Journal of Data Science **4**(2006), 413-424

Exploring the Use of Subpopulation Membership in Bayesian Hierarchical Model Assessment

Guofen Yan[1] and J. Sedransk[2]
[1] *University of Virginia and* [2] *Case Western Reserve University*

Abstract: We investigate whether the posterior predictive p-value can detect unknown hierarchical structure. We select several common discrepancy measures (i.e., mean, median, standard deviation, and χ^2 goodness-of-fit) whose choice is not motivated by knowledge of the hierarchical structure. We show that if we use the entire data set these discrepancy measures do not detect hierarchical structure. However, if we make use of the subpopulation structure many of these discrepancy measures are effective. The use of this technique is illustrated by studying the case where the data come from a two-stage hierarchical regression model while the fitted model does not include this feature.

Key words: Checking function, discrepancy, posterior predictive p-value, test statistic.

1. Introduction

In fitting Bayesian hierarchical models, an inevitable task is to decide how many hierarchical levels should be included in the model. Often a hierarchical model that is fit does not account for all of the stages actually present because these stages were not apparent, a priori. The objective of our research is to find diagnostic methods that are effective in detecting such missing structure.

This work was motivated by an increasing number of Bayesian hierarchical model applications, especially those involving spatial areal units. For example, in areal based sample surveys such as the National Health Interview Survey, there is uncertainty about the number and type of hierarchical levels that are appropriate (Malec, Sedransk, Moriarity and LeClere 1997). In this survey each primary sampling unit is a single county or group of counties. Within each primary sampling unit, groups of households are aggregated into areal units and sampled. A major question is whether it is sufficient to include in the model only the effects due to these primary sampling units, or to also include other, more homogeneous, geographical units such as census tracts. Should one use individual county effects rather than primary sampling unit effects in the model?

Other relevant geographical units may not be ones that are readily identified a priori.

In this paper we investigate the ability of the posterior predictive p-value to detect unknown hierarchical structure. Since there is very limited research in this area we study the simple, but important case where the data come from a two-stage hierarchical regression model while the fitted model does not include this feature. We examine a set of commonly used test statistics whose choice is not motivated by knowledge of the hierarchical structure; i.e., the mean, median, variance (standard deviation) and χ^2 goodness-of-fit measure. We show that these test statistics are ineffective when all of the data are used, but this changes when subpopulation membership is also considered.

We describe the data and models in Section 2, introduce our method in Section 3 and present the results in Section 4. There is a summary in Section 5.

2. Data and Models

The basic data that we used are a set of white female breast cancer incidence rates for U.S. counties for 1995-6. There are ten age classes (0-4, 5-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85 and up) and twelve regions which are the nine Census divisions with three of them split to achieve greater homogeneity of rates (Pickle, Mungiole, Jones and White 1996, p.6). Let d_{kij} and n_{kij} denote, respectively, the number of deaths and population at risk for age class j in health service area (HSA) i and region k. Define

$$\begin{aligned} r_{kij} &= d_{kij}/n_{kij}\ , \\ r^*_{kij} &= \begin{cases} r_{kij} & \text{if} \quad r_{kij} > 0 \\ 10^{-6} & \text{if} \quad r_{kij} = 0, \end{cases} \\ y_{kij} &= \ln r^*_{kij}\ . \end{aligned}$$

The data that we used for our analysis were observations generated from the hierarchical model defined in (2.1) which is a modified version of one fit by Dr. Linda Pickle of the Division of Cancer Control and Population Sciences, National Cancer Institute. There are 156 HSAs belonging to the twelve regions defined above, and seven age classes (25-34, 35-44, ..., 85 and up). The 1092 observations (156 HSAs and seven age classes per HSA) were generated from the model,

$$\begin{gathered} y_{kij} \mid \beta_{ki}, \phi_0 \overset{ind}{\sim} N(\ln \lambda_{kij}, \phi_0) \qquad (2.1) \\ \ln \lambda_{kij} = x'_j\, \beta_{ki}\ , \\ \beta_{ki} \mid \beta_0, D_0 \overset{iid}{\sim} N(\beta_0, D_0)\ , \\ k = 1, \ldots, 12,\ i = 1, \ldots, n_k,\ j = 1, \ldots, 7\ , \end{gathered}$$

where $x_j' = (1, a_j, a_j^2, a_j^3)$ with a_j coded as $3, 4, \ldots, 9$ for the seven age classes, (25-34, 35-44, ..., 85 and up), $\beta_{ki} = (\beta_{ki,0}, \beta_{ki,1}, \beta_{ki,2}, \beta_{ki,3})'$ the regression coefficients of age effect for HSA i and region k, and β_0=$(\beta_{00}, \beta_{01}, \beta_{02}, \beta_{03})'$ =$(-14.49494, 3.09401, -0.34612, 0.01288)^{'}$ which were the values from Dr. Pickle's fitted model. The variance matrix D_0 was taken to be diagonal with entries (0.25, 0.09, 0.0025, 0.0001). We examined three values of ϕ_0, i.e., 0.1, 1.1 and 5. Since the results for $\phi_0 = 0.1$, 1.1 and 5 were similar, we only present the results for $\phi_0 = 5$ in this paper.

