ACT Research Report Series 97-6

Empirical Bayes Estimates of Parameters from the Logistic Regression Model

Walter M. Houston

David J. Woodruff

ACT

August 1997

For additional copies write:
ACT Research Report Series
PO Box 168
Iowa City, Iowa 52243-0168

EMPIRICAL BAYES ESTIMATES OF PARAMETERS FROM THE LOGISTIC REGRESSION MODEL

Walter M. Houston
David J. Woodruff

ABSTRACT

Maximum likelihood and least-squares estimates of parameters from the logistic regression model are derived from an iteratively reweighted linear regression algorithm. Empirical Bayes estimates are derived using an m-group regression model to regress the within-group estimates toward common values. The m-group regression model assumes that the parameter vectors from m groups are independent, and identically distributed, observations from a multivariate normal "prior" distribution. Based on asymptotic normality of maximum likelihood estimates, the posterior distributions are multivariate normal. Under the assumption that the parameter vectors from the m groups are exchangeable, the hyperparameters of the common prior distribution are estimated using the EM algorithm. Results from an empirical study of the relative stability of the empirical Bayes and maximum likelihood estimates are consistent with those reported previously for the m-group regression model. Estimators that use collateral information from exchangeable groups to regress within-group parameter estimates toward a common value are more stable than estimators calculated exclusively from within-group data.

EMPIRICAL BAYES ESTIMATES OF PARAMETERS FROM THE LOGISTIC REGRESSION MODEL

Logistic regression is used in many areas of substantive interest in the social and biological sciences to model the conditional expectation (probability) of a binary dependent variable as a function of an observed (or latent) vector of covariates. In some applications, the parameters of the logistic regression function must be estimated from small samples, which can result in parameter estimates with large sampling variability. For example, ACT uses logistic regression to model the probability of success within specific college courses. Using current estimation procedures, a minimum sample size of 45 within-course observations is required. Further, because the estimation algorithm for the parameters of the logistic regression model is iterative, parameter estimates based on small samples may fail to converge, or converge to local, rather than global, stationary points.

One way to stabilize parameter estimates is to use collateral information from "exchangeable" groups (Lindley, 1971) to refine the within-group estimates. Bayesian m-group regression models have been shown to increase the prediction accuracy and stability of the parameter estimates, relative to estimates that ignore collateral information.

In m-group regression models, the parameter vectors within each of $\underline{m}$ exchangeable groups are assumed to be independent and identically distributed observations with a common probability density function. When this common distribution is treated as a "prior" distribution, posterior distributions and associated inferences follow from standard Bayesian theory (DeGroot, 1970). The effect of the prior distribution is to regress the within-group maximum likelihood estimates toward common values. The extent of the regression effect is inversely related to the precision of the within-group estimates; as the precision of the within-group estimate decreases, the empirical Bayes estimate moves closer to the prior mean. As the precision increases, the maximum likelihood and empirical Bayes estimates converge. Estimators that regress within-group parameter estimates toward common values (often referred to as "regressed" or "shrinkage" estimators) are also found in classical theory (James & Stein, 1961; Evans & Stark, 1996).

The parameters of the prior distribution are often referred to as "hyperparameters", to distinguish them from the parameters of the logistic regression function. There are different methods for estimating the hyperparameters. A fully Bayesian analysis requires a prior density for the hyperparameters (Novick, Jackson, Thayer, and Cole, 1972). The posterior density of the parameter vector is found by integrating the joint posterior density of the parameters and hyperparameters with respect to the hyperparameters. Since these integrals are seldom expressible in a closed form, numerical integration procedures are often required. Markov chain sampling schemes, such as the Gibbs sampler (Tanner,1993), are currently under investigation as an alternative to the numerical methods used previously.

Empirical Bayesian approaches (Braun, Jones, Rubin, & Thayer, 1983; Houston & Sawyer, 1988; Rubin, 1980), in contrast, derive maximum likelihood estimates of the hyperparameters using the EM algorithm (Dempster, Laird, & Rubin; 1977). The estimated hyperparameters are used in calculating the prior and posterior densities for the within-group parameters.

The first section of the paper is concerned with parameter estimation and model fit for the logistic regression model within a single group. An iteratively reweighted linear regression algorithm is used to derive maximum likelihood and least-squares estimates of the parameter vector of the logistic regression function. A chi-square test of the hypothesis of model fit is derived, based on the mean squared error (MSE) from the weighted linear regression model used in the estimation algorithm. The use of a linear regression algorithm makes many of the diagnostic procedures developed for the linear model applicable, with a modification of MSE, to the logistic model, as well.

In the next section, collateral information across m exchangeable groups is used to refine the within-group maximum likelihood estimates. Empirical Bayes estimates are derived from a Bayesian m-group regression model. In this model, the parameter vectors of the logistic regression function are assumed to be independent and identically distributed observations from a multivariate normal distribution.

In the final section of the paper, an application involving college course placement data (grades in specific college courses and standardized achievement test scores) is presented. An empirical study uses these data to compare the stability of the empirical Bayes and maximum likelihood estimates.

The Logistic Regression Model

Model Specification

Let Y_i denote a binary random variable, and let $\mathbf{x}_i$ denote a (p x 1) vector of covariates, for subject i (i = 1 to n). The logistic regression model may be expressed as

$$Y_i = p_i\left(\theta,\mathbf{x}_i\right) + \varepsilon_i, \qquad (1)$$

$$= \frac{\exp\left(\mathbf{x}_i^t\,\theta\right)}{1 + \exp\left(\mathbf{x}_i^t\,\theta\right)} + \varepsilon_i. \qquad (2)$$

The logistic regression function, $p_i(\theta,\mathbf{x}_i)$, has range 0 to 1, with the parameter vector θ and covariate vector $\mathbf{x}_i$ as arguments. It is assumed that the ε_i (i = 1 to n) are stochastically independent, with $E(\varepsilon_i) = 0$. Conditional on $\mathbf{x}_i$ and θ, the Y_i (i = 1 to n) are independent Bernoulli random variables with mean $p_i = p_i(\theta,\mathbf{x}_i)$ and variance = $p_i * (1 - p_i)$.

Both maximum likelihood and least-squares estimates of the parameter vector θ can be derived using an iteratively reweighted linear regression algorithm.

