# Linear Methods for Regression: Subset Selection

Statistical methods for data analysis – Machine learning

Alberto Castellini
University of Verona

# Motivation

Possible **problems of Least Squares Estimation** (LSE):

- **Prediction accuracy**:
  - Low bias, large variance
  - Can sometimes be improved by shrinking. Sacrifice a bit of bias but reduce variance

- **Interpretation:**
  - Identification of a small subset of variables with the strongest effect

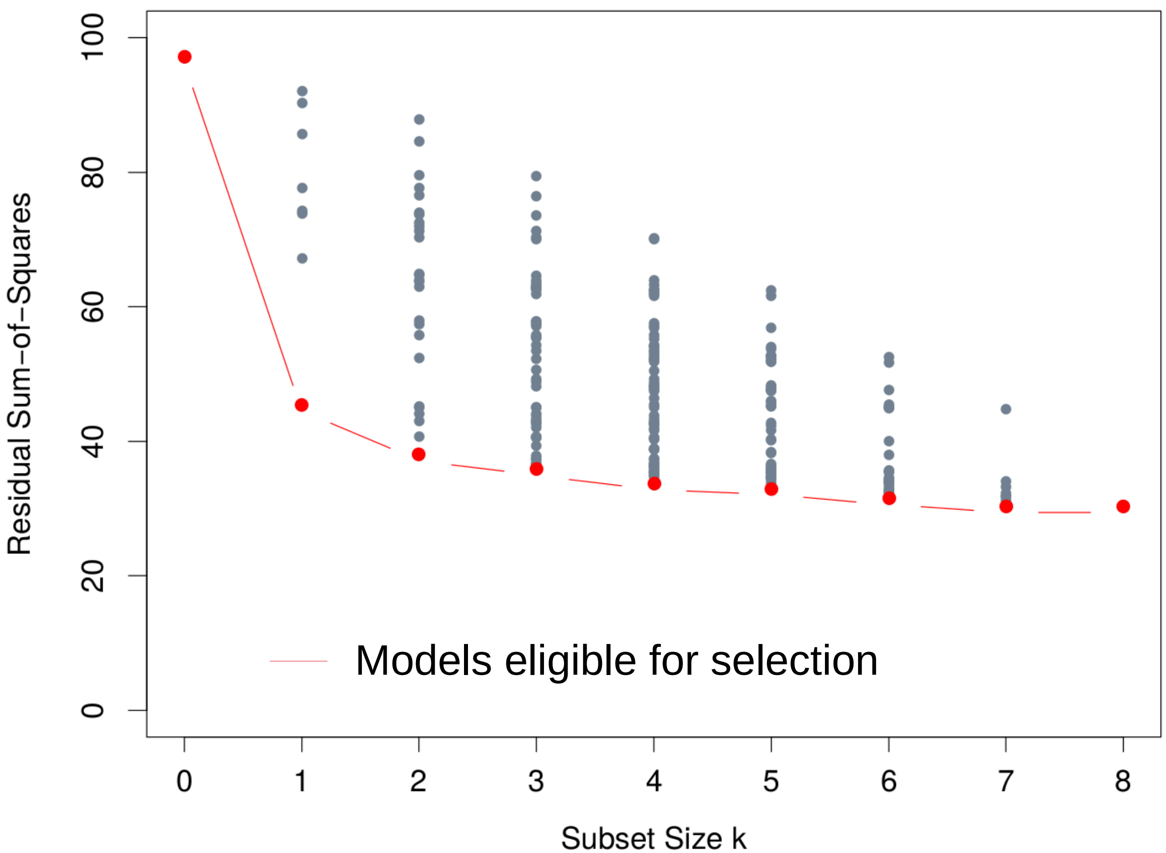
Solution: **Model selection**

- Here we describe different **strategies** to **variable subset selection** with **linear regression.**

- In next lectures **shrinkage** and **dimension-reduction** approaches for controlling variance.

