BAYESIAN M-GROUP REGRESSION: A SURVEY AND AN IMPROVED MODEL

Ivo W. Molenaar and Charles Lewis
Faculteit Sociale Wetenschappen R.U.Groningen
Adapted from Psychometric Society Paper, Uppsala.

0. Summary

When multiple regression equations are to be estimated for m groups which are supposed to be comparable though not identical, both the pooled estimates and m separate least squares estimates per group may be sub-optimal. Lindley, Novick, Jackson and others have advocated a Bayesian estimation procedure in which the estimates would be weighted averages of the separate estimates per group on one hand and some pooled estimate on the other hand, with weights determined essentially by the data. This extension of the Kelley formula for regression to the mean has proven its value in several cross-validation studies (Novick, Jackson, Thayer & Cole, 1972; Lissitz and Schoenfeldt, 1974; Shigemasu, 1976; Jansen, 1977). The modal posterior values for intercepts, slopes and residual variances, however, are not easy to obtain. The procedure outlined by Novick c.s. still poses some numerical and methodological problems. The present paper presents a modified algorithm removing most of the deficiencies. It remains true, however, that m-group regression is an example of a Bayesian model in which it is somewhat difficult to specify a vague prior that would let the data and the collateral information speak for themselves.

1. A simple example

Roelofs and Koppelaar (1978) commented on the data reproduced (with their permission) in Figure 1. For six persons the relation is displayed between blood pressure (y) and muscle tension (x); the subjects have 18, 23, 24, 22, 24 and 17 measurements pairs respectively. For pedagogical reasons we have left out six outlying pairs for subject 1, two for subject 4 and seven for subject 6 from the Roelofs-Koppelaar data.

Suppose it is desired to predict y at $x=x_o$, say, for subject number 3. We would use a least squares regression line based on his 24 (y, x) pairs. Prediction will be inaccurate, because of low sample size and large residual variance. Can the data from the other five subjects be of any help in improving the prediction?

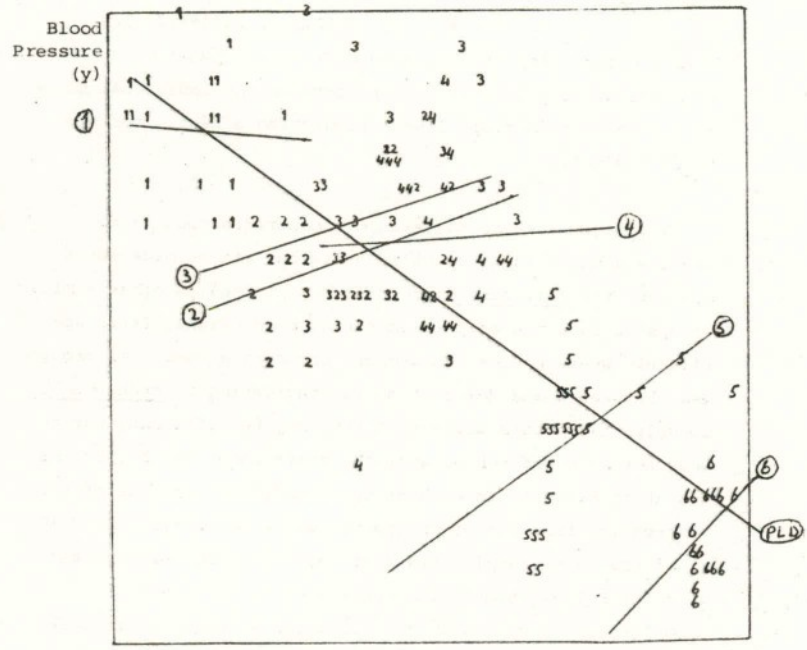

Fig. 1. Relation between Blood Pressure (y) and Muscle Tension (x) for six subjects (numbered 1, 2..., 6) with 18, 23, 24, 22, 24 and 17 observations respectively.

If there were virtually no differences between subjects, the regression line obtained by pooling all 143 observations would certainly be an improvement. One glance at the graph suffices, however, to show that such a pooled estimate would perform poorly in our case. Although the intercepts are certainly different, it looks as if the slopes of the six regression lines within subjects have a lot in common. If y_{ij} denotes the j-th observation of y for the i-th patient, we might fit a model

$$y_{ij}=\alpha_i + \beta\, x_{ij} \qquad (i=1,2,\ldots,6;j=1,2,\ldots n_i)$$

with a common slope β. And even if a common slope is too strong an assumption, we might use a slope between the individual slope β_3 and the common slope β when predicting a next y for the third subject.