Maximum Likelihood Estimation

The natural logarithm of the likelihood function, denoted ln(L), for a sequence of independent Bernoulli random variables with parameters $p_i(\theta,x_i)$, is given by

$$\ln(L) = \sum_{i=1}^{n} \left\{ y_i \mathbf{x}_i^t \boldsymbol{\theta} - \ln\left[1 + \exp\left(\mathbf{x}_i^t \boldsymbol{\theta}\right)\right]\right\}. \tag{3}$$

The gradient of the function ln(L), referred to as the "score" function, is given by

$$s(\boldsymbol{\theta}) = \frac{\partial \ln(L)}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \mathbf{x}_i\left[y_i - p_i(\boldsymbol{\theta},\mathbf{x}_i)\right], \tag{4}$$

so that $s(\theta)$ is a (p x 1) vector-valued function with argument θ. Let

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^t \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}_n^t \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \text{ and } \mathbf{p}(\boldsymbol{\theta},\mathbf{X}) = \begin{bmatrix} p_1(\boldsymbol{\theta},\mathbf{x}_1) \\ \cdot \\ \cdot \\ \cdot \\ p_n(\boldsymbol{\theta},\mathbf{x}_n) \end{bmatrix}, \tag{5}$$

where $\mathbf{X}$ is an (n x p) matrix with the i^{th} row equal to $\mathbf{x}_i^t$, $\mathbf{y}$ is an (n x 1) column vector with the i^{th} element equal to y_i, and $\mathbf{p}(\theta,\mathbf{X})$ is an (n x 1) column vector with the i^{th} element equal to $p_i(\theta,\mathbf{x}_i)$. Using this notation, equation (4) may be rewritten as

$$s(\theta) = \mathbf{X}^t\left[\mathbf{y} - \mathbf{p}(\theta,\mathbf{X})\right]. \tag{6}$$

Maximum likelihood estimates are obtained by finding the roots of the "estimating" equation $s(\theta) = \mathbf{0}$. For linear models, the elements of $\mathbf{p}(\theta,\mathbf{X})$ are linear functions of θ; consequently, the estimating equation can be solved analytically. For the linear model $\mathbf{y} = \mathbf{X}\theta + \varepsilon$ with i.i.d. normal error terms, $\mathbf{p}(\theta,\mathbf{X})$ is equal to $\mathbf{X}\theta$, with estimating equation given by $\mathbf{X}^t\mathbf{X}\theta = \mathbf{X}^t\mathbf{y}$. Assuming $\mathbf{X}$ has full column rank, the maximum likelihood estimate of θ is equal to $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$.

For nonlinear models, however, the elements of $\mathbf{p}(\theta,\mathbf{X})$ are not expressible as linear functions of θ and iterative procedures are required. Let $\theta^{(h)}$ denote the estimate of θ after iteration $\underline{h}$. The Gauss-Newton method uses a first-order Taylor series to approximate the function $\mathbf{p}(\theta,\mathbf{X})$ by the plane tangent to the surface of $\mathbf{p}(\theta,\mathbf{X})$ at the point $\theta^{(h)}$. The approximation is used in the estimating equation to update the parameter estimate from $\theta^{(h)}$ to $\theta^{(h+1)}$.

The first-order Taylor series expansion of the function $\mathbf{p}(\theta,\mathbf{X})$, about the point $\theta^{(h)}$, is given by

$$\mathbf{p}(\boldsymbol{\theta},\mathbf{X}) \approx \mathbf{p}(\boldsymbol{\theta}^{(h)},\mathbf{X}) + \mathbf{P}(\boldsymbol{\theta}^{(h)},\mathbf{X}) * (\boldsymbol{\theta} - \boldsymbol{\theta}^{(h)}), \tag{7}$$

$$\text{where } \mathbf{P}(\boldsymbol{\theta}^{(h)},\mathbf{X}) = \begin{bmatrix} \dfrac{\partial\, p_1(\boldsymbol{\theta},\mathbf{x}_1)}{\partial\, \boldsymbol{\theta}^t} \\ \cdot \\ \cdot \\ \cdot \\ \dfrac{\partial\, p_n(\boldsymbol{\theta},\mathbf{x}_n)}{\partial\, \boldsymbol{\theta}^t} \end{bmatrix} = \begin{bmatrix} p_1(\boldsymbol{\theta}^{(h)},\mathbf{x}_1)\left[1-p_1(\boldsymbol{\theta}^{(h)},\mathbf{x}_1)\right]\mathbf{x}_1^t \\ \cdot \\ \cdot \\ \cdot \\ p_n(\boldsymbol{\theta}^{(h)},\mathbf{x}_n)\left[1-p_n(\boldsymbol{\theta}^{(h)},\mathbf{x}_n)\right]\mathbf{x}_n^t \end{bmatrix} \tag{8}$$

is the (n x p) Jacobian matrix evaluated at $\theta = \theta^{(h)}$, and $\mathbf{p}(\theta^{(h)},\mathbf{X})$ denotes an (n x 1) vector with the i^{th} element equal to $p_i^{(h)} = p_i(\theta^{(h)},x_i) = \exp(x_i{}^t\theta^{(h)}) / [1 + \exp(x_i{}^t\theta^{(h)})]$. Define an (n x n) diagonal matrix $\mathbf{W}^{(h)}$, with the i^{th} diagonal element given by $w_i^{(h)} = p_i^{(h)} * (1 - p_i^{(h)})$. Using this notation, $\mathbf{P}(\theta^{(h)},\mathbf{X})$ may be written as $\mathbf{W}^{(h)}\mathbf{X}$.

Using the approximation given in expression (7), the estimating equation $s(\theta) = \mathbf{0}$ may be written

$$\mathbf{X}^t\left[\mathbf{y}-\mathbf{p}\left(\theta^{(h)},\mathbf{X}\right)\right] = \mathbf{X}^t\mathbf{W}^{(h)}\mathbf{X}\left(\theta-\theta^{(h)}\right). \tag{9}$$

Let

$$\mathbf{z}^{(h)} = \left(\mathbf{W}^{(h)}\right)^{-1}\left[\mathbf{y} - \mathbf{p}(\boldsymbol{\theta}^{(h)},\mathbf{X})\right] + \mathbf{X}\boldsymbol{\theta}^{(h)}, \tag{10}$$

so that $\mathbf{z}^{(h)}$ denotes an (n x 1) column vector with the i^{th} element given by

Substituting equation (10) into equation (9) yields

$$z_i^{(h)} = \frac{\left(y_i - p_i^{(h)}\right)}{p_i^{(h)} * \left(1 - p_i^{(h)}\right)} + \ln \frac{p_i^{(h)}}{1 - p_i^{(h)}}. \tag{11}$$

$$\mathbf{X}^t\mathbf{W}^{(h)}\mathbf{z}^{(h)} = \mathbf{X}^t\mathbf{W}^{(h)}\mathbf{X}\theta. \tag{12}$$

Equation (12) represents the estimating equation for the weighted linear regression of $z_i^{(h)}$ on $\mathbf{x}_i$, with weight $w_i^{(h)}$.

