

Stat GR 5205 Lecture 10

Jingchen Liu

Department of Statistics
Columbia University

Algorithmic approach

- ▶ Forward selection
- ▶ Backward deletion
- ▶ Stepwise regression

Pros and cons

- ▶ Easy to implement
- ▶ Less computation
- ▶ In consistency

Likelihood-based criteria

- ▶ Akaike information criterion (AIC)

$$n \log(\hat{\sigma}^2) + 2p.$$

Derive AIC.

- ▶ Bayesian information criterion

$$n \log(\hat{\sigma}^2) + p \log(n)$$

General form

- ▶ Akaike information criterion (AIC)

$$-2 \log[L(\hat{\theta})] + 2p \approx 2E[\log L(\theta)]|_{\theta=\hat{\theta}}$$

- ▶ Bayesian information criterion

$$-2 \log[L(\hat{\theta})] + p \log(n)$$

Delimma

- ▶ Too few variables (missing the true predictor) – bias.
- ▶ Too many variables – variance.

Delimma

- ▶ Too few variables (missing the true predictor) – bias.
- ▶ Too many variables – variance.

Mallows' C_p

- ▶ Let $\mu_i = E(y|x_i)$.
- ▶ Mean squared error

$$E[(\hat{y}_i - \mu_i)^2|x_i] = E^2(\hat{y}_i - \mu_i|x_i) + \text{Var}(\hat{y}_i - \mu_i|x_i)$$

- ▶ Total mean squared error

$$\sum_{i=1}^n E[(\hat{y}_i - \mu_i)^2|x_i] = \sum_{i=1}^n E^2(\hat{y}_i - \mu_i|x_i) + \sum_{i=1}^n \text{Var}(\hat{y}_i - \mu_i|x_i)$$

Mallows' C_p

- ▶ Let $\mu_i = E(y|x_i)$.
- ▶ Mean squared error

$$E[(\hat{y}_i - \mu_i)^2|x_i] = E^2(\hat{y}_i - \mu_i|x_i) + \text{Var}(\hat{y}_i - \mu_i|x_i)$$

- ▶ Total mean squared error

$$\sum_{i=1}^n E[(\hat{y}_i - \mu_i)^2|x_i] = \sum_{i=1}^n E^2(\hat{y}_i - \mu_i|x_i) + \sum_{i=1}^n \text{Var}(\hat{y}_i - \mu_i|x_i)$$

Mallows' C_p

- ▶ Let $\mu_i = E(y|x_i)$.
- ▶ Mean squared error

$$E[(\hat{y}_i - \mu_i)^2|x_i] = E^2(\hat{y}_i - \mu_i|x_i) + \text{Var}(\hat{y}_i - \mu_i|x_i)$$

- ▶ Total mean squared error

$$\sum_{i=1}^n E[(\hat{y}_i - \mu_i)^2|x_i] = \sum_{i=1}^n E^2(\hat{y}_i - \mu_i|x_i) + \sum_{i=1}^n \text{Var}(\hat{y}_i - \mu_i|x_i)$$

Mallows' C_p

$$C_p = \frac{SSE}{\hat{\sigma}_f^2} - (n - 2p) = p + (n - p) \frac{\hat{\sigma}^2 - \hat{\sigma}_f^2}{\hat{\sigma}_f^2}$$

Comparison

- ▶ AIC
- ▶ BIC
- ▶ C_p

Example

- ▶ Response variable: log-survival time
- ▶ Covariates: blood clotting score, prognostic index, enzyme function test score, liver function test score, age, gender, alcohol use (none, moderate, heavy)
- ▶ AIC

Example

Forward

Start: AIC=-75.7

logsurvival ~ 1

	Df	Sum of Sq	RSS	AIC
+ enzyme	1	5.4762	7.3316	-103.827
+ liver	1	5.3990	7.4087	-103.262
+ progind	1	2.8285	9.9792	-87.178
+ heavy	1	1.7798	11.0279	-81.782
+ score	1	0.7763	12.0315	-77.079
+ gender	1	0.6897	12.1180	-76.692
<none>			12.8077	-75.703
+ age	1	0.2691	12.5386	-74.849
+ alcohol	1	0.2052	12.6025	-74.575

