Variable (Model) Selection

Chapter 5 of SPRM discusses some modeling strategies and metrics. We have discussed some of the latter already.

- Thus far we have mostly worked with example problems where predictor variables were identified in advance. Often in modeling, we may end up having several candidate models, all of which pass the usual hypothesis tests. How do we pick the best model?
- Variable selection is the process of choosing a subset of all available predictors. Depending on the modeling goal, we might choose differently, but in any case, we are interested in a model we can interpret or justify with respect to the problem at hand
- We have already considered model comparison (among nested models) via the F-tests, R^2, adjusted R^2, and other less well-defined, more subjective criteria

Variable (Model) Selection

- There are additional measures, such as likelihood-based criteria (AIC, BIC) for non-nested models, but no metric is universally “best”.

- When there are many variables available, however, comparing individual models by any means can become overwhelming. If there are p predictor variables, there are 2^p possible models (i.e, if there are 10 variables, we would have over 1000 candidate models, not including those with any interaction effects.). This also assumes just one predictor form ($X, \sqrt{X}$, etc) for each predictor.

- Thus, the critical questions then becomes: how to choose good models in an intelligent and efficient way. This applies perhaps more to exploratory modeling but formalizing model selection in all problems is useful, lest we appear to be ‘data-dredging’ or ‘fishing’ for results

Practical Model vs. the Ideal

- **To set criteria for models relative to one another, we need to consider the ideal world where there is a 'correct' model.** What are consequences of including unnecessary variables or excluding necessary variables relative to this model?

  For example, assume we have the most general model as follows (has q candidate predictors)

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq} + \epsilon_i$$

- **For this model, one of two conditions might hold (indicating the correct model)**
  1. All predictors have non-zero β - all are predictors
  2. Model should have $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ non-zero but $\beta_{p+1}, \ldots \beta_q$ are zero, or these associated Xs are not needed

Model vs. Ideal: Consequences

- **We fit one model or the other, not knowing the 'true' state of nature. Then, what are consequences of**

  A. **Instead of using all q predictors, we use only use $p < q$ predictors, and fit:**

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

- If we fit a reduced model (omitting some X), we actually decrease variance on the β coefficients remaining in the model, but the estimates of β can be said to be biased if model (1) above is the correct model. Recall bias here means $E(\hat{\beta}) \neq \beta$ for the predictors in the model

- **The predictions $\hat{y}$ are also biased.** In fact, for the model, the Mean Square Error (MSE) $=$ bias^2 $+$ variance. Bias is generally considered bad but we may have good reason to tolerate it a bit in some more advanced methods. Note that bias is unobservable practically.

- **Why are estimates said to be biased? A:** There is still ''signal' in the residuals that would have been explained by omitted Xs

Model vs. Ideal: Consequences

- **What are consequences if**

  B. **We include all q predictors**

- **If we fit the full model, when we in fact should omit some Xs, we increase variance for all β coefficients in the model,** that is, estimating the extra $\beta_{p+1}, \ldots, \beta_q$ adds (unnecesary) variance. We also increase variance in prediction.

- **In practice bias is always unknown** - we don't know the true state of nature, and cannot parse the MSE into bias vs. variance. We might tolerate bias (through omitting Xs) to have reduced variance on β estimates and $\hat{y}$. In many situations, there will still be smaller MSE, due to the trade-offs that occur

\- Note: these consequences don't hold if X are orthogonal (rarely met, but this is why low correlation among Xs is a good property)

Purposes of Modeling

SPRM Section 5.3

- **An overarching consideration in model selection relates the the purpose and intended use. This might be:**
  - **Descriptive, Exploratory** (with respect to understanding relationships):
    Here, we use the model to search for fundamental relationships. Typically start simply with only a few essential variables, and then choose variables and combinations of variables to build forward. Consideration of why variables should be in the model is required.
  - **Predictive:**
    The focus is here on the high predictive ability of the model. We want predictions to be realistic and close to the sample data. Less consideration is given to which and how many variables are required, included, etc - 'black-box' modeling to some extent.

- **Explanatory:**
  The goal here is to describe the process in a realistic and interpretable way. Lots of thinking required about which variables are important to have in the model. Parsimony is generally sought (smallest model that is complete). Confounding, effect modification must be thoroughly addressed

- **These purposes are not mutually exclusive, and ideally we'd like a little bit of all of these properties in our modeling strategy**

Variable Selection Procedures

- Given the large number of candidate models in many case, we may wish to specify a variable selection strategy *a priori* .
- Thinking through the problem carefully and specifying how to add/remove specific variables is itself a strategy (and one that should be employed!), but with many predictors this is not always clear
- Thus, we may want to use one of the following 'objective' approaches (SPRM Section 5.5)

Variable Selection Procedures

A. **Forward Selection:**

1. Begin with null or empty model (no predictors), add predictor with highest simple correlation with Y. A significance level for entry is established here and applied throughout
2. Add in the predictor that has the highest partial correlation with Y after adjusting for the X variable added above.
3. With new model, return to Step 2, looking at the remaining predictors via the same criterion. Enter those meeting the significance criterion.

