Preface

Part I

Intro

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")
```

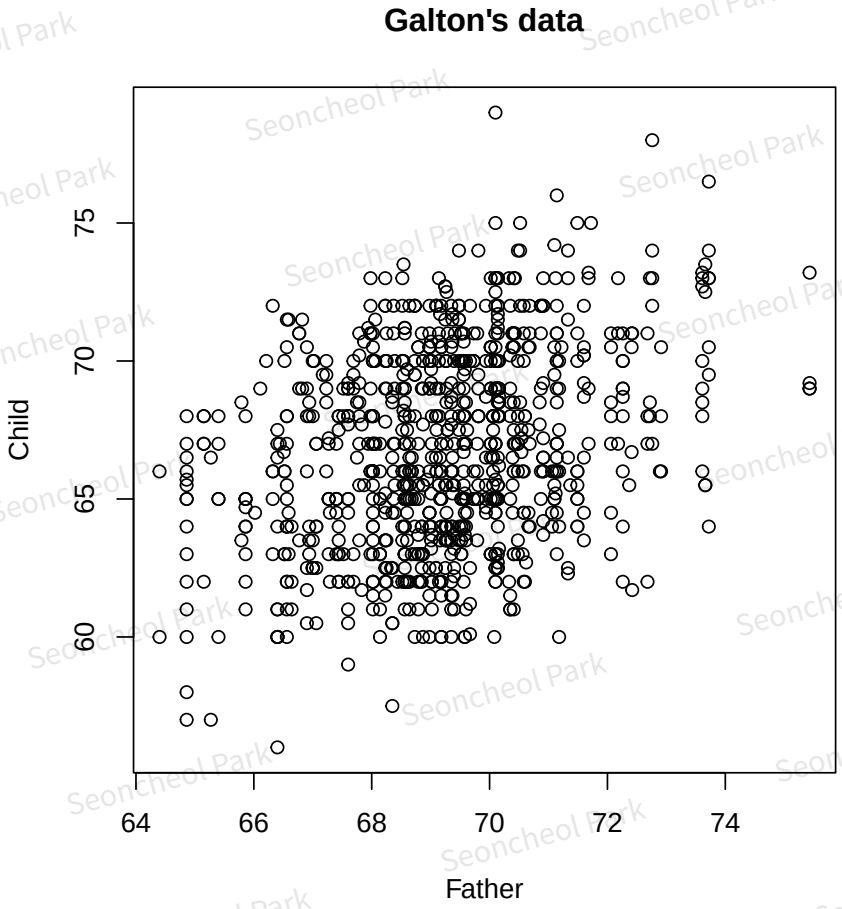

Figure 1.1: Figure: Galton’s dataset

Part 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
$$E(Y|X=x) = \beta_0 + \beta_1 x$$
2. $Y|x$ follows a normal distribution
3. Constant variance
$$\text{Var}(Y|X=x) = \sigma^2 < \infty$$
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**

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

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

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

Q. What is the **best** fit?

Gauss가 제안한 방식은 다음의 **ordinary least squares (OLS)**이다.

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

Exercise 2.1 (Least absolute deviation (LAD)). **Least absolute deviation (LAD)**에 대해 조사해보자.

위의 식을 풀기 위해 각각을 β_0, β_1 로 미분 후 0이 되는 $\hat{\beta}_0, \hat{\beta}_1$ 을 찾는 전략을 이용하게 되는데, 여기서 **정규방정식(normal equation)**을 얻게 된다.

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

Part III

Linear Model Asymptotics

3 Asymptotic Theory for Least Squares

Theorem 3.1 (Random sampling assumption (Hansen (2022) Definition 1.2))**.** The variables (Y_i, X_i) are a **random sample** if they are mutually independent and identically distributed (i.i.d.) across $i = 1, \dots, n$.

Theorem 3.2 (Best linear predictor 관련 assumption (Hansen (2022) Assumption 2.1))**.**

1. $E[Y^2] < \infty$
2. $E\|X\|^2 < \infty$
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. (Hansen (2022) Assumption 7.1)

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

3.1 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. (Hansen (2022) 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 probability 함을 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.

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

(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

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

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

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

Stochastic order notation으로 다음과 같이 쓸 수 있다.

