MODEL ASSESSMENT PLOTS FOR LOGISTIC REGRESSION WITH MULTILEVEL COVARIATES

Iain Pardoe, Department of Decision Sciences, Charles H. Lundquist College of Business, 1208 University of Oregon, Eugene, OR 97403–1208

Key Words: Bayesian methodology; Diagnostic; Graphical method; Hierarchical model; Model criticism; Random effect

Abstract

Residual plots are traditionally used to assess the fit of a regression model, yet can be difficult to interpret when the response variable is binary. This difficulty becomes compounded when covariates have a hierarchical or multilevel structure. An alternative graphical procedure is proposed for visualizing goodness of fit in such settings. The methodology is illustrated with an analysis of individual-level and county-level effects on sentencing practices across the U.S.

1 Introduction

Pardoe and Cook (2002) described a graphical technique for assessing the fit of a logistic regression model, called a "Bayes marginal model plot" (BMMP). This article describes an extension of the BMMP methodology to hierarchical logistic regression. Section 2 describes the dataset used to illustrate the methodology, while the hierarchical model used is outlined in Section 3. Section 4 concerns assessment of the model using BMMPs, while Section 5 contains a discussion.

2 Application: U.S. Imprisonment

In 2001, the U.S. imprisoned its citizens at a rate of 472 per 100,000 (Beck, Karberg, and Harrison, 2002), six to twelve times higher than in other western countries. Furthermore, there is large variation in imprisonment levels within the U.S.; e.g., in 2001, Louisiana's rate per 100,000 residents was 795, while Maine's was 126. Studies of differences in prison use among the states have found various factors to play a key role, including level of violent crime (Greenberg and West, 2001), percent of the population that is African American (McGarrell, 1993), and geographic region—Southern states appear to punish more severely (Michalowski and Pearson, 1990). Other studies examining aggregate punishment variation using a county as the unit of analysis have found unemployment in urban counties and violent crime (McCarthy, 1990), and percent of the population that is African American and Southern region (Weidner and Frase, 2001) to be significantly related to prison use.

By contrast, most sentencing studies focus on individuals, whereby effects of case characteristics, criminal history, and demographics are determined. However, effects of individual-level variables may vary according to the cultural, political, economic, and social contexts in which courts operate (Dixon, 1995). Studies of pooled statewide data have found several contextual variables to have an effect on sentencing, e.g., level of unemployment and crime rate (Myers and Talarico, 1987) and racial composition (Steffensmeier, Kramer, and Streifel, 1993). However, these studies use conventional logistic regression which does not correctly account for individual-level effects that vary according to a jurisdiction's cultural context and organizational constraints (Mears, 1998; Britt, 2000). To properly account for the multilevel nature of individual-level covariates and county-level contextual covariates, hierarchical modeling is more appropriate.

There has been only limited use of hierarchical modeling in criminal justice research. Rountree, Land, and Miethe (1994) used a hierarchical model for intra-city neighborhood differences in victimization risk, while Wooldredge, Griffin, and Pratt (2001) compared hierarchical and conventional models for the impact of prison and inmate characteristics on misconduct. Britt (2000) investigated whether social context and racial disparities affected punishment decisions in Pennsylvania counties for 1991-1994. Controlling for urbanization, racial threat, economic threat, and crime control, punishment severity varied by race across jurisdictions, but measures of social context explained little of this variation.

Pardoe, Weidner, and Frase (2002) analyzed data from the Bureau of Justice Statistics' State Court Processing Statistics program, a biennial collection of data on felony defendants in state courts in 39 of the 75 most populous U.S. counties. Information collected includes demographic characteristics, criminal history, and details of pretrial processing, disposition, and sentencing of felony defendants. Pardoe et al. linked individual-level data for 8,446 felony convictions in 17 states during May 1998 to county-level variables using the Federal Information Processing Standards code, and fit the hierarchical logistic regression model described in Section 3. Given the

lack of consensus regarding determinants of variation in prison use, and the limited use of hierarchical models in this area, it is vitally important to assess the fit of this model before it is to be used to inform policy.

