Selecting the best regression equation

Problem: We have one independent variable Y and k predictor variables X_1, X_2,…, X_k. The goal is to find the "best" subset of the predictor variables. "Best" is defined by several criteria and has a great deal to do with the reason that we undertake this best model selection in the first place.

There are two broad categories when looking for a best-fitting subset of all possible predictors:

1. *Reliability*. If we want to predict Y, by using a combination of the predictors (i.e., emphasize in $E(Y| X_1, X_2,\ldots, X_k.)$), then the model that best predicts Y is said to be *reliable*. In this case, the main goal is accurate estimation of Y, and not so much the particulars of the model itself.
2. *Validity*. When the emphasis is on assessing the relationship between Y and some of the predictors after accounting for the presence of all others (as in the case of a disease and exposure to pollutants, while controlling for demographic or socioeconomic factors), then we emphasize on prediction of the regression coefficients searching for a *valid* regression model.

Steps in selecting the best regression model

1. Specify the maximum model to be considered. That is identify *all possible predictors*
2. Specify the selection criterion
3. Specify a strategy for selecting predictors (variables)
4. Conduct the specified analysis
5. Evaluate the reliability of the chosen model

Step 1: Specifying the maximum model

Starting the model selection process from the maximum model is desirable because:

1. The model contains all conceivable predictors
2. The model contains all higher order (polynomial, such as $(\text{AGE})^2$, log(WGT), etc.) terms and interactions among terms (GENDER $\times$ WGT).
3. The model contains all possible control variables (variables of secondary interest that can modify the values of the variables of interest and thus must be taken into account)
4. "Over-fitting" the model (including irrelevant variables with zero population regression coefficients) does not introduce *bias*[1] in the model, but "under-fitting" (omitting relevant variables) does. However, "over-fitting" can introduce numerical instability, as some predictors may be correlated (thus (**X'X**) will be of rank $r<k$ and $(\mathbf{X}'\mathbf{X})^{-1}$ will not exist). Thus, care must be taken when determining the maximum model.

How large should the maximum model be?

Note that if the number of predictors and intercept $k+1=n$ the number of observations, then the error degrees of freedom will be 0. We know from mathematics that we can fit a $(k+1)^{th}$ expression through k points *exactly*. Thus, no matter what predictors we choose, $R^2 = \frac{SSY - SSE}{SSY} = \frac{SSY - 0}{SSY} = 1.0$ regardless of whether the regression model is reasonable or not.

The weakest requirement is for e.d.f. (error degrees of freedom) = n-k-1$\geq$10.

Another suggested “rule-of-thumb” is $n\geq 5k$ or even $n\geq 10k$. Thus, if we have 50 observations, the largest model should be between k=5 (since $n\geq$(5)(10)=50) and k=10 (thus, $n\geq$(5)(10)=50).

[1] Bias η is a deviation of the expected value of an estimator from the population value ($E(\hat{Y}) = \bar{Y} + \eta$). When η=0, then the estimator is called *unbiased*. In that case, $E(\hat{Y}) = \bar{Y}$

Step 2: Specifying the (model) selection criterion

Consider the full model as $Y = \beta_o + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_{p+1} X_{p+1} + \cdots \beta_k X_k + \varepsilon$ and the reduced model as $Y = \beta_o + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$, with $p \leq k$. Denote $SSE(p)$ the error sum of squares of the reduced (p-variable) model, and $SSY = \sum \left(Y_i - \overline{Y}\right)^2$ is the total corrected sum of squares.

Then candidates for model selection criteria are as follows:

1. The squared multiple correlation $R^2\left(Y | X_1, X_2, \cdots X_p\right) = 1 - \frac{SSE(p)}{SSY}$. This is maximum in the maximum model
2. The F statistic comparing the full and restricted models $F_p = \frac{\left[SSE(p) - SSE(k)\right]/(k-p)}{MSE(k)}$, where $MSE(k)=SSE(k)/(n-k-1)$.
3. The variance (which we would like to minimize) in the p-variable model $MSE(p) = \frac{SSE(p)}{n-p-1}$.
4. The Mallows's C_p statistic which equals $C_p = \frac{SSE(p)}{\hat{\sigma}} - [n - 2(p+1)]$, where $\hat{\sigma}$ is the best estimate of the variance of Y, and usually we take $\hat{\sigma} = MSE(k)$ the mean square error of the full model.

