


07. 제임스-스타인 추정과 리지 회귀

Created By	 성우 천
Last Edited	@Jul 17, 2019 4:35 PM
Tags	CASI

들어가며

피셔의 MLE는 논란의 여지는 있지만, 여전히 주요한 기법으로 사용된다.

MLE의 장점

- 거의 불편항이다.
- 거의 최저분산이다.
- 자동으로 작동한다.

But, 동시에 수백~수천의 모수추정에서 비변항이란 거의 달성할 수 없는 목표이다.

→ 단 몇개의 모수에서 제임스-스타인 추정은 이를 극적으로 보여준다.

→ 이는 축소추정의 시작으로 '전체 성능 향상'을 위해 의도적인 편향을 첨가한다는 아이디어!

7.1 제임스-스타인 추정기

Assumption) 베이즈 상황에서 관측치 x 로부터 단일모수 μ 를 추정한다고 가정하자.

$$\mu \sim \mathcal{N}(M, A) \quad \text{and} \quad x|\mu \sim \mathcal{N}(\mu, 1), \quad (7.1)$$

이 경우에, μ 는 다음과 같은 사후분포를 갖는다.

$$\mu|x \sim \mathcal{N}(M + B(x - M), B) \quad [B = A/(A + 1)] \quad (7.2)$$

위 식은 (5.21) 식에서 유도된 식이다.

$$\mu|x \sim \mathcal{N}\left(M + \frac{A}{A + \sigma^2}(x - M), \frac{A\sigma^2}{A + \sigma^2}\right). \quad (5.21)$$

- 편의상의 가정

$$\sigma^2 = 1, \frac{A}{(A + 1)} = B$$

이때, μ 에 대한 베이즈 추정은 다음과 같다.

$$\hat{\mu}^{\text{Bayes}} = M + B(x - M), \quad (7.3)$$

또한, 다음의 제곱오차 기댓값을 갖는다.

$$E \left\{ (\hat{\mu}^{\text{Bayes}} - \mu)^2 \right\} = B, \quad (7.4)$$

이것을 MLE μ 의 제곱오차 기댓값과 비교해보자.

$$E \left\{ (\hat{\mu}^{\text{MLE}} - \mu)^2 \right\} = 1. \quad (7.5)$$

→ (7.1)에서 $A = 1$ 이라면, $B = 1/2$ 이다. 즉, Bayes μ 는 MLE μ 에 비해 절반의 위험을 갖는다.

다차원에서의 확장

(7.1)이 N개의 독립적인 경우로 존재하는 상황에서는 어떨까?

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_N)' \quad \text{and} \quad \mathbf{x} = (x_1, x_2, \dots, x_N)', \quad (7.6)$$

with

$$\mu_i \sim \mathcal{N}(M, A) \quad \text{and} \quad x_i | \mu_i \sim \mathcal{N}(\mu_i, 1), \quad (7.7)$$

이때도 동일한 계산이 적용된다.

즉, Bayes μ 의 리스크는 MLE μ 가 갖는 리스크의 단지 B배이다.

만약 M과 A(혹은 B)를 알고있다면 이러한 추정도 괜찮다.

M과 A를 모른다면?

이 값들을 주어진 데이터를 통하여 추정하는 것을 시도해 볼 수 있다.

한계적으로 (7.7)은 다음을 생성한다.

$$x_i \stackrel{\text{ind}}{\sim} \mathcal{N}(M, A + 1). \quad (7.11)$$

그렇다면, $\hat{M} = \bar{x}$ 는 M에 대한 불편추정이다.

또한, B의 불편추정량은 다음과 같다. (N>3일때)

$$\hat{B} = 1 - (N - 3)/S \quad \left[S = \sum_{i=1}^N (x_i - \bar{x})^2 \right] \quad (7.12)$$

제임스-스타인 추정은 7.3 식의 플러그인 버전이다.

$$\hat{\mu}_i^{\text{JS}} = \hat{M} + \hat{B} (x_i - \hat{M}) \quad \text{for } i = 1, 2, \dots, N, \quad (7.13)$$

- '경험적 베이즈'라는 용어가 특히 적절하게 보인다. 우리는 베이즈 모델에서 베이즈 추정을 이끌어내고, 이는 모든 데이터 x 로부터 경험적으로 추정한 것이기 때문이다.