Thus, the iteratively reweighted linear regression algorithm to derive the within-group maximum likelihood estimate of θ, may be summarized as follows:

1. For observation i (i = 1 to n), calculate

$$p_i^{(h)} = \frac{\exp\left(\mathbf{x}_i^t\theta^{(h)}\right)}{1 + \exp\left(\mathbf{x}_i^t\theta^{(h)}\right)}, \tag{13}$$

$$z_i^{(h)} = \frac{y_i - p_i^{(h)}}{p_i^{(h)} * \left(1 - p_i^{(h)}\right)} + \ln \frac{p_i^{(h)}}{1 - p_i^{(h)}}, \tag{14}$$

and

$$w_i^{(h)} = p_i^{(h)} * \left(1 - p_i^{(h)}\right). \tag{15}$$

2. Regress $z_i^{(h)}$ on $\mathbf{x}_i$ with weight $w_i^{(h)}$ to update the parameter estimate as

$$\theta^{(h+1)} = \left(\mathbf{X}^t\mathbf{W}^{(h)}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{W}^{(h)}\mathbf{z}^{(h)}. \tag{16}$$

3. Iterate between step 1 and step 2 until the change from $\theta^{(h)}$ to $\theta^{(h+1)}$ is less than the convergence criterion.

Let $\hat{\theta}_{ML} = \theta^{(b)}$, where $\hat{\theta}_{ML}$ denotes the maximum likelihood estimate of θ and $\theta^{(b)}$ denotes the estimate of θ following convergence at iteration b. Given certain regularity conditions on the likelihood function, maximum likelihood estimates are asymptotically normal, unbiased, and efficient, with covariance matrix equal to the inverse of Fisher's information matrix (Wilks, 1961). For the ln(L) function in equation (3), the information matrix is equal to $\mathbf{X}^t\mathbf{W}\mathbf{X}$. Therefore, asymptotically,

$$\left(\hat{\theta}_{ML} | \theta\right) \sim N_p \left[\theta, \left(\mathbf{X}^t\mathbf{W}_{(ML)}\mathbf{X}\right)^{-1} \right], \tag{17}$$

where $\mathbf{W}_{(ML)}$ is a diagonal matrix with the i^{th} diagonal element given by $p_i(\hat{\theta}_{ML},\mathbf{x}_i) * [1 - p_i(\hat{\theta}_{ML},\mathbf{x}_i)]$ and $p_i(\hat{\theta}_{ML},\mathbf{x}_i)$ denotes the logistic regression function evaluated at $\theta = \hat{\theta}_{ML}$.

Least-Squares Estimation

The least-squares estimate of $\boldsymbol{\theta}$ (Gallant, 1987) minimizes the real-valued function SSE($\boldsymbol{\theta}$) = $[\ \mathbf{y} - \mathbf{p}(\theta,\mathbf{X})\ ]^t * [\ \mathbf{y} - \mathbf{p}(\theta,\mathbf{X})\ ]$. Setting the gradient of SSE(θ) equal to $\mathbf{0}$ results in the equation

$$\left[\mathbf{P}(\theta,\mathbf{X})\right]^t * \left[\ \mathbf{y} - \mathbf{p}(\theta,\mathbf{X})\ \right] = \mathbf{0}, \tag{18}$$

where $\mathbf{P}(\theta,\mathbf{X}) = \mathbf{W}\mathbf{X}$ is the Jacobian matrix defined in (8). The least-squares estimate of θ, denoted $\hat{\theta}_{LS}$, is such that the residuals $\mathbf{y} - \mathbf{p}(\hat{\theta}_{LS}, \mathbf{X})$ are orthogonal to the columns of the Jacobian matrix, evaluated at $\theta = \hat{\theta}_{LS}$.

Using the tangent plane approximation of $\mathbf{p}(\theta,\mathbf{X})$ in equation (18), and the transformation given in equation (10), results in the equation

$$\mathbf{X}^t\left[\mathbf{W}^{(h)}\right]^2\mathbf{z}^{(h)} = \mathbf{X}^t\left[\mathbf{W}^{(h)}\right]^2\mathbf{X}\theta. \tag{19}$$

Equation (19) represents the estimating equation for the weighted linear regression of $z_i^{(h)}$ on $\mathbf{x}_i$, with weight $w_i^{(h)} = [p_i^{(h)} * (1 - p_i^{(h)})]^2$. Thus, the least-squares estimate $\hat{\theta}_{LS}$ is derived from the iteratively reweighted regression algorithm, defined previously for the maximum likelihood estimates, but with the weights squared. Gallant (1987) shows that the asymptotic distribution of $(\hat{\theta}_{LS} \mid \theta)$ is multivariate normal with mean θ and covariance matrix $\{[\mathbf{P}(\theta,\mathbf{X})]^t[\mathbf{P}(\theta,\mathbf{X})]\}^{-1}$. Thus, for the logistic regression model,

$$\left(\hat{\theta}_{LS} \middle| \theta\right) \sim N_p\left[\theta,\left(\mathbf{X}^t\mathbf{W}_{(LS)}\mathbf{X}\right)^{-1}\right], \tag{20}$$

where $\mathbf{W}_{(LS)}$ is an (n x n) diagonal matrix with the i^{th} diagonal element given by $\{p_i(\hat{\theta}_{LS},\mathbf{x}_i) * [1 - p_i(\hat{\theta}_{LS},\mathbf{x}_i)]\}^2$.

