Empirical Bayes Crop Prediction

George Casella
Laura Strang

Cornell University

BU-953-M

December, 1987
March, 1988

Research supported by Hatch Grant Number NYS-151402.

Summary

Empirical Bayes methods can provide improved estimation and prediction in problems that can be modeled using underlying similarity between seemingly different problems. Many agricultural problems, in which similar experiments are done in different regions, can be modeled with empirical Bayes techniques. The purpose of this paper is to illustrate a simple empirical Bayes analysis of a crop yield prediction problem.

1. Introduction.

The prediction of yield, in tons per acre, is of major concern in the grape industry. The amount of tonnage produced will play an important role in determining to what uses the grapes will be put (grape juice, jelly, frozen products, etc.). A reasonable predictor of this yield is a count of the number of grape clusters appearing in June preceding the harvest. linear regressions of tons/acre on these June cluster counts show good predictive power. However, use of empirical Bayes methods can further improve upon these predictions. It is the purpose of this paper to explain and illustrate this empirical Bayes improvement.

Data obtained are tons/acre and June cluster counts of concord grapes grown in six geographic areas in the United States, (and are given in Table 1 . Standard regression practice would be to compute six different regressions (one for each area), and to predict yield independently in each of the six areas. Empirical Bayes methods, however, allow one to take advantage of any similarities among the six regions, and to use these similarities in improving the predictions. The empirical Bayes model does not view this situation as six separate regressions, but rather as six realizations of a possibly similar regression problem.

This paper is organized as follows. In Section 2, an introduction to empirical Bayes methods is given, and in Section 3, these methods are applied to model the crop yield data. Section 4 contains details about the numerical results. In particular, there is the standard regression analysis, the empirical Bayes analysis, and a comparison of the results. Section 5 contains a general discussion about the empirical Bayes methods shown here, and other, more general methods.

2. Empirical Bayes Methods Illustrated.

When a number of similar problems are to be analyzed, it is often possible to improve estimation in each problem by taking advantage of the common structure.

One of the simplest settings for an empirical Bayes analysis, one that will serve as a starting point for the regression problem, is the following: Suppose that we have k problems in which we are to estimate the mean, θ_i, of a population, using a sample mean $\bar{Y}_i$ based on n observations, where we assume that

$$\bar{Y}_i \sim \text{normal } (\theta_i, \sigma^2/n), \qquad i = 1,\cdots,k \tag{2.1}$$

where σ^2 is known. (This is the common one-way analysis of variance.) The usual estimator of θ_i is $\bar{Y}_i$ and, although this estimator has many optimality properties, it can be improved upon.

Since we believe that the problems are related, we can use this information in a hierarchical (or Bayesian) model, and assume that the θ_i's themselves come from a common population, that is

$$\theta_i \sim \text{normal } (\mu, \tau^2), \qquad i = 1,\cdots,k\ . \tag{2.2}$$

A strict Bayesian model would also assume that μ and τ^2 are known, an assumption that we make for the moment.

Doing a Bayesian analysis of (2.1) and (2.2), as, for example, found in Berger (1985), we would calculate the Bayes estimator of θ_i as

$$\hat{\theta}_i^B = \left(\frac{\sigma^2}{\sigma^2+n\tau^2}\right)\mu + \left(\frac{n\tau^2}{\sigma^2+n\tau^2}\right)\bar{Y}_i \tag{2.3a}$$

$$= \mu + \left(1 - \frac{\sigma^2}{\sigma^2+n\tau^2}\right)(\bar{Y}_i - \mu)\ , \tag{2.3b}$$

where (2.3b) follows from (2.3a) by algebraic manipulations. Notice that $\hat{\theta}_i^B$ is a weighted average of the prior (common) estimate of θ_i, μ, and the sample estimate, $\bar{Y}_i$. The weights used in the average are dependent on the respective variances, and they work in a manner that gives the most weight to the estimate with the smallest variance. An empirical Bayes estimator is derived from a model similar to the Bayes model, but the parameter values in the prior distribution are not specified. Instead, they are estimated using the marginal distribution of the $\bar{Y}_i$s.