Variable Selection Procedures

B. **Backward Selection:**

1. Begin with all predictors in the model. Remove weakest one (smallest t statistic) with significance level for removal established at this step and applied throughout

2. Re-assess remaining predictors and remove according to criterion above.

3. Stop when there are no more predictor variables to remove

Variable Selection Procedures

C. **Stepwise Selection: (Foreword (FW) version)**

1. Begin with null model, add predictor as in forward selection, with significance level specified
2. Add in the next predictor as in forward selection. At this stage consider omitting first predictor according to criteria for backward stepwise procedure. A separate significance level may be used for removing variables versus adding.
3. Proceed with adding and removing variables as above
4. Stop when no more adding/removing criteria are satisfied

Variable Selection Procedures

– **A Common (manual) approach (- typical univariable to multivariable analysis strategy)**

1. Examine variables one at a time in relation to response, choose those meeting significance level specified
2. Put all in multivariable model, assess significance at typical criterion
3. Proceed with removing variables that are now nonsignificant
4. Stop when no more nonsignificant predictors

This approach is thought of as a way of rigorously screening important predictors (meeting univariate and multivariable model significance criteria required). Perhaps not the best strategy . . . unless approached carefully

Variable Selection: Missing Values for Predictors

An important related issue is incompleteness of data for predictors, a common feature in observational data and sometimes even in carefully controlled experiments.

- To appropriately contrast models, one would need to work with the intersection of X variables that have non-missing values. This is naturally implied in backward selection at the outset.
- If the analysis cohort is not fixed at this or some value, then the number of observations/cases n will change over the modeling process. Predictor variable effects may change solely due to being based on different sets of observations (i.e. datasets).
- On the other hand, using the intersection of complete data discards information, decreases statistical power, and can lead to biased estimates
- *Missing data* techniques are needed - imputation methods, etc

Variable Selection: Example

From C&H Text Table 3.3, Supervisor performance data: The data consists of six candidate predictors in relation to an overall performance measure for supervisors.

```
 X1  Handling of employee complaints
 X2  Allowance of special privileges
 X3  Opportunity to learn new things
 X4  Raises based on performance
 X5  Criticism of poor performance
 X6  Rate of advancing to better jobs
```

```
. corr x1-x6
(obs=30)
         |       x1       x2       x3       x4       x5       x6
---------+------------------------------------------------------
      x1 |   1.0000
      x2 |   0.5583   1.0000
      x3 |   0.5967   0.4933   1.0000
      x4 |   0.6692   0.4455   0.6403   1.0000
      x5 |   0.1877   0.1472   0.1160   0.3769   1.0000
      x6 |   0.2246   0.3433   0.5316   0.5742   0.2833   1.0000
```

- There is moderately high correlation among some predictors (it is a good idea to check this before applying any automated modeling procedures, due to problems introduced by multicollinearity)

```
. reg  y x1 x2 x3 x4 x5 x6

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  6,    23) =   10.50
       Model |  3147.96634     6  524.661057           Prob > F      =  0.0000
    Residual |  1149.00032    23  49.9565359           R-squared     =  0.7326
-------------+------------------------------           Adj R-squared =  0.6628
       Total |  4296.96667    29  148.171264           Root MSE      =   7.068

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .6131876   .1609831     3.81   0.001     .2801687    .9462066
          x2 |  -.0730501   .1357247    -0.54   0.596    -.3538181    .2077178
          x3 |   .3203321   .1685203     1.90   0.070    -.0282787     .668943
          x4 |   .0817321   .2214777     0.37   0.715    -.3764293    .5398936
          x5 |   .0383814   .1469954     0.26   0.796    -.2657018    .3424647
          x6 |  -.2170567   .1782095    -1.22   0.236    -.5857111    .1515977
       _cons |   10.78708   11.58926     0.93   0.362    -13.18713    34.76128
```

```
--------------------------------------------------------------------------------

.* check collinearity
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
          x4 |      3.08    0.324862
          x1 |      2.67    0.374945
          x3 |      2.27    0.440326
          x6 |      1.95    0.512403
          x2 |      1.60    0.624652
          x5 |      1.23    0.814260
-------------+----------------------
    Mean VIF |      2.13
```

- Some significant predictors, fairly high R^2, collinearity not a problem (next page)

- Assuming there are not other model assumption problems, we can proceed to determine which sub-model might be best

Variable Selection: Example
checking collinearity?

