RESEARCH REPORT SERIES
(Computing #2008-2)

Computation of Empirical Bayes Estimates Using Single Level Mixed Models

Robert Creecy

Statistical Research Division
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: August 4, 2008

Computation of Empirical Bayes Estimates Using Single Level Mixed Models

Robert Creecy *
U.S. Census Bureau, Washington D.C.

August 4, 2008

Abstract

This paper is a description of the algorithms and code that were developed in R to compute empirical Bayes estimates. The setting is that some variables that have direct survey estimates for some set of geographic areas or domains are given along with estimates of the variance of those variables. In addition, a set of variables that are correlated with the variable of interest is given. An empirical Bayes estimate is then a weighted average of the direct estimate and a regression estimate. It turns out that the empirical Bayes method in this setting is equivalent to a single level mixed model with known variances.

1 Introduction

This paper was motivated by a computational problem that arose in the project 'Exploration of the Use of Empirical Bayes Procedures for Estimating Changes in Occupancy Rate and Persons per Household,' conducted by staff in the Statistical Research Division (see [Weidman (2008)]). State level and national level empirical Bayes estimates were sought for each county in the state or U.S. A difficulty arose because SAS PROC MIXED was unable to run the national models at all, and SAS also did not converge for several of the state models. A custom R program was written to compute the parameter estimates for the models using maximum likelihood and an EM (expectation maximization) algorithm.

2 Models

Using the notation from [Weidman (2008)], let θ_i be the true value for the ith county, $i = 1 \ldots k$, and

*This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the Census Bureau.

$$Y_i = \theta_i + \eta_i \tag{1}$$

be a direct estimate of θ_i from a survey, where the η_i's are independent sampling errors with $E(\eta_i|\theta_i) = 0$ and $V(\eta_i|\theta_i) = V_i$.

In addition, given a vector $z_i' = (z_{i1}, z_{i2}, \ldots, z_{ir})$ of r model variables available for each county i,

$$\theta_i = z_i'\beta + \epsilon_i \tag{2}$$

is a linear regression model, where β is a vector of unknown coefficients, $E(\epsilon_i) = 0$, and $V(\epsilon_i) = A_i$ with the ϵ_i 's independent.

Then, substituting (2) into (1), a model for the direct estimate can be written as

$$Y_i = \theta_i + \eta_i = z_i'\beta + \epsilon_i + \eta_i \tag{3}$$

This can be recognized as a mixed model where the β are the coefficients of the fixed effects z_i and the ϵ_i and η_i are the coefficients for the random effects.

This model can be translated to the notation of [Searle (1992)]. In that notation a mixed model is written as:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{4}$$

The $n \times 1$ vector of observations $\mathbf{y}$ is modeled with both fixed effects, β and random effects $\mathbf{u}$. Here, $\mathbf{X}$ is an $n \times p$ known model matrix derived from the predictor variables, β is a $p \times 1$ vector of fixed effect parameters and $\mathbf{e}$ is an error vector. The mean vector is then $\mu = \mathbf{X}\beta$. The $n \times q$ random effects model matrix $\mathbf{Z}$ can take a variety of forms in different types of mixed models, but it is often an incidence matrix, that is a matrix of 0-1 values describing the relationship between the ith observation and the qth random effect. For this project, there is one random effect for each county, so $\mathbf{u}$ is an $n \times 1$ vector and $\mathbf{Z} = \mathbf{I_n}$ is simply the identity matrix of size n.

The distributional assumptions of the model given in equation (4) are

$$E(\mathbf{e}) = 0 \quad \text{and} \quad var(\mathbf{e}) = \sigma_e^2\mathbf{I_n} \tag{5}$$

$$E(\mathbf{u}) = 0 \quad \text{and} \quad var(\mathbf{u_i}) = \sigma_i^2\mathbf{I_{q_i}} \tag{6}$$

where q_i is the number of levels of the ith random effect. In this case $q_i = 1$ for all i since there is a separate random effect for each observation.

The following table makes the correspondence between the notation of [Weidman (2008)] and [Searle (1992)] explicit

True value	θ_i	$\mu_i + e_i$
Dependent variable	Y_i	y_i
Number observations	k	n
Number of dependent variables	r	p
Independent variables	z_i	x_i
Fixed effects	β	β
Residual error	ϵ_i	e_i
Residual variance	A_i	σ_e^2
Random effects	η_i	u_i
Random effects variances	V_i	σ_i^2

Note that the terms ϵ_i or e_i, that I am calling residual error in analogy to simple linear regression, can also be thought of as county random effects, and are described that way in [Weidman (2008)]. Also, a more general form of the mixed model described in [Weidman (2008)] allows the variance A_i of each residual ϵ_i to be different; in the estimation , this variance is assumed to be constant. ($A_i = A = \sigma_e^2$).

