SMALL AREA ESTIMATES OF OVERWEIGHT PREVALENCE USING THE THIRD NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY (NHANES III)

Donald Malec (National Center for Health Statistics), William Davis and Xin Cao (Klemm Analysis Group)
Donald Malec, National Center for Health Statistics, 6525 Belcrest Rd., Rm. 915, Hyattsville, MD 20782

Key Words: body mass index, NHANES III, hierarchical model, Markov Chain Monte-Carlo

Abstract

Using a hierarchical model, we estimate the overweight prevalence for U.S. adults using NHANES III data. We provide a model-based justification for the use of the statistical weights for estimation through subsampling. We compare our model-based estimates with design-based estimates at the national level and obtain agreement. Also, we display the model-based prevalence estimate by state.

1. Introduction

There is a continuing need to assess health status, health practices and health resources at both the national and subnational level. Estimates of these health items help determine the demand for health care and the access individuals have to it. Although the NCHS personal interview surveys can provide much of this information at the national level, little can be provided for states and counties because of excessive field costs. The need for subnational health statistics exists, however, because health and health care characteristics are known to vary geographically. Also, health care planning often takes place at the state and local level.

One alternative approach for producing subnational estimates has been to, effectively, increase the sample size by utilizing models defined across the subnational areas (e.g., see Ghosh and Rao, 1994). A challenge has been to use models realistic enough to produce accurate estimates. Towards this end, hierarchical models (models which include geographic variation among rates) have been adapted to small area estimation. With the availability of Markov Chain Monte-Carlo (MCMC) methods, estimates (and estimates of precision) can be made that account for all model errors. Given current resources, model-based estimates can be made for subnational levels. At a minimum, a measure of the geographic variability of health characteristics can be determined in order to decide which health variables should be included in future surveys for subnational statistics.

In this paper we present a methodology for making subnational estimates using hierarchical models that take sample selection into account. We illustrate the methodology by estimating the adult overweight prevalence by state using data from NHANES III. The methodology is general and is especially useful for producing subnational estimates which, at a national level, may coincide with design-based estimates.

1.1 NHANES III: Survey Design

NHANES III is a stratified, multi-level design that was conducted during the years 1988-1994 in two phases (phase 1: 1988-91 and phase 2: 1991-94). Sampled persons provide health and dietary information through a questionnaire and also through a physical exam (including labwork) and a dental exam. The resulting sample of approximately 40,000 persons was selected to represent the civilian, non-institutional population of the United States and provide national characteristics and nutrition status for the entire population and a number of age, race and ethnic subgroups.

Although not excluded from the target sample, small numbers of Black and Mexican-American persons were included in previous NHANES. Therefore, reliable estimates of their health and nutritional status could not be obtained for their subgroups. To resolve this problem, NHANES III was designed to include a large sample of these two largest minority groups of the U.S. population.

1.2 Overweight prevalence in U.S.

Overweight is associated with a number of adverse health outcomes including mortality (Troiano *et al.*, 1996) and has become an increasing problem in the U.S. Kuczmarski *et al.* (1994) documented the recent increased prevalence of overweight in the U.S. adult population. Overweight is typically defined in terms of body mass index (BMI) which is defined by

$$\text{BMI} = \text{Weight}/\text{Height}^2 \qquad (1)$$

When BMI is expressed in the units kg/m^2, overweight is defined as >27.8 for men and >27.3 for women. These are the sex-specific 85th percentile of BMI for men and women aged 20 through 29 from NHANES II (1976-80).

2. Estimation Methodology

In this section, we give a general method for estimating prevalence at a subnational level and specify the model that we used to make state estimates for overweight prevalence.

2.1 Description of Finite Population

We define:

- i: County Indicator, i=1,...,H
- d: characteristics of individual (i.e., age by race by sex by phase classes)
- j: Individual ID, nested within county and characteristic
- Y_{idj}: A characteristic of interest for individual, i,d,j.