# *Best-subset selection*

- Finds for each **k={0,1,2,…,p}** the subset of size *k* that gives **smaller Residual Sum of Squares.**
- **Leaps and bounds** procedure (Furnival and Wilson, 1974): feasible for p as large as 30 or 40.
- RSSs of all subset models for the prostate cancer example:



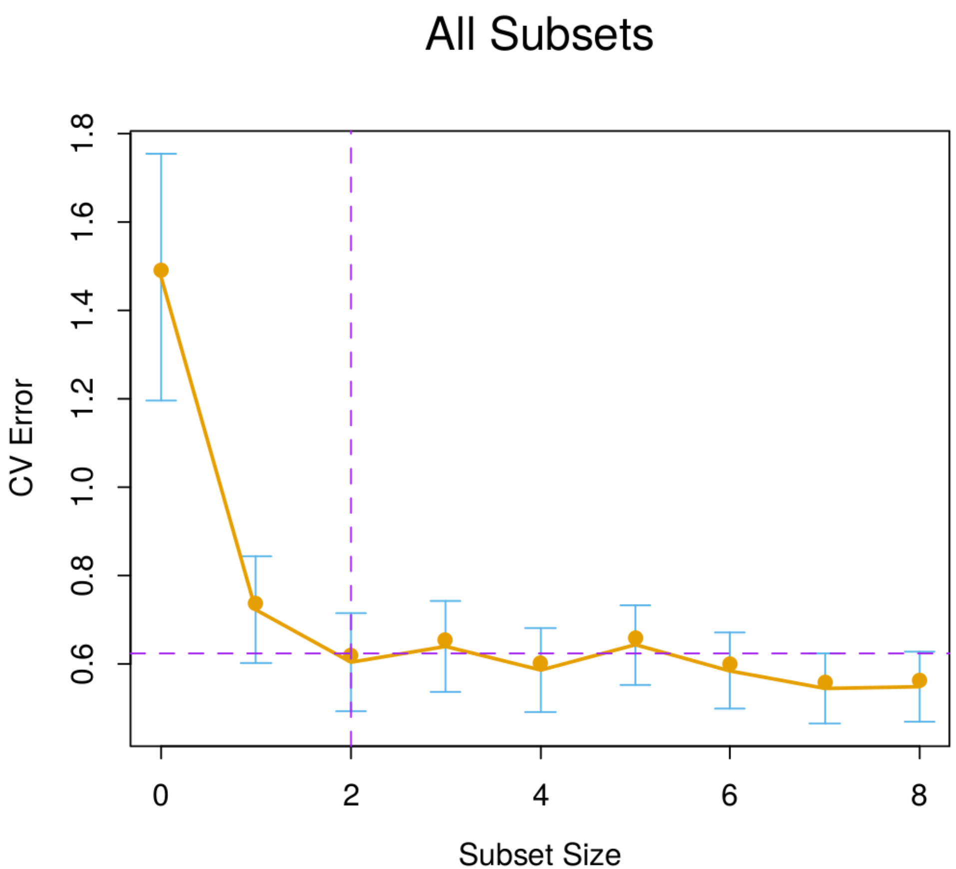

- Best subset of size 2 need not include the variables in the best subset of size 1
- Best-subset curve is necessary decreasing. It cannot be used to select the subset size k

# General idea of best-subset selection

Use the training data to produce a sequence of models varying in complexity and indexed by a single parameter.

- Cross-validation and the AIC criterion (presented in next lectures) can be used to estimate the best parameter k.

| Term | LS | Best Subset |
|---|---|---|
| Intercept | 2.465 | 2.477 |
| lcavol | 0.680 | 0.740 |
| lweight | 0.263 | 0.316 |
| age | −0.141 | |
| lbph | 0.210 | |
| svi | 0.305 | |
| lcp | −0.288 | |
| gleason | −0.021 | |
| pgg45 | 0.267 | |
| Test Error | 0.521 | 0.492 |
| Std Error | 0.179 | 0.143 |



