Glmnet

One of the best algorithms in machine learning today.

Outline of Talk

- Background
  - Ordinary Least Squares Regression (and logistic regression)
  - Limitations of OLS
  - Introduce regularization – (coefficient shrinkage)
- Principles of Operation
- Examples of Results

Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.

Glmnet – Background - modern linear regression – a lot has changed in 200 years

Ordinary Least Squares Regression – First described by Gauss 1794
Seeks to fit a straight line through data so as to minimize sum squared error.

Example: How do men's annual salaries depend on their height?

Here's some data: Men's average annual salaries (reported 2003, inflation adjusted to 2000)

Average Salary: $40,426

Standard Deviation of Salary: $34,334

Dependence on Height (from "blink" by Malcolm Gladwell)

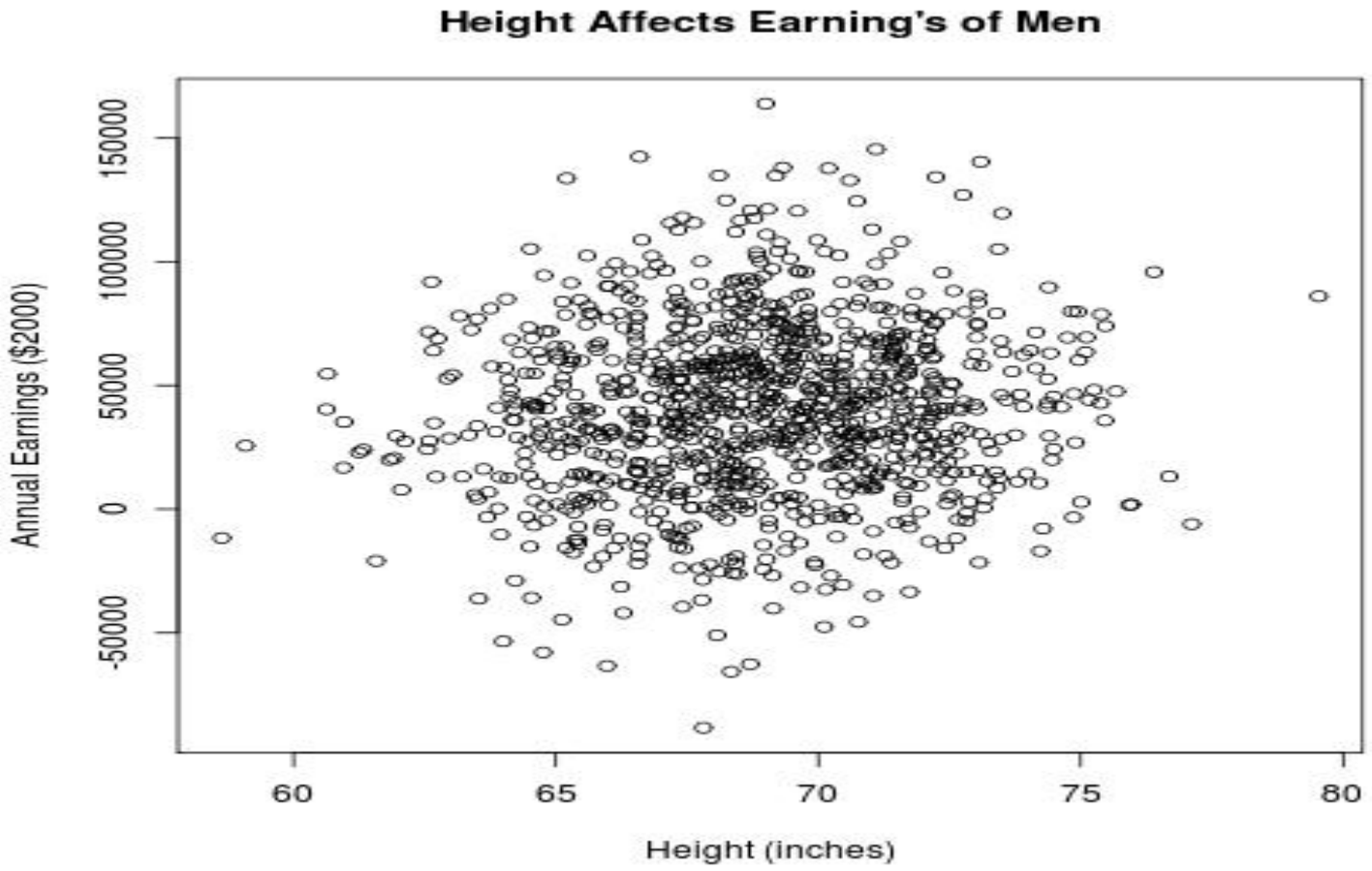
$789/yr in earnings per inch in height.

