A Beginner’s Guide to Linear Models

Seoncheol Park

Seoncheol Park

Contents

Preface

열심히 공부합시다.

I

Intro

Seoncheol Park

1 Introduction

- **회귀분석(regression analysis)**: 설명변수와 반응변수 사이의 함수관계를 알아내는 통계적 방법
- 용어의 역사: Galton의 `Regression toward the mean`란 말에서부터 유래함

1.1 Galton's data

Q. 아버지와 아들 사이의 키 상관관계?

```
library(HistData)
xx = GaltonFamilies$midparentHeight
yy = GaltonFamilies$childHeight

plot(xx, yy, xlab="Father", ylab="Child", main="Galton's data")
```

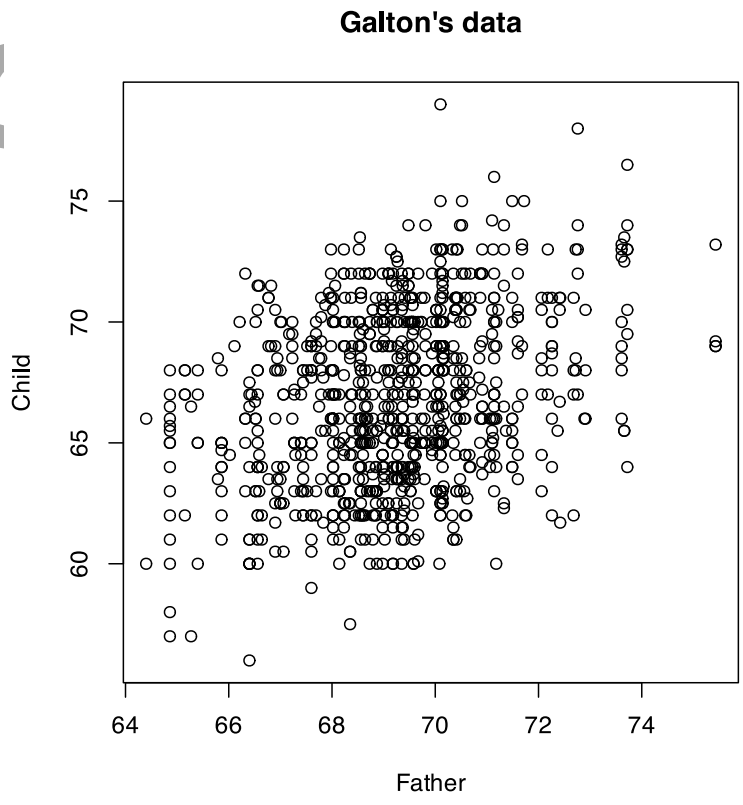

Figure 1.1: Figure: Galton's dataset

Definition 1 테스트용

테스트 정의이다.

II Simple Linear Regression

2 Simple Linear Regression

2.1 Regression analysis

- **Goal**: Find a linear relationship between an explanatory variable (X) and a response variable (Y)
- **Assumptions**

1. Linearity

$$E(Y \mid X = x) = \beta_0 + \beta_1 x \tag{2.1}$$

2. $Y \mid x$ follows a normal distribution
3. Constant variance

$$\text{Var}\ (Y \mid X = x) = \sigma^2 < \infty \tag{2.2}$$

4. Explanatory variable X is a fixed variable (not random)
5. Response variable Y is a random variable with measurement error $\varepsilon \sim (0, \sigma^2)$

- **Simple linear regression model**

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{or} \quad Y = \beta_0 + \beta_1 X + \varepsilon \tag{2.3}$$

2.2 Ordinary least squares (OLS)

With n data points $(x_i, y_i)_{i=1}^n$, our goal is to find the **best** linear fit of the data

$$(x_i, \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i)_{i=1}^n \tag{2.4}$$

Q. What is the **best** fit?
Gauss가 제안한 방식은 다음의 **ordinary least squares (OLS)**이다.

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.5}$$

> **Exercise 2** — **Least absolute deviation (LAD)**
>
> **Least absolute deviation (LAD)**에 대해 조사해보자.

위의 식을 풀기 위해 각각을 β_0, β_1로 미분 후 0이 되는 $\hat{\beta}_0, \hat{\beta}_1$을 찾는 전략을 이용하게 되는데, 여기서 **정규방정식(normal equation)**을 얻게 된다.

$$\begin{cases} -\frac{2}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) & = 0 \\ -\frac{2}{n} \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) & = 0 \end{cases} \tag{2.6}$$

III Nonlinear and Nonparametric Models

Seoncheol Park

3 Boosting

3.1 Boosting: 개요

- Boosting의 가장 큰 특징: base learner를 sequentially하게 fitting함
- Base learner로는 weak learner를 사용: tree를 예로 들면 한 번 정도 split한 tree를 base learner로 사용

https://www.uio.no/studier/emner/matnat/math/STK-IN4300/h22/slides/lect10_modified.pdf