Comments:

1. The multiple squared correlation has several drawbacks, the most significant of which is that it never decreases, even when irrelevant predictors are added. In fact it always attains its maximum. Some use an alternative form called adj. $R_p^2 = \frac{SSE/(n-p-1)}{SSY/(n-1)} = 1-\left[(1-R^2)\frac{(n-1)}{(n-p-1)}\right]$ for this reason. Note that adj. R_p^2 adjusts R^2 for the size of each model.
2. The F_p statistic measures whether the relative change in the error by adding a number of k-p predictors. It should be compared to a tail of an $F_{(k-p),(n-k-1)}$ distribution.
3. Notice that if we consider the best estimate of the variance of Y as $\hat{\sigma} = MSE(k)$, then the Mallow's C_p statistic will be exactly $C_p = k+1$ for the full model. Thus, if a restricted p-variable model has variance that is close to $MSE(k)$, then $C_p \approx p+1$. This is a criterion to identify candidate smaller models. For example, a good candidate 3-variable model would be one with $C_p \approx 4$.
4. All these criteria contain the same information as $F_p = \frac{\left(R_k^2 - R_p^2\right)/(k-p)}{\left(1-R_k^2\right)/(n-k-1)}$, and $C_p=(k-p)F_p+(2p-k+1)$.

CAUTION!!!

SINCE A LARGE NUMBER OF TESTS ARE CARRIED OUT AT EACH STEP, NO-ONE KNOWS THE TRUE SIGNIFICANCE LEVEL (α-LEVEL, TYPE-I ERROR RATE) OF THE TESTS

Step 3: Strategies for selection of the best model

1. *All possible regressions procedure*. This procedure requires fitting of all the possible regression models, and then deciding which one to choose, based on one of the criteria that were mentioned earlier. Note that if each predictor can either be present or absent from each model, there are 2^k-1 combination (minus the model with no predictors. If k=10, there are 1,023 possible models.

Step 3: Strategies for selection of the best model (continued)

2. *Backward-elimination procedure.* This procedure is implemented as follows:

i. A maximum p-value for *removal* is pre-specified.

ii. The maximum model is fitted.

iii. All the partial-*F* statistics (variables-added-last, or Type III) are computed.

iv. If the highest p-value (corresponding to the least significant variable) is larger than the p-value for removal, then the corresponding variable is removed.

v. If no variable is removed the process is stopped, and the remaining variables are declared as the "optimum" model. If a variable is removed, then the resulting (reduced) model is fitted as the maximum model (step ii), and the process repeats until no other variables can be removed.

Step 3: Strategies for selection of the best model (continued)

3. *Forward-selection procedure.* This procedure is implemented as follows:

i. A maximum p-value for *entry* is pre-specified.

ii. Fit each (individual) variable.

iii. All model (simple linear regression) *F* statistics are computed.

iv. If the lowest p-value of the Type-III *F* test[2] (corresponding to the most significant variable) is larger than the p-value for entry, stop. If the lowest p-value is smaller than the entry p-value then enter this variable.

v. For the remaining variables not yet in the model, compute their Type-III partial *F* statistics controlling for all variables in the model. Then go to step iv and repeat the process until no variables can be entered.

[2] Note that in the first case (simple linear regression) the Type-III *F* test is equivalent to the model (overall) *F* test as there are no other variables in the model

Step 3: Strategies for selection of the best model (continued)

4. *Stepwise regression procedure.* This procedure is a version of the forward selection that allows re-examination of the fitted model at each step. It is implemented as follows:

i. A maximum p-value for *removal* and a minimum for entry are pre-specified. These do not have to be equal, but $p_r \geq p_e$ to assure that no variable removed from the model can be entered again in the same step (leading to an infinite loop).

ii. All partial Type-III *F* tests for each candidate variable not in the model adjusted for all variables present in the model[3] are computed. If the lowest p value (corresponding to the most significant candidate variable) is *lower* than the entry p-value, that variable is entered.

iii. The partial *F* statistics of all variables in the model after step ii are computed. If the highest p-value (corresponding to the least significant variable) is *higher* than the p value for removal, the variable is removed.

iv. The model is refitted and step iii is repeated, until no more variables can be removed.

v. The process then goes to step ii, and continues until no variables can be added or removed.

