Illustrating Empirical Bayes Methods

George Casella
Cornell University

BU-973-MB

July, 1991
Revised January, 1992

Research supported by National Science Foundation Grant No. DMS 91-00839 and National Security Agency Grant No. 90F-073.

Table of Contents

Abstract

Empirical Bayes methods have found increasing use in statistical analyses. These methods allow for modelling of complicated systems, and provide a mechanism for obtaining parameter estimates. They are based on Bayesian models, but employ alternate estimation techniques. In this article these methods are explained and illustrated with many examples taken from real situations.

1. What Is Empirical Bayes

Empirical Bayes (EB) is a term that has many meanings, reflecting different approaches to solving problems. It can describe a methodology for both estimation and inference, an important distinction. In a majority of applications, statistical inference tends to be made using frequentist (classical) statistics. This is the type of inference that is applicable to a "long-run" interpretation. That is, inference about a process is to a (usually imagined) sequence of replicated experiments. In contrast, the inference from Bayesian statistics is targeted to the data at hand. Each of these approaches has its strengths and weaknesses, and EB methods attempt to borrow the strengths from each approach.

The general EB approach can be pictured as a compromise between classical and Bayesian approaches, since EB methods sit "in between" classical and Bayesian statistics, borrowing pieces from each. Although this is an oversimplification, it does allow us to see where EB methods fit among more standard ones. The basic methodology is to model a situation using standard Bayesian techniques, and derive parameter estimates for quantities of interest. This uses one of the strengths of Bayesian statistics, the ability to derive estimates for all parameters of interest in a model. These Bayesian estimates, however, will usually contain some unknown quantities, parameters which the Bayesian specifies. This is where the empirical Bayesian separates from the Bayesian. Instead of specifying these unknown quantities, the empirical Bayesian estimates them, and substitutes these estimates into the Bayes quantities. Thus, the empirical Bayesian uses Bayes techniques for modelling, but alternate techniques for constructing estimates from these models. Then for the inferential stage of the analysis, EB can take many directions. Most often, the inference from an EB analysis will be a frequentist inference, in rare cases it can be a Bayesian inference. Morris [25] also defines an EB inference, which is, as expected, a hybrid of frequentist and Bayesian inference. We will see an example of this later.

Here we will mainly be concerned with describing how EB methodology is useful for deriving statistical procedures, and will not directly address their formal evaluation. It is fair to say that most good EB procedures perform well against most criteria, but of course, each case should be checked separately. The abundance of uses found by EB, however, is strong evidence that these procedures perform admirably.

Within EB methodology, the EB approach can be split into two distinct types. These are *parametric* and *nonparametric* EB. Nonparametric EB was the original EB formulation (Robbins [27]) However, in recent years the parametric formulation has found numerous applications, especially in small sample situations. Here we will concentrate on parametric EB techniques. Techniques of nonparametric EB, while quite powerful, are more suitable for large sample analyses.

This tutorial is composed of both theoretical explanations and illustrative examples. All examples have been taken from the literature, and illustrate EB analyses of real data sets. The tutorial

presents some basics underlying the EB approach in Sections 3 and 4, which should provide enough background to understand the principles behind the methodology. In Section 6 the theory is developed a bit further, allowing for a deeper understanding and appreciation of the later examples and methods. A shorter and much less detailed introduction to EB methodology is given in [6], which could serve as a companion to this tutorial.

2. Examples

Before plunging into the more technical aspects of EB estimation, two short examples are presented where EB methods can be extremely useful.

2.1 Interspecies Extrapolation

DuMouchel and Harris [12] investigated interspecies extrapolation of dose-response experiments. They were analyzing data on the effect of different environmental agents (e.g., engine emissions, cigarette smoke) on different species (e.g., mice, humans). Because each species was not exposed to each agent, they were particularly interested in models that would allow inter-species extrapolation. They used the model

$$y_{ij} = \theta_{ij} + \epsilon_{ij} \qquad i = 1, \ldots, k$$

$$\theta_{ij} = \mu + \alpha_i + \gamma_j + \delta_{ij}, \qquad j = 1, \ldots, n_i$$

where

y_{ij} = observed dose-response slope (log), of species i exposed to environmental agent j
θ_{ij} = true dose-response slopes
μ = overall mean
α_i = species-specific effect
γ_j = agent-specific effect

Notice that even though each (i,j) combination has its own dose-response slope (θ_{ij}), the θ_{ij}s are assumed to have a common (simpler) underlying structure. Using EB estimation methods, this model allows for interspecies extrapolation. For example, the effect of diesel engine emissions on humans can be estimated (even though such data do not exist) by extrapolating from the effect of diesel engine emissions on mice. We will return to this example later.

2.2 Selenium in non-fat milk powder

Eberhardt, et al. [13] describe statistical procedures for combining independent estimates of means, procedures that have use in the process of certifying Standard Reference Materials at the

National Bureau of Standards. For example, the concentration of selenium in non-fat milk powder is measured in four different ways, as shown in the following data set.

Table 2.1
Selenium in non-fat milk powder (units are ng/g)

Analytical method	n_i	$\bar{X}_i$	S_i
Atomic absorption spectrometry (hydride generation)	8	105.0	9.258
Neutron activation (instrumental)	12	109.75	4.555
Neutron activation (radiochemical)	14	109.5	1.652
Isotope dilution mass spectrometry (Spark source)	8	113.25	5.800

Eberhardt et al. [13] use the model

$$X_{ij} = \mu + r_i + \epsilon_{ij} \qquad \begin{matrix} i = 1, \ldots, k \\ j = 1, \ldots, n_i \end{matrix}$$

where X_{ij} is the j^{th} observation from group i, μ is the common mean, r_i is a bias term, and ϵ_{ij} is random error. They then describe a number of estimation approaches, highlighting a minimax approach.

A simple Bayesian model that would be a first step in an EB analysis is

$$\begin{aligned} X_{ij} &= \mu_i + \epsilon_{ij} \qquad && i = 1, \ldots, k \\ & && j = 1, \ldots, n_i \\ \mu_i &= \mu + \delta_{ij}\,, \end{aligned} \tag{2.1}$$

incorporating an underlying common mean in a hierarchical model. Starting from this model an EB analysis can be developed in a rather straightforward manner, as we will see later.

3. Notation and Statistical Formulation

We now describe the EB problem in some generality, also explaining some necessary notation and terminology. The general problem is to make an inference about an unknown parameter θ (which could be a vector) based on observing a sample of n observations $y_1, \ldots, y_n$ whose sampling distribution is described by the known function $f(y_1, \ldots, y_n|\theta)$. In Example 2.1 we have for species i, a sample $(y_{i1}, \ldots, y_{in_i})$ with each y_{ij} having sampling distribution $f(y_{ij} \mid \theta_{ij})$, and all y_{ij}'s are

independent.

In classical statistics it is assumed that there is one true value of θ, and we try to estimate that value. In contrast, the Bayesian view is that θ itself is a random variable whose variability can be described by the distribution $\pi(\theta)$. This distinction, between treating the unknown θ as either fixed or random, is a premier difference between Bayesian and classical statistics. (Discussions of the implications of these assumptions can be found in the book by Press [26] or, at a more advanced level, in Berger [2].) In brief, the classical view is that there is some fixed (unknown) value of the parameter that is driving a process and, hence, its value is reflected in the data we see. The Bayesian view is that this underlying parameter varies (often about *its* mean). The data we see also give information about the underlying parameter, but now there are two sources of variation present in the model. There is the random variation in the observations (as in the classical model), but there is also variation in the observations caused by the variability of the underlying parameter. Thus, the Bayesian model immediately yields a second source of variability to be accounted for. The prior distribution $\pi(\theta)$ is then used to parameterize this second source of variability.

Once $\pi(\theta)$ is specified, the Bayesian combines this information with the data, and updates $\pi(\theta)$ using Bayes rule (see, for example, [7]). The updated prior is called a posterior distribution, denoted $\pi(\theta|y_1, \ldots, y_n)$, and is given by the formula

$$\pi(\theta|y_1, \ldots, y_n) = \frac{f(y_1, \ldots, y_n|\theta)\ \pi(\theta)}{\int f(y_1, \ldots, y_n|\theta)\ \pi(\theta)\ d\theta}. \tag{3.1}$$

The posterior distribution now becomes the basis of all inference for the Bayesian statistician. For example, a Bayesian point estimator of θ is given by the mean of $\pi(\theta|y_1, \ldots, y_n)$, that is,

$$E(\theta|y_1, \ldots, y_n) = \int \theta\pi(\theta|y_1, \ldots, y_n)d\theta \tag{3.2}$$

is a Bayesian point estimator.