3.2 AdaBoost

3.3 Gradient boosting

부스팅 공부할 만한 자료: https://mlcourse.ai/book/topic10/topic10_gradient_boosting.html

3.3.1 L^2 boosting

- Reference: https://mdporter.github.io/DS6030/lectures/boosting.pdf

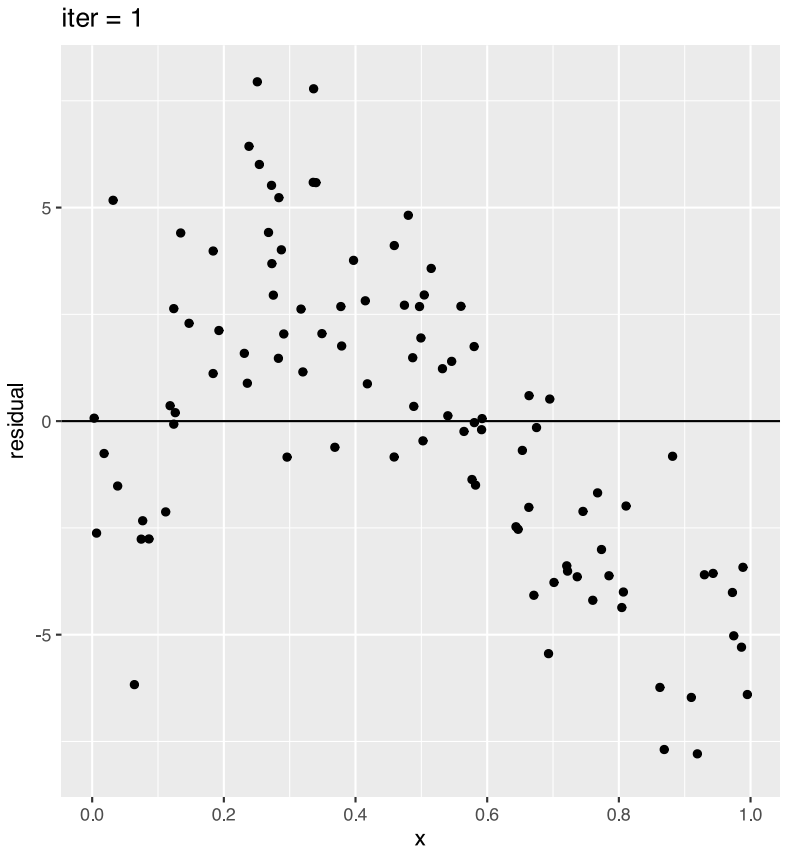

Figure 3.1: L2 boosting

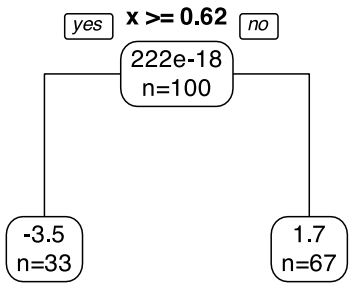

Figure 3.2: L2 boosting

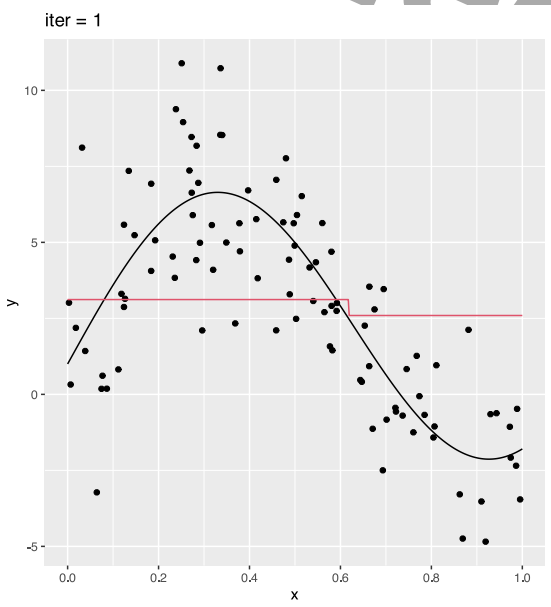

Figure 3.3: L2 boosting

3.3.2 Distributional gradient boosting

- Distirbutional gradient boosting

4 커널회귀

4.1 RKHS

어떤 $n \times p$ 행렬 A가 있을 때 이것의 column space를 $C(A)$라고 하자.
Ronald Christensen은 아래 $C(XX^T) = C(X)$의 결과를 **the fundamental theorem of reproducing kernel Hilbert spaces**라고 부른다.

Definition 3 **Two column spaces are equiv**

For any matrix X, $C(XX^T) = C(X)$.

Proof. Clearly $C(XX^T) \subset C(X)$, so we need to show that $C(X) \subset C(XX^T)$. Let $x \in C(X)$. Then $x = Xb$ for some b. Write $b = b_0 + b_1$, where $b_0 \in C(X^T)$ and $b1 \perp C(X^T)$. Clearly, $Xb_1 = 0$, so we have $x = Xb_0$. But $b_0 = X^T d$ for some d; so $x = Xb_0 = XX^T d$ and $x \in C(XX^T)$.

Definition 4 **Equivalent Linear Models**

If $Y = X_1\beta_1 + e_1$ and $Y = X_2\beta_2 + e_2$ are two models for the same dependent variable vector Y, the models are **equivalent** if $C(X_1) = C(X_2)$.
Since $C(X) = C(XX^T)$, this implies that the linear models $Y = X\beta_1 + e_1$ and $Y = XX^T\beta_2 + e_2$ are equivalent.

RKHS는 p-벡터 x_i를 s-벡터 ϕ_i로 $\phi_i = [\phi_0(x_i), \cdots \phi_{s-1}(x_i)]^T$ 로 변환시킨다. X를 x_i^T들이 행으로 구성된 행렬로 보면 똑같은 논리로 ϕ_i^T가 행으로 구성된 행렬 Φ를 생각할 수 있다. $X^X = [x_i^T x_j]$를 x_i들의 inner products로 만드는 $n \times n$ 행렬로 보면 RKHS는 **reproducing kernel** $R(\cdot, \cdot)$이 존재해

$$\tilde{R} \equiv [R(x_i, x_j)] = [\phi_i^T D(\eta) \phi_j] = \Phi D(\eta) \Phi^T \tag{4.1}$$

