20

Inference After Model Selection

The classical theory of model selection focused on "F tests" performed within Gaussian regression models. Inference after model selection (for instance, assessing the accuracy of a fitted regression curve) was typically done ignoring the model selection process. This was a matter of necessity: the combination of discrete model selection and continuous regression analysis was too awkward for simple mathematical description. Electronic computation has opened the door to a more honest analysis of estimation accuracy, one that takes account of the variability induced by data-based model selection.

Figure 20.1 displays the `cholesterol` data, an example we will use for illustration in what follows: cholestyramine, a proposed cholesterol-lowering drug, was administered to $n = 164$ men for an average of seven years each. The response variable d_i was the ith man's decrease in cholesterol level over the course of the experiment. Also measured was c_i, his compliance or the proportion of the intended dose actually taken, ranging from 1 for perfect compliers to zero for the four men who took none at all. Here the 164 c_i values have been transformed to approximately follow a standard normal distribution,

$$c_i \doteq \mathcal{N}(0, 1). \tag{20.1}$$

We wish to predict cholesterol decrease from compliance. Polynomial regression models, with d_i a Jth-order polynomial in c_i, were considered, for degrees $J = 1, 2, 3, 4, 5$, or 6. The C_p criterion (12.51) was applied and selected a cubic model, $J = 3$, as best. The curve in Figure 20.1 is the OLS (ordinary least squares) cubic regression curve fit to the cholesterol data set

$$\{(c_i, d_i),\ i = 1, 2, \ldots, 164\}. \tag{20.2}$$

We are interested in answering the following question: how accurate is the

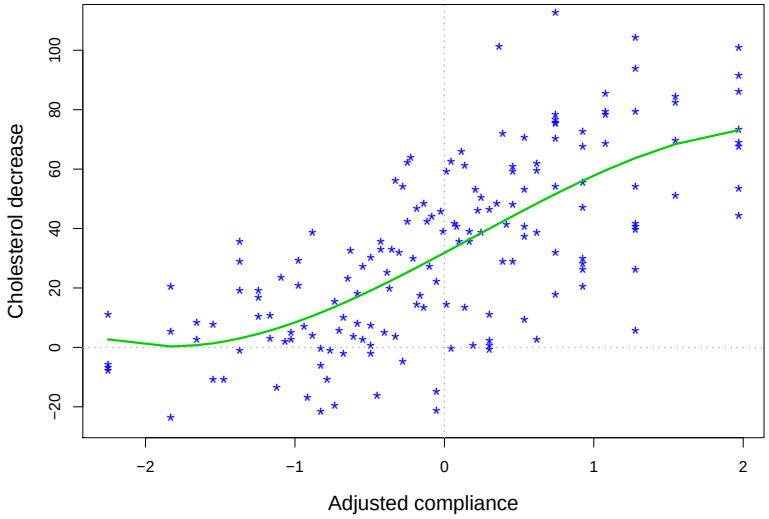

Figure 20.1 `Cholesterol` data: cholesterol decrease plotted versus adjusted compliance for 164 men taking `cholestyramine`. The green curve is OLS cubic regression, with "cubic" selected by the C_p criterion. How accurate is the fitted curve?

fitted curve, taking account of C_p selection as well as OLS estimation? (See Section 20.2 for an answer.)

Currently, there is no overarching theory for inference after model selection. This chapter, more modestly, presents a short series of vignettes that illustrate promising analyses of individual situations. See also Section 16.6 for a brief report on progress in post-selection inference for the lasso.

20.1 Simultaneous Confidence Intervals

In the early 1950s, just before the beginnings of the computer revolution, substantial progress was made on the problem of setting simultaneous confidence intervals. "Simultaneous" here means that there exists a catalog of parameters of possible interest,

$$\boldsymbol{C} = \{\theta_1, \theta_2, \ldots, \theta_J\}, \tag{20.3}$$

and we wish to set a confidence interval for each of them with some fixed probability, typically 0.95, that *all* of the intervals will contain their respective parameters.

As a first example, we return to the **diabetes** data of Section 7.3: $n = 442$ diabetes patients each have had $p = 10$ medical variables measured at baseline, with the goal of predicting **prog**, disease progression one year later. Let $\boldsymbol{X}$ be the 442×10 matrix with ith row x_i' the 10 measurements for patient i; $\boldsymbol{X}$ has been standardized so that each of its columns has mean 0 and sum of squares 1. Also let $\boldsymbol{y}$ be the 442-vector of centered **prog** measurements (that is, subtracting off the mean of the **prog** values).

Ordinary least squares applied to the normal linear model,

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}), \tag{20.4}$$

yields MLE

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{20.5}$$

satisfying

$$\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2 \boldsymbol{V}), \qquad \boldsymbol{V} = (\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{20.6}$$

as at (7.34).

The 95% Student-t confidence interval (11.49) for β_j, the jth component of β, is

$$\hat{\beta}_j \pm \hat{\sigma} V_{jj}^{1/2} t_q^{.975}, \tag{20.7}$$

where $\hat{\sigma} = 54.2$ is the usual unbiased estimate of σ,

$$\hat{\sigma}^2 = \|\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\|^2/q, \qquad q = n - p = 432, \tag{20.8}$$

and $t_q^{.975} = 1.97$ is the 0.975 quantile of a Student-t distribution with q degrees of freedom.

The catalog $\boldsymbol{C}$ in (20.3) is now $\{\beta_1, \beta_2, \ldots, \beta_{10}\}$. The individual intervals (20.7), shown in Table 20.1, each have 95% coverage, but they are not simultaneous: there is a greater than 5% chance that at least one of the β_j values lies outside its claimed interval.

Valid 95% simultaneous intervals for the 10 parameters appear on the right side of Table 20.1. These are the *Scheffé intervals*

$$\hat{\beta}_j \pm \hat{\sigma} V_{jj}^{1/2} k_{p,q}^{(\alpha)}, \tag{20.9}$$

discussed next. The crucial constant $k_{p,q}^{(\alpha)}$ equals 4.30 for $p = 10$, $q = 432$, and $\alpha = 0.95$. That makes the Scheffé intervals wider than the t intervals (20.7) by a factor of 2.19. One expects to pay an extra price for simultaneous coverage, but a factor greater than two induces sticker shock.

Scheffé's method depends on the pivotal quantity

$$Q = \left(\hat{\beta} - \beta\right)' \boldsymbol{V}^{-1} \left(\hat{\beta} - \beta\right) \Big/ \hat{\sigma}^2, \tag{20.10}$$

Table 20.1 *Maximum likelihood estimates $\hat{\beta}$ for 10 diabetes predictor variables* (20.6)*; separate 95% Student-t confidence limits, also simultaneous 95% Scheffé intervals. The Scheffé intervals are wider by a factor of 2.19.*

		Student-*t*		Scheffé	
	$\hat{\beta}$	Lower	Upper	Lower	Upper
`age`	−0.5	−6.1	5.1	−12.7	11.8
`sex`	−11.4	−17.1	−5.7	−24.0	1.1
`bmi`	24.8	18.5	31.0	11.1	38.4
`map`	15.4	9.3	21.6	2.1	28.8
`tc`	−37.7	−76.7	1.2	−123.0	47.6
`ldl`	22.7	−9.0	54.4	−46.7	92.1
`hdl`	4.8	−15.1	24.7	−38.7	48.3
`tch`	8.4	−6.7	23.5	−24.6	41.5
`ltg`	35.8	19.7	51.9	0.6	71.0
`glu`	3.2	−3.0	9.4	−10.3	16.7

which under model (20.4) has a scaled "F distribution,"[1]

$$Q \sim pF_{p,q}. \tag{20.11}$$

If $k_{p,q}^{(\alpha)2}$ is the αth quantile of a $pF_{p,q}$ distribution then $\Pr\{Q \le k_{p,q}^{(\alpha)2}\} = \alpha$ yields

$$\Pr\left\{\frac{\left(\beta-\hat{\beta}\right)' V^{-1}\left(\beta-\hat{\beta}\right)}{\hat{\sigma}^2} \le k_{p,q}^{(\alpha)2}\right\} = \alpha \tag{20.12}$$

for any choice of β and σ in model (20.4). Having observed $\hat{\beta}$ and $\hat{\sigma}$, (20.12) defines an elliptical confidence region $\mathcal{E}$ for the parameter vector β.

Suppose we are interested in a particular linear combination of the coordinates of β, say

$$\beta_c = c'\beta, \tag{20.13}$$

[1] $F_{p,q}$ is distributed as $(\chi_p^2/p)/(\chi_q^2/q)$, the two chi-squared variates being independent. Calculating the percentiles of $F_{p,q}$ was a major project of the pre-war period.