The EB methodology borrows from both Bayesian and classical statistics. Whether or not θ is perceived to be random, EB analysis will start with a Bayesian model in which a prior distribution is specified for the parameter. However, in parametric EB methodology, the most common approach is to specify a *family* of prior distributions indexed by another parameter, called a *hyperparameter.* Thus, we could specify the family of prior distribution $\pi(\theta|\lambda)$, indexed by the hyperparameter λ. Analogous to (3.1), and (3.2), we can calculate the posterior distribution

$$\pi(\theta|y_1, \ldots, y_n, \lambda) = \frac{f(y_1, \ldots, y_n|\theta)\ \pi(\theta|\lambda)}{\int f(y_1, \ldots, y_n|\theta)\ \pi(\theta|\lambda)\ d\theta} \tag{3.3}$$

and posterior mean

$$E(\theta|y_1, \ldots, y_n, \lambda) = \int \theta\pi(\theta|y_1, \ldots, y_n, \lambda)d\theta. \tag{3.4}$$

Now, the "empirical" in empirical Bayes comes into use. In the EB methodology, the hyperparameter λ is now estimated by $\hat{\lambda}$. Using this estimate we can calculate $\pi(\theta|y_1, \ldots, y_n, \hat{\lambda})$, $E(\theta|y_1, \ldots, y_n, \hat{\lambda})$, and any other quantity that is needed. These estimates, which are Bayesian in form but use the data to estimate hyperparameters, are empirical Bayes estimates.

Using a family of priors indexed by a hyperparameter places EB squarely between classical and Bayesian models. In a formal Bayesian model only one prior is used, and a classical model, in which there is no specified prior, is mathematically equivalent to allowing the prior distribution to be anything. So by using a parameterized class of priors, EB models are somewhat more specific than classical models, but less specific than Bayesian models. (Models with classes of priors are also used in *robust* Bayesian analysis. See Berger [2] for a discussion of this methodology.)

The use of a hyperparameter plays an important role, as can be seen in Example 2.2. In expression (2.1) the hyperparameter is μ, a common mean level of all groups. The importance of this is that the EB model (or the Bayes model) specifies a structure between the groups, and tries to estimate this structure. This level of modelling is not present in a classical model.

Before going further, a word of caution is needed. Estimates such as $E(\theta|y_1, \ldots, y_n, \hat{\lambda})$ based on (3.4) will, in many cases, be very respectable point estimates of θ. Analogous calculations of variances, for example $\mathrm{Var}(\theta|y_1, \ldots, y_n, \hat{\lambda})$ have an undesirable property of being underestimates. This is because the variability of $\hat{\lambda}$ as an estimate of λ is not accounted for. Morris [25] was one of the first to address this point, and also to develop good EB variance estimates. More recently, some general formulas for variance approximations have been developed and implemented to obtain good EB variance estimates and confidence intervals [4, 5, 19, 21]. Some of these calculations are illustrated in later examples.

Although there are many ways to estimate λ, most of them are based on using the *marginal* distribution of the data, $m(y_1, \ldots, y_n|\lambda)$. This is given by

$$m(y_1, \ldots, y_n|\lambda) = \int f(y_1, \ldots, y_n|\theta)\, \pi(\theta|\lambda) d\theta, \tag{3.5}$$

which is the denominator of (3.3). Using $m(y_1, \ldots, y_n|\lambda)$, techniques such as maximum likelihood estimation [7] can be used to obtain estimates of λ (as in [18]).

Keeping these cautions in mind, EB estimation proceeds with $\hat{\lambda}$ in place of λ, and can be the basis of inference within classical or Bayesian statistics.

4. Detailed Calculation in a Simple Case

To illustrate some EB calculations, we use a very simple situation of Bernoulli trials (coin tossing). We start with a Bayesian formulation of the problem.

4.1 Bayesian Coin Tossing

Toss a coin n times, and let p be the (unknown) probability of a head. (Observe n Bernoulli trials with success probability p). Let y denote the observed number of heads (successes), then the sampling distribution of y, f(y|p) is the binomial distribution

$$f(y|p) = \binom{n}{y} p^y(1 - p)^{n-y} \qquad y = 0, 1, \ldots, n, \tag{4.1}$$

where $\binom{n}{y} = n!/y!(n-y)!$ is a binomial coefficient. We next consider a simple prior distribution on p:

$$\pi(p) = 6p(1-p), \tag{4.2}$$

which is symmetric about $\frac{1}{2}$.

This choice of prior is mainly for convenience, as its form will simplify the ensuing calculations. It is a *conjugate prior* density for the binomial. Conjugate priors always greatly simplify calculations. If we were to consider $\pi(p)$ on more practical grounds, it is a symmetric prior, reflecting that we have no prior opinion as to which side of $\frac{1}{2}$ the parameter p lies. Also, it has prior standard deviation $\approx .22$. It is pictured in Figure 4.1.

We can calculate the posterior distribution of p given y, $\pi(p|y)$, as

$$\begin{aligned} \pi(p|y) &= \frac{f(y|p)\pi(p)}{\int_0^1 f(y|p)\pi(p)dp} \\ &= \frac{\left[\binom{n}{y} p^y(1-p)^{n-y}\right][6p(1-p)]}{\int_0^1 \left[\binom{n}{y} p^y(1-p)^{n-y}\right][6p(1-p)]\, dp}. \end{aligned} \tag{4.3}$$

The denominator of (4.3) is m(y), the marginal distribution of y (there is no hyperparameter here) and is given by

$$\begin{aligned} m(y) &= \int_0^1 6\binom{n}{y} p^{y+1}(1 - p)^{n-y+1}dp \\ &= 6\binom{n}{y} \frac{\Gamma(y+2)\Gamma(n-y+2)}{\Gamma(n+4)}, \end{aligned} \tag{4.4}$$

a distribution known as the *beta-binomial.* (The notation $\Gamma(a)$ denotes the gamma function $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$.)

Continuing with our calculation, we obtain the posterior distribution of p given y as

$$\pi(p|y) = \frac{\left[\binom{n}{y} p^y(1-p)^{n-y}\right] 6p(1-p)}{6\binom{n}{y} \frac{\Gamma(y+2)\Gamma(n-y+2)}{\Gamma(n+4)}},$$

$$= \frac{\Gamma(n+4)}{\Gamma(y+2)\Gamma(n-y+2)}\ p^{y+1}(1-p)^{n-y+1}, \tag{4.5}$$

which is a form of the *beta* distribution.

As mentioned before, the posterior distribution contains all information for Bayesian inference. Thus, if there is interest in a point estimate of p, a Bayes point estimator is the mean of $\pi(p|y)$, given by

$$\begin{aligned} E(p|y) &= \int_0^1 p\pi(p|y)dp \\ &= \frac{y+2}{n+4}, \end{aligned} \tag{4.6}$$

using (4.5) for $\pi(p|y)$.

We can compare the Bayes estimate to a classical estimate of p, the maximum likelihood estimator. This is the observed success rate, y/n, and is denoted by $\hat{p} = y/n$. With some algebra, we can write

$$E(p|y) = \frac{y+2}{n+4} = \left(\frac{n}{n+4}\right)\hat{p} + \left(1 - \frac{n}{n+4}\right)\left(\frac{1}{2}\right), \tag{4.7}$$

a weighted average of the classical estimate and the prior mean, with the weights dependent on the sample size. Thus, the Bayes estimate is a combination of the classical and prior estimates, with weights that reflect the amount of information (that is, variance or sample size) in the respective estimators.

If we perform $n = 50$ Bernoulli trials and observe $y = 35$ successes, we get a classical estimate of p, $\hat{p} = \frac{35}{50} = .7$, and a Bayes estimate of p, $E(p|y) = \frac{50}{54}(.7) + \frac{4}{54}(.5) = .685$, which pulls the sample estimate toward the prior. We can also form interval estimates for the classical and Bayes estimators. A simple classical 95% (approx.) confidence interval is give by

$$\hat{p} \pm 2\left(\frac{\hat{p}(1-\hat{p})}{n}\right)^{\frac{1}{2}} = .7 \pm .13 = (.57, .83).$$

A Bayes credible interval can be computed from $\pi(p|y)$. For $y = 35$ we have

$$\pi(p|y = 35) = \frac{\Gamma(54)}{\Gamma(37)\Gamma(17)}\ p^{36}(1-p)^{16},$$

which is the beta (37, 17) distribution. Based on this distribution we can calculate a Bayes 95% credible region (.56, .80), which is done by allocating 2.5% to each tail of the posterior (see Figure 4.1).

[Figure 4.1 about here]

Note that the inferences from the classical and Bayesian approach are very different. One reason for calling the frequentist interval a "confidence" interval and calling the Bayesian interval a "credible" interval is to highlight this difference. The 95% classical guarantee is that in 95% of all experiments,

the procedure $\hat{p} \pm 2(\hat{p}(1\text{-}\hat{p})/n)^{\frac{1}{2}}$ will cover the true value of p. In any one realization, however, we do not know if p has been covered. In contrast, the 95% Bayes guarantee is that the probability is 95% that p lies between .56 and .80. That is, for the *particular data observed,* we specify a 95% coverage probability.

These inferences again show an essential difference between Bayesian and classical inference. The coverage of the classical interval applies to a long series of replications of the experiment, allowing us to state that in 95% of all experiments the true value of p will be covered. Experimenters tend to like this inference, since it lends credibility to the repeatability of an experiment. However, this interpretation must be somewhat tempered by the fact that no experiment is truly repeated, as there is always some difference in experiments. For the most part, with careful experimentation, the classical inference should be close to the actual behavior of a process. In contrast, there is no repeatability assumed for the Bayesian inference. Starting from the prior probability distribution, the Bayesian constructs the posterior distribution to make an inference for the particular experiment done. There is no inference to a series of replicated experiments. One can say that the Bayesian is only concerned with getting the inference correct for the particular experiment that was performed, and has no interest in a (usually imagined) series of possible replications.

