SEARCH STRATEGIES:
Stepwise regression
Guided regression

Stepwise regression

One commonly used procedure that I strongly recommend against using is stepwise regression, which can be done either forward or backward. Because this method is so commonly used, however, we need to know what it is.

Forward: Looks at all the candidates for predictor variables. It starts with the model that has constant only, then adds in first the X with the highest correlation with Y. At each step, it adds the variable that increases $R^2$ the most.

Backward: Looks at the model that includes all variables. It drops first the X whose deletion will cause the smallest drop in $R^2$.

When does the procedure stop adding or dropping? Usually we use a .05 level of significance. The computer tests each variable and sees if, if dropped (or added) the change in $R^2$ is significant. This is equivalent to an F-test with one degree of freedom in the numerator and n-K df in the denominator, where K-1 is the number of predictor variables in the larger model.

The .05 level of significance for an F-test with 1 df in the numerator is the equal to $t^2$ with the same degrees of freedom as in the denominator of the F test. Therefore, if we have a very large sample, the cut-off point for the F statistic would be $1.96^2 = 3.92$. An example of backward stepwise regression and then forward stepwise regression follows, using the dataset on income. Here I treated education as the original continuous (years of education) variable. I also included interaction terms for race and education (edb= ed*black and edh= ed*hisp) and for female and education.