Using these data, generate some numbers for a plot.

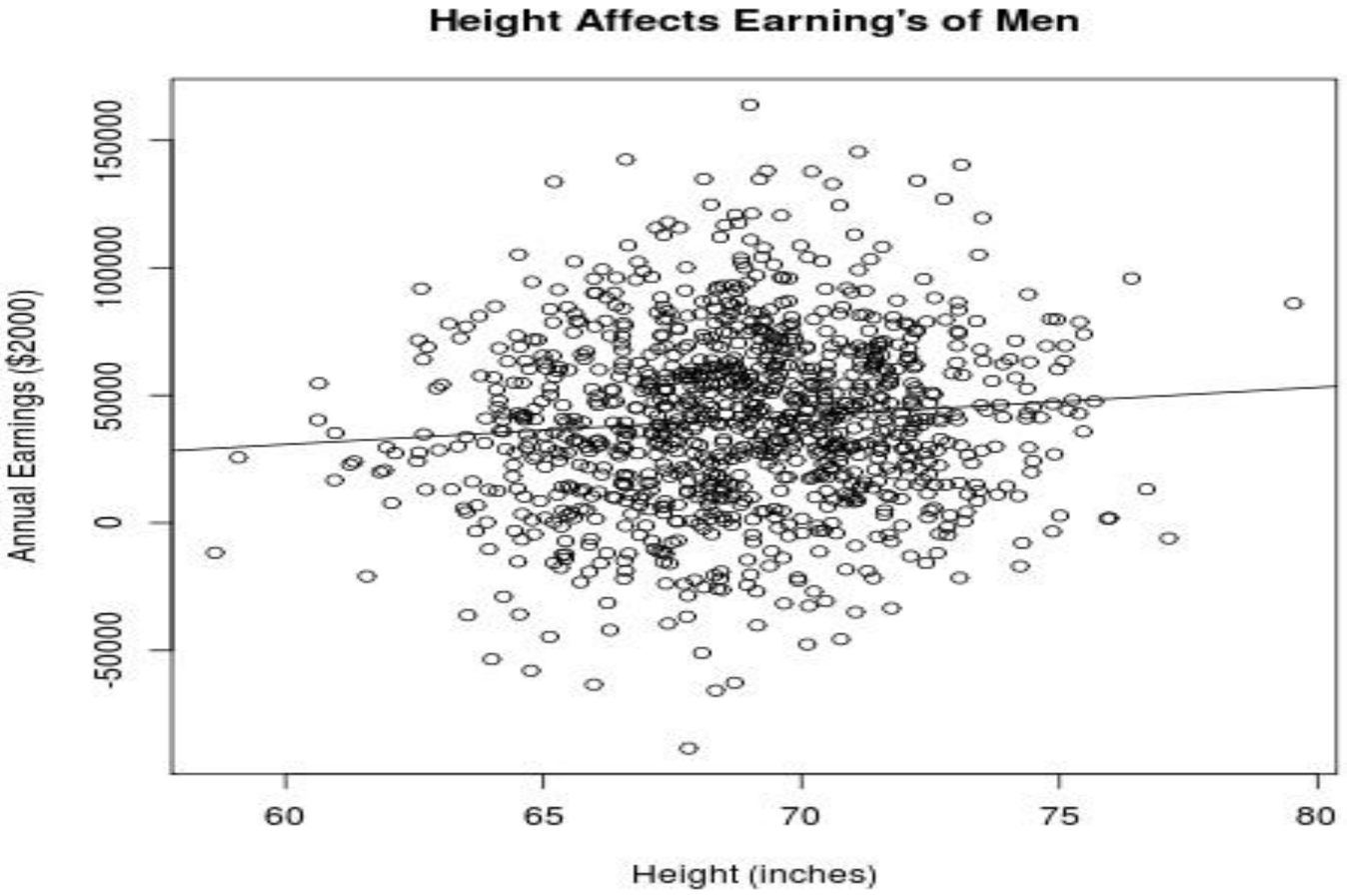
Refs. GAO Congressional Report, Nov 20, 2003, http://www.gao.gov/htext/d0435.html