- Collinearity relates to the degree of correlation between predictors. if two predictors are highly correlated, one will ”stand in’ for the other. This can cause confusion in model selection strategies, and even misleading results.
- The Variance Inflation Factor (VIF) is defined as follows:

$$VIF_j = \frac{1}{1 - R_j^2}, \quad j = 1, \ldots, p$$

- Here, R_j^2 is the r-squared value for a model predicting covariate X_j with all the other predictors. The higher the value is, the larger the inflation factor becomes. Values over 10 may indicate too much correlation

Variable Selection: Example

- We can always fit all $2^6 = 64$ models, (not recommended).
- With some theory/context to guide us, we might follow a specific strategy. For example, starting with the above model, retaining specific factors by design, and sequentially removing others, then inspecting residuals, etc. This would be tractable here, but in models with larger number of predictors, still may be a lengthy process.
- Alternatively, we can proceed with the various automated variable selection procedures. These always should be used with CAUTION.
- The different variable selection procedures (forward, backward, stepwise) are typically implemented via a single module in computer packages, as forward and backward procedures can be thought of as variations on the stepwise approach

Variable Selection: Example

- The help file for Stata's stepwise procedure, which can be used with numerous modeling approaches (not just linear regression):

```
. help sw
-----------------------------------------------------------------------------------------

Stepwise estimation

   sw cmd regress [var1 ...] [weight] [if exp] [in range], { pr(#) | pe(#) | pr(#)
                    pe(#) } [ forward lr hier lockterm1 cmd_options ]

predict after sw behaves the same as predict after the particular estimation command;
 see help for the particular estimation command for details.

sw performs stepwise estimation, the flavor of which is determined by the options:

        pr(#)                 backward selection
        pr(#) hier            backward hierarchical selection
        pr(#) pe(#)           backward stepwise

        pe(#)                 forward selection
        pe(#) hier            forward hierarchical selection
```

```
       pr(#) pe(#) forward   forward stepwise

pr(#) specifies the significance level for removal from the model; terms with
p>=pr() are eligible for removal.

pe(#) specifies the significance level for addition to the model; terms with
p<pe() are eligible for addition.

forward specifies the forward-stepwise method when both pr() and pe() are also
specified.  Specifying both pr() and pe() without forward results in backward
stepwise.  Note that specifying only pr() results in backward selection and
specifying only pe() results in forward selection.
-------------------------------------------------------------------------------------------
```

- Our stepwise procedure would be called ‘forward stepwise’ here

Variable Selection: Example

Example – forward selection:

The probability to enter option, pe, was set to .99 (*only* for illustrative purposes, we typically set at more conventional level) so that all predictor variables will enter based on statistical significance according to criteria described earlier

```
. sw regress y x1 x2 x3 x4 x5 x6, pe(.99)
                      begin with empty model
p = 0.0000 <=  0.9900  adding   x1
p = 0.1278 <=  0.9900  adding   x3
p = 0.2082 <=  0.9900  adding   x6
p = 0.5616 <=  0.9900  adding   x2
p = 0.6426 <=  0.9900  adding   x4
p = 0.7963 <=  0.9900  adding   x5

  Source |       SS       df       MS                Number of obs =      30
---------+------------------------------             F(  6,    23) =   10.50
   Model |  3147.96634     6  524.661057             Prob > F      =  0.0000
Residual |  1149.00032    23  49.9565359             R-squared     =  0.7326
---------+------------------------------             Adj R-squared =  0.6628
```

```
   Total |  4296.96667    29  148.171264               Root MSE      =   7.068
------------------------------------------------------------------------------
       y |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      x1 |   .6131876   .1609831      3.809   0.001       .2801687    .9462065
      x3 |   .3203321   .1685203      1.901   0.070      -.0282787    .6689429
      x6 |  -.2170567   .1782095     -1.218   0.236       -.585711    .1515977
      x2 |  -.0730501   .1357247     -0.538   0.596       -.353818    .2077178
      x4 |   .0817321   .2214777      0.369   0.715      -.3764293    .5398936
      x5 |   .0383814   .1469954      0.261   0.796      -.2657018    .3424647
   _cons |   10.78708   11.58926      0.931   0.362      -13.18713    34.76128
------------------------------------------------------------------------------
```

At stage 1, Stata fits all models with just one variable, and picks the model whose variable has the smallest p-value.

If that p-value is smaller than pe (in this case yes) then it fits the 2-variable models, and chooses the model which has the smallest p-value for the second variable, as long as that p-value is $<$ pe.

The procedure is repeated until adding any other remaining variables would have the added variable's p-value being $>$ pe.