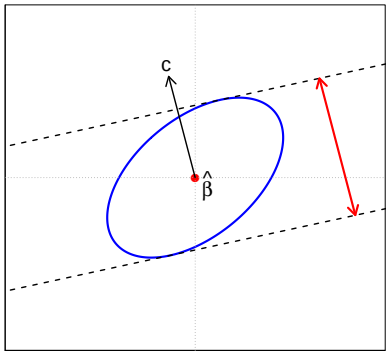

Figure 20.2 Ellipsoid of possible vectors β defined by (20.12) determines confidence intervals for $\beta_c = c'\beta$ according to the "bounding hyperplane" construction illustrated. The red line shows the confidence interval for β_c if c is a unit vector, $c'\boldsymbol{V}c = 1$.

where c is a fixed p-dimensional vector. If β exists in $\mathcal{E}$ then we must have

$$\beta_c \in \left[\min_{\beta\in\mathcal{E}}(c'\beta),\ \max_{\beta\in\mathcal{E}}(c'\beta)\right], \tag{20.14}$$

†1 which turns out† to be the interval centered at $\hat{\beta}_c = c'\hat{\beta}$,

$$\beta_c \in \hat{\beta}_c \pm \hat{\sigma}(c'\boldsymbol{V}c)^{1/2}k_{p,q}^{(\alpha)}. \tag{20.15}$$

(This agrees with (20.9) where c is the jth coordinate vector $(0,\ldots,0,1,0,\ldots,0)'$.) The construction is illustrated in Figure 20.2.

Theorem (Scheffé) *If $\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2\boldsymbol{V})$ independently of $\hat{\sigma}^2 \sim \sigma^2\chi_q^2/q$, then with probability α the confidence statement* (20.15) *for $\beta_c = c'\beta$ will be simultaneously true for all choices of the vector c.*

Here we can think of "model selection" as the choice of the linear combination of interest $\theta_c = c'\beta$. Scheffé's theorem allows "data snooping": the statistician can examine the data and *then* choose which θ_c (or many θ_c's) to estimate, without invalidating the resulting confidence intervals.

An important application has the $\hat{\beta}_j$'s as independent estimates of efficacy for competing treatments—perhaps different experimental drugs for the same target disease:

$$\hat{\beta}_j \overset{\text{ind}}{\sim} \mathcal{N}(\beta_j, \sigma^2/n_j), \qquad \text{for } j = 1, 2, \ldots, J, \tag{20.16}$$

the n_j being known sample sizes. In this case the catalog $\boldsymbol{C}$ might comprise all pairwise differences $\beta_i - \beta_j$, as the statistician tries to determine which treatments are better or worse than the others.

The fact that Scheffé's limits apply to *all* possible linear combinations $c'\beta$ is a blessing and a curse, the curse being their very large width, as seen in Table 20.1. Narrower simultaneous limits† are possible if we restrict the catalog $\boldsymbol{C}$, for instance to just the pairwise differences $\beta_i - \beta_j$. †2

A serious objection, along Fisherian lines, is that the Scheffé confidence limits are *accurate* without being *correct*. That is, the intervals have the claimed overall frequentist coverage probability, but may be misleading when applied to individual cases. Suppose for instance that $\sigma^2/n_j = 1$ for $j = 1, 2, \ldots, J$ in (20.16) and that we observe $\hat{\beta}_1 = 10$, with $|\hat{\beta}_j| < 2$ for all the others. Even if we looked at the data before singling out $\hat{\beta}_1$ for attention, the usual Student-t interval (20.7) seems more appropriate than its much longer Scheffé version (20.9). This point is made more convincingly in our next vignette.

A familiar but pernicious abuse of model selection concerns multiple hypothesis testing. Suppose we observe N independent normal variates z_i, each with its own *effect size* μ_i,

$$z_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1) \qquad \text{for } i = 1, 2, \ldots, N, \tag{20.17}$$

and, as in Section 15.1, we wish to test the null hypotheses

$$H_{0i} : \mu_i = 0. \tag{20.18}$$

Being alert to the pitfalls of simultaneous testing, we employ a false-discovery rate control algorithm (15.14), which rejects R of the N null hypotheses, say for cases $i_1, i_2, \ldots, i_R$. (R equaled 28 in the example of Figure 15.3.)

So far so good. The "familiar abuse" comes in then setting the usual confidence intervals

$$\mu_i \in \hat{\mu}_i \pm 1.96 \tag{20.19}$$

(95% coverage) for the R selected cases. This ignores the model selection process: the data-based selection of the R cases must be taken into account in making legitimate inferences, even if R is only 1 so multiplicity is not a concern.

This problem is addressed by the theory of *false-coverage control*. Suppose algorithm $\mathcal{A}$ sets confidence intervals for R of the N cases, of which

r are actually false coverages, i.e., ones not containing the true effect size μ_i. The false-coverage rate (FCR) of $\mathcal{A}$ is the expected proportion of non-coverages

$$\text{FCR}(\mathcal{A}) = E\{r/R\}, \tag{20.20}$$

the expectation being with respect to model (20.17). The goal, as with the FDR theory of Section 15.2, is to construct algorithm $\mathcal{A}$ to control FCR below some fixed value q.

The BY_q algorithm[2] controls FCR below level q in three easy steps, beginning with model (20.17).

1 Let p_i be the p-value corresponding to z_i,

$$p_i = \Phi(z_i) \tag{20.21}$$

for left-sided significance testing, and order the $p_{(i)}$ values in ascending order,

$$p_{(1)} \le p_{(2)} \le p_{(3)} \le \cdots \le p_{(N)}. \tag{20.22}$$

2 Calculate $R = \max\{i : p_{(i)} \le i \cdot q/N\}$, and (as in the BH_q algorithm (15.14)–(15.15)) declare the R corresponding null hypotheses false.

3 For each of the R cases, construct the confidence interval

$$\mu_i \in z_i \pm z^{(\alpha_R)}, \qquad \text{where } \alpha_R = 1 - Rq/N \tag{20.23}$$

($z^{(\alpha)} = \Phi^{-1}(\alpha)$).

Theorem 20.1 *Under model* (20.17)*, BY_q has FCR $\le q$; moreover, none of the intervals* (20.23) *contain $\mu_i = 0$.*

A simulated example of BY_q was run according to these specifications:

$$\begin{aligned} N = 10{,}000, \quad q = 0.05, \quad z_i \sim \mathcal{N}(\mu_i, 1) \\ \mu_i = 0 \qquad \text{for } i = 1, 2, \ldots, 9000, \\ \mu_i \sim \mathcal{N}(-3, 1) \qquad \text{for } i = 9001, \ldots, 10{,}000. \end{aligned} \tag{20.24}$$

In this situation we have 9000 null cases and 1000 non-null cases (all but 2 of which had $\mu_i < 0$).

Because this is a simulation, we can plot the pairs (z_i, μ_i) to assess the BY_q algorithm's performance. This is done in Figure 20.3 for the 1000 non-null cases (the green points). BY_q declared $R = 565$ cases non-null, those having $z_i \le -2.77$ (the circled points); 14 of the 565 declarations

[2] Short for "Benjamini–Yekutieli;" see the chapter endnotes.

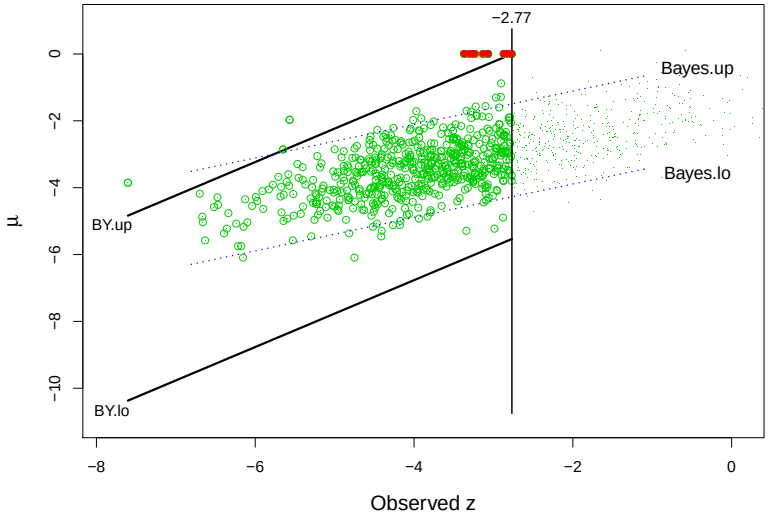

Figure 20.3 Simulation experiment (20.24) with $N = 10{,}000$ cases, of which 1000 are non-null; the green points (z_i, μ_i) are these non-null cases. The FDR control algorithm BH_q ($q = 0.05$) declared the 565 circled cases having $z_i \leq -2.77$ to be non-null, of which the 14 red points were actually null. The heavy black lines show BY_q 95% confidence intervals for the 565 cases, only 17 of which failed to contain μ_i. Actual Bayes posterior 95% intervals for non-null cases (20.26), dotted lines, have half the width and slope of BY_q limits.

were actually null cases (the red circled points), giving false-discovery proportion $14/565 = 0.025$. The heavy black lines trace the BY_q confidence limits (20.23) as a function of $z \leq -2.77$.