이때의 JS μ 의 리스크를 계산해보자.

$$E \left\{ \left\| \hat{\mu}^{\text{JS}} - \mu \right\|^2 \right\} = NB + 3(1 - B). \quad (7.14)$$

→ $N = 20$ 이라고 $A = 10$ 이라면?

- $\text{Lisk}(\text{J-S}) = 11.5$
- $\text{Lisk}(\text{Bayes}) = 10$
- $\text{Lisk}(\text{MLE}) = 20$

MLE의 옹호론자들은 이 부분에서 놀랍지 않다고 반응할 것.

why? 베이즈 모델(7.7)은 모수가 중심점 M 주변에 군집화되지만, MLE에는 그러한 가정이 없기 때문에 그 정도로 잘 작동하지 않을 것이라는 주장!

→ 제임스-스타인 정리는 베이즈 가정이 없다고 해서, MLE가 더 나은 것은 아니라는 사실을 증명.

James–Stein Theorem *Suppose that*

$$x_i | \mu_i \sim \mathcal{N}(\mu_i, 1) \quad (7.15)$$

independently for $i = 1, 2, \dots, N$, with $N \geq 4$. Then

$$E \left\{ \left\| \hat{\mu}^{\text{JS}} - \mu \right\|^2 \right\} < N = E \left\{ \left\| \hat{\mu}^{\text{MLE}} - \mu \right\|^2 \right\} \quad (7.16)$$

이때, 좌변은 다음과 같다.

$$NB + 3(1 - B) < N$$

$$N(B - 1) < 3(B - 1)$$

N 은 4이상의 가정으로, 식 (7.16)은 참이된다.

7.2 야구 선수들

제임스-스타인의 정리는 JS mu가 MLE mu 보다 얼마나 우세한지에 대해서는 알려주지 않는다.

그러나 적절한 환경에서는 그 개선이 상당하다.

야구선수 18명에 대한 최초 90회 타격의 자료(MLE)로, 평균 타율(TRUTH)을 추정해보자.

Player	MLE	JS	TRUTH	x
1	.345	.283	.298	11.96
2	.333	.279	.346	11.74
3	.322	.276	.222	11.51
4	.311	.272	.276	11.29
5	.289	.265	.263	10.83
6	.289	.264	.273	10.83
7	.278	.261	.303	10.60
8	.255	.253	.270	10.13
9	.244	.249	.230	9.88
10	.233	.245	.264	9.64
11	.233	.245	.264	9.64
12	.222	.242	.210	9.40
13	.222	.241	.256	9.39
14	.222	.241	.269	9.39
15	.211	.238	.316	9.14
16	.211	.238	.226	9.14
17	.200	.234	.285	8.88
18	.145	.212	.200	7.50

테이블의 정보는 다음과 같다.

- MLE 열은 MLE로 추정한 (평균) 타율이다.
- JS 열은 J-S로 추정한 타율이다.
- TRUTH 열은 실제 이후 370타석의 타율이다.
- x열은 아크사인 변환으로 얻은 x값들이다.

이 시점에서 J-S추정을 위한 2가지의 방법을 확인해보자.

1. 정규분포 근사

2. 아크사인 변환

여기서는 아크사인 변환을 시행하여 J-S 추정을 사용해보도록 하자. ($n = 90$)

→ 아크사인 변환 : 확률값(혹은 비율)값을 정규화 시키며, 분산을 안정화 시킨다.

$$x_i = 2(n + 0.5)^{1/2} \sin^{-1} \left[\left(\frac{np_i + 0.375}{n + 0.75} \right)^{1/2} \right], \quad (7.21)$$

해당 방법을 통하여 각 JS mu hat을 얻고, 이를 다시 이항크기로 반환한다.

