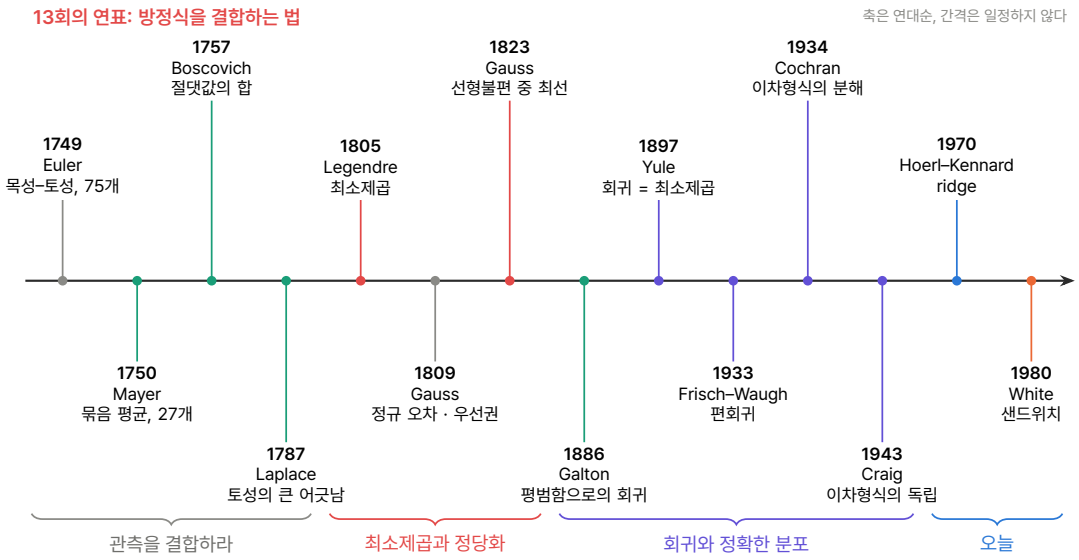
수리통계학 × 통계학의 역사

13회 강의노트

최소제곱
방정식 27개로 미지수 3개를

1749 → 1943

미지수보다 방정식이 많을 때 관측들을 어떻게 결합할 것인가. 18세기 천문학자들은 이 문제를 묶음 평균과 절댓값으로 풀다가, 1805년 Legendre의 최소제곱에서 하나의 답을 얻었다. 이 노트는 최소제곱을 사영의 기하로 다시 세우고, Gauss–Markov 정리로 그 최적성을, Cochran과 Craig의 정리로 정규 오차 아래의 정확한 분포를 증명한 뒤, Galton과 Yule의 회귀로 넘어간다.



다루는 것 선형모형과 결합의 문제 · 정규방정식과 QR · 사영행렬과 지렛대 · Gauss–Markov 정리(BLUE) · 정규표본의 이차형식 · Cochran과 Craig의 정리 · 계수의 t와 일반선형가설의 F · Frisch–Waugh–Lovell · 이분산·공선성·Yule–Simpson · 경사하강과 ridge

원자료 Mayer의 달 칭동 방정식 27개(1750) · Boscovich의 자오선 호 5개(1757) · Laplace의 토성 방정식 24개(1787) · Galton의 부모–자녀 키(1886) · Pearson–Lee(1903) · Berkeley 대학원 입학(1973)

노트북 `sessions/13_least_squares/13_least_squares.ipynb` — 절·정리 번호가 같다

워크북 `13_least_squares_workbook.pdf` (손으로 푸는 34문항) · 풀이판 `_workbook_sol.pdf`

선수 지식 1–12회 (특히 5회 다변량 정규·정리 D, 6회 Student의 정리·t·F, 11회 완비충분통계량, 12회 가능도비 검정)

이 노트를 쓰는 법

이 노트는 다른 책 없이 읽도록 썼다. 쓰이는 정의와 정리는 모두 이 노트나 앞 회차의 노트에서 증명한다. 이 과정의 범위를 넘어 증명하지 않는 사실은 **증명 없이 쓰는 사실** 상자에 따로 밝힌다.

각 절은 **역사 → 정의·정리 → 증명 → 그림 → 연습문제** 순서다. 핵심 정리의 증명은 번호 붙은 단계 ①②③으로 쓰고, 단계마다 오른쪽 끝에 근거(앞 정리의 번호나 정의)를 단다. 읽을 때는 근거를 가리고 "왜 이 단계가 성립하는가"를 먼저 말해 본다. 계수나 가역성처럼 따로 확인할 조건은 **조건 확인** 단락에 밝힌다. 나머지 정리는 **증명 전략**만 적고, 증명을 소문항으로 나눈 유도형 연습으로 돌린다. 이 연습은 건너뛰지 않고 손으로 쓴다.

근거에 붙은 [T1]–[T8]은 별책 **증명 도구 상자**의 도구 번호다. 이 회차에서는 [T1] 원소 추적(열공간과 영공간), [T2] 귀류, [T4] 부등식과 준정치 순서, [T6] 적률생성함수의 계산을 쓴다.

연습문제는 소절 끝에 모여 있다. 난이도는 ★ 계산·확인, ★★ 증명·응용, ★★★ 도전이다. 모든 문제의 풀이가 부록 A에 있고, 문제 끝의 **풀이 ○○쪽**을 누르면 그 쪽으로 간다. 손으로 푸는 문제는 따로 **워크북**(계산 드릴, 빈칸 증명, 증명 문제, 퀄형)에 모았다.

상자의 뜻

> **정의** 대상을 처음 정하는 문장. 왼쪽 선이 보라다.

> **정리** 증명할 주장(정리·보조정리·따름정리·명제). 왼쪽 선이 빨강이다.

> **예제** 숫자를 넣어 끝까지 계산해 보는 곳. 왼쪽 선이 파랑이다.

> **가정 주의** 빼면 정리가 무너지는 가정. 왼쪽 선이 주황이다.

> **증명 없이 쓰는 사실** (예) 이 과정에서 증명하지 않고 쓰는 결과. 무엇이 참인지, 왜 그럴듯한지, 어디서 증명되는지를 밝힌다.

> **역사** 그 개념이 태어난 사람과 문제. 회색 선 옆의 작은 글씨다.

> **연습** 소절 끝의 연습문제. 가는 청록 선이고, 끝에 풀이가 있는 쪽 번호가 붙는다.

색의 뜻 수식과 그림에서 보라는 모수, 청록은 데이터·확률변수, 파랑은 통계량, 빨강은 핵심 결론이다. 노트북과 같은 색이다.

차례

0 들어가며 1

0.1 1–12회에서 가져오는 것 2

0.2 표기 3

1 결합의 문제 (1749–1799) 4

1.1 선형모형 4

1.2 Mayer의 묶음 평균 5

1.3 Boscovich의 절댓값 (1757) 6

2 최소제곱과 사영의 기하 9

2.1 정규방정식 9

2.2 사영행렬 11

2.3 계산: QR 분해 13

3 Gauss–Markov 정리 16

3.1 최소제곱추정량의 적률 16

3.2 최선의 선형불편추정량 17

3.3 정규 오차라면: 모든 불편추정량 가운데 최선 19

3.4 상관된 오차: 일반화 최소제곱 20

4 정규 오차 아래의 분포 — Cochran과 Craig 22

4.1 이차형식이 χ^2일 때 22

4.2 두 이차형식이 독립일 때 23

4.3 Cochran의 정리 25

4.4 계수의 분포와 t 26

4.5 일반선형가설의 F 검정 27

5 회귀의 탄생 — Galton에서 Yule까지 30

5.1 회귀는 최선의 선형예측이다 30

5.2 다른 변수를 고정한다는 것 32

6 원자료로 — 회귀 요약표의 모든 숫자 35

6.1 요약표의 해부 35

6.2 Mayer (1750) — 세 방법의 비교 36

6.3 Laplace의 토성 (1787) 36

7 부숴보기 — 가정이 깨질 때 38

7.1 분산이 다르면 — 샌드위치 38

7.2 한 점이 모든 것을 정하면 — 지렛대와 영향 39

7.3 설명변수가 서로 닮으면 — 공선성 40

7.4 변수 하나를 빼면 — Yule–Simpson의 역설 41

8 오늘의 최소제곱 — 경사하강과 weight decay 43

8.1 경사하강의 속도는 조건수가 정한다 43

8.2 모수가 관측보다 많으면 — 최소노름 해 43

8.3 Ridge와 weight decay 44

9 정리와 역사 노트 47

9.1 한 장 요약 47

9.2 Stigler의 명명 법칙 48

9.3 7 Pillars 지도 48

9.4 다음 회 — 분산분석과 실험설계 (1885 → 1954) 48

A 연습문제 풀이 50

0

들어가며

> **여는 질문.** 미지수가 3개인데 관측 방정식이 27개다. 방정식들은 서로 조금씩 어긋나서 모두를 만족하는 해가 없다. 27개를 어떻게 하나로 결합해야 하는가? 그리고 어떤 결합이 "가장 좋은가"?

1750년 괴팅겐의 천문학자 Tobias Mayer는 달 표면의 분화구 Manilius의 위치를 27일 동안 쟀다. 달은 지구를 향해 같은 면을 보이지만 조금씩 흔들린다(칭동). 그 흔들림을 정하는 세 개의 각을 알고 싶었다. 관측 하나가 세 각의 일차식 하나를 주므로 방정식은 27개였다. 세 개만 골라 풀면 답이 나오지만, 어느 세 개를 고르느냐에 따라 답이 달라진다. 한 해 전 Leonhard Euler는 목성과 토성의 운동에서 같은 종류의 문제를 만났다. 미지수 8개에 방정식 75개였다. Euler는 그 방정식들을 결합하지 못했다. 서로 다른 조건에서 얻은 관측을 더하면 오차가 쌓일 뿐이라고 보았기 때문이다.

Mayer는 반대로 생각했다. 27개를 9개씩 세 묶음으로 나누어 묶음마다 더하면, 미지수 3개의 방정식 3개가 된다. 더하는 동안 오차가 서로 상쇄된다는 것이 그의 직관이었다. 1757년 Boscovich는 다른 원리를 내놓았다. 보정의 절댓값 합을 최소로 하자는 것이다. 1805년 Legendre는 보정의 **제곱**합을 최소로 하자고 제안했다. 이것이 최소제곱이다. 4년 뒤 Gauss는 자신이 1795년부터 이 방법을 써 왔다고 주장했고, 1823년에는 정규분포를 가정하지 않고도 최소제곱이 선형불편추정량 가운데 분산이 가장 작다는 것을 증명했다. 이 결과가 Gauss–Markov 정리다.

이 노트는 그 순서를 따라간다. 1절은 결합의 문제를 선형모형으로 쓰고 Mayer와 Boscovich의 방법을 해부한다. 2절은 최소제곱을 사영의 기하로 세운다. 3절은 Gauss–Markov 정리, 4절은 정규 오차 아래의 정확한 분포(Cochran과 Craig의 정리, t와 F)를 증명한다. 5절은 Galton과 Yule에서 태어난 회귀를 다룬다. 6절은 원자료로 회귀 요약표의 모든 숫자를 직접 계산하고, 7절은 가정이 깨질 때를, 8절은 신경망의 선형층과 weight decay 속의 최소제곱을 본다.

학습 목표

1. 관측의 결합을 선형모형 $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$로 쓰고, Mayer의 묶음 평균과 Boscovich의 절댓값 방법이 어떤 추정량인지 계산한다.
2. 정규방정식과 사영행렬(hat matrix)의 성질을 증명하고, 최소제곱을 QR 분해로 계산한다.

3. Gauss–Markov 정리를 증명하고, 정규 오차 아래에서는 최소제곱추정량이 모든 불편추정량 가운데 최선임을 보인다.
4. 정규표본의 이차형식에 관한 Cochran과 Craig의 정리를 증명하고, 그로부터 계수의 t 검정과 일반선형가설의 F 검정을 유도한다.
5. 회귀 요약표의 모든 숫자를 직접 계산하고, 이분산·지렛대·공선성·Yule–Simpson 역설에서 무엇이 깨지는지 안다.

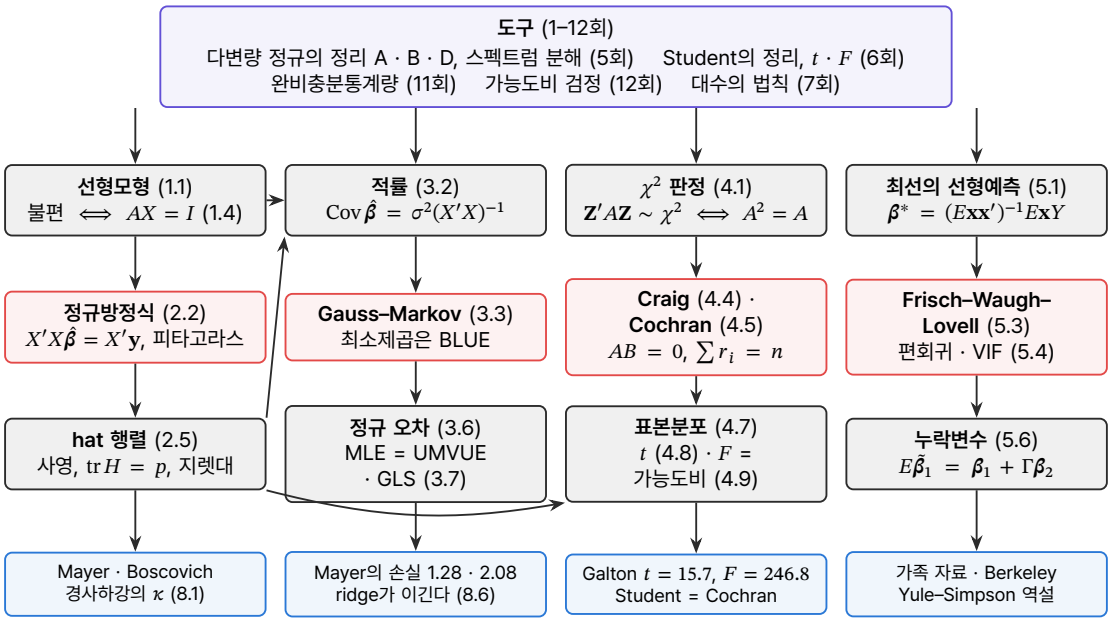


그림 0.1 이번 회차 정리들의 의존 관계. 화살표는 "증명에 쓰인다"는 뜻이다.

0.1 1–12회에서 가져오는 것

앞 회차	내용
5회 보조정리 3.1 · 3.2	선형변환의 평균과 공분산, 대칭행렬의 스펙트럼 분해
5회 정리 3.8 · 3.11 · 3.17	정리 A(정규의 선형변환은 정규), 정리 B(정규에서 무상관은 독립), 정리 D(이차형식은 χ^2)
5회 정리 1.5 · 따름정리 3.19	Gauss의 특성화(평균이 최선이면 오차는 정규), 이변량 정규의 회귀식
6회 정리 1.1 · 2.2, 정의 3.1 · 4.1	표본분산의 $n-1$, Student의 정리, t 분포와 F 분포
11회 정리 3.3 · 4.5	Lehmann–Scheffé 정리, 지수족의 완비충분통계량
12회 정의 3.1, 9회 정리 5.2	가능도비 검정, Wilks 정리

0.2 표기

관측은 n개, 모수는 p개다. **설계행렬** X는 $n \times p$의 상수 행렬이고 i번째 행을 $\mathbf{x}_i'$로 쓴다. 회귀 문헌의 관행대로 반응 벡터는 소문자 굵은 $\mathbf{y} = (y_1, \dots, y_n)'$로 쓰고, 확률벡터로도 같은 글자를 쓴다. $\mathbf{1}$은 모든 성분이 1인 벡터, I_n은 단위행렬이다. $\mathrm{col}(X)$는 X의 열들이 생성하는 $\mathbb{R}^n$의 부분공간, $\mathrm{rank}\, X$는 그 차원이다. $\|\mathbf{v}\|^2 = \mathbf{v}'\mathbf{v}$이고 tr는 대각합이다. 대칭행렬 A, B에 대해 $A \succeq B$는 $A - B$가 양의 준정치(모든 $\mathbf{c}$에 대해 $\mathbf{c}'(A-B)\mathbf{c} \geq 0$)라는 뜻이다.

1

결합의 문제 (1749–1799)

> **역사** 18세기 천문학과 측지학은 같은 벽에 부딪혔다. 이론은 몇 개의 상수로 정해지는 식을 주고, 관측은 그 식에 넣을 숫자를 준다. 관측이 미지수보다 많으면 방정식들이 서로 어긋난다. Euler(1749)는 목성–토성 문제의 75개 방정식 앞에서 멈췄다. Mayer(1750)는 방정식을 묶어 더했고, Boscovich(1757)는 보정의 절댓값 합을 최소로 했으며, Laplace(1787)는 Mayer와 비슷한 결합으로 토성의 "큰 부등"을 풀었다. 이 반세기 동안 "관측을 결합한다"는 생각 자체가 자리 잡았다. Stigler는 이것을 통계학의 첫째 기둥 **Aggregation**의 탄생으로 본다.

1.1 선형모형

정의 1.1 (선형모형) n개의 관측이

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad E\boldsymbol{\varepsilon} = \mathbf{0}, \qquad \operatorname{Cov}\boldsymbol{\varepsilon} = \sigma^2 I_n$$

을 만족한다고 하자. X는 $n \times p$의 상수 행렬이고 $\operatorname{rank} X = p \le n$이다(**완전 계수**). $\boldsymbol{\beta} \in \mathbb{R}^p$와 $\sigma^2 > 0$은 모르는 모수다. 오차에 대한 두 조건(평균 0, 서로 무상관이고 분산이 같음)을 **Gauss–Markov 가정**이라 한다. 정규분포는 가정하지 않는다.

완전 계수는 "미지수가 실제로 p개"라는 뜻이다. 어떤 열이 다른 열들의 일차결합이면 그 열에 해당하는 모수는 자료로 가려낼 수 없다. $p = 1$, $X = \mathbf{1}$이면 $y_i = \mu + \varepsilon_i$, 곧 한 양을 n번 잰 모형이다. 1–12회의 "iid 표본"은 이 모형의 가장 단순한 경우다.

예제 1.2 (Mayer의 27개 방정식) Mayer가 i번째 날 잰 Manilius의 위치에서 계산한 값을 y_i(도)라 하면, 칭동 이론은

$$y_i = -\beta + \alpha\, x_{i2} + (\alpha \sin\theta)\, x_{i3} + \varepsilon_i \qquad (i = 1, \dots, 27)$$

를 준다. β는 달 적도에서 잰 Manilius의 위도이고, α와 θ는 달 적도면의 기울기를 정하는 두 각이다. x_{i2}, x_{i3}은 그날 달의 위치에서 계산되는 삼각함수 값이다. 처음 세 날의 값은 다음과

같다.

i	y_i	x_{i2}	x_{i3}
1	-13.167	0.8836	-0.4682
2	-13.133	0.9996	-0.0282
3	-13.200	0.9899	0.1421

θ가 사인 속에 있어서 모형은 (β, α, θ)에 대해 선형이 아니다. 그러나 $\gamma = \alpha \sin\theta$로 바꿔 쓰면 (β, α, γ)의 선형모형이 된다. X의 첫 열은 모두 -1, 둘째·셋째 열은 x_{i2}, x_{i3}이다. γ를 추정한 뒤 $\hat\theta = \arcsin(\hat\gamma/\hat\alpha)$로 돌아간다. 9회 정리 1.3(MLE의 불변성)과 같은 재모수화다.

정의 1.3 (선형불편추정량) $\tilde{\boldsymbol\beta} = A\mathbf{y}$ (A는 $p \times n$ 상수 행렬) 꼴의 추정량을 **선형추정량**이라 한다. 모든 $\boldsymbol\beta$에서 $E\tilde{\boldsymbol\beta} = \boldsymbol\beta$이면 **선형불편추정량**(LUE)이다.

명제 1.4 (선형불편의 조건) 선형추정량 $A\mathbf{y}$가 불편 $\iff AX = I_p$.

증명. $EA\mathbf{y} = AX\boldsymbol\beta$이다(5회 보조정리 3.1). 이것이 모든 $\boldsymbol\beta$에서 $\boldsymbol\beta$와 같으려면 $\boldsymbol\beta$에 표준기저 $\mathbf{e}_1, \dots, \mathbf{e}_p$를 차례로 넣어 AX의 열이 I_p의 열과 같아야 한다. 거꾸로 $AX = I_p$이면 $AX\boldsymbol\beta = \boldsymbol\beta$이다. ∎

1.2 Mayer의 묶음 평균

Mayer는 27개 방정식을 x_{i2}(곧 α의 계수)의 크기에 따라 9개씩 세 묶음으로 나눴다. 묶음 I은 x_{i2}가 1에 가까운 날(1, 2, 3, 6, 9, 10, 11, 12, 27번), 묶음 II는 x_{i2}가 음수인 날(8, 18, 19, 21–26번), 묶음 III은 나머지다. 묶음마다 방정식을 더하면

$$-9\beta + 8.4987\,\alpha - 0.7932\,\gamma = -118.133,$$

$$-9\beta - 6.1404\,\alpha + 1.7443\,\gamma = -140.283,$$

$$-9\beta + 2.7977\,\alpha + 7.9649\,\gamma = -127.533$$

의 세 방정식이 된다. 오차항은 더한 뒤 버린다.

예제 1.5 (묶음 평균은 선형불편추정량이다) 묶음을 나타내는 27×3 행렬 G(i가 묶음 j에 속하면 $G_{ij} = 1$, 아니면 0)를 쓰면 묶음의 합은 $G'\mathbf{y} = G'X\boldsymbol\beta + G'\boldsymbol\varepsilon$이고, Mayer의 해는

$$\tilde{\boldsymbol\beta}_{\mathrm{M}} = (G'X)^{-1}G'\mathbf{y}$$

이다. $A = (G'X)^{-1}G'$로 두면 $AX = I_3$이므로 명제 1.4로 선형불편이다. 위의 세 방정식을 풀면 $\hat\beta = 14.547^\circ = 14^\circ 32.8'$, $\hat\alpha = 1.4958^\circ = 1^\circ 29.7'$, $\hat\gamma = -0.0996$이고 $\hat\theta = \arcsin(\hat\gamma/\hat\alpha) = -3^\circ 49.1'$이다(연습 1.1). 아무 세 방정식을 골라 푸는 것도 G가 선택 행렬인 경우이므로 역시

선형불편이다. 27개에서 3개를 고르는 방법은 $\binom{27}{3} = 2925$가지다.

모든 선형불편추정량은 평균적으로 옳다. 그렇다면 무엇으로 고르나? 답은 분산이다. Mayer는 묶어 더하면 정확도가 결합한 방정식의 수에 비례해 좋아진다고 썼다. 실제로는 표준편차가 그 수의 제곱근에 비례해 줄어든다(연습 1.2). 비례 상수가 틀렸어도, "더하면 좋아진다"는 직관은 옳았다. 3절의 Gauss–Markov 정리는 모든 선형불편추정량 가운데 분산이 가장 작은 것이 최소제곱임을 보인다. Mayer의 해는 그보다 분산이 크다는 것도 거기서 계산한다.

1.3 Boscovich의 절댓값 (1757)

Newton은 지구가 극 쪽으로 납작하다고 예측했다. 그렇다면 위도 1도에 해당하는 자오선 호의 길이는 극으로 갈수록 길어야 한다. 위도 L에서의 길이는 근사적으로

$$a(L) = a_0 + a_1 \sin^2 L$$

이고, 납작한 정도(편평률)는 $f \approx a_1/(3a_0)$이다. 1750년대까지 다섯 곳에서 이 길이를 쟀다. 미지수 2개에 방정식 5개다.

장소	위도	$x = \sin^2 L$	y (toise)	Boscovich 해의 잔차
Quito	0°0′	0	56751	0.0
희망봉	33°18′	0.2987	57037	79.2
로마	42°59′	0.4648	56979	−93.8
파리	49°23′	0.5762	57074	−75.9
Lapland	66°19′	0.8386	57422	90.5

Boscovich는 직선 $y = b_0 + b_1 x$에 두 조건을 걸었다. (i) 보정(잔차)의 합이 0이다. (ii) 그 조건 아래 보정의 절댓값 합이 가장 작다.

명제 1.6 (Boscovich의 해는 가중 중앙값이다) x_i들이 모두 같지는 않다고 하자. 조건 (i)을 만족하는 직선은 무게중심 $(\bar{x}, \bar{y})$를 지난다. 그 위에서 절댓값 합은

$$\sum_{i=1}^{n} |y_i - b_0 - b_1 x_i| = \sum_{i:\, x_i \neq \bar{x}} |x_i - \bar{x}|\,|s_i - b_1| + \sum_{i:\, x_i = \bar{x}} |y_i - \bar{y}|, \qquad s_i = \frac{y_i - \bar{y}}{x_i - \bar{x}}$$

이고, 이를 최소로 하는 b_1은 기울기 s_i들의 **가중 중앙값**(무게 $w_i = |x_i - \bar{x}|$)이다. 곧 s_i를 크기 순으로 늘어놓고 무게를 누적했을 때 누적 무게가 처음으로 전체의 절반 이상이 되는 s_i다.

증명. (i) $\sum (y_i - b_0 - b_1 x_i) = 0$은 $b_0 = \bar{y} - b_1 \bar{x}$와 같다. 대입하면 $y_i - b_0 - b_1 x_i = (y_i - \bar{y}) - b_1 (x_i - \bar{x})$이고, $x_i \neq \bar{x}$인 항은 $|x_i - \bar{x}|\,|s_i - b_1|$로 쓸 수 있다. 나머지 항은 b_1과 무관하다.