Of interest, are estimates of the finite population mean for individual characteristics defined by groupings of 'd', for

local areas defined by county groupings. That is,

$$\theta_{LD} = \sum_{i\in L}\sum_{d\in D}\sum_{j=1}^{N_{id}} Y_{idj} \Big/ \sum_{i\in L}\sum_{d\in D} N_{id} \qquad (2)$$

where L indexes a particular collection of counties (e.g., all counties in a specific state), D is the set of specific subgroups of interest (e.g., all females regardless of age or race) and N_{id} the total civilian, non-institutional household population in county i, subgroup d. The subgroups, d, are defined by a cross classification of the following four discrete variables:

- gender
- race/ethnicity (white non-Hispanic, black non-Hispanic, and Mexican-American)
- phase (1988-91, 1991-94)
- age categories (20-24, 25-29, ...,75-79, 80+)

2.2 The Population Model

Recall that Y_{idj} denotes the status for individual j in group d in county i where i=1,..., H and j=1,...,N_{id}. In this paper, Y_{idj}=1 denotes overweight status as determined by BMI and Y_{idj}=0 denotes either normal or underweight. Within county i and conditional on the p_{id}, the Y_{idj} are assumed to be independent Bernoulli random variables with:

$$Pr(Y_{idj}=1\,|\,p_{id})=p_{id},\ j=1,...,n_{id} \qquad (3)$$

Since there will often be counties with little or no data in subgroup, d, a model expressing the similarities between areas could improve estimation. We use a special case of the Generalized Linear Mixed Model (GLMM) as specified in Breslow and Clayton (1993) with a logit link:

$$logit\{p_{id}\} = \underline{x}_{id}^t\underline{\alpha} + \underline{z}_{id}^t\underline{\beta}_i \qquad (4)$$

where the parameters $\underline{\alpha}$ and $\underline{\beta}_i$ denote the fixed and random effects respectively with

$$\underline{\beta}_i \sim N(\underline{0},\Gamma) \qquad (5)$$

where, conditional on Γ, the $\underline{\beta}_i$'s are independent. The vectors $\underline{x}_{id}$ and $\underline{z}_{id}$ denote the explanatory variables of the fixed and random components of the model, respectively. These variables must be known for all counties and groups for which a sample is selected or a prediction made.

By specifying Γ=0, this model reduces to a Logistic regression model. However, using GLMM, Γ remains unspecified and can be estimated from the data.

2.3 Overweight Status Model

Each of the 156 subgroups d is defined by one of the 6 possible combinations of race and gender and one of the 26 possible combinations of age and phase. We represent this as d=d(c,a) where c labels the race/gender and "a" labels the age/phase. We found the following specification of the model (4)-(5) adequate for overweight status analysis:

$$logit\{p_{id}\} = \alpha_d + \beta_{ic} \qquad (6)$$

where d=d(c,a) and $\underline{\beta}_i^t=(\beta_{i1},...,\beta_{i6})$, where the six components correspond to the race/gender categories.

2.4 Estimation

Letting $\underline{y}_s$ denote the sampled data, we use a Bayesian approach to make inference about θ_{LD}. The posterior mean of θ_{LD}, $E(\theta_{LD}|\underline{y}_s)$, and the posterior variance of θ_{LD}, $V(\theta_{LD}|\underline{y}_s)$, are used. These moments can be shown to be (see, e.g. Malec *et al.* 1993)

$$E(\theta_{LD}|\underline{y}_s)=\frac{\sum_{i\in L}\sum_{(idj)\in s} y_{idj}+\sum_{i\in L}\sum_{d\in D}\{N_{id}-n_{id}\}E(p_{id}|\underline{y}_s)}{\sum_{i\in L}\sum_{d\in D} N_{id}} \qquad (7)$$

and

$$V(\theta_{LD}|\underline{y}_s)=\frac{\sum_{i,d}\{N_{id}-n_{id}\}E\{p_{id}(1-p_{id})|\underline{y}_s\}+V\left(\sum_{i,d}\{N_{id}-n_{id}\}p_{id}|\underline{y}_s\right)}{(\sum_{i,d} N_{id})^2} \qquad (8)$$

where the sums in (8) are over LxD. We determine (7) and (8) numerically using the method described in section 4.

3. Sampling Adjustment for Modeling

In complex surveys, there is a general concern that ignoring the selection of a sample may produce erroneous inferences (see, e.g. Scott, 1977). In principle, incorrect analyses can be avoided by specifying a model that accounts for sample selection (e.g., see Krieger and Pfeffermann (1992)). However, for heavily oversampled designs such as the NHANES III, the level of effort needed to model the effect of the design on each outcome may be great. The approach taken here is to choose a subsample that is free of the limitations of the original sample to guarantee that the design does not cause a selection bias. In particular, a subsample is selected so that the overall selection probability, given samples of fixed size, is equal.

For this application, inference is conditional on demographic characteristics. Hence, subsampling adjustment is only needed within demographic groups. In addition, a subsampling adjustment to eliminate correlation due to PSU selection is not needed because it is modeled by GLMM. Here, we assume that the observation within PSUs are independent. In another context, a more comprehensive subsampling adjustment is used by Hinkins *et al.* (1994), who subsample to obtain a complete SRS from an (originally) stratified sample.

