Lecture 6

Non-parametric Regression and Shrinkage Methods

Reading: Faraway 2014 Chapters 9.5-9.6 and 11.3-11.4; Faraway 2016 Chapters 14.1-14.2, 14.6; 15.2-15.3; ISLR 2021 Chapters 6.2 and 7.3-7.5, 7.7

DSA 8020 Statistical Methods II

Whitney Huang
Clemson University

Agenda

1. **Non-parametric Regression**
2. **Ridge Regression**
3. **LASSO**

Moving Away From Linear Regression

- We have mainly focused on linear regression so far

  Model: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$

  Data: $\boldsymbol{y}$ (response vector); $\boldsymbol{X}$ (design matrix)

  - $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$; $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \underbrace{\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}}_{\boldsymbol{H}\text{: “Hat” matrix}}\boldsymbol{y}$
  - $\hat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1})$

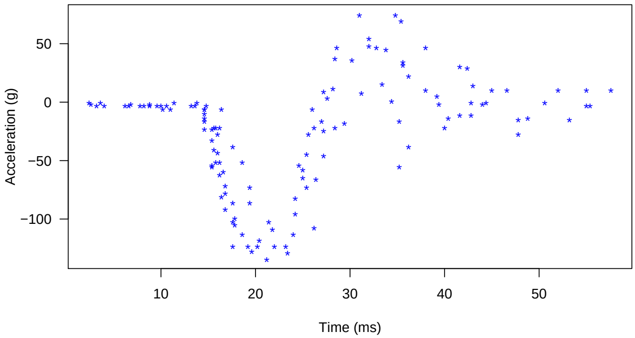
- In this lecture we are going to discuss non-parametric regression modeling

  Model: $y = f(x) + \varepsilon \Rightarrow \mathrm{E}[y|x] = f(x)$

  - The (smooth) function $f(x)$ must be represented somehow
  - The degree of smoothness of $f(x)$ must be made controllable
  - Some means for estimating the most appropriate degree of smoothness from data is required

Data from a Simulated Motorcycle Accident [Silverman, 1985)]

This data set is taken from a simulated motor-cycle crash experiment in order to study the efficacy of crash helmets.



We aim to estimate the smooth regression function $f(x)$

Kernel Estimators

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) y_i = \frac{1}{n}\sum_{i=1}^{n} w_i y_i, \text{ where } w_i = K\left(\frac{x - x_i}{h}\right)/h.$$

$K(\cdot)$ is a **kernel** where $\int K(x)\,dx = 1$, and h is the **bandwidth**, also known as the **smoothing parameter** in this context.

Two choices are required for implementing the kernel estimator:

- Kernel: It is desirable that the kernel provides both **smoothness** and **compactness**. An example is the Epanechnikov kernel

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2) & |x| < 1 \\ 0 & \text{otherwise} \end{cases}$$

- Smoothing parameter: If too small, the estimator will be too rough; if too large, it will smooth out important features

Representing a Smooth Function using Basis Functions

- Basis function representation: $f(x) = \sum_{j=1}^{J} b_j(x)\beta_j$
- There are many basis functions to choose from: Polynomials, Fourier Series, Radial Basis Functions...
- We are going to focus on splines: piecewise polynomials joined together to make a single smooth curve

An Example of a Cubic Spline Function

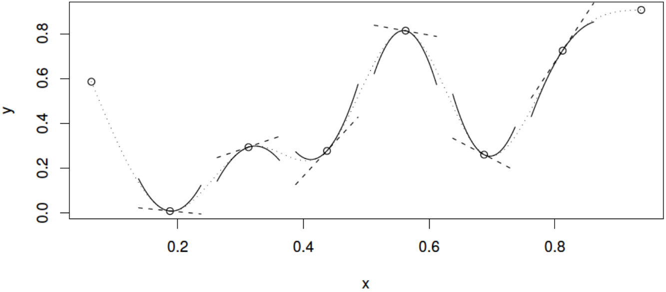


Figure 3.3 *A cubic spline is a curve constructed from sections of cubic polynomial joined together so that the curve is continuous up to second derivative. The spline shown (dotted curve) is made up of 7 sections of cubic. The points at which they are joined (○) (and the two end points) are known as the* knots *of the spline. Each section of cubic has different coefficients, but at the knots it will match its neighbouring sections in value and first two derivatives. Straight dashed lines show the gradients of the spline at the knots and the curved continuous lines are quadratics matching the first and second derivatives at the knots: these illustrate the continuity of first and second derivatives across the knots. This spline has zero second derivatives at the end knots: a '*natural *spline'.*