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

3.2 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.

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

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.

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

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

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

(Hansen (2022) Assumption 7.2)

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

$\Omega < \infty$임을 보이려면 jl번째 원소 $E[X_j X_l e^2]$이 유한함을 보이면 될 것이다. Properties of Linear Projection Model (Hansen (2022) Theorem 2.9.6) (If $E|Y|^r < \infty$ and $E|X|^r < \infty$ for $r \geq 2$, then $E|e|^r < \infty$)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.

$$|E[X_j X_l e^2]| \leq E|X_j X_l e^2| = E[|X_j||X_l|e^2].$$

Stochastic order notation으로 다음과 같이 쓸 수 있다.

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

4 Asymptotic Theory for Quantile Regression

4.1 Basics

Check function

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

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

Part IV

Nonlinear and Nonparametric Models

5 Boosting

5.1 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

5.2 AdaBoost

5.3 Gradient boosting

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

5.3.1 L^2 boosting

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

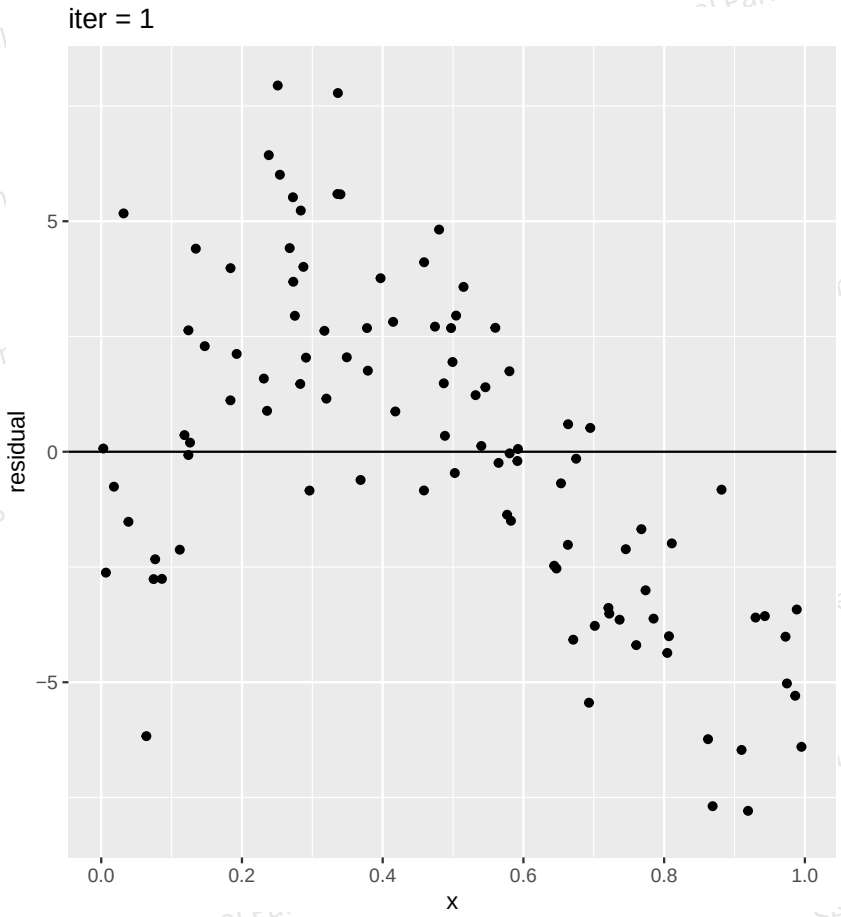

Figure 5.1: L2 boosting

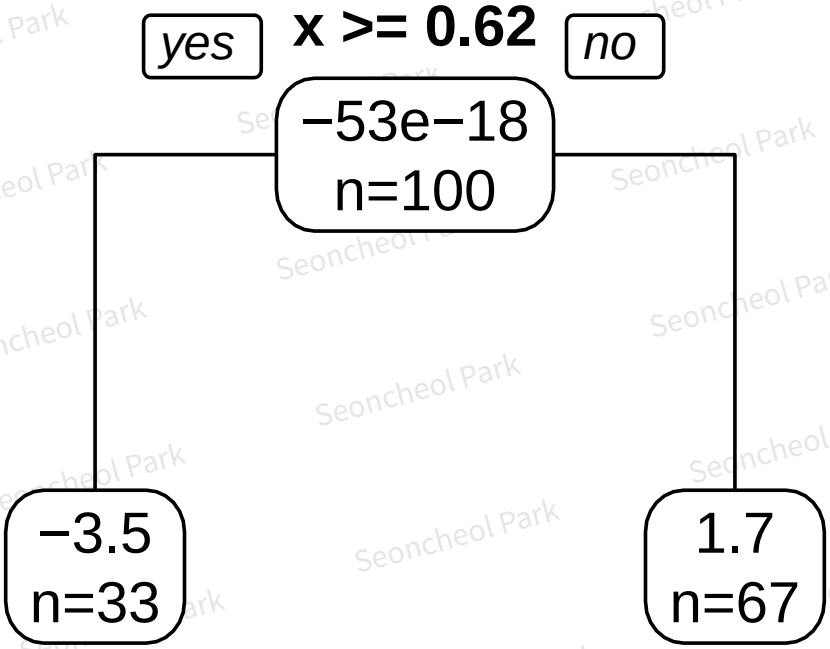

Figure 5.2: L2 boosting

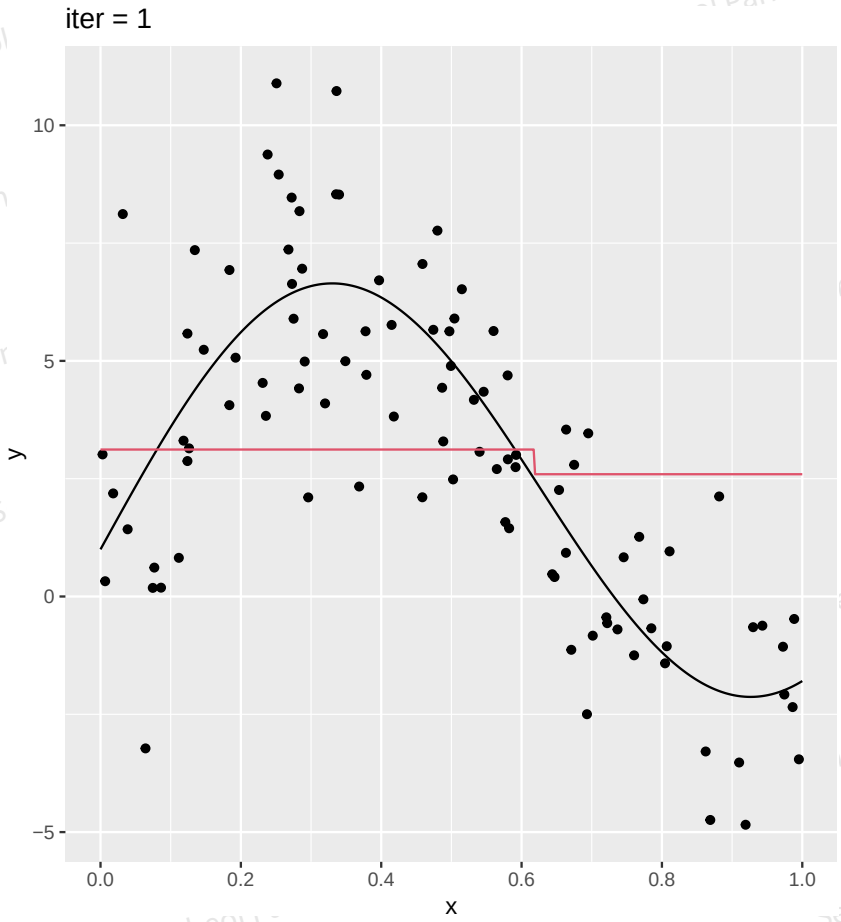

Figure 5.3: L2 boosting

5.3.2 Distributional gradient boosting

- Distirbutional gradient boosting

6 커널회귀

6.1 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 6.1 (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 6.2 (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

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

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

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

를 적합하는 것은 다음의 비모수모형

$$Y = \Phi\beta + e$$

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.

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

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

Theorem 6.1 (Hilbert space가 RKHS가 되기 위한 조건)**.** A Hilbert space is a RKHS iff the evaluation functionals are continuous.

6.2 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.