Regression Model Building -- we've talked about two kinds..

Descriptive Modeling

- explication of the relationships between the set of predictors and the criterion variable (90-95%)
- completeness is the goal -- often requiring the combination of information from simple correlations and various multiple regression models

Predictive Modeling

- are going to compute y' values to make decisions about people (5-10%)
- efficiency is the goal, want the "best model" with the "fewest predictors"

Interpretation of simple correlations, of full models, and of comparisons of nested and non-nested models are commonly combined to yield ***descriptive modeling*** -- remember, the interest is in completely explicating the patterns of relationship between the criterion variable and the predictors. Theory often plays a large part, directing the variables we choose and the models we compare.

When interested in ***predictive modeling***, we want 1 model that we will use to compute y' values for future individuals. Theory can help, cost-analysis is often important, collinearity is to be avoided (because it reduces efficiency).

The various "automated regression procedures" were designed for predictive model building (not for descriptive modeling).

The four most commonly used "automated" procedures are…

Forward Inclusion -- start with the best predictor and add predictors to get a "best model"

Backward Deletion -- start with a full model and delete predictors to get a "best model"

Forward Stepwise Inclusions -- a combination of the first two

All-subsets Regression -- literally getting all 1-, 2-, 3-, ... k-predictors models for comparison

Forward modeling

Step 1 the first predictor in the model is the "best single predictor"

Select the predictor with the numerically largest simple correlation with the criterion ***-- if it is a significant correlation***

$$r_{y,x1} \text{ vs. } r_{y,x2} \text{ vs. } r_{y,x3} \text{ vs. } r_{y,x4}$$

by using this procedure we are sure that the initial model "works"

Step 2 the next predictor in the model is the one that will "contribute the most" -- with two equivalent definitions

1. The 2-predictor model (including the first predictor) with the numerically largest R^2 ***-- if the R^2 is significant and significantly larger than the r^2 from the first step***

$$R^2_{y.x3,x1} \text{ vs. } R^2_{y.x3,x2} \text{ vs. } R^2_{y.x3,x4}$$

2. Add to the model that predictor with the highest semi-partial correlation with the criterion, controlling the criterion for the predictor already in the model ***-- if the semi-partial is significant***

$$r_{y(x1.x3)} \text{ vs. } r_{y(x2.x3)} \text{ vs. } r_{y.(x4.x3)}$$

by using this procedure we are sure the 2-predictor model "works" and "works better than the 1-predictor model"

All subsequent steps -- the next predictor in the model is the one that will "contribute the most" -- with two equivalent definitions

1. The 2-predictor model (including the first predictor) with the numerically largest R^2 ***-- if the R^2 is significant and significantly larger than the R^2 from the previous step***

$$R^2_{y.x3,x2,x1} \text{ vs. } R^2_{y.x3,x2,x4}$$

2. Add to the model that predictor with the highest semi-partial correlation with the criterion, controlling the criterion for the predictors already in the model ***-- if the semi-partial is significant***

$$r_{y.(x1.x3,x2)} \text{ vs. } r_{y.(x4.x3,x2)}$$

by using this procedure, we are sure that each model "works" and "works better than the one before it"

When to quit ??? When no additional predictor will significantly increase the R^2 (same as when no multiple semi-partial is significant).

Difficulties with the forward inclusion model…

- The major potential problem is "over-inclusion" -- a predictor that contributes to a smaller (earlier) model fails to continue to contribute as the model gets larger (with increased collinearity), but the predictor stays in the model.

- Fairly small "variations" in the correlation matrix can lead to very different final models -- models often differ on two "splits" of the same sample

- The resulting model may not be the "best" -- there may be another model with the same # predictors but larger R^2, etc

All of these problems are exacerbated by increased collinearity !!

Backward Deletion

Step 1 -- start with the full model (all predictors) ***-- if the R^2 is significant.*** Consider the regression weights of this model.

Step 2 -- remove from the model that predictor that "contributes the least"

Delete that predictor with the largest p-value associated with its regression (b) weight -- if that p-value is greater than .05. (The idea is … the predictor with the largest p-value is the one "least likely to not be contributing to the model" in the population)

b_{x1}(p=.08) vs. b_{x2}(p=.02) vs. b_{x3}(p=.02) vs. b_{x4}(p=.27)

by using this procedure, we know that each model works as well as the previous one (R^2 numerically, but not statistically smaller)

On all subsequent steps -- the next predictor dropped from the model is that with the largest (non-significant) regression weight.

b_{x1}(p=.21) vs. b_{x2}(p=.14) vs. b_{x3}(p=.012)

When to quit ?? When all the predictors in the model are contributing to the model.