The major idea of Bayesian m-group regression is to provide a model in which valid decisions can be made about how much the collateral information contained in other similar groups of data can help in improving predictions. It is applicable whenever the researcher considers a number of groups (say between 3 and 25) that he is considering as exchangeable: roughly stated this means that they may be different, but he does not know beforehand what the differences may be. If the amount of data within each group is moderate, the use of the information from the other groups may be beneficial. In our blood pressure example, and also in estimating true scores with the Kelley formula for regression to the mean, each "group" is an individual for which a number of measurements is available. In the remainder of this paper, and in many educational applications, each "group" is a school or school class and we have one observation per individual pupil. The mathematical model can be applied to both situations.

Although detailed results would take too much space, let us report briefly on the blood pressure example here. A modest cross-validation study showed that on all relevant error measures the Bayesian estimates were indeed superior (though not much) to the LS estimates, and both performed far better than the totally pooled estimates.

2. Two versions of the formal model

The main idea for incorporating collateral information is that a first stage of the model, explaining how the data are distributed given the regression parameters, is followed by a second stage, in which the (not directly observable) values of these parameters across groups are treated as a random sample from some distribution which is characterized by unknown hyper-parameters. It turns out to be necessary to specify some, rather vague, information on those hyperparameters in a third stage of the model.

Two such models are summarized below. The old one (to the left) is discussed in more detail by Novick c.s. (1972), Jones & Novick (1972) and related papers. The new one (to the right) was built when convergence problems and robustness problems arose, as explained in sections 3, 4; see also Molenaar (1978) and Molenaar and Lewis (1979). In both models the data for the n_i individuals of the i-th group (i=1, 2, ..., m) consist of a criterion score y_{ij} and scores on ℓ predictors x_{kij} ($k = 1, 2, \ldots, \ell$; $j = 1, 2, \ldots, n_i$). In each $(\ell + 1) \times n_i$ matrix X_i' of predictor scores we include a row of ones for the intercept. For the new model the index set $\{0, 1, \ldots, \ell\}$ is partitioned into two disjoint subsets F (parameters common to all groups) and G (parameters different across groups).

OLD MODEL	NEW MODEL
First stage:	
$y_{ij} \overset{\Lambda}{\sim} \mathcal{N}(\sum_{k=0}^{\ell} \beta_{ki} x_{kij};\ \phi_i)$	$y_{ij} \overset{\Lambda}{\sim} \mathcal{N}(\sum_{f \in F} \beta_f x_{fij} + \sum_{g \in G} \beta_{gi} x_{gij};\ \phi)$
Second stage:	
$(\beta_{0i}, \beta_{1i}, \ldots \beta_{\ell i}) \overset{\Lambda}{\sim} \mathcal{N}(\mu, H^{-1})$	$\beta_f \overset{\sim}{\sim}$ uniform $(-\infty, \infty)$;
$\phi_i \overset{\Lambda}{\sim} \chi^{-2}(\nu, \nu_0^2)$	$\beta_{gi} \overset{\sim}{\sim} \mathcal{N}(\mu_g, \Psi_g)$;
	$\log\phi \overset{\sim}{\sim}$ uniform $(-\infty, \infty)$;

OLD MODEL	NEW MODEL
Third stage:	
μ, ν and $\log \sigma^2 \triangleq$ uniform $(-\infty,\infty)$;	$\mu_g \triangleq$ uniform $(-\infty,\infty)$;
$H \triangleq$ Wishart $(\nu',\Sigma, \ell + 1)$;	$\psi_g \triangleq \chi^{-2}(\nu',\nu'\tau_g)$.
Σ diagonal matrix.	
User should supply (see below):	
ν' (small)	ν' (small)
diagonal elements of Σ	τ_g $(g = 0, 1, \ldots \ell)$
small κ (to prevent divergence)	