"blink", Malcolm Gladwell, excerpt from http://www.gladwell.com/blink.excerpt2.html

Plot of Men's Yearly Earnings

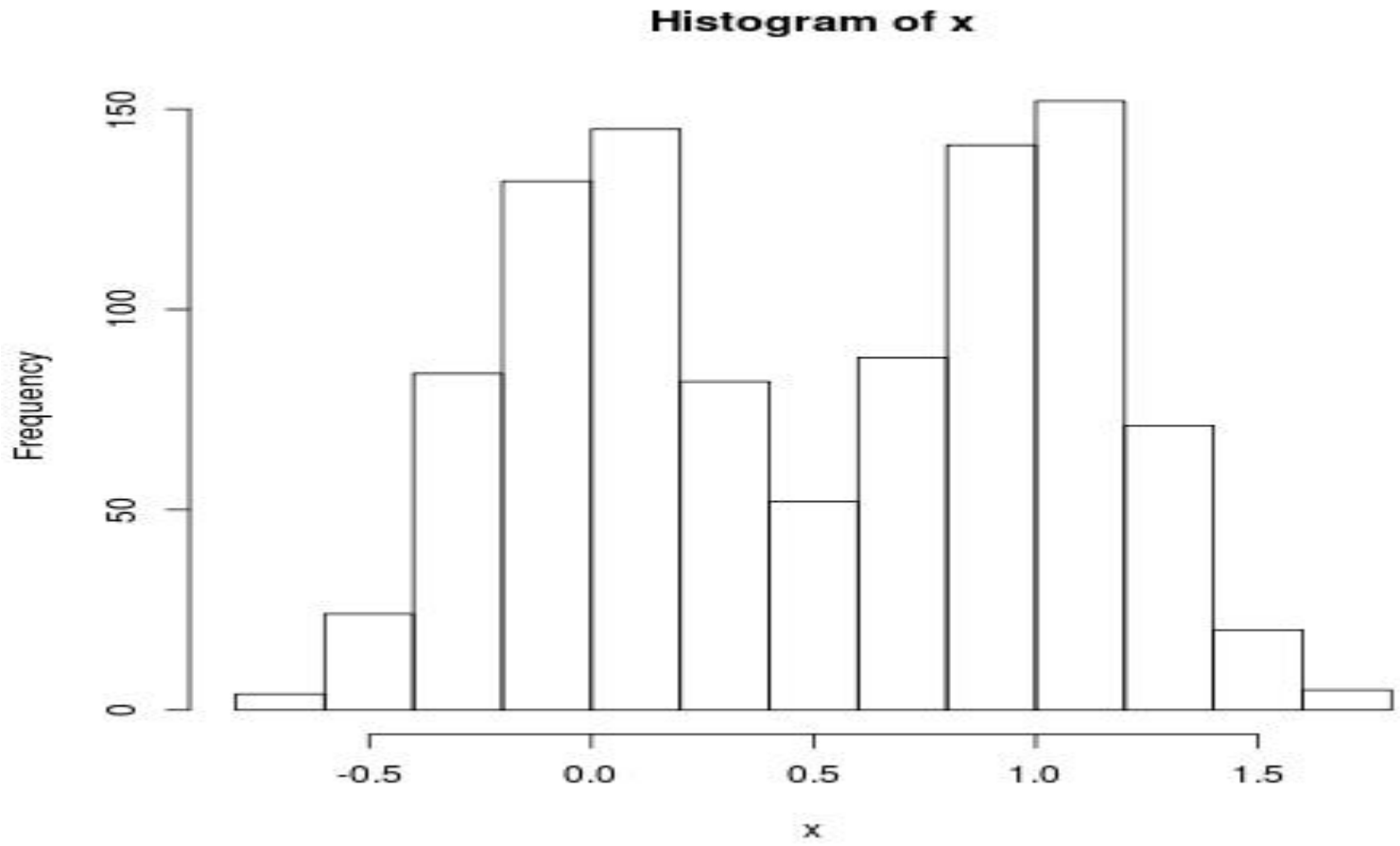


Regression Line Fit to Data

