# 7

# James–Stein Estimation and Ridge Regression

If Fisher had lived in the era of "apps," maximum likelihood estimation might have made him a billionaire. Arguably the twentieth century's most influential piece of applied mathematics, maximum likelihood continues to be a prime method of choice in the statistician's toolkit. Roughly speaking, maximum likelihood provides nearly unbiased estimates of nearly minimum variance, and does so in an automatic way.

That being said, maximum likelihood estimation has shown itself to be an inadequate and dangerous tool in many twenty-first-century applications. Again speaking roughly, unbiasedness can be an unaffordable luxury when there are hundreds or thousands of parameters to estimate at the same time.

The James–Stein estimator made this point dramatically in 1961, and made it in the context of just a few unknown parameters, not hundreds or thousands. It begins the story of *shrinkage estimation*, in which deliberate biases are introduced to improve overall performance, at a possible danger to individual estimates. Chapters 7 and 21 will carry on the story in its modern implementations.

## 7.1 The James–Stein Estimator

Suppose we wish to estimate a single parameter $\mu$ from observation $x$ in the Bayesian situation

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

in which case $\mu$ has posterior distribution

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

as given in (5.21) (where we take $\sigma^2 = 1$ for convenience). The Bayes estimator of $\mu$,

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

has expected squared error

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

compared with 1 for the MLE $\hat{\mu}^{\text{MLE}}=x$,

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

If, say, $A=1$ in (7.1) then $B=1/2$ and $\hat{\mu}^{\text{Bayes}}$ has only half the risk of the MLE.

The same calculation applies to a situation where we have $N$ independent versions of (7.1), say

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

with

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

independently for $i=1,2,\ldots,N$. (Notice that the $\mu_i$ differ from each other, and that this situation is not the same as (5.22)–(5.23).) Let $\hat{\boldsymbol{\mu}}^{\text{Bayes}}$ indicate the vector of individual Bayes estimates $\hat{\mu}_i^{\text{Bayes}}=M+B(x_i-M)$,

$$\hat{\boldsymbol{\mu}}^{\text{Bayes}}=\boldsymbol{M}+B(\boldsymbol{x}-\boldsymbol{M}), \qquad \left[\boldsymbol{M}=(M,M,\ldots,M)'\right], \tag{7.8}$$

and similarly

$$\hat{\boldsymbol{\mu}}^{\text{MLE}}=\boldsymbol{x}.$$

Using (7.4) the total squared error risk of $\hat{\boldsymbol{\mu}}^{\text{Bayes}}$ is

$$E\left\{\left\|\hat{\boldsymbol{\mu}}^{\text{Bayes}}-\boldsymbol{\mu}\right\|^2\right\}=E\left\{\sum_{i=1}^{N}\left(\hat{\mu}_i^{\text{Bayes}}-\mu_i\right)^2\right\}=N\cdot B \tag{7.9}$$

compared with

$$E\left\{\left\|\hat{\boldsymbol{\mu}}^{\text{MLE}}-\boldsymbol{\mu}\right\|^2\right\}=N. \tag{7.10}$$

Again, $\hat{\boldsymbol{\mu}}^{\text{Bayes}}$ has only $B$ times the risk of $\hat{\boldsymbol{\mu}}^{\text{MLE}}$.

This is fine if we know $M$ and $A$ (or equivalently $M$ and $B$) in (7.1). If not, we might try to estimate them from $\boldsymbol{x}=(x_1,x_2,\ldots,x_N)$. Marginally, (7.7) gives

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

Then $\hat{M}=\bar{x}$ is an unbiased estimate of $M$. Moreover,

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

unbiasedly estimates $B$, as long as $N > 3$.† The James–Stein estimator is †1
the plug-in version of (7.3),

$$\hat{\mu}_i^{\text{JS}} = \hat{M} + \hat{B}\left(x_i - \hat{M}\right) \qquad \text{for } i = 1, 2, \ldots, N, \tag{7.13}$$

or equivalently $\hat{\boldsymbol{\mu}}^{\text{JS}} = \hat{\boldsymbol{M}} + \hat{B}(\boldsymbol{x} - \hat{\boldsymbol{M}})$, with $\hat{\boldsymbol{M}} = (\hat{M}, \hat{M}, \ldots, \hat{M})'$.

At this point the terminology "empirical Bayes" seems especially apt: Bayesian model (7.7) leads to the Bayes estimator (7.8), which itself is estimated empirically (i.e., frequentistically) from all the data $\boldsymbol{x}$, and then applied to the individual cases. Of course $\hat{\boldsymbol{\mu}}^{\text{JS}}$ cannot perform as well as the actual Bayes' rule $\hat{\boldsymbol{\mu}}^{\text{Bayes}}$, but the increased risk is surprisingly modest. The expected squared risk of $\hat{\boldsymbol{\mu}}^{\text{JS}}$ under model (7.7) is† †2

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

If, say, $N = 20$ and $A = 1$, then (7.14) equals 11.5, compared with true Bayes risk 10 from (7.9), much less than risk 20 for $\hat{\boldsymbol{\mu}}^{\text{MLE}}$.

