Stat 608 Chapter 7

Variable Selection

Introduction

- Goal: Choose the best model using variable selection methods.
- Start by considering the full model containing all m potential predictor variables:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + e$$

- Variable selection methods choose the subset of predictors that is "best".
- **Overfitting**: including too many predictors (model performs as well or worse than simpler models at predicting new data)
- **Underfitting**: including too few predictors

Introduction

- If the goal is interpretation, simpler models are usually preferred. Use a method that chooses fewer models.

- If the goal is prediction, more variables may be acceptable.

Forward, Backward, and Stepwise Subsets

- If there are m variables, there are 2^m possible regression equations. If m is small enough, run all of them (all possible subsets).
- Backward, Forward, and Stepwise selection procedures examine only *some* of the 2^m possible regression equations.
- **Backward elimination**:
  1. All variables are included in the model. The predictor with the largest p-value is deleted (as long as it isn't significant).
  2. The remaining m-1 variables are now in the model. Again, the predictor with the largest p-value is deleted (as long as it isn't significant).
  3. Variables are deleted until all remaining variables are significant.

Forward, Backward, and Stepwise Subsets

- **Forward selection**:
  1. No variables are in the model. All m models with only one predictor are run. The predictor with the smallest p-value is entered in the model (as long as it is significant). Call this variable x_1.
  2. All models with predictors x_1 and only one other predictor are run; of the remaining predictors x_2, …, x_m, the one with the smallest p-value is entered (as long as it is significant).
  3. Variables are entered until no more predictors are significant, given the others already in the model.

Stepwise Subsets

- **Stepwise Selection Procedure:**
  1. Choose α_E and α_R, significance levels to Enter and Remove predictors.
  2. Forward step: No variables are in the model. All models with one predictor are run. The predictor with the smallest p-value is entered into the model, as long as the p-value is less than α_E. Call this variable x_1.
  3. Forward step: All models with predictors x_1 and only one other predictor are run; of the remaining predictors x_2, …, x_p, the one with the smallest p-value is entered, as long as the p-value is less than α_E.
  4. Backward step: Check to see that the p-value for variable x_1 is smaller than α_R. If not, remove it. If so, leave it in.
  5. Take another forward step, attempting to add a third variable.
  6. Continue taking backward and forward steps until adding an additional predictor does not yield a p-value below α_E.

Could α_E be larger than α_R? Vice versa?

Stepwise is a forward selection procedure, except that a variable can be removed once it is in.

Forward, Backward, and Stepwise Subsets

- These procedures only consider some of the predictors, so they do not necessarily find the model that fits the data the best among all possible subsets.
- Forward, backward, and stepwise may not produce the same final model, though they often do.
- If covariance of the predictors = 0, all three produce the same final model.
- These methods are prone to overfitting, but stiff criteria for adding or deleting variables can mitigate this problem.
- Shouldn't we just remove the insignificant terms all at once?
  - Chapter 5: F-Test for model reduction
  - Chapter 7: Algorithms (not hypothesis tests)

Selection Criteria: (1) R^2 -Adjusted

- Adding irrelevant predictor variables to the regression model often increases R^2.
- To compensate, we adjust for the number of predictors:

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$$

- Choose the subset of the predictors that has the highest value of R^2_{adj}. This is equivalent to choosing the subset of the predictors with the lowest value of MSE (mean square error).

Selection Criteria: (2) AIC (Akaike's Information Criterion)

- Based on maximum likelihood estimation
- R uses the calculation:

$$AIC = n \log\left(\frac{RSS}{n}\right) + 2p$$

- Choose the model which makes AIC as small as possible. (By small, we mean close to -∞).

AIC: Derivation

- p = # parameters = k+1
- Suppose all variables are normally distributed; then y | x1, …, xp is normally distributed.

$$f(y_i|x_{1i}, x_{2i}, \ldots, x_{ki}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \{\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}\})^2}{2\sigma^2}\right)$$

- Likelihood estimation is a method of finding the most likely value of the parameter, given the observed data.
- The likelihood function is the joint distribution of the data, given the parameters, flipped.

AIC: Derivation

- Likelihood function:

$$
\begin{aligned}
& L(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2 | Y) \\
= \; & \Pi_{i=1}^{n} f(y_i | x_i) \\
= \; & \Pi_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \{\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}\})^2}{2\sigma^2} \right) \\
= \; & \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \{\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}\})^2 \right)
\end{aligned}
$$

AIC: Derivation

- Maximizing the log likelihood yields the same statistic as maximizing the likelihood.

$$\log\left(L(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2|Y)\right)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \{\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}\})^2$$

- Taking partial derivatives yields estimates for the β's that are the same as the least squares estimates.