Test of Model Fit

A test of the null hypothesis of model fit uses the residual sum of squares (SSE) from the weighted linear regression of z_i on $\mathbf{x}_i$ with weight w_i. The weighted SSE (Draper & Smith, 1981) may be expressed as

$$\mathrm{SSE} = \sum_{i=1}^{n} w_i(z_i - \hat{z}_i)^2 = \sum_{i=1}^{n} \left[\frac{y_i - \hat{p}_i}{[\hat{p}_i * (1 - \hat{p}_i)]^{\frac{1}{2}}} \right]^2, \tag{21}$$

where

$$\hat{p}_i = p_i(\hat{\theta}_{ML}, x_i), \tag{22}$$

and w_i and z_i are evaluated at $\theta = \hat{\theta}_{ML}$. An equivalent expression for the weighted SSE is given by

$$\mathrm{SSE} = (\mathbf{z} - \hat{\mathbf{z}})^t \mathbf{W} (\mathbf{z} - \hat{\mathbf{z}}) = \mathbf{z}^t \mathbf{A} \mathbf{z}, \tag{23}$$

where

$$\mathbf{A} = \mathbf{W} - \mathbf{W}\mathbf{X}(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W} \tag{24}$$

is an (n x n) projection matrix, with rank(A) = $\underline{n}$ - $\underline{p}$, and the elements of **z** and **W** are evaluated at $\theta = \hat{\theta}_{ML}$. Because the y_i (i = 1 to n) in equation (20) are independent random variables with mean $\hat{p}_i = p_i(\hat{\theta}_{ML}, x_i)$ and variance $\hat{p}_i * (1 - \hat{p}_i)$, the (asymptotic) distribution of SSE is chi-square with $\underline{n}$ - $\underline{p}$ degrees of freedom.

Let MSE = SSE / (n - p). The null hypothesis of model fit is rejected (at level α) if

$$(n-p)*\mathrm{MSE} < \chi^2_{(n-p;\alpha/2)} \quad \text{or} \quad (n-p)*\mathrm{MSE} > \chi^2_{(n-p;1-\alpha/2)}, \tag{25}$$

where $\chi^2(n - p; \alpha/2)$ and $\chi^2(n - p; 1-\alpha/2)$ denote the 100*α/2 percentile and 100*(1-α/2) percentile, respectively, of the chi-square distribution with $\underline{n}$ - $\underline{p}$ degrees of freedom. Under the hypothesis of model fit, the expected value of MSE is equal to 1.

Because parameter estimates are derived from a weighted linear regression algorithm, many of the inferential and diagnostic procedures developed for linear regression models are applicable to the logistic regression model, as well. After parameter estimates have converged, calculate z_i and w_i, and examine statistics from the weighted linear regression of z_i on $\mathbf{x}_i$ with weight w_i. Note, however, that linear regression models make no assumptions about the magnitude of the residual error variance, and estimate it by MSE. The logistic regression model assumes that the residual error variance from the regression of z_i on $\mathbf{x}_i$ with weight w_i is equal to one. So, for example, standard linear model theory uses MSE*$(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}$ to estimate the covariance matrix of $\hat{\theta}_{ML}$, whereas the correct estimate for the logistic regression model is $(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}$. MSE should be set equal to 1 when applying procedures developed for linear models to the logistic regression model.

The Empirical Bayesian M-Group Regression Model

Consider the logistic regression model in (2) for group j (j = 1 to m). Let θ_j denote the parameter vector of the logistic regression function and let $\hat{\theta}_{ML(j)}$ denote the maximum likelihood estimate for group j. Calculation of the posterior density of θ_j requires the sampling distribution of $(\hat{\theta}_{ML(j)} \mid \theta_j)$ and a prior density for θ_j. The density conjugate to the asymptotic normality of within-group maximum likelihood estimates is p-variate normal with hyperparameters $\boldsymbol{\mu}$ and Σ. If the m groups are exchangeable, then the parameter vectors θ_j are assumed to be independent and identically distributed observations from a common "prior" distribution, given by

$$(\boldsymbol{\theta}_j \mid \boldsymbol{\mu},\boldsymbol{\Sigma}) \sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma}),\ j = 1,2,\ldots,m. \tag{26}$$

The assumption of exchangeability enables one to estimate the hyperparameters from the data, and to use the estimates to calculate the posterior density of θ_j.

For the conjugate prior distribution in (26) and the distribution of the within-group maximum likelihood estimates in expression (17), the posterior distribution of $(\theta_j \mid \boldsymbol{\mu}, \Sigma, \hat{\theta}_{ML(j)})$ is p-variate normal with hyperparameters

$$E\left(\theta_j \mid \boldsymbol{\mu},\Sigma,\hat{\theta}_{ML(j)}\right) = \left[\Sigma^{-1} + \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)_j\right]^{-1}\left[\Sigma^{-1}\boldsymbol{\mu} + \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)_j\hat{\theta}_{ML(j)}\right], \tag{27}$$

and

$$\mathrm{COV}\left(\theta_j \mid \boldsymbol{\mu},\Sigma,\hat{\theta}_{ML(j)}\right) = \left[\Sigma^{-1} + \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)_j\right]^{-1}. \tag{28}$$

Note that the posterior mean is a weighted average of $\boldsymbol{\mu}$ and $\hat{\theta}_{ML(j)}$, with weights equal to the precision of the prior density (Σ^{-1}), and the precision of the maximum likelihood estimate $(\mathbf{X}^t\mathbf{W}\mathbf{X})_j$, respectively. The posterior precision is the sum of the sample precision and the prior precision.

The EM algorithm (Dempster, Laird, & Rubin; 1977) is used to derive maximum likelihood estimates of the hyperparameters. Let $\boldsymbol{\mu}_{ML}$ and Σ_{ML} denote the hyperparameter estimates derived from the EM algorithm. In the current application of the EM algorithm, the unobservable parameter vectors θ_j (j=1,2,...,m) are treated as "missing" data. If the θ_j were observed, then a set of "complete" data sufficient statistics for estimating the hyperparameters $\boldsymbol{\mu}$ and Σ is given by

$$\sum_{j=1}^{m}\boldsymbol{\theta}_j \text{ and } \sum_{j=1}^{m}\boldsymbol{\theta}_j\boldsymbol{\theta}_j^t. \tag{29}$$