Step: AIC=-103.83
 logsurvival ~ enzyme

	Df	Sum of Sq	RSS	AIC
+ progind	1	3.01908	4.3125	-130.48
+ liver	1	2.20187	5.1297	-121.11
+ score	1	1.55061	5.7810	-114.66
+ heavy	1	1.13756	6.1940	-110.93
<none>			7.3316	-103.83
+ gender	1	0.25854	7.0730	-103.77
+ age	1	0.23877	7.0928	-103.61
+ alcohol	1	0.06498	7.2666	-102.31

Step: AIC=-130.48

logsurvival ~ enzyme + progind

	Df	Sum of Sq	RSS	AIC
+ heavy	1	1.46961	2.8429	-150.99
+ score	1	1.20395	3.1085	-146.16
+ liver	1	0.69836	3.6141	-138.02
+ alcohol	1	0.22632	4.0862	-131.39
+ age	1	0.16461	4.1479	-130.59
<none>			4.3125	-130.48
+ gender	1	0.08245	4.2300	-129.53

COLUMBIA UNIVERSITY

IN THE CITY OF NEW YORK

• • •

Step: AIC=-163.83

logsurvival ~ enzyme + progind + heavy + score
+ gender + age

	Df	Sum of Sq	RSS	AIC
<none>			2.0052	-163.83
+ alcohol	1	0.033193	1.9720	-162.74
+ liver	1	0.002284	2.0029	-161.90

Example

Backward

Start: AIC=-160.77

```
logsurvival ~ score + progind + enzyme + liver
              + age + gender + alcohol + heavy
```

	Df	Sum of Sq	RSS	AIC
- liver	1	0.00129	1.9720	-162.74
- alcohol	1	0.03220	2.0029	-161.90
- age	1	0.07354	2.0443	-160.79
<none>			1.9707	-160.77
- gender	1	0.08415	2.0549	-160.51
- score	1	0.31809	2.2888	-154.69
- heavy	1	0.84573	2.8165	-143.49
- progind	1	2.09045	4.0612	-123.72
- enzyme	1	2.99085	4.9616	-112.91

Step: AIC=-162.74

```
logsurvival ~ score + progind + enzyme
              + age + gender + alcohol + heavy
```

	Df	Sum of Sq	RSS	AIC
- alcohol	1	0.0332	2.0052	-163.834
<none>			1.9720	-162.736
- age	1	0.0876	2.0596	-162.389
- gender	1	0.0971	2.0691	-162.141
- score	1	0.6267	2.5988	-149.833
- heavy	1	0.8446	2.8166	-145.486
- progind	1	2.6731	4.6451	-118.471
- enzyme	1	5.0986	7.0706	-95.784

Step: AIC=-163.83

```
logsurvival ~ score + progind + enzyme
              + age + gender + heavy
```

	Df	Sum of Sq	RSS	AIC
<none>			2.0052	-163.834
- age	1	0.0768	2.0820	-163.805
- gender	1	0.0977	2.1029	-163.265
- score	1	0.6282	2.6335	-151.117
- heavy	1	0.9002	2.9055	-145.809
- progind	1	2.7626	4.7678	-119.064
- enzyme	1	5.0801	7.0853	-97.672

Example

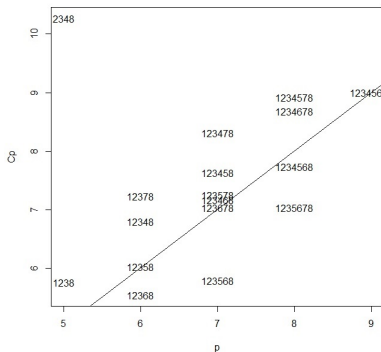


Figure: 1. score, 2. progind, 3. enzyme, 4. liver, 5. age, 6. gender, 7. alcohol, 8. heavy

Least Absolute Shrinkage and Selection Operator(LASSO)

Tibshirani (1996, JRSS B)

- ▶ Observation: soft-thresholding
- ▶ The LASSO estimator

$$\hat{\beta} = \arg \min_{\beta} \sum_i (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_1$$

Least Absolute Shrinkage and Selection Operator(LASSO)

Tibshirani (1996, JRSS B)

- ▶ Observation: soft-thresholding
- ▶ The LASSO estimator

$$\hat{\beta} = \arg \min_{\beta} \sum_i (y_i - x_i^{\top} \beta)^2 + \lambda \|\beta\|_1$$

