Stat GR 5205 Lecture 11

Jingchen Liu

Department of Statistics
Columbia University

Weighted least-squares

- The setting

$$y_i = x_i^\top \beta + \sigma_i \varepsilon_i$$

where

$$\varepsilon_i \sim N(0, 1)$$

- Transformation

$$\frac{y_i}{\sigma_i} = \frac{x_i^\top}{\sigma_i} \beta + \varepsilon_i$$

Weighted least-squares

- General form

$$Y = X\beta + \varepsilon$$

where

$$\varepsilon \sim N(0, \Sigma)$$

- Write $\Sigma^{1/2}\delta$, where $\delta \sim N(0, I)$

$$\Sigma^{-1/2}Y = \Sigma^{-1/2}X\beta + \delta$$

- The estimator

$$\hat{\beta} = [X^\top \Sigma^{-1} X]^{-1} X \Sigma^{-1} Y$$

Weighted least-squares

- Variance of least-square estimator

$$Var(\hat{\beta}) = (X^\top X)^{-1} X^\top \Sigma X (X^\top X)^{-1}$$

- Robust (Sandwich) variance estimator

$$(X^\top X)^{-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 x_i x_i^\top (X^\top X)^{-1}$$

The penalized likelihood

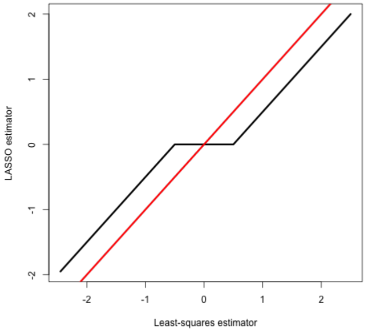


Figure: LS estimator versus LASSO estimator

The penalized likelihood

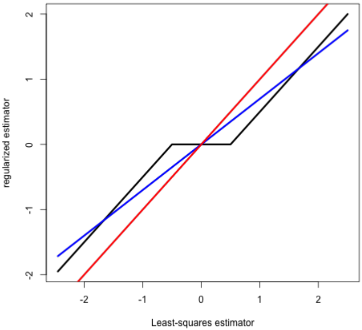


Figure: LS estimator, LASSO estimator, and ridge regression

LASSO and ridge regression

- The LASSO estimator

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

- Ridge regression

$$\hat{\beta} = \arg\min_{\beta} \sum_i (y_i - x_i^\top \beta)^2 + \lambda \|\beta\|_2$$

Computation

- Convex function
- LASSO penalty is convex
- Optimization

Solution path

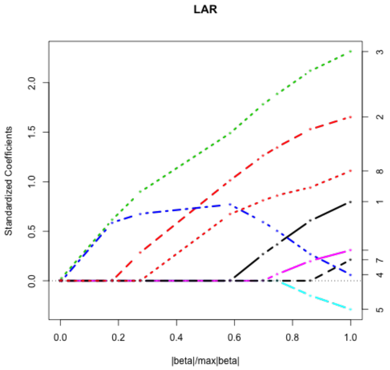


Figure: Solution path

	enzyme	liver	progind	heavy	score	gender	age	alcohol
Var	3	4	2	8	1	6	5	7
Step	1	2	3	4	5	6	7	8

	Df	Rss	Cp
0	1	12.8077	240.4521
1	2	12.6758	239.4407
2	3	8.2207	139.7107
3	4	6.4838	102.0520
4	5	3.0563	25.7884
5	6	2.4719	14.4428
6	7	2.3059	12.6526
7	8	2.0606	9.0527
8	9	1.9707	9.0000

- LASSO

```
     enzyme liver progind heavy score gender age alcohol
Var       3     4        2     8     1      6   5       7
Step      1     2        3     4     5      6   7       8
```

