

Lecture 12: Ridge Regression, LARS, Logistic Regression

Nancy R. Zhang

Statistics 203, Stanford University

February 18, 2010

Exploring the model space

- 1 Forward selection:
 - 1 Start with null model.
 - 2 Repeat: add variable with the most significant F-test.
 - 3 End when no variable has F-test p-value $< \alpha$.
- 2 Backward elimination:
 - 1 Start with full model.
 - 2 Repeat: delete variable with the least significant F-test.
 - 3 End when all variables have F-test p-value $< \alpha$.
- 3 Forward + Backward: Same as forward procedure, with option of deleting a variable at each step.
- 4 All subsets: possible when number of possible predictors is small (< 20).

Model Shrinkage Methods

- 1 Bias variance trade off:

$$EPE = \sigma^2 + (\text{Model bias})^2 + \text{Model variance}$$

$$MSE \equiv E[\beta - \hat{\beta}]^2 = \text{Bias}(\hat{\beta}) + \text{Var}(\hat{\beta}).$$

- 2 Mallows C_p statistic:

$$C_p = SSE + 2p\hat{\sigma}^2.$$

The second term is a “penalty” for model size.

- 3 Today: penalties based on $\hat{\beta}$.
 - 1 Ridge regression
 - 2 LASSO

Ridge Regression

Ridge was developed first. It is based on the idea of constrained minimization:

$$\text{Minimize: } \sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^p X_{ij}\beta_j \right)^2$$

$$\text{Subject to: } \sum_{j=1}^p \beta_j^2 < C.$$

By the Lagrange multiplier method, this is equivalent to:

$$\text{Minimize: } \sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^p X_{ij}\beta_j \right)^2 + \lambda_C \sum_{j=1}^p \beta_j^2.$$

The second term is a penalty that depends on $\|\beta\|^2$.

Ridge Regression

Ridge regression:

$$\text{Minimize: } \sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^p X_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2.$$

- 1 In statistics this is also called “shrinkage”: you are shrinking $\|\beta\|^2$ towards 0.
- 2 λ is a shrinkage parameter that you have to choose.
- 3 The Ridge solution $\hat{\beta}_{\text{ridge}}$ is easy to solve, because the above is still a quadratic function in β .

Ridge Solutions

Ridge loss function:

$$f(\beta) = \sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^p X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2.$$

In matrix notation:

$$\begin{aligned} f(\beta) &= (Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta \\ &= \beta'[X'X + \lambda I]\beta - \beta'X'Y - Y'X\beta + Y'Y \end{aligned}$$

Solving $f'(\beta) = 0$ gives you:

$$\hat{\beta}_{\text{ridge}} = (X'X + \lambda I)^{-1}X'Y.$$

Ridge Solutions

- 1 Whereas the least squares solutions $\hat{\beta} = (X'X)^{-1}X'Y$ are unbiased if model is correctly specified, ridge solutions are *biased*

$$E[\hat{\beta}_{\text{ridge}}] \neq \beta.$$

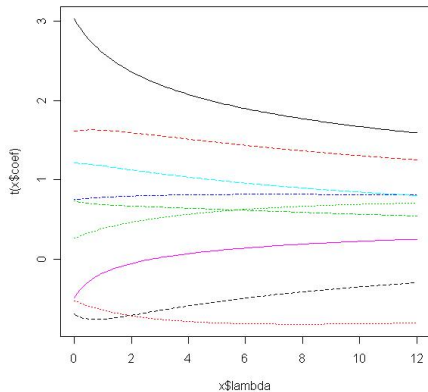
- 2 However, at the cost of bias, Ridge reduces the variance, and thus might reduce MSE.

$$MSE = Bias^2 + Variance$$

- 3 Ridge solutions are hard to interpret, because it is not sparse.

Sparse: some β_j 's are set exactly to 0.

Ridge solutions versus lambda



L_1 penalties

What if we constrain the L_1 norm instead of the Euclidean norm?

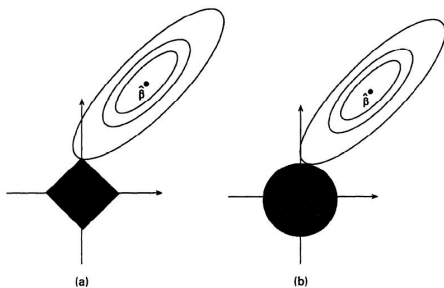
$$\text{Minimize: } \sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^p X_{ij} \beta_j \right)^2$$

$$\text{Subject to: } \sum_{j=1}^p |\beta_j| < C.$$

This is a subtle, but important change.

$$\text{Minimize: } \sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^p X_{ij} \beta_j \right)^2 + \lambda_C \sum_{j=1}^p |\beta_j|.$$

The above is termed Lasso regression (Tibshirani, 1996).