4.2 An Empirical Bayes Approach

In the empirical Bayes approach, we model as a Bayesian, but specify a family of prior distributions. Thus, we might start with the model

$$\begin{aligned} f(y|p) &= \binom{n}{y} p^y(1\text{-}p)^{n\text{-}y} \\ \pi(p|\lambda) &= \frac{\Gamma(2\lambda)}{[\Gamma(\lambda)]^2}\ p^{\lambda\text{-}1}(1\text{-}p)^{\lambda\text{-}1}, \end{aligned} \tag{4.8}$$

where here we have a family of prior distributions. We use a symmetric beta family indexed by the parameter λ, which are all conjugate prior densities.

As specified in (4.8), the model cannot be used for an empirical Bayes analysis for a simple reason: there are more parameters than data. We need to estimate both p and λ, but only have one data value y. To be able to estimate p and λ we need at least two data points. This points out a weakness in empirical Bayes (that it can't be used in very simple situations) but also illustrates a strength of empirical Bayes methods. The strength is that it is well suited for complicated situations, and provides a methodology for obtaining estimators in these situations.

The model in (4.8) can be written in statistical shorthand as the *hierarchical* model

$$\begin{aligned} Y|p &\sim \text{binomial } (n, p) \\ p|\lambda &\sim \text{beta } (\lambda, \lambda) \end{aligned} \tag{4.9}$$

To illustrate an empirical Bayes analysis, we consider a situation of the form

$$\begin{aligned} Y_1|p_1 \sim \text{binomial } (n, p_1) \qquad Y_2|p_2 \sim \text{binomial } (n, p_2) \\ p_1, p_2 \sim \text{beta } (\lambda, \lambda) \end{aligned} \tag{4.10}$$

Here we assume that there are two parameters, p_1 and p_2, that are tied together at some underlying level (since there is a common hyperparameter λ). In the context of coin tossing, we are saying that there are two coins which may have different probabilities of heads (p_1 or p_2), but they come from the same underlying process (with parameter λ).

This is the strength of the empirical Bayes methodology, that it allows us to combine information in such a way as to construct good estimators for each parameter. Using this methodology, we can often find our way through difficult modelling situations and arrive at sensible estimates. In the examples of Section 2, each model combines information in useful ways. In Example 2.1, information is combined across species. That is, the different species represent different problems. In Example 2.2, the information is combined across experiments. This modelling in a hierarchy is often useful, and will usually be a reasonably accurate representation of the process in question. This will be the case whenever the full experiment has parts that have some common thread. Modelling this communality with a parameter allows the data to use this information to possibly achieve better estimates of each part. However, we also want our estimation method to indicate how much weight this "common" part should have. This will help us in constructing good estimates.

Returning to the empirical Bayes model in (4.10), we calculate the posterior distribution of each p_i as

$$\pi(p_i|y_i) = \frac{\Gamma(y_i+2\lambda)}{\Gamma(y_i+\lambda)\Gamma(n-y_i+\lambda)}\ p^{y_i+\lambda-1}(1-p)^{n-y_i+\lambda-1} \qquad i=1, 2,$$

$$= \text{beta } (y_i+\lambda, n-y_i+\lambda)$$

giving posterior expectation

$$E(p_i|y_i, \lambda) = \frac{y_i+\lambda}{n+2\lambda} \qquad i=1, 2.$$

We now calculate the marginal distribution of the y_is, and find

$$m(y_i|\lambda) = \binom{n}{y} \frac{\Gamma(2\lambda)}{[\Gamma(\lambda)]^2}\ \frac{\Gamma(y_i+\lambda)\Gamma(n-y_i+\lambda)}{\Gamma(n+2\lambda)}, \tag{4.11}$$

a beta-binomial distribution. Using this distribution we can estimate λ, and construct our EB

estimates of p_i,

$$E(p_i|y_i, \hat{\lambda}) = \frac{y_i+\hat{\lambda}}{n+2\hat{\lambda}}. \tag{4.12}$$

Notice that our estimate of p_1, $E(p_1|y_1, \hat{\lambda})$ uses information from y_2 (through $\hat{\lambda}$). This is how EB combines information across problems.

Suppose we now observe $y_1 = 35$ and $y_2 = 27$. Using the method of moments on (4.11) (which is computationally easier than maximum likelihood estimation in this case), we obtain an estimate $\hat{\lambda} = 15.205$, which yields EB estimates

$$E(p_1|y_1, \hat{\lambda}) = \frac{35+15.205}{50+30.410} = .624, \quad E(p_2|y_2, \hat{\lambda}) = \frac{27+15.205}{50+30.410} = .525.$$

Note that the EB estimate of p_1 gives more weight to the symmetric prior than the Bayes estimator of (4.7), which estimated to be .685. This is because the additional information ($y_2 = 27$ out of 50) gave more credence to a symmetric model.

We could proceed to estimate an EB variance with $Var(p_i|y_i, \hat{\lambda})$ and even the posterior, with $\pi(p_i|y_i, \hat{\lambda})$. However, as mentioned before, such estimation results in overly optimistic variances and confidence intervals. Instead, we suggest estimating EB variances using more sophisticated methods like those of Morris [25] or of Kass and Steffey [18]. This would lead to a variance estimate of the form

$$\hat{Var}(p_i|y_i, \hat{\lambda}) = \frac{(y_i+\hat{\lambda})(n-y_i+\hat{\lambda})}{(n+\hat{\lambda}+1)(n+2\hat{\lambda})} + C^*, \tag{4.13}$$

where the first term in (4.13) is a direct substitution into the Bayes variance formula. The C^* term is a correction, and can be computed in different ways. The methodology of Morris [25] is detailed in Section 8. For details of the other method, see [18] or [28].

5. Empirical Bayes in the Analysis of Variance

In this section we describe some EB methodology in the analysis of variance (ANOVA) model, a very popular setting for the analysis of experiments. To start, we look at an example.

5.1 A Simple Analysis of Variance

An experiment was run to assess the effect of linseed oil meal on the digestibility of food by steers. The data are from Hsu [17]:

	Treatment				
	1	2	3	4	5
mean	79.23	74.41	74.68	72.35	71.32

std. dev.	6.96	3.77	1.72	4.04	2.57
n	6	6	6	6	6

The measurements are a digestibility coefficient, and the treatments are different amounts of linseed oil meal added to the feed (approximately 1, 2, 3, 4 and 5 kg/animal/day). An ANOVA model for this situation is

$$y_{ij} = \theta_i + \epsilon_{ij} \tag{5.1}$$

where y_{ij} = j^{th} digestibility measurement in i^{th} treatment group, θ_i is the true mean digestibility of the group, and ϵ_{ij} is random error assumed to be normally distributed with mean 0 and variance σ^2 ($\epsilon_{ij} \sim$ normal $(0, \sigma^2)$).

To specify a hierarchical (or Bayes) model, the first step in an empirical Bayes analysis, we must specify a model for the θ_is. The experimenter believes that increasing amounts of linseed oil meal decreases digestibility, and suspects that the θ_is will behave approximately as

$$\theta_i = \alpha + \beta X_i + \delta_i, \tag{5.2}$$

where α and β are parameters (hyperparameters), X_i is the amount of linseed oil meal, and δ_i is random error, $\delta_i \sim$ normal $(0, \tau^2)$. This model specifies a linear trend in the responses, the exact form of which is to be estimated. We thus have the Bayes model

$$\begin{aligned} y_i|\theta_i &\sim \text{normal}(\theta_i, \sigma^2) \\ \theta_i|\alpha,\beta &\sim \text{normal}(\alpha+\beta X_i, \tau^2). \end{aligned} \tag{5.3}$$

Notice that the submodel for the θ_is ties them together. Although there are five θ_is, the submodel describes them to be on a line, which is two-dimensional. The Bayes estimate of θ_i is

$$E(\theta_i|y_i, \alpha, \beta) = \left(\frac{\tau^2}{\sigma^2+\tau^2}\right)(\alpha+\beta X_i) + \left(\frac{\sigma^2}{\sigma^2+\tau^2}\right) y_i, \tag{5.4}$$

and the marginal distribution of y_i is

$$y_i \sim \text{normal}(\alpha+\beta X_i, \sigma^2+\tau^2).$$

As before, we use the marginal distribution of y_i to estimate the unknown prior parameters. We can estimate α and β by performing a simple linear regression of y on X, and use maximum likelihood (or least squares) to estimate σ^2 and τ^2. (More details of these calculations are in the next section). Coding $X_i = i$ we obtain the regression estimates

$$\hat{\theta}_i = 79.66 - 1.76\ X_i \quad (R^2 = .87) \tag{5.5}$$

and substituting estimates into (5.4), the EB estimates are

$$E(\theta_i|y_i, \hat{\alpha}, \hat{\beta}) = .29\hat{\theta}_i + .71 y_i \tag{5.6}$$

To summarize, we have

	Estimates of θ_i				
Treatment	1	2	3	4	5
y_i (cell means)	79.2	74.4	74.7	72.4	71.3
$\hat{\theta}_i$ (regression)	77.9	76.1	74.4	72.6	70.9
empirical Bayes	78.9	74.9	74.6	72.5	71.2

Notice that the empirical Bayes estimates, even though pulled toward the line in (5.5), do not necessarily have to lie on a line. The fact that they do (Figure 5.1) is evidence that the linear submodel for θ_i reasonable. The amount by which the EB estimates are pulled toward the line is data dependent. The stronger the data support the linear model, the more the EB estimates are pulled to the line.