Variable Selection: Analysis Example

- Following C&H, one might use variable entry criteria of $t_{.05,n-p}$ or $|t| > 1$. Using the first one (corresponding to $|t|$ of about 1.5)

- **forward selection**

```
. sw regress y x1 x2 x3 x4 x5 x6, pe(.15)
                      begin with empty model
p = 0.0000 <  0.1500  adding   x1
p = 0.1278 <  0.1500  adding   x3

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  2,    27) =   32.74
       Model |   3042.3177     2  1521.15885           Prob > F      =  0.0000
    Residual |  1254.64897    27  46.4684804           R-squared     =  0.7080
-------------+------------------------------           Adj R-squared =  0.6864
       Total |  4296.96667    29  148.171264           Root MSE      =  6.8168

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .6435176   .1184774     5.43   0.000      .400422    .8866132
          x3 |   .2111918   .1344037     1.57   0.128    -.0645818    .4869655
       _cons |    9.87088   7.061224     1.40   0.174    -4.617554    24.35931
------------------------------------------------------------------------------
```

Variable Selection: Example

- **backward selection** Here, we set the prob(remove) option, `pr`, was set to .33 to correspond to a $|t|$-statistic of 1.0.

```
. sw regress y x1 x2 x3 x4 x5 x6, pr(.33)
                      begin with full model
p = 0.7963 >= 0.3300  removing x5
p = 0.6426 >= 0.3300  removing x4
p = 0.5616 >= 0.3300  removing x2

  Source |       SS       df       MS                  Number of obs =      30
---------+------------------------------               F(  3,    26) =   22.92
   Model |  3117.85753     3  1039.28584               Prob > F      =  0.0000
Residual |  1179.10914    26  45.3503515               R-squared     =  0.7256
---------+------------------------------               Adj R-squared =  0.6939
   Total |  4296.96667    29  148.171264               Root MSE      =  6.7343
------------------------------------------------------------------------------
       y |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      x1 |   .6227297   .1181464      5.271   0.000       .3798763    .8655832
      x6 |  -.1869508   .1448537     -1.291   0.208      -.4847019    .1108003
      x3 |    .312387   .1541997      2.026   0.053      -.0045751    .6293491
   _cons |   13.57774     7.5439      1.800   0.084      -1.928967    29.08445
-------------------------------------------------------------------------------
```

Variable Selection: Example

Comments:

- All models are about the same (forward, backward, and just choosing on our own). Note that $X1$ and $X3$ would always be included, but coefficients, p-values, etc a bit different in each.

- We should pay attention to β coefficients to note inconsistencies or nonsensical results. Here, these are not radically different model to model (should not be, as vif was low).

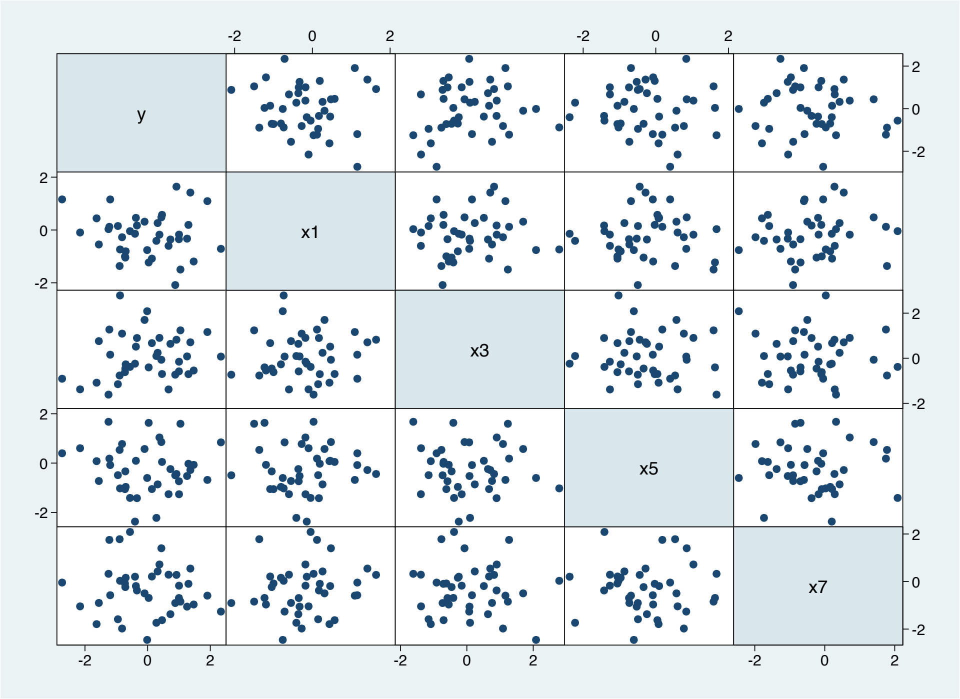
- **Caution:** `sw` should not be done mechanically, without careful consideration of results. We must always apply context and common sense to this approach. The most frequent criticism of automated procedures relates to the fact that they will arrive at answers, whether correct or not. To illustrate:

Variable Selection: A Simulated Data Example

- **A simple simulation experiment:** - Generate a moderately small dataset of completely random response Y and corresponding large number of predictors X. Run different stepwise procedures.

```
.  set obs 40

. set seed 44132254
. gen y=invnorm(uniform())

. gen x1=invnorm(uniform())

. for num 2/15: gen xX=invnorm(uniform())

->  gen x2=invnorm(uniform())
->  gen x3=invnorm(uniform())
->  gen x4=invnorm(uniform())
->  gen x5=invnorm(uniform())
->  gen x6=invnorm(uniform())
->  gen x7=invnorm(uniform())
. . .-
>  gen x14=invnorm(uniform())
-> gen x15=invnorm(uniform())
```

```
. graph matrix y x1 x3 x5 x7
```



Nothing going on here in terms of Y vs. X relationship. From full model and selection procedures, we obtain the following:

- ordinary approach, test all initially

```
. regress y x1-x15

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F( 15,    24) =    1.47
       Model |  23.8400721    15  1.58933814           Prob > F      =  0.1958
    Residual |   26.018029    24  1.08408454           R-squared     =  0.4782
-------------+------------------------------           Adj R-squared =  0.1520
       Total |  49.8581011    39  1.27841285           Root MSE      =  1.0412

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   -.169245   .2573697    -0.66   0.517      -.70043    .3619399
          x2 |   .0157974   .2589846     0.06   0.952    -.5187206    .5503154
          x3 |  -.1640166   .2333674    -0.70   0.489    -.6456633    .3176301
          x4 |   .5187105   .2476059     2.09   0.047     .0076771    1.029744
          x5 |  -.0623123   .2048912    -0.30   0.764     -.485187    .3605624
          x6 |  -.0997396   .1762514    -0.57   0.577    -.4635046    .2640253
          x7 |  -.4994487    .350675    -1.42   0.167    -1.223206    .2243088
          x8 |   .0897644   .2014728     0.45   0.660    -.3260551    .5055839
          x9 |  -.1801775   .1904875    -0.95   0.354    -.5733245    .2129695
         x10 |   .4787773   .2071367     2.31   0.030     .0512682    .9062863
         x11 |  -.4467917   .2575459    -1.73   0.096    -.9783404     .084757
```

```
         x12 |  -.2609168   .2570423    -1.02   0.320    -.7914259    .2695924
         x13 |  -.2458519   .1920375    -1.28   0.213    -.6421979    .1504942
         x14 |  -.3334932   .3279154    -1.02   0.319    -1.010277    .3432909
         x15 |  -.5789872   .2251485    -2.57   0.017    -1.043671   -.1143034
       _cons |  -.1613593   .2109941    -0.76   0.452    -.5968297    .2741112
------------------------------------------------------------------------------
```

- NS model overall as expected, some βs are ‘significant’, but considering number of parameters/tests, none would pass (for example, if we used $.05/15 = .0033$ for significance). **Now run backward stepwise and forward stepwise procedures**

```
.* BACKWARD procedure

. sw reg  y x1-x15, pr(.2)
                      begin with full model
p = 0.9519 >= 0.2000  removing x2
p = 0.7241 >= 0.2000  removing x5
p = 0.6650 >= 0.2000  removing x8
p = 0.5513 >= 0.2000  removing x6
p = 0.6059 >= 0.2000  removing x1
p = 0.3503 >= 0.2000  removing x3
p = 0.3605 >= 0.2000  removing x12
```

```
p = 0.3166 >= 0.2000  removing x14
p = 0.3473 >= 0.2000  removing x13

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  6,    33) =    3.52
       Model |  19.4624581     6  3.24374302           Prob > F      =  0.0084
    Residual |   30.395643    33  .921080091           R-squared     =  0.3904
-------------+------------------------------           Adj R-squared =  0.2795
       Total |  49.8581011    39  1.27841285           Root MSE      =  .95973

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         x11 |  -.3043337   .2139114    -1.42   0.164    -.7395397    .1308723
         x15 |  -.5838627   .1856534    -3.14   0.004    -.9615774   -.206148
         x10 |   .3292684    .149957     2.20   0.035     .0241785    .6343583
          x4 |   .4684389   .1832929     2.56   0.015     .0955268    .8413511
          x7 |  -.2236951   .1704186    -1.31   0.198    -.5704144    .1230243
          x9 |  -.2142553   .1600246    -1.34   0.190    -.5398277    .1113171
       _cons |  -.2397916   .1655045    -1.45   0.157     -.576513    .0969297
------------------------------------------------------------------------------
```

```
.* FORWARD procedure

. sw reg  y x1-x15, pe(.2)
                      begin with empty model
p = 0.0476 <  0.2000  adding   x10
p = 0.1050 <  0.2000  adding   x15
p = 0.0083 <  0.2000  adding   x4

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  3,    36) =    5.42
       Model |  15.519939     3  5.17331299            Prob > F      =  0.0035
    Residual | 34.3381621    36  .953837837            R-squared     =  0.3113
-------------+------------------------------           Adj R-squared =  0.2539
       Total | 49.8581011    39  1.27841285            Root MSE      =  .97665

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         x10 |   .3516723   .1512455     2.33   0.026     .0449321    .6584125
         x15 |  -.5013255   .1781587    -2.81   0.008    -.8626481   -.1400028
          x4 |   .4964532   .1776622     2.79   0.008     .1361375    .8567688
       _cons |  -.1433977   .1570761    -0.91   0.367    -.4619629    .1751674
------------------------------------------------------------------------------
```