$g(b) = \sum w_i |s_i - b|$는 볼록한 조각별 일차함수이고, s_i들이 아닌 점에서의 기울기는 $g'(b) = \sum_{s_i < b} w_i - \sum_{s_i > b} w_i$이다. 이 값은 b가 커지면서 $-\sum w_i$에서 $+\sum w_i$까지 늘어나고, 처음으로 0 이상이 되는 곳이 최소점이다. 그 점이 누적 무게가 절반을 넘는 s_i, 곧 가중 중앙값이다. ∎

예제 1.7 (다섯 개의 호) 무게중심은 $(\bar{x}, \bar{y}) = (0.4357,\ 57052.6)$이다. 다섯 기울기를 크기 순으로 쓰면 로마 −2525.7 (무게 0.029), 희망봉 113.9 (0.137), 파리 152.3 (0.141), Quito 692.3 (0.436), Lapland 916.8 (0.403)이다. 누적 무게의 비율은 0.025, 0.145, 0.268, 0.648, 1이므로 가중 중앙값은 Quito의 기울기 692.3이다. Boscovich의 직선은 $y = 56751 + 692.3\,x$이고 편평률은 1/246이다. 최소제곱 직선은 $y = 56737.4 + 723.4\,x$ (편평률 1/235)이고, 조건 (i) 없이 절댓값 합만 최소로 하면 Quito와 Lapland를 지나는 $y = 56751 + 800.1\,x$ (1/213)이다. 오늘의 값은 약 1/298이다. 다섯 측정 자체에 계통오차가 있었다(그림 1.1).

가정 주의 표의 희망봉 값 0.2987은 Boscovich가 위도를 33°8′로 잘못 넣어 계산한 것이다. 33°18′이면 0.3014다. 이 오류는 결론을 거의 바꾸지 않는다(최소제곱 편평률 1/235.3 → 1/234.9). 이 노트는 역사적 계산을 재현하려고 원래 값을 쓴다.

조건 (ii)만 쓰는 **최소절댓값**(LAD) 추정은 1799년 Laplace가 알고리즘으로 정리했고, 1978년 Koenker와 Bassett가 분위수 회귀로 일반화했다. 다음 명제는 그 해가 왜 늘 몇 개의 관측을 정확히 지나는지 설명한다.

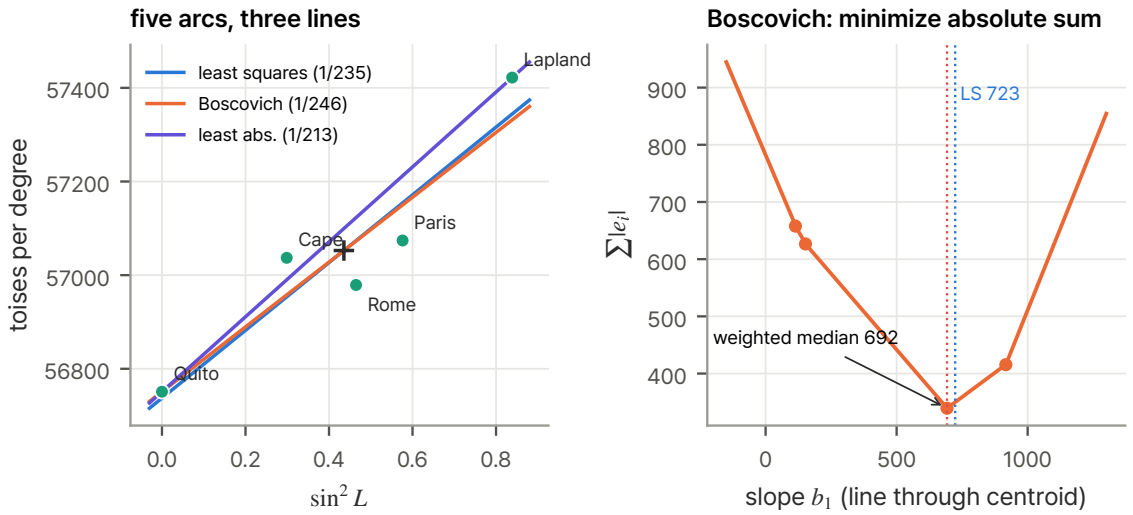


그림 1.1 Boscovich의 다섯 호. 왼쪽: 최소제곱(파랑), Boscovich(주황), 제약 없는 최소절댓값(보라)의 세 직선. 오른쪽: 무게중심을 지나는 직선들의 절댓값 합 $g(b_1)$. 꺾이는 점이 다섯 기울기 s_i이고, 최소점은 가중 중앙값인 Quito의 기울기 692.3이다.

명제 1.8 (최소절댓값 해는 관측 p개를 지난다) $\operatorname{rank} X = p$이면 $\sum_i |y_i - \mathbf{x}_i'\mathbf{b}|$를 최소로 하는 $\mathbf{b}$ 가운데, $\mathbf{x}_i$들이 일차독립인 관측 p개에서 잔차가 0인 것이 있다.

증명. p개의 초평면 $\{\mathbf{b} : \mathbf{x}_i'\mathbf{b} = y_i\}$로 $\mathbb{R}^p$를 나누면, 각 조각 안에서 잔차의 부호가 고정되므로 목적함수는 일차함수다. $\mathbf{x}_i$들이 $\mathbb{R}^p$를 생성하므로 각 조각은 직선을 포함하지 않는 다면체이고, 꼭짓점은 일차독립인 초평면 p개의 교점이다. 목적함수는 0 이상이므로 각 조각에서 아래로 유계이고, 아래로 유계인 일차함수는 그런 다면체 위에서 꼭짓점에서 최솟값을 갖는다(선형계획법의 기본 정리). 전체 최솟값은 어느 조각의 최솟값이므로 어느 꼭짓점에서 달성된다. ∎

다섯 호에서 제약 없는 해가 Quito와 Lapland를 지나는 것, Boscovich의 해가 무게중심과 Quito를 지나는 것이 이 명제의 모습이다. Mayer의 자료에서 최소절댓값 해는 $\hat\beta = 14°34.1'$, $\hat\alpha = 1.515°$이고 세 방정식을 정확히 지난다(6절).

연습문제

연습 1.1 (Mayer의 해 ★) 본문의 묶음 방정식 세 개를 풀어 예제 1.5의 $\hat\beta, \hat\alpha, \hat\gamma, \hat\theta$를 확인하라. 풀이 50쪽

연습 1.2 (정확도는 제곱근으로 ★) $\varepsilon_1, \dots, \varepsilon_m$이 서로 무상관이고 분산이 σ^2이면 평균 $\bar\varepsilon$의 표준편차가 $\sigma/\sqrt{m}$임을 보여라. Mayer의 묶음 하나($m = 9$)를 더해 9로 나누면 표준편차는 몇 분의 1이 되는가? 풀이 50쪽

연습 1.3 (위치모형의 세 방법 ★) $y_i = \mu + \varepsilon_i$ ($X = \mathbf{1}$)에서 (a) 최소제곱 해는 $\bar y$, (b) 최소절댓값 해는 표본 중앙값, (c) Boscovich의 조건 (i)만으로 해가 $\bar y$로 정해짐을 보여라. 풀이 50쪽

연습 1.4 (선형계획법 ★★) 최소절댓값 문제가 선형계획법

$$\min \sum_{i=1}^{n} (u_i + v_i) \quad \text{s.t.} \quad X\mathbf{b} + \mathbf{u} - \mathbf{v} = \mathbf{y}, \;\; \mathbf{u}, \mathbf{v} \ge \mathbf{0}$$

과 같음을 보여라. 최적해에서 각 i마다 $u_i v_i = 0$임을 보여라. 풀이 50쪽

연습 1.5 (세 방정식만 고르기 ★★) Mayer의 방정식에서 세 개를 골라 정확히 풀어 얻는 추정량을 $\tilde{\boldsymbol\beta}_S = X_S^{-1}\mathbf{y}_S$라 하자($X_S$는 고른 세 행). $AX = I_3$인 A를 써서 이것이 선형불편임을 보이고, $\operatorname{Cov}\tilde{\boldsymbol\beta}_S = \sigma^2 (X_S'X_S)^{-1}$임을 보여라. 2925가지 가운데 $\det X_S$가 0에 가까운 것은 왜 나쁜가? 풀이 51쪽

2

최소제곱과 사영의 기하

> **역사** 1805년 Adrien-Marie Legendre는 혜성 궤도를 정하는 책의 부록에 **최소제곱법에 대하여**를 실었다. "제안될 수 있는 모든 원리 가운데 오차의 제곱합을 최소로 하는 것보다 더 일반적이고, 더 정확하며, 더 쓰기 쉬운 것은 없다고 생각한다. 이렇게 하면 오차들 사이에 일종의 평형이 세워져 극단의 오차가 지나친 힘을 쓰지 못하고, 진실에 가장 가까운 상태가 드러난다." 정당화는 없었다. 쓰기 쉽다는 것이 거의 전부였다. 미분하면 미지수 수만큼의 일차방정식이 나오고, 그것은 손으로 풀 수 있었다. 4년 뒤 Gauss는 **천체 운동론**(1809)에서 이 방법을 "1795년부터 써 온 우리의 원리"라고 불렀다. Legendre는 분개했고, 우선권 다툼은 오래 이어졌다.

2.1 정규방정식

정의 2.1 (최소제곱추정량) $S(\mathbf{b}) = \|\mathbf{y} - X\mathbf{b}\|^2 = \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\mathbf{b})^2$을 최소로 하는 $\hat{\boldsymbol{\beta}}$을 **최소제곱추정량**(LSE)이라 한다. $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$을 **적합값**, $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$를 **잔차**, $\mathrm{RSS} = \|\mathbf{e}\|^2$를 잔차제곱합이라 한다.

정리 2.2 (정규방정식) (a) $\hat{\boldsymbol{\beta}}$이 S를 최소로 한다 $\iff$ $X'X\hat{\boldsymbol{\beta}} = X'\mathbf{y}$ (**정규방정식**). (b) 정규방정식은 언제나 해를 갖는다. (c) $\operatorname{rank} X = p$이면 해는 하나뿐이고 $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$이다.

증명. **전략** (a)는 피타고라스 항등식으로 보인다. 정규방정식을 푸는 $\hat{\boldsymbol{\beta}}$에서는 잔차가 $\mathrm{col}(X)$와 직교하므로, 다른 점 $X\mathbf{b}$까지의 거리는 늘어나기만 한다. (b)는 $X'X$와 X의 영공간이 같다는 것에서 나온다.

(a) ⇐

① $X'(\mathbf{y} - X\hat{\boldsymbol{\beta}}) = \mathbf{0}$라 하고 아무 $\mathbf{b}$를 잡으면 $\mathbf{y} - X\mathbf{b} = (\mathbf{y} - X\hat{\boldsymbol{\beta}}) + X(\hat{\boldsymbol{\beta}} - \mathbf{b})$이다. 더하고 빼기 [T4]

② 제곱하면 교차항은 $2(\hat{\boldsymbol{\beta}} - \mathbf{b})'X'(\mathbf{y} - X\hat{\boldsymbol{\beta}}) = 0$이다. 가정

③ 따라서

$$S(\mathbf{b}) = S(\hat{\boldsymbol{\beta}}) + \|X(\mathbf{b} - \hat{\boldsymbol{\beta}})\|^2 \;\geq\; S(\hat{\boldsymbol{\beta}}). \qquad ①②$$

(a) ⇒

① $S(\mathbf{b}) = \mathbf{y}'\mathbf{y} - 2\mathbf{b}'X'\mathbf{y} + \mathbf{b}'X'X\mathbf{b}$는 $\mathbf{b}$의 이차식이고 $\nabla S(\mathbf{b}) = -2X'(\mathbf{y} - X\mathbf{b})$이다.
$\nabla \mathbf{b}'M\mathbf{b} = 2M\mathbf{b}$ (M 대칭)

② 최소점에서 기울기는 $\mathbf{0}$이므로 $X'X\hat{\boldsymbol{\beta}} = X'\mathbf{y}$이다. 극값의 필요조건

(b)

① $X'X\mathbf{v} = \mathbf{0}$이면 $\|X\mathbf{v}\|^2 = \mathbf{v}'X'X\mathbf{v} = 0$이므로 $X\mathbf{v} = \mathbf{0}$이다. 거꾸로는 바로 나오므로 $X'X$와 X의 영공간은 같다. 원소 추적 [T1]

② $\operatorname{rank} X'X = p - \dim \operatorname{null}(X) = \operatorname{rank} X = \operatorname{rank} X'$이다. ①, 차원 정리

③ $\operatorname{col}(X'X) \subseteq \operatorname{col}(X')$이고 두 공간의 차원이 같으므로 $\operatorname{col}(X'X) = \operatorname{col}(X')$이다. ②

④ $X'\mathbf{y} \in \operatorname{col}(X') = \operatorname{col}(X'X)$이므로 $X'X\mathbf{b} = X'\mathbf{y}$를 푸는 $\mathbf{b}$가 있다. ③

(c)

① $\operatorname{rank} X = p$이면 (b)의 ②로 $\operatorname{rank} X'X = p$이므로 $p \times p$ 행렬 $X'X$는 가역이고, 해는 $(X'X)^{-1}X'\mathbf{y}$ 하나다. 가역행렬

■

(a)의 항등식 $S(\mathbf{b}) = S(\hat{\boldsymbol{\beta}}) + \|X(\mathbf{b} - \hat{\boldsymbol{\beta}})\|^2$은 이 회차 전체에서 되풀이된다. 최소제곱은 피타고라스 정리다.

따름정리 2.3 (상수항이 있으면) $\mathbf{1} \in \mathrm{col}(X)$이면(예컨대 X의 한 열이 $\mathbf{1}$이면) $\sum_i e_i = 0$이고, 적합한 평면은 무게중심 $(\bar{\mathbf{x}}, \bar{y})$를 지난다.

증명. 정규방정식은 $X'\mathbf{e} = \mathbf{0}$, 곧 $\mathbf{e}$가 X의 모든 열과 직교한다는 뜻이다. $\mathbf{1} = X\mathbf{a}$이면 $\mathbf{1}'\mathbf{e} = \mathbf{a}'X'\mathbf{e} = 0$이다. $\sum_i (y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}) = 0$을 n으로 나누면 $\bar{y} = \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}$이다. ∎

Boscovich의 조건 (i)(보정의 합이 0)은 최소제곱에서 저절로 성립한다. Legendre가 말한 "평형"이 이것이다.

2.2 사영행렬

보조정리 2.4 (대칭 멱등행렬) P가 $n \times n$ 대칭행렬이고 $P^2 = P$이면

(a) P의 고유값은 0 또는 1이고, $\operatorname{rank} P = \operatorname{tr} P$이다.

(b) $r = \operatorname{rank} P$이면 $U'U = I_r$인 $n \times r$ 행렬 U가 있어 $P = UU'$이고 $\mathrm{col}(P) = \mathrm{col}(U)$이다.

(c) $I - P$도 대칭 멱등이고 $\operatorname{rank}(I-P) = n - r$이다.

(d) 모든 $\mathbf{v}$에 대해 $\mathbf{v} = P\mathbf{v} + (I-P)\mathbf{v}$이고 두 조각은 직교한다. $P\mathbf{v}$는 $\mathrm{col}(P)$에서 $\mathbf{v}$에 가장 가까운 점이다.

증명. **전략** 스펙트럼 분해 $P = V\Lambda V'$(5회 보조정리 3.2)에서 $P^2 = P$를 고유값의 식 $\lambda_i^2 = \lambda_i$로 바꾼다. 나머지는 고유값 1의 고유벡터들로 P를 다시 쓰고 피타고라스를 쓴다. 단계별 증명은 연습 2.7(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 52쪽에 있다.

이런 P를 $\mathrm{col}(P)$ 위로의 **직교사영행렬**이라 한다.

정리 2.5 (hat 행렬) $\operatorname{rank} X = p$일 때 $H = X(X'X)^{-1}X'$라 하자.

(a) H는 대칭 멱등이고 $\mathrm{col}(H) = \mathrm{col}(X)$, $\operatorname{rank} H = \operatorname{tr} H = p$이다. 곧 H는 $\mathrm{col}(X)$ 위로의 직교사영이다.

(b) $\hat{\mathbf{y}} = H\mathbf{y}$, $\mathbf{e} = (I-H)\mathbf{y}$이고 $X'\mathbf{e} = \mathbf{0}$, $\hat{\mathbf{y}}'\mathbf{e} = 0$, $\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{e}\|^2$이다.

(c) $\mathrm{RSS} = \mathbf{y}'(I-H)\mathbf{y}$이고 $I-H$는 $\mathrm{col}(X)^{\perp}$ 위로의 사영으로 계수가 $n-p$이다.

(d) T가 $p \times p$ 가역행렬이면 XT로 만든 hat 행렬도 H이다. 적합값과 잔차는 모수를 어떻게 매기느냐와 무관하다.

증명. **전략** 대칭과 멱등은 곱을 직접 계산해 확인하고, 열공간은 두 방향의 포함으로 보인다 [T1]. 계수는 대각합 $\operatorname{tr} AB = \operatorname{tr} BA$로 구하고 보조정리 2.4를 쓴다. 단계별 증명은 연습 2.8(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 52쪽에 있다.

$\mathbf{y}$는 n차원 공간의 점이고 $\mathrm{col}(X)$는 그 안의 p차원 평면이다. 최소제곱은 $\mathbf{y}$에서 평면까지 수선을

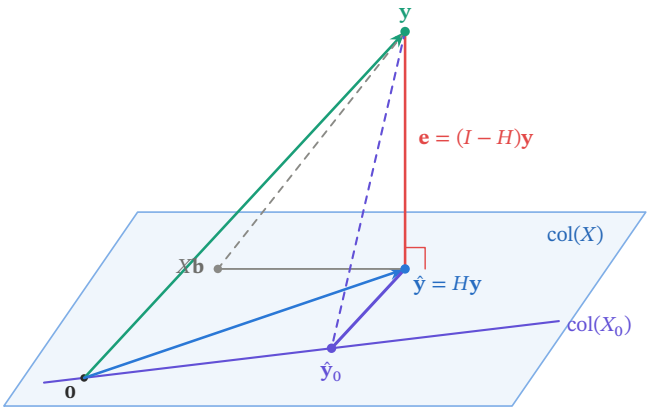


그림 2.1 최소제곱의 기하. 적합값 $\hat{\mathbf{y}} = H\mathbf{y}$는 $\mathbf{y}$를 평면 $\mathrm{col}(X)$에 수직으로 내린 발이고, 잔차 $\mathbf{e}$는 평면과 직교한다. 평면 위의 다른 점 $X\mathbf{b}$까지의 거리는 피타고라스로 $\|\mathbf{y}-X\mathbf{b}\|^2 = \|\mathbf{e}\|^2 + \|\hat{\mathbf{y}}-X\mathbf{b}\|^2$이다. 평면 안의 직선 $\mathrm{col}(X_0)$는 4절의 귀무가설 모형이다. $\mathbf{y}$를 그 위로 내린 $\hat{\mathbf{y}}_0$에 대해 $\mathrm{RSS}_0 - \mathrm{RSS} = \|\hat{\mathbf{y}}-\hat{\mathbf{y}}_0\|^2$이고, 이것이 F 통계량의 분자다.

내리는 일이다(그림 2.1). 5회 6절에서 본 Pearson(1901)의 주성분이 **수직** 거리를 최소화했다면, 최소제곱은 y 방향의 거리를 최소화한다. 그림 속 공간은 관측 하나하나가 축인 공간이지, 산점도의 평면이 아니다.

명제 2.6 (결정계수) $\mathbf{1} \in \mathrm{col}(X)$이면 $\mathrm{TSS} = \sum(y_i-\bar y)^2$, $\mathrm{ESS} = \sum(\hat y_i - \bar y)^2$에 대해

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}, \qquad R^2 := \frac{\mathrm{ESS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = \mathrm{Corr}(\mathbf{y},\hat{\mathbf{y}})^2$$

이다. 여기서 Corr는 두 벡터의 표본상관계수다.

증명. $H\mathbf{1} = \mathbf{1}$이므로 $\hat{\mathbf{y}} - \bar y\mathbf{1} = H(\mathbf{y}-\bar y\mathbf{1})$이고 $\mathbf{e} = (I-H)(\mathbf{y}-\bar y\mathbf{1})$이다. 보조정리 2.4(d)로 $\|\mathbf{y}-\bar y\mathbf{1}\|^2 = \|\hat{\mathbf{y}}-\bar y\mathbf{1}\|^2 + \|\mathbf{e}\|^2$. 따름정리 2.3로 $\hat{\mathbf{y}}$의 평균도 $\bar y$이고, $\mathbf{e} \perp (\hat{\mathbf{y}}-\bar y\mathbf{1})$이므로 $(\mathbf{y}-\bar y\mathbf{1})'(\hat{\mathbf{y}}-\bar y\mathbf{1}) = \mathrm{ESS}$이다. 따라서 $\mathrm{Corr}(\mathbf{y},\hat{\mathbf{y}})^2 = \mathrm{ESS}^2/(\mathrm{TSS}\cdot\mathrm{ESS}) = \mathrm{ESS}/\mathrm{TSS}$. ∎

명제 2.7 (지렛대) $h_{ii} = \mathbf{x}_i'(X'X)^{-1}\mathbf{x}_i$를 관측 i의 **지렛대**(leverage)라 한다. $\hat y_i = \sum_j h_{ij}y_j$이므로 $h_{ii} = \partial\hat y_i/\partial y_i$이다.

(a) $0 \le h_{ii} \le 1$이고 $\sum_i h_{ii} = p$이다.

(b) $\mathbf{1} \in \mathrm{col}(X)$이면 $h_{ii} \ge 1/n$이다.

(c) $\mathrm{Cov}\,\mathbf{e} = \sigma^2(I-H)$이므로 $\mathrm{Var}\,e_i = \sigma^2(1-h_{ii})$이다. 지렛대가 큰 점은 잔차가 작게 나온다.

증명. **전략** $H = HH'$의 대각원소를 비교한다. (b)는 $H - J$ ($J = \mathbf{1}\mathbf{1}'/n$)도 대칭 멱등임을 보여 (a)의 논리를 한 번 더 쓴다. 단계별 증명은 연습 2.9(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 52쪽에 있다.

지렛대의 평균은 p/n이다. Mayer의 자료($p/n = 3/27 = 0.111$)에서 가장 큰 지렛대는 26번 방정식의 0.247이다. 7절에서 지렛대가 1인 점이 무엇을 하는지 본다.

2.3 계산: QR 분해

정규방정식을 곧이곧대로 풀면 수치적으로 위험할 수 있다. 표준 방법은 QR 분해다.

알고리즘 2.8 (QR로 푸는 최소제곱)

1. $X = QR$로 분해한다. Q는 $n \times p$이고 $Q'Q = I_p$, R은 대각원소가 양수인 $p \times p$ 상삼각행렬이다. (X의 열에 Gram–Schmidt를 차례로 쓰면 얻는다. 연습 2.6)
2. $\mathbf{z} = Q'\mathbf{y}$를 계산한다.
3. $R\hat{\boldsymbol{\beta}} = \mathbf{z}$를 아래 행부터 거꾸로 대입해 푼다.
4. $\hat{\mathbf{y}} = Q\mathbf{z}$, $\mathrm{RSS} = \|\mathbf{y}\|^2 - \|\mathbf{z}\|^2$, $(X'X)^{-1} = R^{-1}R^{-\mathsf{T}}$.

왜 맞나. $X'X = R'Q'QR = R'R$이므로 정규방정식은 $R'R\hat{\boldsymbol{\beta}} = R'Q'\mathbf{y}$이고, R'가 가역이므로 $R\hat{\boldsymbol{\beta}} = Q'\mathbf{y}$이다. $H = QR(R'R)^{-1}R'Q' = QQ'$이다.

왜 안전한가. 행렬 M의 조건수 $\kappa(M)$(가장 큰 특이값 / 가장 작은 특이값)는 선형방정식을 풀 때 잃는 유효숫자의 자릿수를 대략 $\log_{10}\kappa$로 알려 준다. $X'X$의 특이값은 X의 특이값의 제곱이므로 $\kappa(X'X) = \kappa(X)^2$이다. 정규방정식은 QR보다 두 배의 자릿수를 잃는다. Galton의 자료(5절)에서 $X = [\mathbf{1},\ \text{부모 키}]$의 $\kappa(X'X)$는 6.8×10^6이다. 부모 키에서 평균을 빼면 3.19가 된다. 키가 평균 68인치 근처에 몰려 있어서 상수 열과 거의 평행하기 때문이다. 8절에서 이 숫자가 경사하강의 속도를 정하는 것을 본다.

예제 2.9 (Mayer의 최소제곱 해) Mayer의 27개 방정식에 최소제곱을 쓰면

$$\hat{\beta} = 14.558^\circ = 14^\circ 33.5',\qquad \hat{\alpha} = 1.5058^\circ = 1^\circ 30.3',\qquad \hat{\gamma} = -0.0719,\qquad \hat{\theta} = -2^\circ 44.3'$$

이고 $\mathrm{RSS} = 0.5689$이다. 묶음 평균(예제 1.5)과 비교하면 β와 α는 1분 안쪽으로 같고, θ는 1도 넘게 다르다. 6절에서 θ의 표준오차가 약 2°임을 보면 이 차이가 놀랍지 않다. 27개의 방정식은 θ를 잘 정하지 못한다.

연습문제

연습 2.1 (단순회귀 ★) $y_i = b_0 + b_1 x_i + \varepsilon_i$에서 정규방정식을 풀어 $\hat{b}_1 = S_{xy}/S_{xx}$, $\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$임을 보여라. 여기서 $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$, $S_{xx} = \sum (x_i - \bar{x})^2$이다. R^2이 표본상관계수의 제곱과 같음도 보여라. 풀이 51쪽

연습 2.2 (평균의 hat 행렬 ★) $X = \mathbf{1}$이면 $H = \mathbf{1}\mathbf{1}'/n$이고 $\hat{\mathbf{y}} = \bar{y}\mathbf{1}$임을 보여라. $\mathrm{RSS}/(n-1)$은 무엇인가? 풀이 51쪽

연습 2.3 (단순회귀의 지렛대 ★) 단순회귀에서 $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$임을 보여라. x가 열 번은 8이고 한 번은 19인 자료($n = 11$)에서 $x = 19$인 점의 지렛대를 구하라. 풀이 51쪽

연습 2.4 (단위를 바꾸면 ★) 키를 인치에서 센티미터로 바꾸면($x \mapsto 2.54x$) 단순회귀의 $\hat{b}_0, \hat{b}_1$, 적합값, R^2은 각각 어떻게 바뀌는가? 정리 2.5(d)로 설명하라. 풀이 51쪽

연습 2.5 (변수를 더하면 ★) X에 열 $\mathbf{z}$를 더한 $[X, \mathbf{z}]$로 최소제곱을 하면 RSS가 늘지 않음을 보여라. 따라서 R^2은 변수를 더할수록 줄지 않는다. 풀이 51쪽

연습 2.6 (Gram–Schmidt와 QR ★★) X의 열 $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p)}$이 일차독립이면, $\mathbf{q}_1 = \mathbf{x}^{(1)}/\|\mathbf{x}^{(1)}\|$, $\mathbf{q}_k \propto \mathbf{x}^{(k)} - \sum_{j<k} (\mathbf{q}_j' \mathbf{x}^{(k)}) \mathbf{q}_j$ (길이 1로 맞춤)로 만든 $Q = [\mathbf{q}_1, \dots, \mathbf{q}_p]$와 어떤 상삼각행렬 R에 대해 $X = QR$임을 보여라. R의 원소는 무엇인가? 풀이 51쪽

연습 2.7 (대칭 멱등행렬의 성질 ★★ 유도형) P가 대칭이고 $P^2 = P$이다. 보조정리 2.4를 다음 순서로 증명하라.