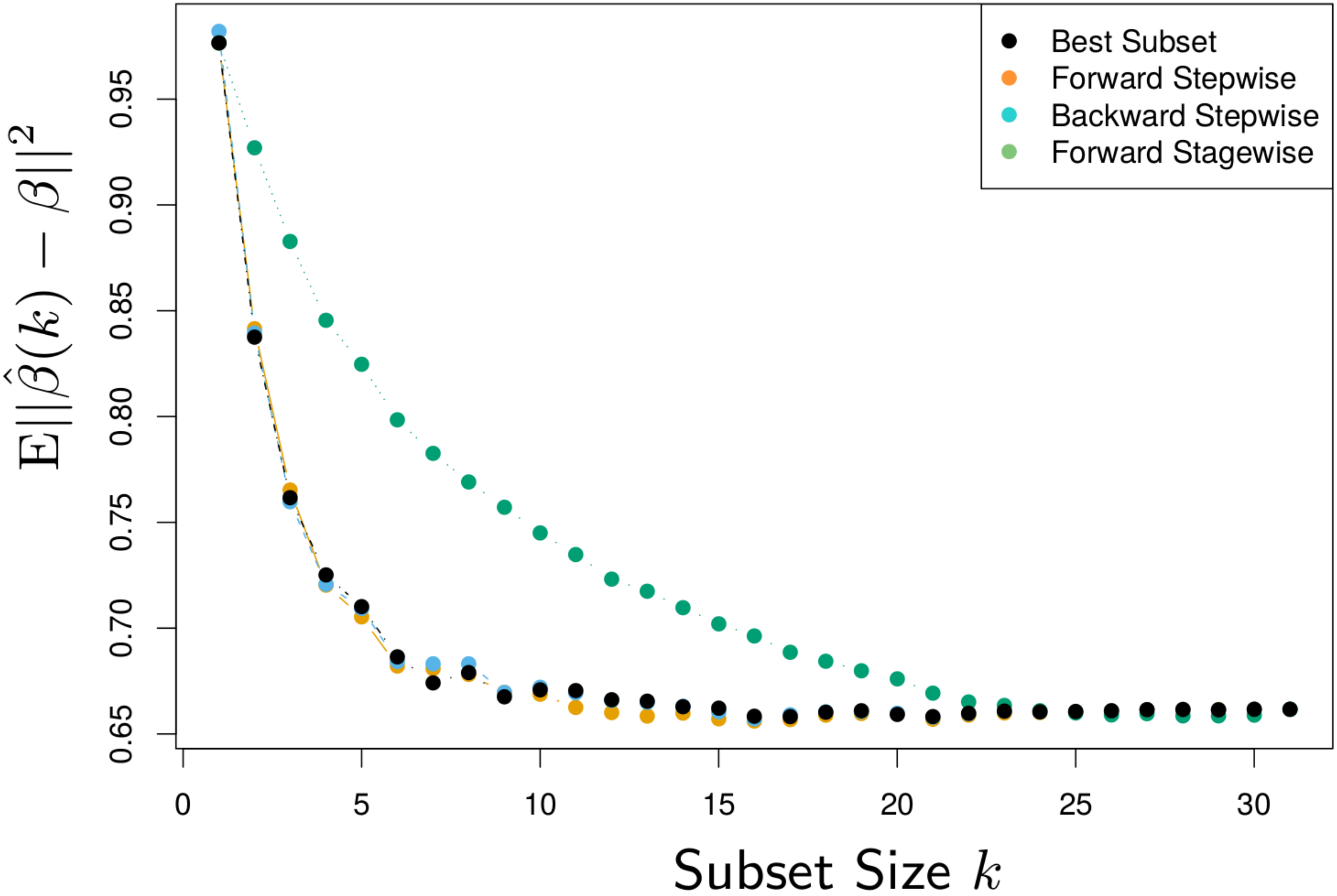
# *Forward-Stepwise Selection*

- Search all possible subsets is **infeasible for large p**, hence we seek a **good path** through them.
- **Forward-stepwise selection:**
  - starts with the **intercept**
  - sequentially **adds** into the model the predictor that **most improve** the fit
  - produces a sequence of models indexed by k, the subset size
  - is a **greedy algorithm**, producing a **nested sequence of models**
  - is **suboptimal** compared to best-subset selection
  - is applicable with **large p**
  - has **lower variance** but perhaps **higher bias**

# *Backward-Stepwise Selection*

- **Backward-stepwise selection:**
  - starts with the **full model**
  - sequentially **deletes** the predictor that has the **least impact** on the fit
  - the **candidate** for dropping is the variable with the **smallest Z-score**
  - can only be used when $N > p$
  - produces a sequence of models indexed by k, the subset size
  - is a **greedy algorithm**, producing a **nested sequence of models**
  - is **suboptimal** compared to best-subset selection
  - is applicable with **large p**
  - has **lower variance** but perhaps **higher bias**

# Comparison




On the prostate cancer example, best-subset, forward and backward selection all gave exactly the same sequence of terms.

# *Hybrid stepwise selection strategies*

- Hybrid stepwise-selection strategies consider **both forward and backward moves** at each step, and **select the “best” of the two**.
- The R function called *step* uses the Akaike (**AIC**) criterion for weighting the choices, i.e., at each step an add or drop is performed that minimizes the AIC score.
- Notice that **standard errors** of coefficients in non-full models are not valid since they do not account for the search process.
  - **Bootstrap** techniques (presented in next lectures) can be used to solve this problem

## *Exercise: Prediction on the prostate cancer dataset*

See text of Exercise 3

# References

[Hastie 2009] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (second edition). Springer. 2009.