Stepwise Variable Selection*

```
> # Illustrate stepwise variable selection.
> # Some of the code is taken from Weibul regression part two.
> # There, I settled on a model with treatment, age and employment status.
> # That was with Weibull regression.
>
> rm(list=ls());  options(scipen=999)
> # install.packages("survival",dependencies=TRUE) # Only need to do this once
> library(survival) # Do this every time
> # install.packages("asaur",dependencies=TRUE) # Only need to do this once
> library(asaur)
> # help(pharmacoSmoking)
> # head(pharmacoSmoking)
> summary(pharmacoSmoking)

       id              ttr            relapse              grp
 Min.   :  1.00   Min.   :  0.00   Min.   :0.000   combination:61
 1st Qu.: 33.00   1st Qu.:  8.00   1st Qu.:0.000   patchOnly  :64
 Median : 67.00   Median : 49.00   Median :1.000
 Mean   : 66.15   Mean   : 77.44   Mean   :0.712
 3rd Qu.: 99.00   3rd Qu.:182.00   3rd Qu.:1.000
 Max.   :130.00   Max.   :182.00   Max.   :1.000
      age           gender          race     employment  yearsSmoking
 Min.   :22.00   Female:81   black   :38   ft   :72   Min.   : 9.00
 1st Qu.:41.00   Male  :44   hispanic: 8   other:39   1st Qu.:22.00
 Median :49.00               other   : 2   pt   :14   Median :30.00
 Mean   :48.84               white   :77              Mean   :30.88
 3rd Qu.:56.00                                       3rd Qu.:39.00
 Max.   :86.00                                       Max.   :56.00
 levelSmoking ageGroup2  ageGroup4  priorAttempts     longestNoSmoke
 heavy:89     21-49:66   21-34:16   Min.   :   0.00   Min.   :   0.0
 light:36     50+  :59   35-49:50   1st Qu.:   1.00   1st Qu.:   7.0
                         50-64:48   Median :   2.00   Median :  90.0
                         65+  :11   Mean   :  12.68   Mean   : 539.7
                                    3rd Qu.:   5.00   3rd Qu.: 365.0
                                    Max.   :1000.00   Max.   :6205.0
> # Make fixed-up data frame called quit
> quit = within(pharmacoSmoking,{
+ DayOfRelapse = Surv(ttr+1,relapse)
+ contrasts(grp) = contr.treatment(2,base=2) # Patch only is reference category
+ colnames(contrasts(grp)) = c('Combo') # Names of dummy vars -- just one
+ # Collapse race categories
+ Race = as.character(race) # Small r race is a factor. This is easier to modify.
+ Race[Race!='white'] = 'blackOther'; Race=factor(Race)
+ }) # Finished making data frame quit
```

```
> everything = coxph(DayOfRelapse ~ grp + age + ageGroup2 + ageGroup4 +
+                                   gender + Race + employment +
+                                   yearsSmoking + levelSmoking + priorAttempts,
+                                   data=quit)
> summary(everything)

Call:
coxph(formula = DayOfRelapse ~ grp + age + ageGroup2 + ageGroup4 +
    gender + Race + employment + yearsSmoking + levelSmoking +
    priorAttempts, data = quit)

  n= 125, number of events= 89

                        coef  exp(coef)   se(coef)      z Pr(>|z|)
grpCombo          -0.6209847  0.5374150  0.2223377 -2.793  0.00522 **
age               -0.0436857  0.9572547  0.0309009 -1.414  0.15744
ageGroup250+       0.6435315  1.9031901  1.1618974  0.554  0.57967
ageGroup435-49     0.2835267  1.3278043  0.4555838  0.622  0.53372
ageGroup450-64    -0.8444256  0.4298041  0.5817233 -1.452  0.14661
ageGroup465+              NA         NA  0.0000000     NA       NA
genderMale         0.0131214  1.0132078  0.2503229  0.052  0.95820
Racewhite         -0.1967183  0.8214220  0.2335774 -0.842  0.39968
employmentother    0.7034399  2.0206918  0.2813184  2.501  0.01240 *
employmentpt       0.6289111  1.8755672  0.3458876  1.818  0.06903 .
