Model Selection and Regularization

Bharat Bhushan Verma

Building a Model

- Which predictor variables to include in the model?
    - Are there any noise inducing variables included in the model?
    - Have you left out any important variable in the model?
- Is linear model the right model?
- What is the purpose of the model? Theory building or Theory testing?

Purpose of Building the Model

- Explanation: Explain the current relation in best functional form
    - SE(estimated y) and SE(Betas) both should be small.
- Control: Sensitivity of the relationship between y and predictors
    - SE(Betas) are very important
- Prediction: Predict previously unseen values in future
    - SE(estimated y) should be small but SE(Betas) are less important.
    - Overfitting needs to be avoided and form of function is not very important.
    - Important to scope the range of values where model is valid.

Steps in Model Building

- Model should be parsimonious and interpretable.
- Start with simplest model
- Consider adding more complexity or interactivity if needed.
- How to determine if model is useful and not overfitting.
  - R^2 tends to keep improving with every additional new predictor.
- Instead, we use
  - $R_{adj}^2 = R^2 - \left(\frac{p-1}{n-p}\right)(1 - R^2)$ results in numerous predictors in models.
  - Information Criteria $AIC = -2\ln(L) + 2p$ is balanced and independent of n.
  - Information Criteria B$IC = -2\ln(L) + pln(n)$ results in less predictors.

Limitations of OLS

- Prediction Accuracy:
  - OLS has least bias.
  - n >> p: OLS yields low variance.
  - n > p: OLS tends to overfit and poor predictions on future observations.
  - P > n: variance is infinite and estimates are non-unique.
- Model Interpretability:
  - OLS tends to produce non-zero coefficient on irrelevant variables.
  - It leads to unnecessarily complex models.

Alternates to OLS

- Subset Selection: Use reduced set of relevant variables.
- Dimension Reduction: Use PCA to extract relevant features.
- Shrinkage: Reduce variance of estimated coefficients.