The first thing to notice is that FCR control has indeed been achieved: only 17 of the declared cases lie outside their limits (the 14 nulls and 3 non-nulls), for a false-coverage rate of $17/565 = 0.030$, safely less than $q = 0.05$. The second thing, however, is that the BY_q limits provide a misleading idea of the location of μ_i given z_i: they are much too wide and slope too low, especially for more negative z_i values.

In this situation we can describe precisely the posterior distribution of μ_i given z_i for the non-null cases,

$$\mu_i | z_i \sim \mathcal{N}\left(\frac{z_i - 3}{2}, \frac{1}{2}\right), \tag{20.25}$$

this following from $\mu_i \sim \mathcal{N}(-3, 1)$, $z_i|\mu_i \sim \mathcal{N}(\mu_i, 1)$, and Bayes' rule (5.20)–(5.21). The Bayes credible 95% limits

$$\mu_i \in \frac{z_i - 3}{2} \pm \frac{1}{\sqrt{2}} 1.96 \tag{20.26}$$

are indicated by the dotted lines in Figure 20.3. They are half as wide as the BY_q limits, and have slope $1/2$ rather than 1.

In practice, of course, we would only see the z_i, not the μ_i, making (20.26) unavailable to us. We return to this example in Chapter 21, where empirical Bayes methods will be seen to provide a good approximation to the Bayes limits. (See Figure 31.5.)

As with Scheffé's method, the BY_q intervals can be accused of being accurate but not correct. "Correct" here has a Bayesian/Fisherian flavor that is hard to pin down, except perhaps in large-scale applications, where empirical Bayes analyses can suggest appropriate inferences.

20.2 Accuracy After Model Selection

The cubic regression curve for the `cholesterol` data seen in Figure 20.1 was selected according to the C_p criterion of Section 12.3. Polynomial regression models, predicting cholesterol decrease d_i in terms of powers ("degrees") of adjusted compliance c_i, were fit by ordinary least squares for degrees $0, 1, 2, \ldots, 6$. Table 20.2 shows C_p estimates (12.51) being minimized at degree 3.

Table 20.2 *C_p table for cholesterol data of Figure 20.1, comparing OLS polynomial models of degrees 0 through 6. The cubic model, degree $= 3$, is the minimizer (80,000 subtracted from the C_p values for easier comparison; assumes $\sigma = 22.0$).*

Degree	C_p
0	71887
1	1132
2	1412
3	667
4	1591
5	1811
6	2758

We wish to assess the accuracy of the fitted curve, taking account of both the C_p model selection method and the OLS fitting process. The bootstrap

is a natural candidate for the job. Here we will employ the nonparametric bootstrap of Section 10.2 (rather than the parametric bootstrap of Section 10.4, though this would be no more difficult to carry out).

The `cholesterol` data set (20.2) comprises $n = 164$ pairs $x_i = (c_i, d_i)$; a nonparametric bootstrap sample $\boldsymbol{x}^*$ (10.13) consists of 164 pairs chosen at random and *with* replacement from the original 164. Let $t(\boldsymbol{x}^*)$ be the curve obtained by applying the C_p/OLS algorithm to the original data set $\boldsymbol{x}^*$ and likewise $t(\boldsymbol{x}^*)$ for the algorithm applied to $\boldsymbol{x}^*$; and for a given point c on the compliance scale let

$$\hat{\theta}_c^* = t(c, \boldsymbol{x}^*) \tag{20.27}$$

be the value of $t(\boldsymbol{x}^*)$ evaluated at compliance $= c$.

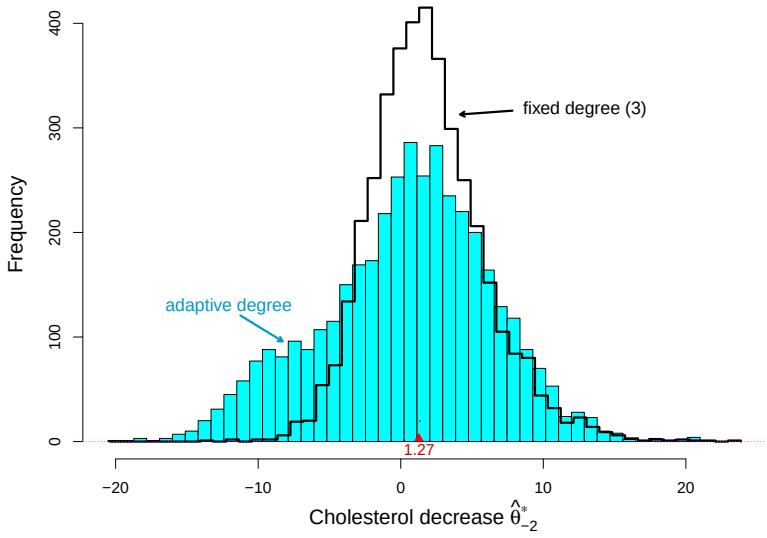

Figure 20.4 A histogram of 4000 nonparametric bootstrap replications for polynomial regression estimates of cholesterol decreases d at adjusted compliance $c = -2$. Solid histogram, adaptive estimator $\hat{\theta}_c^*$ (20.27), using full C_p/OLS algorithm for each bootstrap data set; line histogram, using OLS only with degree 3 for each bootstrap data set. Bootstrap standard errors are 5.98 and 3.97.

$B = 4000$ nonparametric bootstrap replications $t(\boldsymbol{x}^*)$ were generated.[3] Figure 20.4 shows the histogram of the 4000 $\hat{\theta}_c^*$ replications for $c = -2.0$. It is labeled "adaptive" to indicate that C_p model selection, as well as OLS fitting, was carred out anew for each $\boldsymbol{x}^*$. This is as opposed to the "fixed" histogram, where there was no C_p selection, cubic OLS regression always being used.

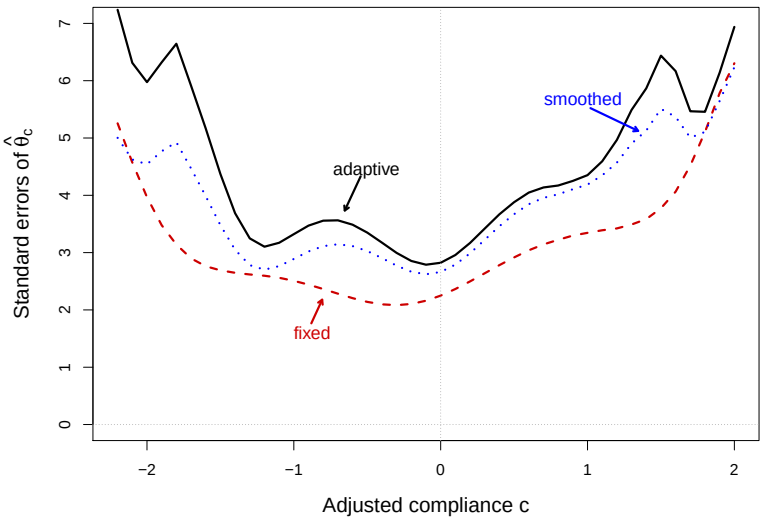

Figure 20.5 Bootstrap standard-error estimates of $\hat{\theta}_c$, for $-2.2 \le c \le 2$. Solid black curve, adaptive estimator (20.27) using full C_p/OLS model selection estimate; red dashed curve, using OLS only with polynomial degree fixed at 3; blue dotted curve, "bagged estimator" using bootstrap smoothing (20.28). Average standard-error ratios: adaptive/fixed = 1.43, adaptive/smoothed = 1.14.

The bootstrap estimate of standard error (10.16) obtained from the adaptive values $\hat{\theta}_c^*$ was 5.98, compared with 3.97 for the fixed values.[4] In this case, accounting for model selection ("adaptation") adds more than 50% to the standard error estimates. The same comparison was made at all values

[3] Ten times more than needed for assessing standard errors, but helpful for the comparisons that follow.