Lack of space forces us to just add that many, hopefully obvious, independence assumptions must be added; they are detailed in the sources already quoted. These sources also tell how integration over the hyperparameters leads to a posterior density for the regression parameters given the data. Up to an additive constant, its logarithm is, for the old model (Lindley, 1970, formula 11):

$$\log p(\{\beta_{ki}\},\{\phi_i\}) =$$
$$-\sum_i (\tfrac{1}{2}n_i + 1)\log\phi_i - \tfrac{1}{2}\sum_{ij}(y_{ij} - \sum_h \beta_{hi}x_{hij})^2/\phi_i \qquad (1)$$
$$-\tfrac{1}{2}(\nu' + m - 1)\log\left|\nu'\sigma_{hk} + \sum_i (\beta_{hi} - \beta_{h.})(\beta_{ki} - \beta_{k.})\right|$$
$$-\tfrac{1}{2}(m + 1)\log\log\{\eta(\theta^{-1} + \kappa)\}.$$

Here θ and η denote the harmonic and geometric mean of the set $\{\phi_i\}$, respectively, $\beta_{h.}$ denotes the mean across i of β_{hi} and $|a_{hk}|$, say, denotes the determinant of an $(\ell + 1) \times (\ell + 1)$ matrix A with elements a_{hk}. For ℓ predictors and m groups, (1) is a function of $(\ell + 2)m$ parameters. Its maximization leads to the desired posterior modal estimates, but it poses some problems.

3. Problems of the old model

The computer programs made available by Novick c.s. seek the maximum of (1) by the following iterative procedure. An initial set of estimates should be computed first; one might take the least squares estimates per group, the least squares estimates for the pooled sample or the so-called model II estimates, see below. Equating the derivatives of (1) with respect to β_{hi} to zero, for fixed i, leads to a set of equations which are linear in β_{hi} if one temporarily considers $\beta_{h.}$ (h = 0, 1, ... ℓ), ϕ_i and the determinant as fixed. They are successively solved for each i; after updating means and determinant this is repeated twice. Next the updated values for all β_{hi} are used to obtain new ϕ_i by equating the derivative of (1) with respect to ϕ_i to zero; such equations are linear in $1/\phi_i$ provided that η, θ and all β_{hi} are temporarily considered as fixed. This whole process is called one iteration cycle, and such cycles should be repeated until the increase per cycle of the function (1) has become negligible.

This algorithm has been used in several applications mentioned in section 0, but not without problems:

(a) very slow convergence;
(b) non-robustness against choice of prior values for ν' and σ_{hh};
(c) non-robustness against choice of initial estimates;
(d) suboptimal determination of the mean value $\beta_{h.}$ for regression parameters for which almost total regression takes place.

4. Improvements and simplifications

As described in Molenaar (1978), convergence can be speeded by the insertion of "leaps" after a user-specified number of iteration cycles. A "leap" is the prediction of an asymptotic value by fitting a geometric series to the two consecutive differences between the parameter values obtained at the last three cycles, (Aitken extrapolation), with a special provision for cases when this series would

lead to an obviously wrong result. Such leaps were first calculated for each of the $(\ell + 2)m$ individual parameters. Later residual variances, and variances across groups of the regression parameters, turned out to be very stable across cycles. As the same was true for the z-scores obtained by standardization across groups of the parameters of individual groups, leaps are now carried out for the means across groups of regression parameters only. They are very successful in reducing the number of cycles needed for convergence.

The deficiencies (c) and (d) above are related to (b) in the following sense. It is obvious that the first line of (1) would be maximized by the least squares (LS) values. The second line is maximized by bringing the determinant as close to zero as possible. When the user has supplied some small values for $\nu'\sigma_{hh}$, this is achieved by linear dependence among the m-vectors β_h $(h = 0, 1, \ldots \ell)$. Now as soon as the estimated values of β_{hi} for some h lie very close together (almost total regression), a change in their deviations from the mean $\beta_{h.}$ has almost no further influence on the residual sum of squares in the first line of (1), and thus it is used to make the determinant decrease. In other words, it pays to let the $(\ell + 1)$-variate normal distribution of the β_h degenerate into a lower-dimensional one. Although the positive value of $\nu'\sigma_{hh}$ prevents complete degeneration, the algorithm based on the old model is deficient: because of the group-by-group calculation of new $\{\beta_{hi}\}$ a change in $\beta_{hi}-\beta_{h.}$ has far more effect on the log posterior density than a change in the mean $\beta_{h.}$, and the optimal value for $\beta_{h.}$ is never found for indices h with small variance across groups. As the empirical results for some datasets did indeed show such undesirable behavior, the revised program uses a common value across groups for any regression parameter for which the prior variance, or the calculated variance across groups beyond cycle 2, is less than some user-specified bound TAUMIN. This bound should be so small that the effect of further changes on the residual sum of squares, be it the observed one or in cross-validation, is almost negligible.