Individual-level variables: Y = 1 if offender received a prison sentence, 0 for a jail or non-custodial sentence; IAGE = offender's age in years; IMAL = 1 for men, 0 for women; IBLK = 1 for African American, 0 otherwise; type of offense is measured with five dummy variables based on the most serious conviction charge: ICVS = murder, rape or robbery, i.e. a "more severe" violent offense, ICVM = assault or other violent crime, i.e. a "less severe" violent offense, ICTR = drug trafficking, ICDR = drug possession, ICPR = burglary or theft, i.e. a property offense (the reference charge category includes weapons, driving-related, and other public order offenses); ICJS = 1 if offender's criminal justice status was active at the time of the offense, 0 otherwise; IPFE = 1 if offender had one or more prior felony convictions, 0 otherwise; IPMI indicates prior misdemeanors similarly; IDET = 1 if offender was detained after being charged, 0 if released; IREV = 1 if offender's pretrial release was revoked, 0 otherwise; IBAD = 1 if offender was arrested while on pretrial release but release was not revoked; ITRI = 1 if offender was convicted by trial, 0 if convicted by plea.

County-level variables: CARR = county's arrest rate per 10,000 residents in 1998, a proxy measure for a county's level of crime; CUNR = county's unemployment rate for 1998. CBLP = a census estimate of the percentage of the county's population that was African American in 1998. CSTH = 1 if the county is located in a Southern state, 0 otherwise.

After excluding data with missing information, 3,672 individuals from 32 counties in 15 states remained. Pardoe et al. (2002) conducted a full analysis incorporating imputation of missing data; only complete data are used for the analyses considered in this article.

3 Hierarchical Logistic Regression

A hierarchical logistic regression model, also referred to in the literature as a multilevel model, can account for lack of independence across levels of nested data (i.e., individuals nested within counties). Conventional regression assumes that all experimental units (in this case, individuals) are independent in the sense that any variables affecting prison sentencing prevalence have the same effect in all counties. Hierarchical modeling relaxes this assumption and allows these variables' effects to vary across counties. One way to do this uses a generalization of the model of Wong and Mason (1985). First, the usual logistic regression model is fit to n_j individuals within each of $J = 32$ counties. The number of individuals in each county ranged from 14 to 456, with total number of individuals $I = \sum_{j=1}^{32} n_j = 3{,}672$. For the ith individual in the jth county, observe a dichotomous response,

$$\mathrm{Y}_{ij} = \begin{cases} 1 & \text{for a prison sentence} \\ 0 & \text{for a jail or non-custodial sentence} \end{cases}$$

$\mathrm{Y}_{ij}|p_{ij} \sim$ Bernouilli(p_{ij}), where $p_{ij} = \Pr(\mathrm{Y}_{ij} = 1)$, and

$$\mathrm{logit}(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \boldsymbol{X}_i^T \boldsymbol{\beta}_j \qquad (1)$$

where $\boldsymbol{X}_i$ represents measurements on K individual-level variables and $\boldsymbol{\beta}_j$ consists of K regression coefficients (specific to the jth county). Next, since each β-coefficient is likely to be related across counties, assume that each one can be explained by up to L county-level variables,

$$\boldsymbol{\beta}_j = \boldsymbol{G}_j \boldsymbol{\eta} + \boldsymbol{\alpha}_j \qquad (2)$$

where $\boldsymbol{G}_j$ is a $K \times M$ block-diagonal matrix of measurements on L county-level variables, $\boldsymbol{\eta}$ consists of M regression coefficients, and $\boldsymbol{\alpha}_j$ is a $K \times 1$ vector of county-level errors. In particular, the kth row of $\boldsymbol{G}_j$ contains a non-zero block with a one for an intercept together with the county-level variables used to explain the kth β-coefficient. Thus, M is $K \times L$ if all county-level variables are used to explain each β-coefficient, or less than this otherwise. Combining (1) and (2) leads to

$$\mathrm{logit}(p_{ij}) = \boldsymbol{X}_i^T \boldsymbol{G}_j \boldsymbol{\eta} + \boldsymbol{X}_i^T \boldsymbol{\alpha}_j \qquad (3)$$

Conventionally, the η-parameters in (3) are fixed effects (they have no j-subscript and represent the same effect over all counties) while the α-parameters are random effects (they have a j-subscript and represent different effects across counties). The presence of both types of effects makes (3) a mixed model. Suppressing the county-level errors so that (3) becomes a fixed effects model and amenable to standard regression requires assuming that individual-level effects are the same across counties, an assumption unlikely to be satisfied in practice.