[4] The latter is not the usual OLS assessment, following (8.30), that would be appropriate for a parametric bootstrap comparison. Rather, it's the nonparametric one-sample bootstrap assessment, resampling pairs (x_i, y_i) as individual sample points.

of the adjusted compliance c. Figure 20.5 graphs the results: the adaptive standard errors averaged 43% greater than the fixed values. The standard 95% confidence intervals $\hat{\theta}_c \pm \widehat{\text{se}} \cdot 1.96$ would be roughly 43% too short if we ignored model selection in assessing $\widehat{\text{se}}$.

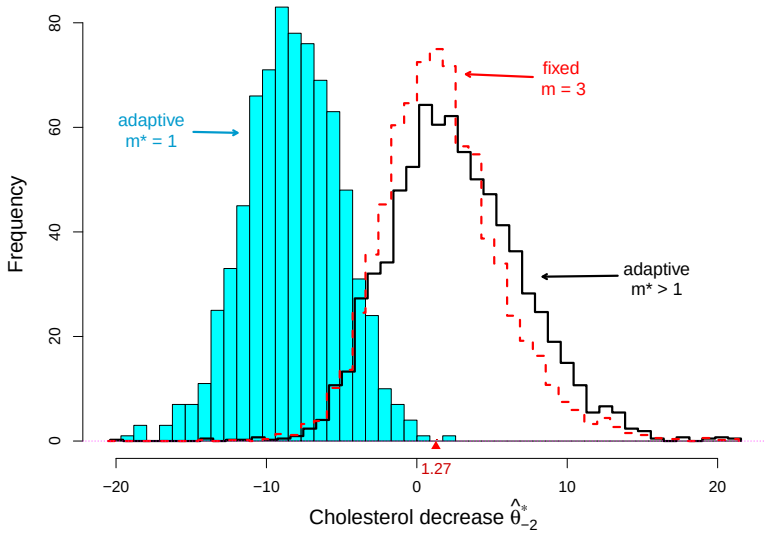

Figure 20.6 "Adaptive" histogram of Figure 20.4 now split into 19% of 4000 bootstrap replications where C_p selected linear regression ($m^* = 1$) as best, versus 81% having $m^* > 1$. $m^* = 1$ cases are shifted about 10 units downward. (The $m^* > 1$ cases resemble the "fixed" histogram in Figure 20.4.) Histograms are scaled to have equal areas.

Having an honest assessment of standard error doesn't mean that $t(c, \boldsymbol{x}^*)$ (20.27) is a good estimator. Model selection can induce an unpleasant "jumpiness" in an estimator, as the original data vector $\boldsymbol{x}$ crosses definitional boundaries. This happened in our example: for 19% of the 4000 bootstrap samples $\boldsymbol{x}^*$, the C_p algorithm selected linear regression, $m^* = 1$, as best, and in these cases $\hat{\theta}^*_{-2.0}$ tended toward smaller values. Figure 20.6 shows the $m^* = 1$ histogram shifted about 10 units down from the $m^* > 1$ histogram (which now resembles the "fixed" histogram in Figure 20.4).

Discontinuous estimators such as $t(c, \boldsymbol{x})$ can't be Bayesian, Bayes posterior expectations being continuous. They can also suffer frequentist difficulties,† including excess variability and overly long confidence intervals. †3

Bagging, or *bootstrap smoothing*, is a tactic for improving a discontinuous estimation rule by averaging (as in (12.80) and Chapter 17).

Suppose $t(\boldsymbol{x})$ is any estimator for which we have obtained bootstrap replications $\{t(\boldsymbol{x}^{*b}), b = 1, 2, \ldots, B\}$. The bagged version of $t(\boldsymbol{x})$ is the average

$$s(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} t(\boldsymbol{x}^{*b}). \tag{20.28}$$

The letter s here stands for "smooth." Small changes in $\boldsymbol{x}$, even ones that move across a model selection definitional boundary, produce only small changes in the bootstrap average $s(\boldsymbol{x})$.

Averaging over the 4000 bootstrap replications of $t(c, \boldsymbol{x}^*)$ (20.27) gave a bagged estimate $s_c(\boldsymbol{x})$ for each value of c. Bagging reduced the standard errors of the C_p/OLS estimates $t(c, \boldsymbol{x})$ by about 12%, as indicated by the green dotted curve in Figure 20.5.

Where did the green dotted curve come from? All 4000 bootstrap values $t(c, \boldsymbol{x}^{*b})$ were needed to produce the single value $s_c(\boldsymbol{x})$. It seems as if we would need to bootstrap the bootstrap in order to compute $\widehat{\text{se}}[s_c(\boldsymbol{x})]$. Fortunately, a more economical calculation is possible, one that requires only the original B bootstrap computations for $t(c, \boldsymbol{x})$.

Define

$$N_{bj} = \#\{\text{times } x_j \text{ occurs in } \boldsymbol{x}^{*b}\}, \tag{20.29}$$

for $b = 1, 2, \ldots, B$ and $j = 1, 2, \ldots, n$. For instance $N_{4000,7} = 3$ says that data point x_7 occurred three times in nonparametric bootstrap sample $\boldsymbol{x}^{4000}$. The B by n matrix $\{N_{bj}\}$ completely describes the B bootstrap samples. Also denote

$$t^{*b} = t(\boldsymbol{x}^{*b}) \tag{20.30}$$

and let cov_j indicate the covariance in the bootstrap sample between N_{bj} and t^{*b},

$$\text{cov}_j = \frac{1}{B}\sum_{b=1}^{B}(N_{bj} - N_{\cdot j})(t^{*b} - t^{*\cdot}), \tag{20.31}$$

where dots denote averaging over B: $N_{\cdot j} = \frac{1}{B}\sum_b N_{bj}$ and $t^{*\cdot} = \frac{1}{B}\sum_b t^{*b}$.

†4 **Theorem 20.2** † *The infinitesimal jackknife estimate of standard error*

(10.41) *for the bagged estimate* (20.28) *is*

$$\widehat{\text{se}}_{\text{IJ}}\,[s_c(\boldsymbol{x})] = \left(\sum_{j=1}^{n} \text{cov}_j^2\right)^{1/2}. \tag{20.32}$$

Keeping track of N_{bj} *as we generate the bootstrap replications* t^{*b} *allows us to compute* cov_j *and* $\widehat{\text{se}}[s_c(\boldsymbol{x})]$ *without any additional computational effort.*

We expect averaging to reduce variability, and this is seen to hold true in Figure 20.5, the ratio of $\widehat{\text{se}}_{\text{IJ}}[s_c(\boldsymbol{x})]/\widehat{\text{se}}_{\text{boot}}[t(c,\boldsymbol{x})]$ averaging 0.88. In fact, we have the following general result.

Corollary *The ratio* $\widehat{\text{se}}_{\text{IJ}}[s_c(\boldsymbol{x})]/\widehat{\text{se}}_{\text{boot}}[t(c,\boldsymbol{x})]$ *is always* ≤ 1.

The savings due to bagging increase with the nonlinearity of $t(\boldsymbol{x}^*)$ as a function of the counts N_{bj} (or, in the language of Section 10.3, in the nonlinearity of $S(\boldsymbol{P})$ as a function of $\boldsymbol{P}$). Model-selection estimators such as the C_p/OLS rule tend toward greater nonlinearity and bigger savings.

Table 20.3 *Proportion of 4000 nonparametric bootstrap replications of* C_p*/OLS algorithm that selected degrees* $m = 1, 2, \ldots, 6$*; also infinitesimal jackknife standard deviations for proportions* (20.32)*, which mostly exceed the estimates themselves.*

	$m=1$	2	3	4	5	6
proportion	.19	.12	.35	.07	.20	.06
$\widehat{\text{sd}}_{\text{IJ}}$	.24	.20	.24	.13	.26	.06

The first line of Table 20.3 shows the proportions in which the various degrees were selected in the 4000 cholesterol bootstrap replications, 19% for linear, 12% for quadratic, 35% for cubic, etc. With $B = 4000$, the proportions seem very accurate, the binomial standard error for 0.19 being just $(0.19 \cdot 0.81/4000)^{1/2} = 0.006$, for instance.

Theorem 20.2 suggests otherwise. Now let t^{*b} (20.30) indicate whether the bth bootstrap sample $\boldsymbol{x}^*$ made the C_p choice $m^* = 1$,

$$t^{*b} = \begin{cases} 1 & \text{if } m^{*b} = 1 \\ 0 & \text{if } m^{*b} > 1. \end{cases} \tag{20.33}$$

The bagged value of $\{t^{*b},\ b = 1, 2, \ldots, B\}$ is the observed proportion

0.19. Applying the bagging theorem yielded $\widehat{\text{se}}_{\text{IJ}} = 0.24$, as seen in the second line of the table, with similarly huge standard errors for the other proportions.

