CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 45
Best Subset Regression

Chris A. Mack
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

Subset Regression

- Given a list of possible regressor variables, including possible interactions:
  - Define the “full” model as the model that includes every regressor and every interaction we care to consider
  - A “subset model” is a model with one or more terms missing
- Which subset model (or full model) is the best?

Testing Individual Terms

- For any individual model term, does the confidence interval for the parameter include 0?
  - The studentized parameter $b_k/SE(b_k)$ is student-t distributed with $n - p$ degrees of freedom
  - Calculate the p-value for a hypothesis test with H_0: $\beta_k = 0$
- If a term is not statistically significant, we generally leave it out of the model

Partial F-Test: Comparing Models

- Consider two models, where model 1 is a sub-set of model 2
  - Is it useful to include the extra variables of model 2 in the regression?
  - Are the added covariates informative?
- Consider this hypothesis test:
  - H_0: added predictor variables have all $\beta_j = 0$
  - H_A: at least one added predictor variable has $\beta_j \neq 0$

Partial F-Test: Comparing Models

- We want to compare the SSE (or chi square for a weighted regression) for these two models:

This is called the extra sum of squares (model 1 is a sub-set of model 2)

$$F = \frac{(SSE_1 - SSE_2)/(p_2 - p_1)}{SSE_2/(n - p_2)} = \frac{(R_2^2 - R_1^2)/(p_2 - p_1)}{(1 - R_2^2)/(n - p_2)}$$

- This statistic is F-distributed with $p_2 - p_1, n - p_2$ degrees of freedom (p = # of adjustable parameters in the model)

Partial F-Test Example

- Full Model: $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
- We suspect that $\beta_1 = \beta_3$
- Create a subset model with $\tilde{x}_1 = x_1 + x_3$

$$\hat{y} = \beta_0 + \beta_1 \tilde{x}_1 + \beta_2 x_2$$

- The Partial F-test of the subset model to the full model is equivalent to testing the null hypothesis that $\beta_1 = \beta_3$

Mallows C_p

- Compare a full model (with k adjustable model parameters) to a subset model (with p adjustable model parameters)

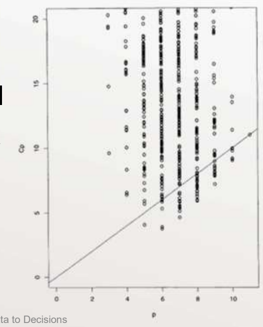
an estimate for χ^2

$$C_p = \frac{SSE_p}{\hat{\sigma}_\varepsilon^2} - n + 2p = \frac{SSE_p}{MSE_k} - n + 2p$$

- If the subset model is correct, then C_p should be about equal to p (smaller is better)
- Only useful when $n >> k$

Mallows C_p

- For moderate k, calculate C_p for every subset model
- If $C_p < p$, we have a candidate model



Colin L. Mallows, "More Comments on Cp", *Technometrics*, **37**(4), 362-372 (Nov 1995).

Likelihood Ratio

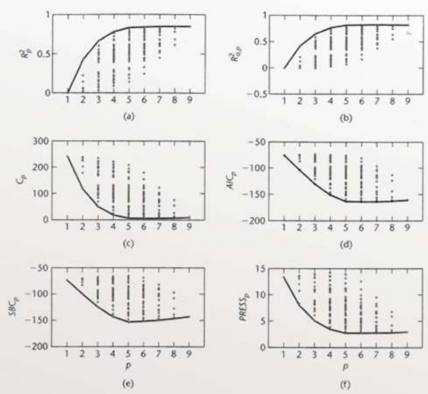
- For nested models
  - Let L_k = likelihood for full k-parameter model
  - Let L_p = likelihood for p-subset model ($L_k > L_p$)

$$G^2 = -2\ln\left(\frac{L_p}{L_k}\right)$$

- For large samples, if the subset model is correct then G^2 is approximately chi square distributed with $DF = k - p$
- If G^2 is too large, we reject the subset model

362 Part Two Multiple Linear Regression

FIGURE 9.5 Plot of Variable Selection Criteria with All Eight Predictors—Surgical Unit Example.



Full Search

From Kutner, Nachtsheim, Neter, and Li, "Applied Linear Statistical Models', Fifth edition, McGraw-Hill (2005).

Lecture 45: What have we learned?

- Understand the use of the partial F-test for subset models
- How is Mallow's C_p used to evaluate subset models?
- How is the likelihood ratio used to evaluate subset models?