Mixed models can be fit using specialized software such as "MLwiN" (Rasbash et al., 2000) and "HLM" (Raudenbush, Bryk, Cheong, and Congdon, 2001). Alternatively, by putting the model into a Bayesian framework, the distinction between fixed and random effects disappears (since all effects are now considered random), and the hierarchical structure is explicitly accounted for in the analysis. Pardoe et al. (2002) followed this Bayesian route, giving $\boldsymbol{\eta}$ a flat (uninformative) prior while specifying an exchangeable prior for the county-level errors, $\boldsymbol{\alpha}_j \sim N(\mathbf{0}, \boldsymbol{\Gamma}^{-1})$, where $\mathbf{0}$ is a K-vector of zeros and $\boldsymbol{\Gamma}^{-1}$ is a $K \times K$ covariance matrix. A hyper-prior distribution was specified for the inverse covariance matrix, $\boldsymbol{\Gamma} \sim$ Wishart$(\boldsymbol{R}, K)$, where $\boldsymbol{R}$ can be considered

a prior estimate of $\mathbf{\Gamma}^{-1}$ based on K observations, and, to represent vague prior knowledge, degrees of freedom for the Wishart distribution was set as small as possible to be K (the rank of $\mathbf{\Gamma}$). $\boldsymbol{R}$ was set to have values ten along the diagonal and zero elsewhere (sensitivity analysis, discussed by Pardoe et al., confirmed that the choice of $\boldsymbol{R}$ has little effect on the results).

The software package WinBUGS (Spiegelhalter, Thomas, and Best, 1999) was used to generate posterior samples for $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}_j$; this free software enables Bayesian analysis of complex statistical models using Gibbs sampling, a Markov chain Monte Carlo (MCMC) technique. The first model considered included $K \times L$ = 16 $\times$ 5 = 80 η-coefficients. WinBUGS ran four chains for 5,000 iterations, discarding the first 2,000 samples from each to leave 12,000 posterior samples for η. Summary statistics for these samples indicated that many η-coefficients were estimated with considerable imprecision. In particular, 11 η-coefficients corresponding to interactions of individual-level and county-level variables had posterior standard deviations at least five times the absolute value of their posterior means; these interactions were excluded from subsequent models.

The number of model terms continued to be reduced in this way, with interactions that demonstrated little ability to explain individual-level coefficients within counties removed. To preserve hierarchy and aid interpretation, no main effects (individual-level and county-level variables by themselves) were removed. As the model was simplified, the number of posterior samples was increased to improve estimation of means and standard deviations. Nine iterations of this procedure (each taking about 24 hours of computing time) produced a final model of just 38 terms, with all interactions having posterior standard deviations no more than the absolute value of their posterior means. After running four chains for 14,500 iterations, trace plots showed a good degree of mixing and MCMC convergence diagnostics indicated convergence. The model appeared to provide a good compromise between, on the one hand, parsimoniously describing the dependence of sentence type on individual and county covariates and, on the other hand, inadvertently excluding potentially important terms.

Before interpreting and using posterior samples from this model, underlying assumptions need to be assessed. Posterior samples of county-level errors, $\boldsymbol{\alpha}_j$, are a form of residual, and so lend themselves to the usual kinds of model diagnostics. The fact that they averaged close to zero across counties is reassuring, but unsurprising. More open to doubt are the normality and exchangeability assumptions. However, normal probability plots revealed no strong abnormalities, and plotting posterior means of the $\boldsymbol{\alpha}_j$ against county-level covariates also revealed no worrisome patterns (plots not shown). Nevertheless, such diagnostics seem insufficient to assess the fit of a model of such complexity. Section 4 describes use of an alternative graphical diagnostic procedure.