We fit a non-hierarchical model to the data,

$$y_{kij} \mid \beta, \phi \overset{ind}{\sim} N(x_j'\beta, \phi) \; . \tag{2.2}$$

Both ϕ and β are assumed unknown with a locally uniform prior, $\pi(\beta, \phi) \propto$ constant.

Note that the fitted model, (2.2), combines the two stages in the actual hierarchical model, (2.1), but does not account for the correlation and heterogeneous variances in the observations that were generated from (2.1).

3. Methodology Using The Posterior Predictive p-Value

The notion of using posterior predictive assessment was first introduced by Guttman (1967), formal Bayesian definitions were given by Rubin (1984) and Geisser (1993), and the procedure was extended by Meng (1994). Denoting a test statistic or checking function by T, the posterior predictive p-value using T is (Rubin 1984; Gelman, Meng and Stern 1996)

$$p_{\text{post}} \quad = \quad P(T(\tilde{y}) \geq T(y^{obs}) \mid y^{obs}) \tag{3.1}$$

where the probability in (3.1) uses the density function

$$p(\tilde{y}|y^{obs}) = \int p_1(\tilde{y}|\beta, \phi)\, p_2(\beta|\phi, y^{obs})\, p_3(\phi|y^{obs})\, d\beta\, d\phi \tag{3.2}$$

where y^{obs} is a vector of observed data, $\tilde{y}$ a vector of predicted data, and given β and ϕ, y^{obs} and $\tilde{y}$ are assumed to be independent. The main idea is to compare a function of the vector of observed values, $T(y^{obs})$, with the distribution (under the assumed model) of the same function of the predicted values, $T(\tilde{y})$. If $T(y^{obs})$ falls in the tails of the distribution of $T(\tilde{y})$ the validity of the assumed model should be questioned. Extreme values of p_{post} (either very large or small) indicate a lack of fit of the model.

When the posterior predictive distribution of $T(\tilde{y})$ given the data is complicated, one can use a sampling-based method to obtain an estimate of (3.1). Under

the fitted non-hierarchical model, (2.2), $p_1(\tilde{y}|\beta,\phi)$, $p_2(\beta|\phi,y^{obs})$ and $p_3(\phi|y^{obs})$ in (3.2) are

$$\begin{aligned} \tilde{y}|\beta,\phi &\sim N(X\beta, I\phi) \\ \beta|\phi,y^{obs} &\sim N(\hat{\beta},(X'X)^{-1}\phi) \\ \phi|y^{obs} &\sim \text{scaled}-\text{Inv}\ \chi^2(\nu,s^2) \end{aligned} \tag{3.3}$$

where $\nu = N-q-2$, $s^2 = (y^{obs}-X\hat{\beta})'(y^{obs}-X\hat{\beta})/\nu$, $\hat{\beta} = (X'X)^{-1}X'y^{obs}$, N is the total sample size, X is the design matrix, and $q=4$, the number of regression parameters. The scaled-Inv $\chi^2(\nu,s^2)$ distribution in (3.3) has pdf, $p_3(\phi|y^{obs}) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} s^{\nu}\phi^{-(\nu/2+1)}e^{-\nu\, s^2/(2\phi)}$. Obtain $\tilde{y}$ by drawing ϕ from $p_3(\phi|y^{obs})$, β from $p_2(\beta|\phi,y^{obs})$ and $\tilde{y}$ from $p_1(\tilde{y}|\beta,\phi)$, then compute $T(\tilde{y})$ using $\tilde{y}$. For example, if T is the mean, $T(\tilde{y}) = \sum_{k,i,j}^{N}\tilde{y}_{kij}/N$. Repeating these steps independently M times yields $\{T(\tilde{y}^{(1)}),\ldots,T(\tilde{y}^{(M)})\}$, which estimates the distribution of $T(\tilde{y})$ given y^{obs}. Finally, (3.1) can be estimated by

$$\hat{p}_{\text{post}} = \frac{1}{M}\sum_{r=1}^{M} I[T(\tilde{y}^{(r)}) \geq T(y^{obs})]\,. \tag{3.4}$$

Meng (1994) proposed an extension of the posterior predictive p-value method that allows a discrepancy measure, $D(y;\theta)$, to depend on both the data and parameters θ. Then, the posterior predictive p-value is defined as

$$p_{\text{post}} = P\{D(\tilde{y},\theta) \geq D(y^{obs},\theta) \mid y^{obs}\}$$

where the probability is with respect to $p(\tilde{y},\theta|y^{obs})$.