The binomial standard errors are *internal*, saying how quickly the bootstrap resampling process is converging to its ultimate value as $B \to \infty$. The infinitesimal jackknife estimates are *external*: if we collected a new set of 164 data pairs (c_i, d_i) (20.2) the new proportion table might look completely different than the top line of Table 20.3.

Frequentist statistics has the advantage of being applicable to any algorithmic procedure, for instance to our C_p/OLS estimator. This has great appeal in an era of enormous data sets and fast computation. The drawback, compared with Bayesian statistics, is that we have no guarantee that our chosen algorithm is best in any way. Classical statistics developed a theory of *best* for a catalog of comparatively simple estimation and testing problems. In this sense, modern inferential theory has not yet caught up with modern problems such as data-based model selection, though techniques such as *model averaging* (e.g., bagging) suggest promising steps forward.

20.3 Selection Bias

Many a sports fan has been victimized by selection bias. Your team does wonderfully well and tops the league standings. But the next year, with the same players and the same opponents, you're back in the pack. This is the *winner's curse*, a more picturesque name for selection bias, the tendency of unusually good (or bad) comparative performances not to repeat themselves.

Modern scientific technology allows the simultaneous investigation of hundreds or thousands of candidate situations, with the goal of choosing the top performers for subsequent study. This is a setup for the heartbreak of selection bias. An apt example is offered by the prostate study data of Section 15.1, where we observe statistics z_i measuring patient–control differences for $N = 6033$ genes,

$$z_i \sim \mathcal{N}(\mu_i, 1), \qquad i = 1, 2, \ldots, N. \tag{20.34}$$

Here μ_i is the *effect size* for gene i, the true difference between the patient and control populations.

Genes with large positive or negative values of μ_i would be promising targets for further investigation. Gene number 610, with $z_{610} = 5.29$, at-

tained the biggest z-value; (20.34) says that z_{610} is unbiased for μ_{610}. Can we believe the obvious estimate $\hat{\mu}_{610} = 5.29$?

"No" is the correct selection bias answer. Gene 610 has won a contest for bigness among 6033 contenders. In addition to being *good* (having a large value of μ) it has almost certainly been *lucky*, with the noise in (20.34) pushing z_{610} in the positive direction—or else it would not have won the contest. This is the essence of selection bias.

False-discovery rate theory, Chapter 15, provided a way to correct for selection bias in simultaneous hypothesis testing. This was extended to false-coverage rates in Section 20.1. Our next vignette concerns the realistic estimation of effect sizes μ_i in the face of selection bias.

We begin by assuming that an effect size μ has been obtained from a prior density $g(\mu)$ (which might include discrete atoms) and then $z \sim \mathcal{N}(\mu, \sigma^2)$ observed,

$$\mu \sim g(\cdot) \quad \text{and} \quad z|\mu \sim \mathcal{N}(\mu, \sigma^2) \tag{20.35}$$

(σ^2 is assumed known for this discussion). The marginal density of z is

$$\begin{aligned} f(z) &= \int_{-\infty}^{\infty} g(\mu)\phi_\sigma(z-\mu)\, d\mu, \\ \text{where } \phi_\sigma(z) &= (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2}\frac{z^2}{\sigma^2}\right). \end{aligned} \tag{20.36}$$

Tweedie's formula[†] is an intriguing expression for the Bayes expectation of μ given z. [†5]

Theorem 20.3 *In model* (20.35)*, the posterior expectation of* μ *having observed* z *is*

$$E\{\mu|z\} = z + \sigma^2 l'(z) \qquad \textit{with } l'(z) = \frac{d}{dz}\log f(z). \tag{20.37}$$

The especially convenient feature of Tweedie's formula is that $E\{\mu|z\}$ is expressed directly in terms of the marginal density $f(z)$. This is a setup for empirical Bayes estimation. We don't know $g(\mu)$, but in large-scale situations we can estimate the marginal density $f(z)$ from the observations $\boldsymbol{z} = (z_1, z_2, \ldots, z_N)$, perhaps by Poisson regression as in Table 15.1, yielding

$$\hat{E}\{\mu_i|z_i\} = z_i + \sigma^2 \hat{l}'(z_i) \qquad \text{with } \hat{l}'(z) = \frac{d}{dz}\log \hat{f}(z). \tag{20.38}$$

The solid curve in Figure 20.7 shows $\hat{E}\{\mu|z\}$ for the prostate study data,

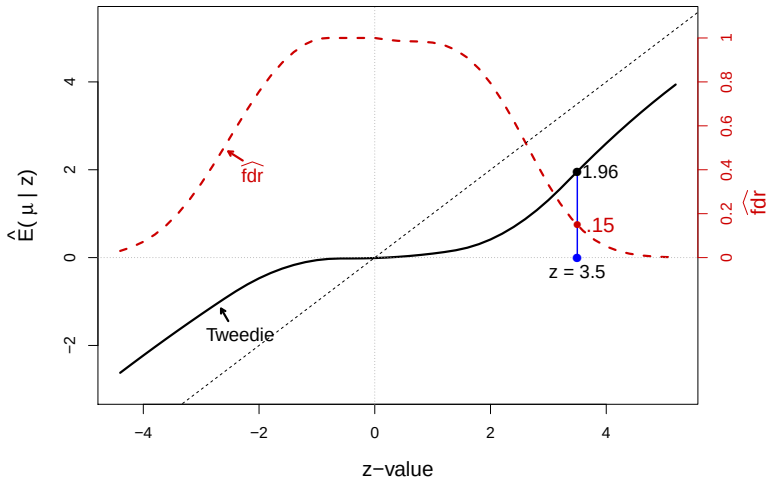

Figure 20.7 The solid curve is Tweedie's estimate $\hat{E}\{\mu|z\}$ (20.38) for the `prostate` study data. The dashed line shows the local false-discovery rate $\widehat{\text{fdr}}(z)$ from Figure 15.5 (red scale on right). At $z = 3.5$, $\hat{E}\{\mu|z\} = 1.96$ and $\widehat{\text{fdr}}(z) = 0.15$. For gene 610, with $z_{610} = 5.29$, Tweedie's estimate is 4.03.

with $\sigma^2 = 1$ and $\hat{f}(z)$ obtained using fourth-degree log polynomial regression as in Section 15.4. The curve has $E\{\mu|z\}$ hovering near zero for $|z_i| \leq 2$, agreeing with the local false-discovery rate curve $\widehat{\text{fdr}}(z)$ of Figure 15.5 that says these are mostly null genes.

$\hat{E}\{\mu|z\}$ increases for $z > 2$, equaling 1.96 for $z = 3.5$. At that point $\widehat{\text{fdr}}(z) = 0.15$. So even though $z_i = 3.5$ has a one-sided p-value of 0.0002, with 6033 genes to consider at once, it still is not a sure thing that gene i is non-null. About 85% of the genes with z_i near 3.5 will be non-null, and these will have effect sizes averaging about 2.31 ($= 1.96/0.85$). All of this nicely illustrates the combination of frequentist and Bayesian inference possible in large-scale studies, and also the combination of estimation and hypothesis-testing ideas in play.

If the prior density $g(\mu)$ in (20.35) is assumed to be normal, Tweedie's formula (20.38) gives (almost) the James–Stein estimator (7.13). The corresponding curve in Figure 20.7 in that case would be a straight line passing through the origin at slope 0.22. Like the James–Stein estimator, ridge regression, and the lasso of Chapter 16, Tweedie's formula is a shrinkage estimator. For $z_{610} = 5.29$, the most extreme observation, it gave

$\hat{\mu}_{629} = 4.03$, shrinking the maximum likelihood estimate more than one σ unit toward the origin.

Bayes estimators are immune to selection bias, as discussed in Sections 3.3 and 3.4. This offers some hope that Tweedie's empirical Bayes estimates might be a realistic cure for the winners' curse. A small simulation experiment was run as a test.

