ORIGIN ≡ 0

Building Regression Models using Automated Serial Procedures

W. Stein

In exploring fit of a linear model relating a dependent variable Y with a somehow determined "maximally useful" design matrix X (comprised of a column of 1's plus some set of independent variables X_i), computer automated procedures are often used to identify interesting subsets of X_i's (i.e., usually less than a full design with p columns). To do this, the, computer algorithims either consider *all possible subsets* (if the set of all X_i's are sufficiently small to allow computation), or *serially built or stepwise subsets* (with an individual X_i added or subtracted during each step of an automated process). In either case, a criterion for admission or exclusion from a preferred subset needs to be calculated for each model fit. Examples below comes from Chapter 9 of Kuter et al. (KNNL) A*pplied Linear Statistical Models* 5th Edition.

Example:

Surgical Unit Example KNNL Table 9.1

K := READPRN("c:/2008LinearModelsData/SurgicalUnit.txt")

$Y := \ln\left(K^{\langle 8\rangle}\right)$

$X1 := K^{\langle 0\rangle}$

$X2 := K^{\langle 1\rangle}$

$X3 := K^{\langle 2\rangle}$

$X4 := K^{\langle 3\rangle}$

$n := \text{length}(Y) \qquad n = 54$

$i := 0 .. n-1$

$ii := 0 .. n-1$

$OV_i := 1$

$I := \text{identity}(n)$

$J_{i,ii} := 1$

K =

	0	1	2	3	4	5	6	7	8	9
0	6.7	62	81	2.59	50	0	1	0	695	6.544
1	5.1	59	66	1.7	39	0	0	0	403	5.999
2	7.4	57	83	2.16	55	0	0	0	710	6.565
3	6.5	73	41	2.01	48	0	0	0	349	5.854
4	7.8	65	115	4.3	45	0	0	1	2343	7.759
5	5.8	38	72	1.42	65	1	1	0	348	5.852
6	5.7	46	63	1.91	49	1	0	1	518	6.25
7	3.7	68	81	2.57	69	1	1	0	749	6.619
8	6	67	93	2.5	58	0	1	0	1056	6.962
9	3.7	76	94	2.4	48	0	1	0	968	6.875
10	6.3	84	83	4.13	37	0	1	0	745	6.613
11	6.7	51	43	1.86	57	0	1	0	257	5.549
12	5.8	96	114	3.95	63	1	0	0	1573	7.361
13	5.8	83	88	3.95	52	1	0	0	858	6.754
14	7.7	62	67	3.4	58	0	0	1	702	6.554
15	7.4	74	68	2.4	64	1	1	0	809	6.695

$X := \text{augment}(OV, X1, X2, X3, X4)$ **< design matrix**

$p := \text{cols}(X) \qquad p = 5$

Least Squares Estimation of the Regression Parameters:

$$b := \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot Y$$

$$b = \begin{pmatrix} 3.8519333 \\ 0.0837389 \\ 0.012671 \\ 0.0156272 \\ 0.0320559 \end{pmatrix}$$

< vector of regresion coefficients

Fitted Values & Hat Matrix H:

$Y_h := X \cdot b$

$H := X \cdot \left(X^T \cdot X\right)^{-1} \cdot X^T$ **< nXn hat matrix**

Residuals:

$e := Y - Y_h$ **< residuals**

ANOVA Table:

Sum of Squares: **Degrees of Freedom:** **Mean Squares:**

$SSR := Y^T \cdot \left[H - \left(\frac{1}{n}\right) \cdot J\right] \cdot Y$ $SSR = (9.7204)$ $df_R := p - 1$ $df_R = 4$ $MSR := \frac{SSR}{df_R}$ $MSR = (2.4301)$

$SSE := Y^T \cdot (I - H) \cdot Y$ $SSE = (3.0841)$ $df_E := n - p$ $df_E = 49$ $MSE := \frac{SSE}{df_E}$ $MSE = (0.0629)$

