Using a Mixture Model for Multiple Imputation in the Presence of Outliers

Michael R. Elliott

Assistant Professor of Biostatistics

Department of Biostatistics

University of Michigan School of Public Heath

Assistant Research Scientist

Institute for Social Research

International Policy on Health Conference Statistics

Philadelpia, PA, January 17, 2008.

Funded by NHLBI Grant R01-HL-068987-01

Overview

- Motivating example: Healthy for Life Project
  - Masking
  - Swamping
- Mixture model for outliers
- Application
  - Model assessment
  - Effect of outliers on obesity assessment
- Discussion/Future Work

Healthy for Life Project

- To ascertain the prevalence of pediatric obesity in medically underserved areas, the Healthy For Life Survey obtained data from a probability sample of children using Health Resource and Service Administration (HSRA) supported Community Health Centers at least once during calendar year 2001 (Stettler et al. 2005).
- Compute body-mass index (BMI) and Box-Cox transform as a function of age and gender (Cole 1990); if BMI "z-score" exceeds 95^{th} percentile of reference population, child is classified as obese.
- Abstract height and weight during last visit to the health clinic in 2001.

Healthy for Life Project: Missing Height Data

- One-fourth of height data missing.
  - Height measured only sporadically; less likely to be observed among older children and children seen more frequently at the clinic.
- Use multiple imputation to reduce bias and inefficiency associated with a complete-case analysis (Stettler et al. 2005).
  - Potentially problematic: data overdispersed and included incorrectly recorded or abstracted elements.
  - Failure to account for abstraction errors may cause insufficient standardization between centers to be interpreted as unequal risk for pediatric obesity.
- Standardization in multi-center studies is expensive; propose analytic alternative to outlier correction when extensive training impossible.

Outlier Detection

- Using standard methods such as consideration of Mahalanobis distance to identify multiple outliers in multivariate data is problematic (Campbell 1980, Rousseeuw and van Zomeren 1990; Hadi 1992).
  - "Masking" prevents identification of outliers when a small cluster of observations inflates the empirical covariance matrix.
  - "Swamping" can make some observations appear to be outliers when true outliers pull the empirical covariance matrix away from non-outlier observations.

Removing Outliers via Multiple Imputation

- Goal is to develop method to assess distribution of population after removing outliers likely due to clerical errors.
- Mixture model defined by latent classes that have common means but differing covariances
  - "Clerical error class" is class with the largest covariance matrix determinant.
  - Multiple imputation imputes both item missingness and latent classes.
  - Subjects assigned to the clerical error class at a given imputation are dropped before the complete-data analysis of the observed and imputed data.

Accounting for Complex Sample Design

- Include design variables in mean model
- Consider association between posterior distribution of latent class membership and probability of selection
- Utilize standard design-based analyses at the complete-data stage of analysis to further enhance robustness.
- Use of MI to compute obesity estimates relies more heavily on the empirical distribution of the data than a fully model-based approach.

Previous Work

- Methods for simultaneous assessing outliers and accounting for missing data in a multiple imputation framework include Little and Smith (1987), Little (1988), Penny and Jollife (1999), and Ghosh-Dastidar and Schafer (2003, 2006).
- Similar to Ghosh-Dastidar and Schafer (2003, 2006) "multiple-edit-multiple imputation" (MEMI) model.
  - Embed the clerical error class in a larger mixture model
  - Consider AIC, BIC, and posterior predictive distribution p-values to select among differing class sizes considered
  - Develop the model in a fashion to more explicitly account for complex sample design.

Complete Data Mixture Model: Likelihood

$$\boldsymbol{Z}_i \mid C_i = k \sim N_q(\boldsymbol{\mu}_i, \Sigma_k)$$

$$C_i \sim MULTI(1, p_1, \ldots, p_K)$$

where $\boldsymbol{Z}_i$ is a q-dimensional outcome of interest, $\mu_{ij} = \boldsymbol{x}_i^T \boldsymbol{\beta}_j$, $j = 1, \ldots, q$, $\mid \Sigma_1 \mid < \cdots < \mid \Sigma_K \mid$.