- A hundred data sets $\boldsymbol{z}$, each of length $N = 1000$, were generated according to a combination of exponential and normal sampling,

$$\mu_i \overset{\text{ind}}{\sim} e^{-\mu} \quad (\mu > 0) \quad \text{and} \quad z_i|\mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1), \tag{20.39}$$

  for $i = 1, 2, \ldots, 1000$.
- For each $\boldsymbol{z}$, $\hat{l}(z)$ was computed as in Section 15.4, now using a natural spline model with five degrees of freedom.
- This gave Tweedie's estimates

$$\hat{\mu}_i = z_i + \hat{l}'(z_i), \qquad i = 1, 2, \ldots, 1000, \tag{20.40}$$

  for that data set $\boldsymbol{z}$.
- For each data set $\boldsymbol{z}$, the 20 largest z_i values and the corresponding $\hat{\mu}_i$ and μ_i values were recorded, yielding the

$$\begin{aligned} &\textit{uncorrected differences} \quad z_i - \mu_i \\ \text{and} \quad &\textit{corrected differences} \quad \hat{\mu}_i - \mu_i, \end{aligned} \tag{20.41}$$

  the hope being that empirical Bayes shrinkage would correct the selection bias in the z_i values.
- Figure 20.8 shows the 2000 (100 data sets, 20 top cases each) uncorrected and corrected differences. Selection bias is quite obvious, with the uncorrected differences shifted one unit to the right of zero. In this case at least, the empirical Bayes corrections have worked well, the corrected differences being nicely centered at zero. Bias correction often adds variance, but in this case it hasn't.

Finally, it is worth saying that the "empirical" part of empirical Bayes is less the estimation of Bayesian rules from the aggregate data than the application of such rules to individual cases. For the prostate data we began with no definite prior opinions but arrived at strong (i.e., *not* "uninformative") Bayesian conclusions for, say, μ_{610} in the prostate study.

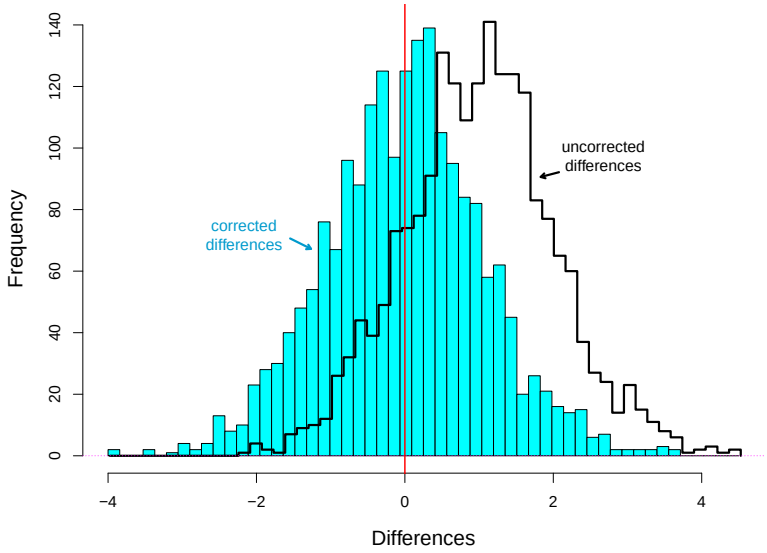

Figure 20.8 Corrected and uncorrected differences for 20 top cases in each of 100 simulations (20.39)–(20.41). Tweedie corrections effectively counteracted selection bias.

20.4 Combined Bayes–Frequentist Estimation

As mentioned previously, Bayes estimates are, at least theoretically, immune from selection bias. Let $\boldsymbol{z} = (z_1, z_2, \ldots, z_N)$ represent the prostate study data of the previous section, with parameter vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)$. Bayes' rule (3.5)

$$g(\boldsymbol{\mu}|\boldsymbol{z}) = g(\boldsymbol{\mu})\, f_{\boldsymbol{\mu}}(\boldsymbol{z})/f(\boldsymbol{z}) \tag{20.42}$$

yields the posterior density of $\boldsymbol{\mu}$ given $\boldsymbol{z}$. A data-based model selection rule such as "estimate the μ_i corresponding to the largest observation z_i" has no effect on the likelihood function $f_{\boldsymbol{\mu}}(\boldsymbol{z})$ (with $\boldsymbol{z}$ fixed) or on $g(\boldsymbol{\mu}|\boldsymbol{z})$. Having chosen a prior $g(\boldsymbol{\mu})$, our posterior estimate of μ_{610} is unaffected by the fact that $z_{610} = 5.29$ happens to be largest.

This same argument applies just as well to any data-based model selection procedure, for instance a preliminary screening of possible variables to include in a regression analysis—the C_p choice of a cubic regression in Figure 20.1 having no effect on its Bayes posterior accuracy.

There is a catch: the chosen prior $g(\boldsymbol{\mu})$ must apply to the entire parameter vector $\boldsymbol{\mu}$ and not just the part we are interested in (e.g., μ_{610}). This is

feasible in one-parameter situations like the stopping rule example of Figure 3.3. It becomes difficult and possibly dangerous in higher dimensions. Empirical Bayes methods such as Tweedie's rule can be thought of as allowing the data vector $\boldsymbol{z}$ to assist in the choice of a high-dimensional prior, an effective collaboration between Bayesian and frequentist methodology.

Our chapter's final vignette concerns another Bayes–frequentist estimation technique. Dropping the boldface notation, suppose that $\mathcal{F} = \{f_\alpha(x)\}$ is a multi-dimensional family of densities (5.1) (now with α playing the role of μ), and that we are interested in estimating a particular parameter $\theta = t(\alpha)$. A prior $g(\alpha)$ has been chosen, yielding posterior expectation

$$\hat{\theta} = E\{t(\alpha)|x\}. \tag{20.43}$$

How accurate is $\hat{\theta}$? The usual answer would be calculated from the posterior distribution of θ given x. This is obviously the correct answer if $g(\alpha)$ is based on genuine prior experience. Most often though, and especially in high-dimensional problems, the prior reflects mathematical convenience and a desire to be uninformative, as in Chapter 13. There is a danger of circular reasoning in using a self-selected prior distribution to calculate the accuracy of its own estimator.

An alternate approach, discussed next, is to calculate the *frequentist* accuracy of $\hat{\theta}$; that is, even though (20.43) is a Bayes estimate, we consider $\hat{\theta}$ simply as a function of x, and compute its frequentist variability. The next theorem leads to a computationally efficient way of doing so. (The Bayes and frequentist standard errors for $\hat{\theta}$ operate in conceptually orthogonal directions as pictured in Figure 3.5. Here we are supposing that the prior $g(\cdot)$ is unavailable or uncertain, forcing more attention on frequentist calculations.)

For convenience, we will take the family $\mathcal{F}$ to be a p-parameter exponential family (5.50),

$$f_\alpha(x) = e^{\alpha' x - \psi(\alpha)} f_0(x), \tag{20.44}$$

now with α being the parameter vector called μ above. The $p \times p$ covariance matrix of x (5.59) is denoted

$$\boldsymbol{V}_\alpha = \mathrm{cov}_\alpha(x). \tag{20.45}$$

Let Cov_x indicate the posterior covariance given x between $\theta = t(\alpha)$, the parameter of interest, and α,

$$\mathrm{Cov}_x = \mathrm{cov}\{\alpha, t(\alpha)|x\}, \tag{20.46}$$

a $p \times 1$ vector. Cov_x leads directly to a frequentist estimate of accuracy for $\hat{\theta}$.

†6 **Theorem 20.4** † *The delta method estimate of standard error for* $\hat{\theta} = E\{t(\alpha)|x\}$ (10.41) *is*

$$\widehat{\text{se}}_{\text{delta}}\left\{\hat{\theta}\right\} = \left(\text{Cov}_x' \, \boldsymbol{V}_{\hat{\alpha}} \, \text{Cov}_x\right)^{1/2}, \tag{20.47}$$

where $\boldsymbol{V}_{\hat{\alpha}}$ *is* $\boldsymbol{V}_\alpha$ *evaluated at the MLE* $\hat{\alpha}$.

The theorem allows us to calculate the frequentist accuracy estimate $\widehat{\text{se}}_{\text{delta}}\{\hat{\theta}\}$ with hardly any additional computational effort beyond that required for $\hat{\theta}$ itself. Suppose we have used an MCMC or Gibbs sampling algorithm, Section 13.4, to generate a sample from the Bayes posterior distribution of α given x,

$$\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(B)}. \tag{20.48}$$

These yield the usual estimate for $E\{t(\alpha)|x\}$,

$$\hat{\theta} = \frac{1}{B}\sum_{b=1}^{B} t\left(\alpha^{(b)}\right). \tag{20.49}$$

They also give a similar expression for $\text{cov}\{\alpha, t(\alpha)|x\}$,

$$\text{Cov}_x = \frac{1}{B}\sum_{b=1}^{B}\left(\alpha^{(b)} - \alpha^{(\cdot)}\right)\left(t^{(b)} - t^{(\cdot)}\right), \tag{20.50}$$

$t^{(b)} = t(\alpha^{(b)})$, $t^{(\cdot)} = \sum_b t^{(b)}/B$, and $\alpha^{(\cdot)} = \sum_b \alpha^{(b)}/B$, from which we can calculate[5] $\widehat{\text{se}}_{\text{delta}}(\hat{\theta})$ (20.47).