A defender of maximum likelihood might respond that none of this is surprising: Bayesian model (7.7) specifies the parameters $\mu_i$ to be clustered more or less closely around a central point $M$, while $\hat{\boldsymbol{\mu}}^{\text{MLE}}$ makes no such assumption, and cannot be expected to perform as well. Wrong! Removing the Bayesian assumptions does not rescue $\hat{\boldsymbol{\mu}}^{\text{MLE}}$, as James and Stein proved in 1961:

**James–Stein Theorem** *Suppose that*

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

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

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

*for all choices of* $\boldsymbol{\mu} \in \mathcal{R}^N$. (The expectations in (7.16) are with $\boldsymbol{\mu}$ fixed and $\boldsymbol{x}$ varying according to (7.15).)

In the language of decision theory, equation (7.16) says that $\hat{\boldsymbol{\mu}}^{\text{MLE}}$ is *inadmissible*:† its total squared error risk exceeds that of $\hat{\boldsymbol{\mu}}^{\text{JS}}$ no matter †3
what $\boldsymbol{\mu}$ may be. This is a strong frequentist form of defeat for $\hat{\boldsymbol{\mu}}^{\text{MLE}}$, not depending on Bayesian assumptions.

The James–Stein theorem came as a rude shock to the statistical world of 1961. First of all, the defeat came on MLE's home field: normal observations with squared error loss. Fisher's "logic of inductive inference," Chapter 4, claimed that $\hat{\mu}^{\text{MLE}} = x$ was the obviously correct estimator in the univariate case, an assumption tacitly carried forward to multiparameter linear

regression problems, where versions of $\hat{\boldsymbol{\mu}}^{\text{MLE}}$ were predominant. There are still some good reasons for sticking with $\hat{\boldsymbol{\mu}}^{\text{MLE}}$ in low-dimensional problems, as discussed in Section 7.4. But shrinkage estimation, as exemplified by the James–Stein rule, has become a necessity in the high-dimensional situations of modern practice.

## 7.2 The Baseball Players

The James–Stein theorem doesn't say by how much $\hat{\boldsymbol{\mu}}^{\text{JS}}$ beats $\hat{\boldsymbol{\mu}}^{\text{MLE}}$. If the improvement were infinitesimal nobody except theorists would be interested. In favorable situations the gains can in fact be substantial, as suggested by (7.14). One such situation appears in Table 7.1. The batting averages[1] of 18 Major League players have been observed over the 1970 season. The column labeled **MLE** reports the player's observed average over his first 90 at bats; **TRUTH** is the average over the remainder of the 1970 season (370 further at bats on average). We would like to predict **TRUTH** from the early-season observations.

The column labeled **JS** in Table 7.1 is from a version of the James–Stein estimator applied to the 18 MLE numbers. We suppose that each player's **MLE** value $p_i$ (his batting average in the first 90 tries) is a binomial proportion,

$$p_i \sim \text{Bi}(90, P_i)/90. \tag{7.17}$$

Here $P_i$ is his *true average*, how he would perform over an infinite number of tries; $\texttt{TRUTH}_i$ is itself a binomial proportion, taken over an average of 370 more tries per player.

At this point there are two ways to proceed. The simplest uses a normal approximation to (7.17),

$$p_i \doteq \mathcal{N}(P_i, \sigma_0^2), \tag{7.18}$$

where $\sigma_0^2$ is the binomial variance

$$\sigma_0^2 = \bar{p}(1-\bar{p})/90, \tag{7.19}$$

with $\bar{p} = 0.254$ the average of the $p_i$ values. Letting $x_i = p_i/\sigma_0$, applying (7.13), and transforming back to $\hat{p}_i^{\text{JS}} = \sigma_0 \hat{\mu}_i^{\text{JS}}$, gives James–Stein estimates

$$\hat{p}_i^{\text{JS}} = \bar{p} + \left[1 - \frac{(N-3)\sigma_0^2}{\sum(p_i - \bar{p})^2}\right](p_i - \bar{p}). \tag{7.20}$$

[1] Batting average = # hits /# at bats, that is, the success rate. For example, Player 1 hits successfully 31 times in his first 90 tries, for batting average $31/90 = 0.345$. This data is based on 1970 Major League performances, but is partly artificial; see the endnotes.

**Table 7.1** *Eighteen baseball players;* `MLE` *is batting average in first 90 at bats;* `TRUTH` *is average in remainder of 1970 season; James–Stein estimator* `JS` *is based on arcsin transformation of MLEs. Sum of squared errors for predicting* `TRUTH`*:* `MLE` *.0425,* `JS` *.0218.*

| 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 |

A second approach begins with the *arcsin transformation*

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

$n = 90$ (column labeled `x` in Table 7.1), a classical device that produces approximate normal deviates of variance 1,

$$x_i \doteq \mathcal{N}(\mu_i, 1), \tag{7.22}$$

where $\mu_i$ is transformation (7.21) applied to $\texttt{TRUTH}_i$. Using (7.13) gives $\hat{\mu}_i^{\text{JS}}$, which is finally inverted back to the binomial scale,

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

Formulas (7.20) and (7.23) yielded nearly the same estimates for the baseball players; the `JS` column in Table 7.1 is from (7.23). James and Stein's theorem requires normality, but the James–Stein estimator often

works perfectly well in less ideal situations. That is the case in Table 7.1:

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

In other words, the James–Stein estimator reduced total predictive squared error by about 50%.