AIC: Derivation

- Substituting LS estimates, we have:

$$\log\left(L(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2|Y)\right)$$
$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}RSS$$

- Then we can solve for the (biased) estimate of the variance of the errors:

$$\hat{\sigma}^2_{MLE} = \frac{RSS}{n}$$

AIC: Derivation

- Akaike found a relationship between information and likelihood theory.
- Idea: measure information lost when model *g* is used to model *f*. This information is estimated by a quantity proportional to:

$$log(L(\hat{\theta}|Y)) - (p+1)$$

- We get the final form after multiplying the above by -2 and ignoring terms that do not depend on RSS or p.

Selection Criteria: (3) AIC_C (AIC Corrected)

- Corrects for bias when n small or p large compared to n. (AIC tends to overfit; the penalty for model complexity is not strong enough.)
- Converges to AIC as n increases.

$$AIC_C = AIC + \frac{2(p+2)(p+3)}{n-p-3}$$

- Choose the model which makes AIC_C as small as possible.

- IMPORTANT NOTE: The formula above is correct; the textbook is incorrect on page 231. See www.stat.tamu.edu/~sheather/book/docs/Errata.pdf.

Selection Criteria: (4) BIC (Bayesian Information Criterion, aka SBC)

- Based on posterior probability of model, but often used in a frequentist sense.

$$BIC = n \log\left(\frac{RSS}{n}\right) + (p+2)\log(n)$$

- Choose the model which makes BIC as small as possible.
- BIC is similar to AIC except with 2p replaced by p log(n). When $n \geq 8$, $\log(n) \geq 2$, so the penalty term for BIC is larger than the penalty term for AIC. BIC favors simpler models than AIC.

Comparison of Selection Procedures

- R^2_{adj} tends toward over-fitting.
- Pro of AIC and AIC_C: They are "efficient." Asymptotically, the error in prediction from the model using AIC and AIC_C is no different from the error from the best model. Not true of BIC.
- Pro of BIC: The probability it selects the correct model is asymptotically 1. Not true of AIC.
- AIC chooses models too complex when n is large. BIC chooses models too simple when n is small.

Comparison of Selection Procedures

- All possible subsets:
  - If the number of predictors in the model is of fixed size p, all four criteria R^2_{adj}, AIC, AIC_C, and BIC choose the same model.
  - When comparing models with different numbers of predictors, we can get different answers.

- Forward, Backward, and Stepwise:
  - Using other information criteria (AIC, BIC) to select a model is equivalent to using p-values to add and remove variables; the difference is where the algorithm stops.

Reminders

- The regression coefficients obtained after variable selection are biased.
- P-values from these models are generally much smaller than their true values.
- Software treats each column of the design matrix as being completely separate, ignoring relationships in polynomial models and models with interaction terms.

Bridge Data

Subset Size	Predictors	R2adj	AIC	AICC	BIC
1	log(Dwgs)	0.702	-94.90	-94.31	-91.28
2	log(Dwgs), log(Spans)	0.753	-102.37	**-101.37**	**-96.95**
3	log(Dwgs), log(Spans), log(Ccost)	**0.758**	**-102.41**	-100.87	-95.19
4	log(Dwgs), log(Spans), log(Ccost), log(Darea)	0.753	-100.64	-98.43	-91.61
5	log(Dwgs), log(Spans), log(Ccost), log(Darea), log(Length)	0.748	-98.71	-95.68	-87.87

LASSO

- LASSO: Least Absolute Shrinkage and Selection Operator, performs variable selection and parameter estimation simultaneously.
- Constrained Least Squares:

$$\min \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_{1i} + \ldots \beta_p x_{pi}])^2, \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq s$$

for some number s non-negative.

- When s is very large, this is equivalent to the usual least squares estimates for the model.
- When s is small, some of the coefficients are 0, effectively removing them from the model.

