     3Q    Max
-12.59  -8.23   1.50   4.73  17.52

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.568      5.262   22.34  1.6e-10 ***
x4            -0.738      0.155   -4.77  0.00058 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.96 on 11 degrees of freedom
Multiple R-squared:  0.675,Adjusted R-squared:  0.645
F-statistic: 22.8 on 1 and 11 DF,  p-value: 0.000576
```

```r
# step 2
# calculate the partial correlation
resy<-summary(lm(y ~ x4, data=cement))$res
resx1<-summary(lm(x1 ~ x4, data=cement))$res
resx2<-summary(lm(x2 ~ x4, data=cement))$res
resx3<-summary(lm(x3 ~ x4, data=cement))$res
cor(cbind(resy, resx1, resx2, resx3))
```

```
         resy    resx1    resx2   resx3
resy   1.0000  0.95677  0.13021 -0.8951
resx1  0.9568  1.00000 -0.04567 -0.8430
resx2  0.1302 -0.04567  1.00000 -0.4786
resx3 -0.8951 -0.84303 -0.47859  1.0000
```

```r
# add x1
g <-lm(y ~ x1 + x4, data=cement)
summary(g)
```

```
Call:
lm(formula = y ~ x1 + x4, data = cement)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-5.023 -1.474  0.137  1.731  3.770

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.0974     2.1240    48.5  3.3e-13 ***
x1            1.4400     0.1384    10.4  1.1e-06 ***
x4           -0.6140     0.0486   -12.6  1.8e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.73 on 10 degrees of freedom
Multiple R-squared:  0.972,Adjusted R-squared:  0.967
F-statistic:  177 on 2 and 10 DF,  p-value: 1.58e-08

# step 3
# calculate the partial correlation
resy<-summary(lm(y ~ x1 + x4, data=cement))$res
resx2<-summary(lm(x2 ~ x1 + x4, data=cement))$res
resx3<-summary(lm(x3 ~ x1 + x4, data=cement))$res
cor(cbind(resy, resx2, resx3))

         resy   resx2   resx3
resy   1.0000  0.5986 -0.5657
resx2  0.5986  1.0000 -0.9624
resx3 -0.5657 -0.9624  1.0000

# add x2
g <-lm(y ~ x1 + x2 + x4, data=cement)
summary(g)


Call:
lm(formula = y ~ x1 + x2 + x4, data = cement)

Residuals:
   Min     1Q Median     3Q    Max
-3.092 -1.802  0.256  1.282  3.898

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   71.648     14.142    5.07  0.00068 ***
x1             1.452      0.117   12.41  5.8e-07 ***
x2             0.416      0.186    2.24  0.05169 .
x4            -0.237      0.173   -1.37  0.20540
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.31 on 9 degrees of freedom
Multiple R-squared:  0.982,Adjusted R-squared:  0.976
F-statistic:  167 on 3 and 9 DF,  p-value: 3.32e-08

# final model is y ~ x1 + x2 + x4
```

4 Backward elimination

```
# step 1
g <- lm(y ~ ., data=cement)
summary(g)


Call:
lm(formula = y ~ ., data = cement)

Residuals:
   Min     1Q Median     3Q    Max
-3.175 -1.671  0.251  1.378  3.925

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.405     70.071    0.89    0.399
x1             1.551      0.745    2.08    0.071 .
x2             0.510      0.724    0.70    0.501
x3             0.102      0.755    0.14    0.896
x4            -0.144      0.709   -0.20    0.844
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.45 on 8 degrees of freedom
Multiple R-squared:  0.982,Adjusted R-squared:  0.974
F-statistic:  111 on 4 and 8 DF,  p-value: 4.76e-07

# step 2
# delete x3
g <- update(g, . ~ . - x3)
summary(g)


Call:
```

```
lm(formula = y ~ x1 + x2 + x4, data = cement)

Residuals:
   Min     1Q Median     3Q    Max
-3.092 -1.802  0.256  1.282  3.898

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   71.648     14.142    5.07  0.00068 ***
x1             1.452      0.117   12.41  5.8e-07 ***
x2             0.416      0.186    2.24  0.05169 .
x4            -0.237      0.173   -1.37  0.20540
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.31 on 9 degrees of freedom
Multiple R-squared:  0.982,Adjusted R-squared:  0.976
F-statistic:  167 on 3 and 9 DF,  p-value: 3.32e-08

# step 3
# delete x4
g <- update(g, . ~ . - x4)
summary(g)


Call:
lm(formula = y ~ x1 + x2, data = cement)

Residuals:
   Min     1Q Median     3Q    Max
 -2.89  -1.57  -1.30   1.36   4.05

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.5773     2.2862    23.0  5.5e-10 ***
x1            1.4683     0.1213    12.1  2.7e-07 ***
x2            0.6623     0.0459    14.4  5.0e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.41 on 10 degrees of freedom
Multiple R-squared:  0.979,Adjusted R-squared:  0.974
F-statistic:  230 on 2 and 10 DF,  p-value: 4.41e-09