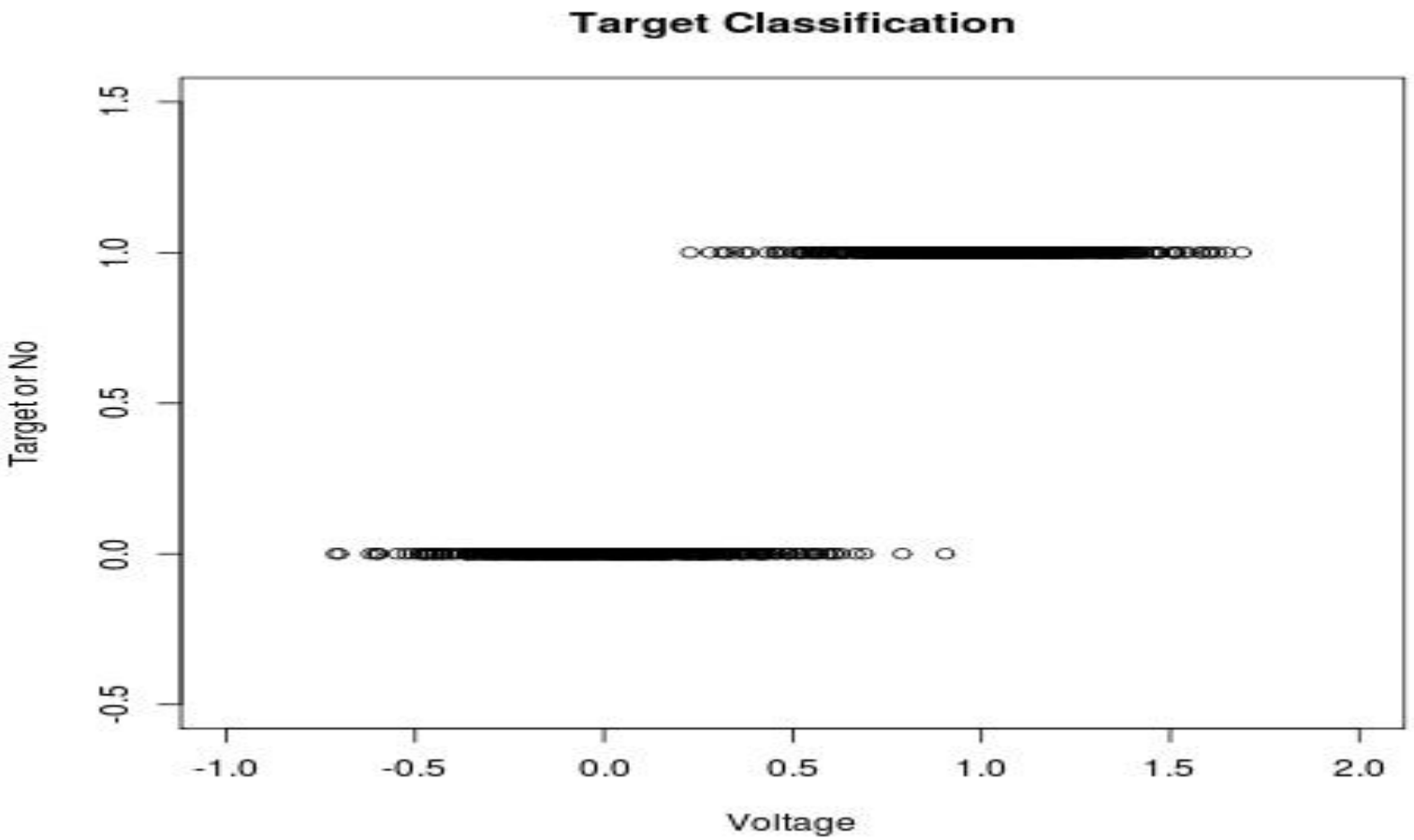


Regression for Classification Problems

Example: Target Detection – Detector yields 1 volt if target is in view, 0 volts otherwise. Additive shot noise.