- Mean of each subject depends on p covariates $\boldsymbol{x}_i$, and a covariance given by his or her latent variance class membership given by C_i.
- Class K is the "clerical error" class with the largest variability.
  - Assume that responses with clerical errors have the same mean but larger variability than other responses.

Complete Data Mixture Model: Priors

$$p(\boldsymbol{\beta}) \sim N(0, V_\beta)$$

$$p(\Sigma_k) \sim INV - WISHART(2, S_k), \;\; k = 1, \ldots, K$$

$$p(p_1, \ldots, p_K) \sim DIRICHLET(1, \ldots, 1)$$

Missing Data

- C_i are missing for all subjects
- Allow some components of $\boldsymbol{Z}_i$ to be missing under missing at random (MAR) assumption (Rubin 1978): conditional on the observed elements of $\boldsymbol{Z}_i$, the missingness status of the elements of $\boldsymbol{Z}_i$ is unrelated to their value.

Model Estimation

- Gibbs sampler data augmentation algorithm (impute missing elements of $\boldsymbol{Z}_i$ and the completely unobserved C_i at each step of the algorithm).

Multiple Imputation

Take m independent draws of $\boldsymbol{Z}^{comp}$ given by replacing the missing elements of $\boldsymbol{Z}$ with their imputed values, analyze using standard complete data procedures, and combine (Rubin 1987):

$$\hat{Q} = m^{-1} \sum_{t=1}^{m} Q\left(\boldsymbol{Z}^{comp(t)}\right).$$

where

$$V^{1/2}(\hat{Q} - Q) \sim t_{\nu}$$

Multiple Imputation

for

$$
\begin{aligned}
V &= U + (1 + m^{-1})B \\
U &= m^{-1} \sum_{t=1}^{m} \widehat{\mathrm{Var}} \left(Q \left(\boldsymbol{Z}^{comp(t)} \right) \right) \\
B &= (m-1)^{-1} \sum_{t=1}^{m} (\hat{Q} - Q \left(\boldsymbol{Z}^{comp(t)} \right))^2 \\
\nu &= (m-1) \left[1 + \frac{U}{(1 + m^{-1})B} \right]^2
\end{aligned}
$$

Delete subjects assigned to the Kth latent class when computing $Q(\boldsymbol{Z}^{comp(t)})$.

Multiple Imputation: Accounting for Complex Sample Design

- Include covariates for sample design in mean regression $\rightarrow$ no need to incorporate the sample design further if C_i independent of selection probability.
- If latter fails, minimize effect of model misspecification by using case-weighted estimates of $Q\left(\boldsymbol{Z}^{comp(t)}\right)$ and Taylor Series linearization estimates for $\widehat{\text{Var}}\left(Q\left(\boldsymbol{Z}^{comp(t)}\right)\right)$ (Woodruff 1971).
  - "Uncongeniality" (Meng 1994): analyst assumes more than the imputer.

Application to the Healthy for Life Project

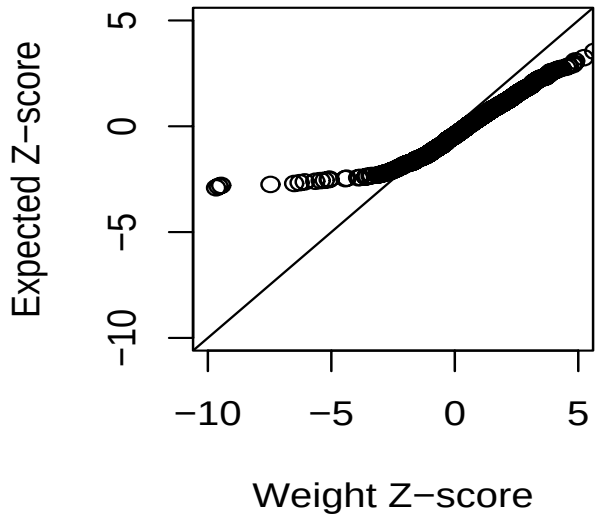
- Probability sample of children aged 2-11 served at one of 141 HRSA-supported Community Health Centers in NJ, NY, PR, VI, DE, DC, MD, PA, VA, and WV between 1/1/01 and 12/31/01.
- Stratified sample of 30 centers, with second-stage sample of approximately 100 children/center stratified by age (2-5 vs. 6-11).
- Inverse probability-of-selection case weights were post-stratified to known age group-region (US mainland urban, suburban, and rural, Puerto Rico (PR) urban and non-urban, and New York City Chinatown) totals.