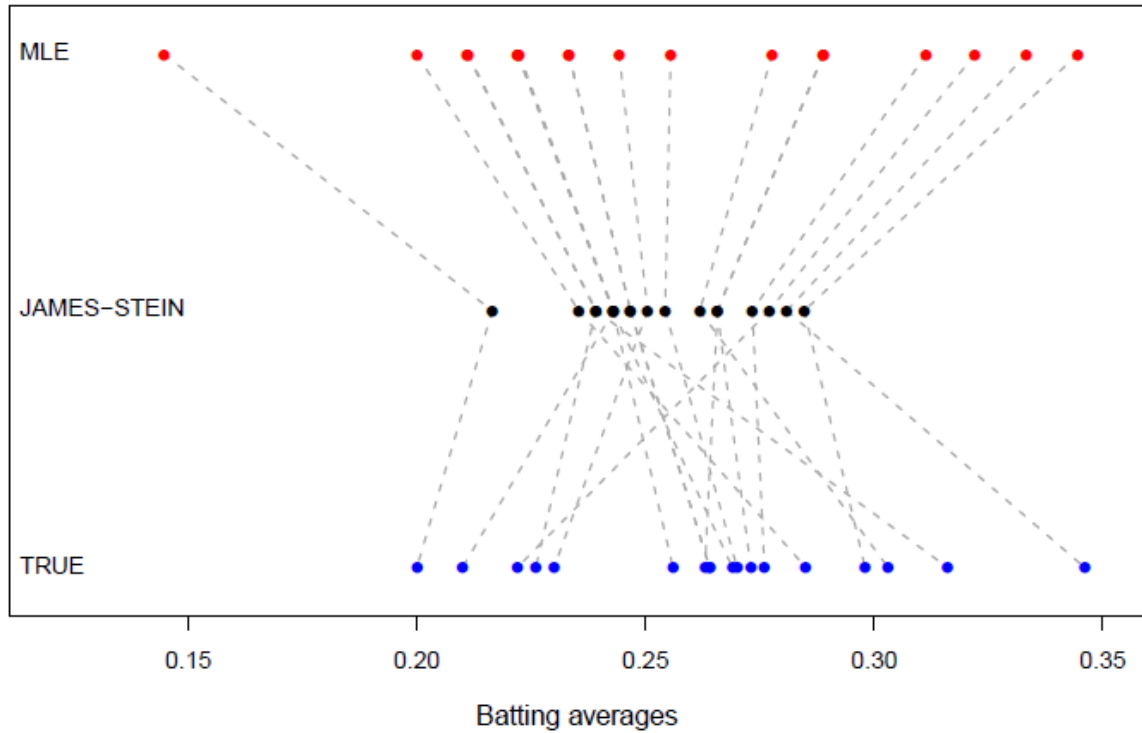
$$\hat{p}_i^{\text{JS}} = \frac{1}{n} \left[\frac{n + 0.75}{n + 0.5} \left(\frac{\sin \hat{\mu}_i^{\text{JS}}}{2} \right)^2 - 0.375 \right]. \quad (7.23)$$

이렇게 해서 얻어진 MLE, JS 추정값의 리스크를 비교해보면

$$\sum_{i=1}^{18} (\text{MLE}_i - \text{TRUTH}_i)^2 = 0.0425 \quad \text{while} \quad \sum_{i=1}^{18} (\text{JS}_i - \text{TRUTH}_i)^2 = 0.0218. \quad (7.24)$$

→ J-S 추정은 전체 예측 제공 오차를 50% 가까이 줄이는 모습을 보인다.

그렇다면, 어떻게 추정이 되었는지 확인해보자.



이렇게 제임스-스타인 법칙은 축소추정을 설명해준다.

그렇다면 왜 이런 축소가 작동하게 되는 것일까?

이는 J-S의 기저인 베이즈 모델에서 초래된다.

(7.8) 모형에서 확인해보며, $M = 0$ 으로 설정해 계산의 편의를 더하도록 하자.

$$E \left\{ \sum_{i=1}^N x_i^2 \right\} = N(A + 1) \quad \text{compared with} \quad E \left\{ \sum_{i=1}^N \mu_i^2 \right\} = NA. \quad (7.25)$$

좌측은 MLE의 제곱합이며, 우측은 이론상의 제곱합을 계산한 것이다.

식을 확인해보면 MLE의 제곱합은 이론상 제곱합을 N 만큼 초과하는 과산포의 성질을 갖는다.

그러나 J-S의 경우 베이즈모형에 기대어, 다음과 같은 제곱합을 갖게된다.

$$E \left\{ \sum_{i=1}^N \left(\hat{\mu}_i^{\text{Bayes}} \right)^2 \right\} = NB^2(A + 1) = NA \frac{A}{A + 1}, \quad (7.26)$$

→ 이는 기존의 NA값을 $A/(A+1)$ 만큼 과축소 시키게 된다.

7.3 리지 회귀

선형 회귀모형에서 β 의 추정은 MLE에 기초하고 있다.