As usual, attaching standard errors to the EB estimates is not trivial. However, it should be mentioned that there is a quick, conservative approximation. That is to use the usual standard errors and confidence intervals together with the EB estimates. Most often, intervals constructed in this way will be conservative. For example, the usual 90% t-interval centered at an EB estimate will most often have coverage probability greater than 90%. There is a small chance that the interval will have coverage probability less than 90%, but this is slight enough not to cause worry. So attaching the usual standard errors to the EB estimates is a simple, conservative tactic. It is also possible to use more sophisticated methods and obtain simultaneous EB confidence statements that yield improvements over the classical Scheffé intervals. Unfortunately, the simultaneity of the confidence statement leads to rather wide intervals that are often not used in practice. Details of the EB simultaneous construction will not be given here, but can be found in [10].

5.2 Extrapolation

The paper of DuMouchel and Harris [12], already mentioned in Section 2, provides a nice illustration of how the EB model can be used to bridge together similar problems. By doing so, conclusions can be extrapolated to different populations. Of course, we must be wary of any extrapolation, since in doing this puts ultimate faith in the model. That is, since extrapolation involves making inferences where no data have been collected, the only basis for the inference is the model. Moreover, this lack of data means we cannot check the adequacy of our model in these regions, hence we must have "ultimate faith." With that disclaimer, EB provides a nice methodology for extrapolation.

The DuMouchel and Harris [12] model is

$$\begin{aligned} y_{ij} &= \theta_{ij} + \epsilon_{ij} \\ \theta_{ij} &= \mu + \alpha_i + \gamma_j + \delta_{ij} \end{aligned} \tag{5.7}$$

where the terms are all defined in Section 2.1. The model for y_{ij} is known as a "cell means" model, since it relates the data y_{ij} to the cell it comes from. In contrast, the model for the θ_{ij} is a "no interaction" model. That is, each θ_{ij} is related only to a row effect (α_i) and a column effect (γ_j). Hence, knowledge of the individual cell is not needed to estimate θ_{ij}. This is the feature of the model that allows for extrapolation.

A schematic diagram of the data analyzed in [12] (somewhat abridged) is

Species	Roofing Tar Emissions	Coke Oven Emissions	Diesel Engine Emissions	Gas Engine Emissions	Benzo Pyrene	Cigarette Smoke
Human	X	X	O	O	O	X
Mice	X	X	X	X	X	X
Hamster	X	X	X	X	X	X

where X = Data Present and O = Data Absent.

The goals were both to provide estimates for cells with no data and to improve precision of estimates (using posterior standard deviations (SD)). Both goals were to be accomplished by modelling the data as having common underlying structure, and borrowing "ensemble strength" through EB models to help improve estimates.

A portion of their results, relating lung cancer risk in humans, is summarized below. They presented not only EB estimates, but also Bayes estimates and maximum likelihood estimates.

Roofing Tar	Estimate (log slope)	Posterior SD	
Orig. Data	.50	1.41	Estimates change
Bayes	.12	1.02	greatly because of
EB	.12	1.01	high SD of
MLE	-.01	.70	original estimate
Coke Oven			
Orig. Data	1.48	.41	Estimates do not
Bayes	1.38	.02	change much because
EB	1.38	.01	of small SD of
MLE	1.30	.70	original estimate
Diesel Engine			
Bayes	-.46	1.45	These values are
EB	-.46	1.40	extrapolated from the
MLE	-.57	.80	analysis. There is no data on humans exposed to diesel engine fumes. The extrapolation results in a high SD.

The methodology of DuMouchel and Harris provides a modelling structure for reasonable extrapolation. Although we should not put ultimate faith in the extrapolated dose-response slopes for humans exposed to diesel engine fumes, these estimates do give us some idea of the risks associated with this exposure (subject, of course, to the approximate correctness of the model).

6. A Deeper Look at Empirical Bayes

In this section we illustrate some of the underlying statistical theory of EB analyses. Although the material in this section is important in fully understanding the mechanisms of EB, the section can be skipped without serious consequences.

A general form of an ANOVA model is

$$y_i \sim \text{normal}\ (\theta_i, \sigma^2) \qquad i=1, \ldots, p \tag{6.1}$$

all statistically independent. Here we can think of y_i as observed ANOVA cell means, θ_i as true cell means. For now we assume σ^2 is a known number (to ease exposition). This assumption is easily relaxed in practice, where σ^2 can be replaced by an estimate $\hat{\sigma}^2$.

In an EB model we assume an underlying structure for the θ_is, which leads us to consider the extension of (6.1)

$$\begin{aligned} y_i &= \theta_i + \epsilon_i \qquad i = 1, \ldots, p \\ \theta_i &= Z_i'\beta + \delta_i \end{aligned} \tag{6.2}$$

where $\epsilon_i \sim$ normal $(0, \sigma^2)$ and $\delta_i \sim$ normal $(0, \tau^2)$ are the associated errors. The vector Z_i contains covariates that (hopefully) link the θ_is together (as in Section 5.1, for example, where the Z_is would reflect the amount of linseed oil meal in the feed),

Notice that there is no subscript on the vector β, as this is the EB part of the model, the part that ties things together. We would like to model β to have as small a dimension as possible, modelling a strong underlying structure into the θ_is. This would allow the EB estimates to offer the most improvement.

Although we won't consider it here, the model (6.2) can be further generalized to

$$\begin{aligned} y_i &= X_i'\theta_i + \epsilon_i \\ \theta_i &= Z_i'\beta + \delta_i, \end{aligned} \tag{6.3}$$

where X_i are covariates for the y_is, with only a slight increase in algebraic effort. Calculations for this case are given in detail in [12] or [23].

A typical set of distributional assumptions for the model (6.2) is

$$\begin{aligned} y_i|\theta_i &\sim \text{normal } (\theta_i, \sigma^2), \qquad i=1, \ldots, p \\ \theta_i|\beta &\sim \text{normal } (Z_i'\beta, \tau^2), \\ \beta_i &\sim \text{uniform } (-\infty, \infty), \end{aligned} \tag{6.4}$$

where Z_i = r × 1 vector of known predictor variables, τ^2 = unknown variances, β = r × 1 vector of unknown regression coefficients. Using matrix notation we can write this model as

$$\begin{aligned} Y|\theta &\sim \text{normal } (\theta, \sigma^2 I), \\ \theta|\beta &\sim \text{normal } (Z'\beta, \tau^2 I), \\ \beta_i &\sim \text{uniform } R^r, \end{aligned} \tag{6.5}$$

More complicated covariance structures than those given in (6.5) are also possible, again bringing an increased amount of algebraic effort.

As outlined in Section 3, we can perform a number of distributional calculations from the model (6.5). Although these are more complicated, they reflect the same underlying principles as given before. Our ultimate goal is to calculate a posterior distribution for θ, $\pi(\theta|y)$, which will provide the basis for our EB inference.

Briefly, the prior distribution on θ, $\pi(\theta)$ is given by

$$\pi(\theta) = \int \pi(\theta|\beta) d\beta, \tag{6.6}$$

where $\pi(\theta|\beta) \sim$ Normal $(Z'\beta, \tau^2 I)$ from (6.5). After completing this integration, we can calculate the

posterior $\pi(\theta|y)$, as

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta}, \tag{6.7}$$

where the denominator is the marginal distribution of y. These calculations are somewhat lengthy, but when completed we find that $\pi(\theta|y)$ is a multivariate normal distribution with

$$\text{mean} = \frac{\tau^2}{\sigma^2+\tau^2}\left(I + \frac{\sigma^2}{\tau^2}H\right)Y$$

$$\text{covariance matrix} = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\left(I - \frac{\sigma^2}{\sigma^2+\tau^2}H\right)^{-1} \tag{6.8}$$

where

$$H = Z(Z'Z)^{-1}Z'.$$

Thus, the Bayes estimator of Y, the posterior mean, is

$$E(\theta|y) = \frac{\tau^2}{\sigma^2+\tau^2}\left(I + \frac{\sigma^2}{\tau^2}H\right)Y = HY + \left(1 - \frac{\sigma^2}{\sigma^2+\tau^2}H\right)(Y - HY) \tag{6.9}$$

(The quantity HY can be obtained by regressing Y on Z using the marginal model $Y = Z\beta + \text{error}$.)