Define the standardized residual,

$$e_{kij} = \frac{y_{kij} - E(y_{kij}|\theta)}{\{\text{var}(y_{kij}|\theta)\}^{1/2}}$$

where, in our example, $\theta = (\beta,\phi)$. Then choosing D as a function of the set of e_{kij} corresponding to the N observations

$$p_{\text{post}} = P(D(\tilde{e};\beta,\phi) \geq D(e^{obs};\beta,\phi) \mid y^{obs})\,. \tag{3.5}$$

Repeated sampling from p_3, p_2 and p_1 in (3.3) yields $\{\phi^{(r)},\beta^{(r)},\tilde{y}^{(r)} : r = 1,\ldots,M\}$. For each posterior draw, $(\phi^{(r)},\beta^{(r)},\tilde{y}^{(r)})$, we compute $\tilde{e}^{(r)}_{kij} = \{\tilde{y}^{(r)}_{kij} - E(y_{kij}|\beta^{(r)},\phi^{(r)})\}/\{\text{var}(y_{kij}|\beta^{(r)},\phi^{(r)})\}^{1/2}$ and

$$e^{obs(r)}_{kij} = \{y^{obs}_{kij} - E(y_{kij}|\beta^{(r)},\phi^{(r)})\}/\{\text{var}(y_{kij}|\beta^{(r)},\phi^{(r)})\}^{1/2}$$

for the N subjects where $E(y_{kij}|\beta^{(r)},\phi^{(r)}) = x'_j\,\beta^{(r)}$ and $\text{var}(y_{kij}|\beta^{(r)},\phi^{(r)}) = \phi^{(r)}$ from the fitted model (2.2). Then, calculate $D(\tilde{e}^{(r)};\beta^{(r)},\phi^{(r)})$ and $D(e^{obs(r)};\beta^{(r)},\phi^{(r)})$ using the N values of $\tilde{e}^{(r)}_{kij}$ and $e^{obs(r)}_{kij}$; for example, if D is the mean,

$$D(e^{obs(r)};\beta^{(r)},\phi^{(r)}) = \sum_{k,i,j} e^{obs(r)}_{kij}/N.$$

With M independent replications, there are $\{D(\tilde{e}^{(r)};\beta^{(r)},\phi^{(r)}),\ D(e^{obs(r)};\beta^{(r)},\phi^{(r)}) : r = 1,\ldots,M\}$. Finally, (3.5) can be estimated by

$$\hat{p}_{\text{post}} = \frac{1}{M}\sum_{r=1}^{M} I[D(\tilde{e}^{(r)};\beta^{(r)},\phi^{(r)}) \geq D(e^{obs(r)};\beta^{(r)},\phi^{(r)})]\,. \tag{3.6}$$

For the checking function, T, we examined the mean and standard deviation (SD). The discrepancy measures we examined are the mean, median, variance and χ^2 measure.

4. Results

4.1 Population based analysis

The histograms in Figure 1 are typical of the results for the estimated posterior predictive distributions of the checking functions. Each histogram is a plot of 1000 replications, $\{T(\tilde{y}^{(r)}) : r = 1,\ldots,1000\}$, with $\hat{p}_{\text{post}}$, (3.4), shown in the title and $T(y^{obs})$ indicated by the dashed line. When using the mean or SD as the checking function, the p-values are always near 0.5, indicating that these two are not useful in detecting the wrongly fitted model.