Source: Simon Wood, *Generalized Additive Models*, p. 122, Fig. 3.3

Regression Splines

- Choose J knot points to partition the range of x to form the spline basis $\boldsymbol{X}$
- Techniques from linear regression can be used to carry out estimation and inference
- However, the model fit tends to depend strongly on J, the number of knots, and $\{\xi_j\}_{j=1}^{J}$, the knot locations
  - Few knots: Resulting class of functions may be too restrictive (bias)
  - Many knots: We run the risk of overfitting (variance)

Problems with Regression Splines

- Regression splines are not truly "nonparametric" as the choices regarding J and $\{\xi_j\}_{j=1}^{J}$ are fundamentally parametric choices and have a large effect on the fit
- Model selection (i.e, choosing the degree of smoothing) is not straightforward
- An alternative approach to controlling smoothness is penalization [Green and Silverman 1993]

Controlling Smoothness with Penalization

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

- The first term captures the fit to the data, while the second penalizes curvature
- λ is the smoothing parameter, and it controls the tradeoff between the two terms:
  - $\lambda = 0$ imposes no restrictions and f will therefore interpolate the data
  - $\lambda = \infty$ returning us to ordinary linear regression

Selecting an appropriate λ is crucial

Natural Cubic Splines Solve the Penalized Least Squares!

Theorem: Out of all twice-differentiable functions, the one that minimizes

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

is a natural cubic spline with knots at every unique value of $\{x_i\}$

This penalized approach leads to the framework of smoothing splines, introduced by Grace Wahba to statisticians

Smoothing Splines (Cont'd)

Let $\{N_j\}_{j=1}^n$ denote the collection of natural cubic spline basis functions and $\boldsymbol{N}$ denote the $n \times n$ design matrix consisting of the basis functions evaluated at $\{x_i\}$:

- $f(x) = \sum_{j=1}^n N_j\beta_j$, where $\boldsymbol{N}_{ij} = N_j(x_i) \Rightarrow f(x) = \boldsymbol{N}\boldsymbol{\beta}$

- We can show that the objective function for penalized splines is

$$(\boldsymbol{y} - \boldsymbol{N}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{N}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\beta},$$

  where $\boldsymbol{\Omega}_{jk} = \int N_j''(x)N_k''(x)\,dx$

- The minimizer is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{N}^{\mathrm{T}}\boldsymbol{N} + \lambda\boldsymbol{\Omega}\right)^{-1}\boldsymbol{N}^{\mathrm{T}}\boldsymbol{y}$$

Smoothing Splines are Linear Smoothers

- From last slide we have

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{N}^{\mathrm{T}}\boldsymbol{N} + \lambda\boldsymbol{\Omega}\right)^{-1}\boldsymbol{N}^{\mathrm{T}}\boldsymbol{y}$$

- Therefore we have

$$\hat{\boldsymbol{y}} = \hat{f}(\boldsymbol{x}) = \boldsymbol{N}\left(\boldsymbol{N}^{\mathrm{T}}\boldsymbol{N} + \lambda\boldsymbol{\Omega}\right)^{-1}\boldsymbol{N}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{L}_{\lambda}\boldsymbol{y},$$

  $\Rightarrow$ a linear smoother

- $tr(\boldsymbol{L}_{\lambda})$ is a measure of the effective number of degrees of freedom

Choosing λ by Cross-Validation (CV)

Main idea:
Sequentially leave each observation out and predict it using the rest of the data. Find the λ that gives the best out of sample predictions.

- CV residual:

$$y_i - \hat{y}_{-i} = \frac{y_i - \hat{y}_i}{(1 - \boldsymbol{L}_{\lambda,i,i})}$$

- CV(λ):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{-i})^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{(1 - \boldsymbol{L}_{\lambda,i,i})^2}$$

- Generalized Cross-Validation (GCV):

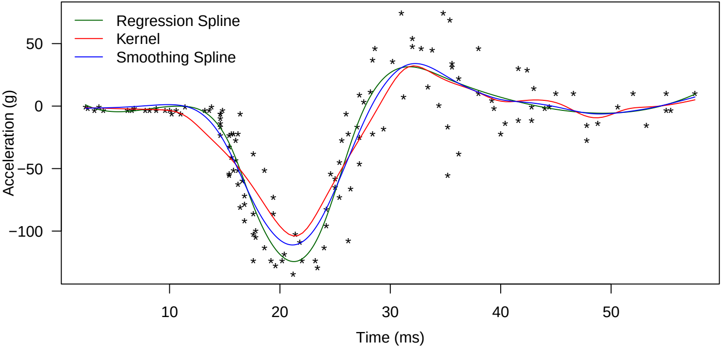
$$\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{(1 - \frac{tr(\boldsymbol{L}_\lambda)}{n})^2}$$