# final model is y ~ x1 + x2
```

5 Stepwise regression

```
attach(cement)

The following objects are masked from cement (position 3):

    x1, x2, x3, x4, y
The following objects are masked from cement (position 5):

    x1, x2, x3, x4, y

cor(cement)

         y      x1      x2       x3       x4
y   1.0000  0.7307  0.8163 -0.53467 -0.82131
x1  0.7307  1.0000  0.2286 -0.82413 -0.24545
x2  0.8163  0.2286  1.0000 -0.13924 -0.97295
x3 -0.5347 -0.8241 -0.1392  1.00000  0.02954
x4 -0.8213 -0.2454 -0.9730  0.02954  1.00000

# step 1
# add x4
g <-lm(y ~ x4, data=cement)
summary(g)


Call:
lm(formula = y ~ x4, data = cement)

Residuals:
   Min     1Q Median     3Q    Max
-12.59  -8.23   1.50   4.73  17.52

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.568      5.262   22.34  1.6e-10 ***
x4            -0.738      0.155   -4.77  0.00058 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.96 on 11 degrees of freedom
Multiple R-squared:  0.675,Adjusted R-squared:  0.645
F-statistic: 22.8 on 1 and 11 DF,  p-value: 0.000576

# step 2
# calculate the partial correlation
```

```r
resy<-summary(lm(y ~ x4, data=cement))$res
resx1<-summary(lm(x1 ~ x4, data=cement))$res
resx2<-summary(lm(x2 ~ x4, data=cement))$res
resx3<-summary(lm(x3 ~ x4, data=cement))$res
cor(cbind(resy, resx1, resx2, resx3))
```

```
         resy    resx1    resx2   resx3
resy   1.0000  0.95677  0.13021 -0.8951
resx1  0.9568  1.00000 -0.04567 -0.8430
resx2  0.1302 -0.04567  1.00000 -0.4786
resx3 -0.8951 -0.84303 -0.47859  1.0000
```

```r
# add x1
g <-lm(y ~ x1 + x4, data=cement)
summary(g)
```

```

Call:
lm(formula = y ~ x1 + x4, data = cement)

Residuals:
   Min     1Q Median     3Q    Max
-5.023 -1.474  0.137  1.731  3.770

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.0974     2.1240    48.5  3.3e-13 ***
x1            1.4400     0.1384    10.4  1.1e-06 ***
x4           -0.6140     0.0486   -12.6  1.8e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.73 on 10 degrees of freedom
Multiple R-squared:  0.972,Adjusted R-squared:  0.967
F-statistic:  177 on 2 and 10 DF,  p-value: 1.58e-08
```

```r
# step 3
# calculate the partial correlation
resy<-summary(lm(y ~ x1 + x4, data=cement))$res
resx2<-summary(lm(x2 ~ x1 + x4, data=cement))$res
resx3<-summary(lm(x3 ~ x1 + x4, data=cement))$res
cor(cbind(resy, resx2, resx3))
```

```
         resy   resx2   resx3
resy   1.0000  0.5986 -0.5657
resx2  0.5986  1.0000 -0.9624
resx3 -0.5657 -0.9624  1.0000
```

```r
# add x2
g <-lm(y ~ x1 + x2 + x4, data=cement)
summary(g)
```

```
Call:
lm(formula = y ~ x1 + x2 + x4, data = cement)

Residuals:
   Min     1Q Median     3Q    Max
-3.092 -1.802  0.256  1.282  3.898

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   71.648     14.142    5.07  0.00068 ***
x1             1.452      0.117   12.41  5.8e-07 ***
x2             0.416      0.186    2.24  0.05169 .
x4            -0.237      0.173   -1.37  0.20540
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.31 on 9 degrees of freedom
Multiple R-squared:  0.982,Adjusted R-squared:  0.976
F-statistic:  167 on 3 and 9 DF,  p-value: 3.32e-08
```

```r
# step 4
# delete x4
g <-lm(y ~ x1 + x2, data=cement)
summary(g)
```

```
Call:
lm(formula = y ~ x1 + x2, data = cement)

Residuals:
   Min     1Q Median     3Q    Max
 -2.89  -1.57  -1.30   1.36   4.05

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.5773     2.2862    23.0  5.5e-10 ***
x1            1.4683     0.1213    12.1  2.7e-07 ***
x2            0.6623     0.0459    14.4  5.0e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.41 on 10 degrees of freedom
Multiple R-squared:  0.979,Adjusted R-squared:  0.974
F-statistic:  230 on 2 and 10 DF,  p-value: 4.41e-09

# final model is y ~ x1 + x2
```