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

이때, X 는 알려진 $n \times p$ 구조의 행렬이고, β 는 미지의 p 차원 모수 벡터이다.

우리는, X 와 y 를 통하여 β 를 추정하는 것이 목적이며, 이 과정에서 최소제곱법을 사용한다.

$$\hat{\beta} = \arg \min_{\beta} \{\|\mathbf{y} - \mathbf{X}\beta\|^2\}. \quad (7.31)$$

축소 추정의 일종인 리지

선형 모형의 커다란 장점은 n 이 아무리 크더라도 미지의 모수 개수를 p 개로 줄여준다는 것이다.

그러나, 요즘의 응용에서 p 의 증가에 따라서 고차원 불편 추정의 한계에 직면하게 되었다.

이때, β 의 추정을 개선하기 위한 리지회귀는 축소 기법의 일종이다.

베타를 추정하는 과정에서 penalty항을 부여하여, 축소를 진행한다.

$$\hat{\beta}(\lambda) = (S + \lambda I)^{-1} X' y = (S + \lambda I)^{-1} S \hat{\beta} \quad (7.36)$$

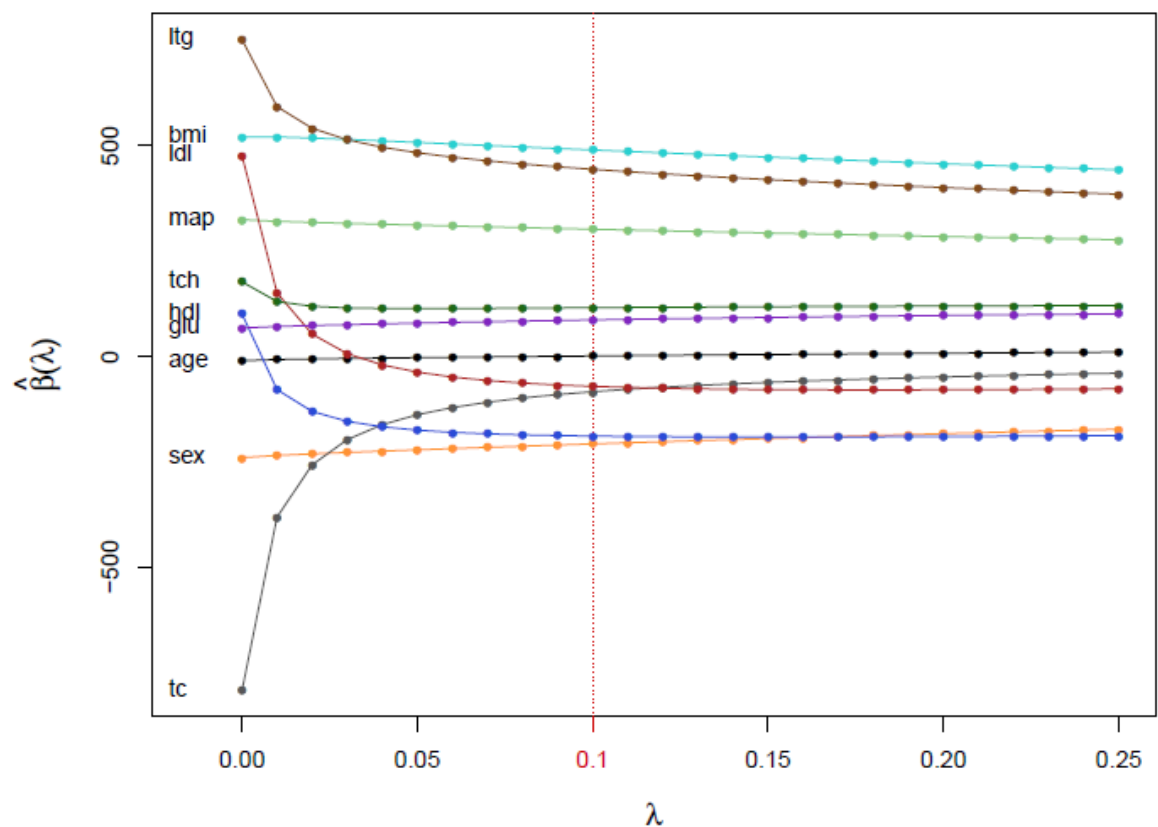
당뇨병 데이터를 통한 이해

age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu	prog
59	1	32.1	101	157	93.2	38	4	2.11	87	151
48	0	21.6	87	183	103.2	70	3	1.69	69	75
72	1	30.5	93	156	93.6	41	4	2.03	85	141
24	0	25.3	84	198	131.4	40	5	2.12	89	206
50	0	23.0	101	192	125.4	52	4	1.86	80	135
23	0	22.6	89	139	64.8	61	2	1.82	68	97
36	1	22.0	90	160	99.6	50	3	1.72	82	138
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

우리는 열가지 척도를 통해 1년후의 질병 경과 수치인 'prog'값을 예측하고 싶다.