The EB estimate is obtained by replacing the unknown quantity $\left(1 - \frac{\sigma^2}{\sigma^2+\tau^2}\right)$ by an estimate based on the marginal (unconditional on θ) distribution of Y. Marginally, the quadratic form

$$\frac{Y'(I - H)Y}{\sigma^2+\tau^2}$$

has a chi-squared distribution with p-r degrees of freedom, which implies its expected value is

$$E\left[\frac{(p - r - 2)\sigma^2}{Y'(I - H)Y}\right] = \frac{\sigma^2}{\sigma^2+\tau^2}, \tag{6.10}$$

so we have an unbiased estimator of $\sigma^2/(\sigma^2+\tau^2)$ based on the marginal distribution. Substituting into (6.9), we obtain our empirical Bayes estimator of θ,

$$E(\theta|Y, \hat{\tau}^2) = HY + \left[1 - \frac{(p - r - 2)\sigma^2}{Y'(I - H)Y}\right](Y - HY). \tag{6.11}$$

The form of (6.11) is quite interesting, and bears comment. The piece in square brackets is the EB part of the estimator, and acts as a weight, yielding an EB estimate of the vector θ that is a weighted average of Y and HY. Estimating θ with Y would be most desirable if we believed only in the first line of (6.2) (the full model), and estimating θ with $HY = Z(Z'Z)^{-1}Z'Y$ would be most

desirable if we believed only in the second line of (6.2) (the submodel). Thus, the EB estimate is a weighted average of full model and submodel estimates. (Recall that the submodel ties together the θ_is, and allows EB to result in an improvement.)

However, there is another important feature of the EB estimate. The weight function (in square brackets in (6.11)) depends on the data through the quantity $Y'(I\text{-}H)Y$. This quantity is the residual sum of squares from the regression of Y on Z, hence it is a measure of how well the submodel fits the data (large values of $Y'(I\text{-}H)Y$ imply the submodel does not fit the data, small values imply that it does). Moreover, if $Y'(I\text{-}H)Y$ is large (so the data do not support the model), the EB estimator is very close to Y, the full model estimate. Conversely, if $Y'(I\text{-}H)Y$ is small (and the data do support the submodel), the EB estimator is close to HY, the submodel estimate. Thus, the EB estimate allows the data to select the appropriate model, and produces an estimate which is close to optimal for that model.

The performance of $E(\theta|Y, \hat{\tau})$ of (6.11) can be improved with the slight modification

$$E^{+}(\theta|Y, \hat{\tau}) = HY + \left[1 - \frac{(p - r - 2)\sigma^2}{Y'(I - H)Y}\right]^{+} (Y - HY), \tag{6.12}$$

where the superscript "+" denotes that the quantity in square brackets is replaced by zero if it is negative. This stops the estimator from pulling past the submodel.

For the more general case of unknown σ^2 and n_i observations in cell i the theory remains essentially unchanged, except that the algebra gets more difficult. Using a similar development, an empirical Bayes estimator of θ is

$$E(\theta|Y, \hat{\tau}, \hat{\sigma}) = GY + \left[1 - \frac{\frac{\nu}{\nu+2}(p - r - 2)\hat{\sigma}^2}{Y'(I - G)'D(I - G)Y}\right]^{+} (Y - GY) \tag{6.13}$$

where D is a diagonal matrix with entries $(n_1, n_2, \ldots, n_p)$, $G = Z(Z'DZ)^{-1}Z'D$, $\hat{\sigma}^2$ = estimate of σ^2 from the full model, and ν = error df.

7. More Examples of Empirical Bayes Analyses

In this section we briefly present five applications of EB methods, illustrating the breadth of influence that these methods have had. There are numerous other applications of EB in the literature, too many to mention here. The field is continually expanding, and a quick perusal through some statistical methodology journals will almost certainly contain a new EB application.

7.1 Estimation of Means

Using the model (2.1), which is similar to models discussed by Eberhardt et al. [13], we can employ the methods of the previous section to obtain empirical Bayes estimates of selenium concentration. Starting from the model (2.1), we can write the following model for the means $\bar{X}_i$:

$$\begin{aligned} \bar{X}_i &= \mu_i + \epsilon_i \qquad\qquad i = 1, \ldots, 4 \\ \mu_i &= \mu + \delta_i \end{aligned} \tag{7.1}$$

where ϵ_i has a normal distribution with mean zero and variance σ^2/n_i, $\epsilon \sim$ normal $(0, \sigma^2/n_i)$, n_i = number of observations in X_i, and $\delta_i \sim$ normal $(0, \tau^2)$. Thus we have a hierarchical model that specifies the individual selenium means (μ_i) have a common underlying value (μ). The observed differences are due to two sources of error that we have modelled with normal distributions.

As written in (7.1), the model does not appear to be in the form of Section 6 because the first stage variances (the variances of $\bar{X}_i$) are not all equal. However, formula (6.13) is applicable, and we will use that to estimate the selenium concentrations. To use formula (6.13) we identify

$$(\bar{X}_1, \ldots, \bar{X}_4)' = (105.0,\ 109.75,\ 109.5,\ 113.25)$$

$$(1,\ 1,\ 1,\ 1)' = Z$$

$$(n_1, n_2, n_3, n_4) = (8,\ 12,\ 14,\ 8) = \text{diagonal of D.}$$

This leads to

$$G = \frac{1}{N}\begin{pmatrix} n_1 & n_2 & n_3 & n_4 \\ n_1 & n_2 & n_3 & n_4 \\ n_1 & n_2 & n_3 & n_4 \\ n_1 & n_2 & n_3 & n_4 \end{pmatrix}$$

where $N = \sum_{i=1}^{4} n_i = 42$, and

$$Y'(I-G)'D(I-G)Y = \sum_{i=1}^{4} n_i(\bar{X}_i - \bar{\bar{X}})^2$$

with $\bar{\bar{X}} = \sum_{i=1}^{4} n_i\bar{X}_i/N$. Also, we calculate the pooled variance by

$$\hat{\sigma}^2 = \sum_{i=1}^{4} (n_i - 1)S_i^2/(N-4) = 28.925 \tag{7.2}$$

with $\nu = N - 4 = 38$ degrees of freedom. Finally, since $p = 4$ and $r = 1$, we have

$$\hat{B} = 1 - \frac{\frac{\nu}{\nu+2}(p-r-2)\hat{\sigma}^2}{\sum_{i=1}^{4} n_i(\bar{X}_i - \bar{\bar{X}})^2} = 1 - \frac{\frac{38}{40}(28.925)}{275.036} = .900 \tag{7.3}$$

This yields empirical Bayes estimates

$$\begin{aligned}
\hat{\mu}_1 &= 109.429 + .9(105.0 - 109.429) = 105.443 \\
\hat{\mu}_2 &= 109.429 + .9(109.75 - 109.429) = 109.718 \\
\hat{\mu}_3 &= 109.429 + .9(109.5 - 109.429) = 109.493 \\
\hat{\mu}_4 &= 109.429 + .9(113.25 - 109.429) = 112.868
\end{aligned} \tag{7.4}$$

Note that the EB estimates are not very different from the $\bar{X}_i$. This is because the submodel, which specifies that the μ_i's have common mean μ, is not strongly supported by the data. In particular, $\bar{X}_1$ and $\bar{X}_4$ are far from the grand mean. Thus, although the theory says that the EB estimate is an improvement over the usual estimate, in this case the practical difference is small.

7.2 Growth Curves

Strenio, et al. [29] applied EB methods to estimation of growth curves. They modelled growth for an individual as a polynomial in age, using a second-stage model that relates individuals (using covariates). Calculations are done in general, using some of the Lindley-Smith (1972) formulas. However, here we will just do an example. Let y_{it} = weight of i^{th} rat at t^{th} week. Then we have

$$E(y_{it}|\beta_{1i}, \beta_{2i}) = \beta_{1i} + \beta_{2i}(t\text{-}1),$$

where β_{1i} and β_{2i} are regression coefficients for the growth curve of individual i. In the submodel,

$$\begin{aligned}
E(\beta_{1i}|\gamma_{11}, \gamma_{12}) &= \gamma_{11} + \gamma_{12}x_i, \\
E(\beta_{2i}|\gamma_{21}, \gamma_{22}) &= \gamma_{21} + \gamma_{22}x_i,
\end{aligned}$$

where the γ_{ij}s are unknown regression coefficients, and x_i = mother's weight (covariate).

The EB estimate of β_{1i} is a linear combination of

$$\hat{\beta}_{1i} \text{ (from regression of } i^{th} \text{ individual on time)}$$

and

$$\hat{\gamma}_{11} + \hat{\gamma}_{12}x_i \text{ (from regression of } \hat{\beta}_{1i} \text{ on } x_i).$$

The EB estimate of β_{2i} similarly obtained. The growth curve (a line in this case) is thus a linear combination of the curve for the individual and the curve for the ensemble. Note that the submodel estimates, $\hat{\gamma}_{11}$ and $\hat{\gamma}_{12}$, are based on data that is summed over time, hence use all of the information. This model was applied to data on growth in rats, and an illustration of the results is given in Figure 7.1.

[Figure 7.1 About Here.]

Note how the EB growth curve pulls the individuals curve toward the population average. This results in a reduced variance (and hence more reliable estimates) of the individual curves.

7.3 Contingency Tables

EB methods can be applied to the analysis of contingency table data, proving especially useful with large, sparse tables. The usual contingency table can be described by

$$x_{ij} = \text{observed frequency in cell (i, j)}$$

where x_{ij} has a multinomial distribution with expectation Np_{ij}, $\sum_{i,j}x_{ij} = N$, $\sum_{i,j}p_{ij} = 1$. It is common to use a log-linear model for the cell probabilities, given by

$$\log p_{ij} = \mu_0 + \mu_{1i} + \mu_{2j} + \mu_{12ij} \cdot \tag{7.5}$$

(μ_{1i}: row effect; μ_{2j}: column effect; μ_{12ij}: interaction effect)

EB methods for contingency tables now place prior distributions on the μ's and estimate unknown prior parameters from the marginal distribution of the x_{ij}'s. A popular method (Bishop, Feinberg and Holland [3]) is to use a Dirichlet prior, which yields estimates of $Np_{ij} = n_{ij}$ of the form

$$\hat{n}_{EB} = \hat{w}_{ij} + (1 - \hat{w})\hat{n}_m,$$

where $\hat{n}_m$ is the maximum likelihood estimator under independence, $\hat{w}_{ij}$ is an estimate based on the log-linear model, and $\hat{w}$ is estimated from the marginal distribution.