Comments:

1. Only the all-possible-regressions method is guaranteed to provide the complete information and thus suggest the best model given the desired criteria. Note that the "best" model is a consequence of the goals of the experimenter and non-statistical factors (such as cost, interpretation, etc.)
2. No other model-selection method is guaranteed to identify the best model. Both backward-elimination and forward-selection method, depend a great deal on the ordering of the variables, and different starting maximum models (in the case of the forward-selection, the specified maximum model will determine the order of the addition of the variables) may produce different "optimum" models.
3. The stepwise-selection method protects from some of the drawbacks of the other two, but is itself affected by the ordering of the variables in the (maximum) model.
4. Thus, the only chance that these methods have to identify a good restricted model, is that the investigator has some prior knowledge of the relative significance of the predictor variables.

---

[3] As above, when the first variable is entered, the Type-III partial F test is equivalent to the model (overall regression) F test.

Comments : "Chunk-wise" regression

5. Because some variables are non-significant individually, but are very significant when taken together, or because we want to consider certain variables together in the model, we can consider adding (or removing) variables in "chunks". This is not a novel or difficult concept. The only change in the backward, forward and stepwise methods is that instead of a Type-III partial *F* test, we will be computing a Type-III *multiple* partial *F* test (of the group or chunk of variables considered adjusted for the variables already present in the model).

Step 4: Conducting the analysis

After the "best" model has been identified, then the statistical analysis must be performed. All the goodness-of-fit and model-checking procedures that we mentioned earlier must be performed, to assure that the model is adequate for the data at hand. **Note that most computer-based model-selection routines do not check the model assumptions for each candidate model, so the "best" model may end up violating some modeling assumptions (normality, homoskedacity, etc.)!**

Step 5: Reliability of the chosen model: The split-sample approach.

To evaluate the reliability of the model, i.e., to assess whether the model will do just as good a job predicting future data sets) as it does with the data set at hand. The following method can be used:

1. We split the original data set in two parts: (i) a *training* sample and (ii) a *validation* sample.
2. The "best" model is identified, by working on the training sample. The regression is then performed, and the regression equation $\hat{Y}_1 = \hat{\beta}_o + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$ is computed, where $\left\{\hat{Y}_{i1}\right\}_{i=1}^{n_1}$ is the vector of n_1 fitted Y's from the training sample, along with $R^2(1) = R^2\left(Y_1 | X_1, X_2, \cdots X_p\right) = r^2\left(Y_1, \hat{Y}_1\right)$
3. Using $\hat{\beta}_o, \hat{\beta}_1, \ldots, \hat{\beta}_p$ we compute $\left\{\hat{Y}^*_{i2}\right\}_{i=1}^{n_2}$, the vector of fitted Y's from the validation sample **based on the regression coefficients calculated from the training sample** and $R^2_*(2) = r^2\left(Y_2, \hat{Y}^*_2\right)$ the *cross-validation correlation*
4. The quantity $R^2(1) - R^2_*(2)$ is called the *shrinkage on cross-validation*. Shrinkage values less than 0.10 are indicative of a reliable model, while values close to 0.90 are problematic.

Comments on the split-sample approach.

1. There are several methods used to assure that an appropriate training and validation sample is obtained. Only few studies can accomplish both roles (the size of the study must be large).
2. If the study is large enough, then the two samples can be determined by methods ranging from the naïve, to the extremely complex. A computer program will most likely be necessary. Some methods of assignment are as follows:

i. Random sampling: Each observation is included in the training sample with probability $p = \frac{n_1}{n_1 + n_2}$.

ii. If categorical variables (such as race or gender) are important, then assignment may be *stratified*. Stratification means that both data sets will have similar representation of observation with the same combinations of the strata of interest.

iii. The pair-matching technique. In this approach, observations with similar levels of certain variables of interest are assigned at random to one or the other sample.