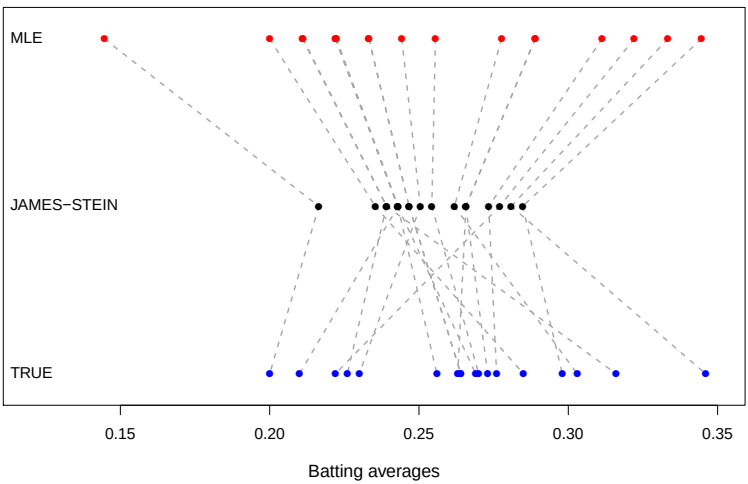


**Figure 7.1** Eighteen baseball players; top line MLE, middle James–Stein, bottom true values. Only 13 points are visible, since there are ties.

The James–Stein rule describes a *shrinkage estimator*, each MLE value $x_i$ being shrunk by factor $\hat{B}$ toward the grand mean $\hat{M} = \bar{x}$ (7.13). ($\hat{B} = 0.34$ in (7.20).) Figure 7.1 illustrates the shrinking process for the baseball players.

To see why shrinking might make sense, let us return to the original Bayes model (7.8) and take $M = 0$ for simplicity, so that the $x_i$ are marginally $\mathcal{N}(0, A+1)$ (7.11). Even though each $x_i$ is unbiased for its parameter $\mu_i$, as a group they are "overdispersed,"

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

The sum of squares of the MLEs exceeds that of the true values by expected amount $N$; shrinkage improves group estimation by removing the excess.

In fact the James–Stein rule *overshrinks* the data, as seen in the bottom two lines of Figure 7.1, a property it inherits from the underlying Bayes model: the Bayes estimates $\hat{\mu}_i^{\text{Bayes}} = Bx_i$ have

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

overshrinking $E(\sum \mu_i^2) = NA$ by factor $A/(A+1)$. We could use the less extreme shrinking rule $\tilde{\mu}_i = \sqrt{B}x_i$, which gives the correct expected sum of squares $NA$, but a larger expected sum of squared estimation errors $E\{\sum(\tilde{\mu}_i - \mu_i)^2|\boldsymbol{x}\}$.

The most extreme shrinkage rule would be "all the way," that is, to

$$\hat{\mu}_i^{\text{NULL}} = \bar{x} \qquad \text{for } i = 1, 2, \ldots, N, \tag{7.27}$$

NULL indicating that in a classical sense we have accepted the null hypothesis of no differences among the $\mu_i$ values. (This gave $\sum(P_i - \bar{p})^2 = 0.0266$ for the baseball data (7.24).) The James–Stein estimator is a data-based rule for compromising between the null hypothesis of no differences and the MLE's tacit assumption of no relationship at all among the $\mu_i$ values. In this sense it blurs the classical distinction between hypothesis testing and estimation.

## 7.3 Ridge Regression

Linear regression, perhaps the most widely used estimation technique, is based on a version of $\hat{\boldsymbol{\mu}}^{\text{MLE}}$. In the usual notation, we observe an $n$-dimensional vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ from the linear model

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}. \tag{7.28}$$

Here $\boldsymbol{X}$ is a known $n \times p$ *structure matrix*, $\beta$ is an unknown $p$-dimensional parameter vector, while the *noise vector* $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$ has its components uncorrelated and with constant variance $\sigma^2$,

$$\boldsymbol{\epsilon} \sim (\boldsymbol{0}, \sigma^2\boldsymbol{I}), \tag{7.29}$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix. Often $\boldsymbol{\epsilon}$ is assumed to be multivariate normal,

$$\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \tag{7.30}$$

but that is not required for most of what follows.