Because the distribution in (26) is of the regular exponential family, the expectation step of the EM algorithm calculates the expected value of the complete data sufficient statistics with respect to the conditional distribution of the missing data, given the observed data and current hyperparameter estimates. In the present application, this conditional distribution is the posterior distribution of $(\theta_j \mid \boldsymbol{\mu}^{(k)}, \Sigma^{(k)}, \hat{\theta}_{ML(j)})$, where $\boldsymbol{\mu}^{(k)}$ and $\Sigma^{(k)}$ denote estimates of the hyperparameters after the k^{th} iteration. Thus, at the expectation step, and for groups $\underline{j} = 1$ to $\underline{m}$, calculate

$$\mathbf{a}_j = E\left(\theta_j \mid \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \hat{\theta}_{\mathbf{ML(j)}}\right) = \left[\boldsymbol{\Sigma}^{(\mathbf{k})-1} + \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)_j\right]^{-1}\left[\boldsymbol{\Sigma}^{(\mathbf{k})-1}\boldsymbol{\mu}^{(\mathbf{k})} + \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)_j\hat{\theta}_{\mathbf{ML(j)}}\right], \tag{30}$$

and

$$\mathbf{B}_j = E\left(\theta_j\theta_j^t \mid \boldsymbol{\mu}^{(k)}, \Sigma^{(k)}, \hat{\theta}_{\mathbf{ML(j)}}\right) = \left[\left(\Sigma^{(k)}\right)^{-1} + \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)_j\right]^{-1} + \mathbf{a}_j\mathbf{a}_j^t, \tag{31}$$

where $\mathbf{a}_j$ is the posterior mean in equation (27) evaluated at the current hyperparameter estimates.

At the maximization step, the hyperparameter estimates are updated as

$$\boldsymbol{\mu}^{(k+1)} = \frac{1}{m}\sum_{j=1}^{m} \mathbf{a}_j, \tag{32}$$

and

$$\Sigma^{(k+1)} = \frac{1}{m}\sum_{j=1}^{m} \mathbf{B}_j - \boldsymbol{\mu}^{(k+1)}\boldsymbol{\mu}^{(k+1)t}. \tag{33}$$

The EM algorithm iterates between the expectation step, given in (30) and (31), and the maximization step, given in (32) and (33), until the hyperparameter estimates stabilize.

Because the distributions of $(\theta_j \mid \boldsymbol{\mu}^{(k)}, \Sigma^{(k)})$ and $(\hat{\theta}_{ML(j)} \mid \theta_j)$ are both normal, the marginal distribution of $(\hat{\theta}_{ML(j)} \mid \boldsymbol{\mu}^{(k)}, \Sigma^{(k)})$ is given by

$$\left(\hat{\theta}_{\mathbf{ML(j)}} \middle| \boldsymbol{\mu}^{(\mathbf{k})}, \Sigma^{(\mathbf{k})}\right) \sim N_p \left[\boldsymbol{\mu}^{(\mathbf{k})}, \left(\mathbf{X}^t\mathbf{WX}\right)_j^{-1} + \Sigma^{(\mathbf{k})}\right]. \tag{34}$$

In a key result, Dempster, Laird, & Rubin (1977) show that at each iteration of the EM algorithm, the updated hyperparameter estimates, from $\boldsymbol{\mu}^{(k)}$ and $\Sigma^{(k)}$ to $\boldsymbol{\mu}^{(k+1)}$ and $\Sigma^{(k+1)}$, increase the product, across groups, of the ln-likelihood function for the marginal distribution $(\hat{\theta}_{ML(j)} \mid \boldsymbol{\mu}, \Sigma)$. Thus, $\boldsymbol{\mu}_{ML}$ and Σ_{ML} are often referred to as 'marginal' maximum likelihood estimates.

The empirical Bayes estimate, denoted $\hat{\theta}_{BAYES(j)}$, is the mean of the posterior distribution of θ_j, evaluated at $\boldsymbol{\mu} = \boldsymbol{\mu}_{ML}$ and $\Sigma = \Sigma_{ML}$. Thus,

$$\hat{\theta}_{\mathbf{BAYES(j)}} = \left[\Sigma_{\mathbf{ML}}^{-1} + \left(\mathbf{X}^t\mathbf{WX}\right)_j\right]^{-1}\left[\Sigma_{\mathbf{ML}}^{-1}\boldsymbol{\mu}_{\mathbf{ML}} + \left(\mathbf{X}^t\mathbf{WX}\right)_j\hat{\theta}_{\mathbf{ML(j)}}\right], \tag{35}$$

where $\mathbf{W}$ is evaluated at $\theta_j = \hat{\theta}_{BAYES(j)}$.

Application

In support of its user institutions, ACT offers a Course Placement Service (ACT,1994) to assist colleges and universities in making course placement decisions for their first-year students. Placement decision rules, based on one or more test variables, are evaluated in terms of the estimated proportions of correct and incorrect decisions. Logistic regression is used to model the probability of success in the course (defined, for example, as a course grade of a 'B' or better) as a function of test score. Estimated conditional probabilities of success from the fitted logistic regression function are combined with the marginal distribution of test scores to estimate the proportion of correct and incorrect decisions for a given decision rule.

One source of error variance in the estimated proportions is variance associated with estimating the parameters of the logistic regression function. Because the logistic regression function is specific to each course at each institution, small sample sizes are not uncommon. To control error variance in the parameter estimates, and thus in the estimated proportions of correct and incorrect decisions, a minimum sample size of 45 is required. To the extent they reduce error variance in estimating the logistic regression parameters, Bayesian estimation procedures might permit a reduction in the sample size required to achieve an acceptable level of precision.

In 1991, at the request of the Oklahoma State Board of Regents, ACT was asked to establish common placement rules for several college courses, including English Composition, College Algebra, American History, and Physics. Schools with 10 or more within-course observations were included in the analysis. To help satisfy the exchangeability assumption, 2-year and 4-year institutions were analyzed separately. Semester and end-of-year course grade data for the Fall 1991 incoming freshman classes at 14 colleges and universities were matched by social security number against ACT Assessment history files to append ACT Assessment test scores to the course grade data. All subsequent analyses were based on these matched data.