가 ϕ_i들의 $n \times n$ inner product matrix이며 $D(\eta)$가 positive definite diagonal matrix가 됨을 말해준다. $D(\eta)$가 positive definite diagonal matrix이므로 PA책 Theorem B.22에 의해 $D(\eta) = QQ^T$인 정방행렬 Q가 존재할 것이고 the fundamental theorem of reproducing kernel Hilbert spaces에 따라 s가 유한하면 $C[\Phi D(\eta) \Phi^T] = C(\Phi)$일 것이다. 따라서 rk 모형

$$Y = \tilde{R}\gamma + e \tag{4.2}$$

를 적합하는 것은 다음의 비모수모형

$$Y = \Phi\beta + e \tag{4.3}$$

를 적합하는 것과 같다. 즉 rk 모형은 $\beta = D(\eta)\Phi^T\gamma$로 reparametrization한 것이다. 특별히 rk 모형을 이요해 예측하는 것은 다음과 같이 하면 된다.

$$\hat{y}(x) = [R(x, x_1), ..., R(x, x_n)]\hat{\gamma}. \tag{4.4}$$

Φ를 가지고 linear structure를 적합하는 것이나 $n \times n$ 행렬 $\tilde{R}$을 이용해 적합하는 것이나 같을 것이고 이를 **kernel trick**이라 한다.

Theorem 5 **Hilbert space가 RKHS가 되기 위한 조건**

A Hilbert space is a RKHS iff the evaluation functionals are continuous.

4.2 Kernel Trick

Kernel trick의 가장 큰 장점은 알려진 함수 $R(\cdot,\cdot)$을 쓰므로 $\tilde{R}$을 만들어내기 쉽다는 것이다. 반대로 $\phi_j(\cdot)$ 함수들에서 s를 specify하는 것은 시간이 더 걸릴 것이다.

또한 $n \times s$행렬 Φ는 s가 크면 이상해지는데, $\tilde{R}$은 항상 $n \times n$이 되어 s가 너무 커질때 이상해지거나 s가 너무 작을때 단순화되는 것을 막아준다.

$s \geq n$이고 x_i들이 distinct (같은 값을 갖는 x들이 없다는 뜻)라면 $\tilde{R}$은 $n \times n$ 이고 rank n인 행렬이며 이것은 saturated model (데이터 수 만큼 모수가 있는 모형)을 만든다. LS estimate는 fitted value가 obs와 같은 자료를 만들 것이며 d.f는 0이 될 것이다. 즉 overfitting이 있는 것인데, 그래서 보통 kernel trick은 penalized (regularized) estimation과 같이 사용하게 된다.

$s \geq n$일 때에는 다른 $R(\cdot,\cdot)$을 선택한다 하더라도, 같은 $C(\tilde{R})$을 주어 같은 모형을 주는 셈이 된다. 즉 같은 least squares fits를 준다. 그러나 parametrization을 다르게 하고 거기에 penalty를 주는 방식 (ridge, LASSO 등)으로 다른 fitted value를 만들어낼 수 있다.

사용하려고 하는 ϕ_j 함수들을 다 알고 있을 경우, rk를 쓰는 이득이 없다. 그러나 ϕ_j를 다루기 어렵거나 $s = \infty$일 경우에는 rks가 도움이 될 것이다.

다음은 많이 쓰이는 rks들을 정리해 놓았다. $\| u - v \|$에만 의존하는 rk들을 **radial basis function** rk라고 부른다.

Names	$R(u,v)$
Polynomial of degree d	$(1+u^Tv)^d$
Polynomial of degree d	$b(c+u^Tv)^d$
Gaussian (radial basis)	$\exp(-b \| u-v \|^2)$
Sigmoid (hyperbolic tangent)	$\tanh\ (bu^Tv+c)$
Linear spline (u,v scalars)	$\min(u,v)$
Cubic spline (u,v scalars)	$\max(u,v)\min^2(u,v)/2 - \min^3(u,v)/6$
Thin plate spline (2 dimensions)	$\| u-v \|^2 \log(\| u-v \|)$

표에 있는 것들 중 hyperbolic tangent는 $\tilde{R}$을 not nonnegatively definite한 것들로 줄 수도 있어 실제로 rk는 아니다. 그러나 u에 대해서 연속인 어떤 $R(u,v)$든지

$$f(x) = \sum_{j=1}^{n} \gamma_j R(x, x_j) \tag{4.5}$$

와 같은 형태의 모형 적합을 유도해낼 수 있기 때문에 이러한 것들을 쓰는 것도 설득력이 있다.

Rk의 아이디어는 함수를 small support를 이용해 근사하는 방법들, 즉 1차원에서 s_*개의 집합으로 나누고 line의 partition을 만들어내는 wavelet, B-spline 등의 방법과 비교할 수 있다. 이러한 방법들은 당연히 p차원에서 s_*^p개의 dimension을 생각해야 하고 고차원에서 다루기 어렵게 된다. 그러나 kernel method에서는 각 data point에 커널을 적합하는 셈이므로 p가 크게 커져도 괜찮다.

4.3 Kernel Trick과 SVM

함수 안에 dot product가 있으면 kernel trick을 쓸 수 있다고 한다. 이러한 것들 중 대표적인 것이 SVM이다. SVM의 objective function은 다음과 같다.

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j, \quad \text{s.t.} \sum_i \alpha_i y_i = 0. \tag{4.6}$$

이 objective 함수 안에는 dot product $\boldsymbol{x}_i^T \boldsymbol{x}_j$가 들어 있고 kernel trick을 쓸 수 있어 SVM이 강력해진다.