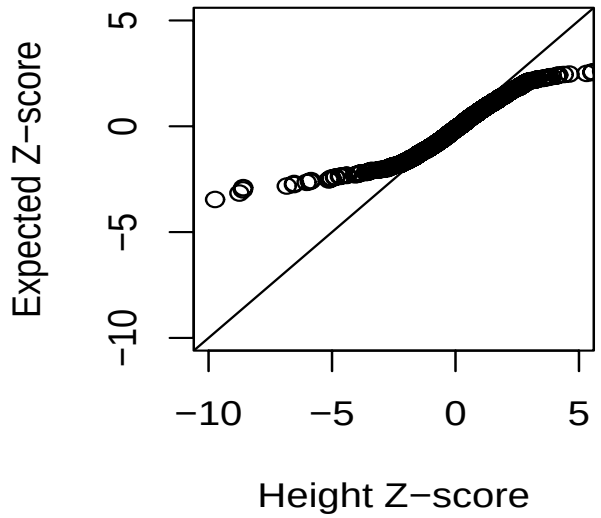
Application to the Healthy for Life Project

- Dropped 373 cases because of unknown age, gender, or both height and weight information; additional 3 cases dropped because of unknown weight information (to simplify analysis). 2,474 cases remained, of which 606 were missing height data.
- Improve normality approximation via "z-score" transformation (Weiss et al. 2004):

$$Z_{ij} = \frac{(Y_{ij}/M_{ij})^{L_{ij}} - 1}{L_{ij} S_{ij}},\ i, = 1, \ldots, n\ j = 1, 2$$

where Y_{i1} and Y_{i2} are the raw weights (kg) and height (m) measures, and $L_{ij} = L_j(A_i, G_i)$, $M_{ij} = M_j(A_i, G_i)$, and $S_{ij} = S_j(A_i, G_i)$ are known population parameters that are functions of the age A_i and gender G_i (Cole 1990).


Expected Z−score
5
0
−5
−10
−10
−5
0
5
Weight Z−score


Expected Z−score
5
0
−5
−10
−10
−5
0
5
Height Z−score

Modeling the Healthy for Life Project data

- $\boldsymbol{x}_i$ consists of age group-by-center dummy variables, to accommodate within-center correlation systematic association between BMI and the probability of selection.
- Restrict $\rho_k = \frac{\sigma_{12k}}{\sigma_{11k}\sigma_{22k}} \equiv \rho$ for $k = 1, \ldots, K-1$.
- Assume

$$V_\beta = 1000 I_2$$

$$p(\log \sigma_{jjk}) \overset{ind}{\sim} N(0, 4)\ j = 1, 2, k = 1, \ldots, K-1$$

$$p(\rho) \sim U(-1, 1)$$

$$S_K = 5 I_2$$

Results of Model Fit

- Both AIC and BIC suggest that the 3-class model provides the best fit to the data.

	p_k	σ^2_{11k}	σ^2_{22k}	ρ_k
k=1	$.912_{.873,.936}$	$1.43_{1.35,1.55}$	$1.14_{1.04,1.24}$	$.70_{.67,.72}$
k=2	$.072_{.049,.106}$	$3.88_{2.40,6.07}$	$12.34_{7.01,18.83}$	$.70_{.67,.72}$
k=3	$.015_{.007,.029}$	$37.48_{21.14,83.88}$	$29.23_{15.23,64.03}$	$.92_{.63,.98}$