Leonard [22] uses a Bayesian approach with normal prior distributions on the μ_{ij}s, and uniform distributions on the hyperparameters. Laird [19] uses an EB approach, employing normal and uniform distributions in a different way, and estimating variances from the marginal distribution. In either case, much computation is involved in getting estimates.

7.4 ANOVA with Unequal Variances

An important assumption in the analysis of variance (ANOVA) is that the variances for each cell are the same. This is reflected in models like (6.1) or (6.2), where it is assumed that each cell has variance σ^2 (not dependent on i). If the equal variance assumption is relaxed, we get a model of the form

$$y_i \sim \text{normal } (\theta_i, \sigma_i^2). \tag{7.6}$$

Although it may not be immediately obvious, the model (7.6) is even more general than the model leading to the EB estimator in (6.13), where it is assumed that there are different numbers of observations per cell (n_i), but each cell has the same variance parameter σ^2.

Efron and Morris [15] apply the model (7.6) to EB estimation of toxoplasmosis rates in 36 cities in El Salvador. They use the model

$$\begin{aligned} y_i &\sim \text{normal } (\theta_i, \sigma_i^2) \qquad i = 1, \ldots 36 \\ \theta_i &\sim \text{normal } (0, \tau^2) \end{aligned} \tag{7.7}$$

where y_i = observed prevalence rate in city i, θ_i = true prevalence in city i, σ_i^2 = variance of the observation in city i, and τ^2 is a common prior variance. The EB estimate of θ_i is

$$E(\theta|\hat{\sigma}_i, \hat{\tau}) = \frac{\hat{\sigma}_i^2}{\hat{\sigma}_i^2+\hat{\tau}^2} y_i, \tag{7.8}$$

where $\hat{\sigma}_i^2$ and $\hat{\tau}^2$ are estimates from the marginal distribution. In the unequal variance case these estimates are more difficult to calculate, but there are simple algorithms for doing so ([15], [25]).

The results are quite interesting, and a selection of the estimates is

Selected Estimates and
Empirical Bayes Estimates of
Toxoplasmosis Prevalence Rates

City	X_i	σ_i	EB_i
1	.293	.304	.035
4	.152	.115	.075
5	.139	.081	.092
8	.098	.087	.062
13	.035	.056	.028
21	-.034	.073	-.024
25	-.098	.068	-.072
28	-.138	.063	-.106
29	-.156	.077	-.107
31	-.241	.106	-.128
32	-.294	.179	-.083
33	-.296	.064	-.225

Notice how the X_i values with large variance get changed more, in particular note the change in estimates for cities 1 and 32. City 33 has virtually the same X_i value as city 32, but a much smaller variance. Hence, its EB estimate changes very little. If we had assumed $\sigma_i^2 = \sigma^2$, each X_i would have been changed by an equal, smaller amount.

Thus, we see how the EB estimation procedure tends to trust individual observations with smaller variance, and not change them much. However, if the individual estimate has a higher variance, the EB estimate tends to trust the overall information more, and changes the estimate substantially.

7.5 Regression Equations

Allus et al. [1] describe a series of response surface experiments to investigate the effect of metals on the growth of plants. They did two experiments, each with three metals. Their methods, and the EB modification that follows, can be easily extended to N metals.

A model for plant growth as a function of three metals is, from Allus et al. [1],

$$\begin{aligned} y = {} & b_0 + b_1x_1 + b_2x_2 + b_3x_3 \\ & + b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2 \\ & + b_{21}x_1x_2 + b_{31}x_1x_3 + b_{32}x_2x_3 \ , \end{aligned} \tag{7.9}$$

where y = measure of plant growth (e.g. root length), b_0 is a constant, b_i and b_{ii} measures the effect of metal i, and b_{ij} measures the interaction effect of metals i and j.

In models such as (7.9), it is common to inquire about the interaction terms. Specifically, there is usually interest in the tenability of the submodel which excludes interactions, which could be of the form

$$\begin{aligned} y = {} & b_0 + b_1x_1 + b_2x_2 + b_3x_3 \\ & + b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2 \end{aligned} \tag{7.10}$$

The submodel (7.10) is only one of many that could be used, each submodel being suggested by some aspect of the underlying process. For example, Allus et al. [1] found in their experiments that some b_{ij} are considerably smaller in magnitude than the other coefficients, suggesting a no-interaction submodel. Also, with other metals it was noted that they were not very toxic to plants at the levels studied, suggesting another submodel.

It is generally acknowledged that pre-testing for the appropriate model, then using that model, is a non-optimal strategy. However, EB methods allow us to incorporate submodel information into the full model estimation. The resulting EB estimators for the coefficients will be closest to whichever model is best supported by the data. Specifically, adapting the formulas of Section 6, the EB estimate of a coefficient b_i is given by

$$\hat{b}_i^{EB} = \hat{b}_i^S + [1 - \hat{B}]^+(\hat{b}_i^F - \hat{b}_i^S) \tag{7.11}$$

where the superscripts "S" and "F" refer to the submodel and full model. These estimates are calculated by performing ordinary regressions on the models (7.9) and (7.10). The shrinkage factor $\hat{B}$ is given by

$$\hat{B} = \frac{(p - r - 2)\hat{\sigma}^2}{\sum_i (\hat{b}_i^F - \hat{b}_i^S)^2 + (\hat{b}_{ii}^F - \hat{b}_{ii}^S)^2} \ . \tag{7.12}$$

The shrinkage factor $\hat{B}$ has an interesting interpretation, one that holds in general. The denominator of $\hat{B}$ is actually the sum of squares for testing the hypothesis of no interaction. Furthermore, $\hat{\sigma}^2$, which is the residual mean square from the full regression, is the error term used to test such an hypothesis. In fact, $\hat{B}$ can be written

$$\hat{B} = \frac{(p-r-2)}{(p-r)} \frac{1}{F} \ , \tag{7.13}$$

where F is the calculated F statistic for testing H_0: no interaction. Comparing (7.13) with (7.11) we see that $\hat{b}_i^{EB}$ will be closest to the estimate (either $\hat{b}_i^S$ or $\hat{b}_i^F$) that the F test best supports.

The form of (7.13) tends to hold in a wide variety of cases. That is, the shrinkage factor is proportional to the F statistic that tests the viability of the submodel.

8. Estimation of Empirical Bayes Variances

A conservative approach to attaching variance estimates to EB point estimates is to use the usual estimates of variance. For example, a standard variance estimate for a maximum likelihood estimate (standard output from many statistical packages) will serve as a conservative estimate of the variance of the EB estimate. Of course, we would like to do better. Since the EB estimates supposedly have improved precision, this should be reflected in our ability to attach smaller variances to our EB estimates. We know these variance estimates will not be as small as the "naive" ones, obtained by direct substitution (as mentioned in Section 3), but we hope that good EB variance estimates will be smaller than the classical ones.

Constructing good EB variance estimates, and thus good EB confidence intervals has been the subject of much research. There have been theoretical investigations ([9], [10]), approximations ([18], [25]), and application of computer-intensive techniques on construction EB confidence intervals ([4], [5]. 20]). Here, we describe the approach of Morris [25], one of the first to construct EB estimates. The concerns addressed by Morris are the fundamental ones in EB variance estimation, and are concerns of all later work.

Recall the EB anova Model as in (6.4)

$$\begin{aligned} y_i|\theta_i &\sim \text{normal } (\theta_i, \sigma^2/n), \\ \theta_i|\beta &\sim \text{normal } (Z_i'\beta, \tau^2), \qquad i=1, \ldots, p \\ \beta &\sim \text{uniform} \end{aligned} \tag{8.1}$$

where y_i = cell means based on n observations per cell, θ_i are the parameters of interest, the true cell means, $\beta = r \times 1 (r < p)$ vector of unknown regression coefficients, Z_i = vector of covariates for cell i, and the variances are σ^2 and τ^2. Under this setup, the Bayes estimator of θ_i is

$$\hat{\theta}_{Bi} = \hat{\theta}_i + (1 - B)(y_i - \hat{\theta}_i), \tag{8.2}$$

where

$$B = \sigma^2/(\sigma^2 + n\tau^2), \quad \hat{\theta}_i = Z_i'\hat{\beta}, \quad \text{and} \quad \hat{\beta} = (Z'Z)^{-1}Z'y.$$

The EB estimate of B, using arguments similar to the previous ones is

$$\hat{B} = (p - r - 2)\left(\frac{v}{v+2}\right)\frac{\hat{\sigma}^2}{n}/\sum(y_i - \hat{\theta}_i)^2 , \tag{8.3}$$

where we have used the operation "+" as described in (6.12). The variance of the Bayes estimator, $\hat{\theta}_{Bi}$, is given by

$$\text{Var}(\hat{\theta}_{Bi}) = \frac{\sigma^2}{n}(1 - B). \tag{8.4}$$

Morris [25] notes that, for the variance of EB estimator, we must also account for increase in variance due to both estimating B and estimating β, and suggests using

$$\text{Variance}(\hat{\theta}_{EB_i}) = \frac{\sigma^2}{n}\left(1 - \frac{p-r}{p}\hat{B}\right) + v(\hat{B})(y_i - \hat{\theta}_i)^2 \tag{8.5}$$

where $v(\hat{B}) = \frac{2}{p-r-2}\hat{B}^2$.