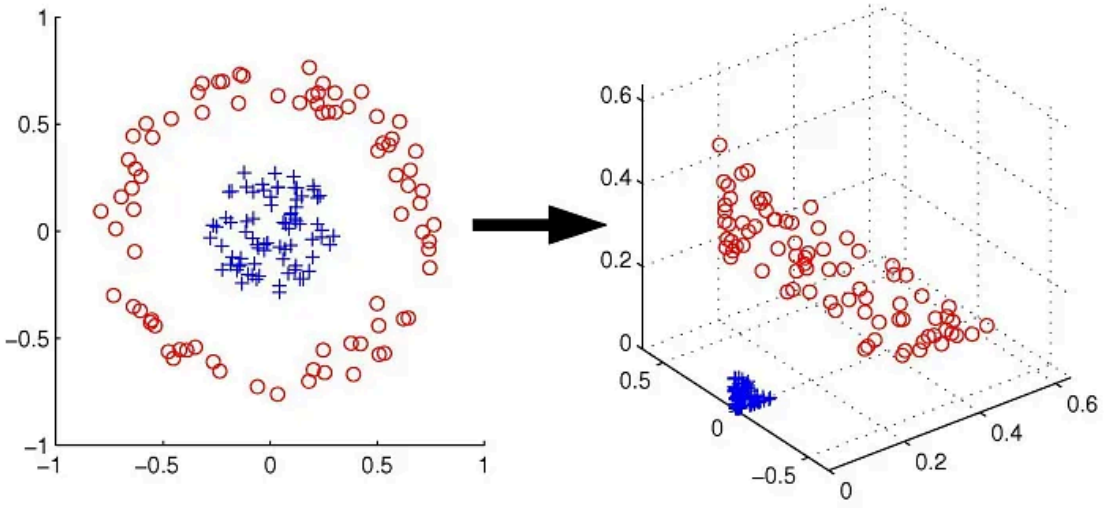

Figure 4.1: Figure: Example of a labeled data inseparable in 2-Dim is separable in 3-Dim.

위 그림에서, 원래 자료 $\boldsymbol{x} = \{x_1, x_2\}$는 2차원에 있는데 이것은 inseparable 하지만 변환

$$\Phi(\boldsymbol{x}) \to (x_1^2, x_2^2, \sqrt{2}x_1x_2) \tag{4.7}$$

를 이용해 오른쪽 그림과 같이 바꾸면 분리가능하다.

앞선 Φ 변환을 이용했을 때 3D 공간에서 decision boundary는 다음과 같다.

$$\beta_0 + \beta_1 x_1^2 + \beta_2 x_2^2 + \beta_3\sqrt{2}x_1x_2 = 0. \tag{4.8}$$

만약 로지스틱과 같은 회귀를 이용한다면 위의 식과 같은 모형을 쓸 것이다. 그러나 SVM에서는 kernel trick을 이용해 decision boundary를 만들 수 있다. 이를 위해 $\langle\Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j)\rangle$의 dot product를 찾아야 한다.

일단 이것을 하려면

- Φ를 정의해야 하고
- Φ 변환 계산시 3×2의 계산
- 그리고 dot product를 계산하는 데 3번 해서

총 9번의 계산이 필요하다. 그러나 만약 커널 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle\boldsymbol{x}_i, \boldsymbol{x}_j\rangle^2$을 이용한다면, 변환 Φ를 찾을 필요도 없고, 그냥 2차원 공간에서 바로 고차원의 similarity measure (dot product)를 만들어낼 수 있다.

로 만들어낼 수 있다. 이것을 하려면

- K는 정의해야 하나
- 두 번째 식까지 두 번의 operation
- 마지막 식에서 제곱을 위해 한 번의 operation

3번의 계산이 필요하다.

다른 예제를 보자. Decision boundary를 다음과 같이 정하였다.

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5\sqrt{2}x_1x_2 = 0. \tag{4.9}$$

$\Phi(\boldsymbol{x}) \to (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$와 같이 5차원 변환 Φ를 이용해 구하는 방법은 총 16번의 계산을 필요로 한다.

그러나 커널 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle 1 + \langle\boldsymbol{x}_i, \boldsymbol{x}_j\rangle\rangle^2$를 이용하면 세 번의 계산으로 된다고 한다.

4.3.1 무한차원에서의 kernel trick

앞선 논리를 그대로 적용하면, kernel trick은 infinite space에서도 유사도를 잴 수 있게 해준다. Gaussian Kernel (RBF), exponential kernel, Laplace kernel 등이 실제로 그러한 역할을 한다. Gaussian kernel은 다음과 같다.

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\Big(-\frac{\| \boldsymbol{x}_i - \boldsymbol{x}_j \|^2}{2\sigma^2}\Big) \tag{4.10}$$

$\sigma = 1$로 두면 위의 Gaussian kernel은 $C = \exp\Big(-\frac{1}{2}\| \boldsymbol{x}_i \|^2\Big)\exp\Big(-\frac{1}{2}\| \boldsymbol{x}_j \|^2\Big)$으로 두었을 때

$$\exp\Big(-\frac{1}{2}\| \boldsymbol{x}_i - \boldsymbol{x}_j \|^2\Big) = C\Big\{1 - \underbrace{\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{1!}}_{\text{1st order}} + \underbrace{\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle^2}{2!}}_{\text{2nd order}} - \underbrace{\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle^3}{3!}}_{\text{3rd order}} + \cdots\Big\} \tag{4.11}$$

이러한 표현은 무한차원으로 확장 가능하며, Gaussian kernel이 무한차원에서의 유사도를 찾을 수 있게 해준다.

물론 커널 기반 방법도 커널을 먼저 정해줘야 하며, cross-validation 등을 이용해 커널 함수를 정하거나 또는 커널에 쓰이는 조율모수를 정하게 된다.

