1 How to choose a model? Chapter 10.

- Fit all possible regressions, and choose a model which optimizes some criterion.
  - Pick a model with maximizes adjusted R^2 among the models under consideration. For a model with p parameters,

    $$R^2_{Adj,p} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2_p)$$

    where R^2_p is the usual R^2 for that model.
  - R^2 is not a good criterion to use, as it will always be maximized for the largest model considered.
  - Where $MSE(p)$ is the error mean square for the model with p regression parameters, choose the model for which $MSE(p)$ is minimized. On page 334 of the book, it is shown that minimizing $MSE(p)$ is equivalent to maximizing $R^2_{Adj,p}$.
  - For a particular model with p parameters, for which the likelihood is L_p, the Akaike Information Criterion is $AIC_p = -2log(L_p) + 2p$. Pick the model which minimizes AIC over the collection of models considered.

    It is shown on page 336 that

    $$AIC = nlog\left(\frac{SSE_p}{n}\right) + plog(n)$$

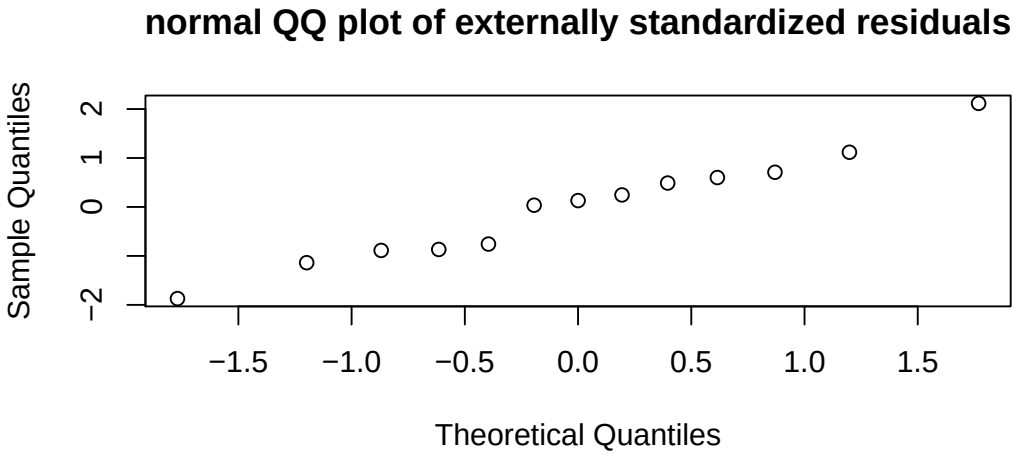
Using library leaps to perform all subsets regression

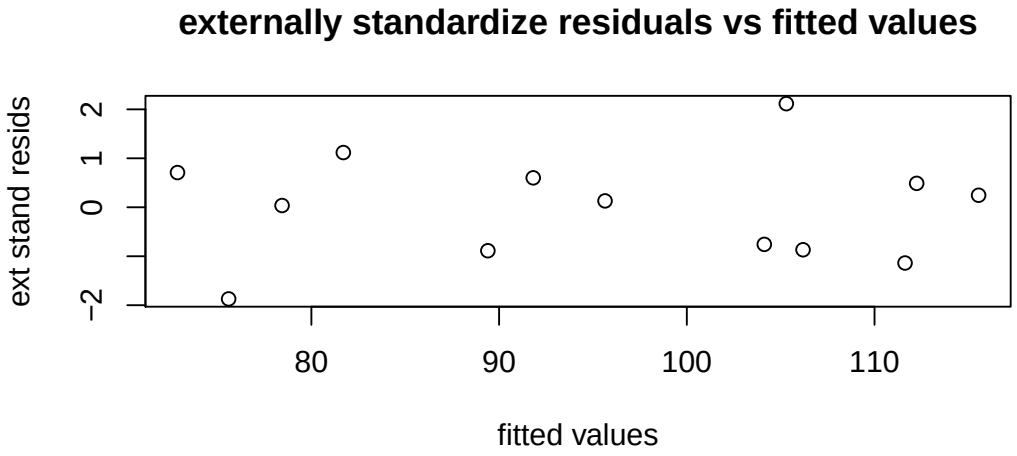
```
> library(leaps)
> data=read.csv("http://bsmith.mathstat.dal.ca/stat3340/Data/cement.csv",header=T)
> data=as.matrix(data[,-1])
> leaps.out=leaps(x=data[,-1],y=data[,1],method="adjr2")
> leaps.out

$which
      1     2     3     4
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)" "1"           "2"           "3"           "4"

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

$adjr2
 [1] 0.6449549 0.6359290 0.4915797 0.2209521 0.9744140 0.9669653 0.9223476
 [8] 0.8164305 0.6160725 0.4578001 0.9764473 0.9763796 0.9750415 0.9637599
[15] 0.9735634
```