Comparing Lasso and Ridge, from Tibshirani (1996).

Lasso

Lasso loss function is no longer quadratic, but is still convex:

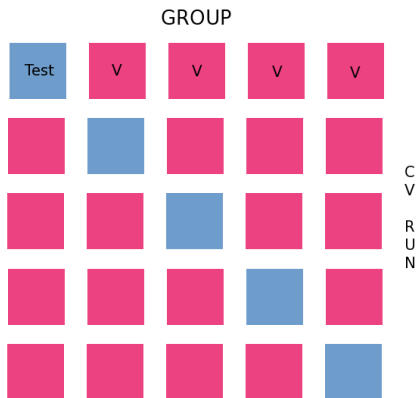
$$\text{Minimize: } \sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^p X_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j|.$$

- 1 Unlike Ridge, there is no analytic solution for the LASSO.
- 2 Efron et al. (2002) gave an efficient algorithm `lars` to solve the Lasso.
- 3 Lasso solutions are sparse.

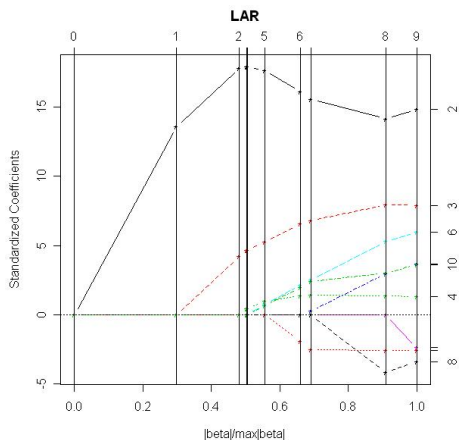
Selecting the shrinkage parameter λ

λ can be selected based on any of the model selection criteria we have discussed.

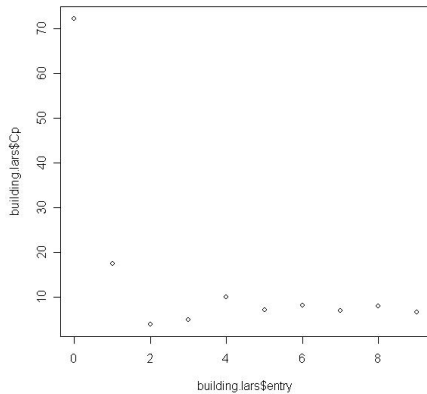
- 1 C_p included in the output of `lars`.
- 2 Cross-validation (`cv.lars`).



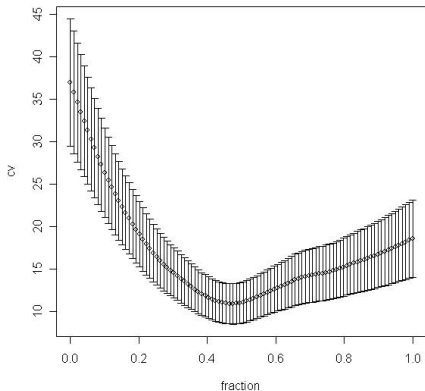
Lars output



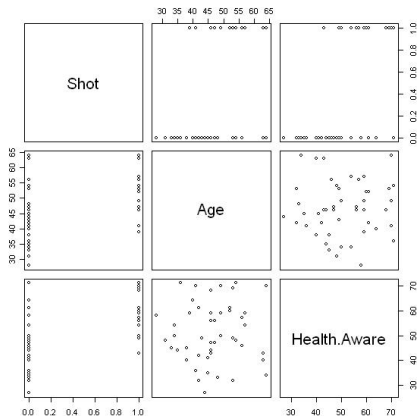
Lars C_p



Lars cross validation prediction error



Binary Response - Flu Shot Example



A clinic sent fliers to its clients to encourage everyone, but especially older persons, to get a flu shot in time for protection against an expected flu epidemic.

- 1 50 clients randomly sampled
- 2 Y : did they get flu shot?
- 3 Predictor variables: Age, health awareness.

Binary outcomes

Previously, we have dealt only with Gaussian models:

- 1 Continuous response.
- 2 Errors assumed to be approximately Gaussian

$$Y \sim N(\mu, \sigma^2), \quad \mu = X\beta.$$

- 3 Variance of errors don't depend on mean.

$$\sigma^2 \text{ constant.}$$

Binary outcomes:

- 1 Response is 0 or 1.
- 2 $E(\text{Response})$ is restricted in $[0, 1]$ interval.
- 3 The concept of "Gaussian error" does not apply here.