$SSTO := Y^T \cdot \left[I - \left(\frac{1}{n}\right) \cdot J\right] \cdot Y$ $SSTO = (12.8045)$ $df_T := n - 1$ $df_T = 53$ $MSTO := \frac{SSTO}{df_T}$ $MSTO = (0.2416)$

Criteria for Model Selection:

Sum of Squares Error:

$SSE_0 = 3.0841$

< one looks for little further drop in SSE for a model with some subset of X_i's

Coefficient of Multiple Determination (R^2):

$R_{sq} := 1 - \frac{SSE_0}{SSTO_0}$ $R_{sq} = 0.75914$

< one looks for little further rise in R^2 for a model with some subset of X_i's

Adjusted Coefficient of Multiple Determination:

$R_{sqa} := 1 - \frac{MSE_0}{MSTO_0}$ $R_{sqa} = 0.7395$

< one looks for little further rise in R_{adj}^2 for a model with some subset of X_i's

Mallow's C_p Criterion:

$Cp := \frac{SSE_0}{MSE_0} - (n - 2 \cdot p)$ $Cp = 5$

< one seeks subsets of X_i's with small Cp value and Cp near p

Akaike's Information Criterion (AIC):

$AIC := n \cdot \ln(SSE_0) - n \cdot \ln(n) + 2 \cdot p$ $AIC = -144.5872$

< one seeks minimum AIC values

Schwartz's Bayesian Criterion

$SBC := n \cdot \ln(SSE_0) - n \cdot \ln(n) + \ln(n) \cdot p$ $SBC = -134.6423$

< one seeks minimum SBC values

PRESS Criterion:

$d_i := \frac{e_i}{1 - H_{i,i}}$ **< KNNL Eq.10.21a**

$PRESS := \sum d^2$ $PRESS = 4.0687$

< one seeks minimum PRESS values

d =

	0
0	-0.0036
1	-0.1179
2	0.0057
3	-0.1866
4	0.5662
5	-0.1497
6	0.3067
7	0.2629
8	0.2395
9	0.2212
10	-0.2782
11	-0.2632
12	-0.1203
13	-0.1453
14	0.1217

Prototype in R:

Calculating the Criteria:

```
#SELECTION CRITERIA FOR SERIAL MODELS
#READ STRUCTURED DATA TABLE WITH NUMERIC CODED FACTOR
K=read.table("c:/2008LinearModelsData/SurgicalUnitR.txt")
K
attach(K)
options(digits=6)
Y=log(Y)
#VIEWING DATA
DATA=cbind(lnY,X1,X2,X3,X4)
DATA
#VIEWING CORRELATION MATRIX
cor(DATA)
```

```
> #VIEWING CORRELATION MATRIX
> cor(DATA)
         lnY         X1         X2         X3       X4
lnY 1.000000  0.2461879  0.4699432  0.6538855 0.649263
X1  0.246188  1.0000000  0.0901197 -0.1496341 0.502416
X2  0.469943  0.0901197  1.0000000 -0.0236054 0.369026
X3  0.653885 -0.1496341 -0.0236054  1.0000000 0.416425
X4  0.649263  0.5024157  0.3690256  0.4164245 1.000000
```

```
#FITTING LINEAR MODEL
FM=lm(Y~X1+X2+X3+X4)
#SETTING UP MATRIX ALGEBRA OBJECTS AND HAT MATRIX
n=length(Y)
I=diag(n)
J=matrix(nrow=n,ncol=n,1)
X=model.matrix(FM)
p=ncol(X)
H=X%*%solve(t(X)%*%X)%*%t(X) #HAT MATRIX
#CALCULATING SUMS OF SQUARES
SSR=t(Y)%*%(H-((1/n)*J))%*%Y
SSR
SSE=t(Y)%*%(I-H)%*%Y
SSE
SSTO=t(Y)%*%(I-(1/n)*J)%*%Y
SSTO
#COMPARING WITH anova() OUTPUT
anova(FM)
```

```
> SSR
        [,1]
[1,] 9.72042
> SSE
        [,1]
[1,] 3.08409
> SSTO
        [,1]
[1,] 12.8045
```

```
> anova(FM)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
X1         1  0.777   0.777  12.344 0.000962 ***
X2         1  2.590   2.590  41.156 5.34e-08 ***
X3         1  6.329   6.329 100.549 1.84e-13 ***
X4         1  0.024   0.024   0.388 0.536270
Residuals 49  3.084   0.063
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#CALCULATING CRITERIA FOR MODEL SELECTION
#SUM OF SQUARES ERROR
SSE[1]
#COEFFICIENT OF MULTIPLE DETERMINATION
Rsq=1-(SSE[1]/SSTO[1])
Rsq
summary(FM)$r.squared #ALTERNATE CALCULATION
#ADJUSTED COEFFICIEINT OF MULTIPLE DETERMINATION
MSE=SSE[1]/(n-p)
MSTO=SSTO[1]/(n-1)
Rsqa=1-(MSE/MSTO)
Rsqa
summary(FM)$adj.r.squared #ALTERNATE CALCULATION
#MALLOW'S Cp
Cp=SSE[1]/MSE -(n-2*p)
Cp
require(wle) #DOWNLOAD {wle} PACKAGE FROM CRAN WEBSITE
mle.cp(FM)  #DON'T KNOW MUCH ABOUT THIS ONE, BUT INTERESTING!
#AKAIKE'S INFORMATION CRITERION AIC
AIC=n*log(SSE[1])-n*log(n)+2*p
AIC
extractAIC(FM)  #REPORTS: (equivalent df, AIC)  see ?extractAIC
#SCHWARTZ'S BAYESIAN CRITERION SBC
SBC=n*log(SSE[1])-n*log(n)+log(n)*p
SBC
extractAIC(FM,k=log(n))
#PRESS CRITERION
e=residuals(FM)
d=e/(1-diag(H))  #KNNL Eq. 10.21a
diag(H)        #MAIN DIAGONAL OF H AS A VECTOR
hatvalues(FM)    #ALTERNATE CALCULATION
USING BUILT-IN FUNCTION & FM
PRESS=sum(d^2)
PRESS
```

```
> SSE[1]
[1] 3.08409
> Rsq
[1] 0.75914
> summary(FM)$r.squared #ALTERNATE
CALCULATION
[1] 0.75914
> Rsqa
[1] 0.739478
> summary(FM)$adj.r.squared
#ALTERNATE CALCULATION
[1] 0.739478
```

```
> Cp
[1] 5
> mle.cp(FM)  #DON'T KNOW MUCH ABOUT THIS ONE, BUT
INTERESTING!
Call:
mle.cp(formula = FM)
Mallows Cp:
     (Intercept) X1 X2 X3 X4   cp
[1,]           1  1  1  1  0 3.39
[2,]           1  1  1  1  1 5.00
Printed the first  2  best models
```

```
> AIC
[1] -144.587
> extractAIC(FM)
[1]    5.000 -144.587
> SBC
[1] -134.642
> extractAIC(FM,k=log(n))
[1]    5.000 -134.642
> PRESS
[1] 4.06875
```