3 Maximum Likelihood Estimation

Searle ([Searle (1992)] in Chapter 6 derives equations for both maximum likelihood estimation (MLE) and restricted maximum liklihood estimation (REML) of the parameters in equation (4) . These equations are the basis for the derivation of EM algorithms described in section 8.3 of Searle. The EM algorithm for the MLE, which is what has been implemented, is illustrated as an example for the 1-way random model in section 8.6. It is a modification of the method described in section 8.6 that takes into account that the random effect variances (V_i or σ_i^2) are assumed to be known. Another variation from the method described in section 8.6 is that an arbitrary model matrix $\mathbf{X}$ ia allowed, not just the constant $\mathbf{X} = 1$ for the mean.

By simplifying and extending the notation for the algorithm in Searle, the following iterative algorithm is obtained:

Input: $\mathbf{y}, \mathbf{X}, \mu^{(\mathbf{0})}, \sigma_{\mathbf{i}}^{\mathbf{2}}, \sigma_{\mathbf{e}}^{\mathbf{2(0)}}$ (all known except $\sigma_e^{2(0)}$ and $\mu^{(0)}$ which are provided as initial (step m=0) estimates of the residual variance and mean vector $\mu = \mathbf{X}\beta$ respectively. In practice, $\sigma_e^{2(0)} = \text{mean}(\sigma_i^2)$ and $\mu^{(0)} = 0$) are used.

At step (m+1), re-estimate the parameters $\beta^{(m+1)}$ and $\sigma_e^{2(m+1)}$ based upon the data and the values of the parameters at the previous step, (m). Also, update estimates of the mean vector $\mu^{(m+1)}$, the random effects vector $\mathbf{u}^{(\mathbf{m+1})}$ and the residual vector, $\mathbf{e}^{(\mathbf{m+1})}$.

$$\beta^{(\mathbf{m+1})} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mu^{(\mathbf{m})}) \quad (7)$$

$$\mu^{(\mathbf{m+1})} = \mathbf{X}'\beta^{(\mathbf{m+1})} \quad (8)$$

$$\mathbf{u}^{(\mathbf{m+1})} = \frac{(\mathbf{y} - \mu^{(\mathbf{m+1})})\sigma_i^2}{(\sigma_i^2 + \sigma_e^{2(m)})} \quad (9)$$

$$\mathbf{e}^{(\mathbf{m+1})} = \mathbf{y} - \mu^{(\mathbf{m+1})} - \mathbf{u}^{(\mathbf{m+1})} \quad (10)$$

$$\sigma_{\mathbf{e}}^{\mathbf{2(m+1)}} = \frac{\mathbf{e}'^{(\mathbf{m+1})}\mathbf{e}^{(\mathbf{m+1})}}{\sum_{i=1}^{n} \frac{\sigma_e^{2(m)}}{(\sigma_i^2 + \sigma_e^{2(m)})}} \quad (11)$$

Iterations proceed until convergence, which is defined as when the absolute relative change of the estimated residual variance changes by less than some tolerance (set to 10^{-8}).

$$\frac{\left|\sigma_e^{2(m+1)} - \sigma_e^{2(m)}\right|}{\sigma_e^{2(m)}} \leq 10^{-8}$$

It may happen that the MLE for σ_e^2 is 0, so iterations also stop if the estimated residual variance becomes less than some epsilon (set to 10^{-12}).

It can be noted that the same algorithm can be used for REML estimation with a simple change of the denominator of equation (11).

The R code that implements the above algorithm is listed in Appendix A. The primary algorithm is `itermix`, but all the code is included. It can be found at `/home/creec003/myR/weidmanmixed/Src` on the SRD research1 computer.

References

[Weidman (2008)] Weidman, L., Malec, D. and Creecy, R. (2008). Exploration of the Use of Empirical Bayes Procedures for Estimating Changes in Occupancy Rate and Persons per Household (Statistical Research Division Technical Report). U.S. Census Bureau.

[Searle (1992)] Searle, S. R., Casella, G. and McCulloch, C. E. (1992). Variance Components. John Wiley and Sons Inc., New York.