yearsSmoking       0.0092857  1.0093290  0.0184825  0.502  0.61538
levelSmokinglight -0.0180956  0.9820671  0.2553175 -0.071  0.94350
priorAttempts      0.0003749  1.0003750  0.0011305  0.332  0.74018
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
grpCombo             0.5374     1.8608    0.3476    0.8309
age                  0.9573     1.0447    0.9010    1.0170
ageGroup250+         1.9032     0.5254    0.1952   18.5563
ageGroup435-49       1.3278     0.7531    0.5437    3.2429
ageGroup450-64       0.4298     2.3266    0.1374    1.3441
ageGroup465+             NA         NA        NA        NA
genderMale           1.0132     0.9870    0.6203    1.6549
Racewhite            0.8214     1.2174    0.5197    1.2983
employmentother      2.0207     0.4949    1.1642    3.5072
employmentpt         1.8756     0.5332    0.9522    3.6945
yearsSmoking         1.0093     0.9908    0.9734    1.0466
levelSmokinglight    0.9821     1.0183    0.5954    1.6198
priorAttempts        1.0004     0.9996    0.9982    1.0026

Concordance= 0.658  (se = 0.032 )
Likelihood ratio test= 28.7  on 12 df,   p=0.004
Wald test            = 27.05  on 12 df,   p=0.008
Score (logrank) test = 28.02  on 12 df,   p=0.005>
> # Fit the restricted model: Restricted by H0
> rest1 =  survreg(DayOfRelapse ~ grp + age + gender + Race + employment ,
+                  dist='weibull', data=quit)
```

Automatic variable selection in R is based on the Akaike information criterion (AIC). The AIC is a measure of how “bad” a model is, based on information theory. Higher minus log likelihood is bad, and lots of predictor variables is bad.

$$\mathrm{AIC} = 2k - 2\log L(\widehat{\theta})$$

```
backwards = step(everything) # Backwards elimination is the default

Start:  AIC=767.61
DayOfRelapse ~ grp + age + ageGroup2 + ageGroup4 + gender + Race +
    employment + yearsSmoking + levelSmoking + priorAttempts


Step:  AIC=767.61
DayOfRelapse ~ grp + age + ageGroup4 + gender + Race + employment +
    yearsSmoking + levelSmoking + priorAttempts

                Df    AIC
- gender         1 765.61
- levelSmoking   1 765.61
- priorAttempts  1 765.70
- yearsSmoking   1 765.87
- Race           1 766.31
- ageGroup4      3 767.27
<none>             767.61
- age            1 767.67
- employment     2 770.85
- grp            1 773.58

Step:  AIC=765.61
DayOfRelapse ~ grp + age + ageGroup4 + Race + employment + yearsSmoking +
    levelSmoking + priorAttempts

                Df    AIC
- levelSmoking   1 763.61
- priorAttempts  1 763.71
- yearsSmoking   1 763.87
- Race           1 764.31
- ageGroup4      3 765.27
<none>             765.61
- age            1 765.73
- employment     2 768.96
- grp            1 771.58

Step:  AIC=763.61
DayOfRelapse ~ grp + age + ageGroup4 + Race + employment + yearsSmoking +
    priorAttempts

                Df    AIC
- priorAttempts  1 761.71
- yearsSmoking   1 761.91
- Race           1 762.32
- ageGroup4      3 763.29
<none>             763.61
- age            1 763.78
- employment     2 766.96
- grp            1 769.60

Step:  AIC=761.71
DayOfRelapse ~ grp + age + ageGroup4 + Race + employment + yearsSmoking

               Df    AIC
- yearsSmoking  1 760.03
- Race          1 760.39
- ageGroup4     3 761.43
<none>             761.71
- age           1 761.83
- employment    2 764.96
- grp           1 767.60
```

```
Step:  AIC=760.03
DayOfRelapse ~ grp + age + ageGroup4 + Race + employment

             Df    AIC
- Race        1 758.55
- age         1 759.85
<none>          760.03
- ageGroup4   3 760.07
- employment  2 763.25
- grp         1 766.27

Step:  AIC=758.55
DayOfRelapse ~ grp + age + ageGroup4 + employment

             Df    AIC
- ageGroup4   3 758.28
- age         1 758.42
<none>          758.55
- employment  2 761.52
- grp         1 764.85

Step:  AIC=758.28
DayOfRelapse ~ grp + age + employment

             Df    AIC
<none>          758.28
- employment  2 762.48
- grp         1 764.18
- age         1 767.24

> # backwards = step(everything,trace=0) would suppress step by step output.