Using (2.1) and (2.2), a standard calculation will show that the marginal distribution of $\bar{Y}_i$ (unconditional on θ_i) is given by

$$\bar{Y}_i \sim \text{normal}\ (\mu,\ \sigma^2/n + \tau^2) \qquad i = 1,\cdots,k\ . \tag{2.4}$$

Thus, marginally, the $\bar{Y}_i$s are identically distributed. Based on the model in (2.4):

$$\mu \text{ is estimated by } \bar{\bar{Y}} = \frac{1}{k}\sum_{i=1}^{k} \bar{Y}_i \tag{2.5}$$

$$\sigma^2/n+\tau^2 \text{ is estimated by } S^2 = \sum_{i=1}^{k} (\bar{Y}_i-\bar{\bar{Y}})^2\ .$$

Furthermore, we use the facts that, if we take expected values according to (2.4), we have

$$E[\bar{\bar{Y}}] = \mu$$

$$E\left[\frac{(k-3)\sigma^2/n}{S^2}\right] = \frac{\sigma^2}{\sigma^2+n\tau^2}\ . \tag{2.6}$$

Combining (2.6) and (2.3b) gives an empirical Bayes estimator of θ_i:

$$\hat{\theta}_i^{EB} = \bar{\bar{Y}} + \left[1 - \frac{(k-3)\sigma^2/n}{S^2}\right]\left(\bar{Y}_i - \bar{\bar{Y}}\right)\ . \tag{2.7}$$

We make two additional modifications to (2.7) to arrive at a final version of an empirical Bayes estimator. First, it has to be demonstrated (Efron and Morris, 1972), that (2.7) is uniformly improved upon if the quantity in square brackets is never allowed to be negative. (This keeps the empirical Bayes estimator between $\bar{\bar{Y}}$ and $\bar{Y}_i$.) We do this with the positive-part function, defined by $[a]^+ = \text{maximum } \{0,a\}$.

The second modification has to do with σ^2. In practice, this quantity is unknown, but can be estimated, in the i^{th} problem, by (say) s_i^2, the sample variance. Since the k problems are modeled with a common variance, we estimate σ^2 with a pooled estimator

$$s^2 = \frac{1}{k} \sum_{i=1}^{k} s_i^2 .$$

With these two modifications, we use (2.7) to arrive at an empirical Bayes estimator for this problem:

$$\hat{\theta}_i^{EB} = \bar{\bar{Y}} + \left[1 - \frac{(k-3)s^2}{nS^2}\right]^+ (\bar{Y}_i - \bar{\bar{Y}}) \tag{2.8}$$

The fact that $\hat{\theta}_i^{EB}$ of (2.8) is a better estimator than $\bar{Y}_i$ has been the topic of an enormous amount of research, dating from Lindley's (1962) discussion of Stein's (1962) paper, with one of the more recent contributions being Casella and Hwang (1987), who also investigate the confidence interval question, and find that domination also obtains in that case. For a more intuitive discussion of empirical Bayes estimators, along with some simple illustrations, see Casella (1985).

One bit of intuition that can be particularly helpful in understanding empirical Bayes estimators, and is reasonably analogous to the above model in the case of homoscedasticity, is the following. Starting with the model (2.1), we may hypothesize that the θ_is are equal, i.e., that the hypothesis $H: \theta_1 = \theta_2 = \cdots = \theta_k$ is true. If so, we would estimate θ_i with $\bar{\bar{Y}}$, as given in (2.5). The estimator (2.8) uses this information, in that it "pulls" $\bar{Y}_i$ toward $\bar{\bar{Y}}$, with a weight dependent on $S^2 = \sum (\bar{Y}_i - \bar{\bar{Y}})^2$, a quantity that measures the plausibility of H. Notice that this reasoning is quite different from classical reasoning in which we want to reject a hypothesis. Here we are looking to hypothesize a model that we believe is true.

3. An Empirical Bayes Regression Model.

An empirical Bayes regression model follows details similar to those outlined in Section 2, but necessarily has a few more complications.

We assume that there are k regression problems

$$Y_{ij} = \alpha_i + \beta_i x_{ij} + \varepsilon_{ij} \qquad \begin{matrix} j = 1,\cdots,n_i \\ i = 1,\cdots,k \end{matrix} \tag{3.1}$$

where, for each i, $\varepsilon_{ij} \sim$ normal $(0,\sigma^2)$, x_{ij}, $j = 1,\cdots,n_i$ are fixed constants, and α_i and β_i are unknown parameters.