Non-parametric Regression Fits for Motorcycle Data

Regression Spline: 10 degrees of freedom quantile knot

Smoothing Spline: the amount of smoothness is estimated from the data by GCV

Kernel Regression: K: Epanechnikov kernel and $h = 5$



Generalized Additive Models for Multiple Predictors

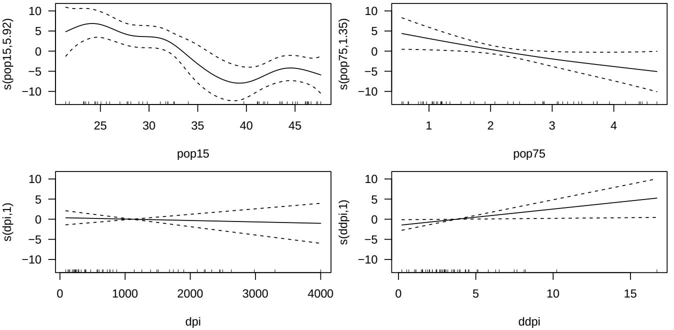
General non-parametric regression models

$$y = f(x_1, x_2, \cdots, x_p) + \varepsilon$$

suffer from the “curse of dimensionality”

Generalized Additive Models

$$y = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \varepsilon$$



Shrinkage Methods

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

$x_1, x_2, \cdots, x_{p-1}$ are the predictors.

Question: What if we have too many predictors (i.e., p is "large")?

- Explanation can be difficult due to collinearity
- Can lead to overfitting by using too many predictors

We will look at two methods, namely Ridge regression and LASSO, that allow us to "shrink" the information contained in all the predictors into a more useful form

Ridge Regression [Hoerl & Kennard, 1970]

Ridge regression assumes that the regression coefficients (after normalization) should not be very large

- The ridge regression estimate chooses the $\boldsymbol{\beta}$ that minimizes:

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

  where $\lambda \geq 0$ is a **tuning parameter** to be determined via cross-validation

- The ridge regression estimates:

$$\hat{\beta}_{\text{ridge}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

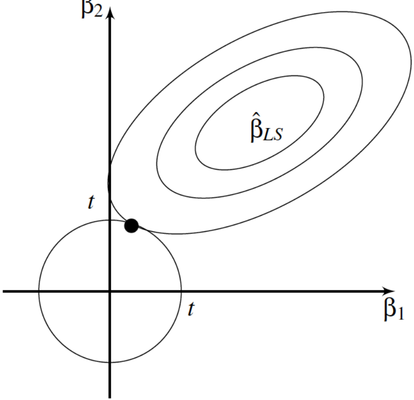
- Ridge regression is particularly effective when the model matrix is collinear

Graphical Illustration of Ridge Regression

Estimation of ridge regression can also be solved by choosing β to minimize

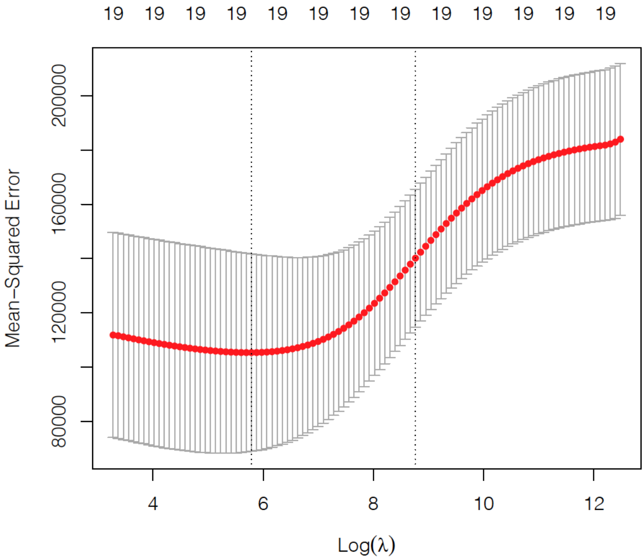
$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij})^2$$

subject to $\sum_{j=1}^{p} \beta_j^2 \leq t^2$



Source: p. 175, Fig. 11.9 *Linear Models with* R, Faraway, 2014

Choosing λ via Cross-Validation



Least Absolute Shrinkage and Selection Operator (LASSO) Tibshirani, 1996

LASSO assumes the effects are **sparse** in that the response can be explained by a small number of predictors with the rest having no effect