4 Bayes Marginal Model Plots

Cook and Weisberg (1997) proposed the use of "marginal model plots" (MMPs) to assess the goodness of fit of a regression model. Extending their rationale to hierarchical regression with covariates $\boldsymbol{X}$ measured on units nested in clusters with covariates $\boldsymbol{G}$ leads to:

$$\mathrm{E_F}(Y|\boldsymbol{X},\boldsymbol{G}) = \mathrm{E}_{\widehat{\mathrm{M}}}(Y|\boldsymbol{X},\boldsymbol{G}),\ \begin{cases} \forall\, \boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^K \\ \forall\, \boldsymbol{G} \in \mathcal{G} \subset \mathbb{R}^L \end{cases} \tag{4}$$

$$\iff \mathrm{E_F}(Y|h) = \mathrm{E}_{\widehat{\mathrm{M}}}(Y|h), \quad \forall\, h = h(\boldsymbol{X},\boldsymbol{G}) :\ \mathbb{R}^{K+L} \to \mathbb{R}^1 \tag{5}$$

where $\mathrm{E_F}$ denotes *model-free* expectation, $\mathrm{E}_{\widehat{\mathrm{M}}}$ denotes *model-based* expectation, $\mathcal{X}$ and $\mathcal{G}$ are the sample spaces of $\boldsymbol{X}$ and $\boldsymbol{G}$ respectively, and h is any measurable function of $\boldsymbol{X}$ and $\boldsymbol{G}$. In practice, useful h-functions include fitted values, individual covariates in the model, potential covariates not in the model, linear combinations of covariates, and random linear projections of covariates. Conditional expectations of Y in the logistic regression context correspond to the probabilities p_{ij} in (1).

Ideally, model assessment requires equality (4) to be checked, but when $K + L > 2$ then $\mathrm{E}(Y|\boldsymbol{X},\boldsymbol{G})$ is difficult to visualize. However, if h is univariate, $\mathrm{E}(Y|h)$ can be visualized in a two-dimensional scatterplot, and equality (5) can be checked. So, to assess the relationship between $\mathrm{E_F}(Y|\boldsymbol{X},\boldsymbol{G})$ and $\mathrm{E}_{\widehat{\mathrm{M}}}(Y|\boldsymbol{X},\boldsymbol{G})$, instead compare $\mathrm{E_F}(Y|h)$ and $\mathrm{E}_{\widehat{\mathrm{M}}}(Y|h)$ for various h. $\mathrm{E_F}(Y|h)$ and $\mathrm{E}_{\widehat{\mathrm{M}}}(Y|h)$ can be estimated with non-parametric smooths, e.g. cubic smoothing splines, the former by smoothing Y versus h, the latter by smoothing fitted values (probabilities), $\mathrm{E}_{\widehat{\mathrm{M}}}(Y|\boldsymbol{X},\boldsymbol{G})$, versus h. Superimpose $\widehat{\mathrm{E}}_{\mathrm{F}}(Y|h)$ and $\widehat{\mathrm{E}}_{\widehat{\mathrm{M}}}(Y|h)$ on a plot of Y versus h to obtain a MMP for the mean in the (marginal) direction h. Using the same method and smoothing parameter for both smooths allows their point-wise comparison, since any estimation bias approximately cancels. Smooths that match closely for any function h provide support for the model; otherwise model inadequacy is indicated.

