Model Selection and Validation

Data Collection Strategies

- Controlled Experiments – Subjects (Experimental Units) assigned to X-levels by Experimenter
  - Purely Controlled Experiments – Researcher only uses predictors that were assigned to units
  - Controlled Experiments with Covariates – Researcher has information (additional predictors) associated with units
- Observational Studies – Subjects (Units) have X-levels associated with them (not assigned by researcher)
  - Confirmatory Studies – New (primary) predictor(s) believed to be associated with Y, controlling for (control) predictor(s), known to be associated with Y
  - Exploratory Studies – Set of potential predictors believed that some or all are associated with Y

Reduction of Explanatory Variables

- Controlled Experiments
  - Purely Controlled Experiments – Rarely any need or desire to reduce number of explanatory variables
  - Controlled Experiments with Covariates – Remove any covariates that do not reduce the error variance
- Observational Studies
  - Confirmatory Studies – Must keep in all control variables to compare with previous research, should keep all primary variables as well
  - Exploratory Studies – Often have many potential predictors (and polynomials and interactions). Want to fit parsimonious model that explains much of the variation in Y, while keeping model as basic as possible. Caution: do not make decisions based on single variable t-tests, make use of Complete/Reduced models for testing multiple predictors

Model Selection Criteria – All Possible Regressions

$P-1$ predictors $\Rightarrow$ 2^{P-1} potential models (each variable can be in or out of model)

R_p^2 or SSE_p criterion (Goal: find p so that $\max\left(R_p^2\right)$ or $\min\left(SSE_p\right)$ "flattens out"):

$$R_p^2 = \frac{SSR_p}{SSTO} = 1 - \frac{SSE_p}{SSTO} \qquad p = \# \text{ of parameters in current model}$$

$R_{a,p}^2$ or MSE_p criterion (Goal: find model that maximizes (or close to) $R_{a,p}^2$ and minimizes MSE_p):

$$R_{a,p}^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE_p}{SSTO} = 1 - \frac{\left(SSE_p/(n-p)\right)}{\left(SSTO/(n-1)\right)} = 1 - \frac{MSE_p}{\left(SSTO/(n-1)\right)}$$

Mallow's C_p criterion (Goal: find model with smallest p so that $C_p \le p$):

$$C_p = \frac{SSE_p}{MSE\left(X_1,\ldots,X_{P-1}\right)} - (n-2p)$$

AIC_p and SBC_p criteria (Goal: choose model that minimize these values):

$$AIC_p = n\ln\left(SSE_p\right) - n\ln(n) + 2p \qquad SBC_p = n\ln\left(SSE_p\right) - n\ln(n) + \left[\ln(n)\right]p$$

$PRESS_p$ criterion (Goal: Small values):

$$PRESS_p = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_{i(i)}\right)^2 \qquad \hat{Y}_{i(i)} \equiv \text{fitted value for } i^{th} \text{ case when it was not used in fitting model}$$

Regression Model Building

- Setting: Possibly a large set of predictor variables (including interactions).
- Goal: Fit a parsimonious model that explains variation in *Y* with a small set of predictors
- Automated Procedures and all possible regressions:
  – Backward Elimination (Top down approach)
  – Forward Selection (Bottom up approach)
  – Stepwise Regression (Combines Forward/Backward)

Backward Elimination Traditional Approach

- Select a significance level to stay in the model (e.g. SLS=0.20, generally .05 is too low, causing too many variables to be removed)
- Fit the full model with all possible predictors
- Consider the predictor with lowest *t*-statistic (highest *P*-value).
  – If $P >$ SLS, remove the predictor and fit model without this variable (must re-fit model here because partial regression coefficients change)
  – If $P \leq$ SLS, stop and keep current model
- Continue until all predictors have *P*-values below SLS
- Note: R uses model based criteria: AIC, SBC instead

Forward Selection – Traditional Approach

- Choose a significance level to enter the model (e.g. SLE=0.20, generally .05 is too low, causing too few variables to be entered)
- Fit all simple regression models.
- Consider the predictor with the highest *t*-statistic (lowest *P*-value)
  - If $P \leq$ SLE, keep this variable and fit all two variable models that include this predictor
  - If $P >$ SLE, stop and keep previous model
- Continue until no new predictors have $P \leq$ SLE
- Note: R uses model based criteria: AIC, SBC instead

Stepwise Regression – Traditional Approach

- Select SLS and SLE (SLE<SLS)
- Starts like Forward Selection (Bottom up process)
- New variables must have $P \leq$ SLE to enter
- Re-tests all "old variables" that have already been entered, must have $P \leq$ SLS to stay in model
- Continues until no new variables can be entered and no old variables need to be removed
- Note: R uses model based criteria: AIC, SBC instead (e.g., stepAIC() in MASS)

Model Validation

- When we have a lot of data, we would like to see how well a model fit on one set of data (training sample) compares to one fit on a new set of data (validation sample), and how the training model fits the new data.
- Want data sets to be similar wrt levels of the predictors
- Training set should have at least 6-10 times as many observations than potential predictors
- Models should give similar model fits based on SSE_p, $PRESS_p$, C_p, and MSE_p and regression coefficients
- Mean Square Prediction Error when training model is applied to validation sample:

$$MSPR = \frac{\sum_{i=1}^{n^*}\left(Y_i - \hat{Y}_i\right)^2}{n^*} \qquad \hat{Y}_i = b_0^T + b_1^T X_{i1}^V + \ldots + b_{p-1}^T X_{i,p-1}^V$$