For an example of Theorem 20.4 in action we consider the `diabetes` data of Section 20.1, with x_i' the ith row of $\boldsymbol{X}$, the 442×10 matrix of prediction, so x_i is the vector of 10 predictors for patient i. The response vector $\boldsymbol{y}$ of progression scores has now been rescaled to have $\sigma^2 = 1$ in the normal regression model,[6]

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\beta, \boldsymbol{I}). \tag{20.51}$$

The prior distribution $g(\beta)$ was taken to be

$$g(\beta) = ce^{-\lambda\|\beta\|_1}, \tag{20.52}$$

[5] $\boldsymbol{V}_{\hat{\alpha}}$ may be known theoretically, calculated by numerical differentiation in (5.57), or obtained from parametric bootstrap resampling—taking the empirical covariance matrix of bootstrap replications $\hat{\beta}_i^*$.

[6] By dividing the original data vector $\boldsymbol{y}$ by its estimated standard error from the linear model $E\{\boldsymbol{y}\} = \boldsymbol{X}\beta$.

with $\lambda = 0.37$ and c the constant that makes $g(\beta)$ integrate to 1. This is the "Bayesian lasso prior,"† so called because of its connection to the lasso, (7.42) and (16.1). (The lasso plays no part in what follows). †7

An MCMC algorithm generated $B = 10{,}000$ samples (20.48) from the posterior distribution $g(\beta|\boldsymbol{y})$. Let

$$\theta_i = x_i'\beta, \tag{20.53}$$

the (unknown) expectation of the ith patient's response y_i. The Bayes posterior expectation of θ_i is

$$\hat{\theta}_i = \frac{1}{B}\sum_{b=1}^{B} x_i'\beta. \tag{20.54}$$

It has Bayes posterior standard error

$$\widehat{\text{se}}_{\text{Bayes}}\left(\hat{\theta}_i\right) = \left[\frac{1}{B}\sum_{b=1}^{B}\left(x_i'\beta^{(b)} - \hat{\theta}_i\right)^2\right]^{1/2}, \tag{20.55}$$

which we can compare with $\widehat{\text{se}}_{\text{delta}}(\hat{\theta}_i)$, the frequentist standard error (20.47).

Figure 20.9 shows the 10,000 MCMC replications $\hat{\theta}_i^{(b)} = x_i'\beta$ for patient $i = 322$. The point estimate $\hat{\theta}_i$ equaled 2.41, with Bayes and frequentist standard error estimates

$$\widehat{\text{se}}_{\text{Bayes}} = 0.203 \quad \text{and} \quad \widehat{\text{se}}_{\text{delta}} = 0.186. \tag{20.56}$$

The frequentist standard error is 9% smaller in this case; $\widehat{\text{se}}_{\text{delta}}$ was less than $\widehat{\text{se}}_{\text{Bayes}}$ for all 442 patients, the difference averaging a modest 5%.

Things can work out differently. Suppose we are interested in the posterior cdf of θ_{332} given $\boldsymbol{y}$. For any given value of c let

$$t\left(c, \beta^{(b)}\right) = \begin{cases} 1 & \text{if } x_{322}'\beta^{(b)} \le c \\ 0 & \text{if } x_{332}'\beta^{(b)} > c, \end{cases} \tag{20.57}$$

so

$$\text{cdf}(c) = \frac{1}{B}\sum_{b=1}^{B} t\left(c, \beta^{(b)}\right) \tag{20.58}$$

is our MCMC assessment of $\Pr\{\theta_{322} \le c|\boldsymbol{y}\}$. The solid curve in Figure 20.10 graphs cdf(c).

If we believe prior (20.52) then the curve *exactly* represents the posterior distribution of θ_{322} given $\boldsymbol{y}$ (except for the simulation error due to stopping at $B = 10{,}000$ replications). Whether or not we believe the prior we can use

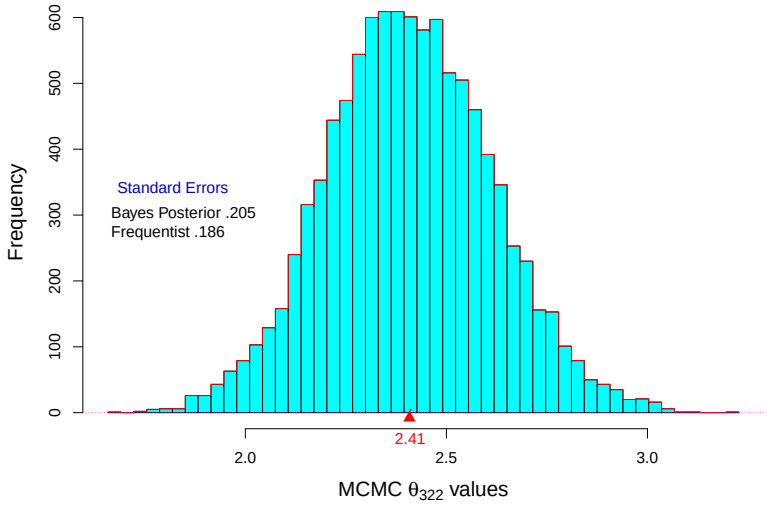

Figure 20.9 A histogram of 10,000 MCMC replications for posterior distribution of θ_{322}, expected progression for patient 322 in the `diabetes` study; model (20.51) and prior (20.52). The Bayes posterior expectation is 2.41. Frequentist standard error (20.47) for $\hat{\theta}_{322} = 2.41$ was 9% smaller than Bayes posterior standard error (20.55).

Theorem 20.4, with $t^{(b)} = t(c, \beta^{(b)})$ in (20.50), to evaluate the frequentist accuracy of the curve.

The dashed vertical red lines show cdf(c) plus or minus one $\widehat{\text{se}}_{\text{delta}}$ unit. The standard errors are disturbingly large, for instance 0.687 ± 0.325 at $c = 2.5$. The central 90% credible interval for θ_{322} (the c-values between cdf(c) 0.05 and 0.95),

$$(2.08, 2.73) \tag{20.59}$$

has frequentist standard errors about 0.185 for each endpoint—28% of the interval's length.

If we believe prior (20.52) then $(2.08, 2.73)$ is an (almost) exact 90% credible interval for θ_{322}, and moreover is immune to any selection bias involved in our focus on θ_{322}. If not, the large frequentist standard errors are a reminder that calculation (20.59) might turn out much differently in a new version of the diabetes study, even ignoring selection bias.

To return to our main theme, Bayesian calculations encourage a disregard for model selection effects. This can be dangerous in objective Bayes

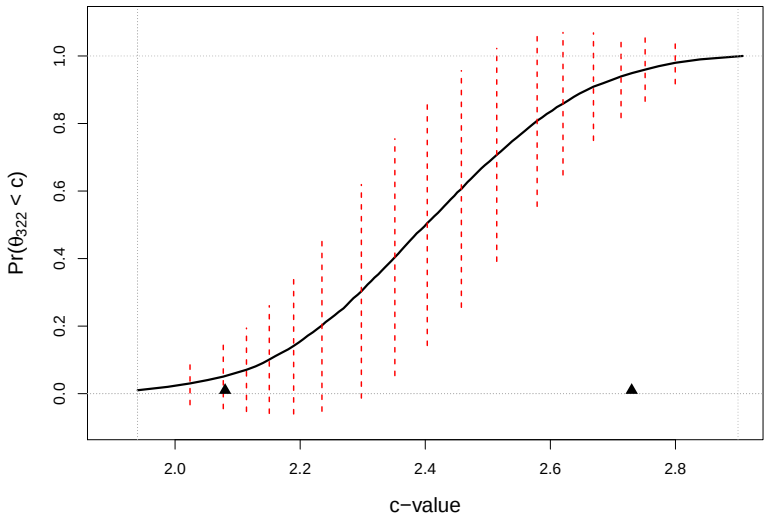

Figure 20.10 The solid curve is the posterior cdf of θ_{322}. Vertical red bars indicate $\pm$ one frequentist standard error, as obtained from Theorem 20.4. Black triangles are endpoints of the 0.90 central credible interval.

settings where one can't rely on genuine prior experience. Theorem 20.4 serves as a frequentist checkpoint, offering some reassurance as in Figure 20.9, or sounding a warning as in Figure 20.10.

20.5 Notes and Details

Optimality theories—statements of best possible results—are marks of maturity in applied mathematics. Classical statistics achieved two such theories: for unbiased or asymptotically unbiased estimation, and for hypothesis testing. Most of this book and all of this chapter venture beyond these safe havens. How far from *best* are the C_p/OLS bootstrap smoothed estimates of Section 20.2? At this time we can't answer such questions, though we can offer appealing methodologies in their pursuit, a few of which have been highlighted here.