However, it can be difficult to judge whether smooths match closely without guidance on model uncertainty. Bayesian model assessment ideas from Gelman, Meng, and Stern (1996) provide one way to visualize this uncertainty. Consider drawing values of $\boldsymbol{\beta}_j$ in (1) from their posterior distributions, and generating a sample of I realizations of Y from the model indexed by these $\boldsymbol{\beta}_j$. Repeat this process a large number m of times and compare the

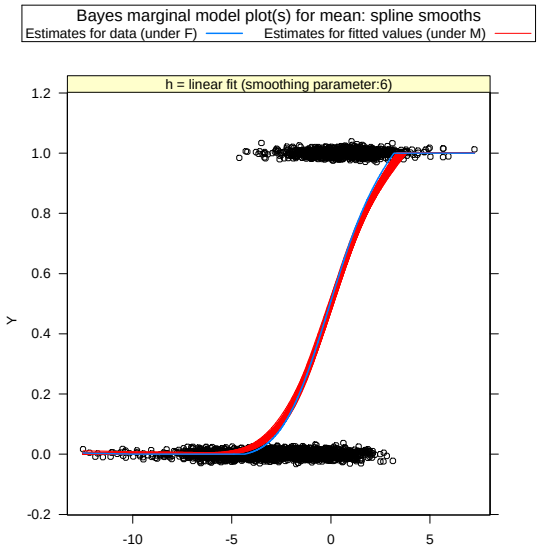

Figure 1: Bayes marginal model plot (BMMP) for the final hierarchical model with $h = \boldsymbol{X}_i^T \boldsymbol{G}_j \hat{\boldsymbol{\eta}} + \boldsymbol{X}_i^T \hat{\boldsymbol{\alpha}}_j$. The data have been jittered to aid visualization of relative density and the smooths are smoothing splines with six effective degrees of freedom.

data Y-values to the m (posterior predictive) realizations from the model. Then, if the data "look like" a typical realization from the model there is no reason to doubt its fit. On the other hand, if the data appear to be very "unusual" with respect to the m model realizations, then the model is called into question. A graphical way to do this is to compare model-free smooths of data Y-values with model-based smooths of predicted probabilities (calculated using sampled $\boldsymbol{\beta}_j$ values). So, in a Bayes marginal model plot (BMMP), instead of superimposing just one model-based smooth, smooths for m model samples are superimposed; $m = 100$ provides good resolution in the plot without excessive computing overhead.

Figure 1 is a BMMP for the final hierarchical model for the imprisonment data with $h = \boldsymbol{X}_i^T \boldsymbol{G}_j \hat{\boldsymbol{\eta}} + \boldsymbol{X}_i^T \hat{\boldsymbol{\alpha}}_j$, where $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\alpha}}_j$ are posterior means. If, for a particular h, the blue model-free smooth lies *substantially outside* the band of red model-based smooths *or* it does not follow the general pattern of the red model-based smooths, then the model is called into question. If, no matter what the function h is, the blue model-free smooth lies *broadly inside* the red model-based band *and* it follows the general pattern of the red model-based smooths, then perhaps the model is a useful one. In Figure 1, the blue smooth of the data passes close to the center of the red band of model-based smooths of $1/(1 + \exp(-\boldsymbol{X}_i^T \boldsymbol{G}_j \boldsymbol{\eta}^* - \boldsymbol{X}_i^T \boldsymbol{\alpha}_j^*))$, where $\boldsymbol{\eta}^*$ and $\boldsymbol{\alpha}_j^*$ are 100 posterior samples. So, there is little indication of lack-of-fit from this plot.

Since fitting a hierarchical logistic model requires substantially more computing time than a conventional (non-hierarchical) model, it is instructive to compare BMMPs