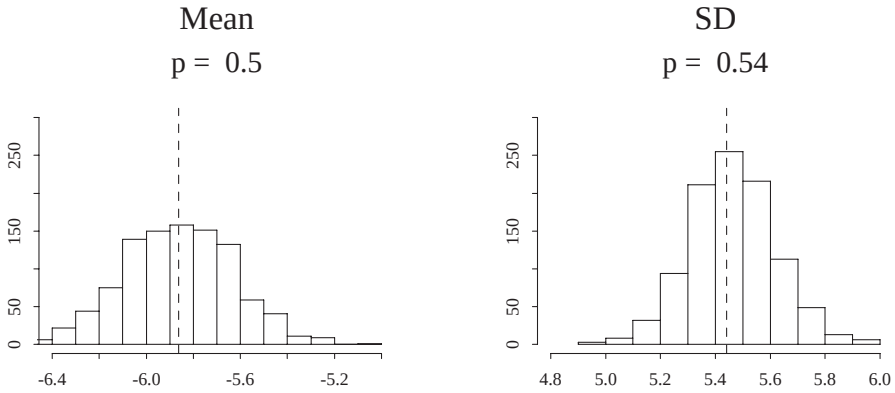

Figure 1: Estimated posterior predictive distributions of the two checking functions, mean and standard deviation (SD). Each histogram plots $\{T(\tilde{y}^{(r)}) : r = 1,\ldots,1000\}$ where a vector of $\tilde{y}^{(r)}$ is drawn independently from $p(\tilde{y}|y^{obs})$. $T(y^{obs})$ is shown by the dashed line, and p in the title is the estimated posterior predictive p-value.

In addition to calculating $\hat{p}_{\text{post}}$, (3.6), for the discrepancy measures, we present scatterplots of $\{D(\tilde{e}^{(r)};\beta^{(r)},\phi^{(r)}),\ D(e^{obs(r)};\beta^{(r)},\phi^{(r)}) : r = 1,\ldots,1000\}$. If the fitted model is consistent with the model used to generate the data we expect the points to be symmetric around the 45° line. Conversely, with an effective checking function and the opposite situation (i.e., the fitted model inconsistent with the model used to generate the data) we would expect significant deviations from symmetry.

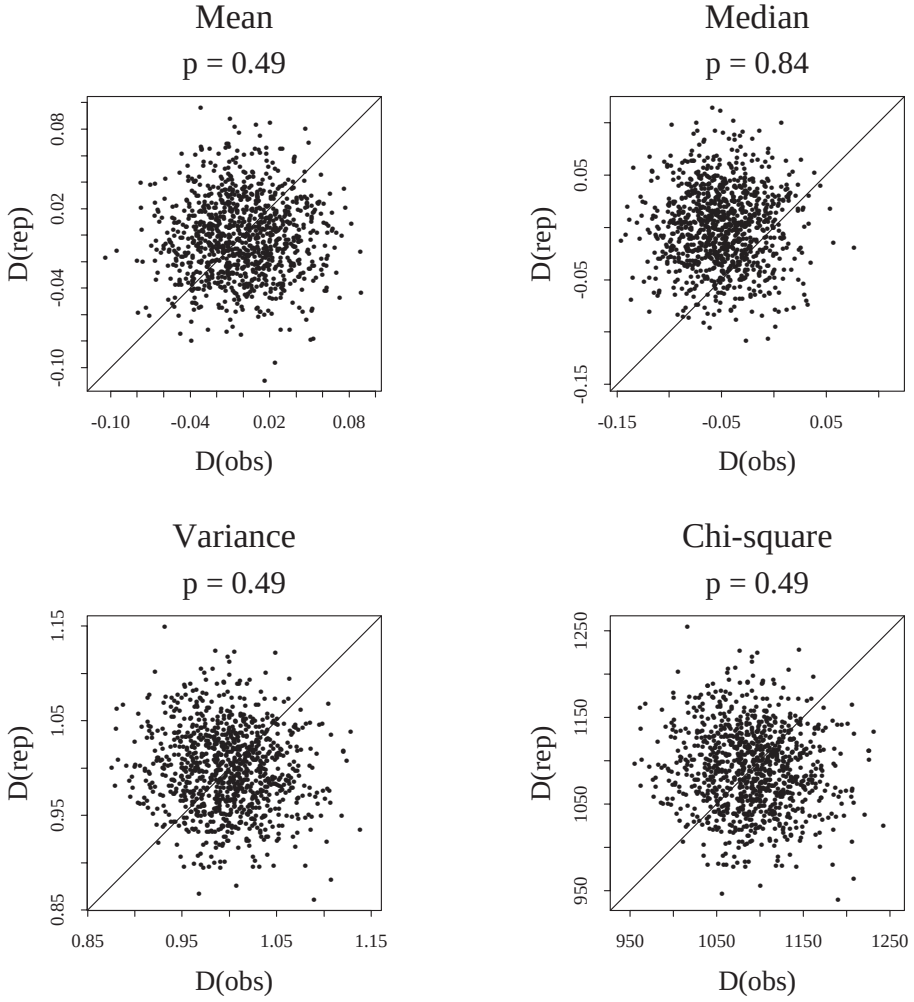

Figure 2: Scatterplots of 1000 pairs of $\{D(\tilde{e}^{(r)};\beta^{(r)},\phi^{(r)}), D(e^{obs(r)};\beta^{(r)},\phi^{(r)})\}$, $r = 1,\ldots,1000$. The D used is identified at the top of each plot and p in the title is the estimated posterior predictive p-value.