The cholestyramine example comes from Efron and Feldman (1991) where it is discussed at length. Data for a control group is also analyzed there.

†1 [p. 398] *Scheffé intervals.* Scheffé's 1953 paper came at the beginning

of a period of healthy development in simultaneous inference techniques, mostly in classical normal theory frameworks. Miller (1981) gives a clear and thorough summary. The 1980s followed with a more computer-intensive approach, nicely developed in Westfall and Young's 1993 book, leading up to Benjamini and Hochberg's 1995 false-discovery rate paper (Chapter 15 here), and Benjamini and Yekutieli's (2005) false-coverage rate algorithm.

Scheffé's construction (20.15) is derived by transforming (20.6) to the case $\boldsymbol{V} = \boldsymbol{I}$ using the inverse square root of matrix $\boldsymbol{V}$,

$$\hat{\gamma} = \boldsymbol{V}^{-1/2}\hat{\beta} \quad \text{and} \quad \gamma = \boldsymbol{V}^{-1/2}\beta \tag{20.60}$$

($(\boldsymbol{V}^{-1/2})^2 = \boldsymbol{V}^{-1}$), which makes the ellipsoid of Figure 20.2 into a circle. Then $Q = \|\hat{\gamma} - \gamma\|^2/\hat{\sigma}^2$ in (20.10), and for a linear combination $\gamma_d = d'\gamma$ it is straightforward to see that $\Pr\{Q \le k_{p,q}^{(\alpha)^2}\} = \alpha$ amounts to

$$\gamma_d \in \hat{\gamma}_d \pm \hat{\sigma}\,\|d\|\,k_{p,q}^{(\alpha)} \tag{20.61}$$

for all choices of d, the geometry of Figure 20.2 now being transparent. Changing coordinates back to $\hat{\beta} = \boldsymbol{V}^{1/2}\hat{\gamma}$, $\beta = \boldsymbol{V}^{1/2}\gamma$, and $c = \boldsymbol{V}^{1/2}d$ yields (20.15).

†2 [p. 399] *Restricting the catalog* $\boldsymbol{C}$. Suppose that all the sample sizes n_j in (20.16) take the same value n, and that we wish to set simultaneous confidence intervals for all pairwise differences $\beta_i - \beta_j$. Tukey's *studentized range* pivotal quantity (1952, unpublished)

$$T = \max_{i \neq j} \frac{\left|\left(\hat{\beta}_i - \hat{\beta}_j\right) - (\beta_i - \beta_j)\right|}{\hat{\sigma}} \tag{20.62}$$

has a distribution not depending on σ or $\boldsymbol{\beta}$. This implies that

$$\beta_i - \beta_j \in \hat{\beta}_i - \hat{\beta}_j \pm \frac{\hat{\sigma}}{\sqrt{n}}\,T^{(\alpha)} \tag{20.63}$$

is a set of simultaneous level-α confidence intervals for all pairwise differences $\beta_i - \beta_j$, where $T^{(\alpha)}$ is the αth quantile of T. (The factor $1/\sqrt{n}$ comes from $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2/n)$ in (20.16).)

Table 20.4 *Half-width of Tukey studentized range simultaneous 95% confidence intervals for pairwise differences $\beta_i - \beta_j$ (in units of $\hat{\sigma}/\sqrt{n}$) for $p = 2, 3, \ldots, 6$ and $n = 20$; compared with Scheffé intervals* (20.15).

p	2	3	4	5	6
Tukey	2.95	3.58	3.96	4.23	4.44
Scheffé	3.74	4.31	4.79	5.21	5.58

Reducing the catalog $\boldsymbol{C}$ from all linear combinations $c'\beta$ to only pairwise differences shortens the simultaneous intervals. Table 20.4 shows the comparison between the Tukey and Scheffé 95% intervals for $p = 2, 3, \ldots, 6$ and $n = 20$.

Calculating $T^{(\alpha)}$ was a substantial project in the early 1980s. Berk *et al.* (2013) now carry out the analogous computations for general catalogs of linear constraints. They discuss at length the inferential basis of such procedures.

†3 [p. 405] *Discontinuous estimators.* Looking at Figure 20.6 suggests that a confidence interval for $\theta_{-2.0}\, t(c, \boldsymbol{x})$ will move far left for data sets x where C_p selects linear regression ($m = 1$) as best. This kind of "jumpy" behavior lengthens the intervals needed to attain a desired coverage level. More seriously, intervals for $m = 1$ may give misleading inferences, another example of "accurate but incorrect" behavior. Bagging (20.28), in addition to reducing interval length, improves inferential correctness, as discussed in Efron (2014a).

†4 [p. 406] *Theorem 20.2 and its corollary.* Theorem 20.2 is proved in Section 3 of Efron (2014a), with a parametric bootstrap version appearing in Section 4. The corollary is a projection result illustrated in Figure 4 of that paper: let $\mathcal{L}(\boldsymbol{N})$ be the n-dimensional subspace of B-dimensional Euclidean space spanned by the columns of the $B \times n$ matrix (N_{bj}) (20.29) and $\boldsymbol{t}^*$ the B-vector with components $t^{*b} - t^{*\cdot}$; then

$$\widehat{\mathrm{se}}_{\mathrm{IJ}}(s) \big/ \widehat{\mathrm{se}}_{\mathrm{boot}}(t) = \left\| \hat{\boldsymbol{t}}^* \right\| \big/ \|\boldsymbol{t}^*\|, \tag{20.64}$$

where $\hat{\boldsymbol{t}}^*$ is the projection of $\boldsymbol{t}^*$ into $\mathcal{L}(\boldsymbol{N})$. In the language of Section 10.3, if $\hat{\theta}^* = S(\boldsymbol{P})$ is very nonlinear as a function of $\boldsymbol{P}$, then the ratio in (20.64) will be substantially less than 1.

†5 [p. 409] *Tweedie's formula.* For convenience, take $\sigma^2 = 1$ in (20.35). Bayes' rule (3.5) can then be arranged to give

$$g(\mu|z) = e^{\mu z - \psi(z)} g(\mu) e^{-\frac{1}{2}\mu^2} \big/ \sqrt{2\pi} \tag{20.65}$$

with

$$\psi(z) = \frac{1}{2} z + \log f(z). \tag{20.66}$$

This is a one-parameter exponential family (5.46) having natural parameter α equal to z. Differentiating ψ as in (5.55) gives

$$E\{\mu|z\} = \frac{d\psi}{dz} = z + \frac{d}{dz} \log f(z), \tag{20.67}$$

which is Tweedie's formula (20.37) when $\sigma^2 = 1$. The formula first appears in Robbins (1956), who credits it to a personal communication from M. K. Tweedie. Efron (2011) discusses general exponential family versions of Tweedie's formula, and its application to selection bias situations.

†6 [p. 414] *Theorem 20.4.* The delta method standard error approximation for a statistic $T(x)$ is

$$\widehat{\text{se}}_{\text{delta}} = \left[(\nabla T(x))' \, \hat{V} \, (\nabla T(x)) \right]^{1/2}, \tag{20.68}$$

where $\nabla T(x)$ is the gradient vector $(\partial T/\partial x_j)$ and $\hat{V}$ is an estimate of the covariance matrix of x. Other names include the "Taylor series method," as in (2.10), and "propagation of errors" in the physical sciences literature. The proof of Theorem 20.4 in Section 2 of Efron (2015) consists of showing that $\text{Cov}_x = \nabla T(x)$ when $T(x) = E\{t(\alpha)|x\}$. Standard deviations are only a first step in assessing the frequentist accuracy of $T(x)$. The paper goes on to show how Theorem 20.4 can be improved to give confidence intervals, correcting the impression in Figure 20.10 that cdf(c) can range outside $[0, 1]$.

†7 [p. 415] *Bayesian lasso.* Applying Bayes' rule (3.5) with density (20.51) and prior (20.52) gives

$$\log g(\beta|\boldsymbol{y}) = -\left\{ \frac{\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2}{2} + \lambda \|\beta\|_1 \right\}, \tag{20.69}$$

as discussed in Tibshirani (2006). Comparison with (7.42) shows that the maximizing value of β (the "MAP" estimate) agrees with the lasso estimate. Park and Casella (2008) named the "Bayesian lasso" and suggested an appropriate MCMC algorithm. Their choice $\lambda = 0.37$ was based on marginal maximum likelihood calculations, giving their analysis an empirical Bayes aspect ignored in their and our analyses.