(a) 스펙트럼 분해 $P = V \Lambda V'$에서 $P^2 = P$가 $\lambda_i^2 = \lambda_i$와 같음을 보이고, $\operatorname{rank} P$가 1인 고유값의 수이며 $\operatorname{tr} P$와 같음을 보여라.

(b) 고유값 1에 해당하는 V의 열을 모은 U로 $P = UU'$, $U'U = I_r$, $\mathrm{col}(P) = \mathrm{col}(U)$를 보여라.

(c) $(I - P)^2 = I - P$와 $\operatorname{tr}(I - P) = n - r$을 보여라.

(d) $(P\mathbf{v})'(I - P)\mathbf{v} = 0$을 보이고, $\mathbf{w} \in \mathrm{col}(P)$이면 $\|\mathbf{v} - \mathbf{w}\|^2 = \|(I - P)\mathbf{v}\|^2 + \|P\mathbf{v} - \mathbf{w}\|^2$임을 써서 $P\mathbf{v}$가 $\mathrm{col}(P)$에서 $\mathbf{v}$에 가장 가까운 점임을 보여라.

풀이 52쪽

연습 2.8 (hat 행렬의 성질 ★★ 유도형) $\operatorname{rank} X = p$, $H = X(X'X)^{-1}X'$이다. 정리 2.5을 다음 순서로 증명하라.

(a) $H' = H$, $H^2 = H$를 곱을 전개해 확인하라. $HX = X$에서 $\mathrm{col}(X) \subseteq \mathrm{col}(H)$를, $H = X(X'X)^{-1}X'$의 꼴에서 $\mathrm{col}(H) \subseteq \mathrm{col}(X)$를 보여라.

(b) $\operatorname{tr} AB = \operatorname{tr} BA$로 $\operatorname{tr} H = p$를 구하고 보조정리 2.4(a)로 $\operatorname{rank} H = p$를 얻어라.

(c) $\hat{\mathbf{y}} = H\mathbf{y}$, $\mathbf{e} = (I - H)\mathbf{y}$를 보이고, 보조정리 2.4(c)(d)와 정규방정식으로 (b)(c)의 나머지

식을 얻어라.

(d) T가 가역이면 XT의 hat 행렬이 H와 같음을 $(T'X'XT)^{-1} = T^{-1}(X'X)^{-1}T^{-\mathsf{T}}$로 보여라.

풀이 52쪽

연습 2.9 (지렛대의 성질 ★★ 유도형) 명제 2.7를 다음 순서로 증명하라.

(a) $H = HH'$의 (i,i) 원소를 비교해 $h_{ii} = h_{ii}^2 + \sum_{j\neq i} h_{ij}^2$을 얻고 $0 \le h_{ii} \le 1$을 보여라. $\sum_i h_{ii} = p$는 무엇에서 나오는가?

(b) $J = \mathbf{1}\mathbf{1}'/n$, $\mathbf{1} \in \mathrm{col}(X)$이면 $HJ = JH = J$이고 $H - J$가 대칭 멱등임을 보여라. (a)의 논리를 $H - J$에 써서 $h_{ii} \ge 1/n$을 얻어라.

(c) $\mathbf{e} = (I-H)\boldsymbol{\varepsilon}$를 보이고 5회 보조정리 3.1로 $\mathrm{Cov}\,\mathbf{e} = \sigma^2(I-H)$를 얻어라.

풀이 52쪽

3

Gauss–Markov 정리

> **역사** Gauss의 첫 정당화(1809)는 순환논법이었다. 그는 평균이 가장 그럴듯한 값이 되는 오차분포가 정규임을 보였고(5회 정리 1.5), 정규 오차를 가정하면 최소제곱이 가장 그럴듯한 값을 준다고 결론지었다. 이듬해 Laplace의 중심극한정리가 오차가 작은 원인들의 합이면 정규라는 근거를 댔다. 그러나 Gauss는 만족하지 않았다. 1821–1823년의 **오차가 가장 작은 관측의 결합 이론**에서 그는 정규분포를 버렸다. 오차의 평균이 0이고 분산이 같기만 하면, 최소제곱이 모든 선형불편 결합 가운데 분산이 가장 작다는 것을 증명한 것이다. 1900년 Markov가 교과서에서 이 증명을 다시 썼고, Neyman(1934)이 그것을 "Markov의 정리"라 부르면서 오늘의 이름이 굳었다. Aitken(1935)은 이 정리를 행렬로 쓰고 상관된 오차로 넓혔다.

3.1 최소제곱추정량의 적률

보조정리 3.1 (이차형식의 기댓값) 확률벡터 $\mathbf{u}$의 평균이 $\boldsymbol{\mu}$, 공분산이 Σ이고 A가 상수 행렬이면 $E\,\mathbf{u}'A\mathbf{u} = \operatorname{tr}(A\Sigma) + \boldsymbol{\mu}'A\boldsymbol{\mu}$이다.

증명. $\mathbf{u}'A\mathbf{u}$는 1×1이므로 자기 대각합이고 $\operatorname{tr}(\mathbf{u}'A\mathbf{u}) = \operatorname{tr}(A\mathbf{u}\mathbf{u}')$이다. 대각합과 기댓값은 바꿀 수 있고 $E\mathbf{u}\mathbf{u}' = \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}'$이므로 $E\,\mathbf{u}'A\mathbf{u} = \operatorname{tr}(A\Sigma) + \operatorname{tr}(A\boldsymbol{\mu}\boldsymbol{\mu}') = \operatorname{tr}(A\Sigma) + \boldsymbol{\mu}'A\boldsymbol{\mu}$. ∎

정리 3.2 (최소제곱추정량의 적률) 정의 1.1의 모형에서

(a) $E\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, $\operatorname{Cov}\hat{\boldsymbol{\beta}} = \sigma^2(X'X)^{-1}$.

(b) $E\mathbf{e} = \mathbf{0}$, $\operatorname{Cov}\mathbf{e} = \sigma^2(I-H)$, 그리고 $\hat{\boldsymbol{\beta}}$과 $\mathbf{e}$는 무상관이다.

(c) $E\,\mathrm{RSS} = (n-p)\sigma^2$이다. 따라서 $s^2 = \mathrm{RSS}/(n-p)$는 σ^2의 불편추정량이다.

증명. **전략** $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}$와 $\mathbf{e}$를 오차 $\boldsymbol{\varepsilon}$의 선형변환으로 쓰고 선형변환의 평균·공분산 공식(5회 보조정리 3.1)을 쓴다. RSS는 $\boldsymbol{\varepsilon}$의 이차형식이므로 보조정리 3.1를 쓴다.

① $L = (X'X)^{-1}X'$로 두면 $\hat{\boldsymbol{\beta}} = L(X\boldsymbol{\beta}+\boldsymbol{\varepsilon}) = \boldsymbol{\beta} + L\boldsymbol{\varepsilon}$이다. — 정리 2.2(c), $LX = I$

② $E\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$이고 $\operatorname{Cov}\hat{\boldsymbol{\beta}} = L(\sigma^2 I)L' = \sigma^2(X'X)^{-1}$이다. — 5회 보조정리 3.1, $LL' = (X'X)^{-1}$

③ $\mathbf{e} = (I-H)\mathbf{y} = (I-H)\boldsymbol{\varepsilon}$이다. — $(I-H)X = 0$, 정리 2.5(a)

④ $E\mathbf{e} = \mathbf{0}$이고 $\operatorname{Cov}\mathbf{e} = \sigma^2(I-H)(I-H)' = \sigma^2(I-H)$이다. — 5회 보조정리 3.1, $I-H$ 대칭 멱등

⑤ $\operatorname{Cov}(\hat{\boldsymbol{\beta}}, \mathbf{e}) = L(\sigma^2 I)(I-H)' = \sigma^2 (X'X)^{-1}(X' - X'H) = 0$이다. $X'H = (HX)' = X'$

⑥ $\mathrm{RSS} = \mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'(I-H)\boldsymbol{\varepsilon}$이다. ③, $I-H$ 대칭 멱등

⑦ $E\,\mathrm{RSS} = \sigma^2 \operatorname{tr}(I-H) = \sigma^2(n-p)$이다. 보조정리 3.1, 정리 2.5(c)

■

$p=1$, $X=\mathbf{1}$이면 (c)는 6회 정리 1.1의 "$n-1$로 나누기"다. 일반적으로 $n-p$는 잔차가 사는 공간 $\operatorname{col}(X)^{\perp}$의 차원이다. 모수 p개를 맞추면서 잔차가 움직일 수 있는 방향 p개를 썼다.

3.2 최선의 선형불편추정량

정리 3.3 (Gauss–Markov) 정의 1.1의 모형에서 $\tilde{\boldsymbol{\beta}} = A\mathbf{y}$를 아무 선형불편추정량이라 하고 $A_0 = (X'X)^{-1}X'$, $D = A - A_0$이라 하자. 그러면

$$\operatorname{Cov}\tilde{\boldsymbol{\beta}} - \operatorname{Cov}\hat{\boldsymbol{\beta}} = \sigma^2 DD' \;\succeq\; 0.$$

따라서 모든 $\mathbf{c} \in \mathbb{R}^p$에 대해 $\operatorname{Var}(\mathbf{c}'\tilde{\boldsymbol{\beta}}) \geq \operatorname{Var}(\mathbf{c}'\hat{\boldsymbol{\beta}})$이다. 최소제곱추정량은 **최선의 선형불편추정량**(BLUE)이다.

증명. **전략** 아무 선형불편추정량의 계수행렬 A를 최소제곱의 A_0와 그 차 D로 나눈다. 불편성에서 $DX = 0$이 나오고, 그러면 공분산이 피타고라스처럼 두 조각의 합으로 갈라진다.

① $AX = I$이고 $A_0X = I$이므로 $DX = 0$이다. 명제 1.4

② $DA_0' = DX(X'X)^{-1} = 0$이고 $A_0D' = (DA_0')' = 0$이다. ①

③ 공분산을 전개한다.

$$\begin{aligned} \operatorname{Cov}\tilde{\boldsymbol{\beta}} &= \sigma^2 AA' = \sigma^2 (A_0 + D)(A_0 + D)' && \text{5회 보조정리 3.1} \\ &= \sigma^2\big(A_0A_0' + DD'\big) && \text{②로 교차항 0} \\ &= \operatorname{Cov}\hat{\boldsymbol{\beta}} + \sigma^2 DD' && A_0A_0' = (X'X)^{-1},\ \text{정리 3.2(a)} \end{aligned}$$

④ $\mathbf{c}'DD'\mathbf{c} = \|D'\mathbf{c}\|^2 \geq 0$이므로 $\sigma^2 DD' \succeq 0$이고 $\operatorname{Var}(\mathbf{c}'\tilde{\boldsymbol{\beta}}) - \operatorname{Var}(\mathbf{c}'\hat{\boldsymbol{\beta}}) = \sigma^2\|D'\mathbf{c}\|^2 \geq 0$이다.

준정치의 정의 [T4]

■

증명은 2절의 피타고라스를 추정량의 공간에서 한 번 더 쓴 것이다. 아무 선형불편추정량은 최소제곱추정량에 그것과 "직교하는" 잡음 $D\mathbf{y}$를 더한 것이다.

따름정리 3.4 (BLUE의 여러 얼굴) 선형불편추정량 $\tilde{\boldsymbol{\beta}}$에 대해

(a) 성분마다 $\operatorname{Var}\tilde{\beta}_j \ge \operatorname{Var}\hat{\beta}_j$이고, $\det\operatorname{Cov}\tilde{\boldsymbol{\beta}} \ge \det\operatorname{Cov}\hat{\boldsymbol{\beta}}$, $\operatorname{tr}\operatorname{Cov}\tilde{\boldsymbol{\beta}} \ge \operatorname{tr}\operatorname{Cov}\hat{\boldsymbol{\beta}}$이다.

(b) $\operatorname{Var}(\mathbf{c}'\tilde{\boldsymbol{\beta}}) = \operatorname{Var}(\mathbf{c}'\hat{\boldsymbol{\beta}})$이면 $\mathbf{c}'\tilde{\boldsymbol{\beta}} = \mathbf{c}'\hat{\boldsymbol{\beta}}$이다. 곧 $\mathbf{c}'\boldsymbol{\beta}$의 BLUE는 하나뿐이다.

증명. (a) 성분은 $\mathbf{c} = \mathbf{e}_j$로 둔 경우다. $B = \operatorname{Cov}\hat{\boldsymbol{\beta}} \succ 0$, $A = \operatorname{Cov}\tilde{\boldsymbol{\beta}} \succeq B$이면 $B^{-1/2}AB^{-1/2} \succeq I$이므로 그 고유값이 모두 1 이상이고, $\det A/\det B = \det(B^{-1/2}AB^{-1/2}) \ge 1$이다. 대각합은 성분의 합이다. (b) 차이가 $\sigma^2\|D'\mathbf{c}\|^2 = 0$이면 $\mathbf{c}'D = \mathbf{0}'$이고 $\mathbf{c}'\tilde{\boldsymbol{\beta}} = \mathbf{c}'(A_0 + D)\mathbf{y} = \mathbf{c}'\hat{\boldsymbol{\beta}}$이다. ∎

예제 3.5 (Mayer는 얼마나 잃었나) Mayer의 묶음 추정량은 $A = (G'X)^{-1}G'$ 인 선형불편추정량이다(예제 1.5). 공분산 $\sigma^2 AA'$와 $\sigma^2(X'X)^{-1}$의 대각원소의 비는 (β, α, γ) 순으로

$$1.28, \qquad 1.15, \qquad 2.08$$

이고 행렬식의 비는 2.37이다. 비는 σ^2과 무관하므로 자료 없이 설계만으로 정해진다. β를 최소제곱만큼 정확히 정하려면 Mayer의 방법으로는 관측이 28% 더 필요했고, γ(곧 θ)에는 두 배가 넘게 필요했다. 그림 3.1의 오른쪽이 이것을 신뢰타원으로 보여 준다. 최소제곱의 타원은 Mayer의 타원 안에 들어간다. $\operatorname{Cov}\tilde{\boldsymbol{\beta}} \ge \operatorname{Cov}\hat{\boldsymbol{\beta}}$의 그림이 그것이다.

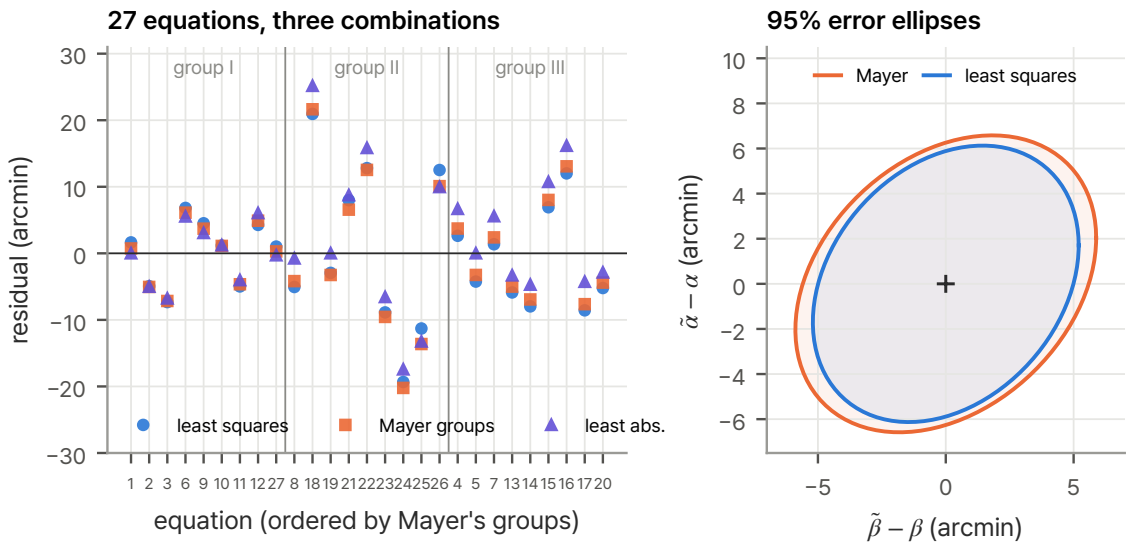


그림 3.1 Mayer의 27개 방정식. 왼쪽: 세 방법의 잔차. 최소제곱(파랑)은 큰 잔차를 줄이고, 최소절댓값(보라)은 세 방정식에서 잔차가 정확히 0이다. 묶음 평균(주황)은 최소제곱과 비슷하다. 오른쪽: 추정오차 $(\tilde{\beta}-\beta,\ \tilde{\alpha}-\alpha)$의 95% 타원(각의 분, σ는 최소제곱의 s로 둠). 최소제곱의 타원이 Mayer의 타원 안에 있다. 정리 3.3의 $\operatorname{Cov}\tilde{\boldsymbol{\beta}} \ge \operatorname{Cov}\hat{\boldsymbol{\beta}}$를 그린 것이다.

가정 주의 Gauss–Markov 정리의 "최선"은 **선형**이고 **불편**인 추정량 가운데서다. (1) 비선형 추정량은 이길 수 있다. 오차가 Laplace 분포 $\frac{1}{2}e^{-|u|}$이면 위치모형에서 중앙값의 점근분산은 $1/(4f(0)^2 n) = 1/n$으로(8회 정리 1.4), 평균의 분산 $2/n$의 절반이다. 최소절댓값 추정은 선형이 아니다. (2) 치우친 추정량은 평균제곱오차에서 이길 수 있다. 8절의 ridge와 16회의 Stein 추정량이 그렇다. (3) 등분산과 무상관이 깨지면 최소제곱은 더 이상 최선이 아니다(따름정리 3.7).

3.3 정규 오차라면: 모든 불편추정량 가운데 최선

정리 3.6 (정규 선형모형) $\mathbf{y} \sim \mathcal{N}_n(X\boldsymbol{\beta}, \sigma^2 I)$, $\operatorname{rank} X = p$라 하자.

(a) $\boldsymbol{\beta}$의 MLE는 $\hat{\boldsymbol{\beta}}$이고 σ^2의 MLE는 RSS/n이다.

(b) $(X'\mathbf{y},\ \mathbf{y}'\mathbf{y})$는 $(\boldsymbol{\beta}, \sigma^2)$의 완비충분통계량이다.

(c) $\hat{\boldsymbol{\beta}}$과 s^2은 **모든** 불편추정량 가운데 분산이 가장 작다(UMVUE). $\mathbf{c}'\hat{\boldsymbol{\beta}}$도 $\mathbf{c}'\boldsymbol{\beta}$의 UMVUE이다.

증명. (a) 로그가능도는 $\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\mathbf{y} - X\boldsymbol{\beta}\|^2$이다. σ^2이 무엇이든 $\boldsymbol{\beta}$에 대한 최대는 $\|\mathbf{y} - X\boldsymbol{\beta}\|^2$의 최소, 곧 $\hat{\boldsymbol{\beta}}$이다. 그 다음 $-\frac{n}{2}\log\sigma^2 - \mathrm{RSS}/(2\sigma^2)$을 σ^2에 대해 최대로 하면 RSS/n이다.

(b) 밀도를 펼치면

$$f(\mathbf{y}) = \exp\Big\{\Big(\frac{\boldsymbol{\beta}}{\sigma^2}\Big)' X'\mathbf{y} - \frac{1}{2\sigma^2}\,\mathbf{y}'\mathbf{y} - \frac{\|X\boldsymbol{\beta}\|^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2)\Big\}$$

이다. 자연모수 $(\boldsymbol{\beta}/\sigma^2,\ -1/(2\sigma^2))$는 $\mathbb{R}^p \times (-\infty, 0)$, 곧 $\mathbb{R}^{p+1}$의 열린 집합 전체를 움직이고 자연통계량은 $(X'\mathbf{y}, \mathbf{y}'\mathbf{y})$이다. 11회 정리 4.5로 완비충분이다.

(c) $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$와 $\mathrm{RSS} = \mathbf{y}'\mathbf{y} - (X'\mathbf{y})'(X'X)^{-1}(X'\mathbf{y})$는 이 통계량의 함수이고, 정리 3.2로 $\hat{\boldsymbol{\beta}}$과 s^2은 불편이다. Lehmann–Scheffé(11회 정리 3.3)로 UMVUE다. ∎

정규 오차에서 최소제곱은 MLE이고, 선형이든 아니든 불편추정량 가운데 최선이다. Gauss의 1809년 논리(정규 ⇒ 최소제곱)와 1823년 논리(정규가 아니어도 선형 가운데 최선)가 이렇게 만난다.

3.4 상관된 오차: 일반화 최소제곱

따름정리 3.7 (Aitken의 일반화 최소제곱 (1935)) $\operatorname{Cov}\boldsymbol{\varepsilon} = \sigma^2 V$이고 V가 알려진 양정치 행렬이면

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = (X'V^{-1}X)^{-1}X'V^{-1}\mathbf{y}, \qquad \operatorname{Cov}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \sigma^2(X'V^{-1}X)^{-1}$$

가 BLUE이다. 보통의 최소제곱은 여전히 불편이지만 공분산은 $\sigma^2(X'X)^{-1}X'VX(X'X)^{-1}$이고, 이는 $\operatorname{Cov}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ 이상이다.

증명. **전략** $V^{-1/2}$를 곱해 오차의 공분산을 $\sigma^2 I$로 바꾼 모형에 정리 3.3를 쓴다. 두 모형의 선형불편추정량이 같은 모임임을 확인하면 비교 대상이 같다. 단계별 증명은 연습 3.6(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 54쪽에 있다.

가중 최소제곱은 V가 대각인 경우다. y_i가 m_i개 관측의 평균이면 $\operatorname{Var} y_i = \sigma^2/m_i$이므로 무게 m_i로 제곱합 $\sum m_i(y_i - \mathbf{x}_i'\mathbf{b})^2$을 최소로 한다. Pearson–Lee의 도수표(5절)는 이렇게 맞춘다. 무게가 틀리면(이분산을 모르고 무시하면) 공분산의 공식이 틀린다. 7절의 주제다.

연습문제

연습 3.1 (단순회귀의 분산 ★) 단순회귀에서 $\operatorname{Var}\hat{b}_1 = \sigma^2/S_{xx}$, $\operatorname{Var}\hat{b}_0 = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$, $\operatorname{Cov}(\hat{b}_0, \hat{b}_1) = -\sigma^2\bar{x}/S_{xx}$임을 정리 3.2로 보여라. $\bar{x}$가 0에서 멀면 $\hat{b}_0$과 $\hat{b}_1$은 어떻게 상관되는가? 풀이 53쪽

연습 3.2 (평균은 BLUE ★) $y_i = \mu + \varepsilon_i$에서 선형불편추정량 $\sum w_i y_i$의 조건이 $\sum w_i = 1$임을 보이고, 그 가운데 분산이 가장 작은 것이 $\bar{y}$임을 직접(라그랑주 승수 없이) 보여라. 풀이 53쪽

연습 3.3 (묶음의 평균 ★) 한 양 μ를 k개 실험실이 쟀다. 실험실 j는 m_j번 재서 평균 $\bar{y}_j$를 보고한다(각 측정의 분산 σ^2, 모두 독립). 따름정리 3.7로 μ의 BLUE가 $\sum m_j\bar{y}_j / \sum m_j$임을 보이고, 이것이 모든 원래 측정의 평균임을 확인하라. 풀이 53쪽

연습 3.4 (Mayer의 분산 ★★) 예제 3.5의 비 1.28, 1.15, 2.08과 2.37을 계산하라(노트북). 왜 γ의 손실이 가장 큰가? 묶음 III의 γ 계수(7.9649)와 나머지 두 묶음의 γ 계수를 보고 설명하라. 풀이 53쪽

연습 3.5 (치우친 추정량이 이길 때 ★★) $y_i = \mu + \varepsilon_i$, $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$에서 $c\bar{y}$ $(0 < c < 1)$의 평균제곱오차가 $c^2\sigma^2/n + (1-c)^2\mu^2$임을 보여라. 이를 최소로 하는 c를 구하고, 그때의 평균제곱오차가 $\bar{y}$의 것보다 작음을 보여라. 이 c를 실제로 쓸 수 없는 이유는 무엇인가? (16회 Stein의 예고다.) 풀이 53쪽

연습 3.6 (일반화 최소제곱 ★★ 유도형) 따름정리 3.7를 다음 순서로 증명하라.

(a) 양정치 V의 제곱근 $V^{1/2}$(5회 보조정리 3.2)로 $\mathbf{y}^* = V^{-1/2}\mathbf{y}$, $X^* = V^{-1/2}X$를 두면 $\mathbf{y}^* = X^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$, $\operatorname{Cov}\boldsymbol{\varepsilon}^* = \sigma^2 I$임을 보여라. $\operatorname{rank} X^* = p$인 이유도 적어라.

(b) 원래 모형의 선형불편추정량 $A\mathbf{y}$ $(AX = I)$와 변환된 모형의 선형불편추정량 $A^*\mathbf{y}^*$ $(A^*X^* = I)$가 $A^* = AV^{1/2}$로 하나씩 대응함을 보여라.

(c) 변환된 모형에 정리 3.3를 쓰고 BLUE $(X^{*\mathsf{T}}X^*)^{-1}X^{*\mathsf{T}}\mathbf{y}^*$를 원래의 $X, V, \mathbf{y}$로 다시 써라.

(d) 보통 최소제곱의 공분산 $\sigma^2(X'X)^{-1}X'VX(X'X)^{-1}$을 5회 보조정리 3.1로 구하고, 그것이 $\operatorname{Cov}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ 이상인 이유를 (b)(c)로 설명하라.