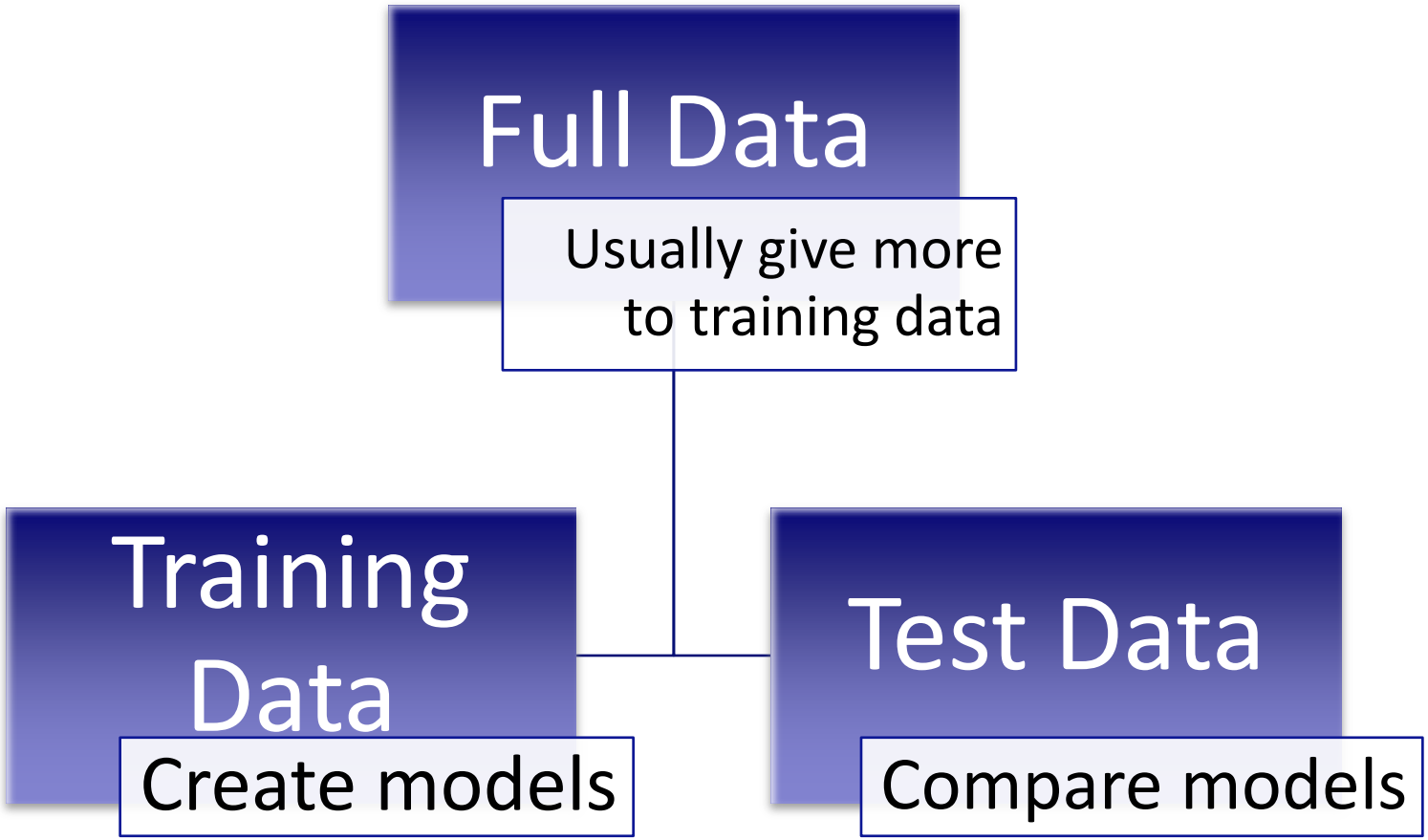
More ideas for variable reduction

- Principal Components: Linear Combinations of the original variables
- Factor Analysis: Sort of PC's. More interpretable.
- Variable Clustering: Each cluster is combination of some variables.

Assessing the Predictive Ability of Regression Models

- Since regression coefficients are biased and p-values are generally much smaller than their true values, we need another approach:
- Split the data, and see how well models built on one part predict the other part not being used to build the model.

Assessing the Predictive Ability of Regression Models



Assessing the Predictive Ability of Regression Models

- Ideally, the training and test data sets will be similar with respect to:
  - Univariate distributions of each of the predictors and response
  - Multivariate distributions of all variables
  - Means, variances, other moments
  - Outliers

- Usually, splitting the data is done randomly. However, especially in small data sets, the above criteria are not always met.