Although (8.5) appears to be formidable, it is composed of very sensible pieces. The first term mimics the Bayes variance of (8.4), where $\frac{p-r}{p}\hat{B}$ is substituted for B. (The factor $\frac{p-r}{p}$ is for bias correction.) The second term in (8.5) is the important one, as it estimates the variance of $\hat{B}$. In the Bayes variance (8.4) it is assumed that B is a fixed, known constant, hence B has no variance itself. The estimate $\hat{B}$, however, does have a variance. If we just use the first term of (8.5) as a variance estimate we are ignoring the inherent variance of $\hat{B}$ as an estimation of B. This is what the second term in (8.5) addresses.

Note that the EB variance estimates for each cell can be different. In fact, the variance for cell i increases as the cell mean y_i moves further from the submodel, and decreases as the cell mean moves closer to the submodel. This will result in some EB variances being larger than the classical estimates, and others being smaller.

8.1 Variance Estimation for the Steer Data

For the steer data of Section 5.1, the classical standard deviation estimate is $\hat{\sigma}/\sqrt{n} = 1.24$, and the EB standard deviations, using equation (8.5), are

	Treatment				
	1	2	3	4	5
EB Standard Deviation	1.23	1.32	1.13	1.13	1.14

In contrast to the variance estimate from the usual analysis of variance, the EB variance estimates differ according to how closely the treatment mean follows the submodel (and, of course, how well the factor $\hat{B}$ is estimated). The EB standard deviation of the cell mean from treatment 2 is the highest, as that cell mean was further from the submodel.

8.2 Variance estimation for the Selenium Data

For the selenium data of Sections 2.1 and 7.1, we can also attach EB variance estimates to our individual mean estimates. Here, however, we have different numbers of observations in each mean. Adapting the variance formula (8.5) to the EB estimates of (7.4), we have

$$\text{Variance } (\hat{\mu}_i) = V_i^2 = \frac{\hat{\sigma}^2}{n_i}\left(1 - \frac{4-1}{4}\hat{B}\right) + \frac{2}{4-3}\left(\bar{X}_i - \bar{\bar{X}}\right)^2, \tag{8.6}$$

where $\bar{X}_i$ and $\bar{\bar{X}}$ are the cell means and grand mean, n_i is the number of observations in $\bar{X}_i$, $\hat{\sigma}^2$ is the pooled error of (7.2) and $\hat{B}$ is the shrinkage factor of (7.3). Substituting into (8.6) yields

	Group			
	1	2	3	4
EB Mean	105.443	109.718	109.493	112.868
EB St. Dev.	5.740	.975	.824	4.983

Again, the largest variance estimates are for those means farthest from the submodel.

It should be emphasized that the EB estimates are valid whether or not the submodel is true. Data that are consistent with the submodel will have smaller EB variances, since the submodel ties things together and effectively gives us more observations. Data that are inconsistent with the submodel will have larger variances, since "submodel pooling" occurs and the EB estimates revert back to usual estimates. In fact, they can yield higher variances than usual estimates, since we are penalized for estimating a submodel that doesn't apply. This can be seen in the selenium data by looking at groups 1 and 2. Group 2 closely follows the submodel, so the EB standard deviation (.975) is smaller than the usual standard deviation ($\sqrt{\hat{\sigma}^2/12} = \sqrt{28.925/12} = 1.553$). However, Group 1 does not agree with the submodel and its EB standard deviation (5.740) is larger than the usual ($\sqrt{\hat{\sigma}^2/8} = \sqrt{28.925/8} = 1.901$).

Eberhardt et al. [13] are most interested in obtaining an estimate of the overall mean selenium concentration. That is, they want to combine the individual estimates of mean selenium concentration and provide a method for minimax combining, which leads to a robust estimate of the overall mean. Their method is based on combining the usual estimates of the means and variances, $\bar{X}_i$ and $\hat{\sigma}^2/n_i$, and

can be adapted to use the EB estimates given above.

A simple EB combined estimate can be calculated using the usual weighting scheme of reciprocal variances. That is, a combined EB estimate of the overall mean is

$$\hat{\mu}_{EB} = \frac{\sum_{i=1}^{4} \hat{\mu}_i / V_i^2}{\sum_{i=1}^{4} 1/V_i^2} = 109.589 .$$

This estimate is similar to the usual combined estimate, and slightly smaller than the Eberhardt et al. minimax estimate.

9. Discussion

The true test of the worth of a statistical procedure is its longevity, and it seems that empirical Bayes has passed that test in admirable fashion. The first empirical Bayes analyses started to appear in the early 1970s and, as of 1991, the methodology is still finding new uses and applications.

For the most part we have illustrated empirical Bayes as a modelling and estimation technique, two aspects of statistics in which empirical Bayes has proven most useful. What has been referred to as EB models are also called hierarchical models, since they allow the modelling of a process in hierarchical layers. By imposing such a structure, it is often the case that complicated situations can be modelled with relatively simple layers. Moreover, the empirical Bayes strategy leads to specification of simpler submodels which can result in estimates with improved precision, and increased ability to extrapolate.

Once EB is applied to any situation other than the most simple, calculation of estimates can become formidable. The explosion in computer-intensive statistical methods, which have resulted in greatly increased calculational ability, has widened the applicability of EB methods. Indeed, almost any EB application has accomplished its calculations through use of the EM Algorithm [11] and more recently researchers have employed bootstrapping algorithms [14] to aid in the calculation of EB confidence intervals ([4], [5], [20]). Even newer computational methods such as the data augmentation algorithm [30] or the Gibbs sampler ([8], [16]) will certainly find uses in the increased applicability of empirical Bayes methodology.

References

1. Allus, M.A., Brereton, R.G., and Nickless, G. (1988). The Effect of Metals on the Growth of Plants: the Use of Experimental Design and Response Surfaces in a Study of the Influence of Tl, Cd, Zn, Fe and Pb on Barley Seedlings. *Chemometrics and Intelligent Laboratory Systems* 3, 215-231.

2. Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis, Second Edition.* New York: Springer-Verlag.

3. Bishop, Y.M., Feinberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Mass.: MIT Press.

4. Carlin, B.P. and Gelfand, A.E. (1991). A sample reuse method for accurate empirical Bayes confidence intervals. *J. Roy. Statist. Soc., Series B* **53**, 189-200.

5. Carlin, B.P. and Gelfand, A.E. (1990). Approaches for empirical Bayes confidence intervals. *J. Amer. Statist. Assoc.* **85,** 105-114.

6. Casella, G. (1985). An introduction to empirical Bayes data analysis. *American Statistician* **39**, 83-87.

7. Casella, G. and Berger, R.L. (1990). *Statistical Inference.* Monterey: Wadsworth and Brooks/Cole.

8. Casella, G. and George, E.I. (1991). Explaining the Gibbs sampler. Technical Report BU-1098-MA, Biometrics Unit, Cornell University. To appear in the *American Statistician.*

9. Casella, G. and Hwang, J.T. (1983). Empirical Bayes confidence sets for the mean of a multivariate normal distribution. *J. Amer. Statist. Assoc.* **78**, p. 688-698.

10. Casella, G. and Hwang, J.T. (1987). Employing vague prior information in the construction of confidence sets. *Journal of Multivariate Analysis* **21**, p. 79-104.

11. Dempster, A.P., Laird, N.M. and Rubin, D.R. (1977). Maximum likelihood from incomplete data via the EM Algorithm (with discussion). *J. Roy. Statist. Society, Series B* **39**, 1-38.

12. DuMouchel, W.M. and Harris, J.E. (1983). Bayes methods for combining the results of cancer studies in humans and other species (with discussion). *J. Amer. Statist. Assoc.* **78**, p. 293-315.

13. Eberhardt, K.R., Reeve, C.P. and Spiegelman, C.H. (1989). A Minimax Approach to Combining Means, with Practical Examples. *Chemometrics and Intelligent Laboratory Systems* 5, 129-148.

14. Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. NSF-CBMS Monograph 38. Philadelphia: SIAM.

15. Efron, B. and Morris, C. (1975). Analysis using Stein's estimator and its generalization. *J. Amer. Statist. Assoc.* **63,** p. 311-319.

16. Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85,** 398-409.

17. Hsu, J.C. (1982). Simultaneous inference with respect to the best treatment in block designs. *J. Amer. Statist. Assoc.* **77**, 461-467.

18. Kass, R.E. and Steffey, D. Approximate Bayesian inference in conditionally independent hierarchical models (Parametric empirical Bayes models). *J. Amer. Statist. Soc.* **84,** 717-726.

19. Laird, N.M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika* **65**, p. 581-590.

20. 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,** p. 739-757.

21. Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-974.

22. Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables. *J. Roy. Statist. Soc., Ser. B* **37**, p.23-37.

23. Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimation for the linear model (with discussion). *J. Roy. Statist. Soc., Ser. B* **34**, p. 1-41.

24. Maritz, J. (1970). *Empirical Bayes methods.* Monograph Series on Applied Probability and Statistics. London: Meuthen.

25. Morris, C.N. (1983). Parametric Empirical Bayes inference: Theory and applications (with discussion). *J. Amer. Statist. Assoc.* **78,** p. 47-65.

26. Press, S.J. (1989). *Bayesian Statistics.* New York: John Wiley.

27. Robbins, H.E. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35**, 1-20.

28. Searle, S.R., Casella, G. and McCulloch, C.E. (1991). *Variance Components.* New York: John Wiley.

29. Strenio, J.F., Weisberg, H.I., and Bryk, A.S. (1983). Empirical Bayes estimation of individual growth-curve parameters and their relationship to covariates. *Biometrics* **39,** p. 71-86.

30. Tanner, M.A. and Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82,** 528-550.

Figure 4.1 Bayes prior and posterior for binomial example of Section 4.1. The solid line is $\pi(p)$ and the dashed line is $\pi(p|y)$.

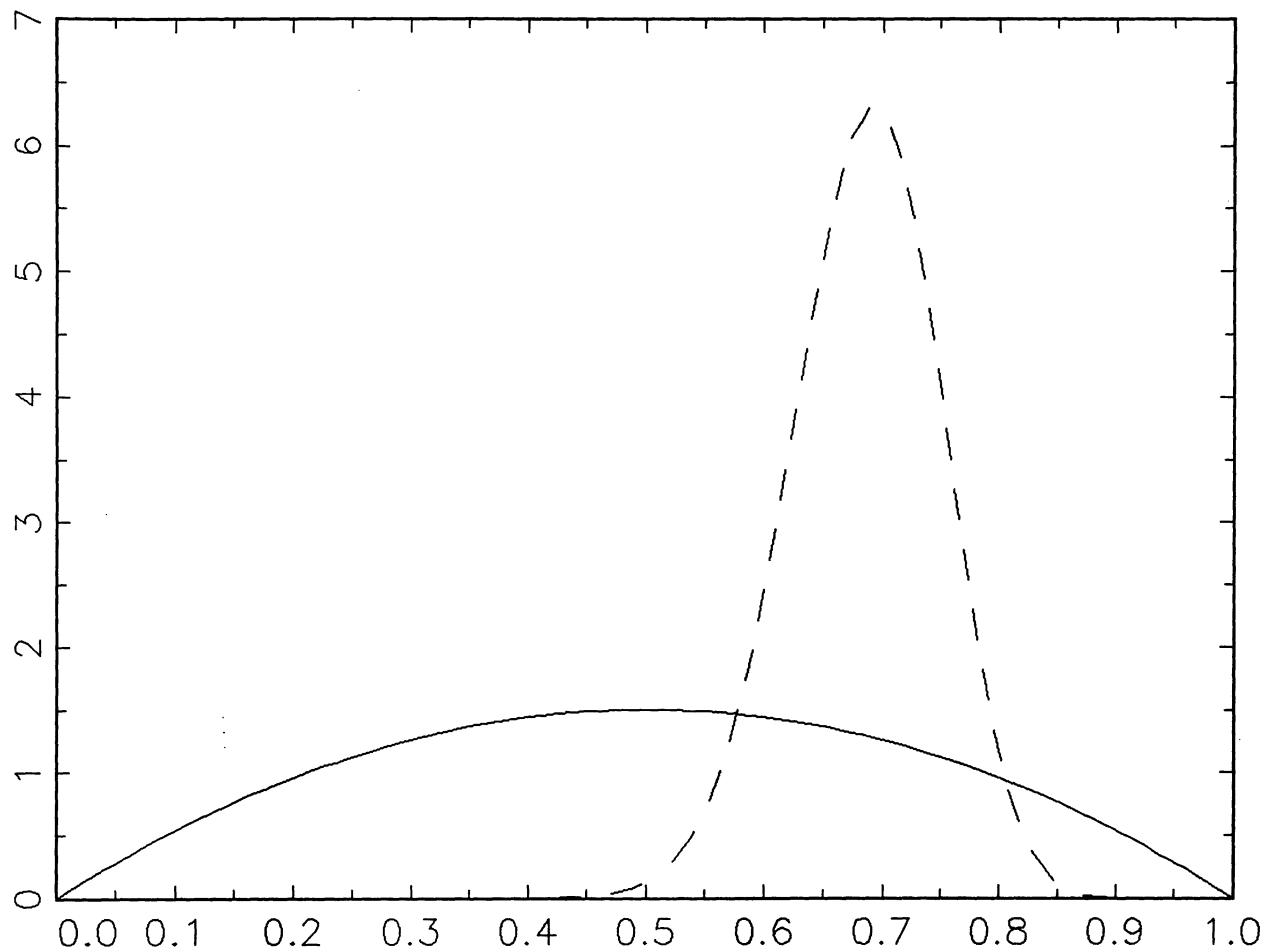

Figure 5.1 Estimates of the cell means for the steer data, Section 5.1. The line represents the submodel $\hat{\theta} = 79.66 - 1.76X$. The crosses are the cell means, and the squares are the EB estimates.

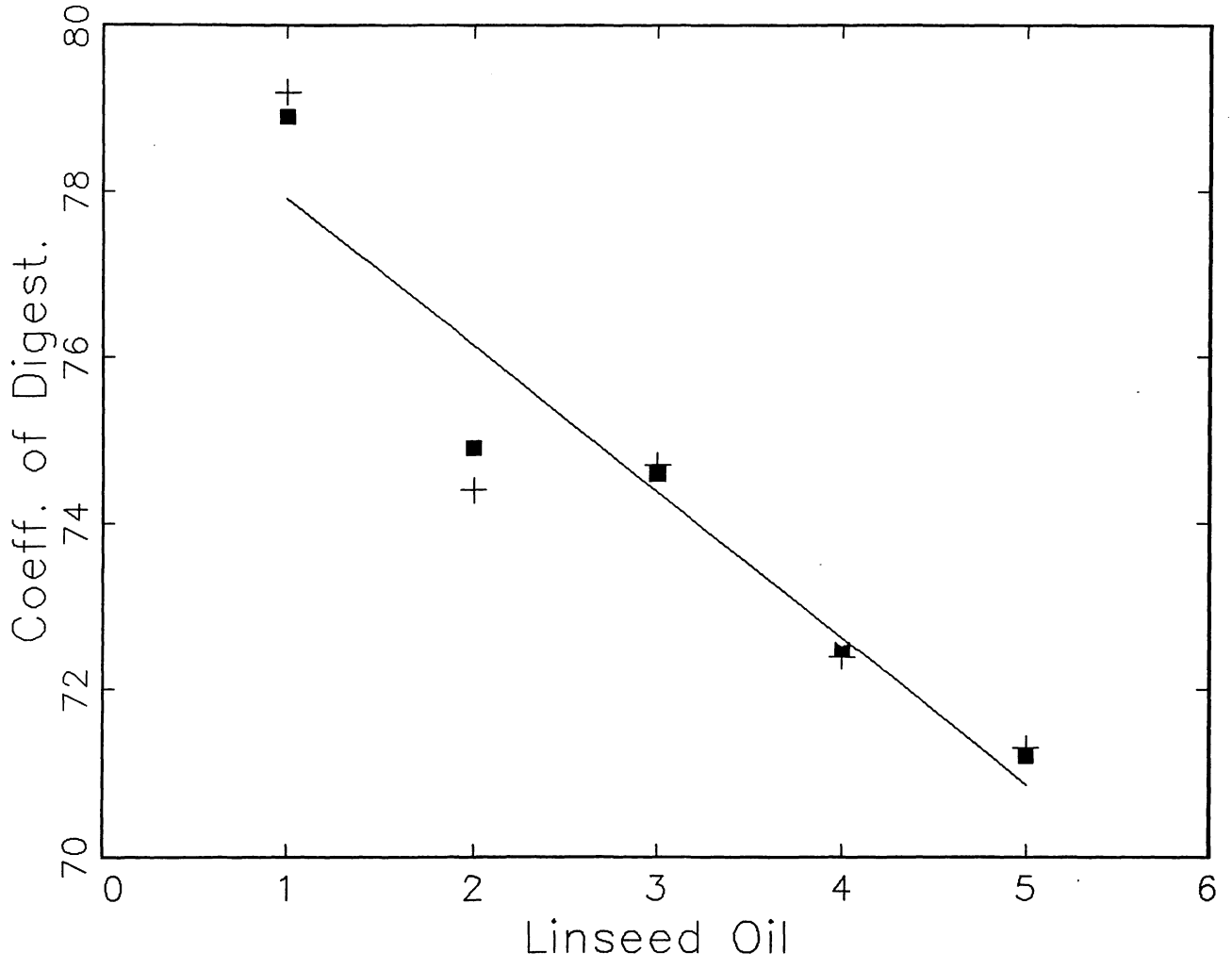

Figure 7.1 Typical growth curve for Strenio, et al. data. Lines pictured are for rat #8. The solid line is the submodel and the long-dashed line is the individual estimate. The EB estimate is the short-dashed line.