The *least squares estimate* $\hat{\beta}$, going back to Gauss and Legendre in the early 1800s, is the minimizer of the total sum of squared errors,

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|y - X\beta\|^2 \right\}. \tag{7.31}$$

It is given by

$$\hat{\beta} = S^{-1} X' y, \tag{7.32}$$

where $S$ is the $p \times p$ inner product matrix

$$S = X'X; \tag{7.33}$$

$\hat{\beta}$ is unbiased for $\beta$ and has covariance matrix $\sigma^2 S^{-1}$,

$$\hat{\beta} \sim \left(\beta, \sigma^2 S^{-1}\right). \tag{7.34}$$

In the normal case (7.30) $\hat{\beta}$ is the MLE of $\beta$. Before 1950 a great deal of effort went into designing matrices $X$ such that $S^{-1}$ could be feasibly calculated, which is now no longer a concern.

A great advantage of the linear model is that it reduces the number of unknown parameters to $p$ (or $p + 1$ including $\sigma^2$), no matter how large $n$ may be. In the kidney data example of Section 1.1, $n = 157$ while $p = 2$. In modern applications, however, $p$ has grown larger and larger, sometimes into the thousands or more, as we will see in Part III, causing statisticians again to confront the limitations of high-dimensional unbiased estimation.

*Ridge regression* is a shrinkage method designed to improve the estimation of $\beta$ in linear models. By transformations[†] we can *standardize* (7.28) so that the columns of $X$ each have mean 0 and sum of squares 1, that is,

†4

$$S_{ii} = 1 \qquad \text{for } i = 1, 2, \ldots, p. \tag{7.35}$$

(This puts the regression coefficients $\beta_1, \beta_2, \ldots, \beta_p$ on comparable scales.) For convenience, we also assume $\bar{y} = 0$. A ridge regression estimate $\hat{\beta}(\lambda)$ is defined, for $\lambda \geq 0$, to be

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

(using (7.32)); $\hat{\beta}(\lambda)$ is a shrunken version of $\hat{\beta}$, the bigger $\lambda$ the more extreme the shrinkage: $\hat{\beta}(0) = \hat{\beta}$ while $\hat{\beta}(\infty)$ equals the vector of zeros.

Ridge regression effects can be quite dramatic. As an example, consider the diabetes data, partially shown in Table 7.2, in which 10 prediction variables measured at baseline—**age**, **sex**, **bmi** (body mass index), **map** (mean arterial blood pressure), and six blood serum measurements—have

**Table 7.2** *First 7 of $n = 442$ patients in the diabetes study; we wish to predict disease progression at one year "prog" from the 10 baseline measurements* **age**, **sex**, . . . , **glu**.

| 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 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

been obtained for $n = 442$ patients. We wish to use the 10 variables to predict **prog**, a quantitative assessment of disease progression one year after baseline. In this case $\boldsymbol{X}$ is the $442 \times 10$ matrix of standardized predictor variables, and $\boldsymbol{y}$ is **prog** with its mean subtracted off.

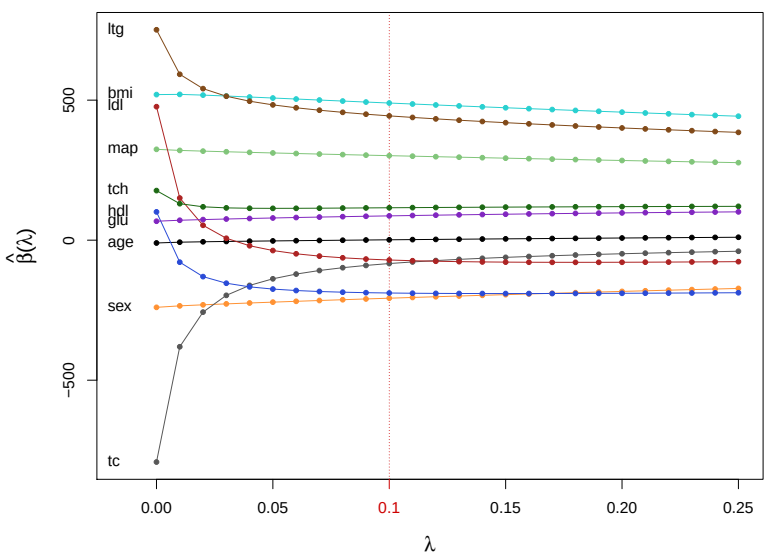


**Figure 7.2** Ridge coefficient trace for the standardized diabetes data.

**Table 7.3** *Ordinary least squares estimate $\hat{\beta}(0)$ compared with ridge regression estimate $\hat{\beta}(0.1)$ with $\lambda = 0.1$. The columns sd(0) and sd(0.1) are their estimated standard errors. (Here $\sigma$ was taken to be 54.1, the usual OLS estimate based on model (7.28).)*

| | $\hat{\beta}(0)$ | $\hat{\beta}(0.1)$ | 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 |

Figure 7.2 vertically plots the 10 coordinates of $\hat{\beta}(\lambda)$ as the ridge parameter $\lambda$ increases from 0 to 0.25. Four of the coefficients change rapidly at first. Table 7.3 compares $\hat{\beta}(0)$, that is the usual estimate $\hat{\beta}$, with $\hat{\beta}(0.1)$. Positive coefficients predict increased disease progression. Notice that `ldl`, the "bad cholesterol" measurement, goes from being a strongly positive predictor in $\hat{\beta}$ to a mildly negative one in $\hat{\beta}(0.1)$.