Each scatterplot in Figure 2 shows 1000 pairs of $\{D(\tilde{e}^{(r)};\beta^{(r)},\phi^{(r)}),\ D(e^{obs(r)};\beta^{(r)},\phi^{(r)})\}$ for the discrepancy measure indicated at the top of each plot. Using

the mean, median, variance and χ^2 measure as discrepancy measures the points are symmetric around the 45 degree line which is inappropriate because the data are from a model, (2.1), which is different from the fitted model, (2.2). Similarly, the p-values are not extreme, which is not appropriate for this case. These results indicate that these common choices, the mean, median, variance and χ^2 measure, are not useful. Since the mean is a sufficient statistic in the fitted model (2.2), the unsatisfactory results for the mean are not surprising.

4.2 Subpopulation based analysis

We still consider the usual (standardized) residuals having the form

$$e_{kij} = \frac{y_{kij} - E(y_{kij}|\beta, \phi)}{\{\text{var}(y_{kij}|\beta, \phi)\}^{1/2}},$$

but consider two discrepancy measures, $D(e; \beta, \phi)$ and $D(e_j; \beta, \phi)$, the former corresponding to using all of the data and the latter limited to the data from subpopulation j. Assuming that there are S mutually exclusive and exhaustive subpopulations, there are two types of posterior predictive p-value, i.e.,

$$p_{\text{post}} = P(D(\tilde{e}; \beta, \phi) \geq D(e^{obs}; \beta, \phi)|y^{obs}) \tag{4.1}$$

and

$$p_{\text{post}(j)} = P(D(\tilde{e}_j; \beta, \phi) \geq D(e_j^{obs}; \beta, \phi)|y^{obs}), \quad j = 1, \ldots, S\,. \tag{4.2}$$

The p-value in (4.2) is conditional on all of the data, but only part of the data is used for checking the agreement between the fitted model and actual data. However, since there are S values of $p_{\text{post}(j)}$ corresponding to S subpopulations that cover the entire population, using the set $\{p_{\text{post}(j)} : j = 1, \ldots, S\}$ provides an overall assessment of the model. If the fitted model is compatible with the data, we expect the set of values, $\{p_{\text{post}(j)} : j = 1, \ldots, S\}$, to be nearly equal, and similar patterns in the scatterplot of $D(\tilde{e}; \beta, \phi)$ and $D(e^{obs}; \beta, \phi)$ (using all of the data) as that in the scatterplots of $D(\tilde{e}_j; \beta, \phi)$ and $D(e_j^{obs}; \beta, \phi)$ for $j = 1, \ldots, S$.

We illustrate this idea in our example. In Figure 2 we have seen that the usual posterior predictive p-values using the mean, median, variance and χ^2 measure do not detect the wrongly fitted model. It is clear that the primary subpopulations in this example are the seven age classes. The overall variance discrepancy measure is

$$\text{var}(e; \beta, \phi) = \sum_k \sum_i \sum_j (e_{kij} - \bar{e})^2/(N-1) \tag{4.3}$$

where N is the total number of observations (over HSAs and age classes). The second measure is the variance based on the data in age class j; i.e.,

$$\text{var}(e_j;\beta,\phi) = \sum_k \sum_i (e_{kij} - \bar{e}_j)^2/(n^* - 1) \tag{4.4}$$

where $\bar{e}_j = \sum_k \sum_i e_{kij}/n^*$ and n^* denotes the total number of observations in age class j.

