Statistical learning - Regression

Rita Almeida

PhD course 3045, VT 2018

Linear regression

Describes the relationship between an outcome variable Y and a set of predictor variables Xs:

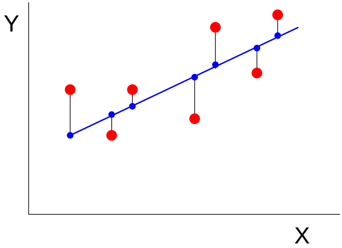
$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- This model is usually fitted by using least squares.

Minimizing the residual sum of squares:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$$

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$



Prediction accuracy and interpretability

Alternative fittings can be better.

- Prediction accuracy
  - Bias: if the relations are approximately linear, least squares have low bias.
  - Variance:
    - if $n \gg p$ least squares estimates have low variance;
    - if n not much larger than p least squares estimates can have large variance, leading to overfitting and poor predictions.
  - Methods constraining or shrinking the estimates can perform much better.

- Interpretability
  - Several Xs might not be associated with Y and just add to the complexity of the model.
  - Methods for automatic variable selection can lead to much more interpretable models.

Some alternatives

- Subset selection:
  1. Identify a subset of predictor variables.
  2. Use least squares on the reduced set of variables.

- Dimension reduction: creating new (fewer) predictors from the original ones.
  1. Project the original p predictors to a subspace with dimensions m, with $m < p$.
  2. Use least squares on a model with the new variables.

- Shrinkage or regularization: fitting a model with all p predictors but using a method that shrinks the estimates to 0.
  - Reduces variance of the estimates.
  - Can select variables by shrinking some estimates to 0.

Ridge regression

The least squares procedure estimates the parameters by minimizing the residual sum of squares:

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2$$

Ridge regression minimizes the same function plus a shrinkage penalty:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

with tuning parameter $\lambda \geq 0$.

Or in another formulation, minimizing RSS subject to:

$$\sum_{j=1}^{p}\beta_j^2 \leq s$$

Ridge regression

- $\lambda = 0$ least square estimates
- $\lambda \to \infty$ all estimates approach 0
- An appropriate λ has to be set!
  - Cross-validation

Note: use standardized predictors.

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

Example:

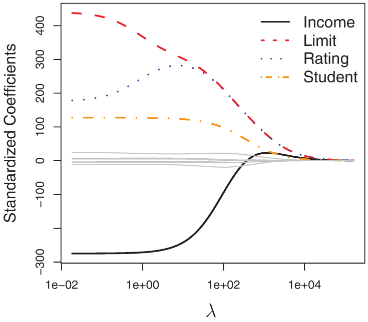


Figure adapted from James et al.

Ridge regression

- Prediction accuracy:
  - RR results in increased bias but decreased variance.
  - RR works well when least squares estimates have high variance.

Simulation, $n = 50, p = 45$

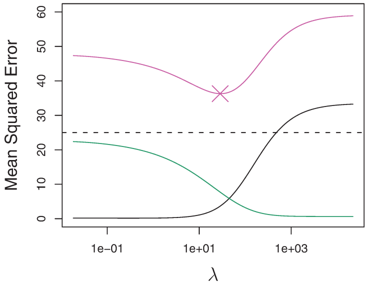
Black: squared bias

Green: variance

Pink: test mean squared error

Dashed: minimum possible MSE

Figure adapted from James et al.



- Interpretability:
  - RR indicates important variables and is computationally faster than variable selection.
  - However the model always keeps all the p variables!

Lasso

Lasso: least absolute shrinkage and selection operator

- Lasso is similar to RR, but it does variable selection.
  - Some parameters will be shrunk to 0.

Lasso minimizes RSS plus a different shrinkage penalty:

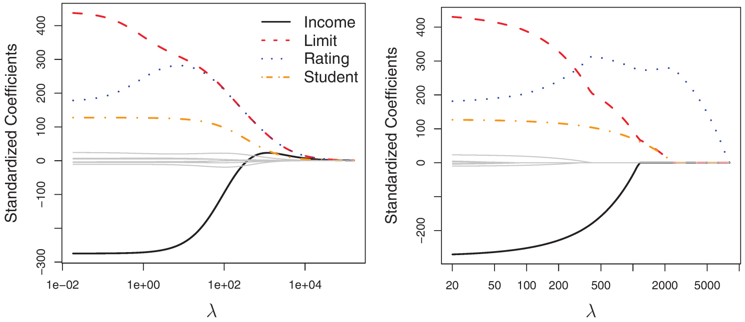
$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

with tuning parameter $\lambda \geq 0$.

Or in another formulation, minimizing RSS subject to:

$$\sum_{j=1}^{p} |\beta_j| \leq s$$

Ridge regression versus Lasso



- Lasso is expected to be better when a small number of predictors have large effects and the rest has almost no effect.
- Ridge regression is expected to be better when many predictors are all weakly related to the outcome variable.

Schizophrenia Research: Cognition 11 (2018) 1–5

Contents lists available at ScienceDirect

Schizophrenia Research: Cognition

journal homepage: www.elsevier.com/locate/scog

Research Paper

Model selection and prediction of outcomes in recent onset schizophrenia patients who undergo cognitive training

Ian S. Ramsay[a,*], Sisi Ma[b], Melissa Fisher[a], Rachel L. Loewy[c], J. Daniel Ragland[d], Tara Niendam[d], Cameron S. Carter[d], Sophia Vinogradov[a]

[a] University of Minnesota, Department of Psychiatry, United States
[b] University of Minnesota, Department of Medicine, United States
[c] University of California, San Francisco, Department of Psychiatry, United States
[d] University of California, Davis, Department of Psychiatry, United States

Example

Question: what are the predictors of training improvements?

- Computerized training:
  - aiming at improving speed and accuracy auditory processing, while performing auditory and verbal working-memory tasks;
  - 8 weeks, 1 hour/day, 5 days/week;
  - adaptive training to 80-85% accuracy.
- Subjects: 43 individuals with recent onset of schizophrenia

Example - regression analysis

Predictor variables:

- GFR: Global functioning role
- GFS: Global functioning social
- Strauss: social contact, hospitalizations, engagement in school or work
- FSIQ: verbal reasoning

Measure	Linear regression			
	Estimate	Std. error	*t*-Value	*p*-Value
(Intercept)	− 0.84	1.49	− 0.56	0.58
Global cognition	− 0.21	0.13	− 1.67	0.11
Symptoms	− 0.001	0.01	− 0.28	0.78
GFR	− 0.02	0.04	− 0.6	0.55
GFS	− 0.09	0.07	− 1.25	0.22
Strauss	0.06	0.05	1.18	0.25
Duration	− 0.001	0.004	− 0.203	0.84
Age	− 0.02	0.03	− 0.56	0.58
FSIQ	− 0.003	0.01	− 0.41	0.68
Gender	0.18	0.16	1.1	0.28
Education	0.13	0.05	2.53	0.02

- Regression analysis with all the variables - model not significant.

Example - Lasso analysis

- λ determined by 10-fold cross validation.
- Reduced and more predictive model.

Measure	A) LASSO	B) Linear regression			
	Coefficient	Estimate	Std. error	*t*-Value	*p*-Value
(Intercept)	− 0.74	− 1.58	0.58	− 2.72	0.01
Global cognition	− 0.16	− 0.25	0.08	− 2.94	0.01*
Gender	0.05	0.18	0.13	1.37	0.18
Education	0.07	0.12	0.04	3.05	0.004*