There is a Bayesian rationale for ridge regression. Assume that the noise vector $\boldsymbol{\epsilon}$ is normal as in (7.30), so that

$$\hat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2 \boldsymbol{S}^{-1}\right) \tag{7.37}$$

rather than just (7.34). Then the Bayesian prior

$$\beta \sim \mathcal{N}_p\left(\mathbf{0}, \frac{\sigma^2}{\lambda}\boldsymbol{I}\right) \tag{7.38}$$

makes

$$E\left\{\beta|\hat{\beta}\right\} = (\boldsymbol{S} + \lambda \boldsymbol{I})^{-1}\boldsymbol{S}\hat{\beta}, \tag{7.39}$$

the same as the ridge regression estimate $\hat{\beta}(\lambda)$ (using (5.23) with $M = 0$, $\boldsymbol{A} = (\sigma^2/\lambda)\boldsymbol{I}$, and $\boldsymbol{\Sigma} = (\boldsymbol{S}/\sigma^2)^{-1}$). Ridge regression amounts to an increased prior belief that $\beta$ lies near 0.

†5 The last two columns of Table 7.3 compare the standard deviations† of $\hat{\beta}$ and $\hat{\beta}(0.1)$. Ridging has greatly reduced the variability of the estimated

regression coefficients. This does *not* guarantee that the corresponding estimate of $\boldsymbol{\mu} = \boldsymbol{X}\beta$,

$$\hat{\boldsymbol{\mu}}(\lambda) = \boldsymbol{X}\hat{\beta}(\lambda), \tag{7.40}$$

will be more accurate than the ordinary least squares estimate $\hat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\beta}$. We have (deliberately) introduced bias, and the squared bias term counteracts some of the advantage of reduced variability. The $C_p$ calculations of Chapter 12 suggest that the two effects nearly offset each other for the diabetes data. However, if interest centers on the coefficients of $\beta$, then ridging can be crucial, as Table 7.3 emphasizes.

By current standards, $p = 10$ is a small number of predictors. Data sets with $p$ in the thousands, and more, will show up in Part III. In such situations the scientist is often looking for a few interesting predictor variables hidden in a sea of uninteresting ones: the prior belief is that most of the $\beta_i$ values lie near zero. Biasing the maximum likelihood estimates $\hat{\beta}_i$ toward zero then becomes a necessity.

There is still another way to motivate the ridge regression estimator $\hat{\beta}(\lambda)$:

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

Differentiating the term in brackets with respect to $\beta$ shows that $\hat{\beta}(\lambda) = (\boldsymbol{S} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$ as in (7.36). If $\lambda = 0$ then (7.41) describes the ordinary least squares algorithm; $\lambda > 0$ *penalizes* choices of $\beta$ having $\|\beta\|$ large, biasing $\hat{\beta}(\lambda)$ toward the origin.

Various terminologies are used to describe algorithms such as (7.41): *penalized least squares*; *penalized likelihood*; *maximized a-posteriori probability* (MAP);†and, generically, *regularization* describes almost any method †6
that tamps down statistical variability in high-dimensional estimation or prediction problems.

A wide variety of penalty terms are in current use, the most influential one involving the "$\ell_1$ norm" $\|\beta\|_1 = \sum_1^p |\beta_j|$,

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

the so-called *lasso* estimator, Chapter 16. Despite the Bayesian provenance, most regularization research is carried out frequentistically, with various penalty terms investigated for their probabilistic behavior regarding estimation, prediction, and variable selection.

If we apply the James–Stein rule to the normal model (7.37), we get a different shrinkage rule† for $\hat{\beta}$, say $\tilde{\beta}^{\text{JS}}$, †7

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

Letting $\tilde{\boldsymbol{\mu}}^{\mathrm{JS}} = \boldsymbol{X}\tilde{\beta}^{\mathrm{JS}}$ be the corresponding estimator of $\boldsymbol{\mu} = E\{\boldsymbol{y}\}$ in (7.28), the James–Stein Theorem guarantees that

$$E\left\{\left\|\tilde{\boldsymbol{\mu}}^{\mathrm{JS}} - \boldsymbol{\mu}\right\|^2\right\} < p\sigma^2 \tag{7.44}$$

no matter what $\beta$ is, as long as $p \geq 3$.[2] There is no such guarantee for ridge regression, and no foolproof way to choose the ridge parameter $\lambda$. On the other hand, $\tilde{\beta}^{\mathrm{JS}}$ does not stabilize the coordinate standard deviations, as in the sd(0.1) column of Table 7.3. The main point here is that at present there is no optimality theory for shrinkage estimation. Fisher provided an elegant theory for optimal unbiased estimation. It remains to be seen whether biased estimation can be neatly codified.