> summary(backwards)

Call:
coxph(formula = DayOfRelapse ~ grp + age + employment, data = quit)

  n= 125, number of events= 89

                    coef exp(coef) se(coef)      z Pr(>|z|)
grpCombo        -0.60788   0.54450  0.21837 -2.784  0.00537 **
age             -0.03529   0.96533  0.01075 -3.282  0.00103 **
employmentother  0.70348   2.02077  0.26929  2.612  0.00899 **
employmentpt     0.65369   1.92262  0.32732  1.997  0.04581 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                exp(coef) exp(-coef) lower .95 upper .95
grpCombo           0.5445     1.8365    0.3549    0.8354
age                0.9653     1.0359    0.9452    0.9859
employmentother    2.0208     0.4949    1.1920    3.4256
employmentpt       1.9226     0.5201    1.0122    3.6518

Concordance= 0.638  (se = 0.03 )
Likelihood ratio test= 22.03  on 4 df,   p=0.0002
Wald test            = 21.91  on 4 df,   p=0.0002
Score (logrank) test = 22.48  on 4 df,   p=0.0002
```

```
> # Try forward selection
> nothing = coxph(DayOfRelapse ~ 1, data=quit) # Just the intercept
> summary(nothing)

Call:  coxph(formula = DayOfRelapse ~ 1, data = quit)

Null model
  log likelihood= -386.1533
  n= 125

> forwards = step(nothing,
+ scope=list(lower=formula(nothing),upper=formula(everything)),
direction="forward")

Start:  AIC=772.31
DayOfRelapse ~ 1

                 Df    AIC
+ ageGroup2       1 763.34
+ ageGroup4       3 766.09
+ grp             1 766.32
+ age             1 767.25
+ yearsSmoking    1 771.07
<none>              772.31
+ gender          1 773.57
+ Race            1 773.89
+ employment      2 774.25
+ levelSmoking    1 774.28
+ priorAttempts   1 774.30

Step:  AIC=763.34
DayOfRelapse ~ ageGroup2

                 Df    AIC
+ grp             1 758.09
+ employment      2 762.01
<none>              763.34
+ Race            1 765.06
+ yearsSmoking    1 765.15
+ priorAttempts   1 765.23
+ gender          1 765.27
+ levelSmoking    1 765.34
+ age             1 765.34
+ ageGroup4       2 766.09

Step:  AIC=758.09
DayOfRelapse ~ ageGroup2 + grp

                 Df    AIC
+ employment      2 755.10
<none>              758.09
+ Race            1 759.82
+ yearsSmoking    1 759.84
+ age             1 760.00
+ gender          1 760.05
+ levelSmoking    1 760.07
+ priorAttempts   1 760.08
+ ageGroup4       2 760.31
```

```
Step:  AIC=755.1
DayOfRelapse ~ ageGroup2 + grp + employment

                Df    AIC
<none>             755.10
+ Race           1 756.53
+ age            1 756.63
+ levelSmoking   1 757.05
+ yearsSmoking   1 757.08
+ gender         1 757.09
+ priorAttempts  1 757.09
+ ageGroup4      2 758.42
```

```
> summary(forwards)
Call:
coxph(formula = DayOfRelapse ~ ageGroup2 + grp + employment,
    data = quit)

  n= 125, number of events= 89

                   coef exp(coef) se(coef)      z Pr(>|z|)
ageGroup250+    -0.8803    0.4146   0.2401 -3.666 0.000246 ***
grpCombo        -0.6470    0.5236   0.2188 -2.957 0.003109 **
employmentother  0.6479    1.9114   0.2597  2.495 0.012601 *
employmentpt     0.5051    1.6571   0.3231  1.563 0.117992
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                exp(coef) exp(-coef) lower .95 upper .95
ageGroup250+       0.4146     2.4117    0.2590    0.6638
grpCombo           0.5236     1.9098    0.3410    0.8040
employmentother    1.9114     0.5232    1.1490    3.1798
employmentpt       1.6571     0.6035    0.8797    3.1214

Concordance= 0.646  (se = 0.033 )
Likelihood ratio test= 25.21  on 4 df,   p=0.00005
Wald test            = 24.46  on 4 df,   p=0.00006
Score (logrank) test = 25.26  on 4 df,   p=0.00004
```

```
> # Try a combination of forward and backward
> both = step( nothing,
scope=list(lower=formula(nothing),upper=formula(everything)) )

Start:  AIC=772.31
DayOfRelapse ~ 1

                Df    AIC
+ ageGroup2      1 763.34
+ ageGroup4      3 766.09
+ grp            1 766.32
+ age            1 767.25
+ yearsSmoking   1 771.07
<none>             772.31
+ gender         1 773.57
+ Race           1 773.89
+ employment     2 774.25
+ levelSmoking   1 774.28
+ priorAttempts  1 774.30
```

```
Step:  AIC=763.34
DayOfRelapse ~ ageGroup2

                Df    AIC
+ grp            1 758.09
+ employment     2 762.01
<none>             763.34
+ Race           1 765.06
+ yearsSmoking   1 765.15
+ priorAttempts  1 765.23
+ gender         1 765.27
+ levelSmoking   1 765.34
+ age            1 765.34
+ ageGroup4      2 766.09
- ageGroup2      1 772.31

Step:  AIC=758.09
DayOfRelapse ~ ageGroup2 + grp

                Df    AIC
+ employment     2 755.10
<none>             758.09
+ Race           1 759.82
+ yearsSmoking   1 759.84
+ age            1 760.00
+ gender         1 760.05
+ levelSmoking   1 760.07
+ priorAttempts  1 760.08
+ ageGroup4      2 760.31
- grp            1 763.34
- ageGroup2      1 766.32

Step:  AIC=755.1
DayOfRelapse ~ ageGroup2 + grp + employment

                Df    AIC
<none>             755.10
+ Race           1 756.53
+ age            1 756.63
+ levelSmoking   1 757.05
+ yearsSmoking   1 757.08
+ gender         1 757.09
+ priorAttempts  1 757.09
- employment     2 758.09
+ ageGroup4      2 758.42
- grp            1 762.01
- ageGroup2      1 767.24

> # For backwards, AIC=758.28. For forwards and both, AIC=755.1 -- lower.
```