According to the adjusted R^2 criterion, the model chosen is the one with the largest adjusted R^2, in this case, in this case the model using X_1, X_2 and X_4.

Now check to see if the model is satisfactory.

```
> y=data[,1]; x1=data[,2];x2=data[,3];x3=data[,4];x4=data[,5]
> lmbest=lm(y~x1+x2+x4)
> esresids=rstudent(lmbest) #externally standardized residals
> par(mfrow=c(2,1))
> qqnorm(esresids,main="normal QQ plot of externally standardized residuals")
> plot(fitted(lmbest),esresids,xlab="fitted values",ylab="ext stand resids",
+        main="externally standardize residuals vs fitted values")
```





- If not possible to fit all possible models, use a stepwise procedure to select a model.
    1. Forward selection. Add variables sequentially, using F tests to decide on whether or not to enter a variable.
    2. Backwards elimination. Fit the largest model under consideration, and then use F tests to remove terms from the model.
    3. Full stepwise. Alternate between adding and removing terms, using F tests at each step.

```
> nullmodel=lm(y~1) #include intercept only
> fullmodel=lm(y~x1+x2+x3+x4) #full model has all main effects, no interactions
> step1=step(nullmodel,scope=list(lower=nullmodel, upper=fullmodel), direction="forward")

Start:  AIC=71.44
y ~ 1

       Df Sum of Sq     RSS    AIC
+ x4    1   1831.90  883.87 58.852
+ x2    1   1809.43  906.34 59.178
+ x1    1   1450.08 1265.69 63.519
+ x3    1    776.36 1939.40 69.067
<none>              2715.76 71.444

Step:  AIC=58.85
y ~ x4

       Df Sum of Sq    RSS    AIC
+ x1    1    809.10  74.76 28.742
+ x3    1    708.13 175.74 39.853
<none>              883.87 58.852
+ x2    1     14.99 868.88 60.629

Step:  AIC=28.74
y ~ x4 + x1

       Df Sum of Sq    RSS    AIC
+ x2    1    26.789 47.973 24.974
+ x3    1    23.926 50.836 25.728
<none>              74.762 28.742

Step:  AIC=24.97
y ~ x4 + x1 + x2

       Df Sum of Sq    RSS    AIC
<none>              47.973 24.974
+ x3    1   0.10909 47.864 26.944

> summary(step1)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-3.0919 -1.8016  0.2562  1.2818  3.8982

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.6483    14.1424   5.066 0.000675 ***
x4           -0.2365     0.1733  -1.365 0.205395
x1            1.4519     0.1170  12.410 5.78e-07 ***
x2            0.4161     0.1856   2.242 0.051687 .
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 2.309 on 9 degrees of freedom
Multiple R-squared:  0.9823,        Adjusted R-squared:  0.9764
F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08
```

```
> fullmodel2=lm(y~x1*x2+x1*x3+x1*x4+x2*x3+x2*x4+x3*x4) #full model has all 2 way interacti
> step2=step(fullmodel2,direction="backward")

Start:  AIC=-4.21
y ~ x1 * x2 + x1 * x3 + x1 * x4 + x2 * x3 + x2 * x4 + x3 * x4

        Df Sum of Sq     RSS     AIC
- x1:x2  1    0.0000  1.7310 -6.2115
- x3:x4  1    0.2366  1.9676 -4.5459
- x1:x4  1    0.2615  1.9924 -4.3830
<none>                1.7309 -4.2117
- x2:x3  1    1.2570  2.9879  0.8853
- x1:x3  1    2.4095  4.1404  5.1260
- x2:x4  1   27.7898 29.5207 30.6619

Step:  AIC=-6.21
y ~ x1 + x2 + x3 + x4 + x1:x3 + x1:x4 + x2:x3 + x2:x4 + x3:x4

        Df Sum of Sq    RSS    AIC
<none>                1.731 -6.211
- x3:x4  1     1.337  3.068 -0.769
- x2:x3  1     9.676 11.407 16.300
- x1:x3  1    16.619 18.350 22.481
- x1:x4  1    20.110 21.841 24.745
- x2:x4  1    37.583 39.314 32.386

> summary(step2)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x1:x3 + x1:x4 + x2:x3 +
    x2:x4 + x3:x4)

Residuals:
       1        2        3        4        5        6        7        8
-0.06149  0.23625  0.09178  0.01904 -0.10247  0.36358  0.15830 -0.48573
       9       10       11       12       13
-0.55701 -0.04384  0.60426 -0.63898  0.41630

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -45.954156  26.642567  -1.725  0.18302
x1            4.161641   0.383503  10.852  0.00167 **
x2            0.909468   0.247971   3.668  0.03506 *
x3           -3.038708   1.375568  -2.209  0.11420
x4            0.976112   0.265862   3.672  0.03496 *
x1:x3         0.103762   0.019334   5.367  0.01266 *
x1:x4        -0.051546   0.008731  -5.904  0.00970 **
```

```
x2:x3         0.085416   0.020858   4.095  0.02633 *
x2:x4         0.015151   0.001877   8.071  0.00397 **
x3:x4         0.032597   0.021410   1.522  0.22525
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.7596 on 3 degrees of freedom
Multiple R-squared:  0.9994,        Adjusted R-squared:  0.9975
F-statistic: 522.6 on 9 and 3 DF,  p-value: 0.0001247
```