## 7.4 Indirect Evidence 2

There is a downside to shrinkage estimation, which we can examine by returning to the baseball data of Table 7.1. One thousand simulations were run, each one generating simulated batting averages

$$p_i^* \sim \mathrm{Bi}(90, \mathtt{TRUTH}_i)/90 \qquad i = 1, 2, \ldots, 18. \tag{7.45}$$

These gave corresponding James–Stein (JS) estimates (7.20), with $\sigma_0^2 = \bar{p}^*(1-\bar{p}^*)/90$.

Table 7.4 shows the root mean square error for the MLE and JS estimates over 1000 simulations for each of the 18 players,

$$\left[\sum_{j=1}^{1000}(p_{ij}^* - \mathtt{TRUTH}_i)^2\right]^{1/2} \quad \text{and} \quad \left[\sum_{j=1}^{1000}(\hat{p}_{ij}^{*\mathrm{JS}} - \mathtt{TRUTH}_i)^2\right]^{1/2}. \tag{7.46}$$

As foretold by the James–Stein Theorem, the JS estimates are easy victors in terms of total squared error (summing over all 18 players). However, $\hat{p}_i^{*\mathrm{JS}}$ loses to $\hat{p}_i^{*\mathrm{MLE}} = p_i^*$ for 4 of the 18 players, losing badly in the case of player 2.

Histograms comparing the 1000 simulations of $p_i^*$ with those of $\hat{p}_i^{*\mathrm{JS}}$ for player 2 appear in Figure 7.3. Strikingly, all 1000 of the $\hat{p}_{2j}^{*\mathrm{JS}}$ values lie

[2] Of course we are assumimg $\sigma^2$ is known in (7.43); if it is estimated, some of the improvement erodes away.

**Table 7.4** *Simulation study comparing root mean square errors for MLE and JS estimators* (7.20) *as estimates of* `TRUTH`*. Total mean square errors .0384 (*`MLE`*) and .0235 (*`JS`*). Asterisks indicate four players for whom* `rmsJS` *exceeded* `rmsMLE`*; these have two largest and two smallest* `TRUTH` *values (player 2 is Clemente). Column* `rmsJS1` *is for the limited translation version of* `JS` *that bounds shrinkage to within one standard deviation of the* `MLE`*.*

| 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 |

*below* $\texttt{TRUTH}_2 = 0.346$. Player 2 could have had a legitimate complaint if the James–Stein estimate were used to set his next year's salary.

The four losing cases for $\hat{p}_i^{*\,\mathrm{JS}}$ are the players with the two largest and two smallest values of the `TRUTH`. Shrinkage estimators work against cases that are genuinely outstanding (in a positive or negative sense). Player 2 was Roberto Clemente. A better informed Bayesian, that is, a baseball fan, would know that Clemente had led the league in batting over the previous several years, and shouldn't be thrown into a shrinkage pool with 17 ordinary hitters.

Of course the James–Stein estimates *were* more accurate for 14 of the 18 players. Shrinkage estimation tends to produce better results *in general*, at the possible expense of extreme cases. Nobody cares much about Cold War batting averages, but if the context were the efficacies of 18 new anticancer drugs the stakes would be higher.

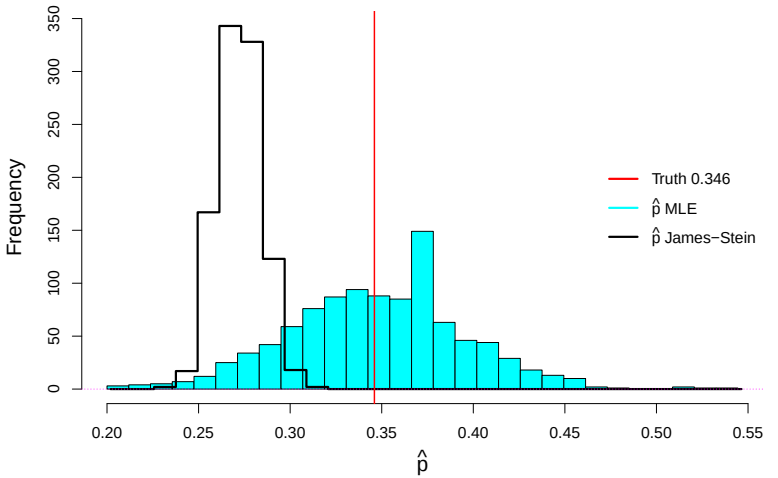


**Figure 7.3** Comparing MLE estimates (solid) with JS estimates (line) for Clemente; 1000 simulations, 90 at bats each.

Compromise methods are available. The `rmsJS1` column of Table 7.4 refers to a *limited translation* version of $\hat{p}_i^{\,\mathrm{JS}}$ in which shrinkage is not allowed to diverge more than one $\sigma_0$ unit from $\hat{p}_i$; in formulaic terms,

$$\hat{p}_i^{\,\mathrm{JS}\,1} = \min\left\{\max\left(\hat{p}_i^{\,\mathrm{JS}}, \hat{p}_i - \sigma_0\right), \hat{p}_i + \sigma_0\right\}. \tag{7.47}$$

This mitigates the Clemente problem while still gaining most of the shrinkage advantages.