Example: The weight data

<u>The “all-possible-regressions method</u>

We use the weight data to explore the best model by the all possible regressions method. Since there are only 3 variables in the model, there are 2^3-1=7 possible models. These are given below:

Model 1: WGT= $\beta_o+\beta_1$HGT+ε

```
. reg wgt hgt

  Source |       SS       df       MS                  Number of obs =      12
---------+------------------------------              F(  1,    10) =   19.67
   Model |  588.922523     1  588.922523               Prob > F      =  0.0013
Residual |  299.327477    10  29.9327477               R-squared     =  0.6630
---------+------------------------------               Adj R-squared =  0.6293
   Total |      888.25    11       80.75               Root MSE      =  5.4711
------------------------------------------------------------------------------
     wgt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     hgt |    1.07223    .241731      4.436   0.001       .5336202    1.610841
   _cons |   6.189849   12.84875      0.482   0.640      -22.43894    34.81864
------------------------------------------------------------------------------
```

Model 2: $WGT=\beta_o+\beta_2 AGE+\varepsilon$

```
. reg wgt age

  Source |       SS       df       MS                 Number of obs =      12
---------+------------------------------              F(  1,    10) =   14.55
   Model |  526.392857     1  526.392857              Prob > F      =  0.0034
Residual |  361.857143    10  36.1857143              R-squared     =  0.5926
---------+------------------------------              Adj R-squared =  0.5519
   Total |      888.25    11       80.75              Root MSE      =  6.0155

------------------------------------------------------------------------------
     wgt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   3.642857   .9551151      3.814   0.003       1.514728    5.770986
   _cons |   30.57143   8.613705      3.549   0.005        11.3789    49.76396
------------------------------------------------------------------------------
```

Model 3: $WGT = \beta_o + \beta_3 AGE2 + \varepsilon$

```
. reg wgt age2

  Source |       SS       df       MS              Number of obs =      12
---------+------------------------------           F(  1,    10) =   14.25
   Model |  521.932047     1  521.932047           Prob > F      =  0.0036
Residual |  366.317953    10  36.6317953           R-squared     =  0.5876
---------+------------------------------           Adj R-squared =  0.5464
   Total |      888.25    11       80.75           Root MSE      =  6.0524

------------------------------------------------------------------------------
     wgt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    age2 |   .2059716   .0545669      3.775   0.004       .0843889    .3275543
   _cons |   45.99764    4.76964      9.644   0.000       35.37022    56.62506
------------------------------------------------------------------------------
```

Model 4: $WGT = \beta_o + \beta_1 HGT + \beta_2 AGE + \varepsilon$

```
. anova wgt hgt age, continuous(hgt age)
                 Number of obs =      12     R-squared     =  0.7800
                 Root MSE      = 4.65984     Adj R-squared =  0.7311

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  692.822607     2  346.411303      15.95     0.0011
                         |
                     hgt |  166.429749     1  166.429749       7.66     0.0218
                     age |  103.900083     1  103.900083       4.78     0.0565
                         |
                Residual |  195.427393     9  21.7141548
              -----------+----------------------------------------------------
                   Total |      888.25    11       80.75
. reg

  Source |       SS       df       MS                    Number of obs =      12
---------+-----------------------------                  F(  2,     9) =   15.95
   Model |  692.822607     2  346.411303                 Prob > F      =  0.0011
Residual |  195.427393     9  21.7141548                 R-squared     =  0.7800
---------+-----------------------------                  Adj R-squared =  0.7311
   Total |      888.25    11       80.75                 Root MSE      =  4.6598

------------------------------------------------------------------------------
     wgt        Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
------------------------------------------------------------------------------
_cons       6.553048   10.94483      0.599   0.564      -18.20587    31.31197
hgt          .722038   .2608051      2.768   0.022       .1320559     1.31202
age         2.050126   .9372256      2.187   0.056      -.0700253    4.170278
------------------------------------------------------------------------------
```