We now use a hierarchical model that reflects the crop yield problem to be modeled. It is thought that in the different growing region (regression problems) that the intercepts are all different but the slopes are similar. This can be seen in Figure 1, which shows a scatterplot of the data for all six areas. We can express this with the hypothesis:

$$H : \beta_1 = \cdots = \beta_k, \qquad \alpha_1,\cdots,\alpha_k \text{ unspecified.} \tag{3.2}$$

Alternatively, we can specify the distributional assumption

$$\beta_i \sim \text{normal } (\beta, \tau_\beta^2), \; i = 1,\cdots,k \tag{3.3}$$

Under the model (3.1) it is customary to estimate α_i and β_i with $\hat{\alpha}_i$ and $\hat{\beta}_i$, the least squares estimators, with sampling distributions.

$$\hat{\alpha}_i \sim n(\alpha_i, \sigma^2_{\alpha_i}) \;,\cdots,\; \hat{\beta}_i \sim n(\beta_i, \sigma^2_{\beta_i}) \;. \tag{3.4}$$

Since $\sigma^2_{\beta_i} = \sigma^2 / \sum_{j=1}^{n_i} (x_{ij}-x_i)^2$ we can, by suitable standardization of the x_{ij}s, assume that $\sigma^2_{\beta_i} = \sigma^2_\beta$ for all i. Following the development in Section 2, a Bayes estimator of β_i is given by

$$\hat{\beta}_i^B = \beta + \left(1 - \frac{\sigma^2_\beta}{\sigma^2_\beta + \tau^2_\beta} \right) (\hat{\beta}_i - \beta) \;. \tag{3.5}$$

Moreover, the marginal expectation for the Y_{ij}s (unconditional on the β_is) is given by

$$E(Y_{ij}) = \alpha_i + \beta x_{ij} \quad , \tag{3.6}$$

showing that β can be estimated by a regression over the entire data set using $\hat{\beta}$, where,

$$\hat{\beta} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{x}})(Y_{ij} - \bar{\bar{Y}}) \; / \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{x}})^2 \quad ,$$

where $\bar{\bar{x}}$ and $\bar{\bar{Y}}$ are the respective grand means. It only remains to estimate the term $\sigma_\beta^2 \; / \; (\sigma_\beta^2 + \tau_\beta^2)$ in (3.5), which can be done using the arguments from the preceding section, and results in the estimate $(k-3)s_\beta^2 \; / \; S_\beta^2$, where

$$s_\beta^2 = \sum_{i=1}^{k} (n_i - 2) s_{\beta_i}^2 \; / \sum_{i=1}^{k} (n_i - 2) \quad ,$$

the pooled sample variance of the β_is, and

$$S_\beta^2 = \sum_{i=1}^{k} n_i (\hat{\beta}_i - \hat{\beta})^2 \quad .$$

Combining all of this with (3.5) and (3.1), and, using the arguments of Section 2, the empirical Bayes regression equation is given by

$$\hat{Y}_{ij}^{EB} = \hat{\alpha}_i + \hat{\beta}_i^{EB} \, x_{ij} \qquad i=1,\cdots,k, \quad j = 1,\cdots,n_i \quad ,$$

where $\hat{\beta}_i^{EB}$ is given by

$$\hat{\beta}_i^{EB} = \hat{\beta} + \left[1 - \frac{(k-3)s_\beta^2}{S_\beta^2} \right]^+ (\hat{\beta}_i - \hat{\beta}) \quad . \tag{3.7}$$

4. Crop Yield Prediction.

Using the techniques outlined in Section 3, an empirical Bayes regression was fitted to the data of Table 1, with Y = yield in tons/acre, and x = June cluster count. In addition, a usual regression analysis, which did a least squares fit independently in each area, was also done. These analyses are compared in Tables 2, 3 and 4.

Tables 2 and 3 show the performance, of the least squares (LS) and empirical Bayes (EB) fits, for the years 1982 and 1983. Note that the predictions were performed as would be done in practice: The data up to year k-1 was used to fit the coefficients that predicted the yield in year k.

It is clear that empirical Bayes is no panacea. The 1982 EB predictions are worse than LS predictions. However, the 1983 EB predictions do prove to be slightly better than the 1983 LS predictions. In Table 4, the performance for six years is summarized, and it can be seen that the empirical Bayes predictions show a slight overall improvement over the least squares predictions.

5. Discussion.

The purpose of this paper was to demonstrate that a simple empirical Bayes model can provide prediction improvement in regression problems. It was seen that, in the crop yield data analyzed, empirical Bayes provided some improvement. It also should be recognized that empirical Bayes may not improve prediction in all cases: It will not cure all ills.

Empirical Bayes methods do, however, enjoy a number of overall optimality properties that will guarantee domination of least squares in the long run. (See, for example, Berger, 1985, or Casella and Hwang, 1987). Thus, in any problem, although EB may lose out to LS in a few instances, it can never lose out too often. The only qualification to this statement is that one must be interested in overall optimality, as we were interested in good predictions for all six regions. If one is only interested in one or two areas, empirical Bayes modeling will not provide any gains; in such cases an individual model is better.