TABLE 1

Empirical Bayes and Maximum Likelihood Parameter Estimates for English Composition at 2-year Institutions

	Estimation procedure			
Number of	Bayes		Maximum likelihood	
observations	Intercept	Slope	Intercept	Slope
91	-5.02	0.28	-6.15	0.34
102	-3.07	0.19	-2.54	0.16
171	-4.54	0.30	-5.78	0.38
61	-4.55	0.28	-5.95	0.37
136	-4.09	0.25	-4.28	0.26
58	-2.61	0.16	-1.42	0.10
247	-3.50	0.23	-3.45	0.23
152	-3.97	0.22	-3.95	0.21
179	-4.21	0.23	-4.20	0.23
82	-3.03	0.18	-1.94	0.13
43	-4.45	0.25	-6.47	0.35
110	-5.92	0.33	-8.46	0.47
245	-5.63	0.28	-7.19	0.35
336	-2.04	0.10	-1.62	0.08

Logistic regression was used to model the probability of success in the course, defined as a course grade of 'B' or better, as a function of ACT Composite score. Table 1 presents empirical Bayes and maximum likelihood estimates of θ_j (j=1,2,...,14) for the English

Composition course at 2-year institutions. Estimates of the hyperparameters of the common prior distribution (μ and Σ) are reported in Table 2.

TABLE 2

Marginal Maximum Likelihood Estimates of the Hyperparameters of the Prior Distribution for English Composition at 2-year Institutions

Regression		Covariance (Σ)	
coefficient	Mean (μ)	Intercept	Slope
Intercept	-4.05	2.07	-0.11
Slope	0.23	-0.11	0.01

The empirical Bayes estimate of θ_j is a weighted average of the prior mean μ, and the within-group maximum likelihood estimate $\hat{\theta}_{ML(j)}$, with weights equal to the prior precision and the precision of $\hat{\theta}_{ML(j)}$, respectively. The appendix reports empirical Bayes and maximum likelihood parameter estimates, and estimates of the hyperparameters of the prior distribution, for the other courses and type of institution (2-year and 4-year).

To compare the relative stability of the empirical Bayes and maximum likelihood estimates, the within-course data at each institution was randomly split in half. Empirical Bayes and maximum likelihood estimates of the logistic regression parameter vector were calculated for each half separately. Let ${}_1\hat{\theta}_{BAYES(j)}$ and ${}_2\hat{\theta}_{BAYES(j)}$ denote the empirical Bayes estimates, and let ${}_1\hat{\theta}_{ML(j)}$ and ${}_2\hat{\theta}_{ML(j)}$ denote the two maximum likelihood estimates, calculated for each half. The Euclidean distance between the estimates from each half was calculated and then summarized across institutions. The Euclidean distances are given by

$$D_{BAYES(j)} = \left[\left({}_1\hat{\theta}_{BAYES(j)} - {}_2\hat{\theta}_{BAYES(j)}\right)' * \left({}_1\hat{\theta}_{BAYES(j)} - {}_2\hat{\theta}_{BAYES(j)}\right)\right]^{\frac{1}{2}}, \tag{36}$$

$$D_{ML(j)} = \left[\left({}_1\hat{\theta}_{ML(j)} - {}_2\hat{\theta}_{ML(j)}\right)' * \left({}_1\hat{\theta}_{ML(j)} - {}_2\hat{\theta}_{ML(j)}\right)\right]^{\frac{1}{2}}. \tag{37}$$

TABLE 3

Euclidean Distances for the Bayes and Maximum Likelihood Estimation Procedures, within Course and Type of Institution.

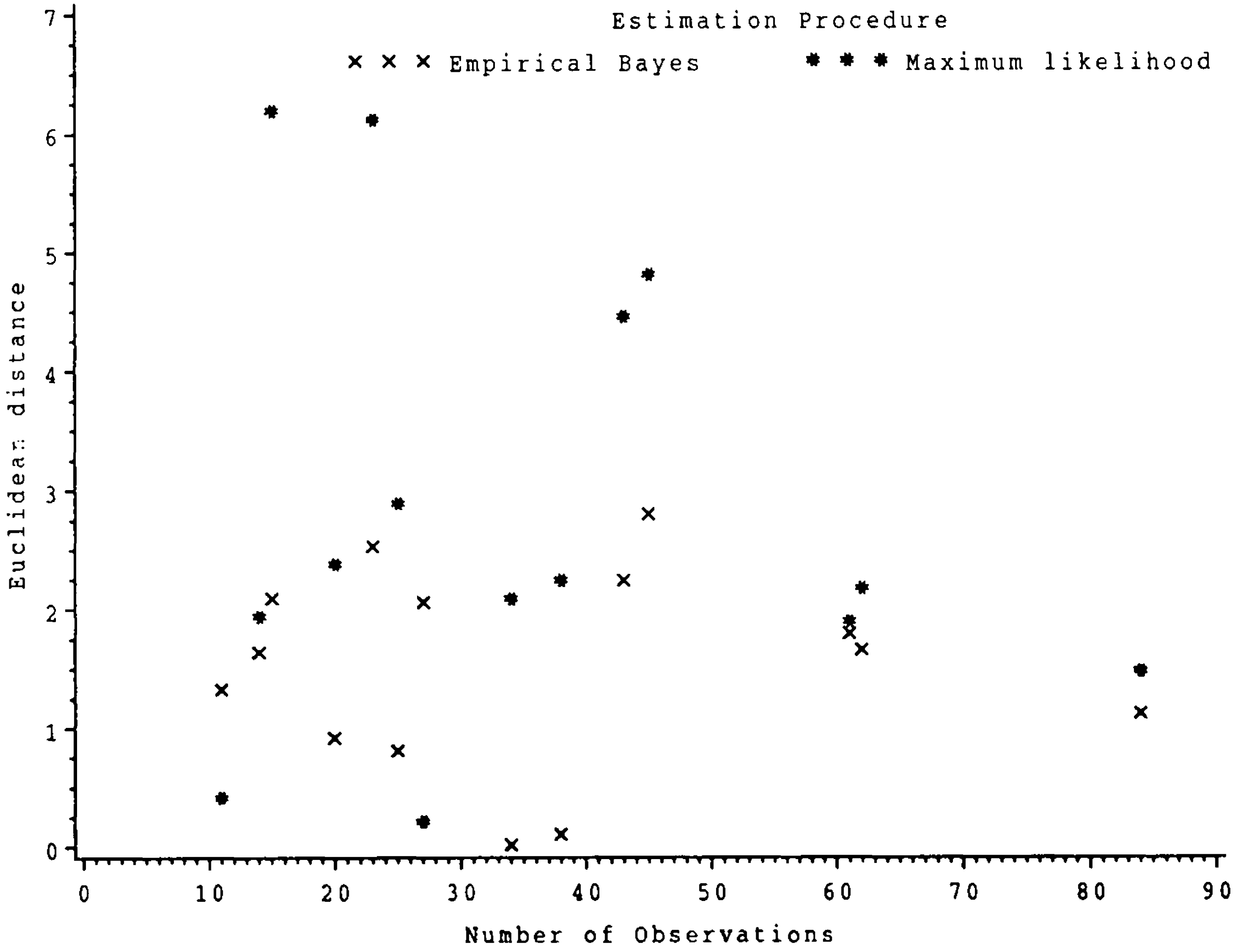
	Distance Measure				
	Number of colleges	D_{BAYES}		D_{ML}	
Course (Type)		Mean	SD	Mean	SD
English (2-year)	14	1.51	0.84	2.80	1.89
English (4-year)	10	0.91	0.64	1.48	1.07
Algebra (2-year)	10	0.67	0.57	4.62	3.74
Algebra (4-year)	8	0.70	0.42	1.32	1.35
History (2-year)	11	2.85	1.35	3.35	2.65
History (4-year)	9	0.71	0.59	3.57	5.01
Physics (2-year)	5	1.23	0.11	2.63	1.03
Physics (4-year)	5	4.51	0.16	9.46	8.11