Model Checking

- Test distributional assumptions using posterior predictive distributions (Gelman, Meng, and Stern, 1996). Consider PPD of $S = n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K} I(C_i = k) \tilde{Z}_{ki}^2$ where

  $$\tilde{Z}_{ki}^2 = \begin{cases} (\boldsymbol{Z}_i - \boldsymbol{\mu}_i)^T \Sigma_k^{-1} (\boldsymbol{Z}_i - \boldsymbol{\mu}_i) \text{ if } Z_{i2} \text{ is observed.} \\ (Z_{i1} - \mu_{i1})^2 / \sigma_{11k}^2 \text{ if } Z_{i2} \text{ is missing.} \end{cases}$$

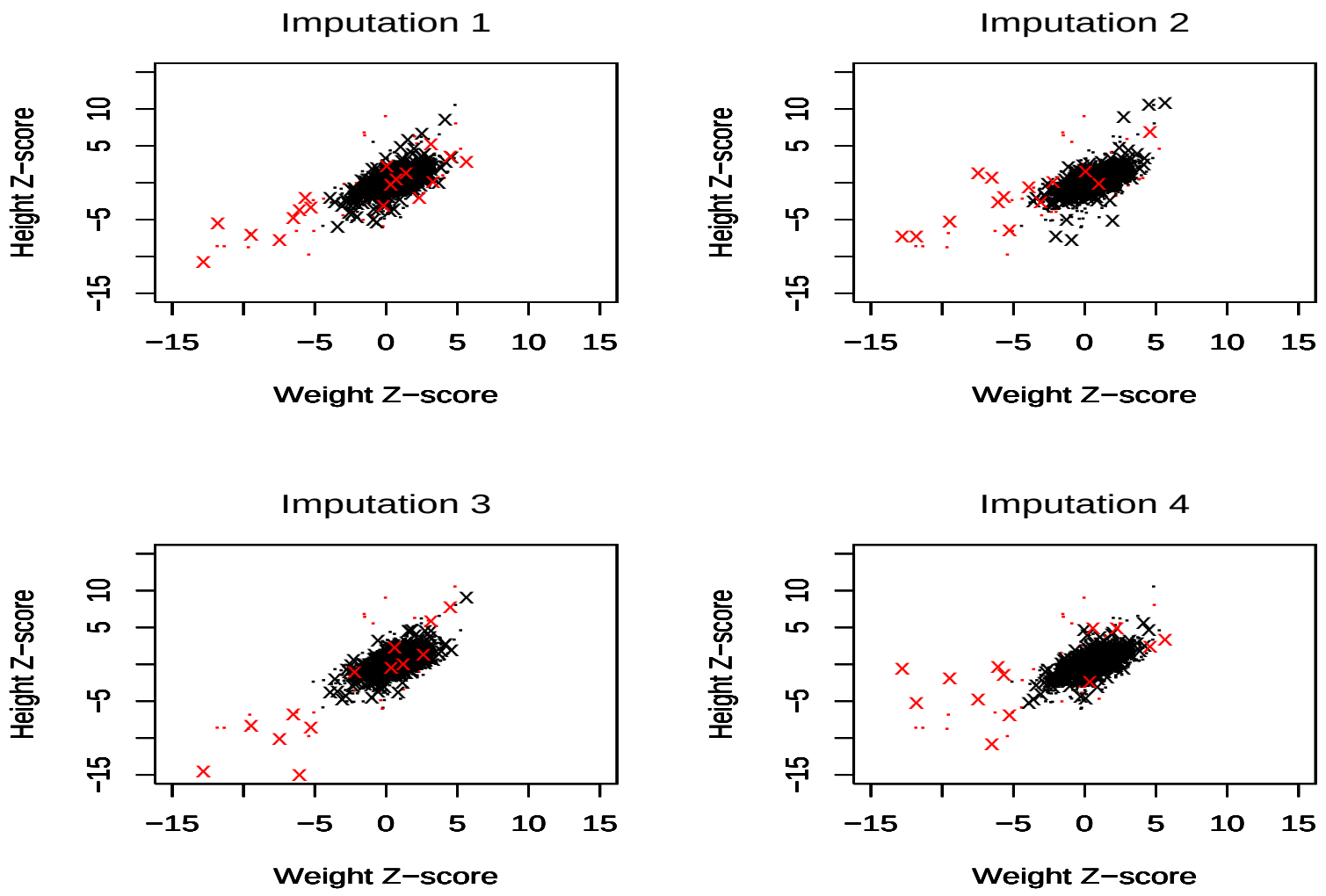
  . If the number of classes is sufficient, the normality assumption within class will hold, at least approximately, and S^{obs} and S^{rep} will correspond.