풀이 54쪽

4

정규 오차 아래의 분포 — Cochran과 Craig

> **역사** Fisher는 1922년 회귀계수의 t 분포를, 1925년 분산분석표를 내놓았다. 그의 논증은 n차원 공간의 기하적 직관에 기대었다. 1934년 케임브리지의 대학원생이던 William Cochran은 정규변수의 제곱합을 이차형식 여럿으로 나눌 때 각 조각이 언제 서로 독립인 χ^2이 되는지를 정리로 증명했다. 1943년 아이오와의 Allen Craig는 두 이차형식이 독립일 조건을 한 줄로 썼다. 두 행렬의 곱이 0이면 된다는 것이다. Craig의 "필요하다" 쪽 증명에는 틈이 있었다. 일본의 Sakamoto(1944)가 따로 같은 정리에 이르렀고, 올바른 증명을 둘러싼 논문이 1980년대까지 이어졌다. Craig는 뒤에 제자 Robert Hogg와 함께 수리통계학의 표준 교과서를 썼다.

이 절 내내 $\mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, I_n)$, 곧 $Z_1, \dots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$이다(5회 따름정리 3.12).

4.1 이차형식이 χ^2일 때

정리 4.1 (이차형식의 χ^2 판정) A가 대칭이면 $\mathbf{Z}'A\mathbf{Z} \sim \chi^2(r) \iff A^2 = A$이고 $\operatorname{rank} A = r$.

증명. **전략** ($\Leftarrow$) 사영을 정규직교 좌표로 쓰면 독립인 표준정규 r개의 제곱합이 된다. ($\Rightarrow$) 고유값으로 쓴 mgf를 $\chi^2(r)$의 mgf와 맞추고, 두 다항식의 근을 비교한다.

$\Leftarrow$

① $A = UU'$, $U'U = I_r$로 쓴다. — 보조정리 2.4(b)

② $\mathbf{W} = U'\mathbf{Z} \sim \mathcal{N}_r(\mathbf{0}, U'U) = \mathcal{N}_r(\mathbf{0}, I_r)$이다. — 5회 정리 3.8

③ $\mathbf{Z}'A\mathbf{Z} = \mathbf{Z}'UU'\mathbf{Z} = \|\mathbf{W}\|^2$은 독립인 표준정규 r개의 제곱합이므로 $\chi^2(r)$이다. — 5회 따름정리 1.9

$\Rightarrow$

① $A = V\Lambda V'$로 쓰고 $\mathbf{W} = V'\mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, I)$로 두면 $\mathbf{Z}'A\mathbf{Z} = \sum_i \lambda_i W_i^2$이다. — 5회 보조정리 3.2, 정리 3.8

② $|t| < 1/(2\max_i |\lambda_i|)$이면

$$E\, e^{t\mathbf{Z}'A\mathbf{Z}} = \prod_{i=1}^{n} E\, e^{t\lambda_i W_i^2} = \prod_{i=1}^{n} (1 - 2t\lambda_i)^{-1/2}. \qquad W_i \text{ 독립, } \chi^2(1)\text{의 mgf}$$

③ 가정으로 이것은 $\chi^2(r)$의 mgf $(1-2t)^{-r/2}$과 같다. 제곱해 뒤집으면 $\prod_i(1-2t\lambda_i)=(1-2t)^r$이 0 근처의 구간에서 성립한다. ②

④ 두 다항식이 한 구간에서 같으므로 모든 t에서 같다. 다항식의 항등

⑤ 오른쪽은 $t=\frac{1}{2}$에서만 0이고 중복도가 r이다. 왼쪽은 $\lambda_i\neq 0$마다 $t=1/(2\lambda_i)$에서 0이다. 따라서 0이 아닌 λ_i는 모두 1이고 정확히 r개다. 근의 비교

⑥ $\Lambda^2=\Lambda$이므로 $A^2=V\Lambda^2V'=A$이고 $\operatorname{rank}A=r$이다. ⑤

조건 확인 $\chi^2(1)$의 mgf $(1-2s)^{-1/2}$은 $s<\frac{1}{2}$에서만 유한하다. $|t|<1/(2\max_i|\lambda_i|)$이면 모든 i에서 $t\lambda_i<\frac{1}{2}$이므로 ②의 각 인수가 유한하고, 독립인 변수의 함수의 곱의 기댓값은 기댓값의 곱이다(3회 정리 3.4(b)를 n개로 되풀이). ■

> **따름정리 4.2** (평균이 있을 때) $\mathbf{y}\sim\mathcal{N}_n(\boldsymbol{\mu},\sigma^2I)$이고 A가 계수 r의 대칭 멱등행렬이며 $A\boldsymbol{\mu}=\mathbf{0}$이면 $\mathbf{y}'A\mathbf{y}/\sigma^2\sim\chi^2(r)$이다.

증명. $\mathbf{y}=\boldsymbol{\mu}+\sigma\mathbf{Z}$이고 $A\boldsymbol{\mu}=\mathbf{0}$이면 $\boldsymbol{\mu}'A=\mathbf{0}'$이므로 $\mathbf{y}'A\mathbf{y}=\sigma^2\mathbf{Z}'A\mathbf{Z}$이다. 정리 4.1. ■

4.2 두 이차형식이 독립일 때

> **보조정리 4.3** (이차형식의 공분산) A,B가 대칭이면 $E\,\mathbf{Z}'A\mathbf{Z}=\operatorname{tr}A$이고 $\operatorname{Cov}(\mathbf{Z}'A\mathbf{Z},\ \mathbf{Z}'B\mathbf{Z})=2\operatorname{tr}(AB)$이다. 특히 $\operatorname{Var}\mathbf{Z}'A\mathbf{Z}=2\operatorname{tr}(A^2)$이다.

증명. 첫 식은 보조정리 3.1다. 표준정규의 독립성과 $EZ^2=1$, $EZ^4=3$에서, 네 첨자에 대해

$$EZ_iZ_jZ_kZ_l=\delta_{ij}\delta_{kl}+\delta_{ik}\delta_{jl}+\delta_{il}\delta_{jk}$$

이다(네 첨자가 모두 같으면 3, 두 쌍으로 나뉘면 1, 그 밖에는 홀수 차수가 남아 0). 따라서

$$E(\mathbf{Z}'A\mathbf{Z})(\mathbf{Z}'B\mathbf{Z})=\sum_{i,j,k,l}a_{ij}b_{kl}EZ_iZ_jZ_kZ_l=\operatorname{tr}A\ \operatorname{tr}B+\sum_{i,j}a_{ij}b_{ij}+\sum_{i,j}a_{ij}b_{ji}=\operatorname{tr}A\ \operatorname{tr}B+2\operatorname{tr}(AB)$$

이고 (B가 대칭이므로 두 합이 모두 $\operatorname{tr}(AB)$), $\operatorname{tr}A\ \operatorname{tr}B$를 빼면 공분산이다. ■

정리 4.4 (Craig의 정리 (1943)) A, B는 $n \times n$ 대칭행렬, L은 $q \times n$ 행렬이다.

(a) $AB = 0$이면 $\mathbf{Z}'A\mathbf{Z}$와 $\mathbf{Z}'B\mathbf{Z}$는 독립이다.

(b) $LA = 0$이면 $L\mathbf{Z}$와 $\mathbf{Z}'A\mathbf{Z}$는 독립이다.

(c) A, B가 양의 준정치이고 $\mathbf{Z}'A\mathbf{Z}$와 $\mathbf{Z}'B\mathbf{Z}$가 독립이면 $AB = 0$이다.

증명. **전략** 이차형식을 0이 아닌 고유값의 좌표 $U_A'\mathbf{Z}$의 함수로 쓴다. 결합정규인 벡터는 무상관이면 독립이므로, 독립을 공분산 행렬의 계산으로 바꾼다. (c)는 거꾸로 공분산 0에서 대각합으로 간다.

준비

① 0이 아닌 고유값만 모아 $A = U_A \Lambda_A U_A'$로 쓴다. U_A의 열은 정규직교이고 Λ_A는 가역인 대각행렬이다. B도 같은 방식으로 U_B, Λ_B를 잡는다. 5회 보조정리 3.2

② $\mathbf{Z}'A\mathbf{Z} = (U_A'\mathbf{Z})'\Lambda_A(U_A'\mathbf{Z})$는 $U_A'\mathbf{Z}$의 함수이고, $AU_A = U_A\Lambda_A$에서 $U_A = AU_A\Lambda_A^{-1}$이다. ①

(b)

① $\operatorname{Cov}(L\mathbf{Z},\ U_A'\mathbf{Z}) = LU_A = LAU_A\Lambda_A^{-1} = 0$이다. 5회 보조정리 3.1, 준비의 ②, $LA = 0$

② $(L\mathbf{Z}, U_A'\mathbf{Z})$는 $\mathbf{Z}$의 선형변환이므로 결합정규이고, 무상관이므로 독립이다. 5회 정리 3.8, 3.11

③ 독립인 두 벡터의 함수 $L\mathbf{Z}$와 $\mathbf{Z}'A\mathbf{Z}$는 독립이다. 준비의 ②, 3회 정리 3.4(a)

(a)

① $\operatorname{Cov}(U_A'\mathbf{Z},\ U_B'\mathbf{Z}) = U_A'U_B = \Lambda_A^{-1}U_A'AB\,U_B\Lambda_B^{-1} = 0$이다. 준비의 ②를 A, B에, $AB = 0$

② (b)와 같은 논리로 $U_A'\mathbf{Z}$와 $U_B'\mathbf{Z}$가 독립이고, 그 함수인 두 이차형식도 독립이다. (b)의 ② ③

(c)

① 독립이면 공분산이 0이므로 $\operatorname{tr}(AB) = 0$이다. 보조정리 4.3

② 양의 준정치 행렬의 제곱근을 쓰면

$$\operatorname{tr}(AB) = \operatorname{tr}\big(B^{1/2}A^{1/2}A^{1/2}B^{1/2}\big) = \big\|A^{1/2}B^{1/2}\big\|_F^2 \qquad \operatorname{tr}MN = \operatorname{tr}NM$$

이다. $\|M\|_F^2 = \operatorname{tr}M'M$은 원소의 제곱합이다.

③ 따라서 $A^{1/2}B^{1/2}=0$이고 $AB=A^{1/2}(A^{1/2}B^{1/2})B^{1/2}=0$이다. ①②

조건 확인 (c)의 ①은 이차형식의 2차 적률이 유한하다는 것을 쓴다. $\mathbf{Z}$의 성분은 모든 차수의 적률을 가지므로 그렇다. 제곱근 $A^{1/2}$, $B^{1/2}$이 있으려면 A, B가 양의 준정치여야 한다. 이 가정을 빼는 일반판은 아래 상자에 있다. ■

> **증명 없이 쓰는 사실** (Craig–Sakamoto 정리) (c)는 양의 준정치라는 조건 없이도 참이다. 대칭행렬 A, B에 대해 $\mathbf{Z}'A\mathbf{Z}$와 $\mathbf{Z}'B\mathbf{Z}$가 독립이면 $AB=0$이다. 증명은 결합 mgf가 곱으로 갈라진다는 조건 $\det(I-2sA-2tB)=\det(I-2sA)\det(I-2tB)$에서 $AB=0$을 이끄는 행렬식의 대수인데, 이 과정의 범위를 넘는다(Driscoll–Krasnicka 1995, Olkin 1997). 이 노트에서 쓰는 이차형식은 모두 사영행렬이므로 (c)로 충분하다.

4.3 Cochran의 정리

정리 4.5 (Cochran의 정리 (1934)) $I_n=A_1+\cdots+A_k$이고 각 A_i는 대칭, $\operatorname{rank}A_i=r_i$라 하자. 그러면 $\mathbf{Z}'\mathbf{Z}=\sum_i \mathbf{Z}'A_i\mathbf{Z}$이고,

$$\mathbf{Z}'A_1\mathbf{Z},\dots,\mathbf{Z}'A_k\mathbf{Z} \text{ 가 서로 독립이고 } \mathbf{Z}'A_i\mathbf{Z}\sim\chi^2(r_i) \quad\Longleftrightarrow\quad r_1+\cdots+r_k=n.$$

증명. **전략** (⇒)는 평균의 비교다. (⇐)는 세 단계다. 계수의 합 조건으로 각 A_i가 멱등임을 보이고, 서로 직교함을 보인 뒤, 모든 사영의 기저를 모은 직교행렬로 $\mathbf{Z}$를 회전한다.

⇒

① 독립인 $\chi^2(r_i)$들의 합은 $\chi^2(\sum_i r_i)$이다. mgf의 곱, 3회 정리 3.4(d)

② 그 합이 $\mathbf{Z}'\mathbf{Z}\sim\chi^2(n)$이므로 평균을 비교하면 $\sum_i r_i=n$이다. $E\chi^2(r)=r$

⇐

① i를 고정하면 $I-A_i=\sum_{j\neq i}A_j$이므로 $\operatorname{rank}(I-A_i)\le\sum_{j\neq i}r_j=n-r_i$이다.
$\operatorname{rank}(M+N)\le\operatorname{rank}M+\operatorname{rank}N$

② 모든 $\mathbf{v}$가 $A_i\mathbf{v}+(I-A_i)\mathbf{v}$이므로 $\operatorname{col}(A_i)+\operatorname{col}(I-A_i)=\mathbb{R}^n$이고 $\operatorname{rank}(I-A_i)\ge n-r_i$이다. [T1]

③ 따라서 $I-A_i$의 영공간, 곧 A_i의 고유값 1의 고유공간의 차원은 r_i이다. ①②, 차원 정리

④ 대칭행렬 A_i의 0이 아닌 고유값은 중복을 세어 r_i개이고 그 가운데 1이 r_i개이므로, 0이 아닌 고유값은 모두 1이다. 곧 $A_i^2=A_i$이다. ③, 스펙트럼 분해

⑤ $\mathbf{v} \in \mathrm{col}(A_i)$이면 $A_i\mathbf{v} = \mathbf{v}$이고, 대칭 멱등인 A_j에서 $\mathbf{v}'A_j\mathbf{v} = \|A_j\mathbf{v}\|^2$이므로

$$\|\mathbf{v}\|^2 = \sum_j \mathbf{v}'A_j\mathbf{v} = \|\mathbf{v}\|^2 + \sum_{j\neq i}\|A_j\mathbf{v}\|^2. \qquad \textstyle\sum_j A_j = I,\ ④$$

따라서 $j \neq i$이면 $A_j\mathbf{v} = \mathbf{0}$, 곧 $A_jA_i = 0$이다.

⑥ $A_i = U_iU_i'$ $(U_i'U_i = I_{r_i})$로 쓰면 $U_j'U_i = U_j'A_jA_iU_i = 0$ $(j \neq i)$이다. $U = [U_1, \dots, U_k]$는 $n \times n$이고 $U'U = I$, 곧 직교행렬이다. 보조정리 2.4(b), ⑤, $\sum r_i = n$

⑦ $\mathbf{W} = U'\mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, I)$의 성분은 독립이고, $\mathbf{Z}'A_i\mathbf{Z} = \|U_i'\mathbf{Z}\|^2$은 $\mathbf{W}$의 i번째 블록의 제곱합이다. 블록들은 겹치지 않으므로 서로 독립인 $\chi^2(r_i)$이다. 5회 정리 3.8, 따름정리 1.9

■

예제 4.6 (Student의 정리는 Cochran의 정리다) $J = \mathbf{1}\mathbf{1}'/n$이라 하면 $I = J + (I-J)$이고 계수는 $1 + (n-1) = n$이다. Cochran으로 $\mathbf{Z}'J\mathbf{Z} = n\bar{Z}^2 \sim \chi^2(1)$과 $\mathbf{Z}'(I-J)\mathbf{Z} = \sum(Z_i - \bar{Z})^2 \sim \chi^2(n-1)$이 독립이다. $\bar{Z} = \frac{1}{n}\mathbf{1}'\mathbf{Z}$ 자체가 $\sum(Z_i - \bar{Z})^2$과 독립인 것은 Craig (b)에서 $L = \frac{1}{n}\mathbf{1}'$, $L(I-J) = \mathbf{0}'$로 나온다. 6회 정리 2.2가 이렇게 한 줄이 된다. 14회의 분산분석은 I를 더 많은 사영으로 나누는 같은 정리다.

4.4 계수의 분포와 t

정리 4.7 (정규 선형모형의 표본분포) $\mathbf{y} \sim \mathcal{N}_n(X\boldsymbol{\beta}, \sigma^2 I)$, $\mathrm{rank}\, X = p$이면

(a) $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p\big(\boldsymbol{\beta},\ \sigma^2(X'X)^{-1}\big)$,

(b) $\mathrm{RSS}/\sigma^2 \sim \chi^2(n-p)$,

(c) $\hat{\boldsymbol{\beta}}$과 RSS는 독립,

(d) $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'X'X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2(p)$.

증명. $\mathbf{Z} = \boldsymbol{\varepsilon}/\sigma \sim \mathcal{N}_n(\mathbf{0}, I)$로 두자. (a) $\hat{\boldsymbol{\beta}}$은 $\mathbf{y}$의 선형변환이므로 정규이고(5회 정리 3.8) 평균과 공분산은 정리 3.2(a). (b) $\mathrm{RSS} = \boldsymbol{\varepsilon}'(I-H)\boldsymbol{\varepsilon} = \sigma^2\mathbf{Z}'(I-H)\mathbf{Z}$이고 $I-H$는 계수 $n-p$의 대칭 멱등(정리 2.5(c))이므로 정리 4.1. (c) $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \sigma L\mathbf{Z}$, $L = (X'X)^{-1}X'$이고 $L(I-H) = 0$이므로 Craig (b). (d) 5회 정리 3.17을 (a)에 쓴다. ■

따름정리 4.8 (계수의 t, 신뢰구간, 예측구간) $s^2 = \mathrm{RSS}/(n-p)$이고 $\mathbf{c} \in \mathbb{R}^p$, $\mathbf{c} \neq \mathbf{0}$이면

$$T = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}} - \mathbf{c}'\boldsymbol{\beta}}{s\sqrt{\mathbf{c}'(X'X)^{-1}\mathbf{c}}} \sim t(n-p).$$

특히 $\mathrm{SE}(\hat{\beta}_j) = s\sqrt{[(X'X)^{-1}]_{jj}}$이고 $\hat{\beta}_j \pm t_{n-p,\alpha/2}\,\mathrm{SE}(\hat{\beta}_j)$는 β_j의 $1-\alpha$ 신뢰구간이다. 새 관측

$y_0 = \mathbf{x}_0'\boldsymbol{\beta} + \varepsilon_0$ ($\varepsilon_0 \sim \mathcal{N}(0,\sigma^2)$, $\mathbf{y}$와 독립)에 대해

$$\mathbf{x}_0'\hat{\boldsymbol{\beta}} \;\pm\; t_{n-p,\alpha/2}\, s\sqrt{1+\mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0}$$

는 $1-\alpha$ 예측구간이다.

증명. 정리 4.7(a)로 $(\mathbf{c}'\hat{\boldsymbol{\beta}}-\mathbf{c}'\boldsymbol{\beta})/(\sigma\sqrt{\mathbf{c}'(X'X)^{-1}\mathbf{c}}) \sim \mathcal{N}(0,1)$이고, (b)(c)로 $(n-p)s^2/\sigma^2 \sim \chi^2(n-p)$가 그것과 독립이다. 6회 정의 3.1로 $t(n-p)$다. 예측구간: $y_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}} = \varepsilon_0 - \mathbf{x}_0'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \sim \mathcal{N}\big(0,\sigma^2(1+\mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0)\big)$이고 s^2과 독립이다(ε_0은 $\mathbf{y}$와 독립이고 $\hat{\boldsymbol{\beta}}$은 RSS와 독립). ∎

$X=\mathbf{1}$이면 이것이 6회 정리 5.2의 t-구간과 6회 정리 8.1의 예측구간이다. 신뢰구간의 폭은 $n\to\infty$에서 0으로 가지만, 예측구간의 폭은 $2t\,\sigma$ 근처에서 멈춘다. 평균은 정확히 알게 되어도 다음 한 사람의 오차 ε_0은 줄지 않는다.

4.5 일반선형가설의 F 검정

정리 4.9 (일반선형가설) C는 계수 q의 $q\times p$ 행렬, $\mathbf{d}\in\mathbb{R}^q$이고 $H_0:\ C\boldsymbol{\beta}=\mathbf{d}$라 하자. $W = C(X'X)^{-1}C'$로 쓰고, $C\mathbf{b}=\mathbf{d}$ 아래 $\|\mathbf{y}-X\mathbf{b}\|^2$을 최소로 하는 $\hat{\boldsymbol{\beta}}_0$와 그때의 잔차제곱합 RSS_0을 생각한다.

(a) $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}} - (X'X)^{-1}C'W^{-1}(C\hat{\boldsymbol{\beta}}-\mathbf{d})$.

(b) $\mathrm{RSS}_0 - \mathrm{RSS} = \|X\hat{\boldsymbol{\beta}} - X\hat{\boldsymbol{\beta}}_0\|^2 = (C\hat{\boldsymbol{\beta}}-\mathbf{d})'W^{-1}(C\hat{\boldsymbol{\beta}}-\mathbf{d})$.

(c) 정규 오차이고 H_0이 참이면

$$F = \frac{(\mathrm{RSS}_0-\mathrm{RSS})/q}{\mathrm{RSS}/(n-p)} \sim F(q,\ n-p).$$

(d) H_0의 가능도비 검정(12회 정의 3.1)의 비는 $\lambda = (\mathrm{RSS}/\mathrm{RSS}_0)^{n/2} = \big(1+\frac{q}{n-p}F\big)^{-n/2}$이므로, 가능도비 검정은 F가 클 때 기각하는 검정이다.

증명. **전략** 제약 아래의 최소화를 2절의 피타고라스로 푼다. $\|\mathbf{y}-X\mathbf{b}\|^2$을 RSS와 "$\hat{\boldsymbol{\beta}}$에서의 거리"로 나누고, 제약을 만족하는 점들 가운데 $\hat{\boldsymbol{\beta}}_0$가 그 거리를 최소로 함을 직교성으로 보인다. 분포는 정리 4.7에서, 가능도비는 σ^2을 먼저 최대화해 얻는다.

(a)(b)

① 모든 $\mathbf{b}$에서 $\|\mathbf{y}-X\mathbf{b}\|^2 = \mathrm{RSS} + (\mathbf{b}-\hat{\boldsymbol{\beta}})'X'X(\mathbf{b}-\hat{\boldsymbol{\beta}})$이다. — 정리 2.2(a)의 항등식

② (a)의 $\hat{\boldsymbol{\beta}}_0$는 $C\hat{\boldsymbol{\beta}}_0 = C\hat{\boldsymbol{\beta}} - WW^{-1}(C\hat{\boldsymbol{\beta}}-\mathbf{d}) = \mathbf{d}$, 곧 제약을 만족하고 $X'X(\hat{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}}) = -C'W^{-1}(C\hat{\boldsymbol{\beta}}-\mathbf{d})$이다. — (a)의 식에 C, $X'X$를 곱한다

③ $C\mathbf{b}=\mathbf{d}$인 아무 $\mathbf{b}$에서

$$(\mathbf{b}-\hat{\boldsymbol{\beta}}_0)'X'X(\hat{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}})=-(C\mathbf{b}-C\hat{\boldsymbol{\beta}}_0)'W^{-1}(C\hat{\boldsymbol{\beta}}-\mathbf{d})=0. \tag{2}$$

④ 따라서 $(\mathbf{b}-\hat{\boldsymbol{\beta}})'X'X(\mathbf{b}-\hat{\boldsymbol{\beta}})=(\mathbf{b}-\hat{\boldsymbol{\beta}}_0)'X'X(\mathbf{b}-\hat{\boldsymbol{\beta}}_0)+(\hat{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}})'X'X(\hat{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}})$이다.

③, 더하고 빼기 [T4]

⑤ 첫 항은 0 이상이고 $\mathbf{b}=\hat{\boldsymbol{\beta}}_0$일 때만 0이므로 $\hat{\boldsymbol{\beta}}_0$가 제약 아래 최소점이고 $\mathrm{RSS}_0-\mathrm{RSS}=\|X(\hat{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}})\|^2$이다.

①④, $X'X>0$

⑥ (a)를 넣으면 $\|X(\hat{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}})\|^2=(C\hat{\boldsymbol{\beta}}-\mathbf{d})'W^{-1}C(X'X)^{-1}C'W^{-1}(C\hat{\boldsymbol{\beta}}-\mathbf{d})=(C\hat{\boldsymbol{\beta}}-\mathbf{d})'W^{-1}(C\hat{\boldsymbol{\beta}}-\mathbf{d})$이다.

$C(X'X)^{-1}C'=W$

(c)

① H_0 아래 $C\hat{\boldsymbol{\beta}}-\mathbf{d}=C(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\sim\mathcal{N}_q(\mathbf{0},\sigma^2W)$이다. 정리 4.7(a), 5회 정리 3.8

② (a)(b)의 ⑥으로 $(\mathrm{RSS}_0-\mathrm{RSS})/\sigma^2\sim\chi^2(q)$이다. 5회 정리 3.17

③ 이것은 $\hat{\boldsymbol{\beta}}$의 함수이므로 RSS와 독립이고 $\mathrm{RSS}/\sigma^2\sim\chi^2(n-p)$이다. 정리 4.7(b)(c)

④ 두 χ^2을 자유도로 나눈 비이므로 $F\sim F(q,n-p)$이다. 6회 정의 4.1

(d)

① $\mathbf{b}$를 고정하고 σ^2에 대해 최대로 하면 가능도는 $\left(2\pi\,\mathrm{RSS}(\mathbf{b})/n\right)^{-n/2}e^{-n/2}$이다.

정리 3.6(a)의 계산

② 제약 아래 최대와 전체 최대의 비는 $\lambda=(\mathrm{RSS}_0/\mathrm{RSS})^{-n/2}$이고, (b)와 F의 정의로 $\mathrm{RSS}_0/\mathrm{RSS}=1+\frac{q}{n-p}F$이다. 12회 정의 3.1

조건 확인 W의 역행렬이 있어야 한다. $(X'X)^{-1}$이 양정치이고 C의 행이 일차독립이므로, $\mathbf{v}\neq\mathbf{0}$이면 $C'\mathbf{v}\neq\mathbf{0}$이고 $\mathbf{v}'W\mathbf{v}=(C'\mathbf{v})'(X'X)^{-1}(C'\mathbf{v})>0$이다. W는 양정치이고 가역이다. ■

그림 2.1에서 $\mathrm{RSS}_0-\mathrm{RSS}=\|\hat{\mathbf{y}}-\hat{\mathbf{y}}_0\|^2$은 두 적합값 사이의 거리다. F는 "귀무가설이 잃는 설명"을 "잔차의 크기"로 잰 것이다. (d)로 Wilks 정리(9회 정리 5.2)의 $-2\log\lambda=n\log(1+\frac{q}{n-p}F)$는 $n\to\infty$에서 qF에 가까워지고 $\chi^2(q)$로 수렴한다. 정규 선형모형에서는 그 근사 대신 정확한 F를 쓸 수 있다.