Two further simplifications in the new model are (a) homoscedasticy not only within but also across groups; (b) independence of the priors for all regression parameters. The log pos-

terior density, again found by integrating out the hyperparameters, is up to an additive constant

$$\log p(\{\beta_{gi}, \beta_f\}, \phi) = -\tfrac{1}{2}(n+2)\log\phi + \\ -\frac{1}{2\phi}\sum_i\sum_j\Big(y_{ij} - \sum_f \beta_f x_{fij} - \sum_g \beta_{gi} x_{gij}\Big)^2 + \\ -\tfrac{1}{2}(m+\nu'-1)\sum_g \log\Big\{\nu'\tau_g + \sum_i (\beta_{gi} - \beta_{g\cdot})^2\Big\}. \qquad (2)$$

Here $n = \sum_{i=1}^{m} n_i$ denotes the total sample size.

It is instructive to compare (2) to (1). The first term is simplified because of $\phi_i = \phi$; moreover there is no final term involving geometric and harmonic means of ϕ_i. Denoting the middle term as $-\frac{1}{2\phi}Q(\beta)$, it is clear that the modal estimate for ϕ is $\dot{\phi} = Q(\beta)/(n+2)$, and

$$\log p(\{\beta_{gi}, \beta_f\}, \dot{\phi}) = -\tfrac{1}{2}(n+2)\log Q(\beta) \qquad (3)$$

$$+\tfrac{1}{2}(n+2)\log(n+2) - \tfrac{1}{2}(n+2) +$$

$$-\tfrac{1}{2}(m+\nu'-1)\sum_g \log\Big\{\nu'\tau_g + \sum_i (\beta_{gi} - \beta_{g\cdot})^2\Big\}.$$

This makes clear the compromise character of the modal estimates for β. The first term above would be maximized by minimizing $Q(\beta)$, that is by using the least squares estimates. The last term is maximized when $\beta_{gi} = \beta_{g\cdot}$ for each i, but when the variance is less than the bound TAUMIN, the index passes into the set F, and we would end using the pooled estimate. The point is further elaborated in section 5.

The revised computer program maximizes (2) by iteration, amalogous to section 3, but each cycle now consists of an updating of ϕ, an updating of $\{\beta_f | f\in F\}$, an updating of $\{\beta_{gi} | g\in G\}$ and a check whether any index from G should pass into F. It leads to faster convergence , less core requirements, robustness to initial estimates specification and improved determination of (almost) totally regressed parameter values.

A preliminary FORTRAN version is available, but it can only be used when estimates for ν' and τ_g are given. A complete and interactive program in BASIC is being made, for inclusion in the next version of the CADA package by Novick et al.

5. Choice of the prior specification

The Bayesian estimates can always be viewed as a compromise between least squares values and pooled values. Unless one of these extremes is compatible with both the data and the prior information, however, the simultaneous presence of an intercept and k predictors poses an extra problem. Kelley could write $\hat{\tau}_i = \rho X_i + (1 - \rho)\dot{X}.$, and the reliability ρ determines the extent to which regression to the mean occurs. In our regression model, however, this extent will typically differ from parameter to parameter. Not only do we have $k + 1$ different extents of regression, but also each extent, and the best value to regress to, are influenced by the decisions on the other extents. And finally, when the extent was a reliability it could be estimated by one of the standard psychometric methods, but slopes and intercepts are not observable quantities, and this is an extra obstacle in trying to split their variance into true variance and error variance. This has been tried, more or less, by Jackson, Novick and Thayer (1971): they identify the least squares estimates with observed scores and their sampling error with measurement error, and obtain "model II estimates" for the slopes, and an unbiased estimate of the variance across groups of the true slopes. The means of predictor and criterion are then used to get model II estimates of the intercepts, and the sum of squares of these estimates divided by $m - 1$ is used when a prior estimate of the variance across groups of the true intercepts is needed.

In the new model too, the user has to provide prior estimates of the parameter variances across groups (now called τ_h) and of the degrees of freedom determining how much devi-

ation of the actual variance Ψ_h from the estimate τ_h is supposed to be compatible with the model. Earlier publications on one hand suggest that this should be genuine prior information, independent of the data, and that it can be chosen in such a way (e.g. taking $\nu' = 1$) that it has very little influence on the final Bayesian estimates. On the other hand model II estimates, obviously data-dependent, are nearly always used, and it is suggested that the data themselves will determine, almost independently of the prior knowledge, the amount of regression to the mean that is optimal. It is clear from (2), however, that the data only enter in the form of the sum of squared residuals, and that the least squares solution will inevitably come out when we let the data speak alone. In order to have any regression to the mean we must consider the second line.