IV Robust and Quantile Regression

Seoncheol Park

5 Robust Regression

5.1 Robust statistics

Q. F가 오염이 되었을 때 우리가 생각하는 추정량 $\hat{\theta}$가 얼마나 영향을 받을 것인가?

- Contaminated distribution function:

$$F_\varepsilon = \varepsilon\delta_y + (1-\varepsilon)F \tag{5.1}$$

  where
  - δ_y: mass 1 to the point y
  - F: a distribution function
- Influence function of $\hat{\theta}$ at F:

$$IF_{\hat{\theta}}(y, F) = \lim_{\varepsilon \to 0} \frac{\hat{\theta}(F_\varepsilon) - \hat{\theta}(F)}{\varepsilon} \tag{5.2}$$

5.2 Robust Regression

- **Robust regression**: 잡음성이 많은 자료에 사용

6 Quantile Regression

6.1 Check loss function

- Check loss function

$$\rho_\tau(u) = u(\tau - I(u < 0)) \tag{6.1}$$

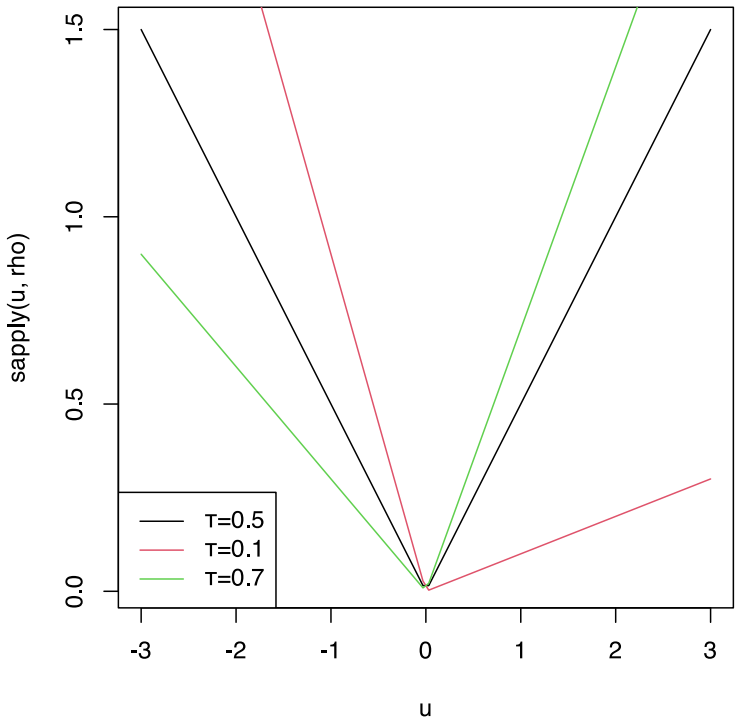

Figure 6.1: Figure: Check loss function.

- Loss function: 값이 클수록 손실이 많음
  - $\tau < 0.5$: $\rho_\tau(u)$는 $u < 0$일 때 큰 가중치
  - $\tau > 0.5$: $\rho_\tau(u)$는 $u > 0$일 때 큰 가중치
- $\rho_\tau(u)$는 $u = 0$에서 미분 불가능

6.2 Estimation

- Objective function

$$R(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\beta) = \sum_{i=1}^{n} \{\ \tau \mid \varepsilon_i \mid I[\varepsilon_i \geq 0] + (1-\tau) \mid \varepsilon_i \mid I[\varepsilon_i < 0]\ \} \tag{6.2}$$

- Directional derivative of R in direction w:

$$\begin{aligned} \nabla R(\beta, w) &= \frac{d}{dt} R(\beta + tw) \mid_{t=0} \\ &= \frac{d}{dt} \end{aligned} \tag{6.3}$$

6.3 Quantile regression as a linear programming

- Linear program: solving

$$\min_{\mathbf{z}} \mathbf{c}^T \mathbf{z} \quad \text{subject to} \quad A\mathbf{z} = \mathbf{b}, \mathbf{z} \geq 0 \tag{6.4}$$

- 선형계획 문제를 풀 때는 $\mathbf{z}$ 부분에 해당하는 값이 양수여야 함. 따라서 분위수 회귀문제를 선형계획으로 풀려면 다음과 같이 ε_i를 slack variable을 이용해 양수와 음수 파트로 나눠야 함.

$$\varepsilon_i = u_i - v_i \tag{6.5}$$

  이때
  - $u_i = \max(0, \varepsilon_i) = \mid \varepsilon_i \mid I[\varepsilon_i \geq 0]$
  - $v_i = \max(0, -\varepsilon_i) = \mid \varepsilon_i \mid I[\varepsilon_i < 0]$

- Then

$$\sum_{i=1}^{n} \rho_\tau(\varepsilon_i) = \sum_{i=1}^{n} \tau u_i + (1-\tau) v_i = \tau \mathbf{1}_n^T \mathbf{u} + (1-\tau) \mathbf{1}_n^T \mathbf{v} \tag{6.6}$$

  여기에서 $\mathbf{u} = (u_1, ..., u_n)^T$, $\mathbf{v} = (v_1, ..., v_n)^T$이다.

- Quantile regression objective function (Equation (6.2)) may be reformulated as a linear program:

$$\min_{\beta, \mathbf{u}, \mathbf{v}} \{\tau \mathbf{1}_n^T \mathbf{u} + (1-\tau) \mathbf{1}_n^T \mathbf{v} \mid \mathbf{y} = \mathbf{X}\beta + \mathbf{u} - \mathbf{v}\} \tag{6.7}$$

- 알고리즘
  - Simplex method
  - Frisch-Newton interior point method
  - Sparse regression quantile fitting

6.3.1 R 코드

- 출처: stackexchange

6.4 Quantile crossing

- 분위수 회귀는 기본적으로 각 τ에 대해 따로 회귀모형을 적합하는 방식이기 때문에 낮은 τ에서의 조건부 분위수 추정값이 높은 τ에서의 조건부 분위수 추정값보다 높은 역전 현상이 발생할 수도 있는데, 이를 **quantile crossing**이라 함
- Quantile crossing을 막으려면 한 번에 conditional distribution 전체를 모델링 해야 하고, GAMLSS 등이 그런 방법을 취하고 있으나, 분포가정을 해야 함

V

Linear Model Asymptotics

Seoncheol Park

7 Asymptotic Theory for Least Squares

기본적인 내용은 B. Hansen [1] 를 따른다.