The next two sections provide a justification for our use of the weights in the modeling process.

3.1 Implementation

We define a subsample, within each age, race/ethnicity, gender, and phase cell, d, as follows: Given that the initial

sample of units: $r_1,...,r_n \in S_d$ were selected with marginal probabilities: $\pi_{r_1},...,\pi_{r_n}$. We could select a subsample: $S_{d2} \subseteq S_d$ with marginal probabilities:

$$\pi^*_{r_i} = \frac{\pi^{-1}_{r_i}}{\max_{j \in S_d} \pi^{-1}_j} \quad , \quad r_i \in S_d \ .$$

For a fixed sample size the overall marginal selection probability that $r_i \in S_{d2}$ is independent of π_{r_i}.

We can now apply the population model outlined in (3)-(5) to S_{d2}. From (3), define

$$f(Y_{idj}|p_{id}) = p_{id}^{Y_{idj}}(1-p_{id})^{1-Y_{idj}}. \qquad (9)$$

The density of subsampled measurements $\underline{y}_s=\{y_{idj}\}$ and the random-effect parameter $\underline{\beta}=(\underline{\beta}_i,...,\underline{\beta}_H)$ given $\underline{\phi}=(\underline{\alpha}, \Gamma)$ and S_{d2} is:

$$f(\underline{y}_s,\underline{\beta}|\underline{\phi}) = \prod_d \prod_{idj \in S_{d2}} f(y_{idj}|p_{id}) f(\underline{\beta}|\Gamma)$$

where

$$f(\underline{\beta}|\Gamma) = f(\{\underline{\beta}_i\}_{i \in S}|\Gamma)\ f(\{\underline{\beta}_i\}_{i \notin S}|\Gamma) \ .$$

This can be specified in terms of S_d as:

$$f(\underline{y}_s,\underline{\beta}|\underline{\phi},\underline{\delta}) = \prod_{idj \in S} (f(y_{idj}|p_{id}))^{\delta_{idj}} f(\underline{\beta}|\Gamma) \qquad (10)$$

with $\underline{\delta}=(\delta_{idj})$ and $\pi^* = (\pi^*_{idj})$, where

$$Prob(\delta_{idj}=1|\pi^*) = \pi^*_{idj} \ . \qquad (11)$$

Given $\underline{y}_s$ and $\underline{\delta}$, the posterior distribution of $\underline{\Omega}=(\underline{\beta},\underline{\phi})$ is

$$f(\underline{\Omega}|\underline{y}_s,\underline{\delta}) = \frac{f(\underline{y}_s,\underline{\beta}|\underline{\phi},\underline{\delta}) f(\underline{\phi})}{\int f(\underline{y}_s,\underline{\beta}|\underline{\phi},\underline{\delta}) f(\underline{\phi})\, d\underline{\Omega}} \qquad (12)$$

where $f(\underline{\phi})$ denotes the prior distribution of $\underline{\phi}$. By selecting a single subsample, S_{d2}, we can obtain the conditional mean of functions of $\underline{\Omega}$ such as p_{id} (as needed in (7)) using (12).

If the marginal selections probabilites of the original sample are different, a subsample may be much smaller than the original. We can rectify some of the inefficiency by using (11) to create a marginal likelihood:

$$f(\underline{y}_s,\underline{\beta}|\underline{\phi}) = \sum_{\underline{\delta}} Prob(\underline{\delta}|\underline{\pi}^*)\ f(\underline{y}_s,\underline{\beta}|\underline{\phi},\underline{\delta}) \qquad (13)$$

and carrying out the inference specified in (7) and (8).

3.2 Approximate Results

Using (13) requires extensive computation such as using an additional data-augmentation scheme. As an approximation, we replaced the sample indicators in (10) by their expected value, giving the approximate density:

$$f(\underline{y}_s,\underline{\beta}|\underline{\phi},\underline{\pi}^*) = \prod_{idj \in S} (f(y_{idj}|p_{id}))^{\pi^*_{idj}} f(\underline{\beta}|\Gamma) \ . \qquad (14)$$

We make inference about θ_{LD} by replacing (10) with (14).