Let us start in (3) with the least squares values. What happens when we pull all β_{hi} inward a little, either to their own mean or to the pooled estimates? The sum of squared residuals $Q(\beta)$ will certainly increase. The question is to what extent this is compensated for by the decrease of $\Sigma(\beta_{gi}-\beta_{g.})^2$. The mediating role of $\nu'\tau_g$ now becomes clear: if it is small compared to the value of $\Sigma(\beta_{gi}-\beta_{g.})^2$ being considered, this decrease will be influential, and if it is large the regression of the β_{gi} has almost no effect on the log posterior density. As the log of $Q(\beta)$ is multiplied by $\frac{1}{2}(n + 2)$, where n denotes total sample size, and the log in the last line of (3) only by $\frac{1}{2}(m + \nu' - 1)$, the other determining factor is how much larger the total sample size is than the number of groups (note that $m + \nu' - 1$ will not differ much from m).

The present authors doubt whether a simple and satisfactory procedure for specification of ν' and τ_g exists for cases where the product for $\nu'\tau_g$ is small; this difficulty also applies to m-group proportions and model II analysis of variance and was anticipated in earlier writings by Lindley c.s. Molenaar and Lewis (1979) will contain a description of our partial solution adopted in the BASIC program.

References.

Jackson, P.H., Novick, M.R.and Thayer, D.T. Estimating regressions in m-groups. Br. J. Math. Statist. Psychol., 1971, 24 129-153.

Jansen, G.G.H. An application of Bayesian Statistical Methods to a Problem in Educational Measurement. Thesis Groningen University, 1977.

Jones, P.K. and Novick, M.R. Implementation of a Bayesian system for prediction in m-groups. ACT Technical Report No. 6. Iowa City, Iowa: The American College Testing Program, 1972.

Lindley, D.V. A Bayesian solution for some educational prediction problems, II. Research Bulletin 69-91, Princeton, New Jersey: Educational Testing Service, 1969.

Lindley, D.V. and Smith, A.F.M. Bayesian estimates for the linear model. Journal of the Royal Statistical Society (Series B), 1972, 34, 1-41.

Lindley, D.V. The estimation of many parameters. In V.P. Godambe and D.A. Sprott (Eds.) Foundations of statistical inference. Toronto: Holt, Rinehart and Winston, 1971, 435-455.

Lissitz, R.W. and Schoenfeldt, L.F. Moderator subgroups for the estimation of educational performance: a comparison of prediction models. American Educational Research Journal, 1974, 11, 63-75, followed by debate, 76-92.

Molenaar, I.W. A fast solution of the Lindley equations for the m-group regression problem. Technical Report 78-3, Iowa Testing Programs, University of Iowa.

Molenaar, I.W. and Lewis, Charles, Bayesian m-group regression: a survey and an improved model, to appear in: MDN, SWS/Ver. voor Statistiek, Rotterdam (Neth.), Winter 1978/1979 issue.

Molenaar, I.W. and Lewis C. m-group regression revisited. To appear in 1979.

Novick, M.R., Jackson, P.H., Thayer, D.T. and Cole, N.S. Applications of Bayesian methods to the prediction of educational performance. ACT Research Report No. 42, Iowa City, Iowa: The American College Testing Programs, 1971.

Novick, M.R., Jackson, P.H., Thayer, D.T. and Cole, N.S. Estimating multiple regressions in m-groups: a cross-validation study. Br. J. Math. Statist. Psychol., 1972, 25, 33-50.

Novick, M.R., Lewis, C. and Jackson, P.H. The estimation of proportions in m-groups. Psychometrika, 1973, 38, 19-46.

Novick, M.R., Isaacs, G.L. and DeKeyrel, D.F. Computer Assisted Data Analysis - 1977. University of Iowa.

Roelofs, J.W. and Koppelaar, H. Case-study Biofeedback (in Dutch), Bulletin Ver. voor Statistiek 9 ,1978, 10-11.

Shigemasu, K. Development and validation of a simplified m-group regression model. Journal of Educational Statistics, 1976, 1, 157-180.

Acknowledgement: The authors are grateful to Melvin Novick for encouragement and support, and to Max Booleman for some computational assistance.