

## 3.2: Variable Selection

Dr. Bean - Stat 5100

### 1 Why Variable Selection

- Up until now, we have focused on trying to make predictions/inference using all the potential explanatory variables we have available to us.
- We now wish to consider several candidate models, ultimately making a judgment as to which model is “best.”
  - Selection is more than an art than it is a science: no “right” decisions, several *wrong* decisions, several “reasonables.”
  - This is an iterative process, that makes it difficult to know when we are “done” (see Figure 1 on last page).
- One element of the model building process involves **selecting a subset** of potential explanatory variables for use in the final model.
  - Follows the Ockham’s razor principle: *entia non sunt multiplicanda praeter necessitatem*

“Entities should not be multiplied without necessity” (i.e. all else equal: simpler answers are better).

(Groups) Why might we prefer simpler models to complex ones? *Should* we prefer simpler models to more complex ones?

- Simpler models are easier to interpret/describe.
- Simpler models are harder to overfit.
- *Conversely*, simpler models may fail to describe a complex problem.

(Groups) Why is variable selection not something we would normally want to use in an experimental setting?

Observational studies are usually searching to find *something* interesting, while in an experiment, we wish to test whether *specific things* are interesting. In experiments, we should have decided beforehand what factors we were going to control for.

### 2 Methods of Variable Selection

How to pick the “best” subset of variables?

- Whenever possible, remove variables based on **context**, which comes with **expertise**.
- Automatic Methods:
  - **All possible regressions:** Consider all possible combinations of predictor variables, select the “best” model according to some measurement criteria.
  - **Stepwise methods:** Take a structured approach that takes a (semi) intelligent search through a subset of all possible models.
  - **Penalized regression:** more in Module 4.

## 2.1 All Possible Regressions

Consider all subsets of predictor variables  $X_1, \dots, X_{p-1}$ .

- Number of subsets of size  $p - 1 = \binom{P-1}{p-1} = \frac{(P-1)!}{(p-1)!(P-p)!}$ .
- Number of subsets of all possible sizes:  $\sum_{p=1}^P \binom{P-1}{p-1} = 2^{P-1}$ .

### 2.1.1 Measures of “goodness”

- R-square - but which model will always have the highest  $R^2$ ?

$$R_p^2 = 1 - \frac{SS_{Error,p}}{SS_{Total}}$$

- Adjusted R-square - balances against # of predictors

$$R_{a,p}^2 = 1 - \frac{n-1}{n-p} \cdot \frac{SS_{Error,p}}{SS_{Total}}$$

As  $p$  increases,  $R_{a,p}^2$  first increases, then decreases

- Mallows’s  $C_p$  - for a certain subset of  $p - 1$  predictors:

$$C_p = \frac{SS_{Error} \text{ from model with } p-1 \text{ predictors}}{MSE \text{ from model with } P-1 \text{ predictors}} + 2p - n$$

When a subset of  $p - 1$  predictors gives unbiased  $\hat{Y}$ ’s,  $E[C_p] \approx p$ .

– so look for model with smallest  $p$  such that  $C_p \approx p$ ,

i.e., want  $C_p \approx \# \text{ predictors} + 1$ .

- Akaike’s information criteria & Schwarz’s Bayesian criterion
  - both penalize larger numbers of predictors (want small):

$$\begin{aligned} AIC_p &= n \log SS_{Error,p} - n \log n + 2p \\ SBC_p &= n \log SS_{Error,p} - n \log n + p \log n \end{aligned}$$

- Prediction sum of squares – based on leave-one-out philosophy ( $\hat{Y}_{i(i)}$ )

$$PRESS_p = \sum_{i=1}^n \left( Y_i - \hat{Y}_{i(i)} \right)^2$$

– look for models with small  $PRESS_p$

## 2.2 Stepwise Selection

Stepwise methods:

- automatically select a model based on some criterion (convenient)
- less satisfactory, do not “guarantee” the “right” model
- best used as “confirmatory” approaches
- three main: backward (okay), forward (worst), stepwise (hybrid)

Backward Elimination – basic algorithm

1. Fit model with all  $P - 1$  predictors
  - (a) Compare each predictor’s individual P-value to some threshold (**slstay**; default in SAS is 0.10)
  - (b) If any predictor’s P-value  $> \mathbf{slstay}$ , drop predictor with largest P-value
2. Repeat with  $P - 2$  predictors
3. Continue until all predictors remaining have P-values below **slstay**

Forward Selection – basic algorithm

1. Find predictor with highest correlation with response
  - (a) Regress response on this predictor
  - (b) Leave predictor in model if P-value is below some threshold (**slentry**; default in SAS is 0.50)
2. Given the previously entered predictor, find the predictor with the highest partial correlation with response
  - (a) Add this predictor to the model
  - (b) Leave in model if P-value is below **slentry**
3. Continue until no more predictors warrant inclusion (P-value of “next” predictor above threshold)

Big problem here: best 2-variable model does not necessarily contain best 1-variable model (first step(s) can throw everything off)

Stepwise Selection – basic algorithm:

1. Take a “forward” step: add “best” predictor with P-value below **slentry** (default 0.15)
2. Take a “backward” step: evaluate all predictors in model and drop the variable with the highest P-value above **slstay** (default 0.15)
3. Iterate “forward” and “backward” steps until model stays the same

Note: in all these automatic stepwise procedures (backward, forward, stepwise), the `slentry` and `slstay` thresholds are deceptive. After the first step (really a hypothesis test), they are not significance levels ( $\alpha$ ), but “conditional” significance levels, which are harder to interpret.