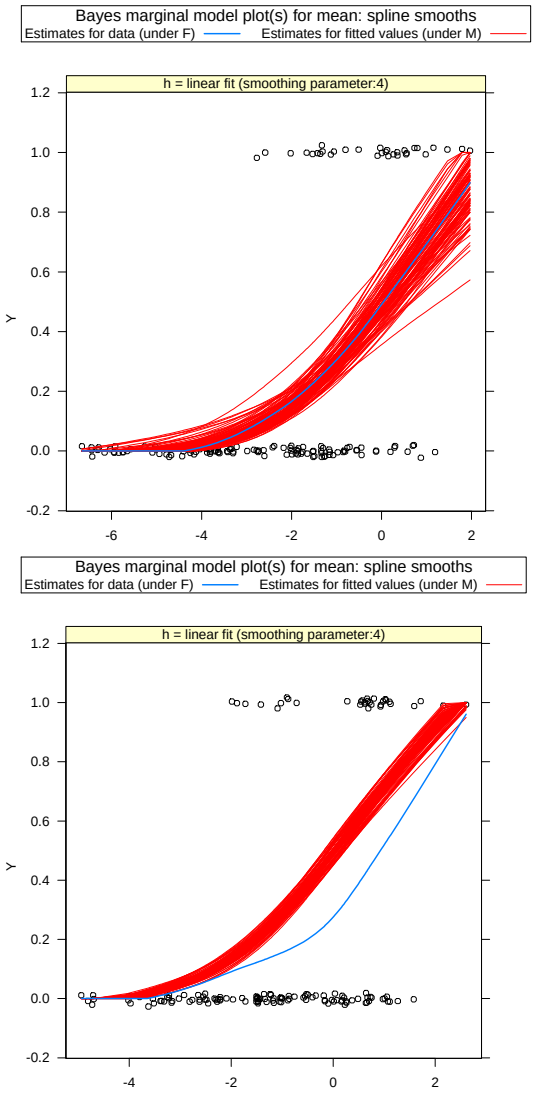

Figure 2: BMMPs for county 10; $h = \boldsymbol{X}_i^T \boldsymbol{G}_j \hat{\boldsymbol{\eta}} + \boldsymbol{X}_i^T \hat{\boldsymbol{\alpha}}_j$ for hierarchical model (upper) and $h = \boldsymbol{X}_i^T \boldsymbol{G}_j \hat{\boldsymbol{\eta}}$ for non-hierarchical model (lower). The smoothing splines have four effective degrees of freedom.

for a non-hierarchical model containing the same terms (main effects and interactions) as the final hierarchical model. A BMMP for such a non-hierarchical model with $h = \boldsymbol{X}_i^T \boldsymbol{G}_j \hat{\boldsymbol{\eta}}$ (plot not shown), is qualitatively very similar to Figure 1. This is unsurprising since both models give similar predictions when averaging across counties.

However, equality (5) should also match for subsets of the data, in particular within counties. Figure 2 contains BMMPs for one of the counties (number 10); the upper plot with $h = \boldsymbol{X}_i^T \boldsymbol{G}_j \hat{\boldsymbol{\eta}} + \boldsymbol{X}_i^T \hat{\boldsymbol{\alpha}}_j$ is for the hierarchical model, the lower plot with $h = \boldsymbol{X}_i^T \boldsymbol{G}_j \hat{\boldsymbol{\eta}}$ is for the non-hierarchical model. Now the non-hierarchical model clearly appears to be inadequate, while the hierarchical model continues to display no lack-of-fit. A similar as-

