

Lecture 7: Linear regression variable selection

CSI 5v93: Introduction to machine learning

Baylor University
Computer Science Department

Dr. Greg Hamerly
<http://cs.baylor.edu/~hamerly/>

CSI 5v93: Introduction to machine learning, Lecture 7 – p. 1/16

Announcements

- Homework 2 due February 8th – see updates on web

CSI 5v93: Introduction to machine learning, Lecture 7 – p. 2/16

Questions?

CSI 5v93: Introduction to machine learning, Lecture 7 – p. 3/16

Chapter 3: Linear methods for regression

- 3.1 – Introduction
- 3.2 – Linear regression models and least squares
- 3.3 – Multiple regression from simple univariate regression
- 3.4 – Subset selection and coefficient shrinkage
- 3.5 – Computational considerations

CSI 5v93: Introduction to machine learning, Lecture 7 – p. 4/16

Subset selection and coefficient shrinkage (3.4)

There are two reasons why least-squares estimates are not totally satisfactory:

- prediction accuracy – many parameters means that the model may have *high variance*; by reducing the number of parameters, we may reduce the variance (but increase the bias). This may help the model to do better on unseen data.
- interpretability – a model with many parameters is not as easy to understand as a model with few parameters.

Your book introduces two different methods of reducing the number of parameters:

- subset selection – choosing only a subset of the variables
- coefficient shrinkage – causing the parameters of less important variables to go to zero

Subset selection

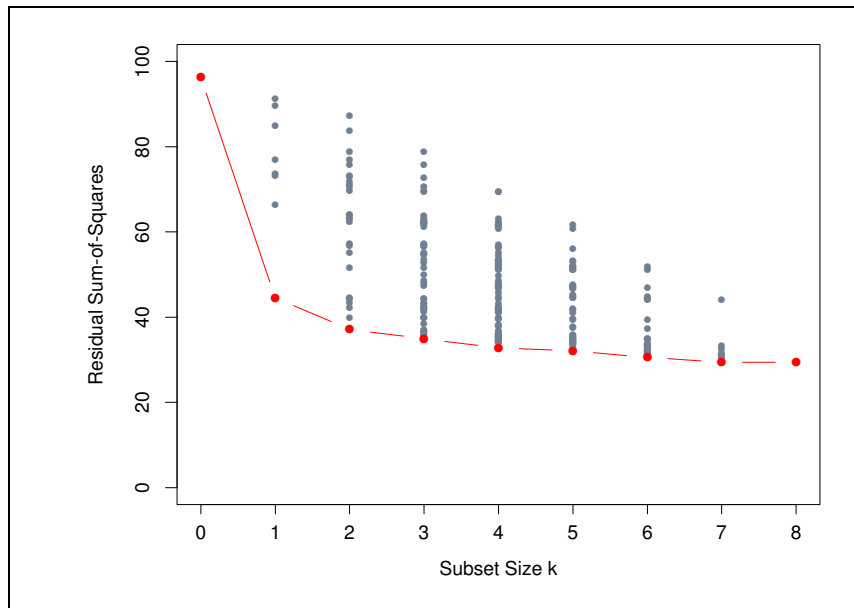
Consider subsets of size 1, 2, 3, ..., d of d variables, and choose the subset that gives a small error.

Considering all possible subsets is computationally infeasible for large d ; the number of subsets goes up exponentially.

Choosing the subset

The RSS will always be smallest for the model with the most parameters (all d parameters), so this is not very useful.

RSS for different subsets for the cancer-prediction problem:



CSI 5v93: Introduction to machine learning, Lecture 7 – p. 7/16

Stepwise selection

Rather than consider every possible subset, stepwise selection adds one variable at a time.

Forward stepwise selection:

- start with one parameter (the intercept)
- add the next best parameter (using the F statistic to determine the best)
- repeat until the change in the F statistic is not significant (with some probability, e.g. 95%)

Backwards stepwise selection is similar, but works in reverse.

CSI 5v93: Introduction to machine learning, Lecture 7 – p. 8/16

Issues with stepwise selection

What are some drawbacks with stepwise selection? Is forward or backward easier? Do you expect they will produce the same results?

Coefficient shrinkage

Subset selection eliminates variables completely by testing if their coefficients are sufficiently close to zero.

This is a combinatorially difficult problem to try all subsets.

An alternative is to “shrink” the coefficient values so that variables become less important.

This can turn the combinatorial, discrete problem of subset selection into a smoother problem which can be optimized easily.

This also solves the problem of correlated variables whose coefficients negate one another and have high variance.

Ridge regression

Ridge regression adds a squared penalty to each coefficient:

$$\begin{aligned} &RSS(\beta) + \lambda \sum_{j=1}^d \beta_j^2 \\ &= \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^d x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^d \beta_j^2 \end{aligned}$$

So we want to choose the β that minimizes the above:

$$\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^d x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^d \beta_j^2 \right\}$$

Ridge regression penalty

$$\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^d x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^d \beta_j^2 \right\}$$

Note that the β_0 is not penalized. Why?