We have not yet discussed the area of confidence intervals, but these important quantities deserve some mention. In the area of simultaneous inference, e.g., inference about all six areas, the EB model is superior. Taking the usual LS confidence intervals and recentering them at the EB estimates will provide a uniformly superior interval (in terms of coverage probability). Moreover, it is even possible to reduce the size of the EB intervals and still produce a superior interval. For further details about confidence procedures, see Morris (1983), Casella and Hwang (1987), or Laird and Louis (1987).

One final note. We have tried to illustrate a simple empirical Bayes model, but there are many more sophisticated EB models that can provide greater improvements than obtained here. For example, we assumed that the variances were equal in the six areas, which is quite a strong assumption.

There are methods for dealing with the unequal variance model, methods that will give greater improvement than obtained here. Moreover, we choose a relatively simple hypothesis, given in (3.2), to shrink toward. More careful modeling would lead to better hypotheses, possibly involving shrinkage of the α_is, and more effective EB estimators. For some illustrations of this, see the references in Casella (1985). Also, Morris (1983) contains applications and theoretical discussions.

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, second edition. Springer-Verlag, New York.

Casella, G. (1985). An Introduction to Empirical Bayes Data Analysis. *American Statistician* **39**, 83-87.

Casella, G., and Hwang, J. T. (1987). Employing Vague Prior Information in the Construction of Confidence Sets. *Journal of Multivariate Analysis* **21**, 79-104.

Efron, B., and Morris, C. N. (1972). Limiting the Risk of Bayes and Empirical Bayes Estimators, Part II: The Empirical Bayes Case. *Journal of the American Statistical Association* **67**, 130-139.

Laird, N. M., and Louis, T. A. (1987). Empirical Bayes Confidence Intervals Based on Bootstrap Samples (with discussion). *Journal of the American Statistical Association* **82**, 739-750.

Lindley, D. V. (1962). Discussion of Professor Stein's Paper. *Journal of the Royal Statistical Society, Series B* **24**, 265-296.

Morris, C. N. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association* **78**, 47-65.

Stein, C. (1962). Confidence Sets for the Mean of a Multivariate Normal Distribution (with discussion). *Journal of the Royal Statistical Society, Series B* **24**, 265-296.

Table 1. Raw Data

	Area 1		Area 2		Area 3	
Year	Tons/Acre	Clus.Count	Tons/Acre	Clus.Count	Tons/Acre	Clus.Count
71	5.6	116.37	5.3	114.44	4.9	107.83
72	2.6	102.89	3.0	106.95	2.9	114.02
73	3.2	82.77	3.5	96.93	3.3	99.81
74	4.5	110.68	4.9	114.54	4.1	116.08
75	4.2	97.50	4.5	120.10	4.4	114.51
76	5.2	115.88	5.4	131.71	5.2	98.53
77	2.7	80.19	2.8	76.82	2.5	59.67
78	4.8	125.24	5.2	143.06	4.5	102.81
79	4.9	116.15	4.8	141.37	4.9	118.14
80	4.7	117.36	4.9	141.26	5.0	108.04
81	4.1	93.31	5.6	130.49	4.1	112.75
82	4.4	107.46	4.4	121.04	3.8	126.04
83	5.4	122.30	6.8	149.10	4.7	123.92

	Area 4		Area 5		Area 6	
Year	Tons/Acre	Clus.Count	Tons/Acre	Clus.Count	Tons/Acre	Clus.Count
71	4.5	112.50	3.2	95.20	7.6	177.00
72	2.2	114.30	3.1	–	5.6	124.00
73	1.3	56.20	2.3	75.00	5.8	149.00
74	2.9	89.40	2.4	84.50	6.2	116.00
75	3.5	112.80	3.8	87.60	6.7	135.00
76	1.0	39.30	3.0	91.00	6.8	142.00
77	2.1	52.10	3.7	91.00	6.6	127.00
78	4.1	126.80	3.3	92.3	11.4	173.00
79	4.3	104.20	3.3	74.00	5.8	130.00
80	3.9	109.50	2.6	96.00	8.3	169.00
81	4.0	106.70	1.9	67.20	7.9	132.00
82	4.6	124.50	3.3	90.90	8.0	122.00
83	4.8	129.50	3.1	82.00	10.5	184.00

Table 2

1982 Predictions

(Models Use Data up to 1982)