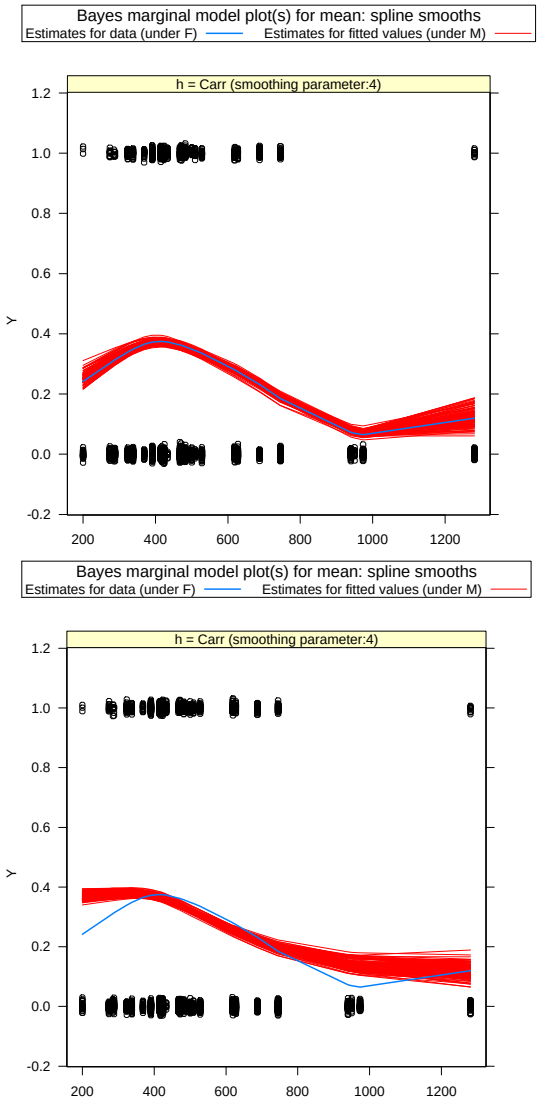

Figure 3: BMMPs with h = CARR for hierarchical model (upper) and non-hierarchical model (lower). The smoothing splines have four effective degrees of freedom.

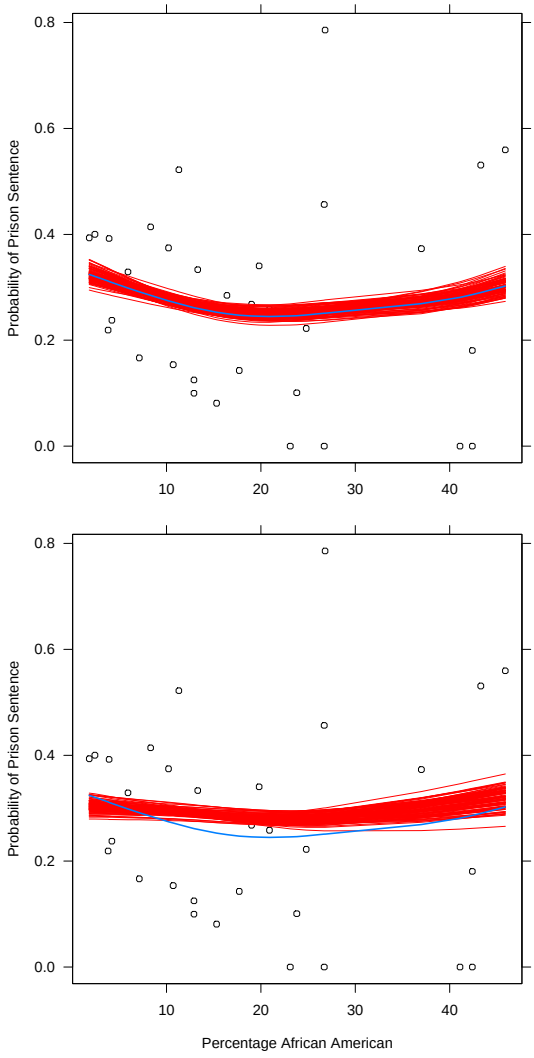

Figure 4: County-level BMMPs with h = CBLP for hierarchical model (upper) and non-hierarchical (lower). The smoothing splines have four effective degrees of freedom.

sessment can be made for comparable BMMPs for the other counties (plots not shown).

Nevertheless, a series of BMMPs for various h-functions should be constructed to gain confidence in any particular model. Since the models differ greatly on how county-level covariates are treated, consider the BMMPs with h = CARR in Figure 3. Again the non-hierarchical model appears inadequate, while the hierarchical model shows promise. Differences are also apparent for the other county-level covariates (plots not shown).

Finally, consider a BMMP from the perspective of a county as the unit of analysis. The data Y-values now become proportions of individuals in the counties sentenced to prison. The model-based probabilities of receiving a prison sentence in each county can be obtained by averaging individual probabilities. County-level BMMPs can then be constructed with h now a function of county-level covariates, $\boldsymbol{G}$, only. BMMPs based on this premise with h = CBLP are shown in Figure 4. Again the hierarchical model seems better than the non-hierarchical one. Differences are also apparent for the other county-level covariates (plots not shown).

In conclusion, the hierarchical model appears to fit well, and certainly improves on the conventional model. Pardoe et al. (2002) discuss results from the hierarchical model fit to the full imprisonment dataset.