이때, 최소 자승 추정을 활용한 beta와 리지 회귀 추정을 통한 beta를 비교하고 싶다.

리지의 모수 lambda의 변화에 따른 추정된 계수의 흐름을 그래프를 통해 확인하자.



이때, $\lambda = 0$ 인 지점은 OLS를 통한 beta의 값을 의미한다.

lambda가 증가함에 따라, beta의 추정값이 감소하는 것을 확인할 수 있다.

표를 통하여 OLS와 리지의 차이점을 다시 확인해보자.

	$\hat{\beta}(0)$	$\hat{\beta}(0.1)$	sd(0)	sd(0.1)
age	-10.0	1.3	59.7	52.7
sex	-239.8	-207.2	61.2	53.2
bmi	519.8	489.7	66.5	56.3
map	324.4	301.8	65.3	55.7
tc	-792.2	-83.5	416.2	43.6
ldl	476.7	-70.8	338.6	52.4
hdl	101.0	-188.7	212.3	58.4
tch	177.1	115.7	161.3	70.8
ltg	751.3	443.8	171.7	58.4
glu	67.6	86.7	65.9	56.6

OLS와 리지로 추정된 beta의 추정값의 표준편차를 비교할 수 있다.

위의 표를 확인해 봤을 때, 리지는 추정된 회귀 계수의 변동성을 상당히 줄여줌을 알 수 있다.

리지의 의미

이는 OLS로 찾은 μ hat이 리지를 통해 찾은 μ hat보다 정확하다는 의미는 아니다.

그러나 관심사가 beta에 집중된다면, 리지는 상당히 중요해진다.

이에 대한 동기를 부여하는 또다른 방법이 있다.

$$\hat{\beta}(\lambda) = \arg \min_{\beta} \{\|y - X\beta\|^2 + \lambda \|\beta\|^2\}. \quad (7.41)$$

괄호 속 항을 beta에 대해 미분하면, 식 (7.36)과 같은 결과를 확인할 수 있다.

또다른 축소 기법 LASSO

리지와 대응하는 또다른 축소기법은, l1 norm을 사용한 LASSO 추정이 있다.

$$\tilde{\beta}(\lambda) = \arg \min_{\beta} \{\|y - X\beta\|^2 + \lambda \|\beta\|_1\}, \quad (7.42)$$

J-S로 추정하기

정규 모델 (3.37)에 J-S법칙을 활용하면, 또다른 J-S beta hat을 얻을 수 있다.

$$\tilde{\beta}^{JS} = \left[1 - \frac{(p-2)\sigma^2}{\hat{\beta}' S \hat{\beta}} \right] \hat{\beta}. \quad (7.43)$$

주석 7

J-S로 찾은 mu hat을 모수에 대한 추정기로 설정하면, J-S 정리는 다음을 보장한다.

$$E \left\{ \|\tilde{\mu}^{JS} - \mu\|^2 \right\} < p\sigma^2 \quad (7.44)$$

7.4 간접 증거 2

축소 추정의 단점을 이전에 소개한 야구 데이터를 통하여 확인해보자.

Player	TRUTH	rmsMLE	rmsJS	rmsJS1
1	.298	.046	.033	.032
2	.346*	.049	.077	.056
3	.222	.044	.042	.038
4	.276	.048	.015	.023
5	.263	.047	.011	.020
6	.273	.046	.014	.021
7	.303	.047	.037	.035
8	.270	.049	.012	.022
9	.230	.044	.034	.033
10	.264	.047	.011	.021
11	.264	.047	.012	.020
12	.210*	.043	.053	.044
13	.256	.045	.014	.020
14	.269	.048	.012	.021
15	.316*	.048	.049	.043
16	.226	.045	.038	.036
17	.285	.046	.022	.026
18	.200*	.043	.062	.048

2~4열은 각각 MLE, J-S, J-S(축소를 제한한 추정) 을 통해 구한 추정값으로 구한 RMSE를 나타낸다.