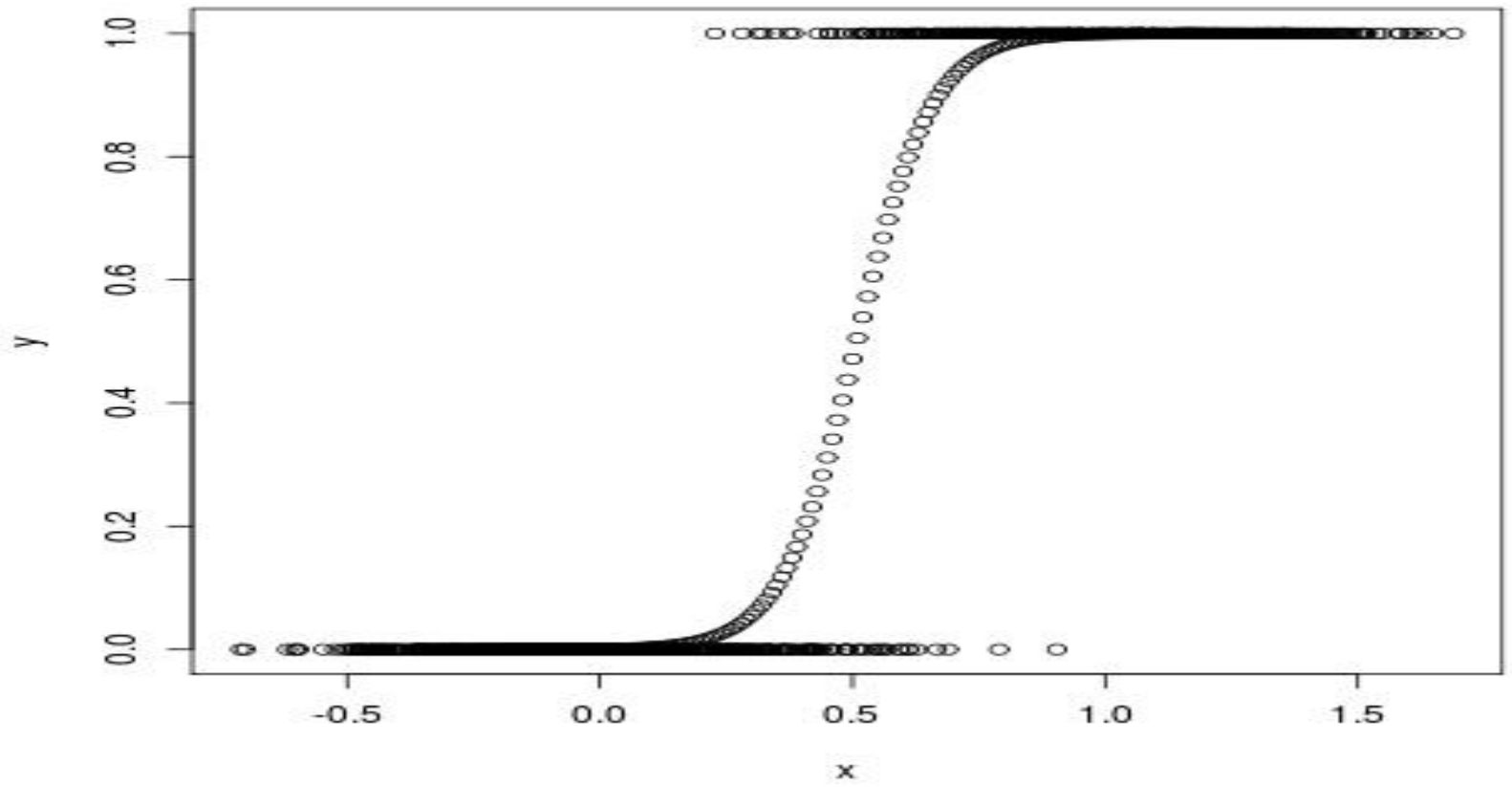


Target Classification with Labels



Could use OLS apparatus – treat class outcome as 0,1 and fit straight line.

Logistic regression



Dealing with Overfitting

These plots showed how it works when we have 1000 data points. –Suppose we only have a few..

- Plots for 2, 10, 100.