> anova(step1,step2)

```
Analysis of Variance Table

Model 1: y ~ x4 + x1 + x2
Model 2: y ~ x1 + x2 + x3 + x4 + x1:x3 + x1:x4 + x2:x3 + x2:x4 + x3:x4
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      9 47.973
2      3  1.731  6    46.242 13.357 0.0287 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

- There can be inconsistencies when using stepwise procedures. Compare model from full stepwise procedure below to backwards elimination model.
- Resulting model depends on input parameters, which govern F tests for adding or removing terms.
- There are multiple testing issues. When looking at the "final" model, tests will not have the right level, and CI's won't have the correct coverage.
- The "lasso" procedure attemps to provide a model selection procedure where the CI's and tests for the selected model are valid.

```
> step(nullmodel,scope=list(lower=nullmodel,upper=fullmodel2),direction="both")

Start:  AIC=71.44
y ~ 1

       Df Sum of Sq     RSS    AIC
+ x4    1   1831.90  883.87 58.852
+ x2    1   1809.43  906.34 59.178
+ x1    1   1450.08 1265.69 63.519
+ x3    1    776.36 1939.40 69.067
<none>              2715.76 71.444

Step:  AIC=58.85
y ~ x4

       Df Sum of Sq     RSS    AIC
+ x1    1    809.10   74.76 28.742
+ x3    1    708.13  175.74 39.853
<none>               883.87 58.852
+ x2    1     14.99  868.88 60.629
- x4    1   1831.90 2715.76 71.444

Step:  AIC=28.74
y ~ x4 + x1

        Df Sum of Sq     RSS    AIC
+ x2     1     26.79   47.97 24.974
+ x3     1     23.93   50.84 25.728
<none>               74.76 28.742
+ x1:x4  1      0.13   74.64 30.720
- x1     1    809.10  883.87 58.852
- x4     1   1190.92 1265.69 63.519

Step:  AIC=24.97
y ~ x4 + x1 + x2
```

```
        Df Sum of Sq    RSS    AIC
<none>                47.97 24.974
+ x2:x4  1      6.43  41.54 25.102
- x4     1      9.93  57.90 25.420
+ x1:x4  1      1.01  46.96 26.696
+ x1:x2  1      0.54  47.43 26.827
+ x3     1      0.11  47.86 26.944
- x2     1     26.79  74.76 28.742
- x1     1    820.91 868.88 60.629

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

Coefficients:
(Intercept)           x4           x1           x2
    71.6483      -0.2365       1.4519       0.4161
```