> **예제 4.10** (두 특별한 경우) (i) $q=1$, $C=\mathbf{e}_j'$, $\mathbf{d}=0$이면 $F=t_j^2$이다(연습 4.4). 계수 하나의 F 검정은 양쪽 t 검정이다. (ii) 상수항이 있고 H_0이 "상수항 말고 모든 계수가 0"이면 $\mathrm{RSS}_0=\mathrm{TSS}$이고
>
> $$F=\frac{R^2/(p-1)}{(1-R^2)/(n-p)}.$$
>
> 회귀 요약표 맨 아래의 F가 이것이다. Galton의 자료(5절)에서 $t=15.71$, $F=246.8=15.71^2$

이다.

연습문제

연습 4.1 (멱등이 아니면 ★) $n=2$에서 $A=\mathrm{diag}(2,0)$과 $A=\mathrm{diag}(1,-1)$일 때 $\mathbf{Z}'A\mathbf{Z}$의 분포를 쓰고, 둘 다 χ^2이 아님을 평균과 분산으로 확인하라. 풀이 54쪽

연습 4.2 (S^2의 분산 ★) $S^2=\mathbf{Z}'(I-J)\mathbf{Z}/(n-1)$에 보조정리 4.3를 써서 $\mathrm{Var}\,S^2=2/(n-1)$을 다시 얻어라(6회 따름정리 2.3). 풀이 54쪽

연습 4.3 (일반 공분산의 Craig ★★) $\mathbf{X}\sim\mathcal{N}_n(\mathbf{0},\Sigma)$, $\Sigma\succ 0$이면 $A\Sigma B=0$일 때 $\mathbf{X}'A\mathbf{X}$와 $\mathbf{X}'B\mathbf{X}$가 독립임을 정리 4.4로 보여라. 풀이 54쪽

연습 4.4 ($F=t^2$ ★) 정리 4.9(b)에서 $C=\mathbf{e}_j'$, $\mathbf{d}=0$이면 $F=\hat\beta_j^2/\mathrm{SE}(\hat\beta_j)^2=t_j^2$임을 보여라. 풀이 54쪽

연습 4.5 (R^2에서 F로 ★) 예제 4.10(ii)의 식을 증명하고, Galton의 $R^2=0.2105$, $n=928$, $p=2$로 F를 계산하라. 풀이 54쪽

연습 4.6 (근사와 정확 ★★) Mayer의 자료($n=27$, $p=3$)에서 계수 하나에 대한 가능도비 검정을 Wilks 근사($-2\log\lambda>3.84$)로 하면 실제 크기는 얼마인가? 정리 4.9(d)로 F의 임계값으로 바꿔 계산하라. 풀이 55쪽

5

회귀의 탄생 — Galton에서 Yule까지

역사 Francis Galton은 최소제곱을 쓰지 않았다. 1886년 **유전되는 키의 평범함으로의 회귀**에서 그는 205쌍의 부모와 그 자녀 928명의 키를 표로 만들고, 부모 키(아버지 키와 어머니 키에 1.08을 곱한 값의 평균)의 칸마다 자녀 키의 중앙값을 읽었다. 그 점들은 기울기가 약 2/3인 직선 위에 있었다. 키가 큰 부모의 자녀는 크지만 부모만큼은 아니었다. Galton은 이것을 "평범함으로의 회귀"라 불렀다(5회 2절). Karl Pearson(1896)이 상관계수의 공식을 주었고, Pearson과 Alice Lee(1903)는 천 가족이 넘는 자료로 부모와 자녀의 회귀를 다시 쟀다. 1897년 George Udny Yule은 결정적인 한 걸음을 디뎠다. 회귀선은 정규분포를 가정하지 않아도 최소제곱으로 맞춘 직선이라는 것이다. 천문학의 최소제곱과 유전학의 회귀가 이렇게 만났다. Yule(1899)은 곧바로 이것을 여러 설명변수로 넓혀 영국의 빈곤율을 분석했다.

5.1 회귀는 최선의 선형예측이다

정리 5.1 (모집단의 최소제곱) 확률벡터 $\mathbf{x} \in \mathbb{R}^p$와 확률변수 Y가 유한한 2차 적률을 갖고 $E\mathbf{x}\mathbf{x}'$가 가역이라 하자.

(a) $E(Y - \mathbf{x}'\mathbf{b})^2$을 최소로 하는 $\mathbf{b}$는 $\boldsymbol{\beta}^* = (E\mathbf{x}\mathbf{x}')^{-1}E\mathbf{x}Y$ 하나다. 오차 $u = Y - \mathbf{x}'\boldsymbol{\beta}^*$는 $E\mathbf{x}u = \mathbf{0}$를 만족한다.

(b) $E[Y \mid \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$인 $\boldsymbol{\beta}$가 있으면 $\boldsymbol{\beta}^* = \boldsymbol{\beta}$이다.

(c) $\mathbf{x} = (1, X)'$이면 $\beta_1^* = \operatorname{Cov}(X,Y)/\operatorname{Var}X$, $\beta_0^* = EY - \beta_1^* EX$이다.

증명. (a) 정리 2.2의 피타고라스를 기댓값으로 한다. $E\mathbf{x}u = E\mathbf{x}Y - E\mathbf{x}\mathbf{x}'\boldsymbol{\beta}^* = \mathbf{0}$이므로 교차항 $2E[(\boldsymbol{\beta}^* - \mathbf{b})'\mathbf{x}u]$가 0이고

$$E(Y - \mathbf{x}'\mathbf{b})^2 = Eu^2 + (\mathbf{b} - \boldsymbol{\beta}^*)'(E\mathbf{x}\mathbf{x}')(\mathbf{b} - \boldsymbol{\beta}^*)$$

이다. $E\mathbf{x}\mathbf{x}'$가 양정치이므로 둘째 항은 $\mathbf{b} = \boldsymbol{\beta}^*$에서만 0이다. (b) 반복기댓값으로 $E\mathbf{x}(Y - \mathbf{x}'\boldsymbol{\beta}) = E\big[\mathbf{x}\,E(Y - \mathbf{x}'\boldsymbol{\beta} \mid \mathbf{x})\big] = \mathbf{0}$이므로 $\boldsymbol{\beta}$가 $(E\mathbf{x}\mathbf{x}')\boldsymbol{\beta} = E\mathbf{x}Y$를 풀고, 해는 하나뿐이다. (c) 두 식을 풀면 된다(연습 2.1의 모집단판). ∎

5회 따름정리 3.19는 이변량 정규에서 조건부 기댓값이 직선임을 보였다. 정리 5.1는 정규가 아니어도 그 직선이 "최선의 직선"이라는 뜻을 갖는다고 말한다. 이것이 Yule의 통찰이다. 표본의

최소제곱 $\hat{\boldsymbol{\beta}} = (\frac{1}{n}\sum \mathbf{x}_i\mathbf{x}_i')^{-1}(\frac{1}{n}\sum \mathbf{x}_i y_i)$은 이 식의 기댓값을 표본평균으로 바꾼 것이므로, 대수의 법칙(7회)으로 $\boldsymbol{\beta}^*$에 수렴한다. 선형모형이 틀려도 최소제곱은 "가장 좋은 선형 근사"를 추정한다.

예제 5.2 (Galton의 928명) 부모 키 x로 자녀 키 y를 최소제곱으로 맞추면

$$\hat{y} = 23.94 + 0.646\,x, \qquad \mathrm{SE}(\hat{b}_1) = 0.041, \qquad R^2 = 0.210$$

이다. 거꾸로 y로 x를 맞춘 기울기는 0.326이다. 두 기울기의 곱 $0.646 \times 0.326 = 0.210$은 r^2이다($r = 0.459$). 두 회귀선은 다른 직선이다. 부모에서 자녀로도, 자녀에서 부모로도 평균 쪽으로 당겨진다(그림 5.1 왼쪽). Pearson–Lee(1903)의 도수표를 무게로 맞추면(따름정리 3.7) 아버지–아들 기울기는 0.519 (1078쌍, $r = 0.514$), 아버지–딸 0.479, 어머니–아들 0.563, 어머니–딸 0.545다. Galton의 기울기가 더 큰 것은 그의 설명변수가 두 부모의 평균이어서 분산이 개인 키의 약 절반이기 때문이다.

가정 주의 **회귀의 오류.** 1933년 Horace Secrist는 **사업에서 평범함의 승리**에서, 1920년 수익이 높았던 회사들이 10년 뒤 평균으로 내려오고 낮았던 회사들이 올라온 것을 보고 경쟁이 모든 회사를 평범하게 만든다고 결론지었다. Hotelling은 서평에서 이것이 상관이 1보다 작으면 반드시 생기는 회귀일 뿐이라고 지적했다. 같은 자료를 거꾸로 보면 1930년의 상위 회사들은 1920년에 평균에 더 가까웠다(연습 5.2). 처치 전에 극단인 집단을 고르면 처치가 없어도 "개선"된다.

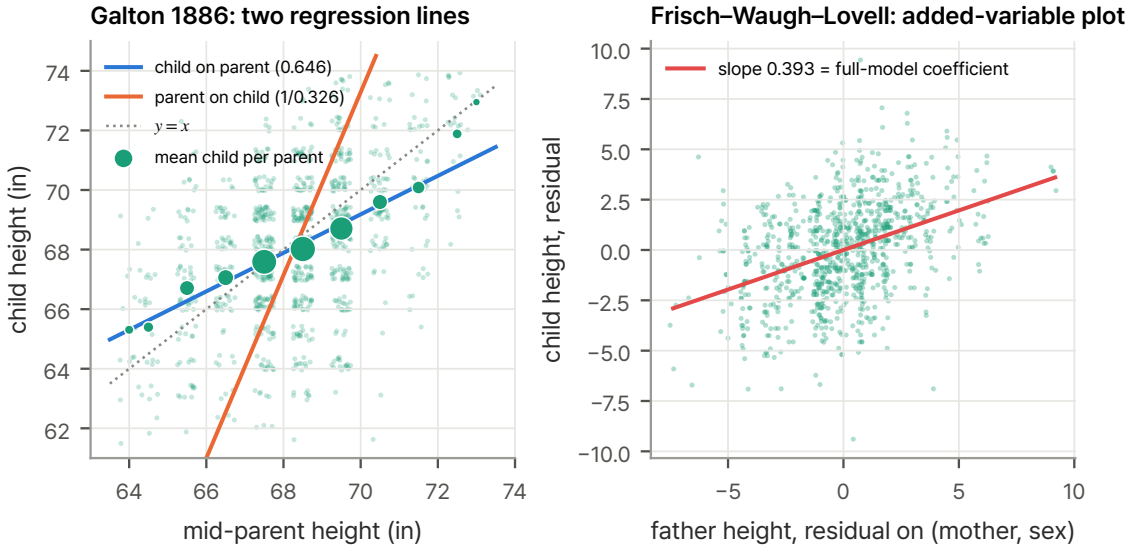


그림 5.1 왼쪽: Galton의 928명(점은 겹치지 않게 조금 흔들었다). 자녀를 부모로 맞춘 직선(파랑, 기울기 0.646)과 부모를 자녀로 맞춘 직선(주황)은 다른 직선이고, 둘 다 $y = x$(점선)보다 평균 쪽으로 기울어 있다. 초록 점은 부모 키마다 자녀 키의 평균(크기는 인원)이다. 오른쪽: 가족 자료의 편회귀 그림. 아버지 키와 자녀 키를 각각 (상수, 어머니 키, 성별)로 맞춘 잔차끼리의 기울기 0.393은 다중회귀의 아버지 계수와 같다(정리 5.3).

5.2 다른 변수를 고정한다는 것

설명변수가 여럿이면 계수 하나는 무엇을 뜻하나? Yule(1907)은 "다른 변수의 영향을 뺀" 편회귀계수라는 답을 주었다. 다음 정리가 그 뜻을 정확히 만든다.

정리 5.3 (Frisch–Waugh–Lovell) $X=[X_1,X_2]$ $(n\times p_1,\, n\times p_2)$이 완전 계수이고 $\hat{\boldsymbol{\beta}}=(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\beta}}_2)$가 $\mathbf{y}$를 X로 맞춘 최소제곱 해라 하자. H_1을 X_1의 hat 행렬, $M_1=I-H_1$이라 하면

$$\hat{\boldsymbol{\beta}}_2=(X_2'M_1X_2)^{-1}X_2'M_1\mathbf{y}$$

이다. 곧 $\hat{\boldsymbol{\beta}}_2$는 "$X_1$로 맞추고 남은 $\mathbf{y}$"를 "X_1로 맞추고 남은 X_2"로 맞춘 계수이고, 두 회귀의 잔차는 같다.

증명. **전략** 전체 회귀의 식 $\mathbf{y}=X_1\hat{\boldsymbol{\beta}}_1+X_2\hat{\boldsymbol{\beta}}_2+\mathbf{e}$에 M_1을 곱해 X_1을 지운다. 남은 식이 $M_1\mathbf{y}$를 M_1X_2로 맞추는 회귀의 정규방정식을 그대로 만족함을 확인한다.

① $\mathbf{y}=X_1\hat{\boldsymbol{\beta}}_1+X_2\hat{\boldsymbol{\beta}}_2+\mathbf{e}$이고 $X_1'\mathbf{e}=\mathbf{0}$, $X_2'\mathbf{e}=\mathbf{0}$이다. 정리 2.2(a)

② $M_1X_1=0$이고, $\mathbf{e}\perp\mathrm{col}(X_1)$이므로 $M_1\mathbf{e}=\mathbf{e}$이다. 정리 2.5(a)(c)

③ ①의 양변에 M_1을 곱하면 $M_1\mathbf{y}=M_1X_2\hat{\boldsymbol{\beta}}_2+\mathbf{e}$이다. ②

④ $(M_1X_2)'\mathbf{e}=X_2'M_1\mathbf{e}=X_2'\mathbf{e}=\mathbf{0}$이다. M_1 대칭, ②①

⑤ ③④는 $\hat{\boldsymbol{\beta}}_2$가 $M_1\mathbf{y}$를 M_1X_2로 맞추는 정규방정식을 풀고 그 잔차가 $\mathbf{e}$라는 뜻이다. 정리 2.2(a)

⑥ $M_1X_2\mathbf{v}=\mathbf{0}$이면 $X_2\mathbf{v}\in\mathrm{col}(X_1)$이고 X가 완전 계수이므로 $\mathbf{v}=\mathbf{0}$이다. 따라서 $X_2'M_1X_2=(M_1X_2)'(M_1X_2)$는 가역이고 해는 하나다. 정리 2.2(c)

■

Frisch와 Waugh(1933)는 시계열에서 추세를 미리 빼고 회귀하는 것과 추세를 변수로 넣는 것이 같은 계수를 준다는 것을 보였다. Lovell(1963)이 일반형으로 썼다. 상수항 열 $X_1=\mathbf{1}$을 빼는 것은 평균을 빼는 것이다. 단순회귀의 기울기가 중심화한 자료의 기울기인 이유가 이것이다.

따름정리 5.4 (계수의 분산과 분산팽창계수) 상수항이 있는 모형에서 x_j를 나머지 열로 맞춘 결정계수를 R_j^2, $S_{jj} = \sum_i (x_{ij} - \bar{x}_j)^2$이라 하면

$$\operatorname{Var}\hat{\beta}_j = \frac{\sigma^2}{S_{jj}(1 - R_j^2)}, \qquad \mathrm{VIF}_j = \frac{1}{1 - R_j^2}.$$

증명. **전략** 정리 5.3로 $\hat{\beta}_j$를 "나머지 열로 맞추고 남은 x_j"에 대한 단순한 계수로 쓴다. 그 분모가 x_j를 나머지로 맞춘 잔차제곱합이고, 명제 2.6로 $S_{jj}(1-R_j^2)$이다. 단계별 증명은 연습 5.5(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 55쪽에 있다.

VIF_j는 다른 변수와의 상관 때문에 $\hat{\beta}_j$의 분산이 몇 배 커지는지 말한다. 7절에서 VIF가 수백인 자료를 본다.

예제 5.5 (Galton의 가족 자료) Galton의 원장부를 옮긴 가족 자료(205가족의 자녀 934명)로 자녀 키를 아버지 키, 어머니 키, 성별(아들이면 1)로 맞추면

$$\hat{y} = 16.52 + 0.393\,(\text{아버지}) + 0.318\,(\text{어머니}) + 5.21\,(\text{아들})$$

이고 표준오차는 차례로 2.73, 0.029, 0.031, 0.142, $R^2 = 0.635$, $s = 2.16$이다. 아버지 계수 0.393은 아버지 키를 (상수, 어머니, 성별)로 맞춘 잔차에 자녀 키의 잔차를 맞춘 기울기와 소수점 열두 자리까지 같다(정리 5.3, 그림 5.1 오른쪽). "아버지와 어머니의 계수가 같다"는 $H_0 : \beta_{\text{아버지}} = \beta_{\text{어머니}}$의 F 검정(정리 4.9, $q = 1$)은 $F = 3.00$, $p = 0.084$다. Galton의 "어머니 키에 1.08을 곱한다"는 처방은 아들과 딸의 평균 키의 비 69.23/64.10 = 1.080과 같다.

가정 주의 한 가족의 자녀들은 독립이 아니다. 같은 부모에게서 같은 유전자와 같은 식탁을 물려받았다. 정의 1.1의 $\operatorname{Cov}\boldsymbol{\varepsilon} = \sigma^2 I$가 깨지므로 위의 표준오차는 실제보다 작다. 가족마다 공통의 효과를 넣는 계층모형은 16회에서 다룬다.

명제 5.6 (누락변수의 치우침) 참 모형이 $\mathbf{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$인데 X_1만으로 맞추면 $\tilde{\boldsymbol{\beta}}_1 = (X_1'X_1)^{-1}X_1'\mathbf{y}$의 기댓값은

$$E\tilde{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1 + \Gamma\boldsymbol{\beta}_2, \qquad \Gamma = (X_1'X_1)^{-1}X_1'X_2$$

이다. Γ는 빠진 변수 X_2를 X_1로 맞춘 계수다. 치우침은 $X_1'X_2\boldsymbol{\beta}_2 = \mathbf{0}$일 때만 0이다.

증명. $E\tilde{\boldsymbol{\beta}}_1 = (X_1'X_1)^{-1}X_1'(X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2)$. $X_1'X_1$이 가역이므로 마지막 문장이 나온다. ∎

빠진 변수가 결과와 관계있고($\boldsymbol{\beta}_2 \neq \mathbf{0}$) 남은 변수와 상관되면($\Gamma \neq 0$), 남은 변수의 계수는 빠진 변수의 몫까지 떠안는다. 7절의 Yule–Simpson 역설이 이것의 극단이다.

연습문제

연습 5.1 (두 회귀선 ★) y를 x로 맞춘 기울기 b_{yx}와 x를 y로 맞춘 기울기 b_{xy}의 곱이 r^2임을 보여라. 두 직선을 같은 (x, y) 평면에 그리면 어느 쪽이 더 가파른가? Galton의 수치로 확인하라. 풀이 55쪽

연습 5.2 (회귀의 오류 ★★) (X, Y)가 평균 μ, 분산 σ^2이 같고 상관이 $\rho < 1$인 이변량 정규라 하자. $E[Y - X \mid X = x] = -(1-\rho)(x-\mu)$와 $E[X - Y \mid Y = y] = -(1-\rho)(y-\mu)$를 보여라. 1920년($X$)의 상위 회사가 1930년($Y$)에 내려오는 것과 1930년의 상위 회사가 1920년에 낮았던 것이 어떻게 동시에 참인가? 풀이 55쪽

연습 5.3 (직선이 놓치는 것 ★★) $X \sim \mathcal{N}(0,1)$, $Y = X^2$이면 정리 5.1의 최선의 직선은 $\beta_0^* = 1$, $\beta_1^* = 0$임을 보여라. Y는 X로 완전히 정해지는데 최선의 직선의 설명력은 0이다. 무엇이 잘못되었나? 풀이 55쪽

연습 5.4 (성별을 빼면 ★★) 가족 자료에서 자녀 키를 Galton의 부모 키(아버지 키와 어머니 키의 1.08배의 평균)만으로 맞추면 기울기 0.637, $R^2 = 0.103$이고, 성별을 더하면 기울기 0.687, $R^2 = 0.633$이다. 명제 5.6로 기울기가 바뀐 방향을 설명하라. (노트북에서 Γ를 계산한다.) 풀이 55쪽

연습 5.5 (계수의 분산과 VIF ★★ 유도형) 따름정리 5.4를 다음 순서로 증명하라. X_1을 j번째 열을 뺀 나머지 열(상수항 포함), $M = I - H_1$이라 하자.

(a) 정리 5.3로 $\hat\beta_j = \mathbf{x}^{(j)\mathsf{T}} M \mathbf{y} / \|M\mathbf{x}^{(j)}\|^2$임을 보여라.

(b) $M\mathbf{y} = M\mathbf{x}^{(j)}\beta_j + M\boldsymbol{\varepsilon}$를 보이고 $\operatorname{Var}\hat\beta_j = \sigma^2 / \|M\mathbf{x}^{(j)}\|^2$을 얻어라.

(c) $\|M\mathbf{x}^{(j)}\|^2$이 x_j를 X_1로 맞춘 잔차제곱합임을 보이고, 명제 2.6로 $S_{jj}(1-R_j^2)$와 같음을 보여라. 왜 X_1에 상수항이 있어야 하는가?

풀이 55쪽

6

원자료로 — 회귀 요약표의 모든 숫자

통계 소프트웨어의 회귀 요약표에는 숫자가 스무 개쯤 있다. 이 절은 그 숫자들이 모두 1–4절의 정리에서 나온다는 것을 보이고, 18세기의 세 문제를 그 표로 다시 읽는다.

6.1 요약표의 해부

알고리즘 6.1 (회귀 요약표) 입력은 X ($n \times p$, 첫 열은 $\mathbf{1}$)와 $\mathbf{y}$다.

1. 알고리즘 2.8로 $\hat{\boldsymbol{\beta}}$, $\mathbf{e}$, RSS, $(X'X)^{-1}$을 구한다. $s^2 = \mathrm{RSS}/(n-p)$.
2. 계수마다 $\mathrm{SE}_j = s\sqrt{[(X'X)^{-1}]_{jj}}$, $t_j = \hat\beta_j/\mathrm{SE}_j$, $p_j = 2P(t_{n-p} > |t_j|)$, 구간 $\hat\beta_j \pm t_{n-p,0.025}\mathrm{SE}_j$ (따름정리 4.8).
3. $R^2 = 1 - \mathrm{RSS}/\mathrm{TSS}$, 수정 $R^2 = 1 - (1-R^2)\frac{n-1}{n-p}$, $F = \frac{R^2/(p-1)}{(1-R^2)/(n-p)}$와 그 p-값 $P(F(p-1, n-p) > F)$ (예제 4.10).
4. 로그가능도 $\hat\ell = -\frac{n}{2}\left(\log\frac{2\pi\,\mathrm{RSS}}{n} + 1\right)$ (정리 3.6(a)의 최대값), $\mathrm{AIC} = -2\hat\ell + 2p$, $\mathrm{BIC} = -2\hat\ell + p\log n$.

수정 R^2은 R^2을 RSS와 TSS의 불편추정량의 비로 바꾼 것이다($1 - s^2/\hat\sigma_y^2$). 변수를 더하면 R^2은 줄지 않지만(연습 2.5) 수정 R^2은 줄 수 있다. AIC와 BIC에서 모수의 수를 p로 세는 것은 Python의 statsmodels 관행이다. R의 `AIC()`는 σ^2까지 $p+1$개로 세어 2만큼 크다. 모형을 비교할 때는 같은 관행끼리 비교하면 된다.

예제 6.2 (Galton의 요약표)

	계수	SE	t	p	95% 구간
상수	23.9415	2.8109	8.517	6.5×10^{-17}	[18.425, 29.458]
부모 키	0.6463	0.0411	15.711	1.7×10^{-49}	[0.566, 0.727]

$n = 928$, $s = 2.2385$, $R^2 = 0.2105$, 수정 $R^2 = 0.2096$, $F = 246.84$ ($p = 1.7\times10^{-49}$)

$\hat\ell = -2063.58$, AIC $= 4131.16$, BIC $= 4140.83$

기울기의 p-값과 F의 p-값이 같은 것은 $F = t^2$이기 때문이다(연습 4.4). 부모 키가 72인치인 가족의 자녀 평균 키의 95% 신뢰구간은 70.47 ± 0.33, 그 가족의 자녀 한 명의 95% 예측구간은

70.47 ± 4.41이다(따름정리 4.8). 평균은 0.3인치 안쪽으로 알지만 한 사람은 4인치 넘게 틀릴 수 있다. $R^2 = 0.21$이 말하는 것이 이것이다.

6.2 Mayer (1750) — 세 방법의 비교

방법	$\hat\beta$	$\hat\alpha$	$\hat\theta$	$\sum e_i^2$	$\sum \lvert e_i \rvert$
최소제곱	14°33.5′	1°30.3′	−2°44.3′	0.5689	3.220
Mayer의 묶음 평균	14°32.8′	1°29.7′	−3°49.1′	0.5767	—
최소절댓값	14°34.1′	1°30.9′	−5°2.8′	—	3.066
최소제곱의 SE	2.1′	2.5′	1°56.8′	$s = 0.154° = 9.2'$	

예제 6.3 (Mayer의 θ는 정해지지 않았다) 최소제곱의 계수 표에서 $\hat\gamma = -0.0719$, SE $= 0.0508$, $t = -1.41$, $p = 0.17$ ($t(24)$)이다. $\gamma = \alpha \sin\theta$이므로 "$\theta = 0$"을 5%에서 기각하지 못한다. $\hat\theta = \arcsin(\hat\gamma/\hat\alpha)$의 표준오차를 델타 방법(7회 정리 5.4)으로 구하면 1.95°다(연습 6.2). 세 방법의 $\hat\theta$가 −2.7°에서 −5.0°까지 흩어진 것은 이 불확실성의 크기와 맞는다. 반면 β와 α는 세 방법이 2분 안쪽에서 일치한다. 관측 하나의 오차 $s = 9.2'$을 27개로 결합해 β를 2.1′의 정밀도로 정했다. 대략 $9.2/\sqrt{27} = 1.8$이다.