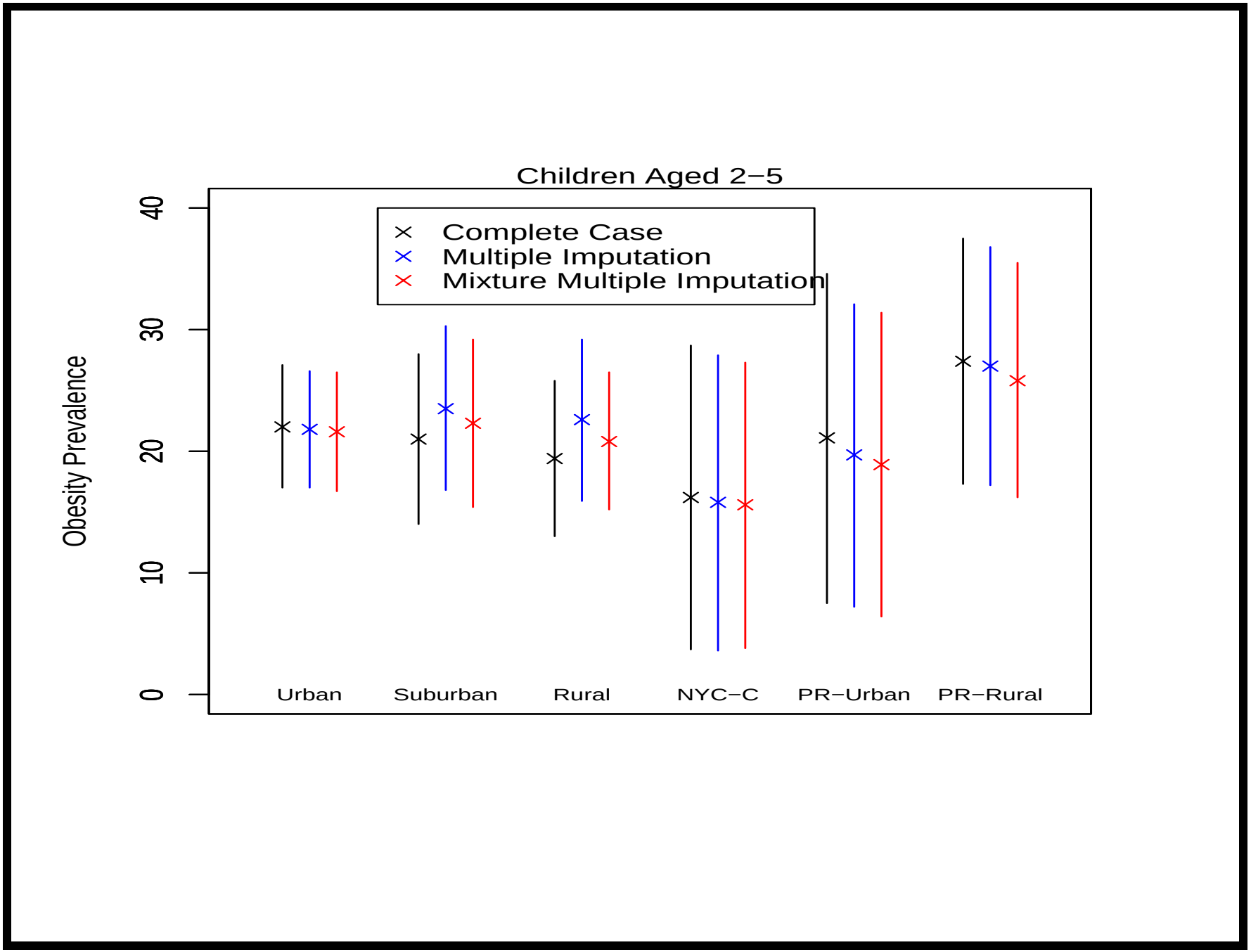
  - $P(S^{obs} < S^{rep}|y) = .46$ for the 3-class model.