5 Discussion

This article has demonstrated how Bayes marginal model plots (BMMPs) can be extended to assessment of hierarchical models containing random effects. Plots can be constructed at different levels of the hierarchy, e.g. with two levels: at the individual level and the cluster level. The example on U.S. imprisonment illustrates the need to use hierarchical models with multilevel covariates.

The methodology is not limited to logistic regression, and is generally applicable to any regression model. References to normal linear and additive model applications can be found in Pardoe and Cook (2002), which also contains further discussion of technical aspects of BMMPs such as calibration and smoothing. S-PLUS and R functions that can be used in conjunction with BUGS and BOA to construct BMMPs are available at:

`http://lcb1.uoregon.edu/ipardoe/research/bmmpsoft.htm`

References

Beck, A. J., J. C. Karberg, and P. M. Harrison (2002). *Prison and Jail Inmates at Midyear 2001 (Bureau of Justice Statistics Bulletin)*. Washington, DC: Bureau of Justice Statistics.

Britt, C. L. (2000). Social context and racial disparities in punishment decisions. *Justice Quarterly 17*, 707–732.

Cook, R. D. and S. Weisberg (1997). Graphics for assessing the adequacy of regression models. *Journal of the American Statistical Association 92*, 490–499.

Dixon, J. (1995). The organizational context of criminal sentencing. *American Journal of Sociology 100*, 1157–1198.

Gelman, A., X.-L. Meng, and H. Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica 6*, 733–807.

Greenberg, D. and V. West (2001). State prison populations and their growth, 1971–1991. *Criminology 39*, 615–654.

McCarthy, S. R. (1990). A micro-level analysis of social structure and social control: Intrastate use of jail and prison confinement. *Justice Quarterly 7*, 326–340.

McGarrell, E. F. (1993). Institutional theory and the stability of a conflict model of the incarceration rate. *Justice Quarterly 10*, 7–28.

Mears, D. P. (1998). The sociology of sentencing: Reconceptualizing decision-making processes and outcomes. *Law and Society Review 32*, 667–724.

Michalowski, R. J. and M. A. Pearson (1990). Punishment and social structure at the state level: A cross-sectional comparison of 1970 and 1980. *Journal of Research in Crime and Delinquency 27*, 52–78.

Myers, M. A. and S. M. Talarico (1987). *The Social Contexts of Criminal Sentencing*. New York: Springer-Verlag.

Pardoe, I. and R. D. Cook (2002). A graphical method for assessing the fit of a logistic regression model. *The American Statistician*. In press.

Pardoe, I., R. R. Weidner, and R. Frase (2002). Sentencing convicted felons in the United States: A Bayesian analysis using multilevel covariates. Technical report, Charles H. Lundquist College of Business, University of Oregon.

Rasbash, J., W. Browne, H. Goldstein, M. Yang, I. Plewis, M. Healy, G. Woodhouse, D. Draper, I. Longford, and T. Lewis (2000). *A User's Guide to MLwiN* (2nd ed.). London: Institute of Education.

Raudenbush, S. W., A. S. Bryk, Y. F. Cheong, and R. Congdon (2001). *HLM 5: Hierarchical Linear and Nonlinear Modeling* (2nd ed.). Chicago: Scientific Software International.

Rountree, P. W., K. C. Land, and T. D. Miethe (1994). Macro-micro integration in the study of victimization: A hierarchical logistic model analysis across Seattle neighborhoods. *Criminology 32*, 387–414.

Spiegelhalter, D. J., A. Thomas, and N. G. Best (1999). *WinBUGS Version 1.2 User Manual*. Cambridge, UK: MRC Biostatistics Unit.

Steffensmeier, D., J. Kramer, and C. Streifel (1993). Gender and imprisonment decisions. *Criminology 31*, 411–446.

Weidner, R. R. and R. Frase (2001). A county-level comparison of the propensity to sentence felons to prison. *International Journal of Comparative Criminology 1*, 1–22.

Wong, G. Y. and W. M. Mason (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association 80*, 513–524.

Wooldredge, J., T. Griffin, and T. Pratt (2001). Considering hierarchical models for research on inmate behavior: Predicting misconduct with multilevel data. *Justice Quarterly 18*, 203–231.