Several Conclusions:

Need methods to

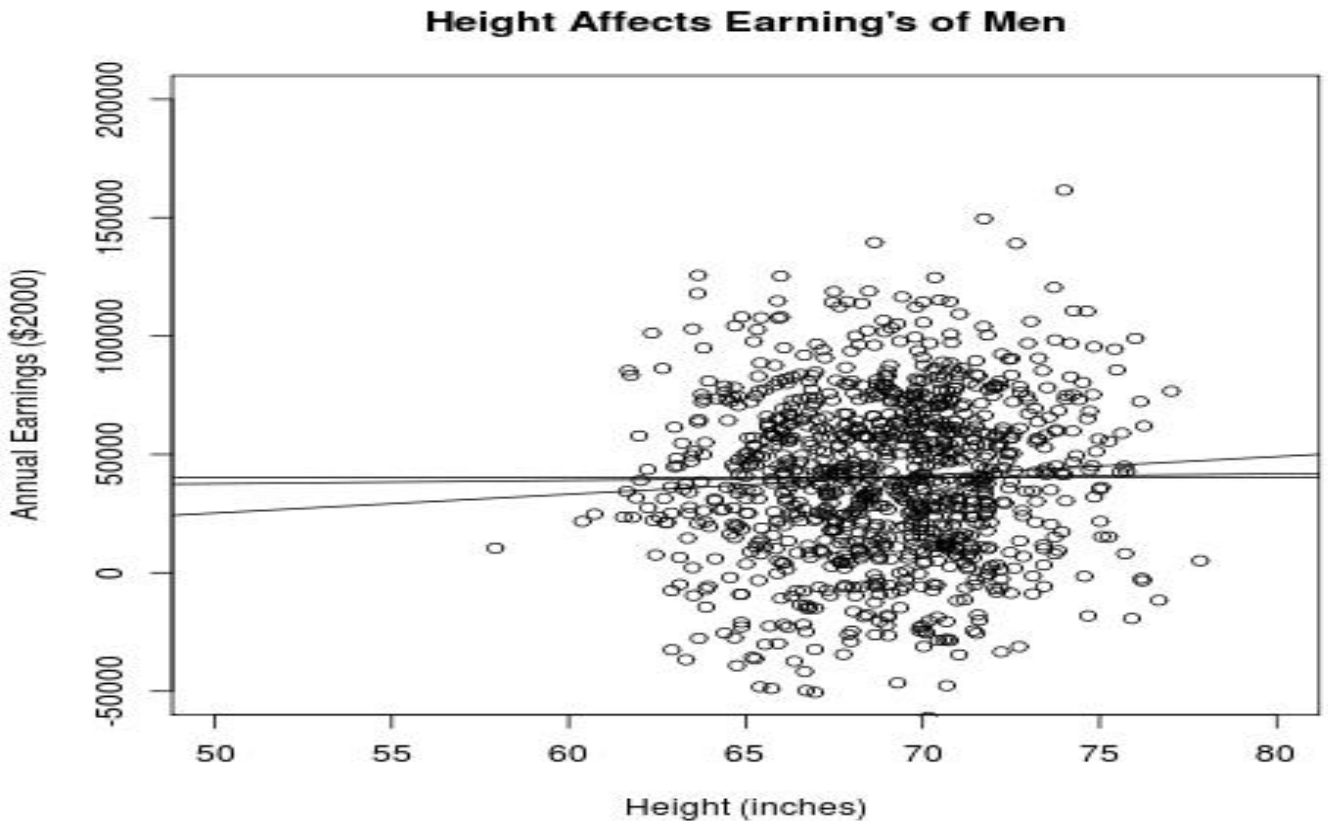
Determine if we're overfitting

Hedge our answers if we are overfitting

More data is better (as long as we can computer answers).

Coefficient shrinkage methods

With ordinary least squares regression, objective is to minimize the sum squared error between actual values and our linear approximation.
Coefficient shrinkage methods add a penalty on coefficients.
Yields a family of regression solutions – from completely insensitive to input data to unconstrained ordinary least squares



Why add this complication?

Eliminate over-fitting
Tune the model to achieve best generalization – best performance on new data.
Pick member of solution family which gives best performance on held out data (i.e. data not included in training set).

A couple of important details:

1. Bias value is not included in the coefficient penalty
2. Inputs (attributes) must be scaled. Usual choice is
   mean = 0.0
   standard deviation = 1.0

Coefficient Penalty Function

With more than one attribute the coefficient penalty function has important effect on solutions.
Some choices of coefficient penalty give control over sparseness of the solutions.
Two types of coefficient penalty are most frequently used:

- Sum of squared coefficients.
- Sum of absolute value of coefficients.

glmnet algorithm incorporates Elastic Net penalty introduced by Zou and Hastie. – Weighted combination of sum squares and sum of absolute values.

Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.