- Using stepwise procedures with the same criteria, two models contrary to each other, with overall F-test significance, arise from this data. ???
- Models are still reasonably consistent (all identify $X4, X10, X15$), and, adjusting for multiple comparisons, there would effectively be no model.
- But, in real-life modeling, we have expectations of relationships, do not always rigorously adjust for multiple comparisons and number of models considered, and so we might accept findings as real
- Also, illustrates why external validation (checking model with data not used to build it) is highly valuable

Modeling with Collinearity Present

Here is a correlation matrix for blood pressure (Y) in relation to six predictors: age, body surface area, weight, duration of hypertension, pulse, and a stress measure

```
. corr bp weight bsa dur pulse stress age
(obs=20)

             |       bp   weight      bsa      dur    pulse   stress      age
-------------+---------------------------------------------------------------
          bp |   1.0000
      weight |   0.9501   1.0000
         bsa |   0.8659   0.8753   1.0000
         dur |   0.2928   0.2006   0.1305   1.0000
       pulse |   0.7214   0.6593   0.4648   0.4015   1.0000
      stress |   0.1639   0.0344   0.0184   0.3116   0.5063   1.0000
         age |   0.6591   0.4073   0.3785   0.3438   0.6188   0.3682   1.0000
```

The correlation between weight and BSA is very high (0.875)

Modeling with Collinearity Present

Checking the VIF:

```
. quietly reg bp weight bsa dur pulse stress age
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
      weight |      8.42    0.118807
         bsa |      5.33    0.187661
       pulse |      4.41    0.226574
      stress |      1.83    0.545005
         age |      1.76    0.567277
         dur |      1.24    0.808205
-------------+----------------------
    Mean VIF |      3.83
```

Weight and BSA have fairly high VIF - are linear functions of the other variables (with each other being the strongest predictors most likely)

Modeling with Collinearity Present

Run the stepwise model:

```
. sw reg bp weight bsa dur pulse stress age, pe(.05)
p = 0.0000 <  0.0500  adding   weight
p = 0.0000 <  0.0500  adding   age
p = 0.0078 <  0.0500  adding   bsa
      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(3, 16)        =    971.93
       Model |  556.943853         3  185.647951   Prob > F        =    0.0000
    Residual |  3.05614729        16  .191009206   R-squared       =    0.9945
-------------+----------------------------------   Adj R-squared   =    0.9935
       Total |         560        19  29.4736842   Root MSE        =    .43705
------------------------------------------------------------------------------
          bp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      weight |   .9058219   .0489895    18.49   0.000     .8019688    1.009675
         age |   .7016201   .0439595    15.96   0.000       .60843    .7948101
         bsa |   4.627399   1.521068     3.04   0.008      1.40288    7.851918
       _cons |  -13.66724   2.646638    -5.16   0.000    -19.27786   -8.056613
------------------------------------------------------------------------------
```

Both weight and BSA are retained, model fit is very good

Modeling with Collinearity Present

What about this model?

```
 reg bp weight   age
      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(2, 17)        =    978.25
       Model |  555.176061         2  277.58803    Prob > F        =    0.0000
    Residual |  4.82393934        17  .283761138   R-squared       =    0.9914
-------------+----------------------------------   Adj R-squared   =    0.9904
       Total |         560        19  29.4736842   Root MSE        =    .53269
------------------------------------------------------------------------------
          bp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      weight |   1.032961   .0311563    33.15   0.000     .9672269    1.098695
         age |   .7082517   .0535141    13.23   0.000      .595347    .8211565
       _cons |  -16.57936   3.007463    -5.51   0.000    -22.92456   -10.23417
------------------------------------------------------------------------------
```

Dropping BSA variable results in very small loss in R^2. Model dropping weight and keeping BSA variable is similar, not quite as good ($R^2 = 0.88$). It is not clear that both should be in the model, depends on questions of interest