- AIC

```
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

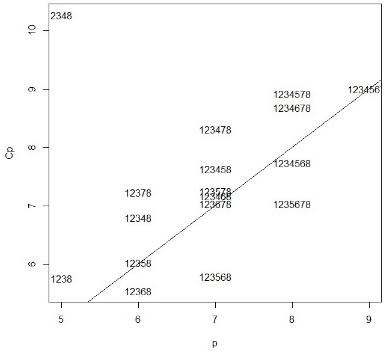


Figure: 1. score, 2. progind, 3. enzyme, 4. liver, 5. age, 6. gender, 7. alcohol, 8. heavy

Comparison

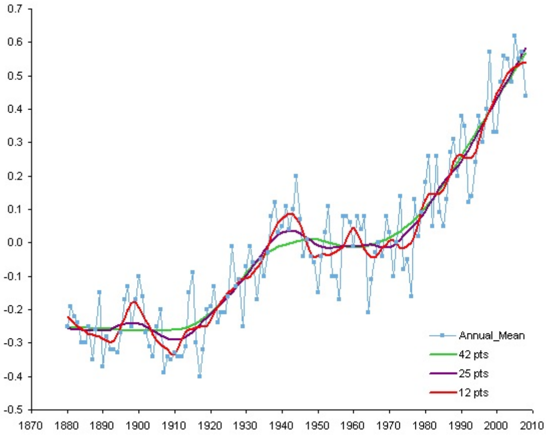
- AIC, BIC, and C_p
- LASSO, a.k.a. L_1 regularized regression
- Sparsity
- Oracle property, $\sqrt{N} \ll \lambda \ll N$, conditions on collinearity
- Other penalty functions

Final comments

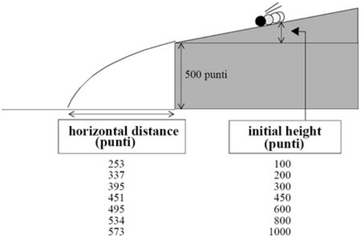
- The results from variable selection only serves as a guideline of research.
- Model selection is not a replacement of science.
- Machine learning

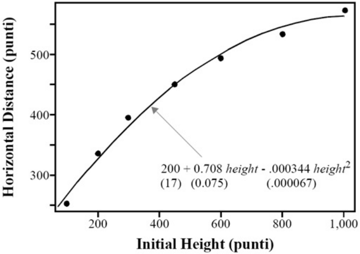
Nonparametric regression

- About nonparametric regression
- Different approaches


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
Annual_Mean
42 pts
25 pts
12 pts

Galileo's experiment





$$distance = \beta_0 + \beta_1 height + \beta_2 height^2 + \varepsilon$$

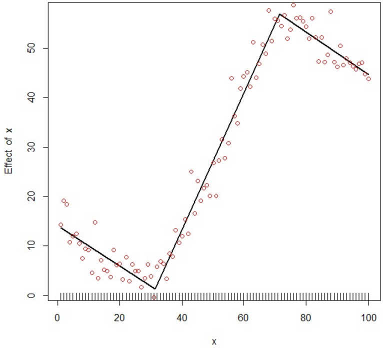
Segment/piecewise regression

$$f(x) \triangleq E(Y|X = x)$$

- $f(x)$ is a piecewise linear function

$$f(x) = \beta_0 + \beta_1(x - x_1)^+ + \beta_2(x - x_2)^+ + \beta_3(x - x_3)^+ \ldots$$

- Variable selection


Effect of x
0
10
20
30
40
50
0
20
40
60
80
100
x

Kernel smoothing

- An estimate

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} y_i I(x_i = x)}{\sum_{i=1}^{n} I(x_i = x)}$$