Hui Zou and Trevor Hastie (2005) Regularization and variable selection via the elastic net, *J. R. Statist. Soc.* B (2005) 67, *Part* 2, *pp.* 301–320.
http://www.stanford.edu/~hastie/Papers/B67.2%20%282005%29%20301-320%20Zou%20&%20Hastie.pdf

Coefficient Penalty Function Choices

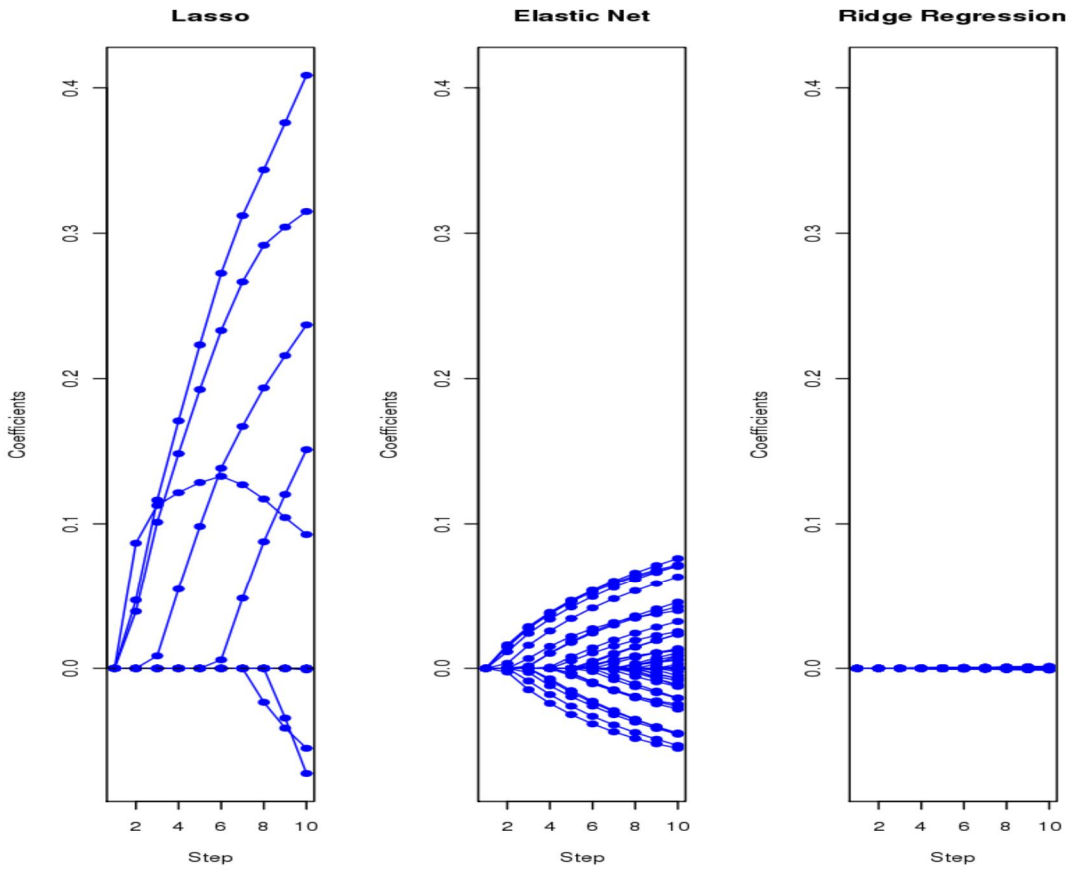


Figure 1: Leukemia data: profiles of estimated coefficients for three methods, showing only first 10 steps (values for λ) in each case. For the elastic net, α = 0:2.

(This figure comes from Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.

Glmnet algorithm

Suppose:

- Coefficients (not including bias) are arranged in a vector w.
- Coefficient penalty is $\lambda P(w)$, where:
  - P(w) is sum squares or sum of absolute values of elements of w (or a combination)
  - $\lambda \geq 0$ is a parameter that determines the severity of the coefficient penalty function

Adding the coefficient penalty changes the minimization problem from

- Minimize(squared fit error) → Minimize(squared fit error + $\lambda P(w)$)

We no longer have Gauss's closed form solution. That's where glmnet algorithm comes in.