Theorem 6 **Random sampling assumption**

The variables (Y_i, X_i) are a **random sample** if they are mutually independent and identically distributed (i.i.d.) across $i = 1, ..., n$.

Theorem 7 **Best linear predictor 관련 assumption**

1. $E[Y^2] < \infty$
2. $E \parallel X \parallel^2 < \infty$
3. $Q_{XX} = E[XX^T]$ is positive definite

이 가정의 처음 두 개는 X, Y가 유한한 평균과 분산, 공분산을 갖음을 의미한다. 세 번째는 Q_{XX}의 column들이 linearly independent하고 역행렬이 존재함을 보장한다.
(Q_{XX}가 positive definite일 때 linearly independence는 찾아볼 것)

위의 random sampling과 finite second moment assumption을 가져간채로 least squares estimation에 대한 assumption을 다시 정리한다. (B. Hansen [1] Assumption 7.1)

1. The variables (Y_i, X_i), $i = 1, ..., n$ are i.i.d.
2. $E[Y^2] < \infty$.
3. $E \parallel X \parallel^2 < \infty$.
4. $Q_{XX} = E[XX^T]$ is positive definite.

7.1 Consistency of Least Squares Estimator

이 절의 목표는 $\hat{\beta}$가 β에 consistent함을

1. weak law of large numbers (WLLN)
2. continuous mapping theorem (CMT)

을 이용해 보이는 것이다. (B. Hansen [1] 7.2)
Derivation을 다음과 같은 요소들로 구성된다.

1. OLS estimatior가 sample moment들의 집합의 연속함수로 표현될 수 있다.
2. WLLN을 이용해 sample moments가 population moments에 converge in probability함을 보인다.
3. CMT를 이용해 연속함수에서 converges in probability가 보존됨을 보장한다

그렇다면 먼저 OLS estimator를 다음과 같이 sample moments $\hat{Q}_{XX} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$와 $\hat{Q}_{XX} = \frac{1}{n}\sum_{i=1}^n X_i Y_i$의 함수로 쓸 수 있다.

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i Y_i\right) = \hat{Q}_{XX}^{-1}\hat{Q}_{XY} \tag{7.1}$$

(Y_i, X_i)가 mutually i.i.d. 라는 가정은 (Y_i, X_i)로 구성된, 예를 들면 $X_i X_i^T$와 $X_i Y_i$가 i.i.d.임을 의미한다. 이들은 또한 앞선 Assumption 7.1에 의해 finite expectation을 갖는다. 이러한 조건 하에서, $n \to \infty$일 때 WLLN은

$$\hat{Q}_{XX} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T \xrightarrow{p} E[XX^T] = Q_{XX}, \quad \hat{Q}_{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \xrightarrow{p} E[XY] = Q_{XY}. \tag{7.2}$$

그 다음 continuous mapping theorem을 써서 $\hat{\beta} \to \beta$임을 보일 수 있다는 것이다. $n \to \infty$일 때,

$$\hat{\beta} = \hat{Q}_{XX}^{-1}\hat{Q}_{XY} \xrightarrow{p} Q_{XX}^{-1}Q_{XY} = \beta. \tag{7.3}$$

Stochastic order notation으로 다음과 같이 쓸 수 있다.

$$\hat{\beta} = \beta + o_p(1). \tag{7.4}$$

7.2 Asymptotic Normality

Asymptotic normality를 다룰 때에는

1. 먼저 estimator를 sample moment의 함수로 쓰는 것으로부터 시작한다.
2. 그리고 그것들 중 하나가 zero-mean random vector의 sum으로 표현될 수 있고 이는 CLT를 적용 가능케 한다.

우선 $\hat{\beta} - \beta = \hat{Q}_{XX}^{-1}\hat{Q}_{Xe}$라고 두자. 그리고 이를 $\sqrt{n}$에 곱하면 다음 표현을 얻을 수 있다.

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i\right). \tag{7.5}$$

즉 normalized and centered estimator $\sqrt{n}(\hat{\beta} - \beta)$는 (1) sample average의 함수 $\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right)^{-1}$과 normalized sample average $\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i\right)$의 곱으로 쓸 수 있다.

그러면 뒷부분은 $E[Xe] = 0$이고 이것의 $k \times k$ 공분산함수를 다음과 같이 둘 수 있다.

$$\Omega = E[(Xe)(Xe)^T] = E[XX^T e^2]. \tag{7.6}$$

그리고 아래 가정에서처럼 $\Omega < \infty$라는 가정 하에 $X_i e_i$는 i.i.d. mean zero, 유한한 분산을 갖고 CLT에 의해

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i \xrightarrow{d} \mathcal{N}(0, \Omega). \tag{7.7}$$

(B. Hansen [1] Assumption 7.2)

1. The variables $(Y_i, X_i), i = 1, ..., n$ are i.i.d.
2. $E[Y^4] \underset{4}{<} \infty$.
3. $E \parallel X \parallel < \infty$.
4. $Q_{XX} = E[XX^T]$ is positive definite.

