Discrete Response Model
Lecture 5

Models for Count Response, Discrete Response Model Evaluation, and Model Selection

datascience@berkeley

Variable Selection

Stepwise Selection

- **Stepwise methods** for variable selection:
  - Use with caution.
  - For datasets with not too many variables (say, no more than a couple hundred), doing EDA is important.
  - For datasets with thousands-plus variables, a selection method is likely needed.
  - Always remember that theory, subject matter knowledge, and contextual information are important.
  - More details are covered in the text.

- Notice that all of these variable selection methods assume a "given a set of variables." In practice, it is common to create additional variables.
- As such, when building a model, one may have to:
  1. Examine the given set of variables.
  2. Consider various transformations of a selected set of variables.
  3. Consider create additional variables.
  4. Select a set of variables among the given, transformed, and the created variables.

LASSO

- The least absolute shrinkage and selection operator (LASSO) (Tibshirani (1996)) has evolved since.
  - Basic idea: Add a penalty to the log-likelihood function and then maximize it to obtain estimates.
  - This penalty is chosen to help extenuate the effects of those explanatory variables that are truly important, while keeping parameter estimates close to 0 for those parameters that are not truly important.
  - The model with the smallest residual deviance is considered to the "best."

The LASSO parameters estimate $\hat{\beta}_{0,LASSO}, \hat{\beta}_{1,LASSO}, \ldots, \hat{\beta}_{p,LASSO}$ maximize

$$log\left(L\left(\beta_0, \beta_1, \ldots, \beta_p | y_1, \ldots, y_n\right)\right) - \lambda \sum_{j=1}^{p} |\beta_j|$$

where λ is a .

Berkeley SCHOOL OF INFORMATION