결합하지 않고 세 방정식만 골라 풀면 어떻게 되나? 2925가지 선택에서 $\hat\beta$의 중앙값은 14.55°로 맞지만, 범위는 −426°에서 104°까지이고, 최소제곱 해의 1분 안쪽에 드는 것은 5.1%뿐이다. Euler가 두려워한 것은 이런 불안정이었다. 결합은 그것을 없앤다.

최소절댓값 해는 1, 5, 19번 방정식을 정확히 지난다(명제 1.8). 그 해의 절댓값 합 3.066은 최소제곱 해의 3.220보다 작고, 제곱합에서는 반대다. 각 방법은 자기 기준에서 최선이다.

6.3 Laplace의 토성 (1787)

목성과 토성의 운동에는 수백 년 주기의 커다란 어긋남이 있었다. Euler가 1749년에 풀지 못한 그 문제다. Laplace는 1785–1787년 두 행성의 공전주기가 거의 5:2의 비라는 데서 그 원인을 찾았고, 1591년부터 1785년까지의 관측 24개로 토성 궤도의 보정값 네 개를 정했다. 방정식은

$$y_i = b_0 + b_1 t_i + b_2 x_{i2} + b_3 x_{i3} + \varepsilon_i \qquad (i = 1, \dots, 24)$$

꼴이다. t_i는 1750년 무렵을 0으로 잰 시간(년), x_{i2}, x_{i3}는 궤도 위치의 사인과 코사인, y_i는 관측과 이론의 차이(각의 분)다. Laplace는 Mayer와 비슷한 결합으로 풀었다.

예제 6.4 (토성의 요약표) 최소제곱 해는 $\hat{b}_0 = 3.43$ (SE 0.27), $\hat{b}_1 = 0.0032$ (0.0040), $\hat{b}_2 = 0.70$ (0.30), $\hat{b}_3 = 0.27$ (0.32)이다. 잔차의 표준편차는 $s = 1.02'$이고, 가장 큰 잔차는 1753년의 $-1.68'$이다. 17–18세기 망원경 관측의 오차가 1분 정도였으므로, 잔차가 그 크기라면 이론이 관측을 설명한다는 뜻이다. 요약표의 $R^2 = 0.25$, $F = 2.24$ ($p = 0.11$)는 여기서 별 의미가 없다. y_i가 이미 "관측 − 이론"이어서, 설명할 것이 거의 남지 않은 것이 이 모형의 성공이기 때문이다. R^2은 모형의 좋고 나쁨이 아니라 $\mathbf{y}$의 변동 가운데 설명된 비율일 뿐이다.

연습문제

연습 6.1 (요약표 읽기 ★) Galton의 요약표에서 기울기의 t, p, 95% 구간을 계수와 SE만으로 다시 계산하라($t_{926,0.025} = 1.9625$). 기울기가 1이라는 가설(부모만큼 자란다)의 t는 얼마인가? 풀이 56쪽

연습 6.2 (θ의 델타 방법 ★★) $\theta = g(\alpha, \gamma) = \arcsin(\gamma/\alpha)$의 기울기 벡터를 구하고, 다변량 델타 방법(7회 정리 5.4)으로 $\mathrm{SE}(\hat{\theta}) \approx \sqrt{\nabla g'\, \widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\gamma})\, \nabla g}$를 쓰라. 노트북의 공분산으로 1.95°를 확인하라. 풀이 56쪽

연습 6.3 (로그가능도와 AIC ★) 알고리즘 6.1의 $\hat{\ell}$ 식을 유도하고 Galton의 $\mathrm{RSS} = 4640.27$, $n = 928$로 $\hat{\ell}$, AIC, BIC를 계산하라. R의 관행이면 AIC는 얼마인가? 풀이 56쪽

7

부숴보기 — 가정이 깨질 때

정의 1.1의 가정은 셋이다. 평균이 $X\boldsymbol{\beta}$이고, 분산이 같고, 오차가 서로 무상관이다. 4절은 여기에 정규성을 더했다. 하나씩 깨 본다.

7.1 분산이 다르면 — 샌드위치

명제 7.1 (이분산 아래의 최소제곱) $\operatorname{Cov}\boldsymbol{\varepsilon} = \Sigma = \operatorname{diag}(\sigma_1^2, \dots, \sigma_n^2)$이면 $\hat{\boldsymbol{\beta}}$은 여전히 불편이고

$$\operatorname{Cov}\hat{\boldsymbol{\beta}} = (X'X)^{-1}\Big(\sum_{i=1}^{n} \sigma_i^2 \mathbf{x}_i \mathbf{x}_i'\Big)(X'X)^{-1}$$

이다. 보통의 공식 $s^2(X'X)^{-1}$은 일반적으로 이것과 다르다.

증명. 불편성은 $E\boldsymbol{\varepsilon} = \mathbf{0}$만 쓴다. 공분산은 5회 보조정리 3.1로 $(X'X)^{-1}X'\Sigma X(X'X)^{-1}$이고 $X'\Sigma X = \sum \sigma_i^2 \mathbf{x}_i \mathbf{x}_i'$이다. ∎

빵 두 장 $(X'X)^{-1}$ 사이에 속 $\sum \sigma_i^2 \mathbf{x}_i \mathbf{x}_i'$가 있어서 **샌드위치**라 부른다. 속의 σ_i^2을 e_i^2으로 바꾼 추정량(HC0)을 Eicker(1963), Huber(1967), White(1980)가 차례로 내놓았다. $n \to \infty$에서 일치추정량이다. 작은 표본에서는 $n/(n-p)$를 곱하거나(HC1) e_i^2을 $e_i^2/(1-h_{ii})^2$로 바꾼다(HC3). 명제 2.7(c)로 $Ee_i^2 \approx \sigma_i^2(1-h_{ii})$이어서 잔차가 오차보다 작기 때문이다.

예제 7.2 (분산이 x^2에 비례할 때) x_i를 0에서 10까지 고르게 100개 잡고 $y_i = 1 + 0.5x_i + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, (0.05x_i^2)^2)$이라 하자. 기울기의 참 표준편차는 명제 7.1로 0.097이다. 보통의 공식으로 추정한 분산의 기댓값은 참 분산의 0.63배이고(연습 7.1), 표준오차로 바꾸면 0.077이다. 95% t-구간의 실제 포함확률은 약 88%로 떨어진다. HC1과 HC3 표준오차로 바꾸면 약 94%와 95%로 돌아온다(노트북, 4만 번 반복). 분산이 큰 점이 x의 양 끝, 곧 기울기를 정하는 데 힘이 큰 곳에 있어서 보통의 공식이 불확실성을 과소평가한다.

7.2 한 점이 모든 것을 정하면 — 지렛대와 영향

명제 7.3 (한 점을 빼고 다시 맞추기) $h_{ii} < 1$일 때 관측 i를 빼고 맞춘 최소제곱 해를 $\hat{\boldsymbol{\beta}}_{(i)}$라 하면

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \frac{(X'X)^{-1}\mathbf{x}_i e_i}{1 - h_{ii}}, \qquad y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(i)} = \frac{e_i}{1 - h_{ii}}.$$

증명. **전략** 관측 하나를 빼면 $X'X$에서 계수 1인 행렬 $\mathbf{x}_i\mathbf{x}_i'$가 빠진다. 그 역행렬을 Sherman–Morrison 공식(연습 7.2)으로 쓰고 전개한다. 단계별 증명은 연습 7.5(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 57쪽에 있다.

빼고 맞춘 잔차 $e_i/(1 - h_{ii})$의 제곱합(PRESS)은 한-점-빼기 교차검증의 오차다. 최소제곱에서는 n번 다시 맞출 필요 없이 한 번의 적합으로 계산된다. 지렛대가 1에 가까우면 그 점을 빼는 순간 해가 크게 움직인다.

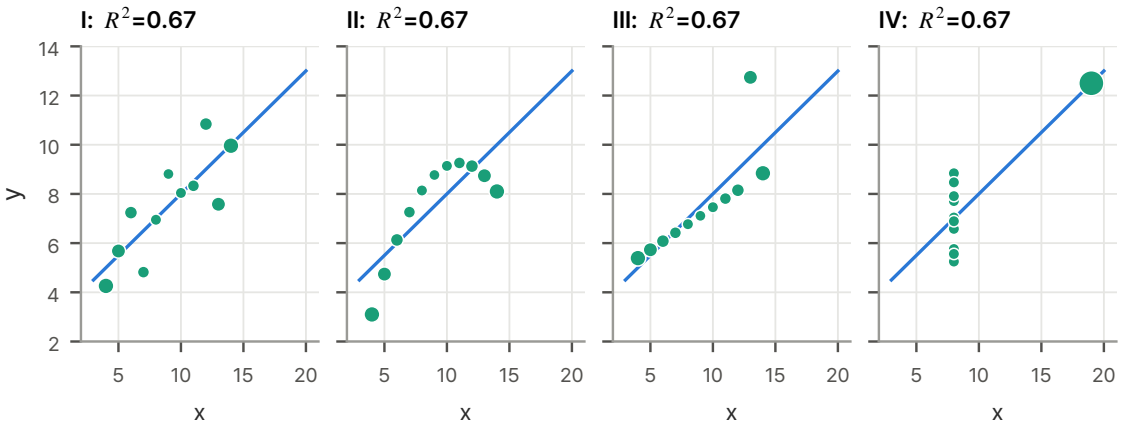


그림 7.1 Anscombe의 네 자료. 최소제곱 직선(파랑)과 R^2이 모두 같다. 점의 크기는 지렛대 h_{ii}에 비례한다. 넷째 자료의 오른쪽 점은 $h_{ii}=1$이다.

예제 7.4 (Anscombe의 네 자료 (1973)) Francis Anscombe는 $n=11$인 자료 네 개를 만들었다. 네 자료 모두 최소제곱 직선이 $\hat{y}=3.00+0.500x$이고 $R^2=0.667$이며 요약표의 숫자가 거의 같다. 그러나 그림 7.1을 보면 첫째만 직선 모형에 맞는다. 둘째는 곡선이고, 셋째는 이상치 하나가 기울기를 끌어올렸으며, 넷째는 x가 열 번 8이고 한 번 19다. 넷째의 $x=19$인 점은 지렛대가 1이어서(연습 2.3) 잔차가 언제나 0이고, 기울기는 그 한 점과 나머지 열 점의 평균을 잇는 직선의 기울기다. 그 점을 빼면 기울기는 정해지지도 않는다. 요약표만 보지 말고 그림을 보라는 것이 Anscombe의 교훈이다.

7.3 설명변수가 서로 닮으면 — 공선성

예제 7.5 (VIF 344) $x_1\sim\mathcal{N}(0,1)$, $x_2=x_1+\mathcal{N}(0,0.05^2)$로 $n=100$개를 만들고 $y=1+x_1+x_2+\varepsilon$, $\varepsilon\sim\mathcal{N}(0,1)$로 맞춘다. 모집단에서는 $r_{12}^2=1/1.0025$이므로 VIF $=401$이고, 한 표본에서는 $r_{12}=0.9985$, VIF $=1/(1-r_{12}^2)=344$였다. 따름정리 5.4로 두 계수의 분산은 상관이 없을 때의 344배이고, 그 표본에서 $\hat\beta_1=4.71$, $\hat\beta_2=-2.87$, 표준오차는 둘 다 2.06이다. 참값은 둘 다 1이다. 두 계수 각각은 거의 아무것도 말하지 않는다. 그러나 합 $\hat\beta_1+\hat\beta_2=1.83$의 표준오차는 0.111이다. 자료는 "두 변수를 함께 올리면 y가 얼마나 오르나"에는 정확히 답하고, "하나만 올리면"에는 답하지 못한다. 두 변수가 따로 움직인 적이 거의 없기 때문이다. 그림 7.2의 가운데가 반복 표본에서의 $(\hat\beta_1,\hat\beta_2)$다. 직선 $\beta_1+\beta_2=2$를 따라 길게 늘어선다.

공선성은 모형의 잘못이 아니라 자료의 한계다. 처방은 두 변수를 따로 움직인 자료를 더 모으거나, 합처럼 정해지는 조합만 보고하거나, 8절의 ridge처럼 치우침을 받아들이는 것이다.

7.4 변수 하나를 빼면 — Yule–Simpson의 역설

예제 7.6 (Berkeley 대학원 입학 (1973)) 1973년 가을 Berkeley 대학원의 큰 여섯 학과에 4526명이 지원했다. 남자의 합격률은 44.5%, 여자는 30.4%다. 합격(1 또는 0)을 여자 표시로 회귀하면(선형확률모형) 계수는 −0.142다. 학과 표시 여섯 개를 넣으면 계수는 +0.018 (SE 0.015)이 된다. 학과 안에서는 차이가 거의 없거나 여자 쪽이 조금 높다. 여자는 합격률이 낮은 학과에 많이 지원했다. 합격률이 6%인 F과, 25%인 E과, 35%인 C과의 여자 비율은 48%, 67%, 65%였고, 합격률이 64%와 63%인 A과와 B과에서는 12%와 4%였다. 명제 5.6의 언어로, 빠진 변수(학과)가 결과(합격)와 관계있고 남은 변수(성별)와 상관되어 성별의 계수가 학과의 몫을 떠안았다(그림 7.2 오른쪽).

역사 Pearson(1899)과 Yule(1903)은 두 집단을 합칠 때 연관의 방향이 뒤집힐 수 있다는 것을 알았다. Simpson(1951)이 이것을 정리했고, 1975년 Bickel·Hammel·O'Connell이 Berkeley 자료로 유명하게 만들었다. 어느 쪽이 "옳은" 계수인지는 통계학만으로 정해지지 않는다. 학과 선택이 성별의 결과인지, 성별과 무관한 다른 원인의 결과인지에 따라 묻는 질문이 다르다. 회귀계수는 "다른 변수를 고정했을 때"의 비교이고, 무엇을 고정할지는 인과의 구조가 정한다.

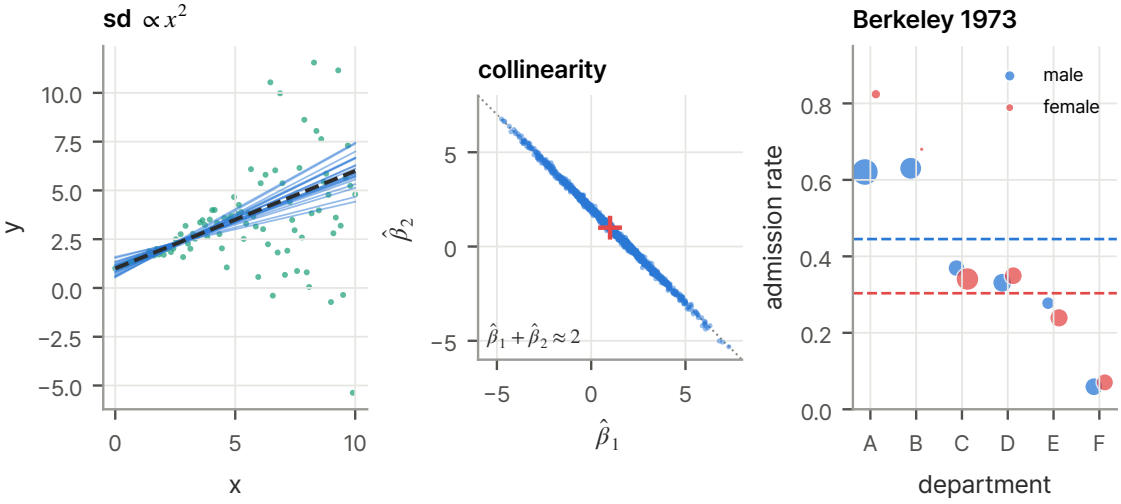


그림 7.2 가정이 깨질 때. 왼쪽: 분산이 x^2에 비례하는 자료와 반복 표본의 적합 직선 20개. 직선의 흔들림은 양 끝의 큰 분산에서 온다. 가운데: 공선성(모집단 VIF = 401) 아래 반복 표본의 $(\hat\beta_1, \hat\beta_2)$. 각각은 크게 흩어지지만 합은 2 근처에 묶여 있다. 오른쪽: Berkeley의 여섯 학과. 학과마다 남녀 합격률(점 크기는 지원자 수)과 전체 합격률(가로선). 전체에서는 여자가 낮지만 학과 안에서는 거의 같다.

연습문제

연습 7.1 (보통 공식의 기댓값 ★★) $\operatorname{Cov}\boldsymbol{\varepsilon} = \Sigma$일 때 보통의 분산 추정량 $s^2[(X'X)^{-1}]_{jj}$의 기댓값이 $\dfrac{\operatorname{tr}\big((I-H)\Sigma\big)}{n-p}[(X'X)^{-1}]_{jj}$임을 보여라(보조정리 3.1). 예제 7.2의 비 0.63을 노트북으로 확인하라. 풀이 56쪽

연습 7.2 (Sherman–Morrison ★) $1-\mathbf{u}'A^{-1}\mathbf{u}\neq 0$이면 $(A-\mathbf{u}\mathbf{u}')\Big(A^{-1}+\dfrac{A^{-1}\mathbf{u}\mathbf{u}'A^{-1}}{1-\mathbf{u}'A^{-1}\mathbf{u}}\Big)=I$임을 확인하라. 풀이 56쪽

연습 7.3 (공선성의 두 표준오차 ★★) $\operatorname{Var}(\hat\beta_1+\hat\beta_2)=\operatorname{Var}\hat\beta_1+\operatorname{Var}\hat\beta_2+2\operatorname{Cov}(\hat\beta_1,\hat\beta_2)$에서, 두 분산이 크더라도 합의 분산이 작을 수 있는 이유를 $\operatorname{Cov}(\hat\beta_1,\hat\beta_2)$의 부호로 설명하라. 상수항이 없고 x_1, x_2가 평균 0, 분산 1, 상관 r인 경우 $(X'X)^{-1}$을 직접 쓰고 $\operatorname{Var}(\hat\beta_1\pm\hat\beta_2)$를 구하라. 풀이 57쪽

연습 7.4 (두 학과의 역설 ★★) 두 학과 P, Q가 있다. P는 남자 80명 중 48명(60%), 여자 20명 중 13명(65%)을, Q는 남자 20명 중 4명(20%), 여자 80명 중 20명(25%)을 뽑았다. 학과마다 여자가 높지만 전체로는 남자가 높음을 보이고, 이것을 명제 5.6의 $\Gamma\beta_2$로 설명하라. 풀이 57쪽

연습 7.5 (한 점을 빼고 다시 맞추기 ★★ 유도형) $A=X'X$, $h_{ii}=\mathbf{x}_i'A^{-1}\mathbf{x}_i<1$이다. 명제 7.3를 다음 순서로 증명하라.

(a) 관측 i를 뺀 설계의 $X'X$와 $X'\mathbf{y}$가 $A-\mathbf{x}_i\mathbf{x}_i'$와 $X'\mathbf{y}-\mathbf{x}_i y_i$임을 보여라.

(b) 연습 7.2의 공식을 $\mathbf{u}=\mathbf{x}_i$에 써서 $\hat{\boldsymbol{\beta}}_{(i)}$를 A^{-1}, $\mathbf{x}_i$, h_{ii}로 써라.

(c) 전개해 $\hat{\boldsymbol{\beta}}_{(i)}=\hat{\boldsymbol{\beta}}-A^{-1}\mathbf{x}_i y_i+\dfrac{A^{-1}\mathbf{x}_i(\hat y_i-h_{ii}y_i)}{1-h_{ii}}$를 얻고, 이것을 $\hat{\boldsymbol{\beta}}-\dfrac{A^{-1}\mathbf{x}_i e_i}{1-h_{ii}}$로 정리하라.

(d) (c)에 $\mathbf{x}_i'$를 곱해 $y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(i)}=e_i/(1-h_{ii})$를 얻어라.

풀이 57쪽

8

오늘의 최소제곱 — 경사하강과 weight decay

신경망의 선형층 `nn.Linear`에 평균제곱오차를 손실로 붙이면 그것이 최소제곱이다. 다른 것은 푸는 방법뿐이다. 정규방정식이나 QR 대신 경사하강을 쓴다. 이 절은 그 차이가 무엇을 낳는지 본다.

8.1 경사하강의 속도는 조건수가 정한다

정리 8.1 (최소제곱의 경사하강) $L(\mathbf{w}) = \frac{1}{2n}\|\mathbf{y} - X\mathbf{w}\|^2$, $A = X'X/n$이고 $\operatorname{rank} X = p$라 하자. A의 고유값을 $0 < \mu_{\min} \le \dots \le \mu_{\max}$, $\kappa = \mu_{\max}/\mu_{\min}$이라 하고 학습률 $\eta = 1/\mu_{\max}$로 $\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \nabla L(\mathbf{w}_k)$를 되풀이하면

$$\mathbf{w}_k - \hat{\boldsymbol{\beta}} = (I - \eta A)^k (\mathbf{w}_0 - \hat{\boldsymbol{\beta}}), \qquad \|\mathbf{w}_k - \hat{\boldsymbol{\beta}}\| \le \left(1 - \frac{1}{\kappa}\right)^k \|\mathbf{w}_0 - \hat{\boldsymbol{\beta}}\|.$$

오차를 ϵ배로 줄이는 데 대략 $\kappa \log(1/\epsilon)$번이 든다.

증명. $\nabla L(\mathbf{w}) = A\mathbf{w} - X'\mathbf{y}/n$이고 정규방정식으로 $X'\mathbf{y}/n = A\hat{\boldsymbol{\beta}}$이므로 $\nabla L(\mathbf{w}) = A(\mathbf{w} - \hat{\boldsymbol{\beta}})$이다. 따라서 $\mathbf{w}_{k+1} - \hat{\boldsymbol{\beta}} = (I - \eta A)(\mathbf{w}_k - \hat{\boldsymbol{\beta}})$. 대칭행렬 $I - \eta A$의 고유값은 $1 - \mu_i/\mu_{\max} \in [0,\ 1 - 1/\kappa]$이므로 그 노름은 $1 - 1/\kappa$ 이하다. $(1 - 1/\kappa)^k \le e^{-k/\kappa}$에서 마지막 문장이 나온다. ∎

예제 8.2 (Galton을 경사하강으로) $X = [\mathbf{1},\ \text{부모 키}]$의 κ는 6.8×10^6이다(2절). 원점에서 출발해 상대오차 10^{-3}에 이르려면 4,721만 번의 반복이 든다. 부모 키에서 평균을 빼면 $\kappa = 3.19$이고 19번이면 된다(그림 8.1 왼쪽). 상수 열과 거의 평행한 열이 손실 함수의 등고선을 길쭉한 골짜기로 만들고, 경사하강은 골짜기의 벽 사이를 오가며 바닥을 따라 아주 느리게 내려간다. 입력을 표준화하는 것, 신경망의 배치 정규화(5회 6절의 백색화)는 이 골짜기를 둥글게 만드는 일이다.

8.2 모수가 관측보다 많으면 — 최소노름 해

$p > n$이면 $\operatorname{rank} X < p$이고 정규방정식의 해가 무수히 많다. 그 해들은 모두 자료를 완벽히 맞춘다(잔차 0). 경사하강은 그 가운데 어느 것을 고르나?

정리 8.3 (경사하강은 최소노름 해로 간다) $\mathbf{w}_0 = \mathbf{0}$에서 출발해 $0 < \eta < 2/\mu_{\max}$로 경사하강을 하면 $\mathbf{w}_k$는 정규방정식의 해 가운데 노름이 가장 작은 것 $\mathbf{w}^\dagger$로 수렴한다.

증명. **1. 행공간에 머문다.** $\nabla L(\mathbf{w}) = X'(X\mathbf{w} - \mathbf{y})/n \in \mathrm{col}(X')$이므로 $\mathbf{w}_0 = \mathbf{0}$에서 시작한 $\mathbf{w}_k$는 모두 $\mathrm{col}(X')$에 있다.

2. 행공간의 해는 하나이고 노름이 최소다. 해 하나를 $\mathbf{w}^*$라 하면 해 전체는 $\mathbf{w}^* + \mathrm{null}(X)$이다. $\mathbb{R}^p = \mathrm{col}(X') \oplus \mathrm{null}(X)$ (직교 분해)로 $\mathbf{w}^* = \mathbf{w}^\dagger + \mathbf{w}_n$이라 쓰면 $X\mathbf{w}_n = \mathbf{0}$이므로 $\mathbf{w}^\dagger$도 해이고, 행공간에 있는 해는 이것 하나다(두 해의 차는 $\mathrm{null}(X) \cap \mathrm{col}(X') = \{\mathbf{0}\}$에 있다). 아무 해 $\mathbf{w} = \mathbf{w}^\dagger + \mathbf{v}$ ($\mathbf{v} \in \mathrm{null}(X)$)에 대해 $\|\mathbf{w}\|^2 = \|\mathbf{w}^\dagger\|^2 + \|\mathbf{v}\|^2$이다.

3. 수렴. 정리 8.1의 증명처럼 $\mathbf{w}_{k+1} - \mathbf{w}^\dagger = (I - \eta A)(\mathbf{w}_k - \mathbf{w}^\dagger)$이고 $\mathbf{w}_k - \mathbf{w}^\dagger \in \mathrm{col}(X')$이다. A는 $\mathrm{col}(X')$를 자기 자신으로 보내고, 거기서의 고유값은 양수 $\mu_i \le \mu_{\max}$이므로 $I - \eta A$의 고유값 $1 - \eta\mu_i$는 절댓값이 1보다 작다. 따라서 $\mathbf{w}_k \to \mathbf{w}^\dagger$. ∎

$\mathbf{w}^\dagger = X^+\mathbf{y}$ (X^+는 Moore–Penrose 역행렬)이다. 아무도 노름을 작게 하라고 시키지 않았는데 경사하강이 스스로 그렇게 한다. 이것을 **암묵적 정칙화**라 부른다. 모수가 자료보다 훨씬 많은 오늘의 신경망이 왜 자료를 외우면서도 새 자료에서 망가지지 않는지 설명하려는 연구가 이 정리에서 출발했다.