여기서 두 번째 조건이 $\Omega < \infty$임을 의미한다. $\Omega < \infty$임을 보이려면 jl번째 원소 $E[X_j X_l e^2]$이 유한함을 보이면 될 것이다. Properties of Linear Projection Model (B. Hansen [1] Theorem 2.9.6) (If $E \mid Y \mid^r < \infty$ and $E \mid X \mid^r < \infty$ for $r \geq 2$, then $E \mid e \mid^r < \infty$)을 이용해 위의 2, 3번 조건에 의해 $E[e^4] < \infty$임을 보일 수 있다. 그러면 expectation inequality에 의해 Ω의 jl번째 원소는 다음과 같이 bounded된다.

$$\mid E[X_j X_l e^2] \mid \leq E \mid X_j X_l e^2 \mid = E[\mid X_j \mid \mid X_l \mid e^2]. \tag{7.8}$$

Cauchy-Schwarz 부등식을 적용하면 다음과 같다.

$$(E[X_j^2 X_\ell^2])^{1/2}(E[e^4])^{1/2} \leq (E[X_j^4])^{1/4}(E[X_\ell^4])^{1/4}(E[e^4])^{1/2} < \infty. \tag{7.9}$$

Theorem 8

앞선 가정은

$$\Omega < \infty \tag{7.10}$$

를 내포하고

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i e_i \xrightarrow{d} \mathcal{N}(0, \Omega) \tag{7.11}$$

as $n \to \infty$.

식 Equation (7.2) , Equation (7.5) , Equation (7.11) 을 함께 쓰면 다음과 같다.

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathbf{Q}_{XX}^{-1} \mathcal{N}(0, \Omega) = \mathcal{N}(0, \mathbf{Q}_{XX}^{-1} \Omega \mathbf{Q}_{XX}^{-1}) \quad \text{as} \quad n \to \infty. \tag{7.12}$$

여기서 마지막 등식은 normal vector의 linear combination이 normal이라는 것에서부터 왔다.

Theorem 9 **Asymptotic normality of least squares estimator**

앞선 가정 하에서, $n \to \infty$일 때

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_\beta) \tag{7.13}$$

where $\mathbf{Q}_{XX} = E[XX^T]$, $\Omega = E[XX^T e^2]$, and

$$\mathbf{V}_\beta = \mathbf{Q}_{XX}^{-1} \Omega \mathbf{Q}_{XX}^{-1}. \tag{7.14}$$

Stochastic order notation으로 다음과 같이 쓸 수 있다.

$$\hat{\beta} = \beta + O_p(n^{-1/2}). \tag{7.15}$$

이는 식 Equation (7.4) 보다 더 강한 조건이라고 한다.

Remark

- 원래 o_p가 더 강한조건이긴 하나 order의 차이가 나서 저렇게 말하는 듯

- $\mathbf{V}_\beta$: **asymptotic covariance matrix** of $\hat{\beta}$,

$$\mathbf{V}_\beta = \mathbf{Q}_{XX}^{-1} \Omega \mathbf{Q}_{XX}^{-1} \tag{7.16}$$

  이며 $\sqrt{n}(\hat{\beta} - \beta)$의 asymptotic distribution의 variance

Remark

- $\mathbf{V}_\beta$의 형태를 보면 Ω가 $\mathbf{Q}_{XX}^{-1}$ 사이에 끼어있는 형태이므로 **sandwich** form이라고 부름

8 Asymptotic Theory for Quantile Regression

8.1 Basics

Check function

$$\rho_\tau(x) = x(\tau - I\{x < 0\}) = \begin{cases} -x(1-\tau), & x < 0 \\ x\tau, & x \geq 0 \end{cases}. \tag{8.1}$$

$$\psi_\tau(x) = \frac{d}{dx}\rho_\tau(x) = \tau - I\{x < 0\}, \quad x \neq 0. \tag{8.2}$$

Theorem 10 **Consistency of quantile regression estimator**

Assume that (Y_i, X_i) are i.i.d., $E \mid Y \mid < \infty$, $E[\| X \|^2] < \infty$, $f_\tau(e \mid x)$ exists and satisfies $f_\tau(e \mid x) \leq D < \infty$, and the parameter space for β is compact. For any $\tau \in (0,1)$ such that

$$\mathbf{Q}_\tau \stackrel{\text{def}}{=} E[XX^T f_\tau(0 \mid X)] > 0 \tag{8.3}$$

then $\hat{\beta}_\tau \xrightarrow{p} \beta_\tau$ as $n \to \infty$.

Theorem 11 **Asymptotic distribution of quantile regression estimator**

In addition to the assumptions of Theorem 10 , assume that $f_\tau(e \mid x)$ is continuous in e, and β_τ is in the interior of the parameter space. Then as $n \to \infty$

$$\sqrt{n}(\hat{\beta}_\tau - \beta_\tau) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_\tau) \tag{8.4}$$

where $\mathbf{V}_\tau = \mathbf{Q}_\tau^{-1}\Omega_\tau\mathbf{Q}_\tau^{-1}$ and $Q_\tau = E[XX^T\psi_\tau^2]$ for $\psi_\tau = \tau - I\{Y < X^T\beta_\tau\}$.

VI Advanced Topics

Seoncheol Park

9 Spatial Linear Models

- Reference: Spatial Linear Models for Environmental Data

9.1 Linear Model

- A **linear model** for a response variable Y postulates that the response is related to the observed values of p explanatory variables $X_1, ..., X_p$ via this equation:

$$Y = X_1\beta_1 + X_2\beta_2 + \cdots + X_p\beta_p + e, \tag{9.1}$$

where

- $\beta_1, ..., \beta_p$: (unknown) parameters (**coefficients**)
- e: a random variable (**error**)
- $X_1\beta_1 + X_2\beta_2 + \cdots + X_p\beta_p$: mean structure
- When n obs $(y_1, \boldsymbol{x}_1), ..., (y, \boldsymbol{x}_n)$ are taken on the response and explanatory variables, the linear model as just defined implied that

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i, \quad (i = 1, ..., n) \tag{9.2}$$

where

- $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ and $e_1, ..., e_n$ are n possibly correlated random variables.
- Using vector and matrix notations:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}. \tag{9.3}$$

- $\boldsymbol{e}$: **error vector**
- $\boldsymbol{\Sigma}$: $\mathrm{Var}\ (\boldsymbol{e})$