Logistic Regression Model

Suppose we have an increasing function $g : (0, 1) \rightarrow (-\infty, \infty)$. It seems reasonable to model the binary responses as follows:

$$P(Y = 1 | X_1, \dots, X_p) = g^{-1}(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p).$$

A popular choice for g is the logit transform:

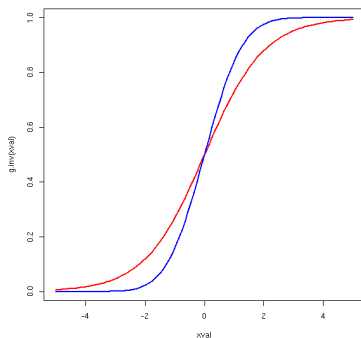
$$g(\pi) = \log \left(\frac{\pi}{1 - \pi} \right).$$

The inverse g^{-1} is

$$g^{-1}(z) = \frac{e^z}{1 + e^z}.$$

Logit link function

$$P(Y = 1|X) = \frac{e^{X\beta}}{1 + e^{X\beta}}$$



Logistic Regression Model

So far, we've given a model for the relationship between the mean of Y and X :

$$\pi(X) = P(Y = 1|X) = \frac{e^{X\beta}}{1 + e^{X\beta}}.$$

Or, equivalently,

$$\text{logit}[P(Y = 1|X)] = \beta X.$$

What about the variance of Y ? Since Y is bernoulli 0-1,

$$\text{Var}(Y) = \pi(X)[1 - \pi(X)].$$

Note that the variance is a function of the mean.

Interpretation of Logistic Model

A commonly used quantity to quantify binary data is the *odds*

$$\text{odds} = \frac{P(Y = 1|X)}{P(Y = 0|X)}.$$

What are the odds of

- 1 rain?
- 2 winning the match?

In the logistic regression model,

$$\log(\text{odds}) = \beta X.$$

The parameter β is the contribution of unit increase in X to the odds.

Maximum likelihood estimation of β

- Data: $\mathbf{X}_i, Y_i \in \{0, 1\}, i = 1, \dots, n$.
- Model: $P(Y_i = 1) = \frac{e^{\mathbf{x}_i\beta}}{1 + e^{\mathbf{x}_i\beta}}$.
- Goal: estimate β , compute confidence intervals, hypothesis testing.

Log-likelihood:

$$\begin{aligned}l(\beta) &= \log \left[\prod_{i=1}^n \left(\frac{e^{\mathbf{x}_i\beta}}{1 + e^{\mathbf{x}_i\beta}} \right)^{Y_i} \left(\frac{1}{1 + e^{\mathbf{x}_i\beta}} \right)^{1 - Y_i} \right] \\ &= \left(\sum_{i=1}^n Y_i \mathbf{x}_i \right) \beta - \sum_{i=1}^n \log(1 + e^{\mathbf{x}_i\beta})\end{aligned}$$

Goal: solve for $\hat{\beta} = \operatorname{argmax}_{\beta} l(\beta)$.

Review - Model fitting (i.e. solving for $\hat{\beta}$)

Fitting can be done by Newton-Raphson:

- 1 Let $u' = (\frac{\delta l(\beta)}{\delta \beta_i})_{i=1, \dots, p}$ be the gradient vector.
- 2 Let H be the Hessian matrix $h_{i,j} = \frac{\delta^2 l(\beta)}{\delta \beta_i \delta \beta_j}$.
- 3 Start with an initial $\beta^{(0)}$, then iterate $\beta^{(t+1)} = \beta^{(t)} - (H^{(t)})^{-1} u^{(t)}$.

The idea is, for each iteration t , to approximate $l(\beta)$ locally by a quadratic:

$$l(\beta) \approx l(\beta^{(t)}) + u^{(t)'}(\beta - \beta^{(t)}) + \frac{1}{2}(\beta - \beta^{(t)})' H^{(t)}(\beta - \beta^{(t)}),$$

and solve for $\delta l(\beta)/\delta \beta \approx u^{(t)} + H^{(t)}(\beta - \beta^{(t)}) = 0$.

For logistic regression model,

$$\beta^{(t+1)} = \beta^{(t)} + \{X' \text{diag}[\pi_i^{(t)}(1 - \pi_i^{(t)})]X\}^{-1} X'(y - \pi^{(t)}).$$

This is equivalent to doing a weighted linear regression at each step.