8.3 Ridge와 weight decay

정의 8.4 (Ridge 추정량) $\lambda > 0$에 대해 $\|\mathbf{y} - X\mathbf{b}\|^2 + \lambda\|\mathbf{b}\|^2$을 최소로 하는

$$\hat{\boldsymbol{\beta}}_\lambda = (X'X + \lambda I)^{-1}X'\mathbf{y}$$

를 **ridge 추정량**이라 한다(Hoerl–Kennard 1970). 상수항은 벌점에서 빼는 것이 보통이어서 설명변수와 반응을 중심화한 뒤에 쓴다.

$X'X + \lambda I$는 언제나 양정치이므로 $p > n$이어도, 공선성이 심해도 해가 하나로 정해진다. 신경망에서 손실에 $\frac{\lambda}{2}\|\mathbf{w}\|^2$을 더하는 것을 **weight decay**라 부른다. 경사하강 한 걸음이 $\mathbf{w} \leftarrow (1 - \eta\lambda)\mathbf{w} - \eta\nabla L(\mathbf{w})$, 곧 매번 가중치를 조금씩 줄이는 꼴이기 때문이다. 손실이 제곱오차이면 그 고정점이 ridge 해다(연습 8.1).

명제 8.5 (ridge는 방향마다 줄인다) $X = UDV'$를 특이값 분해($d_1 \geq \dots \geq d_p > 0$)라 하면

$$\hat{\boldsymbol{\beta}}_\lambda = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} \cdot \frac{\mathbf{u}_j'\mathbf{y}}{d_j}\,\mathbf{v}_j, \qquad \hat{\boldsymbol{\beta}} = \sum_{j=1}^{p} \frac{\mathbf{u}_j'\mathbf{y}}{d_j}\,\mathbf{v}_j.$$

최소제곱 해의 성분을 방향 $\mathbf{v}_j$마다 $d_j^2/(d_j^2+\lambda)$배로 줄인다. 특이값이 작은 방향, 곧 자료가 정보를 적게 가진 방향을 가장 많이 줄인다.

증명. $X'X + \lambda I = V(D^2 + \lambda I)V'$, $X'\mathbf{y} = VDU'\mathbf{y}$이므로 $\hat{\boldsymbol{\beta}}_\lambda = V(D^2+\lambda I)^{-1}DU'\mathbf{y}$이다. 성분으로 쓰면 위의 식이고, $\lambda = 0$이 최소제곱이다. ∎

정리 8.6 (Hoerl–Kennard (1970)) 정의 1.1의 모형에서 $0 < \lambda < \sigma^2 / \max_j \vartheta_j^2$이면 ($\boldsymbol{\vartheta} = V'\boldsymbol{\beta}$)

$$E\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|^2 < E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2.$$

치우친 ridge 추정량이 BLUE를 평균제곱오차에서 이기는 λ가 언제나 있다.

증명. V로 돌려도 노름은 같으므로 $\boldsymbol{\vartheta} = V'\boldsymbol{\beta}$의 좌표에서 계산한다. 최소제곱의 j번째 성분 $\hat{\vartheta}_j = \mathbf{u}_j'\mathbf{y}/d_j$는 평균 ϑ_j, 분산 σ^2/d_j^2이고, ridge의 성분은 $c_j\hat{\vartheta}_j$, $c_j = d_j^2/(d_j^2+\lambda)$이다. 분산과 치우침의 제곱을 더하면

$$M(\lambda) = E\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|^2 = \sum_{j=1}^{p} \frac{\sigma^2 d_j^2 + \lambda^2 \vartheta_j^2}{(d_j^2+\lambda)^2}, \qquad M'(\lambda) = \sum_{j=1}^{p} \frac{2d_j^2(\lambda\vartheta_j^2 - \sigma^2)}{(d_j^2+\lambda)^3}$$

이다. $0 < \lambda < \sigma^2/\max \vartheta_j^2$이면 모든 항이 음수이므로 M은 $[0,\lambda]$에서 감소하고, $M(\lambda) < M(0) = E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$이다. ∎

예제 8.7 (공선성 자료의 ridge) 예제 7.5의 표본(중심화)에서 X의 두 특이값은 13.56과 0.366이다. λ를 0, 1, 10으로 올리면 $(\hat\beta_1, \hat\beta_2)$는 $(4.71, -2.87) \to (1.36, 0.47) \to (0.93, 0.82)$로 참값 $(1,1)$ 쪽으로 모인다(그림 8.1 오른쪽). 줄어드는 것은 거의 전부 작은 특이값의 방향 $\mathbf{v}_2 \approx (1,-1)/\sqrt{2}$, 곧 "두 계수의 차"다. 합의 방향은 거의 건드리지 않는다. 이 설계에서 $\sigma = 1$, $\boldsymbol{\beta} = (1,1)$이면 최소제곱의 평균제곱오차는 7.47이고 ridge는 $\lambda \approx 7$에서 0.010으로, 700배 작다. $\boldsymbol{\beta}$가 거의 전부 큰 특이값의 방향에 있어서 작은 방향을 줄이는 대가(치우침)가 거의 없기 때문이다.

가정 주의 최적의 λ는 모르는 $\boldsymbol{\beta}$에 의존한다(연습 3.5의 c와 같다). 실무에서는 교차검증으로 고른다. 16회에서 Stein은 정규 평균 세 개 이상을 동시에 추정할 때, $\boldsymbol{\beta}$와 무관하게 언제나 최소제곱을 이기는 축소 추정량이 있음을 보인다.

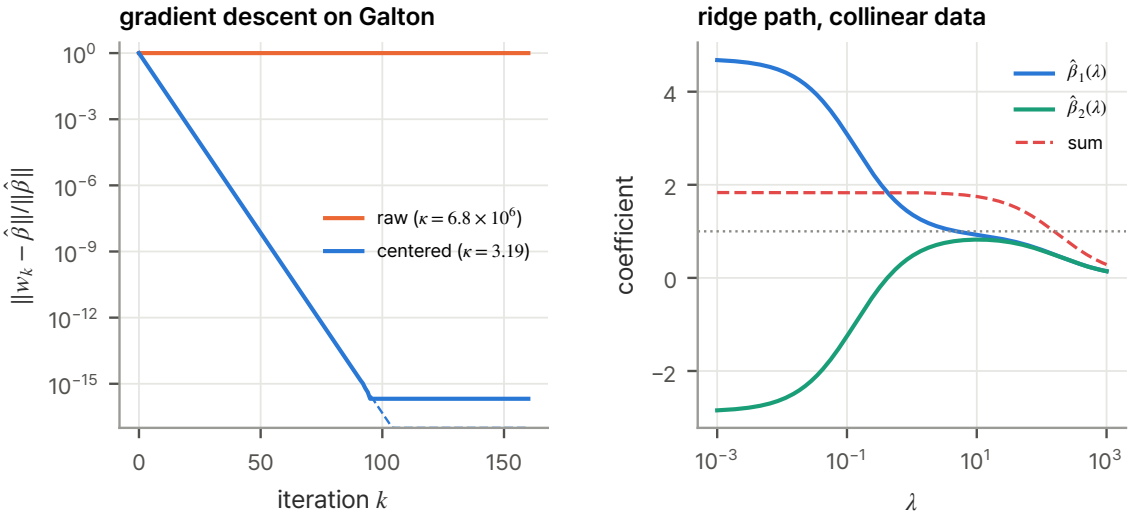


그림 8.1 왼쪽: Galton 자료의 경사하강. 원래의 부모 키($\kappa = 6.8\times10^6$, 주황)로는 오차가 거의 줄지 않고, 중심화하면($\kappa = 3.19$, 파랑) 100번쯤에 기계 정밀도에 이른다. 점선은 정리 8.1의 상계 $(1-1/\kappa)^k$이다. 오른쪽: 공선성 자료의 ridge 경로. λ가 커지면 두 계수가 참값 1(회색 점선) 근처로 모이고, 합(빨강 점선)은 $\lambda \approx 10$까지 거의 그대로다.

연습문제

연습 8.1 (덧붙인 자료 ★) (a) $\tilde X = \begin{pmatrix} X \\ \sqrt{\lambda} I_p \end{pmatrix}$, $\tilde{\mathbf y} = \begin{pmatrix} \mathbf y \\ \mathbf 0 \end{pmatrix}$로 최소제곱을 하면 ridge 해가 나옴을 보여라. (b) weight decay를 붙인 경사하강 $\mathbf w \leftarrow (1-\eta\lambda)\mathbf w - \eta X'(X\mathbf w - \mathbf y)$의 고정점이 $\hat{\boldsymbol\beta}_\lambda$임을 보여라.

풀이 57쪽

연습 8.2 (두 점, 세 모수 ★★) $X = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$, $\mathbf y = (1, 1)'$이다. 정규방정식의 해 전체를 쓰고, 그 가운데 노름이 가장 작은 해가 $(1/3, 1/3, 2/3)$임을 보여라. 원점에서 출발한 경사하강이 이 해로 가는 것을 정리 8.3으로 설명하라.

풀이 57쪽

연습 8.3 (조건수와 중심화 ★★) 단순회귀에서 $A = X'X/n = \begin{pmatrix} 1 & \bar x \\ \bar x & \overline{x^2} \end{pmatrix}$이다. x를 중심화하면 $A = \mathrm{diag}(1,\ s_x^2)$ ($s_x^2 = \frac{1}{n}\sum(x_i - \bar x)^2$)임을 보이고, 중심화하지 않았을 때 $\bar x^2 \gg 1$이면 $\kappa \approx (1 + \bar x^2 + s_x^2)^2 / s_x^2$임을 보여라(고유값의 합과 곱을 쓰라). Galton의 $\bar x = 68.3$, $s_x^2 = 3.19$로 6.8×10^6을 확인하라.

풀이 58쪽

9

정리와 역사 노트

9.1 한 장 요약

정리	내용	조건
정규방정식 (2.2)	$X'X\hat{\boldsymbol{\beta}} = X'\mathbf{y}$, $S(\mathbf{b}) = S(\hat{\boldsymbol{\beta}}) + \|X(\mathbf{b}-\hat{\boldsymbol{\beta}})\|^2$	—
hat 행렬 (2.5)	H는 $\mathrm{col}(X)$ 위로의 사영, $\mathrm{tr}\,H = p$, $\hat{\mathbf{y}} \perp \mathbf{e}$	$\mathrm{rank}\,X = p$
적률 (3.2)	$E\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, $\mathrm{Cov}\,\hat{\boldsymbol{\beta}} = \sigma^2(X'X)^{-1}$, $Es^2 = \sigma^2$	Gauss–Markov 가정
Gauss–Markov (3.3)	선형불편추정량 가운데 최소제곱이 최선(BLUE)	Gauss–Markov 가정
정규 선형모형 (3.6)	최소제곱 = MLE = UMVUE	정규 오차
Aitken (3.7)	$(X'V^{-1}X)^{-1}X'V^{-1}\mathbf{y}$가 BLUE	$\mathrm{Cov}\,\boldsymbol{\varepsilon} = \sigma^2 V$
χ^2 판정 (4.1)	$\mathbf{Z}'A\mathbf{Z} \sim \chi^2(r) \iff A$ 멱등, 계수 r	$\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, I)$
Craig (4.4) · Cochran (4.5)	$AB = 0 \Rightarrow$ 독립, $\sum r_i = n \iff$ 독립인 $\chi^2(r_i)$	$\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, I)$
표본분포 (4.7, 4.8)	$\hat{\boldsymbol{\beta}}$ 정규, $\mathrm{RSS}/\sigma^2 \sim \chi^2(n-p)$, 독립, 계수의 t	정규 오차
일반선형가설 (4.9)	$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p)}$, 가능도비 검정	정규 오차
최선의 선형예측 (5.1)	$\boldsymbol{\beta}^* = (E\mathbf{x}\mathbf{x}')^{-1}E\mathbf{x}Y$	2차 적률
Frisch–Waugh–Lovell (5.3)	편회귀계수 = 잔차끼리의 회귀, $\mathrm{VIF} = 1/(1-R_j^2)$	—
누락변수 (5.6)	$E\tilde{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1 + \Gamma\boldsymbol{\beta}_2$	—
샌드위치 (7.1) · 한 점 빼기 (7.3)	$(X'X)^{-1}X'\Sigma X(X'X)^{-1}$, $e_i/(1-h_{ii})$	—
경사하강 (8.1, 8.3)	속도 $(1-1/\kappa)^k$, 원점에서 출발하면 최소노름 해	—
Hoerl–Kennard (8.6)	평균제곱오차에서 BLUE를 이기는 ridge가 있다	—

최소제곱의 모든 정리는 하나의 그림에서 나왔다. $\mathbf{y}$를 평면 $\mathrm{col}(X)$에 수직으로 내리는 것이다 (그림 2.1). 정규방정식은 "잔차가 평면과 직교한다", Gauss–Markov는 "다른 추정량은 직교하는 잡음을 더한 것", Cochran은 "직교하는 평면들 위의 조각은 독립", F 검정은 "두 평면 위의 발

사이의 거리”다. 14회의 분산분석은 이 그림에 평면을 더 그린다.

9.2 Stigler의 명명 법칙

이름과 발견

- **최소제곱**: 처음 출판한 사람은 Legendre(1805)다. 미국의 Robert Adrain(1808)도 따로 이르렀다. Gauss의 1795년 주장은 그의 편지와 일기로 뒷받침되지만, 출판으로 세상에 알린 것은 Legendre였다(Stigler 1981).
- **Gauss–Markov 정리**: Gauss(1823)의 정리다. Markov(1900)는 그것을 교과서에 옮겼을 뿐인데 Neyman(1934)의 인용으로 이름이 붙었다. Plackett(1949)가 이를 바로잡았다.
- **Cochran의 정리**: Fisher(1925)는 분산분석표에서 특수한 경우를 이미 썼다. Cochran(1934)은 그것을 일반 정리로 증명했다.
- **Craig의 정리**: Craig(1943)의 필요조건 증명에는 틈이 있었다. Sakamoto(1944), Matusita(1949), Ogawa(1949)가 같은 문제를 다뤘고, 그래서 Craig–Sakamoto 정리라고도 부른다.
- **Frisch–Waugh–Lovell**: 편회귀의 생각은 Yule(1907)에 있었다. Frisch와 Waugh(1933)가 시계열의 추세에서, Lovell(1963)이 일반형으로 썼다.
- **Simpson의 역설**: Pearson(1899)과 Yule(1903)이 먼저 보았다. Simpson(1951)은 그 정리를 썼다.
- **Ridge**: 수치해석에서는 Tikhonov(1943)의 정칙화로 먼저 나왔다.
- **회귀**: Galton의 “평범함으로의 회귀”는 유전의 현상을 가리키는 말이었다. Yule이 그것을 최소제곱 적합과 같은 것으로 만든 뒤, 유전과 무관한 모든 최소제곱 적합이 “회귀”라 불리게 되었다.

9.3 7 Pillars 지도

이번 회차는 첫째 기둥 **Aggregation**의 완성이다. 3회의 평균에서 시작한 “관측을 결합하면 정보가 는다”는 생각이, Mayer의 묶음과 Legendre의 최소제곱을 거쳐 Gauss–Markov 정리라는 최적성으로 끝났다. 넷째 기둥 **Regression**도 여기서 칠한다. Galton이 유전에서 본 회귀와 천문학의 최소제곱이 Yule에게서 하나가 되었다. 4절의 F 검정은 셋째 기둥 **Likelihood**가 선형모형에서 정확한 분포를 얻은 모습이다.

9.4 다음 회 — 분산분석과 실험설계 (1885 → 1954)

이번 회차의 설계행렬 X는 주어진 것이었다. 천문학자는 달의 위치를 고를 수 없었다. 14회는 X를 **고르는** 이야기다. Rothamsted의 Fisher는 비료와 품종을 밭에 어떻게 배치할지를 정해 그 효과를 날씨와 땅의 차이에서 떼어 냈다. 수학은 이번 회차와 같다. 처치의 효과는 사영이고, 분산분석표는 Cochran의 정리다. 새로운 것은 무작위화다. 무작위로 배치하면 X와 오차가 상관될 길이 끊기고,

7절의 Yule–Simpson 역설이 설계 단계에서 사라진다.

참고문헌

- Mayer, T. Abhandlung über die Umwälzung des Monds um seine Axe. *Kosmographische Nachrichten und Sammlungen auf das Jahr 1748* (1750). · Boscovich, R. J. De litteraria expeditione per pontificiam ditionem. *Bononiensi Scientiarum et Artium Instituto Atque Academia Commentarii* 4 (1757).
- Legendre, A.-M. *Nouvelles méthodes pour la détermination des orbites des comètes*. Paris, 1805. · Gauss, C. F. *Theoria Motus Corporum Coelestium*. Hamburg, 1809. · *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*. Göttingen, 1823.
- Galton, F. Regression towards mediocrity in hereditary stature. *J. Anthropological Institute* 15 (1886), 246–263. · Pearson, K., Lee, A. On the laws of inheritance in man. *Biometrika* 2 (1903), 357–462.
- Yule, G. U. On the theory of correlation. *JRSS* 60 (1897), 812–854. · An investigation into the causes of changes in pauperism in England. *JRSS* 62 (1899), 249–295. · On the theory of correlation for any number of variables. *Proc. R. Soc. A* 79 (1907), 182–193.
- Frisch, R., Waugh, F. V. Partial time regressions as compared with individual trends. *Econometrica* 1 (1933), 387–401. · Lovell, M. C. Seasonal adjustment of economic time series and multiple regression analysis. *JASA* 58 (1963), 993–1010.
- Cochran, W. G. The distribution of quadratic forms in a normal system. *Proc. Cambridge Phil. Soc.* 30 (1934), 178–191. · Craig, A. T. Note on the independence of certain quadratic forms. *Ann. Math. Statist.* 14 (1943), 195–197. · Driscoll, M. F., Gundberg, W. R. A history of the development of Craig's theorem. *American Statistician* 40 (1986), 65–70.
- Aitken, A. C. On least squares and linear combination of observations. *Proc. R. Soc. Edinburgh* 55 (1935), 42–48. · Plackett, R. L. A historical note on the method of least squares. *Biometrika* 36 (1949), 458–460.
- Anscombe, F. J. Graphs in statistical analysis. *American Statistician* 27 (1973), 17–21. · White, H. A heteroskedasticity-consistent covariance matrix estimator. *Econometrica* 48 (1980), 817–838. · Bickel, P. J., Hammel, E. A., O'Connell, J. W. Sex bias in graduate admissions. *Science* 187 (1975), 398–404.
- Hoerl, A. E., Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12 (1970), 55–67. · Koenker, R., Bassett, G. Regression quantiles. *Econometrica* 46 (1978), 33–50.
- Stigler, S. M. Gauss and the invention of least squares. *Ann. Statist.* 9 (1981), 465–474. · *The History of Statistics*, 1장. Harvard, 1986. · 자료: HistData `Mayer`, `Saturn`, `Galton`, `GaltonFamilies`, `PearsonLee`; datasets `anscombe`, `UCBAdmissions`.

A

연습문제 풀이

각 풀이의 머리에 있는 문제 ○○쪽을 누르면 문제로 돌아간다. 계산 문제의 수치는 노트북에서 같은 계산으로 확인했다.

1절

연습 1.1 문제 8쪽

첫 식에서 둘째 식을 빼면 $14.6391\,\alpha - 2.5375\,\gamma = 22.150$, 셋째 식에서 둘째 식을 빼면 $8.9381\,\alpha + 6.2206\,\gamma = 12.750$이다. 이 둘을 풀면 $\hat{\alpha} = 1.4958$, $\hat{\gamma} = -0.0996$이고, 첫 식에 넣으면 $-9\hat{\beta} = -118.133 - 8.4987\hat{\alpha} + 0.7932\hat{\gamma}$에서 $\hat{\beta} = 14.547$이다. $\hat{\theta} = \arcsin(-0.0996/1.4958) = -3.82^\circ = -3^\circ 49.1'$.

연습 1.2 문제 8쪽

무상관이므로 $\operatorname{Var}\bar{\varepsilon} = \frac{1}{m^2}\sum \operatorname{Var}\varepsilon_i = \sigma^2/m$이고 표준편차는 $\sigma/\sqrt{m}$이다. $m = 9$이면 3분의 1이다. Mayer가 생각한 9분의 1이 아니다.

연습 1.3 문제 8쪽

(a) 정리 2.2에서 $X = \mathbf{1}$이면 $n\hat{\mu} = \sum y_i$. (b) $\sum |y_i - \mu|$의 최소점은 중앙값이다(5회 연습 1.4). (c) 조건 (i)은 $\sum (y_i - b_0) = 0$, 곧 $b_0 = \bar{y}$이고, 미지수가 하나뿐이어서 조건 (ii)가 할 일이 남지 않는다.

연습 1.4 문제 8쪽

$\mathbf{b}$를 고정하면 $u_i - v_i = y_i - \mathbf{x}_i'\mathbf{b}$를 만족하는 $u_i, v_i \ge 0$ 가운데 $u_i + v_i$의 최소는 $|y_i - \mathbf{x}_i'\mathbf{b}|$이고 $u_i = \max(r_i, 0)$, $v_i = \max(-r_i, 0)$일 때 이루어진다. 따라서 두 문제의 최솟값과 최소점의 $\mathbf{b}$가 같다. 최적해에서 $u_i, v_i > 0$이면 둘에서 $\min(u_i, v_i)$를 빼도 제약이 유지되고 목적함수가 작아지므로 모순이다.

연습 1.5 문제 8쪽

E_S를 고른 세 행을 뽑는 3×27 행렬이라 하면 $\mathbf{y}_S = E_S\mathbf{y}$, $X_S = E_S X$이고 $A = X_S^{-1}E_S$에 대해 $AX = X_S^{-1}X_S = I_3$이다. $\operatorname{Cov}\tilde{\boldsymbol{\beta}}_S = \sigma^2 AA' = \sigma^2 X_S^{-1}E_S E_S' X_S^{-\mathsf{T}} = \sigma^2(X_S'X_S)^{-1}$. $\det X_S$가 0에 가까우면 $(X_S'X_S)^{-1}$이 커져 분산이 폭발한다. 2925가지 해의 $\hat\beta$가 $-426°$에서 $104°$까지 흩어진 것이 그 결과다(노트북).

2절

연습 2.1 문제 14쪽

정규방정식은 $nb_0 + b_1\sum x_i = \sum y_i$, $b_0\sum x_i + b_1\sum x_i^2 = \sum x_i y_i$이다. 첫 식에서 $b_0 = \bar y - b_1\bar x$를 둘째 식에 넣으면 $b_1(\sum x_i^2 - n\bar x^2) = \sum x_i y_i - n\bar x\bar y$, 곧 $b_1 S_{xx} = S_{xy}$. $\hat y_i - \bar y = \hat b_1(x_i - \bar x)$이므로 $\mathrm{ESS} = \hat b_1^2 S_{xx} = S_{xy}^2/S_{xx}$이고 $R^2 = S_{xy}^2/(S_{xx}S_{yy}) = r^2$.

연습 2.2 문제 14쪽

$X'X = n$, $X'\mathbf{y} = \sum y_i$이므로 $H = \mathbf{1}\mathbf{1}'/n$, $\hat{\mathbf{y}} = \bar y\mathbf{1}$. $\mathrm{RSS} = \sum(y_i - \bar y)^2$이므로 $\mathrm{RSS}/(n-1) = S^2$이다. $n - p = n - 1$이다.

연습 2.3 문제 14쪽

FWL(정리 5.3)이나 직접 계산으로, 중심화한 설계 $[\mathbf{1},\ x-\bar x]$는 직교하는 두 열이므로 $H = \mathbf{1}\mathbf{1}'/n + (\mathbf{x}-\bar x\mathbf{1})(\mathbf{x}-\bar x\mathbf{1})'/S_{xx}$이다(정리 2.5(d)로 H는 중심화와 무관하다). 대각원소가 $1/n + (x_i-\bar x)^2/S_{xx}$이다. 예에서 $\bar x = 9$, $S_{xx} = 10\cdot 1 + 100 = 110$이므로 $h = 1/11 + 100/110 = 1$이다.

연습 2.4 문제 14쪽

X가 XT, $T = \operatorname{diag}(1,\ 2.54)$로 바뀐다. 정리 2.5(d)로 H, 적합값, 잔차, R^2은 그대로다. $XT\hat{\boldsymbol{\beta}}' = X\hat{\boldsymbol{\beta}}$이어야 하므로 $\hat{\boldsymbol{\beta}}' = T^{-1}\hat{\boldsymbol{\beta}}$, 곧 절편은 같고 기울기는 $\hat b_1/2.54$이다.

연습 2.5 문제 14쪽

$[X, \mathbf{z}]$로 맞춘 최소제곱은 $\|\mathbf{y} - X\mathbf{b} - c\mathbf{z}\|^2$을 $(\mathbf{b}, c)$에 대해 최소로 한다. $c = 0$으로 제한한 최솟값이 원래의 RSS이므로, 제한 없는 최솟값은 그보다 크지 않다. TSS는 그대로이므로 R^2은 줄지 않는다.

연습 2.6 문제 14쪽

만드는 방법에서 $\mathbf{x}^{(k)} = \sum_{j<k}(\mathbf{q}_j'\mathbf{x}^{(k)})\mathbf{q}_j + \|\mathbf{r}_k\|\,\mathbf{q}_k$ ($\mathbf{r}_k$는 빼고 남은 벡터)이다. 일차독립이므로

$\mathbf{r}_k \neq \mathbf{0}$이다. R의 원소를 $r_{jk} = \mathbf{q}_j'\mathbf{x}^{(k)}$ $(j<k)$, $r_{kk} = \|\mathbf{r}_k\| > 0$, $r_{jk} = 0$ $(j>k)$로 두면 $X = QR$이다. $\mathbf{q}_k$들은 귀납법으로 서로 직교하고 길이가 1이다.