**(Individual)** Clearly, it would be better to compare all possible models, rather than a subset (in stepwise methods). Why do stepwise methods even exist?

When  $P$  gets large, fitting  $2^{P-1}$  models quickly becomes unrealistic computationally. Also, chance of a “consensus” among measurement techniques as to which is the best model becomes unlikely.

### 2.3 Remember this...

- In order to have reliable results, we need  $n \gg P$  (often 6\*10 times larger).
- Each described technique measures how well your models fit the data you already have, which might not translate to new data (in production).

We get a sense of how our models perform on new data by:

- Splitting our data into “training” and “test” sets.
- Fit each model using only the training data, then use the model to predict on the test data.
- Calculate the mean square prediction error:

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

**FIGURE 9.1**  
Strategy for  
Building a  
Regression  
Model.

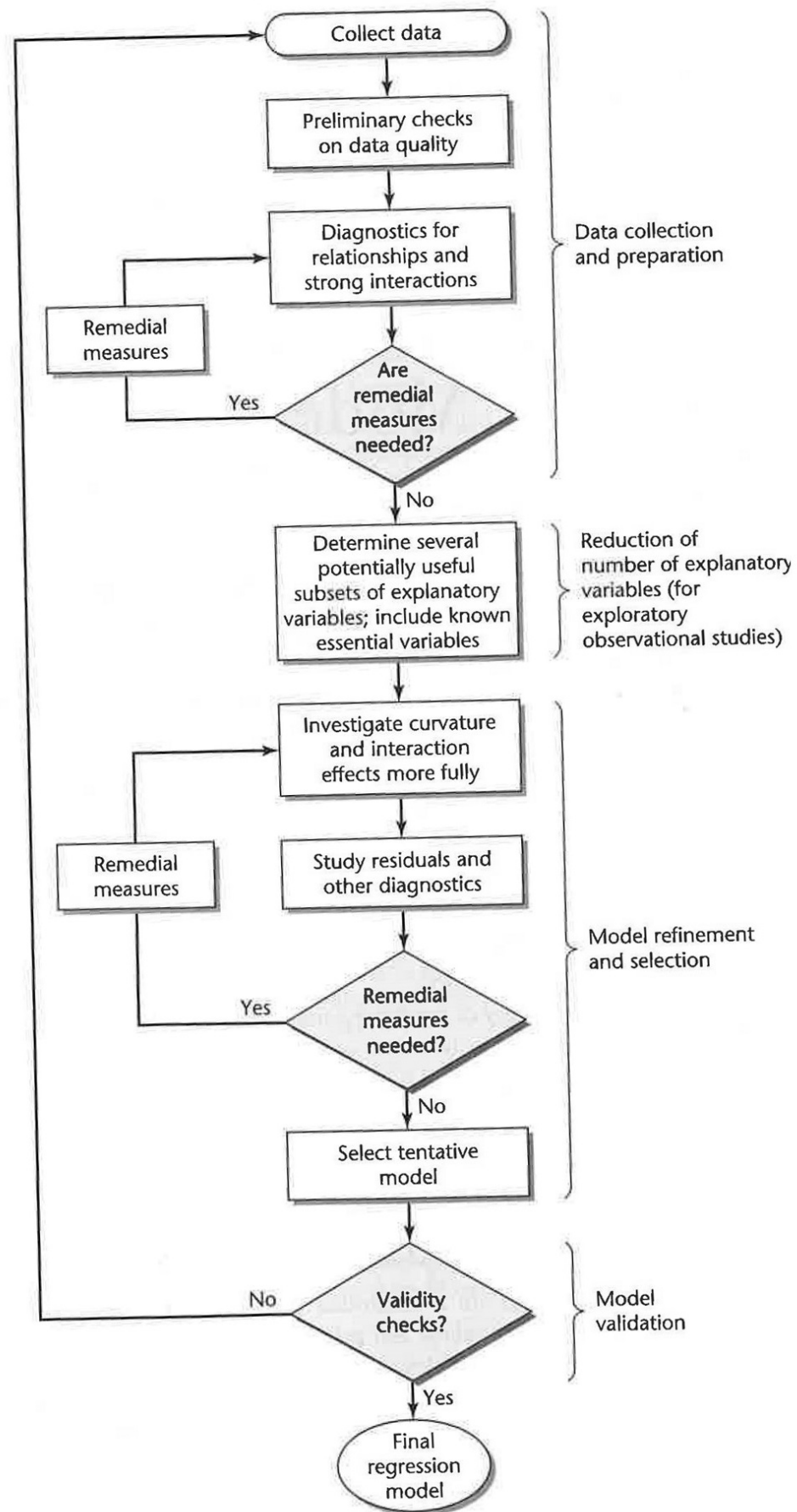


Figure 1: General model for multiple regression model selection (taken from Kutner et. al. (2004)).