The use of indirect evidence amounts to *learning from the experience of others*, each batter learning from the 17 others in the baseball examples. "Which others?" is a key question in applying computer-age methods. Chapter 15 returns to the question in the context of false-discovery rates.

## 7.5 Notes and Details

The Bayesian motivation emphasized in Chapters 6 and 7 is anachronistic: originally the work emerged mainly from frequentist considerations and was justified frequentistically, as in Robbins (1956). Stein (1956) proved the inadmissibility of $\hat{\boldsymbol{\mu}}^{\mathrm{MLE}}$, the neat version of $\hat{\boldsymbol{\mu}}^{\mathrm{JS}}$ appearing in James and Stein (1961) (Willard James was Stein's graduate student); $\hat{\boldsymbol{\mu}}^{\mathrm{JS}}$ is itself inadmissable, being everywhere improvable by changing $\hat{B}$ in (7.13)

to $\max(\hat{B}, 0)$. This in turn is inadmissable, but further gains tend to the minuscule.

In a series of papers in the early 1970s, Efron and Morris emphasized the empirical Bayes motivation of the James–Stein rule, Efron and Morris (1972) giving the limited translation version (7.47). The baseball data in its original form appears in Table 1.1 of Efron (2010). Here the original 45 at bats recorded for each player have been artificially augmented by adding 45 binomial draws, $\text{Bi}(45, \texttt{TRUTH}_i)$ for player $i$. This gives a somewhat less optimistic view of the James–Stein rule's performance.

"Stein's paradox in statistics," Efron and Morris' title for their 1977 *Scientific American* article, catches the statistics world's sense of discomfort with the James–Stein theorem. Why should our estimate for Player A go up or down depending on the other players' performances? This is the question of direct versus indirect evidence, raised again in the context of hypothesis testing in Chapter 15. Unbiased estimation has great scientific appeal, so the argument is by no means settled.

Ridge regression was introduced into the statistics literature by Hoerl and Kennard (1970). It appeared previously in the numerical analysis literature as Tikhonov regularization.

†1 [p. 93] *Formula* (7.12). If $Z$ has a chi-squared distribution with $\nu$ degrees of freedom, $Z \sim \chi^2_\nu$ (that is, $Z \sim \text{Gam}(\nu/2, 2)$ in Table 5.1), it has density

$$f(z) = \frac{z^{\nu/2-1}e^{-z/2}}{2^{\nu/2}\Gamma(\nu/2)} \qquad \text{for } z \geq 0, \tag{7.48}$$

yielding

$$E\left\{\frac{1}{z}\right\} = \int_0^\infty \frac{z^{\nu/2-2}e^{-z/2}}{2^{\nu/2}\Gamma(\nu/2)}\,dz = \frac{2^{\nu/2-1}}{2^{\nu/2}}\frac{\Gamma(\nu/2-1)}{\Gamma(\nu/2)} = \frac{1}{\nu-2}. \tag{7.49}$$

But standard results, starting from (7.11), show that $S \sim (A+1)\chi^2_{N-1}$. With $\nu = N-1$ in (7.49),

$$E\left\{\frac{N-3}{S}\right\} = \frac{1}{A+1}, \tag{7.50}$$

verifying (7.12).

†2 [p. 93] *Formula* (7.14). First consider the simpler situation where $M$ in (7.11) is known to equal zero, in which case the James–Stein estimator is

$$\hat{\mu}_i^{\text{JS}} = \hat{B}x_i \qquad \text{with } \hat{B} = 1-(N-2)/S, \tag{7.51}$$

where $S = \sum_1^N x_i^2$. For convenient notation let

$$\hat{C} = 1-\hat{B} = (N-2)/S \quad \text{and} \quad C = 1-B = 1/(A+1). \tag{7.52}$$

The conditional distribution $\mu_i|\boldsymbol{x} \sim \mathcal{N}(Bx_i, B)$ gives

$$E\left\{\left(\hat{\mu}_i^{\mathrm{JS}} - \mu_i\right)^2 \middle| \boldsymbol{x}\right\} = B + (\hat{C} - C)^2 x_i^2, \tag{7.53}$$

and, adding over the $N$ coordinates,

$$E\left\{\left\|\hat{\boldsymbol{\mu}}^{\mathrm{JS}} - \boldsymbol{\mu}\right\|^2 \middle| \boldsymbol{x}\right\} = NB + (\hat{C} - C)^2 S. \tag{7.54}$$

The marginal distribution $S \sim (A+1)\chi^2_N$ and (7.49) yields, after a little calculation,

$$E\left\{(\hat{C} - C)^2 S\right\} = 2(1 - B), \tag{7.55}$$

and so

$$E\left\{\left\|\hat{\boldsymbol{\mu}}^{\mathrm{JS}} - \boldsymbol{\mu}\right\|^2\right\} = NB + 2(1 - B). \tag{7.56}$$

By orthogonal transformations, in situation (7.7), where $M$ is not assumed to be zero, $\hat{\boldsymbol{\mu}}^{\mathrm{JS}}$ can be represented as the sum of two parts: a JS estimate in $N - 1$ dimensions but with $M = 0$ as in (7.51), and a MLE estimate of the remaining one coordinate. Using (7.56) this gives

$$\begin{aligned} E\left\{\left\|\hat{\boldsymbol{\mu}}^{\mathrm{JS}} - \boldsymbol{\mu}\right\|^2\right\} &= (N-1)B + 2(1-B) + 1 \\ &= NB + 3(1-B), \end{aligned} \tag{7.57}$$

which is (7.14).