전체 제곱 오차 합의 관점에서는 J-S추정이 MLE보다 우수하다는 것을 이전에 확인했다.

그러나, 각 선수들에 대한 RMSE를 확인해보면, 4명의 선수에서 MLE가 더 우세한 모습을 보인다.

이러한 현상을 보인 선수는 최댓값 두 명과 최솟값 두 명이다.

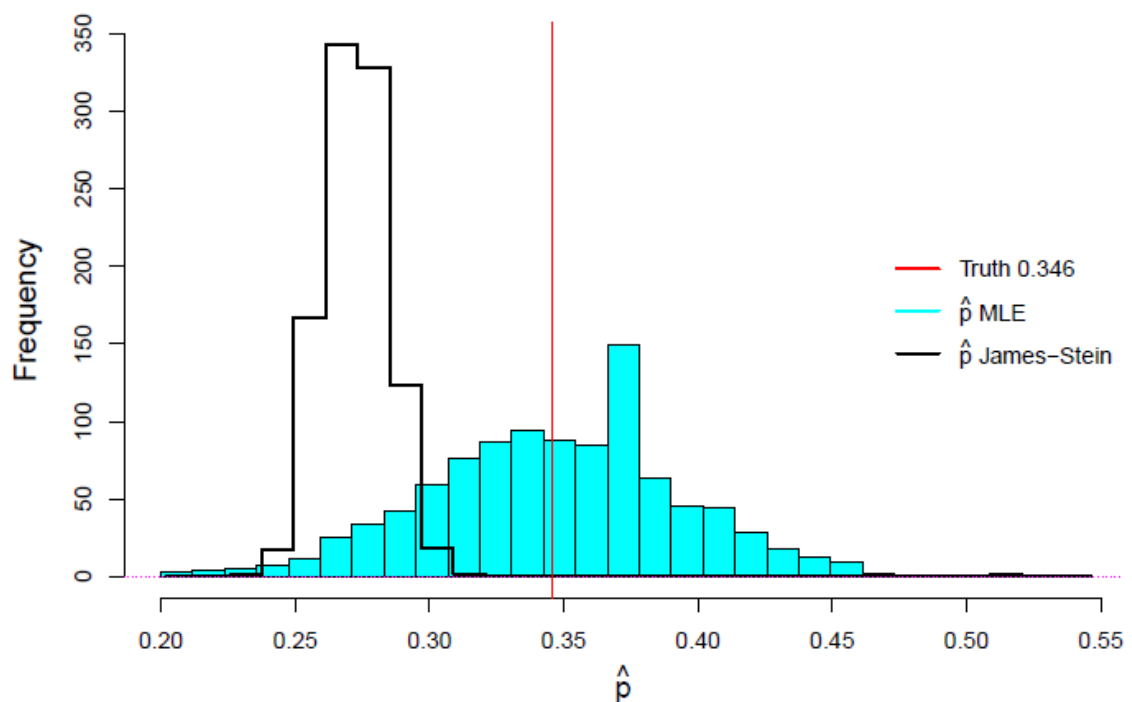
→ 즉 축소 추정은 양 끝값에서는 올바르게 잘 작동하지 않는 모습을 보인다.

뛰어난 선수의 문제(클레멘테의 문제)

표에서 선수 2는 클레멘테라는 선수를 의미하며, 이 선수는 직전 몇년간 타율 상위권을 유지했다.

따라서, 야구 팬들은 이 선수를 축소 집단에 포함시키지 않았을 것이다.

다음은, 클레멘테의 타율에 대한 TRUTH값과 MLE / J-S로 추정된 값의 분포를 보여 준다.



뛰어난 선수에 대하여 지나친 축소를 한 모습을 보인다.

절충안

이러한 현상을 절충하기 위한 방법이 4열에 존재한다.

J-S의 변형 버전으로, 축소가 각 \hat{p} 으로부터 한 단위 σ_0 이상 벗어나지 못하도록 한다.

$$\hat{p}_i^{JS1} = \min \{ \max (\hat{p}_i^{JS}, \hat{p}_i - \sigma_0), \hat{p}_i + \sigma_0 \}. \quad (7.47)$$

이 방법으로, 뛰어난 선수들의 문제를 완화시키는 축소 추정을 할 수 있다.