Some Additional Criteria for Evaluating Models - non-nested models

SPRM 5.4

- We have already discussed the Mean Squared Error as the 'variance' of the whole model, $MSE = \frac{SSE}{n-p-1}$. Among two models, materially smaller MSE would be preferred in general
- The R^2 and adjusted version R^2_{adj}: Latter is useful for *non-nested* models
- Two measures that can be used *whether or not models are nested* relate to the 'information' in the model. The Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) provide measures that balance information extracted from the data (fit) and number of parameters

Some Additional Criteria for Evaluating Models

AIC:

$$AIC = n \log_e(SSE_p/n) + 2p$$

where p is the $total$ number of parameters (intercept included)

BIC:

$$BIC = n \log_e(SSE_p/n) + p \log_e(n)$$

- penalizes more heavily for having lots of parameters p relative to observations n

- These quantities for a given model are not particularly interpretable, but rather in contrasting two models, are useful. Among two candidates, the model with smaller AIC or BIC would be preferred. Models need not be nested, and these quantities take into account number of parameters

Example: Structured vs. Automated Approach

Surgical Unit data: The variables are

- x1: blood clotting measure
- x2: prognostic index
- x3: enzyme measure
- x4: liver function measure
- x5: age
- x6: gender
- y: survival - outcome
- lny: log survival - outcome

- Predict ln(survival time). It is suspected that there may be an interaction effect between $x2$ and $x3$ and thus we generate an interaction term of $x2 \times x3$ for consideration.

- **A systematic (manual) modeling approach,** followed by stepwise procedure

```
. gen x2x3=x2*x3
. reg lny x1 x2 x3 x4 x5 x6 x2x3

      Source |       SS       df       MS              Number of obs =      54
-------------+------------------------------           F(  7,    46) =   22.46
       Model |  9.90882875     7  1.41554696           Prob > F      =  0.0000
    Residual |  2.89889608    46   .06301948           R-squared     =  0.7737
-------------+------------------------------           Adj R-squared =  0.7392
       Total |  12.8077248    53  .241655186           Root MSE      =  .25104
------------------------------------------------------------------------------
         lny |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .0950729   .0296256     3.21   0.002     .0354397    .1547061
          x2 |   .0150846   .0069173     2.18   0.034     .0011608    .0290084
          x3 |   .0177798   .0052646     3.38   0.001     .0071828    .0283768
          x4 |  -.0012851   .0560579    -0.02   0.982    -.1141239    .1115537
          x5 |  -.0049206   .0032525    -1.51   0.137    -.0114675    .0016263
          x6 |   .0627443   .0735302     0.85   0.398    -.0852642    .2107528
        x2x3 |  -.0000244   .0000769    -0.32   0.752    -.0001792    .0001303
       _cons |   3.903096   .5111992     7.64   0.000     2.874105    4.932087
------------------------------------------------------------------------------
.* store some results on this model for later
```

```
. est store A
```