Automated Stepwise Regression in R:

```
#AUTOMATED STEPWISE REGRESSION
FM=lm(Y~X1+X2+X3+X4)
RM=lm(Y~1)

#STEPWISE REGRESSION USING AIC
step(FM,direction="backward")
step(RM,~X1+X2+X3+X4,direction="forward")
step(RM,~X1+X2+X3+X4,direction="both")
```

```
> step(RM,~X1+X2+X3+X4,direction="both")
Start:  AIC=-75.72
Y ~ 1
       Df Sum of Sq     RSS     AIC
+ X3    1      5.47    7.33 -103.81
+ X4    1      5.40    7.41 -103.27
+ X2    1      2.83    9.97  -87.21
+ X1    1      0.78   12.03  -77.10
<none>                12.80  -75.72
Step:  AIC=-103.81
Y ~ X3
       Df Sum of Sq     RSS     AIC
+ X2    1      3.02    4.31 -130.48
+ X4    1      2.20    5.13 -121.09
+ X1    1      1.55    5.78 -114.64
<none>                 7.33 -103.81
- X3    1      5.47   12.80  -75.72
Step:  AIC=-130.48
Y ~ X3 + X2
       Df Sum of Sq     RSS     AIC
+ X1    1      1.20    3.11 -146.16
+ X4    1      0.70    3.61 -138.01
<none>                 4.31 -130.48
- X2    1      3.02    7.33 -103.81
- X3    1      5.66    9.97  -87.21
Step:  AIC=-146.16
Y ~ X3 + X2 + X1
       Df Sum of Sq     RSS     AIC
<none>                 3.11 -146.16
+ X4    1      0.02    3.08 -144.59
- X1    1      1.20    4.31 -130.48
- X2    1      2.67    5.78 -114.64
- X3    1      6.33    9.44  -88.19
Call:
lm(formula = Y ~ X3 + X2 + X1)
Coefficients:
(Intercept)           X3           X2
  X1
     3.7664       0.0164       0.0133
0.0955
```

```
> step(FM,direction="backward")
Start:  AIC=-144.59
Y ~ X1 + X2 + X3 + X4
       Df Sum of Sq     RSS     AIC
- X4    1      0.02    3.11 -146.16
<none>                 3.08 -144.59
- X1    1      0.53    3.61 -138.01
- X2    1      1.89    4.97 -120.82
- X3    1      3.48    6.57 -105.76
Step:  AIC=-146.16
Y ~ X1 + X2 + X3
       Df Sum of Sq     RSS     AIC
<none>                 3.11 -146.16
- X1    1      1.20    4.31 -130.48
- X2    1      2.67    5.78 -114.64
- X3    1      6.33    9.44  -88.19
Call:
lm(formula = Y ~ X1 + X2 + X3)
Coefficients:
(Intercept)           X1           X2           X3
     3.7664       0.0955       0.0133       0.0164
```

```
> step(RM,~X1+X2+X3+X4,direction="forward")
Start:  AIC=-75.72
Y ~ 1
       Df Sum of Sq     RSS     AIC
+ X3    1      5.47    7.33 -103.81
+ X4    1      5.40    7.41 -103.27
+ X2    1      2.83    9.97  -87.21
+ X1    1      0.78   12.03  -77.10
<none>                12.80  -75.72
Step:  AIC=-103.81
Y ~ X3
       Df Sum of Sq     RSS     AIC
+ X2    1      3.02    4.31 -130.48
+ X4    1      2.20    5.13 -121.09
+ X1    1      1.55    5.78 -114.64
<none>                 7.33 -103.81
Step:  AIC=-130.48
Y ~ X3 + X2
       Df Sum of Sq     RSS     AIC
+ X1    1      1.20    3.11 -146.16
+ X4    1      0.70    3.61 -138.01
<none>                 4.31 -130.48
Step:  AIC=-146.16
Y ~ X3 + X2 + X1
       Df Sum of Sq     RSS     AIC
<none>                 3.11 -146.16
+ X4    1      0.02    3.08 -144.59
Call:
lm(formula = Y ~ X3 + X2 + X1)
Coefficients:
(Intercept)           X3           X2           X1
     3.7664       0.0164       0.0133       0.0955
```

#STEPWISE REGRESSION USING SBC
step(FM,direction="backward",k=log(n))
step(RM,~X1+X2+X3+X4,direction="forward",k=log(n))
step(RM,~X1+X2+X3+X4,direction="both",k=log(n))