Model 5: $WGT = \beta_o + \beta_1 HGT + \beta_3 (AGE)^2 + \varepsilon$

```
. anova wgt hgt age2, continuous(hgt age2)

                           Number of obs =      12     R-squared     =  0.7764
                           Root MSE      = 4.69752     Adj R-squared =  0.7267

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  689.649951     2  344.824976      15.63     0.0012
                         |
                     hgt |  167.717904     1  167.717904       7.60     0.0222
                    age2 |  100.727428     1  100.727428       4.56     0.0614
                         |
                Residual |  198.600049     9  22.0666721
              -----------+----------------------------------------------------
                   Total |      888.25    11       80.75
. reg

  Source |       SS       df       MS                  Number of obs =      12
---------+------------------------------               F(  2,     9) =   15.63
   Model |  689.649951     2  344.824976               Prob > F      =  0.0012
Residual |  198.600049     9  22.0666721               R-squared     =  0.7764
---------+------------------------------               Adj R-squared =  0.7267
   Total |      888.25    11       80.75               Root MSE      =  4.6975

------------------------------------------------------------------------------
     wgt        Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
------------------------------------------------------------------------------
_cons       15.11754    11.7969      1.281   0.232       -11.5689    41.80398
hgt         .7259765   .2633306      2.757   0.022       .1302814    1.321672
age2        .1148016   .0537332      2.137   0.061      -.0067513    .2363546
------------------------------------------------------------------------------
```

Model 6: $WGT = \beta_o + \beta_2 AGE + \beta_3 (AGE)^2 + \varepsilon$

```
. anova wgt age age2, continuous(age age2)

                           Number of obs =      12     R-squared     =  0.5927
                           Root MSE      =  6.3401     Adj R-squared =  0.5022

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  526.478508     2  263.239254       6.55     0.0176
                         |
                     age |   4.5464612     1   4.5464612       0.11     0.7443
                    age2 | .085651307     1  .085651307       0.00     0.9642
                         |
                Residual |  361.771492     9  40.1968324
              -----------+----------------------------------------------------
                   Total |      888.25    11       80.75
. reg

  Source |       SS       df       MS                  Number of obs =      12
---------+------------------------------               F(  2,     9) =    6.55
   Model |  526.478508     2  263.239254               Prob > F      =  0.0176
Residual |  361.771492     9  40.1968324               R-squared     =  0.5927
---------+------------------------------               Adj R-squared =  0.5022
   Total |      888.25    11       80.75               Root MSE      =  6.3401

------------------------------------------------------------------------------
     wgt        Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
------------------------------------------------------------------------------
_cons        32.40411   40.72717      0.796   0.447      -59.72715    124.5354
age          3.205364   9.530956      0.336   0.744      -18.35516    24.76589
age2         .0249816   .5411899      0.046   0.964      -1.199275    1.249238
------------------------------------------------------------------------------
```

Model 7: $\beta_o+\beta_1 HGT+\beta_2 AGE+\beta_3 (AGE)^2+\varepsilon$

```
. anova wgt hgt age age2, continuous(hgt age age2)
                           Number of obs =      12     R-squared     =  0.7803
                           Root MSE      =  4.9395     Adj R-squared =  0.6978

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  693.060463     3  231.020154       9.47     0.0052
                         |
                     hgt |  166.581955     1  166.581955       6.83     0.0310
                     age |  3.41051231     1  3.41051231       0.14     0.7182
                    age2 |  .237856856     1  .237856856       0.01     0.9238
                         |
                Residual |  195.189537     8  24.3986921
              -----------+----------------------------------------------------
                   Total |      888.25    11       80.75
. reg
  Source |       SS       df       MS                       Number of obs =      12
---------+------------------------------                    F(  3,     8) =    9.47
   Model |  693.060463     3  231.020154                    Prob > F      =  0.0052
Residual |  195.189537     8  24.3986921                    R-squared     =  0.7803
---------+------------------------------                    Adj R-squared =  0.6978
   Total |      888.25    11       80.75                    Root MSE      =  4.9395

------------------------------------------------------------------------------
     wgt        Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
------------------------------------------------------------------------------
_cons        3.438426   33.61082      0.102   0.921      -74.06826    80.94512
hgt          .7236902   .2769632      2.613   0.031        .085012    1.362368
age          2.776875   7.427279      0.374   0.718      -14.35046    19.90421
age2        -.0417067   .4224071     -0.099   0.924      -1.015779    .9323659
------------------------------------------------------------------------------
```

Summary of results of the all possible regressions method:

Model	No. of variables	Variables used	Estimated coefficients $\hat{\beta}_o$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	Partial F statistics X_1	X_2	X_3	Overall F statistic	R_p^2	$MSE(p)$	C_p
1	1	HGT (X_1)	6.190	1.073			19.67**			19.67**	0.663	29.93	4.27
2	1	AGE (X_2)	30.57		3.64			14.55**		14.55**	0.593	36.18	6.83
3	1	$(AGE)^2$ (X_3)	46.00			0.21			14.25**	14.25**	0.588	36.63	7.01
4	2	HGT, AGE	6.55	0.72	2.05		7.67*	4.79		15.95**	0.780	21.71	2.01
5	2	HGT, $(AGE)^2$	15.12	0.73		0.12	7.60*		4.57	15.63**	0.776	22.07	2.14
6	2	AGE, $(AGE)^2$	32.40		3.21	0.03		0.113	0.002	6.55*	0.593	40.2	8.83
7	3	HGT, AGE, $(AGE)^2$	3.44	0.72	2.78	-0.04	6.827*	0.140	0.010	9.47**	0.780	24.40	4.00

Conclusions:

1. Even though all three single-predictor models (1,2,3) have significant overall *F* tests, the Mallow's C_p statistic clearly implies that these models are inadequate. Recall that for a single-predictor model we would like a C_p close to 2.
2. Another statistic that implies that a single-predictor model may be inadequate is the F_p statistic. For the
3. best model (Model 1), this is $F_p = \frac{\left(R_k^2 - R_p^2\right)/(k-p)}{\left(1-R_k^2\right)/(n-k-1)} = \frac{(0.780-0.663)/(3-1)}{(1-0.780)/(12-3-1)} = 2.127$. We can compare this to an $F_{2,8}$. Then $F_{2,8;0.75} < F_1 < F_{2,8;0.90}$. This means that the full model is not significantly superior from the single-variable model (since the increase in R^2 is not large enough). However, given the size of the data, this is not surprising.
4. By the same criterion, the 3-predictor model is clearly over-fitted, as the change in R^2 is very small. From the two-predictor models, Model 4 should probably be selected based on both its Mallow's C_p statistic and its largest R^2 among two-predictor models. Also, F_2=0.007 implying that the addition of a third variable (`AGE2`) in the model has an infinitesimal effect.

Conclusions (continued):

5. Notice that the F_p tests involve the $MSE(k)$ in the denominator. Notice that

$$F_p = \frac{\left[\left(R_k^2 - R_p^2\right)/(k-p)\right]}{\left[\left(1-R_k^2\right)/(n-k-1)\right]} = \left(\frac{\frac{SSY-SSE(k)}{SSY} - \frac{SSY-SSE(p)}{SSY}}{1-\frac{SSY-SSE(k)}{SSY}}\right)\frac{n-k-1}{k-p}$$

$$= \frac{\left[SSE(p)-SSE(k)\right]/(k-p)}{SSE(k)/(n-k-1)}$$

$$= \frac{\left[SSE(p)-SSE(k)\right]/(k-p)}{MSE(k)}$$

so that this test is equivalent to a (multiple) partial F test. These tests are preferable to the overall F tests, as the latter are computed on the restricted model's $MSE(p)$ (which may or may not be a good estimate of the overall variance). By contrast, these tests are based on the $MSE(k)$ of the full model.

6. Finally, for Model 4, $MSE(2)$=21.71, which is the lowest among all 7 models considered.

The backward elimination procedure

i. Specify the maximum p-value for *removal*. Say 0.10.

ii. The maximum model is fitted. This is Model 7 from above and it is

$$\text{WGT} = 3.438 + (0.724)\text{HGT} + (2.777)\text{AGE} - (0.042)(\text{AGE})^2$$

iii. All the Type-III partial-*F* statistics are computed. These are (from Model 7) F(`HGT|AGE, AGE2`)= 6.83, F(`AGE|HGT, AGE2`)= 0.14 and F(`AGE2|HGT, AGE`)= 0.01.

iv. Since the smallest *F* test (corresponding to `AGE2`) is F=0.10<$F_{1,8;0.90}$=3.46 then `AGE2` is removed.

v. Refit the model (now Model 4). Go to step ii.

ii. The Type-III *F* tests are F(`HGT|AGE`)=7.67, F(`AGE|HGT`)=4.78. The least significant test is that associated with `AGE.` Since F=4.79> $F_{1,9;0.90}$=3.36 the variable is not removed and the algorithm terminates.

