

Multiple Linear Regression IV

Sara López-Pintado

Department of Statistics

Rutgers University

Fall, 2005

Objective: discuss how to build a regression model

- Select the predictor or explanatory variables
- Introduce diagnostic tools for examining the appropriateness of a fitted regression model
- Discuss remedial measures when the model conditions are not met

- Model building processes
 - Data collection and preparation
 - Reduction of explanatory or predictor variables
 - Model refinement and selection
 - Model Validation

When considering exploratory observational studies, if the number of explanatory variables is large then we will wish to reduce the number of explanatory variables.

Reasons for this:

- Regression models with many explanatory variables are difficult to understand and interpret.
- Many highly intercorrelated variables may appear in the model and this increases the sampling variation of the regression coefficients among other things...
- Detracts the model's descriptive ability and doesn't improve the model's predictive ability.

- Goal: identify good subsets of potentially useful explanatory variables to be included in the final regression model and the determination of appropriate functional and interaction relations for these variables.
- No one subset of explanatory variables may be always the best. Depending on the purpose of the investigator different explanatory variables might be selected.
- Elimination of key explanatory variables can seriously damage the explanatory power of the model and lead to biased estimates. (Important omitted variables are called latent variables)
- If too many explanatory variables are included, then this overfitted model will often produce variance of the estimated parameters larger than the ones obtained with a simpler model.

Two different approaches for reducing the number of possible explanatory variables in an exploratory observational study:

1. Consider all possible subsets of explanatory variables that can be developed from the pool of potential predictor variables and identify those subsets that are good with respect to some criterion (useful for sets of explanatory variables that are small or moderate in size).
2. Consists in using automatic search procedures to arrive to a single subset of the explanatory variables: e.g. Stepwise regression (recommended for reductions involving large pools of explanatory variables).

Note: Be careful in fitting the regression model containing the entire set of potential x and then simply dropping those for which t-statistic has small absolute value.

- All possible regression procedures for variable selection
- Consider all possible subsets of x variables and identify for detail examination a few good subsets according to some criterion. (The number of possible subsets is 2^{P-1} , where $P - 1$ is the total number of potential predictor variables).
- This procedure is only reasonable when the number of possible variables in the model is small.
- Different criteria for comparing the regression models may be used: R_p^2 , MSE_p , C_p , $PRESS_p$

- R_p^2 or SSE_p criterion: (these two criterion are equivalent)

$$R_p^2 = 1 - \frac{SSE_p}{SSTO}$$

Note that R_p^2 always increases with p . Find the point where adding a variable is not worthwhile because it leads to a small increase in R_p^2 .

- MSE_p or R_a^2 : (equivalent criterions)

$$R_a^2 = 1 - \left(\frac{n-1}{n-p} \right) \frac{SSE}{SSTO}$$

Find the subset where MSE_p is minimized (or R_a^2 maximized)

- C_p criterion: related with the total mean squared error of the n fitted values for each subset regression model

$$\begin{aligned} E[\widehat{Y}_i - \mu_i]^2 &= \left(E(\widehat{Y}_i) - \mu_i \right)^2 + \sigma^2(\widehat{Y}_i) \\ &= \text{bias}^2 + \text{variance}(\widehat{Y}_i) \end{aligned}$$

where μ_i is the true mean response when the levels of the x variables are those in the i case.

Total mean square error:

$$\sum_{i=1}^n E[Y_i - \mu_i]^2 = \sum_{i=1}^n \left(E(\widehat{Y}_i) - \mu_i \right)^2 + \sum_{i=1}^n \sigma^2(\widehat{Y}_i)$$

Define

$$\Gamma_p = \frac{\sum_{i=1}^n \left(E(\widehat{Y}_i) - \mu_i \right)^2 + \sum_{i=1}^n \sigma^2(\widehat{Y}_i)}{\sigma^2}.$$

The estimated Γ_p is

$$C_p = \frac{SSE_p}{MSE(X_1, \dots, X_{p-1})} - (n - 2p)$$

where SSE_p is the squared sum error for the fitted model with p parameters ($p - 1$ variables).

If $E[\hat{Y}_i] = \mu_i$ (no bias in the model with $p - 1$ variables)

$$E[C_p] \approx p$$

Objective: to identify subsets of x variable for which

1. The C_p value is small
 2. The C_p value is near p
- AIC_p (Akaike's information criterion) (penalizes adding predictors in the model)

$$AIC_p = n \ln SSE_p - n \ln n + 2p$$

- $PRESS_p$ criterion: uses the prediction sum of squares

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

The $PRESS_p$ measure differs from SSE_p in that each fitted value \hat{Y}_i is obtained by deleting the i -th observation from the data

$$PRESS_p = \sum (Y_i - Y_{i(i)})^2$$

Objective: find models with small $PRESS_p$, because they have small prediction errors.

- Drawback of using all possible regression procedure is that it is computationally intensive when the number of variables is high.

- Automatic search procedures for model selection
- Best subset algorithms (methods of obtaining the best subsets according to a specified criterion without having to fit all possible models).
- Stepwise regression methods (very useful when the number of potential explanatory variables is greater than 30)
 - ★ It consists in developing a sequence of regression models, at each step adding or deleting an X variable.
 - ★ The criterion for adding or deleting an X variable can be stated equivalently in terms of error sum of squares reduction, coefficient of partial correlation, t-statistic or F-statistic.
 - ★ A weakness of this procedure is that it identifies a single model as the best one.
 - ★ Sometimes we can use automatic search procedures as a starting point.

- Forward stepwise procedure

We use the F statistic to add or delete a variable

- STEP 1: Fit a simple linear regression model with each $P - 1$ variables. Obtain the F statistic for each variable. The variable with highest F value is the first candidate for the first addition. If the F value exceeds a predetermined level, the X variable is added. Otherwise, the program terminates with no X added to the model.
- STEP 2: Assume X_7 is the variable added in step 1. Fit all regression models with 2 X variables, where X_7 is one of them. For each regression model compute the F statistic:

$$F_k^* = \frac{MSE(X_k/X_7)}{MSE(X_k, X_7)} = \left(\frac{b_k}{S(b_k)} \right)^2$$

The variable X_k with largest F_k^* is the candidate to enter the model. If the value exceeds a predetermine level, the second variable is added, otherwise the program terminates.

- STEP 3: Suppose X_3 is added. Now examine if other X variables in the model should be dropped. At this point only one other X variable is in the model:

$$F_7^* = \frac{MSE(X_7/X_3)}{MSE(X_7, X_3)} = \left(\frac{b_7}{S(b_7)} \right)^2$$

If this F_7^* value falls below a predetermined limit, the variable is dropped from the model, otherwise it is retained.

- STEP 4: Suppose that X_7 is retained in the model. The stepwise regression routine examines which X variable is the next candidate for addition. Then examines, whether any of the variables already in the model should now be dropped, and ...so on until no further X variables can either be added or deleted, at which point the search terminates.

- Other automatic search procedures:
- Forward Selection
- Backward Elimination