New command
```
. sw regress income ed kids famsize female black hisp edb edh edfem, pr(.05)
```
Old command
```
. stepwise income ed kids famsize female black hisp edb edh edfem, backward fstay(3.92)
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 1606</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1.6626e+11</td>
<td>5</td>
<td>3.3251e+10</td>
<td>F(  5, 1600) = 123.96</td>
</tr>
<tr>
<td>Residual</td>
<td>4.2919e+11</td>
<td>1600</td>
<td>268244830</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>5.9545e+11</td>
<td>1605</td>
<td>370994885</td>
<td>R-square = 0.2792</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Adj R-square = 0.2770</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = 16378</td>
</tr>
</tbody>
</table>

| income | Coef.   | Std. Err. | t     | P>|t|    | [95% Conf. Interval] |
|--------|---------|-----------|-------|--------|---------------------|
| ed     | 2839.629| 144.9526  | 19.590| 0.000  | 2555.312            |
|        |         |           |       |        | 3123.946            |
| hisp   | 13444.42| 4312.403  | 3.118 | 0.002  | 4985.864            |
|        |         |           |       |        | 21902.97            |
| edb    | -444.5462| 114.8286  | -3.871| 0.000  | -669.7765           |
|        |         |           |       |        | -219.316            |
| edh    | -1632.396| 376.0756  | -4.341| 0.000  | -2370.049           |
|        |         |           |       |        | -894.7432           |
| edfem  | -1025.567| 63.29707  | -16.202| 0.000  | -1149.721           |
|        |         |           |       |        | -901.4136           |
| _cons  | -9377.727| 1897.624  | -4.942| 0.000  | -13099.82           |
|        |         |           |       |        | -5655.636           |
. stepwise income ed kids famsize female black hisp edb edh edfem, forward
   fstay(3.92)

Adding:  ed    F=    270.5
Adding:  edfem  F=    254.3
Adding:   edb   F=    13.65
Adding:   edh   F=     15.74
Adding:   hisp  F=       9.72

(stepwise)

Source        |       SS       | df       | MS                  Number of obs =  1606
---------------+----------------+----------+--------------------- F(  5,  1600) =  123.96
Model         |  1.6626e+11    | 5        | 3.3251e+10          Prob > F      =  0.0000
Residual      |  4.2919e+11   | 1600     | 268244830           R-square      =  0.2792
---------------+----------------+----------+--------------------- Adj R-square =  0.2770
Total         |  5.9545e+11   | 1605     | 370994885           Root MSE      =   16378

------------------------------------------------------------------
income         |      Coef.   | Std. Err. |       t     | P>|t|      | [95% Conf. Interval]
---------------+----------------+------------+------------+---------+---------------------------
ed          |  2839.629     | 144.9526   | 19.590     | 0.000   | 2555.312-3123.946
hisp       |  13444.42     | 4312.403   | 3.118      | 0.002   | 4985.864-21902.97
edb        | -444.5462     | 114.8286   | -3.871     | 0.000   | -669.7765-219.316
edh       | -1632.396     | 376.0756   | -4.341     | 0.000   | -2370.049-894.7432
edfem    | -1025.567     | 63.29707   |-16.202     | 0.000   | -1149.721-901.4136
_cons     | -9377.727     | 1897.624   | -4.942     | 0.000   | -13099.82-5655.636
------------------------------------------------------------------

In these examples, we ended up with the same model, but frequently the forward and backward procedures do not
give you the same model. They do if all predictor variables are unrelated - i.e. if the correlations are all zero. This
is, however, not the rule in social science investigations - we deal with correlated predictors all the time.

A variant on backward stepwise regression is to reconsider at each stage all variables that were dropped in previous
steps. We can request this type of regression by adding both the fenter and fstay options to our regress command; the
output is unchanged in our case, so is not shown.

. stepwise income ed kids famsize female black hisp edb edh edfem, backward
   fstay(3.92) fenter(3.92)

All possible subsets

Another approach says, try all possible subsets - all models in which there is one predictor, all with 2 predictors,
etc. and choose the one with the highest R² value. Here there is a good chance that small differences in our data
will lead to different models - that by chance alone particular combinations will work. It, like stepwise regression,
ignores all theory and previous results about the issue we are investigating.

Guided regression

The approach that I recommend may be summarized as follows:

1. Look at your data
   look at distributions - one-way and two-way
   think about transformations
   detect unusual values (we'll talk more next time about ways of doing this
   think about whether to drop them and what new information they offer

2. Think about appropriate theoretical models
3. Consider whether predictor variables are so related that you want to worry about multicollinearity

4. **BE SELECTIVE** in your analyses
   Remember, even though computer analysis is fast and relatively cheap, the brain cannot make use of too much data.

5. **USE GUIDED REGRESSION**

The notion of guided regression is best expressed by John Tukey. His advice:

1. When you lots of candidates for inclusion in your regression model, divide them into three groups:
   - **Key variables:** 0-6 variables you have strong theoretical reasons for wanting to include in all regressions
   - **Promising variables:** up to 12 variables that deserve somewhat special attention
   - **The haystack:** motley collection that deserves limited attention

2. Carry out your analysis using only the key variables. Then calculate residuals for Y and all other predictor variables

   e.g. Key variables are $X_1$, $X_2$ and $X_3$
   Other variables are $X_4, \ldots, X_{12}$

   **Residuals:**
   
   \[
   Y' = Y - b_0 - b_1 X_1 - b_2 X_2 - b_3 X_3
   \]
   
   \[
   X_{i4}' = X_{i4} - b_{40} - b_{41} X_1 - b_{42} X_2 - b_{43} X_3
   \]
   
   .
   
   \[
   X_{i12}' = X_{i12} - b_{12,0} - b_{12,1} X_1 - b_{12,2} X_2 - b_{12,3} X_3
   \]

3. Using these residuals, consider the all possible subsets of the promising variables.

   e.g. $X_4$, $X_5$, $X_6$, $X_7$ are promising

   Then calculate all possible models relating $Y'$ to the residuals of these variables:

   \[
   Y'_1 = b_0 + b_4 X_4'
   \]
   \[
   Y'_2 = b_0 + b_5 X_5'
   \]
   \[
   Y'_3 = b_0 + b_6 X_6'
   \]
   \[
   Y'_4 = b_0 + b_7 X_7'
   \]
   \[
   Y'_5 = b_0 + b_4 X_4' + b_5 X_5'
   \]
   \[
   Y'_6 = b_0 + b_4 X_4' + b_6 X_6'
   \]
   
   etc.

Check the $R^2$ values or use F-tests to see which of these variables should remain in the model.

Then calculate new residuals for Y and for the haystack variables.

4. Use stepwise regression, forward or backward or combination, to sift through the haystack.

5. Finally, recalculate a full model using

   key variables
successful promising variables
successful haystack variables

ASSUMPTIONS OF ORDINARY LEAST SQUARES (OLS) REGRESSION

Basic assumptions:

1. The X values are fixed - i.e. we usually don’t treat them as a sample.
2. Errors have zero mean - i.e. $E[\epsilon_i] = 0$

Assumptions 1 and 2 are sufficient to ensure that our estimates of the $\beta_k$ are unbiased. They are not sufficient to prove that the OLS estimates are more efficient than other possible unbiased estimators.

UNBIASED: $E[\beta_k] = \beta_k$
EFFICIENT: $\beta_k$ is more efficient than another unbiased estimator, $\alpha_k$, if $\beta_k$ has smaller variance.

3. Errors have the same constant variance (homoscedasticity)

\[ \text{Var}[\epsilon_i] = \sigma^2 \]
for all $i$.

4. Errors are uncorrelated one with another

\[ \text{Cov}[\epsilon_i, \epsilon_j] = 0 \quad \text{for all } i \neq j \]

When we add assumptions 3 and 4 to 1 and 2, it can be proved that the standard errors of the coefficients are also unbiased, and OLS is more efficient than any other linear unbiased estimator - the estimates are BLUE.

If we also add the assumption that the predictors, the X variables, are not correlated with the error terms, i.e.

\[ \text{Cov}[X_{ik}, \epsilon_i] = 0 \quad \text{for all } i,k \]

then the estimates of the coefficients and their standard errors are consistent, which is defined as follows:

$\beta_k$ is a consistent estimator of $\beta_k$ if the probability that they are very close approaches 1 as the sample size increases toward an infinite number.

5. Errors are normally distributed

\[ \epsilon_i \sim N(0, \sigma^2) \quad \text{for all } i \]

This assumption justifies our use of F tests and t tests, especially with small samples. It also leads to the proof that OLS is more efficient than any other unbiased estimator, linear or not.

These assumptions are summarized in the statement:

We assume the linear model is correct, with normal, independent, and identically distributed (normal i.i.d.) errors.

Much of what we do is to try to check whether these assumptions are met - and what to do when they are not! If these assumptions are violated, other types of estimators (e.g. robust estimators, ridge regression estimators) may be better - more efficient.

HOW DO WE INVESTIGATE WHETHER ASSUMPTIONS ARE VIOLATED?

1. Look at correlation matrices and scatterplot matrices to detect non-linearity and heteroscedasticity
2. Look at plots of residuals vs predicted Y values, again to look for non-linearity and heteroscedasticity
   * can use band regression - later

3. Look for autocorrelation - correlation among the values of the variables for different cases - next time
   easiest example is temperature - in a hot summer, it's likely that successive months will be hot

4. Do the tests for normality

5. See which cases are especially influential - next handout