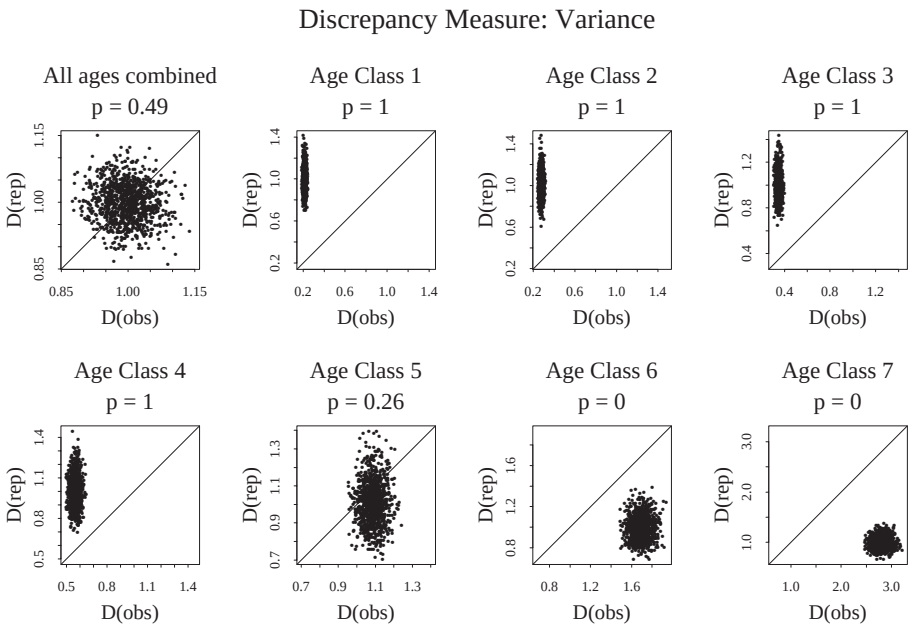

Figure 3: The upper left scatterplot is a scatterplot of 1000 pairs of $\{D(\tilde{e}^{(r)};\beta^{(r)},\phi^{(r)}), D(e^{obs(r)};\beta^{(r)},\phi^{(r)}) : r = 1,\ldots,1000\}$, each of the other plots is a scatterplot of $\{D(\tilde{e}_j^{(r)};\beta^{(r)},\phi^{(r)}), D(e_j^{obs(r)};\beta^{(r)},\phi^{(r)}) : r = 1,\ldots,1000\}$ limited to age class j, where D is the variance. The value of p is the estimated posterior predictive p-value.

The upper left plot in Figure 3 is a scatterplot of $\text{var}(\tilde{e};\beta,\phi)$ and $\text{var}(e^{obs};\beta,\phi)$ where $\text{var}(e;\beta,\phi)$ is defined in (4.3). (This is the same plot as the one for the variance in Figure 2.) As we have seen in Section 4.1 this plot does not show a lack of fit from fitting (2.2) to these data generated from (2.1) (the estimated p is 0.49). Each of the other scatterplots in Figure 3 is restricted to a single age class; i.e., the scatterplots use the variance defined in (4.4). These plots have different characteristics. For the younger ages the points are concentrated in the upper left hand corners while for the older ages the points are concentrated in the lower right hand corners. The values of $p_{\text{post}(j)}$ vary from 1 to 0 as one goes from the younger to the older ages. There are different patterns in the scatterplot

for all ages combined and the scatterplot for any specific age class, as well as among the seven age classes. This technique (i.e., using the variance based on the subpopulation) clearly shows that the fitted model is not concordant with the data while the usual one does not. Figure 4 shows similar patterns for the χ^2 discrepancy measure.

