Chapter 8: Model Building

1. Data Collection
2. Reduction of Explanatory variables
3. Model Refinement and selection
4. Model Validation

Data Collection

1. Controlled experiments
2. Controlled Experiments with supplemental variables
3. Confirmatory Studies
4. Exploratory Observational Studies

Reduction of Explanatory Variables

We start with a large list of potential variables that may affect the response

- However we always want the most economical model
- Overfitting is not always the best option
- Too many variables may be inter-related

The problem is that we select ONE model from a pool of candidate models. And in doing so the first question we need to ask is:
What is to be done with the model?

Some possibilities:

1. Learn something about the system from which the data are taken.
2. Learn which regressors are important, which are not, i.e. variable screening
3. Prediction

Please keep in mind:

Statistics is rarely a substitute for sound scientific knowledge and reasoning. It is more of an AID

All possible regression procedure

Criterions: R_p^2, Adj. R_p^2, $C_{p.}$, $PRESS_p$
Notation: p is the number of predictors in the model (INCLUDING the intercept)

R_p^2 : Multiple Coefficient of Determination

=1- (SS Res)/ (SS Total)

Adj R_p^2: Adjusted Multiple Coefficient of Determination

$$= 1 - \frac{SS\ \text{Res}/(\text{n} - \text{p})}{\text{SS Total}/(\text{n} - 1)} = 1 - \frac{s^2/(n-1)}{SSTotal}$$

$C_{p.}$: Collin Mallows Cp. criterion

$$= Cp = p + \frac{(s^2 - \hat{\sigma}^2)(n-p)}{\hat{\sigma}^2}$$

Idea:

$$\sum_{i=1}^{n} \frac{MSE(\hat{y}(x_i))}{\sigma^2} = \sum_{i=1}^{n} \{ \frac{Var(\hat{y}(x_i)) + [Bias(\hat{y}(x_i))]^2}{\sigma^2} \}$$

It can be shown that:

$$\sum_{i=1}^{n} \{ \frac{Var(\hat{y}(x_i))}{\sigma^2} \} = \sum_{i=1}^{n} x_{1i}'(X_1'X_1)^{-1}x_{1i} = trI_p = p$$

Want Cp=p and one favors the model with the smallest Cp, because it indicates lower bais.

$Press_p$: Predicted Sums of Squares

How well the use of the fitted values for a subset model can predict the observed response Yi

$$PRESSp = \sum_{i=1}^{n}(y_i - \hat{y}_{i,-i})^2$$

It can be shown that, $$PRESS = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Forward Stepwise Selection

Too many predictor variables would make the selection of best subsets too difficult.

Forward Selection (step-wise)

- Put one variable in first.
- Introduce the second one based on partial F-test.
- If significant keep variable, else drop it

Generally require a pre-selected F-in and F-out for a variable to enter or exit a model. The rationale is, following entry a variable has to continue to perform or be eliminated.

This way the model can drop variables once introduced in the model.

Other procedures are Forward and Backward procedures, which use sequential F-tests.

MaxR procedure:

Produces k models one for each number 1- k. This allows the analyst a choice based on pre-conceived notion regarding the SIZE of the final model.

Methodology is similar to Forward Step-wise. At each stage the variable that enters is the one that produces the largest increase in R-sq. No F tests are conducted.

At each stage it allows for replacement of a model regressor (one that is already in the model) with one that has not entered, IF the replacement produces the best R-sq.

Remember:

1. No assurance that the BEST model of a particular size is found.
2. The best model of a particular size is being found in the sense of PRESS.