- LASSO choose $\hat{\beta}$ to minimize:

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

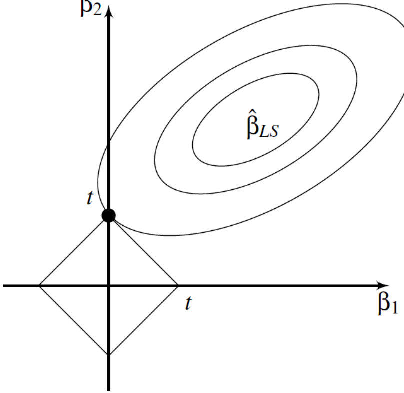
- No explicit solution to this minimization problem
- The penalty term has the effect of forcing some of the coefficient estimates to be zero when the tuning parameter λ is "large" $\Rightarrow$ performs shrinkage and variable selection

Graphical Illustration of LASSO

Estimation of LASSO can also be solved by choosing β to minimize

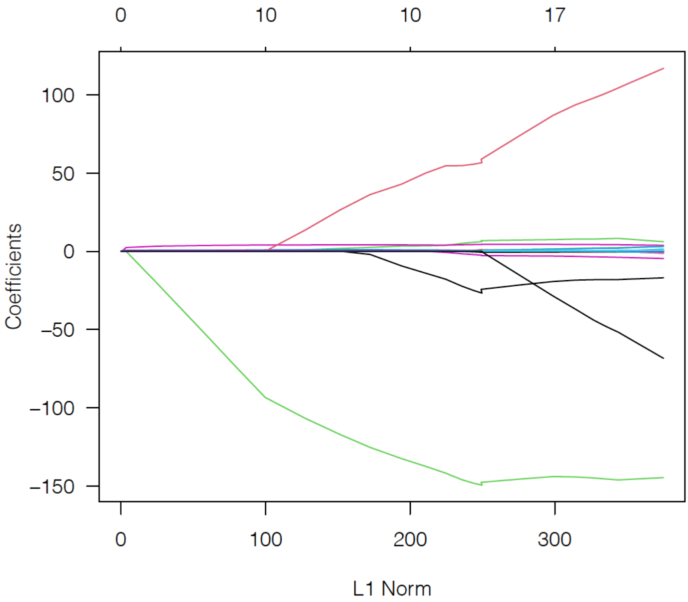
$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p-1}\beta_j x_{ij})^2$$

subject to $\sum_{j=1}^{p} |\beta_j| \leq t$

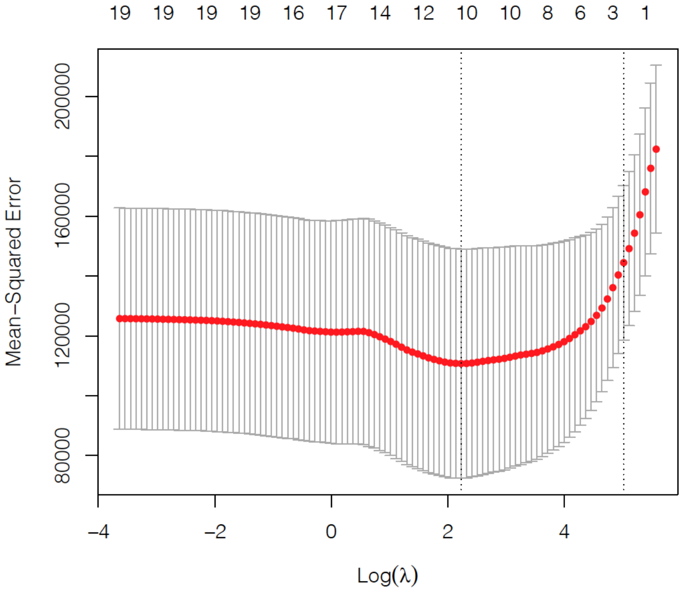


Source: p. 175, Fig. 11.9 *Linear Models with R*, Faraway, 2014

LASSO Path



Selecting λ via Cross-Validation



Summary

These slides cover:

- Non-Parametric Regression
- Ridge Regression
- LASSO

`R` functions to know:

- **Non-Parametric Regression**: `ksmooth` (kernel regression); `bs` (regression splines); `sreg` in the `fields` package (smoothing splines); `gam` (generalized additive models)
- **Ridge Regression/LASSO**: `glmnet` and `cv.glmnet` in the `glmnet` package