Penalized regression for feature selection

Brian Kissmer

USU Department of Biology

Nov. 7th, 2024

Learning objectives

1. Understand the problem of having too many covariates
2. Be able to understand how LASSO regression solves this problem
3. Know how to implement LASSO in R

Today’s outline

1. Over-parameterization and feature selection
2. LASSO regression
3. R packages
4. LASSO regression in R

The problem of too many covariates

Sometimes you can have too many covariates, especially in observational studies

- ➢ Linking climatic factors to demographic patterns
- ➢ Linking genotype to phenotype

The problem of too many covariates

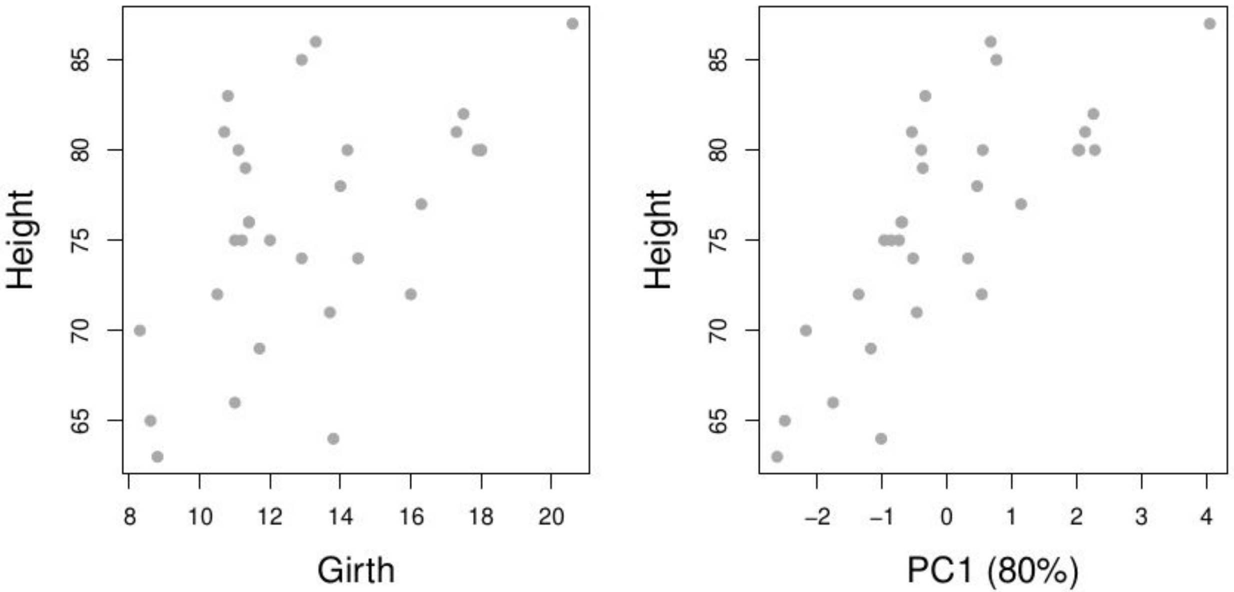
1. r^2 necessarily goes up with more covariates, but predictive power goes down
2. Can at most estimate $N - 1$ regression coefficients (r^2 will be 1.0)
3. With more than $N - 1$ covariates, standard regressions do not work

Solutions to the too many covariates problem

What to do when you get too many covariates:

1. Get rid of some
2. Use an ordination approach to project covariates to a lower-dimensional space
3. Use a step-wise regression
4. Use a form of penalized regression, such as LASSO

Use ordination to reduce number of covariates



PC1 captures 80% of the variation in tree volume, height, and girth; overall measure of ‘tree size’

Stepwise regression to add or remove covariates

- **Forward stepwise:**
  - Start from a simple model and iteratively add covariates that most improve fit
- **Backward stepwise**:
  - Start from a full model (but still fewer than *N* covariates and remove covariates that least improve fit

Penalized or regularized regression

Model fit is a compromise between improving fit and a penalty for more and bigger regression coefficients

- Start with all possible covariates
- “Shrink” some regression coefficients to 0 (remove them)
- Non-zero coefficients are selected as those that matter for the model

Least absolute shrinkage and selection operator (LASSO)

LASSO is a regression analysis method that selects and regularizes (shrinks) coefficients to increase the predictive power of the model

Goodness of fit

$$S = \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki})\right)^2$$

Least absolute shrinkage and selection operator (LASSO)

LASSO is a regression analysis method that selects and regularizes (shrinks) coefficients to increase the predictive power of the model

Penalty

$$\lambda||\beta||_1 = \lambda \sum_{k=0}^{K} |\beta_k|$$

Least absolute shrinkage and selection operator (LASSO)

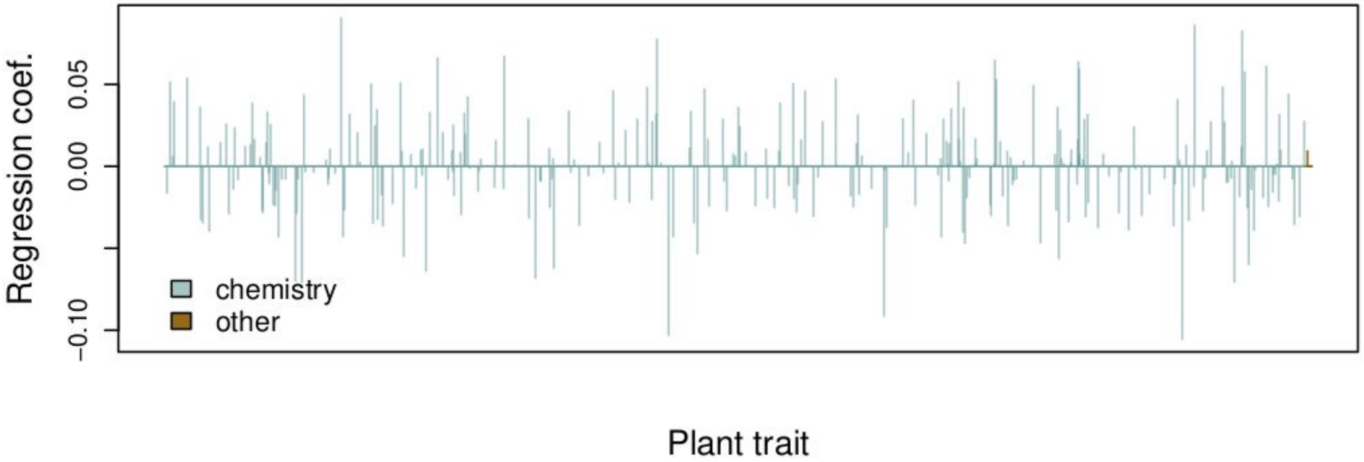
LASSO is a regression analysis method that selects and regularizes (shrinks) coefficients to increase the predictive power of the model

Overall fit

$$\min\left(\frac{1}{n}S + \lambda||\beta||_1\right)$$

LASSO estimates of regression coefficients

Caterpillar survival as a function of 1760 plant traits based on ~1000 data points



~ 200 covariates retained with non-zero effects

How do you estimate the regression coefficients?

λ denotes the strength of the penalty for non-zero regression coefficients

$$\min\left(\frac{1}{n}S + \lambda||\beta||_1\right)$$

We chose a value of λ to maximize prediction accuracy with cross-validation

Hypothesis testing with linear regression models

Divide data into training and testing sets, estimate coefficients from training data set but evaluate performance on test set

LASSO in R

See the handout on installing packages and performing LASSO regression in R