9.2 Spatial Linear Model

- A **spatial linear model** is defined as a linear model for spatial data for which the elements of $\boldsymbol{\Sigma}$ are spatially structured functions of the data sites

9.2.1 Spatial Aitken Model

- The simplest spatial linear model is an extension of the Gauss-Markov model that we call the **spatial Aitken model**:

$$\boldsymbol{\Sigma} = \sigma^2\boldsymbol{R}, \tag{9.4}$$

where

- σ^2: an unknown positive parameter and
- $\boldsymbol{R}$ is a **known** positive definite matrix whose elements are spatially structured functions of the data sites. We refer to $\boldsymbol{R}$ as the **scale-free covariance matrix** of the observations; in some settings, but not all, it is a correlation matrix.
- Estimation: **generalized least squares (GLS)**

Example 12 **A toy example of a spatial Aitken model**

Consider four obs located at the corners of the unit square in $\mathbb{R}^2$, ordered such that the first and last obs are located at opposite corners of the square. Suppose that the model for the obs is given by

$$\boldsymbol{R} = \begin{pmatrix} 1 & e^{-1} & e^{-1} & e^{-\sqrt{2}} \\ e^{-1} & 1 & e^{-\sqrt{2}} & e^{-1} \\ e^{-1} & e^{-\sqrt{2}} & 1 & e^{-1} \\ e^{-\sqrt{2}} & e^{-1} & e^{-1} & 1 \end{pmatrix} \tag{9.5}$$

- **Symmetric**
- **Positive definiteness**: the variance of any linear combinations of obs having this cov matrix, expect the trivial linear combination with all coeffs equal to zero.

9.3 Spatial General Linear Model

- A more flexible spatial linear model
- $\boldsymbol{\Sigma}$ is not fully specified but is given by a **known** spatially structured, matrix-valued parametric function $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is an **unknown** parameter vector.
- Joint parameter space for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ generally taken to be

$$\{(\boldsymbol{\beta}, \boldsymbol{\theta}) : \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m\} \tag{9.6}$$

where

- Θ: the set of vectors $\boldsymbol{\theta}$ for which $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is symmetric and positive definite, or possibly some subset of that set.

Remark 9.1. A spatial Aitken model is a special case of a spatial general linear model, with $\boldsymbol{\theta} \equiv \sigma^2$ and $\boldsymbol{R}$ not functionally dependent on any unknown parameters.

Example 13 **A toy example of a spatial Aitken model (2)**

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sigma^2 \begin{pmatrix} 1 & e^{-1/\alpha} & e^{-1/\alpha} & e^{-\sqrt{2}/\alpha} \\ e^{-1/\alpha} & 1 & e^{-\sqrt{2}/\alpha} & e^{-1/\alpha} \\ e^{-1/\alpha} & e^{-\sqrt{2}/\alpha} & 1 & e^{-1/\alpha} \\ e^{-\sqrt{2}/\alpha} & e^{-1/\alpha} & e^{-1/\alpha} & 1 \end{pmatrix} \tag{9.7}$$

The parameter space for $\boldsymbol{\theta} \equiv (\sigma^2, \alpha)^T$ within which $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is p.d. is

$$\{(\sigma^2, \alpha) : \sigma^2 > 0, \alpha > 0\} \tag{9.8}$$

Small values of α correspond to weak spatial correlation among the obs, and as α increases the spatial correlation among obs become stronger.

10 Gaussian Processes

10.1 Regression analysis

- All finite collection of realizations (n obs) is modeled as having a multivariate normal (MVN) distribution.
- **Mean function**: $\mu(x)$
- **Covariance function**: $\Sigma(x, x')$
- $\Sigma(x, x') = \exp\left\{ - \parallel x - x' \parallel^2 \right\}$
- $\Sigma(x, x) = 1$
- $\Sigma(x, x')$ must be **positive definite**
- Σ_n: covariance matrix (p.d.)

10.2 Splines vs GP

11 PCA and Least Squares

11.1 PCA as least squares problems

PCA can be formulated as follows:

Given m vectors $\boldsymbol{x}_1, ..., \boldsymbol{x}_m \in \mathbb{R}^n$, find matrices $\boldsymbol{U} \in \mathcal{M}_{\mathbb{R}}(k, n)$ and $\boldsymbol{V} \in \mathcal{M}_{\mathbb{R}}(n, k)$ such that

$$\sum_{i=1}^{m} \| \boldsymbol{x}_i - \boldsymbol{V}\boldsymbol{U}\boldsymbol{x}_i \|^2 \tag{11.1}$$

is minimized.

That is, for $k < n$, the vector $\boldsymbol{U}\boldsymbol{x}_i \in \mathbb{R}^k$ is the projection of $\boldsymbol{x}_i$ into a lower-dimensional subspace, and $\boldsymbol{V}\boldsymbol{U}\boldsymbol{x}_i$ is the **reconstructed** original vector. PCA aims to find matrices $\boldsymbol{U}$, $\boldsymbol{V}$ that minimize the reconstruction error as measured by the ℓ^2-norm. It can be shown that, in fact, these matrices are orthogonal and $\boldsymbol{U} = \boldsymbol{V}^T$, so the problem reduces to

$$\arg \min_{V \in \mathcal{M}_{\mathbb{R}}(n,k)} \sum_{i=1}^{m} \| \boldsymbol{x}_i - \boldsymbol{V}\boldsymbol{V}^T\boldsymbol{x}_i \|^2 \tag{11.2}$$

Further manipulations show that $\boldsymbol{V}$ is the matrix whose columns are the eigenvectors corresponding to the k largest eigenvalues of

$$\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^T \tag{11.3}$$

as expected. So indeed, PCA is a least squares method and it is quite sensitive to outliers.

12 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

References

Bibliography

[1] B. Hansen, *Econometrics*. Princeton University Press, 2022, p. 1080.