Note. D_{BAYES} and D_{ML} represent the Euclidean distance between vectors of parameter estimates calculated from within-course data randomly split in half.

Table 3 presents the mean and standard deviation of $D_{BAYES(j)}$ and $D_{ML(j)}$, across the m groups within each course (English Composition, Algebra, American History, and Physics) and type of institution (2-year and 4-year). The mean and standard deviation of $D_{BAYES(j)}$ are consistently less than the mean and standard deviation of $D_{ML(j)}$, for every course and type of institution. These results suggest the effectiveness of the m-group regression procedure in stabilizing parameter estimates, relative to within-group maximum likelihood estimates.

FIGURE 1

Euclidean Distance between Estimated Parameter Vectors as a Function of Sample Size



A plot of the Euclidean distances ($D_{BAYES(j)}$ and $D_{ML(j)}$) versus sample size is presented in Figure 1 for the English Composition course at 2-year institutions. Figure 1 suggests that the m-group regression procedure might permit a reduction in the minimum sample size requirement, without sacrificing the precision of the estimates using the current estimation procedure and sample size requirement. Figure 1 also suggests that the relative advantage of the m-group regression procedure over within-group maximum likelihood estimation, in terms of increased stability in the estimates, is a decreasing function of sample size, with the m-group regression procedure more advantageous as within-group sample sizes decrease.

References

American College Testing. (ACT, 1994). ACT Assessment Course Placement Service,Interpretive Guide. Iowa City, Iowa: ACT, Inc.

Braun, H.T., Jones, D.H., Rubin, D.B., & Thayer, D.T. (1983). Empirical Bayes estimation of coefficients in the general linear model from data of deficient rank. Psychometrika 48, 171-181.

DeGroot, M.H. (1970). Optimal Statistical Decisions. New York: McGraw-Hill.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society (B), 39 1-38.

Draper, N.R. & Smith, H. (1981). Applied Regression Analysis. New York: John Wiley and Sons.

Evans, S.N. & Stark, P.B. (1996). Shrinkage estimators, Skorokhod's problem, and stochastic integration by parts. Annals of Statistics, 24(2), 809-815.

Gallant, A.R. (1987). Nonlinear statistical models. New York: John Wiley and Sons.

Houston, W.M., & Sawyer, R.L. (1988). Central prediction systems for predicting specific course grades. (ACT Research Report No. 88-4). Iowa City, Iowa: ACT, Inc.

James, W., & Stein, C. (1961). Estimation with quadratic loss. Proc. Fourth Berkeley Symp. Math. Statist. Probab. 1, 361-380. University of California Press.

Lindley, D.V. (1971). The estimation of many parameters. In V.P. Godambe & D.A. Sprott (eds.) Foundations of statistical inference. Toronto: Holt, Rinehart, & Wilson.

Novick, M.R., Jackson, P.H., Thayer, D.T., & Cole, N.S. (1972). Estimating multiple regressions in m-groups: a cross-validation study. British Journal of Mathematical and StatisticalPsychology, 75, 33-50.

Rubin, D.B. (1980). Using empirical Bayes techniques in the law school validity studies. Journal of the American Statistical Association, 75, 801-816.

Tanner, M.A. (1993). Tools for Statistical Inference. New York: Springer-Verlag.

Wilks, S.S. (1961). Mathematical Statistics. John Wiley & Sons: New York.

Appendix

Parameter and Hyperparameter Estimates, by Course and Type of Institution.

Table A1 reports the empirical Bayes and maximum likelihood parameter estimates of θ_j (j=1 to 10) for the English Composition course at 4-year institutions. Table A2 reports the estimated hyperparameters (μ and Σ) of the prior distribution for θ_j. Table A3 and Table A4 report parameter and hyperparameter estimates, respectively, for the College Algebra course at 2-year institutions. Table A5 and Table A6 report parameter and hyperparameter estimates, respectively, for the College Algebra course at 4-year institutions. Table A7 and Table A8 report parameter and hyperparameter estimates, respectively, for the American History course at 2-year institutions. Table A9 and Table A10 report parameter and hyperparameter estimates, respectively, for the American History course at 4-year institutions. Table A11 and Table A12 report parameter and hyperparameter estimates, respectively, for the Physics course at 2-year institutions. Table A13 and Table A14 report parameter and hyperparameter estimates, respectively, for the Physics course at 4-year institutions. The number of institutions ranged from 5, for both Physics' courses, to 12, for the American History course at 2-year institutions).