The penalized likelihood

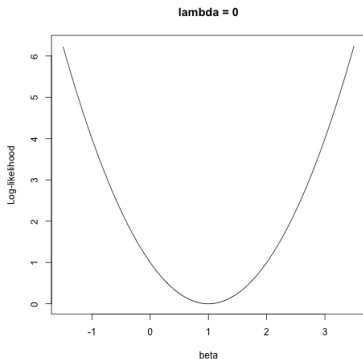


Figure: $(\beta - 1)^2 + \lambda \|\beta\|_1$

The penalized likelihood

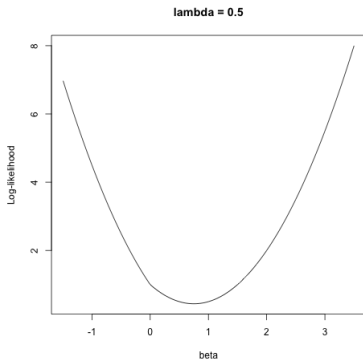


Figure: $(\beta - 1)^2 + \lambda \|\beta\|_1$

The penalized likelihood

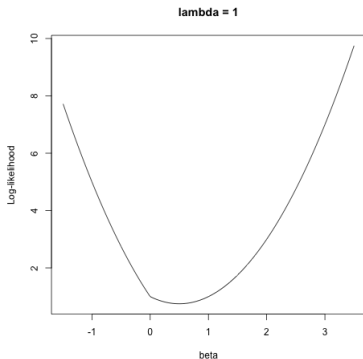


Figure: $(\beta - 1)^2 + \lambda \|\beta\|_1$

The penalized likelihood

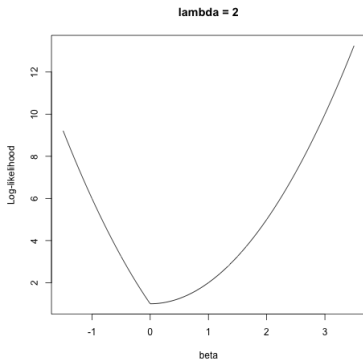


Figure: $(\beta - 1)^2 + \lambda \|\beta\|_1$

The penalized likelihood

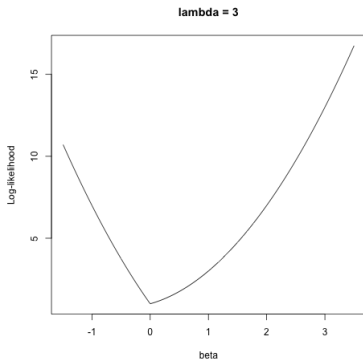


Figure: $(\beta - 1)^2 + \lambda \|\beta\|_1$

The penalized estimator

- ▶ The penalized likelihood

$$(\beta - \hat{\beta})^\top X^\top X (\beta - \hat{\beta}) + \lambda \|\beta - \hat{\beta}\|_1$$

- ▶ Simplified situation

$$(\beta - \hat{\beta})^2 + \lambda \|\beta\|_1$$

The penalized estimator

- ▶ The penalized likelihood

$$(\beta - \hat{\beta})^\top X^\top X (\beta - \hat{\beta}) + \lambda \|\beta - \hat{\beta}\|_1$$

- ▶ Simplified situation

$$(\beta - \hat{\beta})^2 + \lambda \|\beta\|_1$$

The penalized likelihood

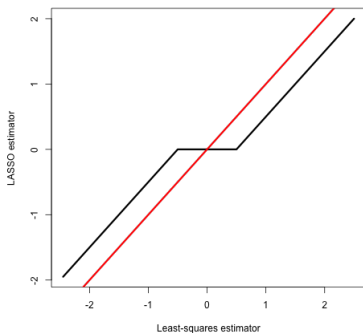


Figure: LS estimator versus LASSO estimator

The penalized likelihood

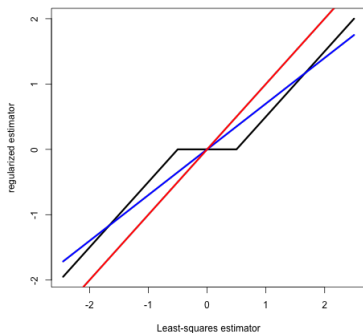


Figure: LS estimator, LASSO estimator, and ridge regression