Difficulties with the backward deletion model…

- The major potential problem is "under-inclusion" -- a predictor that is deleted from a larger (earlier) model would contribute to a smaller model, but isn't "re-included".

- Fairly small "variations" in the correlation matrix can lead to very different final models -- models often differ on two "splits" of the same sample

- The resulting model may not be the "best" -- there may be another model with the same # predictors but larger R^2, etc

All of these problems are exacerbated by increased collinearity !!

Forward Stepwise Modeling

Step 1 the first predictor in the model is the "best single predictor" (same as the forward inclusion model)

Select the predictor with the numerically largest simple correlation with the criterion ***-- if it is a significant correlation***

by using this procedure we are sure that the initial model "works"

Step 2 the next predictor in the model is the one that will "contribute the most" -- with two equivalent definitions (same as the forward inclusion model)

1. The 2-predictor model (including the first predictor) with the numerically largest R^2 ***-- if the R^2 is significant and significantly larger than the r^2 from the first step***
2. Add to the model that predictor with the highest semi-partial correlation with the criterion, controlling the criterion for the predictor already in the model ***-- if the semi-partial is significant***

by using this procedure we are sure the 2-predictor model "works" and "works better than the 1-predictor model"

On all Subsequent steps (each having two parts)

a. -- remove from the model that predictor that "contributes the least" (same as the backward deletion model)

Delete that predictor with the largest p-value associated with its regression (b) weight -- if that p-value is greater than .05. (The idea is … the predictor with the largest p-value is the one "least likely to not be contributing to the model" in the population)

-- if a predictor is deleted, look for a second (third, etc) that should also be deleted, before moving on to part b.

by using this procedure, we are sure that all the predictors in the model are contributing before adding any additional predictors to the model

b. the next predictor in the model is the one that will "contribute the most" (same as for forward inclusion) -- with two equivalent definitions

1. The 2-predictor model (including the first predictor) with the numerically largest R^2 ***-- if the R^2 is significant and significantly larger than the r^2 from the first step***
2. Add to the model that predictor with the highest semi-partial correlation with the criterion, controlling the criterion for the predictor already in the model ***-- if the semi-partial is significant***

by using this procedure we are sure the model with the added predictor "works" and "works better than the model without it"

When to quit ? -- when BOTH of two conditions exist…

1. All predictors included in the model are contributing to it
2. None of the predictors that are not in the model would contribute if they were added.

by using this procedure we avoid both over-inclusion and under-inclusion

Difficulty with the forward inclusion model

- The resulting model may not be the "best" -- there may be another model with the same # predictors but larger R^2
- Assumes that the best model is found by starting with the best single predictor

This problem is exacerbated by increased collinearity !!

All subsets regression

Just what it sounds like… Get all 1-predictor models, all 2-predictor models, all 3-predictor models, etc.

Difficulties with all subsets regression

- generates an awful lot of models (none of which were theoretically determined)
- Many of the models will be very comparable in R^2 but have very different theoretical interpretations
- not available in most stats packages and wearisome to do by hand

All of these problems are exacerbated by increased collinearity !!

Problems of forward, backward and stepwise models …

All of these procedures were designed for construction of non-theoretical models, and they rarely provide direct tests of the sorts of hypotheses we have !!!

Each procedure has a different definition of “best model” that is tied to the particular selection and/or deletion rules that are used.

All decisions about which variables are “potential contributors” are made based on significance tests (not really a problem, but …)

All selection/deletion decisions between two or more “potential contributors” are made NUMERICALLY, and so, may result in very differently interpreted models with small differences in sampling variability.

Taken together, these problem often lead to …

Different procedures (forward, backward, stepwise) often produce different “best models” -- even when starting from the same set of predictors.

Different models “perform less differently” (in terms of R^2) than they appear (based on their different predictor memberships).

So...

All-subsets models often provide a useful “wake-up call” for recognizing the limited performance difference among models that may seem very different.

Finally…

None of these procedures are very useful for DESCRIPTIVE purposes, which usually require careful description of the simple correlations, full model, and various model comparisons to be usefully complete !!

(Classic “Trick”) Question -- When do you not need to even perform a simple regression model to make a prediction based on the single predictor variable?

Hint: Remember that the simple regression formula “just” changes the scale of the predictor into a “mimic” of the criterion -- those folks with higher scores on x will have higher scores on the y’ and those with lower scores on x will have lower scores on y’.

Answer: When the decision rule is to take the top (or bottom) n candidates. Say you want the “five best” candidates based on a single predictor -- the five with the highest x values will be those with the five highest y’ values !!

Will this apply to multiple regression modeling??