The best model is model [`HGT, AGE`], with $\text{WGT} = 6.553 + (0.722)\text{HGT} + (2.050)\text{AGE}$.

The way to instantaneously do this with STATA is to use the `sw` command as follows:

```
. sw  reg wgt hgt age age2, pr(0.10)

                      begin with full model
p = 0.9238 >= 0.1000  removing age2

  Source |       SS       df       MS                  Number of obs =      12
---------+------------------------------               F(  2,     9) =   15.95
   Model |  692.822607     2  346.411303               Prob > F      =  0.0011
Residual |  195.427393     9  21.7141548               R-squared     =  0.7800
---------+------------------------------               Adj R-squared =  0.7311
   Total |      888.25    11       80.75               Root MSE      =  4.6598

------------------------------------------------------------------------------
     wgt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     hgt |    .722038   .2608051      2.768   0.022        .1320559     1.31202
     age |   2.050126   .9372256      2.187   0.056       -.0700253    4.170278
   _cons |   6.553048   10.94483      0.599   0.564       -18.20587    31.31197
------------------------------------------------------------------------------
```

<u>The forward selection procedure</u>

The steps involved here are as follows:

i. A maximum p-value for *entry* is pre-specified. Say 0.10.

ii. By fitting each single variable (Models 1,2 and 3), the most highly correlated with Y variable is `HGT` with squared correlation R^2=0.663.

iii. The overall F statistic is F=19.67 which is significant at the 0.10 level (the criterion for entry)

iv. Since the p-value of the F test is smaller than the p-value for entry, `HGT` is entered.

v. For `AGE` and `AGE2` compute the Type-III F statistics controlling for `HGT.` Then go to step iii.

iii. F(`AGE|HGT`)=4.78 (Model 4) and F(`AGE2|HGT`)=4.56 (Model 5).

iv. Since the p-value of F(`AGE|HGT`)=4.78>$F_{1,9;0.90}$=3.36, variable `AGE` is entered in the model. Go back to step iii.

iii. F(`AGE2|HGT,AGE`)=0.010 (Model 7). Since F(`AGE2|HGT,AGE`)=0.10< $F_{1,8;0.90}$=3.46 `AGE2` is not entered and the algorithm terminates.

The best model is again model [`HGT, AGE`], with $\text{WGT} = 6.553 + (0.722)\text{HGT} + (2.050)\text{AGE}$.

The STATA output for a forward selection procedure is as follows:

```
. sw  reg wgt hgt age age2, pe(0.10)

                      begin with empty model[4]
p = 0.0013 <  0.1000  adding   hgt
p = 0.0565 <  0.1000  adding   age

  Source |       SS       df       MS                  Number of obs =      12
---------+------------------------------               F(  2,     9) =   15.95
   Model |  692.822607     2  346.411303               Prob > F      =  0.0011
Residual |  195.427393     9  21.7141548               R-squared     =  0.7800
---------+------------------------------               Adj R-squared =  0.7311
   Total |      888.25    11       80.75               Root MSE      =  4.6598

------------------------------------------------------------------------------
     wgt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     hgt |    .722038   .2608051      2.768   0.022       .1320559     1.31202
     age |   2.050126   .9372256      2.187   0.056      -.0700253    4.170278
   _cons |   6.553048   10.94483      0.599   0.564      -18.20587    31.31197
------------------------------------------------------------------------------
```

The stepwise selection algorithm

The steps in this routine are as follows:

i. Specify the p-values for entry and removal. Assume p_r=0.20 and p_e =0.10 since $p_r \le p_e$.

ii. Just like in step ii in the forward-selection algorithm above `HGT` is added to the model.

iii. The Type-III F for or `AGE` and `AGE2` are computed controlling for `HGT`. F(`AGE|HGT`)=4.78 (Model 4) and F(`AGE2|HGT`)=4.56 (Model 5). By the same argument as before `AGE` is added to the model.

iv. The Type-III F tests for each candidate variable in the model are computed. These are F(`HGT|AGE`)=7.66>$F_{1,9;0.80}$=1.91 and F(`AGE|HGT`)=4.78>$F_{1,9;0.80}$=1.91 (Model 4). Thus, both variables remain in the model. Now go to step iii.

iii. The partial F statistic F(`AGE2|HGT,AGE`)=0.010 is computed (Model 7). Since 0.10< $F_{1,9;0.90}$=3.46 `AGE2` is not added in the model, and the routine terminates.