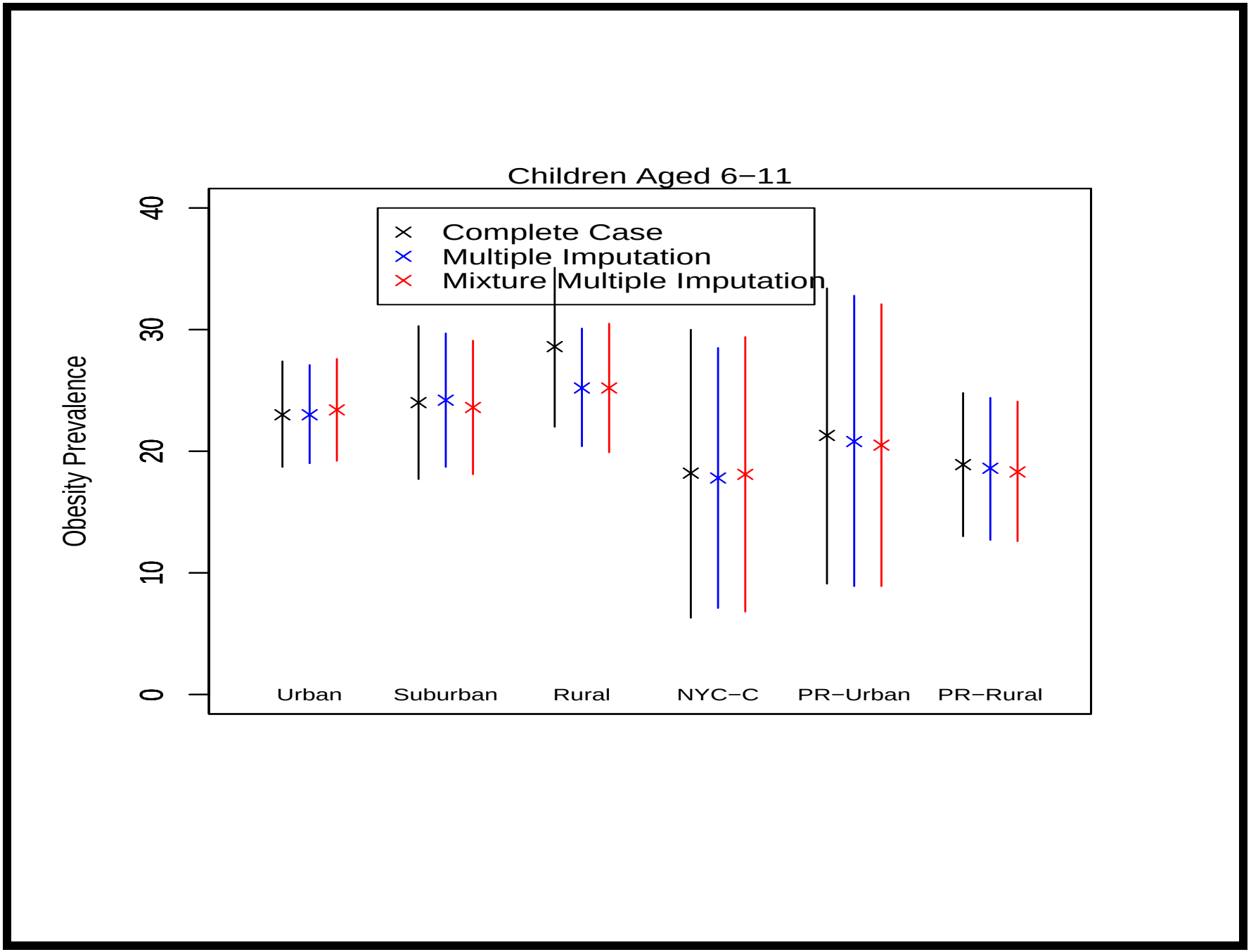
Model Checking

- Spearman correlations between the posterior medians of latent variance class probability membership for the 3-class constrained model $\hat{\pi}_{ki}$ and inverse of the case weight $1/w_i$: -.022 for $k = 1$ (p=.27), .008 for $k = 2$ (p=.69),and .012 for $k = 3$ (p=.54).
  - Suggests that including the centers as fixed effects has been sufficient to remove design effect from the model.