Area	Actual Yield	LS Prediction	EB Prediction
1	4.4	4.5	4.6
2	4.4	4.7	5.2
3	3.8	5.1	5.5
4	4.6	4.5	4.9
5	1.9	2.4	2.7
6	8.0	6.0	5.6
Residual Root Mean Squared		1.006	1.307

Model Coefficients

Area	Intercept	LS Slope	EB Slope
1	- .93	.050	.052
2	.56	.034	.038
3	.69	.035	.038
4	- .42	.039	.042
5	- .05	.037	.040
6	-3.39	.077	.073

$$\hat{\beta}_i^{EB} = .057 + .829(\hat{\beta}_i - .057)$$

Table 3

1983 Predictions

(Models Use Data up to 1982)

Area	Actual Yield	LS Prediction	EB Prediction
1	5.4	5.2	5.4
2	6.8	5.6	6.3
3	4.7	4.7	5.4
4	4.8	4.7	5.1
5	3.3	3.3	3.5
6	10.5	10.5	10.2
Residual Root Mean Squared		.489	.422

Model Coefficients

Area	Intercept	LS Slope	EB Slope
1	- .93	.050	.052
2	.54	.034	.038
3	1.47	.026	.032
4	- .44	.040	.043
5	- .49	.041	.044
6	-1.95	.068	.066

$$\hat{\beta}_i^{EB} = .057 + .821(\hat{\beta}_i - .057) \qquad i=1,\cdots,6$$

Table 4

Residual Root Means Squared

Year	LS Predictions	EB Predictions
1978	1.58	1.16
1979	.62	.57
1980	.91	.65
1981	.76	.92
1982	1.01	1.31
1983	.49	.42
Overall	2.36	2.20

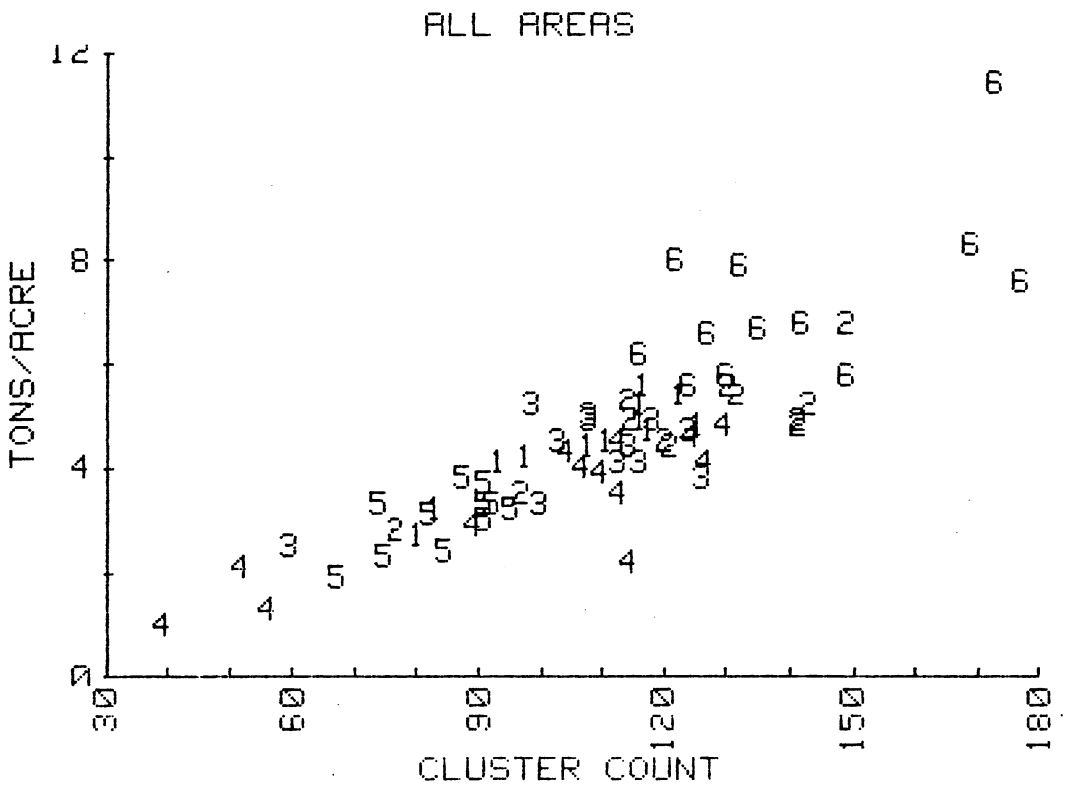

Figure 1. Scatterplot of data for all six areas. Plotting character identifies area.