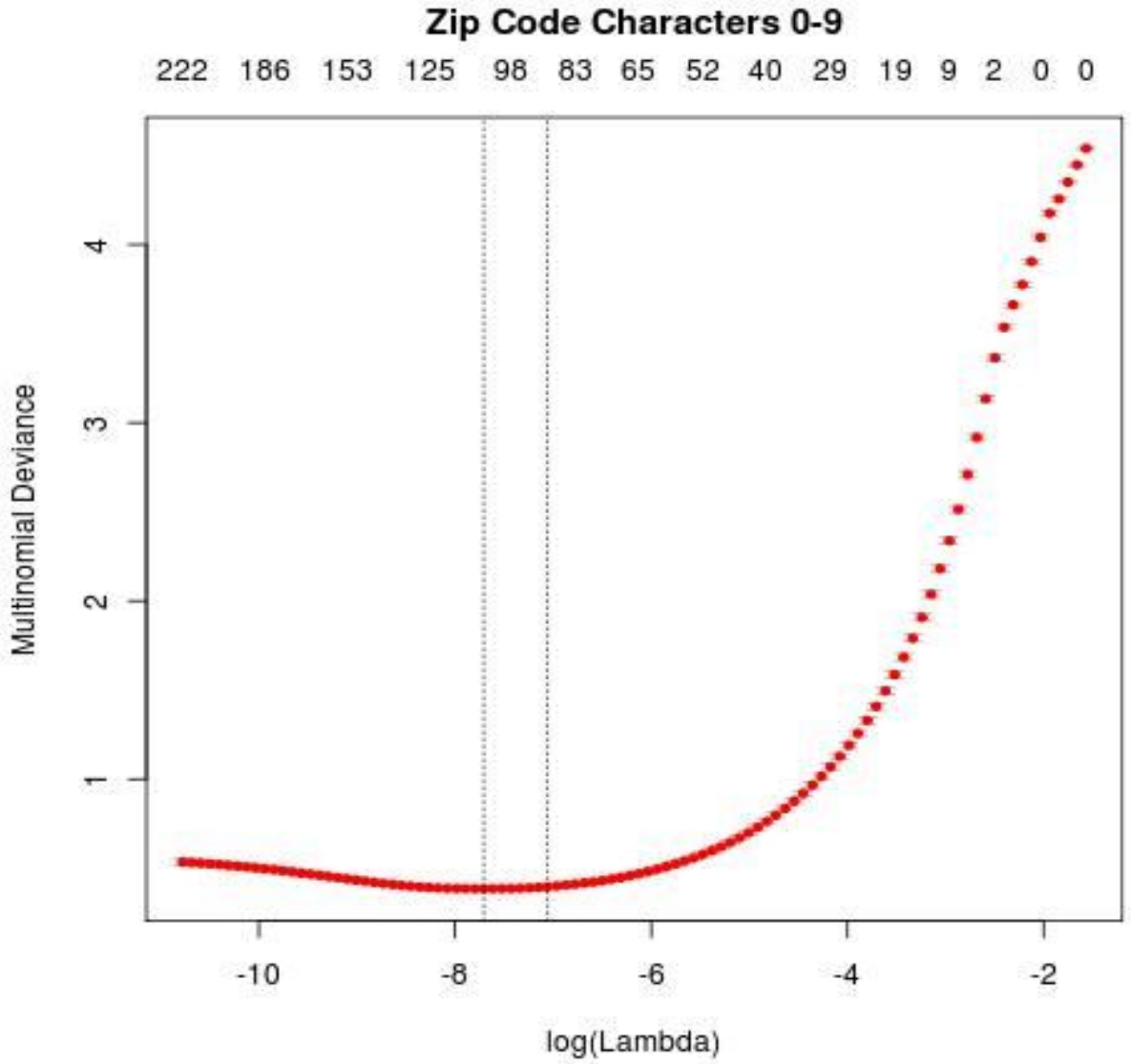
Basic idea of glmnet algorithm:

- If the coefficient penalty is very heavy (λ very large) then:
  - All coefficients = 0.0
- A small change in λ gives a small change in the minimizing value of w.
- Use coordinate descent to track incremental changes in optimum w.

Authors demonstrate speed advantage on real datasets that range from x4 to more than x100

- **-Speed advantage more pronounced on wide attribute spaces and/or large data sets.**

Examples
Zip code data



Advantages of glmnet

Relatively simple implementation – (after training) evaluation requires multiply and sum for each attribute. Even for wide attribute spaces this is manageable in real time – e.g. text processing - spam filtering, POS, NER