Subset Selection: an overview

Statistical Data Mining I

Rachael Hageman Blair

Subset Selection

Why? Where does Least Squares fall short?

- Least squares estimates have low bias but large variance. Prediction accuracy can be improved by trading bias for smaller variance.
- Overall interpretation with so many predictors. It is easier in practice, to focus on smaller subsets of predictors with large effects. We are willing to sacrifice some of the details for simplistic interpretation.

The idea behind subset selection:

- Retain only a subset of variables, and use least squares to fit the reduced models.

  The question is **how to choose the subset ? ….**

Subset Selection

Often Criticized:

- The number of possible subsets can be huge.
- Often there are several ‘good’ models.
- The best X columns may be no better than random variables.
- Bias in the regression coefficients.
- Can rely on a series of F-tests, multiple testing on the same model.

Subset Selection

Best subset selection - Leaps and Bounds algorithm:

- Find the best subset of size $k \in \{0,1,2,\dots,p\}$, which minimizes the RSS.
- Computationally intensive, works well for $p \le 30$.
- Choice of k is somewhat subjective.

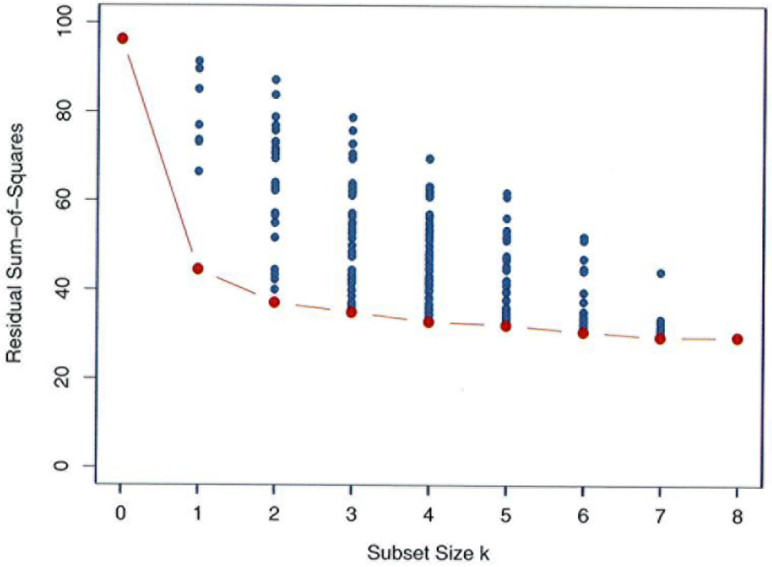

FIGURE 3.5. *All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.*

Subset Selection

Forward Stepwise Selection:

- A greedy algorithm which produces a nested sequence of models.
- Starts with the intercept, and progressively adds to the model to find the best fit for *k* predictors.
- Searches for the next step in the path in a smart way.
- Still need to specify *k*.

Backwards Stepwise Selection:

- Starts with the full model, and sequentially deletes the predictor with the least impact.
- The candidate for dropping is the smallest Z-score.
- Can only be used when $N > p$.

Subset Selection

Forward-Stagewise Selection:

- Computationally slow method.
- The algorithm identifies the variable most correlated with the residual.
- Operates like forward-stepwise, but works with correlation.

Subset Selection

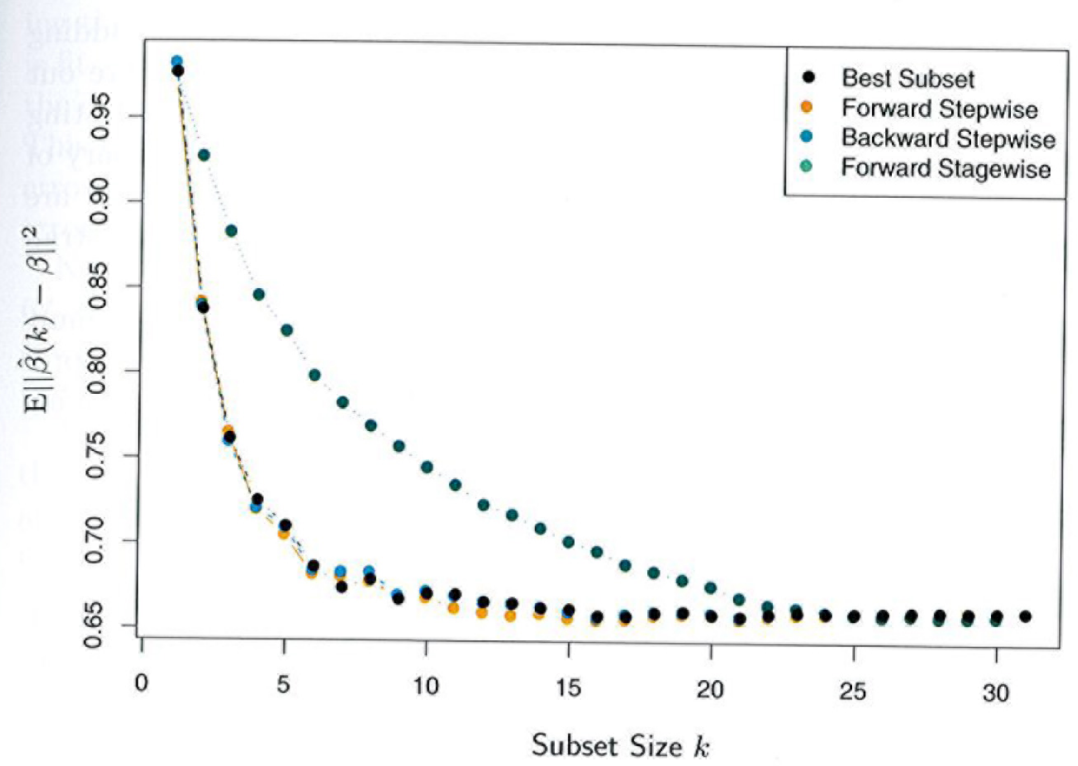

FIGURE 3.6. *Comparison of four subset-selection techniques on a simulated linear regression problem* $Y = X^T\beta + \varepsilon$. *There are* $N = 300$ *observations on* $p = 31$ *standard Gaussian variables, with pairwise correlations all equal to* 0.85. *For* 10 *of the variables, the coefficients are drawn at random from a* $N(0, 0.4)$ *distribution; the rest are zero. The noise* $\varepsilon \sim N(0, 6.25)$, *resulting in a signal-to-noise ratio of* 0.64. *Results are averaged over* 50 *simulations. Shown is the mean-squared error of the estimated coefficient* $\hat{\beta}(k)$ *at each step from the true* β.