연습 2.7 문제 14쪽

(a) 스펙트럼 분해(5회 보조정리 3.2) $P = V\Lambda V'$에서 $P^2 = V\Lambda V'V\Lambda V' = V\Lambda^2 V'$이다. $P^2 = P$이면 $V\Lambda^2 V' = V\Lambda V'$이고 양쪽에 V', V를 곱하면 $\Lambda^2 = \Lambda$, 곧 $\lambda_i^2 = \lambda_i$이고 $\lambda_i \in \{0,1\}$이다. 계수는 0이 아닌 고유값의 수, 곧 1의 개수이고, 그것이 $\operatorname{tr} P = \operatorname{tr}\Lambda = \sum_i \lambda_i$이다. (b) 고유값 1에 해당하는 V의 열 r개를 U로 모으면 $P = V\Lambda V' = UU'$이다. V의 열이 정규직교이므로 $U'U = I_r$이다. $P = UU'$에서 $\operatorname{col}(P) \subseteq \operatorname{col}(U)$이고, $PU = UU'U = U$에서 $\operatorname{col}(U) \subseteq \operatorname{col}(P)$이다. (c) $(I-P)' = I-P$이고 $(I-P)^2 = I - 2P + P^2 = I - P$이다. (a)를 $I-P$에 쓰면 $\operatorname{rank}(I-P) = \operatorname{tr}(I-P) = n - r$이다. (d) $(P\mathbf{v})'(I-P)\mathbf{v} = \mathbf{v}'P(I-P)\mathbf{v} = \mathbf{v}'(P-P^2)\mathbf{v} = 0$이다. $\mathbf{w} \in \operatorname{col}(P)$이면 $P\mathbf{v} - \mathbf{w} \in \operatorname{col}(P)$이고, $\operatorname{col}(P)$의 원소 $P\mathbf{u}$와 $(I-P)\mathbf{v}$의 내적은 $\mathbf{u}'P(I-P)\mathbf{v} = 0$이다. 따라서

$$\|\mathbf{v}-\mathbf{w}\|^2 = \|(I-P)\mathbf{v} + (P\mathbf{v}-\mathbf{w})\|^2 = \|(I-P)\mathbf{v}\|^2 + \|P\mathbf{v}-\mathbf{w}\|^2 \;\geq\; \|\mathbf{v}-P\mathbf{v}\|^2$$

이고 등호는 $\mathbf{w} = P\mathbf{v}$일 때뿐이다.

연습 2.8 문제 14쪽

(a) $X'X$가 대칭이므로 그 역행렬도 대칭이고

$$H' = X\big((X'X)^{-1}\big)'X' = H, \qquad H^2 = X(X'X)^{-1}(X'X)(X'X)^{-1}X' = H$$

이다. $HX = X$이므로 X의 열은 모두 $\operatorname{col}(H)$에 있고, $\operatorname{col}(X) \subseteq \operatorname{col}(H)$이다. 거꾸로 $H\mathbf{v} = X\big((X'X)^{-1}X'\mathbf{v}\big)$이므로 $\operatorname{col}(H) \subseteq \operatorname{col}(X)$이다.

(b) 대각합의 성질 $\operatorname{tr} AB = \operatorname{tr} BA$로

$$\operatorname{tr} H = \operatorname{tr}\big(X\cdot(X'X)^{-1}X'\big) = \operatorname{tr}\big((X'X)^{-1}X'X\big) = \operatorname{tr} I_p = p$$

이고, 보조정리 2.4(a)로 $\operatorname{rank} H = \operatorname{tr} H = p$이다.

(c) $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = X(X'X)^{-1}X'\mathbf{y} = H\mathbf{y}$이고 $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (I-H)\mathbf{y}$이다. 정규방정식이 $X'\mathbf{e} = \mathbf{0}$이고, $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$이므로 $\hat{\mathbf{y}}'\mathbf{e} = \hat{\boldsymbol{\beta}}'X'\mathbf{e} = 0$이다. $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$의 두 조각이 직교하므로 $\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{e}\|^2$이다. $\mathrm{RSS} = \mathbf{e}'\mathbf{e} = \mathbf{y}'(I-H)'(I-H)\mathbf{y} = \mathbf{y}'(I-H)\mathbf{y}$이고, 보조정리 2.4(c)로 $I-H$는 계수 $n-p$의 사영이다. (d) $(XT)'(XT) = T'X'XT$이고 그 역행렬은 $T^{-1}(X'X)^{-1}T^{-\mathsf{T}}$이다. 따라서 $XT\,T^{-1}(X'X)^{-1}T^{-\mathsf{T}}\,T'X' = X(X'X)^{-1}X' = H$이다.

연습 2.9 문제 15쪽

(a) H가 대칭 멱등이므로 $H = H^2 = HH'$이고 (i,i) 원소는 $h_{ii} = \sum_j h_{ij}^2 = h_{ii}^2 + \sum_{j\neq i} h_{ij}^2 \geq h_{ii}^2$이다. $h_{ii} \geq h_{ii}^2$은 $0 \leq h_{ii} \leq 1$과 같다. $\sum_i h_{ii} = \operatorname{tr} H = p$ (정리 2.5(a)). (b) $\mathbf{1} \in \operatorname{col}(X) = \operatorname{col}(H)$

이므로 $H\mathbf{1}=\mathbf{1}$이고 $HJ=H\mathbf{1}\mathbf{1}'/n=J$, $JH=(HJ)'=J$이다. $J^2=\mathbf{1}(\mathbf{1}'\mathbf{1})\mathbf{1}'/n^2=J$이므로 $(H-J)^2=H-HJ-JH+J=H-J$이고 $H-J$는 대칭 멱등이다. (a)와 같은 계산으로 그 대각원소 $h_{ii}-1/n$은 0 이상이다. (c) $(I-H)X=0$이므로 $\mathbf{e}=(I-H)(X\boldsymbol{\beta}+\boldsymbol{\varepsilon})=(I-H)\boldsymbol{\varepsilon}$이다. 5회 보조정리 3.1로 $\operatorname{Cov}\mathbf{e}=(I-H)(\sigma^2 I)(I-H)'=\sigma^2(I-H)$이고 대각원소는 $\sigma^2(1-h_{ii})$이다.

3절

연습 3.1 문제 20쪽

$X'X=\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$의 행렬식은 nS_{xx}이므로 $(X'X)^{-1}=\frac{1}{nS_{xx}}\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$. $\sum x_i^2=S_{xx}+n\bar{x}^2$을 쓰면 세 식이 나온다. 상관은 $-\bar{x}/\sqrt{\bar{x}^2+S_{xx}/n}$이고, $\bar{x}$가 0에서 멀면 -1(또는 $+1$)에 가깝다. 절편은 자료에서 먼 $x=0$에서의 외삽이어서 기울기와 강하게 얽힌다. 중심화하면 상관이 0이 된다.

연습 3.2 문제 20쪽

$E\sum w_i y_i=\mu\sum w_i$이므로 불편 $\iff \sum w_i=1$. $\operatorname{Var}=\sigma^2\sum w_i^2$이고 $\sum w_i^2=\sum(w_i-\frac{1}{n})^2+\frac{2}{n}\sum w_i-\frac{1}{n}=\sum(w_i-\frac{1}{n})^2+\frac{1}{n}\ge\frac{1}{n}$. 등호는 $w_i=1/n$일 때뿐이다.

연습 3.3 문제 20쪽

$\bar{y}_j=\mu+\bar{\varepsilon}_j$, $\operatorname{Var}\bar{\varepsilon}_j=\sigma^2/m_j$이므로 $V=\operatorname{diag}(1/m_1,\dots,1/m_k)$, $X=\mathbf{1}$이다. 따름정리 3.7로 $\hat{\mu}=(\mathbf{1}'V^{-1}\mathbf{1})^{-1}\mathbf{1}'V^{-1}\bar{\mathbf{y}}=\sum m_j\bar{y}_j/\sum m_j$. 분자는 모든 측정의 합이므로 전체 평균이다.

연습 3.4 문제 20쪽

$A=(G'X)^{-1}G'$로 AA'를 계산해 $(X'X)^{-1}$과 대각원소를 비교한다(노트북). 묶음 합은 각 묶음 안에서 x_{i3}의 차이를 버린다. 묶음 I의 x_{i3}은 -0.47에서 0.46까지 퍼져 있지만 합은 -0.79이고, 묶음 II도 -0.91에서 0.94까지 퍼져 있지만 합은 1.74이다. γ에 대한 정보는 묶음 III의 큰 합(7.96)에만 남는다. 최소제곱은 각 방정식의 x_{i3}을 따로 쓰므로 묶음 안의 퍼짐까지 정보로 쓴다. α의 계수는 Mayer가 묶음을 그것으로 나눴으므로 정보가 덜 버려졌다.

연습 3.5 문제 20쪽

$Ec\bar{y}=c\mu$, $\operatorname{Var}c\bar{y}=c^2\sigma^2/n$이므로 평균제곱오차는 $c^2\sigma^2/n+(1-c)^2\mu^2$. c로 미분해 0으로 두면 $c^*=\mu^2/(\mu^2+\sigma^2/n)$이고, 그때의 값은 $\frac{\sigma^2}{n}\cdot\frac{\mu^2}{\mu^2+\sigma^2/n}<\frac{\sigma^2}{n}$이다. c^*가 모르는 μ에 의존하므로 그대로는 쓸 수 없다.

연습 3.6 문제 21쪽

(a) 5회 보조정리 3.2로 양정치 V는 양정치인 대칭 제곱근 $V^{1/2}$를 갖고, $V^{-1/2} = (V^{1/2})^{-1}$이다. $\mathbf{y}^* = V^{-1/2}\mathbf{y} = X^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$, $\boldsymbol{\varepsilon}^* = V^{-1/2}\boldsymbol{\varepsilon}$이고, 5회 보조정리 3.1로 $\operatorname{Cov}\boldsymbol{\varepsilon}^* = V^{-1/2}(\sigma^2 V)V^{-1/2} = \sigma^2 I$이다. $V^{-1/2}$가 가역이므로 $\operatorname{rank} X^* = \operatorname{rank} X = p$이다. (b) $A^* = AV^{1/2}$로 두면 $A\mathbf{y} = AV^{1/2}V^{-1/2}\mathbf{y} = A^*\mathbf{y}^*$이고, $A^*X^* = AV^{1/2}V^{-1/2}X = AX$이다. 따라서 $AX = I \iff A^*X^* = I$이고, $A \mapsto AV^{1/2}$는 일대일 대응이다. 두 모형의 선형불편추정량은 같은 추정량들의 모임이다. (c) 변환된 모형은 정의 1.1의 가정을 만족하므로 정리 3.3로 BLUE는 $(X^{*\mathsf{T}}X^*)^{-1}X^{*\mathsf{T}}\mathbf{y}^* = (X'V^{-1}X)^{-1}X'V^{-1}\mathbf{y}$이고 공분산은 $\sigma^2(X^{*\mathsf{T}}X^*)^{-1} = \sigma^2(X'V^{-1}X)^{-1}$이다. (d) 보통 최소제곱은 $L\mathbf{y}$, $L = (X'X)^{-1}X'$이고 $LX = I$이므로 여전히 불편이다. 5회 보조정리 3.1로 $\operatorname{Cov} L\mathbf{y} = L(\sigma^2 V)L' = \sigma^2(X'X)^{-1}X'VX(X'X)^{-1}$이다. $L\mathbf{y}$는 (b)로 변환된 모형의 선형불편추정량이므로 (c)의 BLUE보다 공분산이 작지 않다(정리 3.3).

4절

연습 4.1 문제 29쪽

$\operatorname{diag}(2,0)$: $2Z_1^2$, 평균 2, 분산 8. 평균이 2인 χ^2은 $\chi^2(2)$이고 분산이 4이므로 아니다. $\operatorname{diag}(1,-1)$: $Z_1^2 - Z_2^2$, 평균 0, 분산 4. 음수 값을 가지므로 χ^2일 수 없다.

연습 4.2 문제 29쪽

$I - J$는 대칭 멱등이므로 $\operatorname{Var}\mathbf{Z}'(I-J)\mathbf{Z} = 2\operatorname{tr}\big((I-J)^2\big) = 2\operatorname{tr}(I-J) = 2(n-1)$. $(n-1)^2$으로 나누면 $2/(n-1)$이다($\sigma = 1$).

연습 4.3 문제 29쪽

$\mathbf{Z} = \Sigma^{-1/2}\mathbf{X}$로 두면 $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, I)$이다. $\tilde{A} = \Sigma^{1/2}A\Sigma^{1/2}$로 쓰면 $\mathbf{X}'A\mathbf{X} = \mathbf{Z}'\tilde{A}\mathbf{Z}$이다. $\tilde{B}$도 같게 두면 $\tilde{A}\tilde{B} = \Sigma^{1/2}A\Sigma B\Sigma^{1/2} = 0$이므로 정리 4.4(a).

연습 4.4 문제 29쪽

$C = \mathbf{e}_j'$이면 $W = [(X'X)^{-1}]_{jj}$이고 $\mathrm{RSS}_0 - \mathrm{RSS} = \hat{\beta}_j^2/W$. $F = \hat{\beta}_j^2/(Ws^2) = \big(\hat{\beta}_j/\mathrm{SE}_j\big)^2 = t_j^2$.

연습 4.5 문제 29쪽

$\mathrm{RSS}_0 = \mathrm{TSS}$이므로 분자는 $\mathrm{ESS}/(p-1) = R^2\mathrm{TSS}/(p-1)$, 분모는 $(1-R^2)\mathrm{TSS}/(n-p)$. Galton: $F = \dfrac{0.21046}{0.78954} \times 926 = 246.8$.

연습 4.6 문제 29쪽

$-2\log\lambda = 27\log(1+F/24) > 3.841$은 $F > 24(e^{3.841/27}-1) = 3.669$와 같다. H_0 아래 $F \sim F(1,24)$이므로 실제 크기는 $P(F(1,24) > 3.669) = 0.067$이다. 정확한 5% 임계값은 4.26이다. $n = 27$에서 Wilks 근사는 1종 오류를 3할 넘게 부풀린다.

5절

연습 5.1 문제 34쪽

$b_{yx} = S_{xy}/S_{xx}$, $b_{xy} = S_{xy}/S_{yy}$이므로 곱은 $S_{xy}^2/(S_{xx}S_{yy}) = r^2$. (x,y) 평면에서 x를 y로 맞춘 직선의 기울기는 $1/b_{xy}$이고 $|1/b_{xy}| = |b_{yx}|/r^2 \geq |b_{yx}|$이므로 더 가파르다. Galton: $0.646 \times 0.326 = 0.210$, $1/0.326 = 3.07$.

연습 5.2 문제 34쪽

5회 따름정리 3.19로 $E[Y \mid X = x] = \mu + \rho(x-\mu)$이므로 $E[Y - X \mid X = x] = -(1-\rho)(x-\mu)$. 대칭이므로 반대 방향도 같다. 두 해의 분산이 같아도 극단은 양쪽 시간 방향으로 모두 평균에 가까워진다. 극단을 고르는 순간 그 해의 우연한 오차까지 함께 고르기 때문이고, 그 오차는 다른 해에 반복되지 않는다. 분산이 줄어든 것이 아니므로 "평범해짐"은 일어나지 않았다.

연습 5.3 문제 34쪽

$\mathrm{Cov}(X, X^2) = EX^3 = 0$이므로 $\beta_1^* = 0$, $\beta_0^* = EX^2 = 1$이다. 최선의 직선은 상수 1이고 설명력은 0이다. 잘못은 모형의 꼴에 있다. $E[Y \mid X] = X^2$은 직선이 아니다. 설계행렬에 x^2 열을 더하면 (모수에 대해서는 여전히 선형모형) 완전히 맞는다.

연습 5.4 문제 34쪽

표본에서도 명제 5.6의 식은 정확한 항등식이다. 짧은 회귀의 기울기 = 긴 회귀의 기울기 $+\hat{\Gamma}\times$(성별 계수)이다. $\hat{\Gamma}$는 성별(아들 1)을 부모 키로 맞춘 기울기로, 노트북에서 -0.0095이다. $0.687 + (-0.0095)(5.215) = 0.637$. 부모 키가 큰 가족에 아들의 비율이 조금 낮았고, 성별을 빼면 그만큼 기울기가 작게 나온다. R^2이 크게 오른 것은 성별이 자녀 키의 변동을 많이 설명하기 때문이다.

연습 5.5 문제 34쪽

(a) 정리 5.3에서 $X_2 = \mathbf{x}^{(j)}$ (한 열)로 두면 $\hat{\beta}_j = (\mathbf{x}^{(j)\mathsf{T}} M \mathbf{x}^{(j)})^{-1} \mathbf{x}^{(j)\mathsf{T}} M \mathbf{y}$이다. M이 대칭 멱등이므로 $\mathbf{x}^{(j)\mathsf{T}} M \mathbf{x}^{(j)} = \|M\mathbf{x}^{(j)}\|^2$이다. (b) $\mathbf{y} = X_1\boldsymbol{\beta}_1 + \mathbf{x}^{(j)}\beta_j + \boldsymbol{\varepsilon}$이고 $MX_1 = 0$이므로 $M\mathbf{y} = M\mathbf{x}^{(j)}\beta_j + M\boldsymbol{\varepsilon}$이다. (a)에 넣으면 $\hat{\beta}_j = \beta_j + \mathbf{x}^{(j)\mathsf{T}} M\boldsymbol{\varepsilon}/\|M\mathbf{x}^{(j)}\|^2$이고, 5회 보조정리 3.1로 $\mathrm{Var}\,\hat{\beta}_j =$

$\sigma^2 \mathbf{x}^{(j)\top} M M \mathbf{x}^{(j)} / \|M\mathbf{x}^{(j)}\|^4 = \sigma^2/\|M\mathbf{x}^{(j)}\|^2$이다. (c) $M\mathbf{x}^{(j)} = (I - H_1)\mathbf{x}^{(j)}$는 x_j를 X_1로 맞춘 잔차이므로 $\|M\mathbf{x}^{(j)}\|^2$은 그 회귀의 RSS다. X_1에 상수항이 있으므로 명제 2.6를 그 회귀에 쓸 수 있고, $\mathrm{TSS} = S_{jj}$이므로 $\mathrm{RSS} = S_{jj}(1 - R_j^2)$이다. 상수항이 없으면 TSS = ESS + RSS가 중심화한 제곱합으로 성립하지 않아 이 식이 맞지 않는다.

6절

연습 6.1 문제 37쪽

$t = 0.6463/0.04114 = 15.71$, $p = 2P(t_{926} > 15.71) = 1.7\times10^{-49}$, 구간 $0.6463 \pm 1.9625\times0.04114 = [0.566,\ 0.727]$. 기울기 1의 t는 $(0.6463 - 1)/0.04114 = -8.60$이다. 자녀는 부모만큼 평균에서 멀어지지 않는다.

연습 6.2 문제 37쪽

$u = \gamma/\alpha$라 하면 $\partial\theta/\partial\alpha = -\gamma/\big(\alpha^2\sqrt{1-u^2}\big)$, $\partial\theta/\partial\gamma = 1/\big(\alpha\sqrt{1-u^2}\big)$. $\hat\alpha = 1.5058$, $\hat\gamma = -0.0719$에서 $\nabla g = (0.0318,\ 0.6649)$ (라디안). 주된 항은 $0.6649 \times 0.0508 = 0.0338$ 라디안 $= 1.94°$이고, 공분산 행렬 전체를 쓰면 $1.95°$이다.

연습 6.3 문제 37쪽

정리 3.6(a)의 $\hat\sigma^2 = \mathrm{RSS}/n$을 로그가능도에 넣으면 $-\frac{n}{2}\log(2\pi\hat\sigma^2) - \frac{n}{2}$이다. $\mathrm{RSS}/n = 5.0003$이므로 $\hat\ell = -464\,(\log 31.418 + 1) = -2063.58$. AIC $= 4127.16 + 4 = 4131.16$, BIC $= 4127.16 + 2\log 928 = 4140.83$. R의 관행이면 AIC $= 4133.16$이다.

7절

연습 7.1 문제 42쪽

$\mathrm{RSS} = \varepsilon'(I - H)\varepsilon$이고 보조정리 3.1로 $E\,\mathrm{RSS} = \mathrm{tr}\big((I - H)\Sigma\big)$이다. $n - p$로 나누고 $[(X'X)^{-1}]_{jj}$를 곱한다. 예제의 설계에서 이것과 명제 7.1의 참 분산의 비가 0.63이다(노트북).

연습 7.2 문제 42쪽

$c = 1 - \mathbf{u}'A^{-1}\mathbf{u}$라 하면 곱은 $I + \frac{\mathbf{u}\mathbf{u}'A^{-1}}{c} - \mathbf{u}\mathbf{u}'A^{-1} - \frac{\mathbf{u}(\mathbf{u}'A^{-1}\mathbf{u})\mathbf{u}'A^{-1}}{c} = I + \mathbf{u}\mathbf{u}'A^{-1}\big(\frac{1-\mathbf{u}'A^{-1}\mathbf{u}}{c} - 1\big) = I$.

연습 7.3 문제 42쪽

$\hat\beta_1$과 $\hat\beta_2$의 공분산은 음수다(두 변수가 양의 상관이면 한쪽 계수가 크게 나올 때 다른 쪽이 작게 나온다). 합의 분산에서 큰 두 분산을 음의 공분산이 거의 상쇄한다. 예에서 $X'X = n\left(\begin{smallmatrix}1 & r\\ r & 1\end{smallmatrix}\right)$, $(X'X)^{-1} = \frac{1}{n(1-r^2)}\left(\begin{smallmatrix}1 & -r\\ -r & 1\end{smallmatrix}\right)$이므로

$$\operatorname{Var}(\hat\beta_1 \pm \hat\beta_2) = \frac{\sigma^2(2 \mp 2r)}{n(1-r^2)} = \frac{2\sigma^2}{n(1 \pm r)}.$$

합은 $r \to 1$에서 σ^2/n으로 작고, 차는 $2\sigma^2/(n(1-r))$로 폭발한다.

연습 7.4 문제 42쪽

전체: 남자 $52/100 = 52\%$, 여자 $33/100 = 33\%$. 학과 표시 Q와 여자 표시 F로 합격을 맞추면 칸의 평균이 $0.60 - 0.40Q + 0.05F$로 정확히 맞으므로 긴 회귀의 여자 계수는 $+0.05$, Q의 계수는 -0.40이다. Γ는 Q를 F로 맞춘 기울기로, 여자 가운데 Q 지원 비율 0.8에서 남자의 0.2를 뺀 0.6이다. 짧은 회귀의 계수는 $0.05 + 0.6 \times (-0.40) = -0.19 = 0.33 - 0.52$이다.

연습 7.5 문제 42쪽

(a) 관측 i를 빼면 $X'X = \sum_k \mathbf{x}_k\mathbf{x}_k'$와 $X'\mathbf{y} = \sum_k \mathbf{x}_k y_k$에서 $k = i$인 항이 빠지므로 $A - \mathbf{x}_i\mathbf{x}_i'$와 $X'\mathbf{y} - \mathbf{x}_i y_i$이다. (b) $h_{ii} = \mathbf{x}_i'A^{-1}\mathbf{x}_i < 1$이므로 연습 7.2의 공식을 쓸 수 있다.

$$\hat{\boldsymbol\beta}_{(i)} = \left(A^{-1} + \frac{A^{-1}\mathbf{x}_i\mathbf{x}_i'A^{-1}}{1-h_{ii}}\right)(X'\mathbf{y} - \mathbf{x}_i y_i).$$

(c) $A^{-1}X'\mathbf{y} = \hat{\boldsymbol\beta}$, $\mathbf{x}_i'A^{-1}X'\mathbf{y} = \hat y_i$, $\mathbf{x}_i'A^{-1}\mathbf{x}_i = h_{ii}$을 써서 전개하면

$$\hat{\boldsymbol\beta}_{(i)} = \hat{\boldsymbol\beta} - A^{-1}\mathbf{x}_i y_i + \frac{A^{-1}\mathbf{x}_i(\hat y_i - h_{ii}y_i)}{1-h_{ii}} = \hat{\boldsymbol\beta} + \frac{A^{-1}\mathbf{x}_i\big(\hat y_i - h_{ii}y_i - (1-h_{ii})y_i\big)}{1-h_{ii}} = \hat{\boldsymbol\beta} - \frac{A^{-1}\mathbf{x}_i e_i}{1-h_{ii}}.$$

(d) (c)에 $\mathbf{x}_i'$를 곱하면 $\mathbf{x}_i'\hat{\boldsymbol\beta}_{(i)} = \hat y_i - h_{ii}e_i/(1-h_{ii})$이므로 $y_i - \mathbf{x}_i'\hat{\boldsymbol\beta}_{(i)} = e_i + h_{ii}e_i/(1-h_{ii}) = e_i/(1-h_{ii})$이다.

8절

연습 8.1 문제 46쪽

(a) $\tilde X'\tilde X = X'X + \lambda I$, $\tilde X'\tilde{\mathbf{y}} = X'\mathbf{y}$이므로 정규방정식이 ridge의 식이다. (b) 고정점에서 $\mathbf{w} = (1-\eta\lambda)\mathbf{w} - \eta X'(X\mathbf{w} - \mathbf{y})$이므로 $\lambda\mathbf{w} = -X'(X\mathbf{w} - \mathbf{y})$, 곧 $(X'X + \lambda I)\mathbf{w} = X'\mathbf{y}$.

연습 8.2 문제 46쪽

$\operatorname{rank} X = 2 = n$이므로 정규방정식의 해는 $X\mathbf{w} = \mathbf{y}$의 해, 곧 $w_1 + w_3 = 1$, $w_2 + w_3 = 1$이다. 해 전체는 $(1-t, 1-t, t)$이고 노름의 제곱 $2(1-t)^2 + t^2$은 $t = 2/3$에서 최소다. 해 $(1/3, 1/3, 2/3) =$

$\frac{1}{3}(1,0,1)+\frac{1}{3}(0,1,1)$은 X의 행공간에 있다. 원점에서 시작한 경사하강은 행공간을 떠나지 않으므로(정리 8.3의 1단계) 그 안의 유일한 해로 간다.

연습 8.3 문제 46쪽

중심화하면 $\bar{x}=0$이므로 $A=\mathrm{diag}(1,s_x^2)$. 중심화하지 않으면 A의 대각합은 $1+\bar{x}^2+s_x^2$, 행렬식은 $\overline{x^2}-\bar{x}^2=s_x^2$이다. $\bar{x}^2\gg 1$이면 큰 고유값은 대각합에 가깝고 작은 고유값은 행렬식/대각합에 가까우므로 $\kappa\approx(1+\bar{x}^2+s_x^2)^2/s_x^2$. Galton: $(1+4666+3.19)^2/3.19=6.8\times10^6$. 중심화하면 $\kappa=3.19$.