> step(RM,~X1+X2+X3+X4,direction="both",k=log(n))

```
Start:  AIC=-73.73
Y ~ 1
       Df Sum of Sq    RSS    AIC
+ X3    1      5.47   7.33 -99.83
+ X4    1      5.40   7.41 -99.29
+ X2    1      2.83   9.97 -83.23
<none>               12.80 -73.73
+ X1    1      0.78  12.03 -73.12
Step:  AIC=-99.83
Y ~ X3
       Df Sum of Sq     RSS     AIC
+ X2    1      3.02    4.31 -124.51
+ X4    1      2.20    5.13 -115.12
+ X1    1      1.55    5.78 -108.68
<none>                 7.33  -99.83
- X3    1      5.47   12.80  -73.73
Step:  AIC=-124.51
Y ~ X3 + X2
       Df Sum of Sq     RSS     AIC
+ X1    1      1.20    3.11 -138.21
+ X4    1      0.70    3.61 -130.05
<none>                 4.31 -124.51
- X2    1      3.02    7.33  -99.83
- X3    1      5.66    9.97  -83.23
Step:  AIC=-138.21
Y ~ X3 + X2 + X1
       Df Sum of Sq     RSS     AIC
<none>                 3.11 -138.21
+ X4    1      0.02    3.08 -134.64
- X1    1      1.20    4.31 -124.51
- X2    1      2.67    5.78 -108.68
- X3    1      6.33    9.44  -82.23
Call:
lm(formula = Y ~ X3 + X2 + X1)
Coefficients:
(Intercept)           X3           X2           X1
     3.7664       0.0164       0.0133       0.0955
```

> step(FM,direction="backward",k=log(n))

```
Start:  AIC=-134.64
Y ~ X1 + X2 + X3 + X4
       Df Sum of Sq     RSS     AIC
- X4    1      0.02    3.11 -138.21
<none>                 3.08 -134.64
- X1    1      0.53    3.61 -130.05
- X2    1      1.89    4.97 -112.87
- X3    1      3.48    6.57  -97.81
Step:  AIC=-138.21
Y ~ X1 + X2 + X3
       Df Sum of Sq     RSS     AIC
<none>                 3.11 -138.21
- X1    1      1.20    4.31 -124.51
- X2    1      2.67    5.78 -108.68
- X3    1      6.33    9.44  -82.23
Call:
lm(formula = Y ~ X1 + X2 + X3)
Coefficients:
(Intercept)           X1           X2
X3
     3.7664       0.0955       0.0133
0.0164
```

> step(RM,~X1+X2+X3+X4,direction="forward",k=log(n))

```
Start:  AIC=-73.73
Y ~ 1
       Df Sum of Sq    RSS    AIC
+ X3    1      5.47   7.33 -99.83
+ X4    1      5.40   7.41 -99.29
+ X2    1      2.83   9.97 -83.23
<none>               12.80 -73.73
+ X1    1      0.78  12.03 -73.12
Step:  AIC=-99.83
Y ~ X3
       Df Sum of Sq     RSS     AIC
+ X2    1      3.02    4.31 -124.51
+ X4    1      2.20    5.13 -115.12
+ X1    1      1.55    5.78 -108.68
<none>                 7.33  -99.83
Step:  AIC=-124.51
Y ~ X3 + X2
       Df Sum of Sq     RSS     AIC
+ X1    1      1.20    3.11 -138.21
+ X4    1      0.70    3.61 -130.05
<none>                 4.31 -124.51
Step:  AIC=-138.21
Y ~ X3 + X2 + X1
       Df Sum of Sq     RSS     AIC
<none>                 3.11 -138.21
+ X4    1      0.02    3.08 -134.64
Call:
lm(formula = Y ~ X3 + X2 + X1)
Coefficients:
(Intercept)           X3           X2           X1

     3.7664       0.0164       0.0133       0.0955
```