Automated Model Selection Procedures -- Searching for "the best" regression model

When we are interested in prediction, we really have two goals for our regression mode: 1) Accuracy – the larger the R² the more accurate will be our y' values and 2) Efficiency – we don't want any unnecessary (and perhaps expensive) predictors in the model. To meet these two (somewhat contradictory) goals we need to identify a set of predictors with two attributes – all the predictors are related to the criterion variable, and the predictors are not strongly related to each other (called "reduced collinearity"). Over the years, there have three commonly used procedures for selecting a regression model with these characteristics from a larger set of predictors.

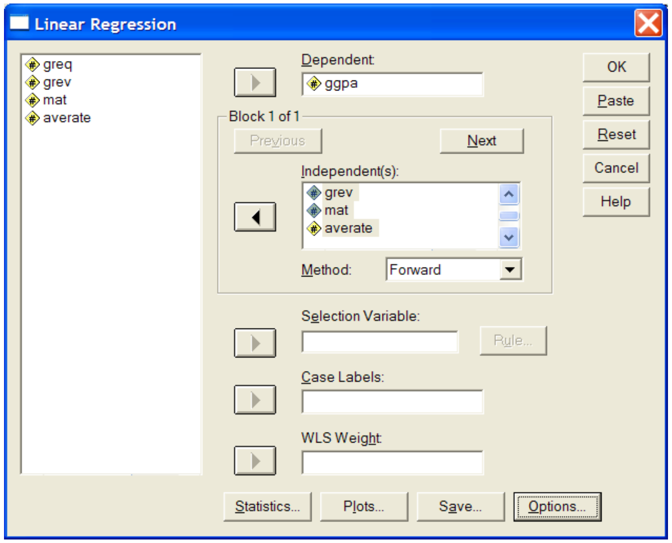
Forward Inclusion-- Start with that predictor having the highest simple correlation, and on each successive step, add that variable which will produce the largest increase in R-square (that with the largest partial correlation), stopping when an additional predictor will not increase R-square significantly.

Backward Deletion -- Start with a full model, on successive steps, delete the predictor that contributes the least to the model (that with the least significant/largest regression weight p-value), stopping when deleting the next variable would produce a significant drop in R-square (when all the variables in the model contribute).

Forward Stepwise Selection -- Think of this one as a combination of forward and backward. Start with that predictor having the highest simple correlation. For the second step, add the variable that will increase R-square the most (the one with the largest partial, but only if the R-square increase is significant). Each successive step has two parts: 1) if any predictor in the model is not contributing, toss it (if more than one, toss the one contributing the least, the one with the largest p-value), 2) if all variables in the model are contributing, then add that variable which will produce the largest increase in R-square (that with the largest partial correlation, but only if the R-square change will be significant). Stop when all the variables in the model are contributing, and when there is no additional predictor that will increase R-square significantly.

Analyze → Regression → Linear

- Move the criterion variable into the "Dependent" window
- More all predictors you are interested in into the "Independent(s) window
- Select the desired automated procedure (forward, backward or stepwise) from the drop-down Method menu
- If doing backward or stepwise, you may want to modify the p-value used to remove a variable from the model. The SPSS default is .10 (to prevent an infinite loop of dropping and then adding the same predictor(s)). Click the "Options" button then change the "Removal" value. Changing to .06 or .055 usually works well.



Linear Regression

greq
grev
mat
averate

Dependent: ggpa

Block 1 of 1

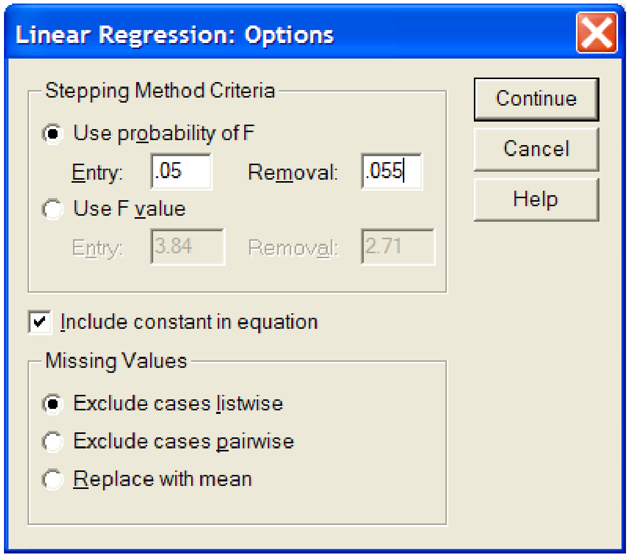
Independent(s): grev, mat, averate

Method: Forward

Selection Variable:

Case Labels:

WLS Weight:



Linear Regression: Options

Stepping Method Criteria

● Use probability of F

Entry: .05 Removal: .055

○ Use F value

Entry: 3.84 Removal: 2.71

☑ Include constant in equation

Missing Values

● Exclude cases listwise

○ Exclude cases pairwise

○ Replace with mean

Forward Selection

```
Equation Number 1    Dependent Variable..   GPA
   Beginning Block Number  1.  Method:  Forward

   Variable(s) Entered on Step Number  1..   GREQ
                                             Analysis of Variance
   Multiple R           .61090                          DF      Sum of Squares      Mean Square
   R Square             .37320      Regression           1             3.89423          3.89423
   Adjusted R Square    .35082      Residual            28             6.54044           .23359
   Standard Error       .48331      F =      16.67143       Signif F =  .0003



   ------ Variables in the Equation -----     ------------- Variables not in the Equation -------------
   Variable              B      T  Sig T      Variable     Beta In  Partial  Min Toler         T  Sig T
   GREQ       7.435521E-03  4.083  .0003      GREV          .38162   .42841     .78992     2.464  .0204
   (Constant)      -.89269  -.863  .3952      MAT           .47699   .58147     .93147     3.714  .0009
                                              AVERATE   1.7618E-03   .00220     .97564      .011  .9910




   Variable(s) Entered on Step Number            2..    MAT
                                  Analysis of Variance
   Multiple R           .66494                        DF      Sum of Squares      Mean Square
   R Square             .43561    Regression           2             6.10562          3.05281
   Adjusted R Square    .55440    Residual            27             4.32904           .16033
   Standard Error       .40042    F =      19.04022       Signif F =  .0000

   --------- Variables in the Equation --------     ------------- Variables not in the Equation -------------
   Variable              B        T  Sig T   ←       Variable     Beta In  Partial  Min Toler         T  Sig T
   GREQ       5.915671E-03    1.660  .0728          GREV          .22567   .29095     .68957     1.551  .1331
   MAT              .03094    3.714  .0009          AVERATE       .03897   .05958     .90416      .304  .7633
   (Constant)     -2.10596   -2.297  .0296

   End Block Number   1   PIN =   .050 Limits reached.
                         (PIN = p-value for input -- i.e., no other variable would contribute)
```

Notice that GREQ is an example of "over inclusion" -- it contributed initially, but doesn't contribute when MAT was added to the model.

Backward Selection

```
 Equation Number 2   Variable(s) Entered on Step Number   1..    AVERATE  2..    GREV   3..    MAT  4..    GREQ
                               Analysis of Variance
Multiple R           .78904                           DF      Sum of Squares      Mean Square
R Square             .62258        Regression          4             6.49639          1.62410
Adjusted R Square    .56219        Residual           25             3.93828           .15753
Standard Error       .39690        F =      10.30968       Signif F =  .0000

------ Variables in the Equation --------
Variable              B          T  Sig T
AVERATE    3.996575E-04       .393  .6978
GREV       1.652321E-03      1.544  .1352
MAT              .02633      2.974  .0064
GREQ       4.772980E-03      2.782  .0101
(Constant)     -2.10804     -2.308  .0296
```