†3 [p. 93] *The James–Stein Theorem.* Stein (1981) derived a simpler proof of the JS Theorem that appears in Section 1.2 of Efron (2010).

†4 [p. 98] *Transformations to form* (7.35). The linear regression model (7.28) is *equivariant* under scale changes of the variables $\boldsymbol{x}_j$. What this means is that the space of fits using linear combinations of the $\boldsymbol{x}_j$ is the same as the space of linear combinations using scaled versions $\tilde{\boldsymbol{x}}_j = \boldsymbol{x}_j / s_j$, with $s_j > 0$. Furthermore, the least squares fits are the same, and the coefficient estimates map in the obvious way: $\hat{\tilde{\beta}}_j = s_j \hat{\beta}_j$.

Not so for ridge regression. Changing the scales of the columns of $\boldsymbol{X}$ will generally lead to different fits. Using the penalty version (7.41) of ridge regression, we see that the penalty term $\|\beta\|^2 = \sum_j \beta_j^2$ treats all the coefficients as equals. This penalty is most natural if all the variables are measured on the same scale. Hence we typically use for $s_j$ the standard deviation of variable $\boldsymbol{x}_j$, which leads to (7.35). Furthermore, with ridge regression we typically do not penalize the intercept. This can be achieved

by *centering* and scaling each of the variables, $\tilde{\boldsymbol{x}}_j = (\boldsymbol{x}_j - \mathbf{1}\bar{x}_j)/s_j$, where

$$\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n \quad \text{and} \quad s_j = \left[\sum (x_{ij} - \bar{x}_j)^2\right]^{1/2}, \tag{7.58}$$

with $\mathbf{1}$ the $n$-vector of 1s. We now work with $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_p)$ rather than $\boldsymbol{X}$, and the intercept is estimated separately as $\bar{y}$.

†5 [p. 100] *Standard deviations in Table 7.3.* From the first equality in (7.36) we calculate the covariance matrix of $\hat{\beta}(\lambda)$ to be

$$\text{Cov}_\lambda = \sigma^2(\boldsymbol{S} + \lambda \boldsymbol{I})^{-1}\boldsymbol{S}(\boldsymbol{S} + \lambda \boldsymbol{I})^{-1}. \tag{7.59}$$

The entries sd(0.1) in Table 7.3 are square roots of the diagonal elements of $\text{Cov}_\lambda$, substituting the ordinary least squares estimate $\hat{\sigma} = 54.1$ for $\sigma^2$.

†6 [p. 101] *Penalized likelihood and MAP.* With $\sigma^2$ fixed and known in the normal linear model $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I})$, minimizing $\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2$ is the same as maximizing the log density function

$$\log f_\beta(y) = -\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \text{ constant}. \tag{7.60}$$

In this sense, the term $\lambda\|\beta\|^2$ in (7.41) *penalizes* the likelihood $\log f_\beta(y)$ connected with $\beta$ in proportion to the magnitude $\|\beta\|^2$. Under the prior distribution (7.38), the log posterior density of $\beta$ given $\boldsymbol{y}$ (the log of (3.5)) is

$$-\frac{1}{2\sigma^2}\left\{\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\|\beta\|^2\right\}, \tag{7.61}$$

plus a term that doesn't depend on $\beta$. That makes the maximizer of (7.41) also the maximizer of the posterior density of $\beta$ given $\boldsymbol{y}$, or the MAP.

†7 [p. 101] *Formula* (7.43). Let $\gamma = (\boldsymbol{S}^{1/2}/\sigma)\beta$ and $\hat{\gamma} = (\boldsymbol{S}^{1/2}/\sigma)\hat{\beta}$ in (7.37), where $\boldsymbol{S}^{1/2}$ is a matrix square root of $\boldsymbol{S}$, $(\boldsymbol{S}^{1/2})^2 = \boldsymbol{S}$. Then

$$\hat{\gamma} \sim \mathcal{N}_p(\gamma, \boldsymbol{I}), \tag{7.62}$$

and the $M = 0$ form of the James–Stein rule (7.51) is

$$\hat{\gamma}^{\text{JS}} = \left[1 - \frac{p-2}{\|\hat{\gamma}\|^2}\right]\hat{\gamma}. \tag{7.63}$$

Transforming back to the $\beta$ scale gives (7.43).