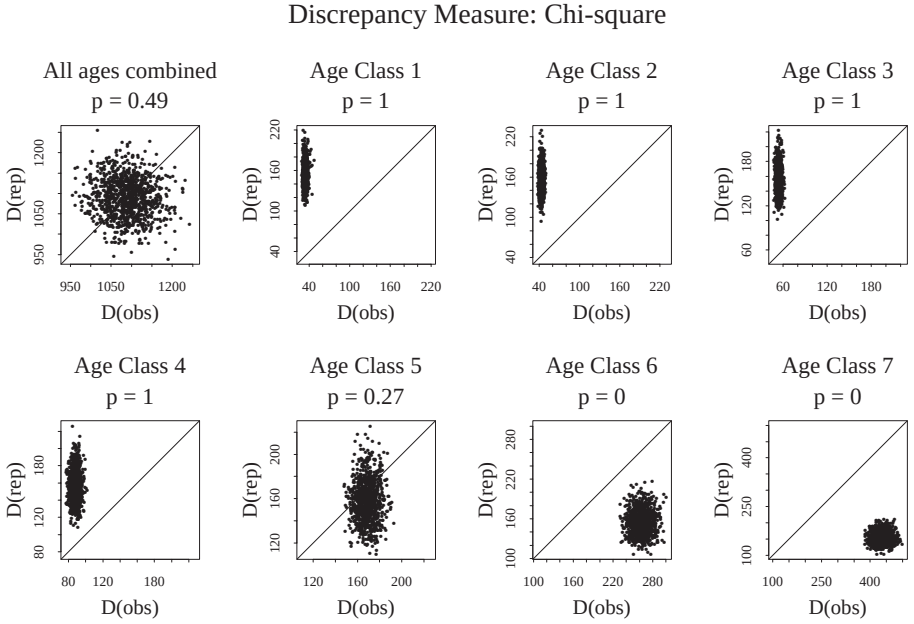

Figure 4: The upper left scatterplot is a scatterplot of 1000 pairs of $\{D(\tilde{e}^{(r)};\beta^{(r)},\phi^{(r)}), D(e^{obs(r)};\beta^{(r)},\phi^{(r)}) : r = 1,\ldots,1000\}$, each of the other plots is a scatterplot of $\{D(\tilde{e}_j^{(r)};\beta^{(r)},\phi^{(r)}), D(e_j^{obs(r)};\beta^{(r)},\phi^{(r)}) : r = 1,\ldots,1000\}$ limited to age class j, where D is the χ^2 discrepancy measure. The value of p is the estimated posterior predictive p-value.

To explain these results we compare (4.1) and (4.2) by evaluating the expected values, $E\{D(\tilde{e};\beta,\phi|y^{obs})\}$ and $E\{D(e^{obs};\beta,\phi|y^{obs})\}$, and comparing these to $E\{D(\tilde{e}_j;\beta,\phi|y^{obs})\}$ and $E\{D(e_j^{obs};\beta,\phi|y^{obs})\}$. Note that these expectations use (3.2) for moments of $\tilde{y}$ given y^{obs} and (2.1) for the actual (not fitted) distribution of y^{obs}. For illustration we use the variance discrepancy measure. Define $\mathrm{var}(e^{obs};\beta,\phi)$ and $\mathrm{var}(\tilde{e};\beta,\phi)$ as the variances corresponding to all of the observed and replicated data, with corresponding definitions for $\mathrm{var}(e_j^{obs};\beta,\phi)$ and $\mathrm{var}(\tilde{e_j};\beta,\phi)$ as limited to age class j.

For the variance based on all of the data it can be shown, after considerable algebraic manipulation, that

$$\begin{aligned} &E\{\mathrm{var}(\tilde{e};\beta,\phi)|\phi_0,\beta_0,D_0\} \\ &= E_{(y^{obs}|\phi_0,\beta_0,D_0)}\, E_{(\beta,\phi|y^{obs})}\, E_{(\tilde{y}|\beta,\phi)}\, \mathrm{var}(\tilde{e};\beta,\phi) = 1 \end{aligned} \tag{4.5}$$

and

$$\begin{aligned} E\{\text{var}(e^{obs};\beta,\phi)|\phi_0,\beta_0,D_0\} &= E_{(y^{obs}|\phi_0,\beta_0,D_0)}\, E_{(\beta,\phi|y^{obs})}\, \text{var}(e^{obs};\beta,\phi) \\ &= (N-3)/(N-1) \doteq 1 \end{aligned} \tag{4.6}$$

The results in (4.5) and (4.6) have shown that, on average, the observed and predicted variance discrepancy measures are equal when using all of the data. This suggests that the usual posterior predictive p-value will yield a moderate p-value even when the fitted model is not appropriate.

For the variance using age class j, it can be shown that

$$\begin{aligned} E\{\text{var}(\tilde{e}_j;\beta,\phi)|\phi_0,\beta_0,D_0\} &= E_{(y^{obs}|\phi_0,\beta_0,D_0)}\, E_{(\beta,\phi|y^{obs})}\, E_{(\tilde{y}|\beta,\phi)}\, \text{var}(\tilde{e}_j;\beta,\phi) \\ &= 1 \end{aligned} \tag{4.7}$$

and

$$\begin{aligned} E\{\text{var}(e_j^{obs};\beta,\phi)|\phi_0,\beta_0,D_0\} &= E_{(y^{obs}|\phi_0,\beta_0,D_0)}\, E_{(\beta,\phi|y^{obs})}\text{var}(e_j^{obs};\beta,\phi) \\ &\doteq \frac{\phi_0 + x_j' D_0 x_j}{\phi_0 + \frac{1}{7}\sum_j x_j' D_0 x_j}\,. \end{aligned} \tag{4.8}$$

Recall that $x_j' = (1, a_j, a_j^2, a_j^3)$ with a_j coded as $3, 4, \ldots, 9$ for the seven age classes 25-34, 35-44, 45-54, $\ldots$, 85 and up. In (4.8) the value of the numerator is the variance of the individual observations in age class j and this quantity increases with older age class j. The denominator is the average variance over the seven age classes. Hence, (4.8) is expected to be less than one, (4.7), for the younger age classes and larger than one for the older age classes. That is, approximately,

$$E(\text{var}(e_j^{obs};\beta,\phi)|y^{obs}) < E(\text{var}(\tilde{e}_j;\beta,\phi)|y^{obs}) \text{ for younger age classes } j$$

and

$$E(\text{var}(e_j^{obs};\beta,\phi)|y^{obs}) > E(\text{var}(\tilde{e}_j;\beta,\phi)|y^{obs}) \text{ for older age classes } j\,.$$

These theoretical results have shown that, on average, the observed variance discrepancy measure is smaller than the predicted discrepancy measure for younger age classes and vice versa for older age classes.