Inference for β

In Gaussian case:

$$\hat{\beta} = (X'X)^{-1}X'Y, \quad Y \sim N(X\beta, \sigma^2 I).$$

Since Gaussian vectors remain Gaussian under linear transforms,

$$\hat{\beta} \sim N(\beta, (X'X)^{-1}\sigma^2).$$

For logistic regression, $\hat{\beta}$ is no longer linear in Y . However, *asymptotically* (i.e. n large), it is Gaussian. It's covariance can be estimated by

$$\widehat{\text{cov}}(\hat{\beta}) = (X' \text{diag}[\hat{\pi}_i(1 - \hat{\pi}_i^{(t)})]X)^{-1}.$$

From the square root of the diagonal elements of the above matrix you can get $\widehat{\text{s.e.}}(\hat{\beta})$.

Wald tests for β

- 1 Confidence intervals for β :

$$[\hat{\beta} - z_{\alpha/2} \widehat{\text{s.e.}}(\hat{\beta}), \hat{\beta} + z_{\alpha/2} \widehat{\text{s.e.}}(\hat{\beta})]$$

- 2 Two sided test $H_0 : \beta = 0$, reject if

$$\left| \frac{\hat{\beta}}{\widehat{\text{s.e.}}(\hat{\beta})} \right| > z_{\alpha/2}$$

- 3 Test of constraint $H_0 : C_{j \times p} \beta_{p \times 1} = h_{j \times 1}$, reject if

$$(C\hat{\beta} - h)'(C\widehat{\text{cov}}(\hat{\beta})C')^{-1}(C\hat{\beta} - h)$$

is larger than $\chi_{j,1-\alpha}^2$.

Assessment of model fit

In linear regression, we used the F -test:

$$F = \frac{[SSE(RM) - SSE(FM)]/[\Delta df]}{SSE(FM)/[n - df(FM)]}.$$

$$F \sim F_{\Delta df, n - df(FM)}.$$

The analogous quantity of SSE for non-linear models is *deviance*:

$$\text{Deviance}(\hat{\beta}) = -2[l(\tilde{\beta}, Y) - l(\hat{\beta}, Y)],$$

where

$l(\cdot, Y)$ is log-likelihood,

$\tilde{\beta}$ is fit of data using saturated model (n predictors).

Assessment of model fit

For Gaussian model,

$$\text{Deviance}(\hat{\beta}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

For Logistic model,

$$\text{Deviance}(\hat{\beta}) = 2 \sum_{i=1}^n \left[Y_i \log \frac{Y_i}{\hat{\pi}_i} + (1 - Y_i) \log \frac{1 - Y_i}{1 - \hat{\pi}_i} \right]$$

Convention: $0 \times \log 0 = 0$.

Nested Chi-squared tests of model fit

As for SSE, the greater the deviance, the poorer the fit. If reduced model (RM) were true, then

$$\text{Deviance}(RM) - \text{Deviance}(FM) \rightarrow \chi_{df(FM)-df(RM)}^2.$$

Thus, reject RM at asymptotic level α if

$$\text{Deviance}(RM) - \text{Deviance}(FM) > \chi_{df(FM)-df(RM), 1-\alpha}^2.$$

Model diagnosis

In linear regression, the standardized residuals were used to diagnose model fit.

$$r_i = y_i - \hat{y}_i, \quad r_i^* = \frac{r_i}{\hat{\sigma}_{r_i}} = \frac{r_i}{\sqrt{1 - \rho_{ii}}}.$$

The analogous quantity here is the Pearson residual,

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}},$$

$$r_i^* = \frac{r_i}{\sqrt{1 - \hat{h}_{ii}}}.$$

where \hat{h}_{ii} are diagonals of

$$\hat{H} = W^{1/2} X (X' W X)^{-1} X' W^{1/2}.$$

$$W = \text{diag}[\pi_i(1 - \pi_i)].$$

Model diagnosis

Another quantity you can use is the deviance residuals:

$$\begin{aligned} \text{Deviance}(\hat{\beta}) &= 2 \sum_{i=1}^n \left[Y_i \log \frac{Y_i}{\hat{\pi}_i} + (1 - Y_i) \log \frac{1 - Y_i}{1 - \hat{\pi}_i} \right] \\ &= 2 \sum \text{observed} \times \log \frac{\text{observed}}{\text{fitted}}. \end{aligned}$$

So let d_i be the contribution of data point i to the above measure of mis-fit:

$$d_i = Y_i \log \frac{Y_i}{\hat{\pi}_i} + (1 - Y_i) \log \frac{1 - Y_i}{1 - \hat{\pi}_i}$$

$$\text{Deviance residual: } \sqrt{|d_i|} \times \text{sign}(y_i - \hat{\pi}_i).$$