---

[4] Notice that STATA carries out the algorithm somewhat differently. The “null” model (the model with only the intercept) is fitted first (at step 0).

To perform a step-wise routine in STATA, we use the following command:

```
. sw  reg wgt hgt age age2,  pr(0.20) pe(0.10) forward
                      begin with empty model
p = 0.0013 <  0.1000  adding   hgt
p = 0.0565 <  0.1000  adding   age

  Source |       SS       df       MS                  Number of obs =      12
---------+------------------------------               F(  2,     9) =   15.95
   Model |  692.822607     2  346.411303               Prob > F      =  0.0011
Residual |  195.427393     9  21.7141548               R-squared     =  0.7800
---------+------------------------------               Adj R-squared =  0.7311
   Total |      888.25    11       80.75               Root MSE      =  4.6598

------------------------------------------------------------------------------
     wgt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     hgt |    .722038   .2608051      2.768   0.022        .1320559     1.31202
     age |   2.050126   .9372256      2.187   0.056       -.0700253    4.170278
   _cons |   6.553048   10.94483      0.599   0.564       -18.20587    31.31197
------------------------------------------------------------------------------
```

Note that STATA cannot accept $p_r \leq p_e$ so we have set p_r=0.20. Also, since STATA can perform a *backward* as well as a *forward* stepwise selection procedure, we have specified `forward` as the option for the type of stepwise selection that we wanted.

A (contrived) example of a "chunk-wise" selection method

If we want to make sure that a number of predictors will be added or subtracted together (say HGT and AGE), then all *F* tests are *multiple* *F* tests. The *F* test that forms the criterion of entry in the (forward selection) or removal (backward elimination) of the variables will be *F*(HGT,AGE) instead of *F*(HGT) or *F*(AGE) as before. In this case, *F*(HGT,AGE)=15.95 (the overall *F* test from Model 4).

All partial *F* tests are similarly *multiple* partial *F* tests. That is, instead of *F*(HGT|AGE2) or and *F*(AGE|AGE2) we must compute *F*(HGT,AGE|AGE2). We use *SSE*(AGE2)=366.32 (Model 3) so,

$$F(\text{HGT, AGE}|\text{ AGE2}) = \frac{[SSE(\text{AGE2}) - SSE(\text{HGT, AGE, AGE2})]/2}{MSE(\text{HGT, AGE, AGE2})} = \frac{[366.32 - 195.19]/2}{24.40} = 3.51$$

Since 3.51>$F_{2,8;0.90}$=3.11, we see that this "chunk" would be entered in all the model selection routines described above.

The STATA command structure that allows us to perform chunk-wise model selection consists of enclosing the variables in the group (chunk) in parentheses. For example, for a (forward) stepwise selection with (HGT and AGE) grouped together we have:

```
. sw reg wgt (hgt age) age2, pr(0.20) pe(0.10) forward
                      begin with empty model
p = 0.0011 <  0.1000  adding   hgt age

  Source |       SS       df       MS                  Number of obs =      12
---------+------------------------------               F(  2,     9) =   15.95
   Model |  692.822607     2  346.411303               Prob > F      =  0.0011
Residual |  195.427393     9  21.7141548               R-squared     =  0.7800
---------+------------------------------               Adj R-squared =  0.7311
   Total |      888.25    11       80.75               Root MSE      =  4.6598

------------------------------------------------------------------------------
     wgt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     hgt |    .722038   .2608051      2.768   0.022       .1320559     1.31202
     age |   2.050126   .9372256      2.187   0.056      -.0700253    4.170278
   _cons |   6.553048   10.94483      0.599   0.564      -18.20587    31.31197
------------------------------------------------------------------------------
```

Giving us the same model as before.