Multiple Imputation




Children Aged 2−5
Complete Case
Multiple Imputation
Mixture Multiple Imputation
Obesity Prevalence
0
10
20
30
40
Urban
Suburban
Rural
NYC−C
PR−Urban
PR−Rural


Children Aged 6−11
Complete Case
Multiple Imputation
Mixture Multiple Imputation
Obesity Prevalence
0
10
20
30
40
Urban
Suburban
Rural
NYC−C
PR−Urban
PR−Rural

Effect of Outliers

- If height data is missing and an older child incorrectly noted as younger, the resulting weight z-score would be extremely large, likely yielding a large BMI after height imputation, and potentially classifying a non-obese child as obese; the reverse is true if a younger child is incorrectly noted as older.
- Since children are more likely than not to be non-obese, the net effect of age transcription errors should be to inflate obesity rates among younger children, and deflate to a much lesser degree obesity rates among older children.
- Analysis of 2.5% and 97.5% quantiles suggested that younger children tended to have large BMI outliers and older children tended to have small BMI outliers, consistent with clerical errors in age.

Simulation Study

- $\boldsymbol{Z}_i \mid C_i = k \sim N_2(0, \Sigma_k), \Sigma_k = \sigma_k \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}$ for $k < K$ and $\Sigma_K = \sigma_K \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, and $n = 500$.
- Delete Z_{i2} under an MAR mechanism, so that large values of Z_{i1} tended to be associated with missing Z_{i2} unless the observation was a member of the "outlier" class.
- Simulation A: $K = 2$, $\sigma_1 = 1$, $\sigma_2 = 100$, $\rho = .5$, $\boldsymbol{p} = (.98\ .02)^T$;
- Simulation B: $K = 4$, $\sigma_1 = .25$, $\sigma_2 = 1$, $\sigma_3 = 9$, $\sigma_4 = 100$, $\rho = .5$, $\boldsymbol{p} = (.225\ .225\ .225\ .10)^T$.

Simulation Study: Target parameters of interest

Mean and 95% confidence interval for

- $p = P(Z_{i2} < Z_{2_{(.9)}})$ where $Z_{2_{(.9)}}$ is the 90th percentile for Z_2
- $\rho_k = \rho$ for $k < K$:

Simulation Study: Results

- When the fraction of outliers was small, both imputation methods correctly estimated the proportion of the Z_{i2} observations above the 90th percentile.
- When the fraction of outliers was large, standard imputation overestimated the fraction belonging to 90th percentile and above by approximately 30 percent. Mixture imputation resulted in correct inference for percentile.
- Standard imputation methods did not sufficiently correct for the bias toward the null in the estimation of the non-outlier correlation when the fraction of outliers was small.
- The estimate of the common correlation was essentially unbiased under all scenarios under the mixture imputation, and the coverage was approximately correct despite the tendency to underestimate the model size, especially with BIC.

Summary

- LC model for variability that simultaneously accounts for missing data and clerical error outliers.
- MI framework allows estimation to proceed using standard design-based methods for complete-case analyses.
  - Including transcription errors in the HSRA analysis lead to modest overestimates of obesity among younger children in selected subregions with higher transcription error rates, but for most subdomains their impact appears to be minimal.
- Method also suggests a class of 5-10% of the population overdispersed by a factor of 2-4; may be of clinical interest as an obesity/malnutrition cluster.

Future Work

- In HSRA application, MAR assumption reasonable. NMAR model (sensitivity analysis as a function of non-identified parameters) is also possible.
- A more fully design-consistent model would cross-classify the dispersion classes by the probability of selection (Elliott and Sammel 2002).
- Model assumes "outlying completely at random"
  - Treat the transcription errors as missing and impute both height and weight z-scores.
- Implement fully Bayesian method to accommodates uncertainty in the number of classes: add a model choice step to the Gibbs routine via a product space search (Carlin and Chib 1995) or reversible jump (Green 1995) step.