- Kernel smoothing takes advantage of the continuity of $f(x)$.
- Kernel estimation

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} y_i K_h(x - x_i)}{\sum_{i=1}^{n} K_h(x - x_i)}$$

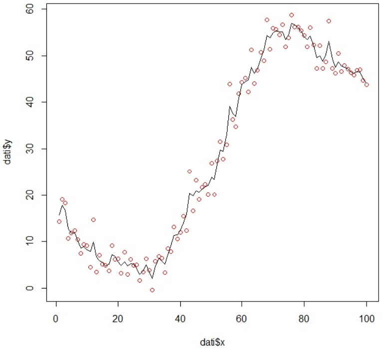


Figure: Gaussian kernel, bandwidth $= 2$

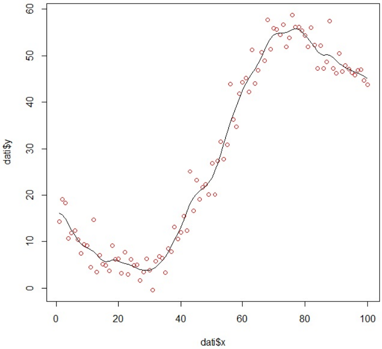


Figure: Gaussian kernel, bandwidth = 5

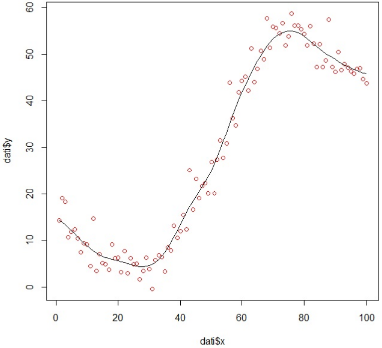


Figure: Gaussian kernel, bandwidth = 10

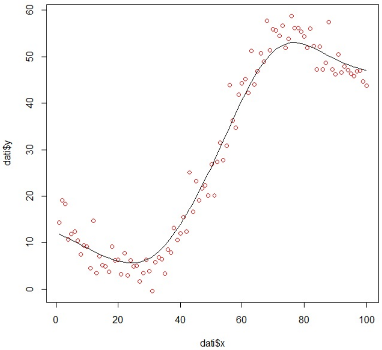


Figure: Gaussian kernel, bandwidth = 20

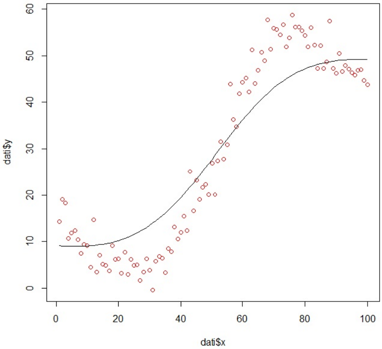


Figure: Gaussian kernel, bandwidth = 50

Smoothing spline

- Least-squares estimate

$$\min \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2$$

- General least-squares estimate

$$\min \sum_{i=1}^{n} (y_i - f(x_i))^2$$

- Regularization

$$\min \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b [f''(x)]^2 dx$$

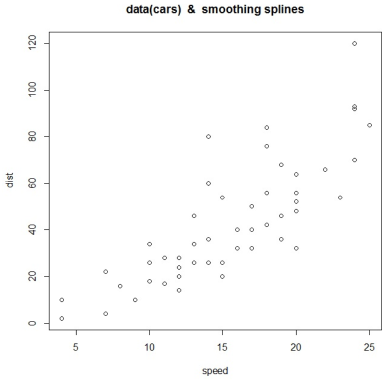

data(cars) & smoothing splines
dist
0
20
40
60
80
100
120
5
10
15
20
25
speed

Figure: Cars

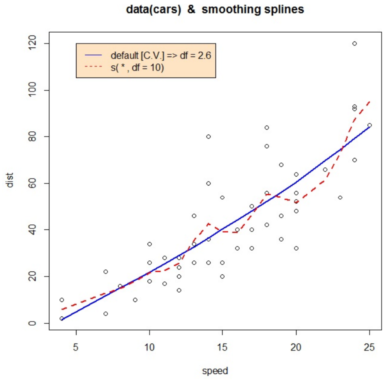


Figure: Cars: spline fitting

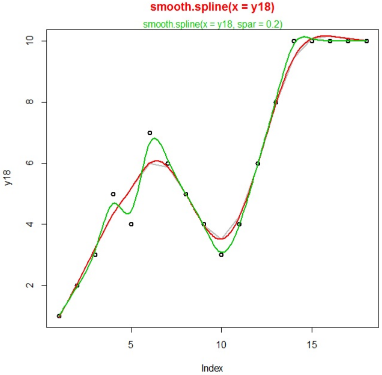


Figure: Simulated data

Nonparametric regression

- Expansion

$$f(x) = \sum_{i=1}^{\infty} c_i B_i(x)$$

- Truncation

$$f(x) = \sum_{i=1}^{n} c_i B_i(x)$$