Notice that while neither AVERATE nor GREV are contributing, AVERATE can be said to be "contributing less" because its regression weight is less likely to be different from 0 (i.e., larger p-value – more likely to be a Type I error)

```
Variable(s) Removed on Step Number           5..    AVERATE
                           Analysis of Variance
Multiple R           .78756                           DF      Sum of Squares      Mean Square
R Square             .62025    Regression          3             6.47208          2.15736
Adjusted R Square    .57643    Residual           26             3.96259           .15241
Standard Error       .39039    F =      14.15524       Signif F =  .0000

------ Variables in the Equation --------      ------------- Variables not in the Equation -------------
Variable              B         T  Sig T        Variable     Beta In  Partial  Min Toler         T  Sig T
GREQ       1.630285E-03     1.551  .1331        AVERATE       .04908   .07832     .68768      .393  .6978
MAT              .02614     3.006  .0058
GREV       4.892734E-03     2.946  .0067
(Constant)     -2.14336    -2.397  .0240


Variable(s) Removed on Step Number           6..    GREQ
                            Analysis of Variance
Multiple R           .76494                           DF      Sum of Squares      Mean Square
R Square             .58513     Regression          2             6.10562          3.05281
Adjusted R Square    .55440     Residual           27             4.32904           .16033
Standard Error       .40042     F =      19.04022       Signif F =  .0000

------- Variables in the Equation --------       ------------- Variables not in the Equation -------------
Variable              B         T  Sig T          Variable     Beta In  Partial  Min Toler         T  Sig T
MAT              .03094     3.714  .0009          GREQ          .12567   .19095     .68957     1.551  .1331
GREV       5.915671E-03     3.784  .0008          AVERATE       .23897   .25958     .90416     2.304  .0433 ←
(Constant)     -2.10596    -2.297  .0296

     End Block Number   2   POUT =     .055 Limits reached.
                           (POUT = p-value to output - i.e., all variables contribute)
```

Notice that AVERATE is an example of "under inclusion"-- it didn't contribute in the full model, but would contribute if added back into this one.

Forward Stepwise Selection

```
Equation Number 3    Dependent Variable..   GPA
Beginning Block Number  1.  Method:  Stepwise         Variable(s) Entered on Step Number  1..   GREQ

                                          Analysis of Variance
Multiple R           .61090                          DF      Sum of Squares      Mean Square
R Square             .37320      Regression           1             3.89423          3.89423
Adjusted R Square    .35082      Residual            28             6.54044           .23359
Standard Error       .48331      F =      16.67143       Signif F =  .0003

------ Variables in the Equation -----     ------------- Variables not in the Equation -------------
Variable              B      T  Sig T      Variable     Beta In  Partial  Min Toler         T  Sig T
GREQ       7.435521E-03  4.083  .0003      GREV          .38162   .42841     .78992     2.464  .0204
(Constant)      -.89269  -.863  .3952      MAT           .47699   .58147     .93147     3.714  .0009
                                           AVERATE   1.7618E-03   .00220     .97564      .011  .9910


Variable(s) Entered on Step Number           2..    MAT
                               Analysis of Variance
Multiple R           .66494                        DF      Sum of Squares      Mean Square
R Square             .43561    Regression          2             6.10562          3.05281
Adjusted R Square    .55440    Residual           27             4.32904           .16033
Standard Error       .40042    F =      19.04022       Signif F =  .0000

--------- Variables in the Equation --------     ------------- Variables not in the Equation -------------
Variable              B        T  Sig T          Variable     Beta In  Partial  Min Toler         T  Sig T
GREQ       5.915671E-03    1.660  .0728          GREV          .22567   .29095     .68957     1.551  .1331
MAT              .03094    3.714  .0009          AVERATE       .03897   .05958     .90416      .304  .7633
(Constant)     -2.10596   -2.297  .0296



Variable(s) Removed on Step Number           3..    GREQ
                               Analysis of Variance
Multiple R           .59887                        DF      Sum of Squares      Mean Square
R Square             .35863    Regression          1             3.56462          3.56462
Adjusted R Square    .31245    Residual           28             6.84125           .24539
Standard Error       .52182    F =      14.52335       Signif F =  .0009

--------- Variables in the Equation --------     ------------- Variables not in the Equation -------------
Variable              B        T  Sig T          Variable     Beta In  Partial  Min Toler         T  Sig T
MAT              .04156    4.248  .0004          GREV          .22567   .31095     .63957     1.651  .1164
(Constant)     -1.34566   -1.876  .0562          AVERATE       .03897   .05958     .80416      .367  .7317
                                                 GREQ          .21321   .27653     .76121     1.660  .0728
End Block Number   1   PIN =     .050 Limits reached.   POUT =     .055 Limits reached.
```

You should notice that the three procedures ended up with three different models!!!!   What might this tell you?