Operational issue: center the data before using ridge regression.

Matrix solution for ridge regression

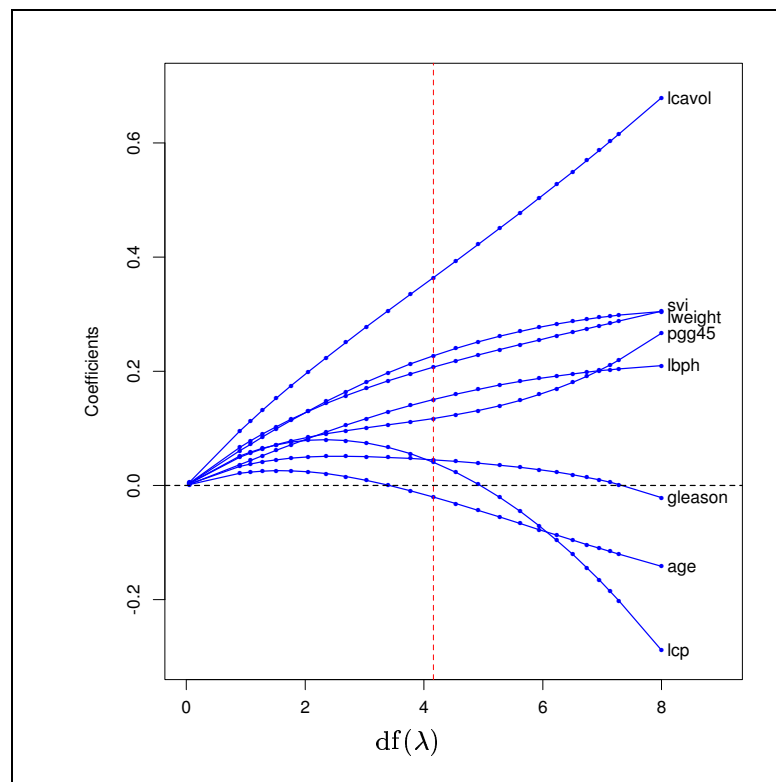
$$\begin{aligned}RSS(\lambda, \beta) &= \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^d x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^d \beta_j^2 \\ &= (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^T \beta\end{aligned}$$

This leads to:

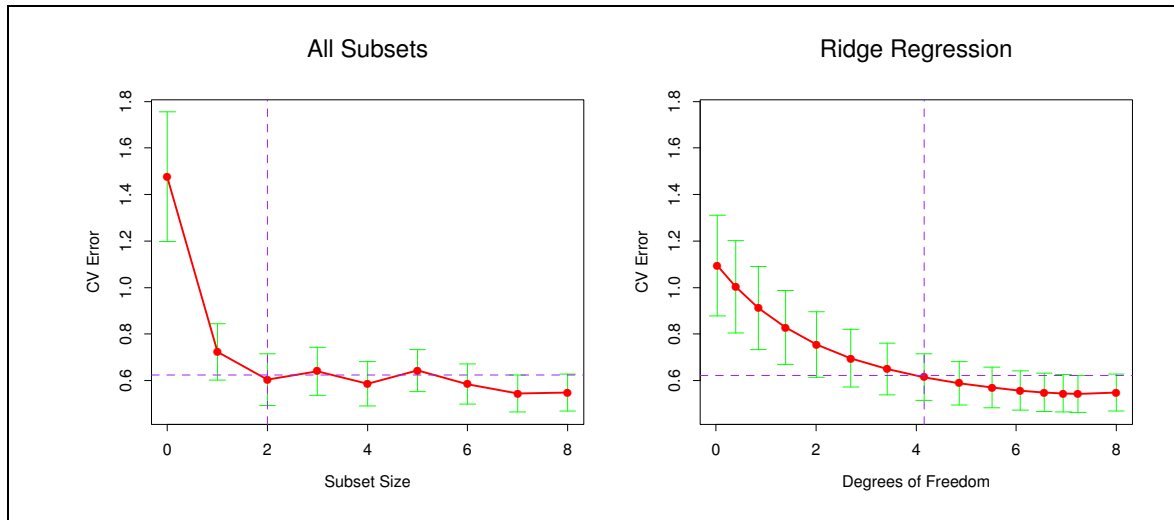
$$\hat{\beta} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

Where \mathbf{I} is the $d \times d$ identity matrix.

Example of ridge regression



Examples of variable selection



Error bars are obtained by cross-validation.

The least complex model within 1 standard error of the best average error is chosen.

CSI 5v93: Introduction to machine learning, Lecture 7 – p. 15/16

2-minute journal

Please write a response to the following on a piece of paper and hand it in immediately. Please make it anonymous (no names). Write about:

- major points you learned today
- areas not understood or requiring clarification

CSI 5v93: Introduction to machine learning, Lecture 7 – p. 16/16