4 Appendix A: R Code

```
‘itermix‘ <-
function(y,x,mu,sig2e,sig2i,maxit=1000,tol=1e-8,sigmin=1e-12,trace=FALSE) {
    n<-length(y)
it <- 1
 sig2enew <- sig2e
sig2e <- 0
p<- dim(x)[2]
    xxi <- solve(t(x) %*% x)
    xy <-t(x) %*% y
    xxix <- xxi %*% t(x)
 while ((it <= maxit) && (abs((sig2e-sig2enew)/sig2e) > tol) && ( sig2enew > sigmin )) {
   if (it != 1) {
     beta <- xxix %*% (y-u)
     mu <- x %*% beta
    }
   sig2e <- sig2enew
   u <- uhat(y,mu,sig2e,sig2i)
   eps <- epshat(y,mu,sig2e,sig2i)
   dfe <- n-sum(sig2e/(sig2e+sig2i))
   sig2enew <- sum(eps^2)/(n-dfe)
   if (trace) {
     cat("sig2e ", sig2enew," dfe ",dfe," change ", sig2enew-sig2e,"
     rel change ",abs((sig2e-sig2enew)/sig2e),"\n")
   }
   it <- it+1
 }
converged <-  abs((sig2e-sig2enew)/sig2e) <= tol
dimnames(beta)[[1]][1] <- "intercept"
list(sig2e=sig2enew,beta=beta,mu=mu,u=u,eps=eps,iters=it,converged=converged)
}

‘uhat‘ <-
  function(y,mu,sig2e,sig2i) {
    (y-mu)*sig2i/(sig2e+sig2i)
  }

‘epshat‘ <-
  function(y,mu,sig2e,sig2i) {
    y-mu-uhat(y,mu,sig2e,sig2i)
  }

setwd('/home/creec003/myR/weidmanmixed/apr08')
source('/home/creec003/myR/weidmanmixed/Src/mixfuns.R',
```

```
      echo=TRUE,max.deparse.length=10000)
  library(Hmisc)
  libname <- '/cenhome/tsay0001/tract_level_plan/april_08/robdata'
  if (libname == '') libname <- '.'
  setname <- ''

charstates <-
c("nat","01", "02", "04", "05", "06", "08", "12", "13", "16", "17", "18",
  "19", "20", "21", "22", "23", "24", "25", "26", "27", "28", "29",
"30", "31", "32", "34", "35", "36", "37", "38", "39", "40",
"41", "42", "45", "46", "47", "48", "49", "50", "51", "53", "54",
"55", "56")

# charstates <- "nat"
for (variable in c('occ','pph')) {
  for (state in charstates) {
   if (state  == "nat") {  setname <- paste(state,"_",variable,"_rev",sep="")}
   else
    {setname <- paste("st",state,"_",variable,"_rev",sep="")}
    cat("Reading SAS data set ",setname,"\n")
    mydat <- sas.get(libname,setname)
    k <- dim(mydat)[2]-2
    n <- dim(mydat)[1]
    x<- matrix(1,n,1)
    if (k > 0) x<- cbind(x,as.matrix(mydat[,1:k,drop=FALSE]))
    y<- as.matrix(mydat[,k+1,drop=FALSE])
    sig2i <- as.matrix(mydat[,k+2,drop=FALSE])
    indepnames <-  names(mydat)[1:k]
    depname <- names(mydat)[k+1]
    varname <- names(mydat)[k+2]
    cat("Independent variables (X): ", indepnames,"\n")
    cat("Dependent variable (y): ", depname,"\n")
    cat(" Sampling variance variable (sig2i): ", varname,"\n")
    mu <- 0
    sig2e<-mean(sig2i)
    res<- try(itermix(y,x,mu,sig2e,sig2i))
    if (res$converged) {
      cat("EM iterations converged in ",res$iter," iterations \n")
      cat("Residual variance (sig2e):",res$sig2e,"\n")
      cat("Beta\n")
      print(t(res$beta))
      mydat[paste(depname,".mu",sep="")]<-res$mu
      mydat[paste(depname,".u",sep="")]<-res$u
      mydat[paste(depname,".eps",sep="")]<-res$eps
      write.table(mydat,file=paste(setname,".dat",sep=""),row.names=FALSE)
      myparams <-  as.data.frame(cbind(state=state,sig2e=res$sig2e,t(res$beta)))
```

```
      write.table(myparams,file=paste(setname,".params",sep=""),
        row.names=FALSE,quote=FALSE)
    } else {
      cat("EM iterations DID NOT converge in ",res$iter," iterations \n")
      cat("Residual variance (sig2e):",res$sig2e,"\n")
      cat("Beta\n")
      print(t(res$beta))
    }
    cat("\n \n \n")
  }
}
```