A Biggner's Guide to Linear Models

Seoncheol Park

2024-05-19

Table of contents

Preface

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")
```

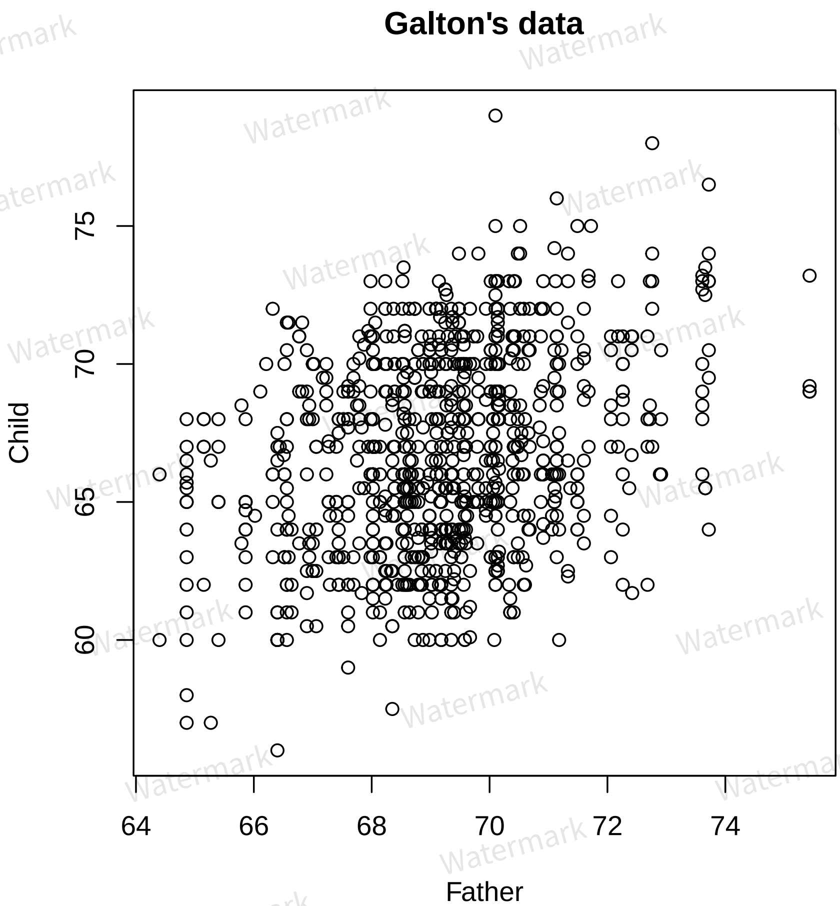

Figure 1.1: Figure: Galton's dataset

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
$$\mathrm{Var}(Y|X=x) = \sigma^2 < \infty$$
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_1로 미분 후 0이 되는 $\hat{\beta}_0, \hat{\beta}_1$을 찾는 전략을 이용하게 되는데, 여기서 **정규방정식(normal equation)**을 얻게 된다.

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

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 $i = 1, \ldots, n$.

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 $(Y_i, X_i), i = 1, \ldots, n$ 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.

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

(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.

$$\sqrt{n}(\hat{\beta}-\beta)=\left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i^T\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_ie_i\right).$$

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

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

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

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

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

(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 $E[X_jX_le^2]$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.

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

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

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

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

5 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

References

Hansen, Bruce. 2022. Econometrics. Princeton University Press.