TABLE A1

Empirical Bayes and Maximum Likelihood Parameter Estimates for English Composition at 4-year Institutions

	Estimation procedure			
Number of observations	Bayes		Maximum likelihood	
	Intercept	Slope	Intercept	Slope
180	-2.37	0.12	-1.95	0.10
266	-3.91	0.22	-4.00	0.22
186	-3.57	0.21	-3.69	0.22
146	-4.43	0.26	-5.85	0.35
192	-3.15	0.19	-3.12	0.19
156	-6.35	0.34	-8.67	0.47
696	-1.54	0.10	-1.35	0.09
34	-3.62	0.21	-4.22	0.24
146	-5.22	0.26	-6.20	0.30
1053	-0.86	0.10	-0.62	0.09

TABLE A2

Marginal Maximum Likelihood Estimates of the Hyperparameters of the Prior Distribution for English Composition at 4-year Institutions

		Covariance (Σ)	
Regression coefficient	Mean (μ)	Intercept	Slope
Intercept	-3.51	3.39	-0.16
Slope	0.20	-0.16	0.01

TABLE A3

Empirical Bayes and Maximum Likelihood Parameter Estimates for Algebra at 2-year Institutions

	Estimation procedure			
	Bayes		Maximum likelihood	
Number of observations	Intercept	Slope	Intercept	Slope
34	-2.95	0.15	-3.50	0.17
49	-2.71	0.17	-3.59	0.24
41	-2.98	0.15	-6.38	0.31
49	-2.87	0.15	-2.61	0.14
72	-2.77	0.16	-1.55	0.10
61	-2.82	0.16	-5.94	0.33
32	-2.86	0.15	-0.24	0.02
26	-2.94	0.15	-4.20	0.20
176	-2.99	0.15	-3.25	0.16
157	-3.06	0.14	-2.94	0.13

TABLE A4

Marginal Maximum Likelihood Estimates of the Hyperparameters of the Prior Distribution for Algebra at 2-year Institutions

		Covariance (Σ)	
Regression coefficient	Mean (μ)	Intercept	Slope
Intercept	-2.89	0.06	-0.01
Slope	0.15	-0.01	0.01

TABLE A5

Empirical Bayes and Maximum Likelihood Parameter Estimates for Algebra at 4-Year Institutions.

	Estimation procedure			
	Bayes		Maximum likelihood	
Number of observations	Intercept	Slope	Intercept	Slope
113	-3.85	0.18	-4.07	0.19
121	-4.67	0.21	-6.39	0.29
102	-4.50	0.20	-5.53	0.25
67	-3.76	0.17	-5.98	0.31
62	-3.84	0.17	-3.97	0.18
549	-4.00	0.16	-3.85	0.15
94	-4.47	0.20	-5.77	0.26
772	-1.67	0.09	-1.29	0.08

TABLE A6

Marginal Maximum Likelihood Estimates of the Hyperparameters of the Prior Distribution for Algebra at 4-year Institutions

		Covariance (Σ)	
Regression coefficient	Mean (μ)	Intercept	Slope
Intercept	-3.85	1.40	-0.06
Slope	0.17	-0.06	0.01

TABLE A7

Empirical Bayes and Maximum Likelihood Parameter Estimates for American History at 2-year Institutions.

	Estimation procedure			
	Bayes		Maximum likelihood	
Number of observations	Intercept	Slope	Intercept	Slope
38	-3.58	0.20	-3.23	0.19
90	-2.75	0.18	-2.34	0.15
121	-2.45	0.18	-3.73	0.27
47	-5.68	0.27	-7.38	0.36
45	-3.24	0.17	-0.96	0.04
84	-3.34	0.21	-4.31	0.27
69	-3.93	0.21	-3.64	0.20
25	-5.34	0.26	-16.32	0.84
34	-5.65	0.26	-5.32	0.23
57	-4.69	0.25	-7.54	0.41
147	-7.35	0.33	-9.60	0.43
288	-4.83	0.23	-4.70	0.23

TABLE A8

Marginal Maximum Likelihood Estimates of the Hyperparameters of the Prior Distribution for American History at 2-year Institutions

		Covariance (Σ)	
Regression coefficient	Mean (μ)	Intercept	Slope
Intercept	-4.40	2.92	-0.11
Slope	0.23	-0.11	0.01

TABLE A9

Empirical Bayes and Maximum Likelihood Parameter Estimates for American History at 4-year Institutions.

	Estimation procedure			
	Bayes		Maximum likelihood	
Number of observations	Intercept	Slope	Intercept	Slope
98	-5.37	0.24	-5.70	0.26
143	-6.48	0.26	-8.82	0.36
34	-5.81	0.25	-8.87	0.37
53	-4.36	0.22	-14.06	0.81
72	-4.97	0.23	-3.71	0.17
109	-4.53	0.23	-8.18	0.42
399	-5.40	0.24	-5.09	0.23
26	-5.60	0.24	-6.52	0.28
66	-4.37	0.22	-2.89	0.16
386	-5.61	0.24	-5.35	0.23

TABLE A10

Marginal Maximum Likelihood Estimates of the Hyperparameters of the Prior Distribution for American History at 4-year Institutions

		Covariance (Σ)	
Regression coefficient	Mean (μ)	Intercept	Slope
Intercept	-5.25	0.58	-0.11
Slope	0.24	-0.01	0.01

TABLE A11

Empirical Bayes and Maximum Likelihood Parameter Estimates for Physics at 2-year Institutions.

	Estimation procedure			
	Bayes		Maximum likelihood	
Number of observations	Intercept	Slope	Intercept	Slope
84	-3.05	0.16	-2.69	0.15
30	-3.07	0.16	-1.84	0.08
41	-3.13	0.16	-3.78	0.16
52	-3.13	0.16	-6.23	0.34
46	-3.05	0.16	-1.63	0.08

TABLE A12

Marginal Maximum Likelihood Estimates of the Hyperparameters of the Prior Distribution for Physics at 2-year Institutions

		Covariance (Σ)	
Regression coefficient	Mean (μ)	Intercept	Slope
Intercept	-3.09	0.08	-0.01
Slope	0.16	-0.01	0.01

TABLE A13

Empirical Bayes and Maximum Likelihood Parameter Estimates for Physics at 4-year Institutions.

	Estimation procedure			
	Bayes		Maximum likelihood	
Number of observations	Intercept	Slope	Intercept	Slope
79	-4.97	0.24	-5.97	0.28
29	-4.85	0.24	-6.30	0.30
47	-4.05	0.20	-2.38	0.14
60	-4.46	0.22	-3.14	0.15
74	-4.66	0.23	-5.18	0.25

TABLE A14

Marginal Maximum Likelihood Estimates of the Hyperparameters of the Prior Distribution for Physics at 4-year Institutions

		Covariance (Σ)	
Regression coefficient	Mean (μ)	Intercept	Slope
Intercept	-4.60	0.47	-0.02
Slope	0.23	-0.02	0.01