The above is our full model, stored it as Model A. Now we look at some other models. Drop least significant predictors other than interaction

```
. reg lny x1 x2 x3 x5 x2x3

      Source |       SS       df       MS              Number of obs =      54
-------------+------------------------------           F(  5,    48) =   32.11
       Model |  9.85956218     5  1.97191244           Prob > F      =  0.0000
    Residual |  2.94816266    48  .061420055           R-squared     =  0.7698
-------------+------------------------------           Adj R-squared =  0.7458
       Total |  12.8077248    53  .241655186           Root MSE      =  .24783

------------------------------------------------------------------------------
         lny |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .0959735   .0217194     4.42   0.000     .0523037    .1396434
          x2 |    .016032   .0067313     2.38   0.021     .0024979    .0295661
          x3 |   .0185256   .0050542     3.67   0.001     .0083634    .0286878
          x5 |  -.0048902   .0030783    -1.59   0.119    -.0110795     .001299
        x2x3 |  -.0000331   .0000749    -0.44   0.660    -.0001838    .0001175
```

```
       _cons |   3.846634   .4977812     7.73   0.000     2.845777     4.84749
------------------------------------------------------------------------------

. set store B

 .* drop interaction and other ~null variable (x5)

. reg lny x1 x2 x3

      Source |       SS       df       MS              Number of obs =      54
-------------+------------------------------           F(  3,    50) =   52.00
       Model |  9.69918607     3  3.23306202           Prob > F      =  0.0000
    Residual |  3.10853876    50  .062170775           R-squared     =  0.7573
-------------+------------------------------           Adj R-squared =  0.7427
       Total |  12.8077248    53  .241655186           Root MSE      =  .24934

------------------------------------------------------------------------------
         lny |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .0954583   .0216921     4.40   0.000     .0518884    .1390283
          x2 |     .01334   .0020347     6.56   0.000     .0092532    .0174268
          x3 |   .0164517   .0016299    10.09   0.000     .0131779    .0197254
       _cons |   3.766176   .2267583    16.61   0.000     3.310718    4.221633
------------------------------------------------------------------------------

. est store C
.
```

```
.
.
.* FOR ILLUSTRATIVE PURPOSES, go too far, drop an important variable

. reg lny x1 x2

      Source |       SS       df       MS              Number of obs =      54
-------------+------------------------------           F(  2,    51) =    9.09
       Model |  3.36505402     2  1.68252701           Prob > F      =  0.0004
    Residual |  9.44267081    51  .185150408           R-squared     =  0.2627
-------------+------------------------------           Adj R-squared =  0.2338
       Total |  12.8077248    53  .241655186           Root MSE      =  .43029

------------------------------------------------------------------------------
         lny |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .0630202   .0370214     1.70   0.095    -.0113034    .1373437
          x2 |    .013129   .0035111     3.74   0.000     .0060801    .0201778
       _cons |    5.23573   .3000092    17.45   0.000     4.633437    5.838024
------------------------------------------------------------------------------
.
. * store results
. est store D
.
.
. *
```

```
.* fit some model not nested in models B through D
.
.  reg lny  x2 x3 x4

      Source |       SS       df       MS              Number of obs =      54
-------------+------------------------------           F(  3,    50) =   42.40
       Model |  9.19359439     3  3.06453146           Prob > F      =  0.0000
    Residual |  3.61413045    50  .072282609           R-squared     =  0.7178
-------------+------------------------------           Adj R-squared =  0.7009
       Total |  12.8077248    53  .241655186           Root MSE      =  .26885

------------------------------------------------------------------------------
         lny |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |   .0110093   .0024043     4.58   0.000     .0061801    .0158385
          x3 |   .0126091   .0019547     6.45   0.000     .0086831    .0165352
          x4 |   .1297686   .0417491     3.11   0.003     .0459131    .2136241
       _cons |    4.40582   .1989817    22.14   0.000     4.006154    4.805487
------------------------------------------------------------------------------

. est store E
```

- **Now contrast all models**

```
. display information-based measures
. est stats A B C D E
-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           A |     54   -37.77142    2.342978      8     11.31404    27.22592
           B |     54   -37.77142     1.88797      6     8.224059    20.15796
           C |     54   -37.77142    .4577638      4     7.084472    15.04041
           D |     54   -37.77142   -29.54156      3     65.08312    71.05007
           E |     54   -37.77142   -3.611096      4     15.22219    23.17813
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

We use minimum of AIC or BIC as our model selection criteria, in addition to other considerations. So, the best is model C.

$$lny_i = \beta_0 + \beta_1 x1 + \beta_2 x2 + \beta_3 x3$$

This model had best R^2 among 'smaller' models, best AIC, BIC

- **How about using stepwise regression `sw`?**

  We specify both `pr` and `pe`. Typically, `pr` should be greater than `pe`.

```
. sw reg lny x1 x2 x3 x4 x5 x6 x2x3, pe(0.05) pr(0.1) forward
                      begin with empty model
p = 0.0000 <  0.0500  adding   x2x3
p = 0.0026 <  0.0500  adding   x1
p = 0.0017 <  0.0500  adding   x3
p = 0.0300 <  0.0500  adding   x2
p = 0.7722 >= 0.1000  removing x2x3

      Source |       SS       df       MS              Number of obs =      54
-------------+------------------------------           F(  3,    50) =   52.00
       Model |  9.69918607     3  3.23306202           Prob > F      =  0.0000
    Residual |  3.10853876    50  .062170775           R-squared     =  0.7573
-------------+------------------------------           Adj R-squared =  0.7427
       Total |  12.8077248    53  .241655186           Root MSE      =  .24934

------------------------------------------------------------------------------
         lny |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |     .01334   .0020347     6.56   0.000     .0092532    .0174268
```

```
          x1 |   .0954583   .0216921     4.40   0.000     .0518884    .1390283
          x3 |   .0164517   .0016299    10.09   0.000     .0131779    .0197254
       _cons |   3.766176   .2267583    16.61   0.000     3.310718    4.221633
------------------------------------------------------------------------------
```

- **In this case, arrives at same 'best model'.** Note: interaction term entered before main effects (a generally undesirable property). Stepwise procedures provide means to constrain the form of model to some extent, preventing this anomoly

Summary: Variable Selection

- Earlier measures and new quantities introduced here (AIC, BIC) are useful, while context and 'domain knowledge' should guide modeling most
- Automated procedures can be useful but must be used cautiously. There is not universal agreement about approaches, although backward stepping procedures seem to be preferred
- Model selection is a large and evolving area of statistics, many more tools available