As we have seen in Figure 3, the 1000 replicate data points, $\{\text{var}(e_j^{obs(r)};\beta^{(r)}, \phi^{(r)}) : r = 1,\ldots,1000\}$, cluster around a value less than one for the younger age classes (e.g., around 0.2 for age class 1) and around a value larger than 1 for the older age classes (e.g., around 2.8 for age class 7). From (4.7) and Figure 3, the 1000 replications of $\text{var}(\tilde{e}_j;\beta,\phi)$ are always centered around 1 for any age class j. Thus, $p_{\text{post}(j)} = P(\text{var}(\tilde{e}_j;\beta,\phi) \geq \text{var}(e_j^{obs};\beta,\phi)|y^{obs})$ is large

for the smaller age classes and small for the older age classes, and it varies with age class. Hence, the $\{\text{var}(\tilde{e}_j^{(r)}; \beta^{(r)}, \phi^{(r)}), \text{var}(e_j^{obs(r)}; \beta^{(r)}, \phi^{(r)}) : r = 1, \ldots, 1000\}$ data points cluster asymmetrically around the 45° line, and the patterns are different for the different age classes. Note that, as expected, $\text{var}(\tilde{e}_j; \beta, \phi)$ has larger variability than $\text{var}(e_j^{obs}; \beta, \phi)$.

The preceding theoretical argument indicates that for this case (i.e., data generated from the hierarchical model while the fitted model does not have this feature), the "usual" posterior predictive p-value, i.e., using all of the data, is ineffective while the posterior predictive p-value using subpopulations is useful. Effective diagnostics should reveal data structures that are not taken into account by the fitted model, and in the hierarchical setting the missing structure is often heterogeneous variation for the subpopulations. Therefore, diagnostics conditioning on important subpopulations may help to detect missing structure in the fitted model. Clearly, in this case, inconsistency in the set of scatterplots using the variance (also χ^2) discrepancy measure reveals the heterogeneous variance in the data across the age classes which the postulated model does not have.

5. Summary

We have shown that the posterior predictive p-value using subpopulation structure can effectively identify hierarchical structure while the usual one (i.e., using all of the data) may not do so.

Sinharay and Stern (2003) study the performance of the mean and standard deviation in assessing the postulated distribution of the parameters at the second stage of their two-stage hierarchical fitted model. In their study, the posterior predictive p-value using the mean or the standard deviation based on all of the data was not effective, a finding similar to ours.

In conclusion, we have shown that the difficult task of identifying hierarchical structure (Yan and Sedransk 2004) may be helped by considering subpopulations separately. Further research is needed to establish the generality of our finding, demonstrated here for a common, important model.

Acknowledgement

The authors are grateful to Dr. Linda Pickle of the Division of Cancer Control and Population Sciences, National Cancer Institute, for providing the model and related data.

References

Geisser, S. (1993). *Predictive Inference: An Introduction*. Chapman and Hall.

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

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society B* **29**, 83-100.

Malec, D., Sedransk, J., Moriarity C., and LeClere, F. (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association* **92**, 815-826.

Meng, X.L. (1994). Posterior predictive p-values. *The Annals of Statistics* **22**, 1142-1160.

Pickle, L., Mungiole, M., Jones, G., and White, A. (1996). *Atlas of United States Mortality*. U.S. Department of Health and Human Services.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* **12**, 1151-1172.

Sinharay, S. and Stern, H. S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference* **111**, 209-221.

Yan, G. and Sedransk, J. (2004). Evaluation of Bayesian Model Diagnostic Techniques for Hierarchical Data. Technical report.

Received May 27, 2005; accepted December 6, 2005.

Guofen Yan
Division of Biostatistics and Epidemiology
Department of Public Health Sciences
University of Virginia
PO Box 800717
Charlottesville, Virginia 22908-0717, USA
gy4g@virginia.edu

J. Sedransk
Department of Statistics
Case Western Reserve University
Cleveland, Ohio 44106, USA
jxs123@case.edu