The density, $\prod_{idj \in S} f(y_{idj}|p_{id})$, specified by (3) is in the exponential family with sufficient statistics $(\{n_{id}\},\{m_{id}\})$, where $m_{id}=\sum_{j=1}^{n_{id}} y_{idj}$. The approximate density, $\prod_{idj \in S} (f(y_{idj}|p_{id}))^{\pi^*_{idj}}$, is also in the exponential family with sufficient statistics $(\{n^*_{id}\},\{m^*_{id}\})$, where $m^*_{id}=\sum_{j=1}^{n_{id}} \pi^*_{idj}\, y_{idj}$ and $n^*_{id}=\sum_{j=1}^{n_{id}} \pi^*_{idj}$. The closure under sampling adjustment of the logistic distribution generalizes to the exponential family including the normal model. If the original sample selection is noninformative, a measure of the loss of information due to the sampling adjustment is $\sum_{i,d} n^*_{id} / \sum_{i,d} n_{id}$.

Our estimation methodology appears related to Folsom and Liu (1994), who make small area estimates using a survey weighted empirical-Bayes model. However, their model assumes that the variances are proportional to the inverse of the sample weights, while our methodology does not require this assumption.

4. Inferential Methodology

A Bayesian analysis requires the specification of a prior distribution for $(\underline{\alpha}, \Gamma)$. To insure that the sample information dominates the inference, we used an overdispersed prior distribution. In particular, we choose the conditional density of $\underline{\alpha}|\Gamma$ to be constant and an inverse Wishart distribution for Γ with one degree of freedom and mean=k I_{6x6}. This prior is dominated by the data but seems to avoid problems with use of vague priors in hierarchical models (Berger (1985, Sec. 4.6.2)).

Since the posterior moments of θ_{LD} are nonlinear functions of $\underline{\Omega}$, and the distribution $f(\underline{\Omega}|\underline{y}_s)$ cannot be expressed in a simple form, numerical evaluation is needed. We used the MCMC methodology to generate R sets of parameters, $\{\underline{\Omega}^{(r)}: r = 1,...,R\}$ from the posterior distribution and evaluate $p^{(r)}_{id}$ for each r.

We used block-at-a-time Metropolis-Hastings algorithm (Hastings, 1970, Chib and Greenberg, 1995) to generate one long run of the chain. The modes and Hessians were searched at each iteration to determine the candidate-generating densities of $\underline{\beta}$ and $\underline{\alpha}$. Conditionally Γ was sampled directly from its inverse Wishart distribution. We also used CODA

software (Best *et al.*, 1995) to perform the output analysis and convergence diagnosis for the chain.

4.1 Selecting County Covariates

We arrived at the model specified by (6) after first evaluating whether county covariates could explain some of the variation of prevalence rates. We selected county level variables to be included in $\underline{x}$ and $\underline{z}$ of (4) with a fixed effect model that included demographic information and county-level covariates from the Area Resource File (1989). We used the model

$$logit\{p_{id}\} = \underline{x}_{id}^t \underline{\alpha} + \underline{w}_i^t \underline{\eta} \qquad (15)$$

where $\underline{w}_i$ is a vector of 30 county level covariates that were thought by subject matter specialists to be related to overweight and $\underline{\eta}$ denotes the associated parameter vector. We give typical examples of county covariates:

- •Education (fractions with educ <9 years or college grad)
- •Economic (unemployment and poverty rate, home value)
- •Demographic (percent rural and urban, pop. density)
- •Labor force (fraction in construction, manufacturing, etc)
- •Health care (number of physicians, hospitals, beds, etc. per capita)

The county covariates were dominated by the demographic variables so we included only a random component in the model (4). The addition of all the county covariates to the demographic ones only increased the R^2 from 0.056 to 0.059.

5. Estimation Results

We illustrate the calculations using two choices of L and D from (2). In section 5.1, we show the estimates made at the national level for demographic subgroups. We compare our model-based estimates with design-based estimates to justify our claim that the results may coincide at a national level. In section 5.2, we show our estimates for all adults within the 50 states and D. C.

5.1 National Estimates by Demographic Subgroups

In this section we compare model and design-based estimates for 78 demographic categories.[1] Based on NHANES III phase 1, Kuczmarski *et al.* (1994) show that the overweight prevalence is highest for ethnic (Black non-Hispanic and Mexican-American) females.

We used the BMI values for all adults (20 and over) who were examined in a Mobile Examination Center (MEC). Of the 16,573 adults who were examined in a MEC, 16,523 had values for both height and weight -- hence BMI. We used all these values for estimation. We used standard expansion estimators to estimate the overweight prevalence for all 78 demographic categories using the MEC examination weights (Mohadjer *et al.* (1996)).

For the model-based estimates, we approximated the selection probabilities of Section 3.1 by the inverse of the MEC examination weights. For both phases, we used the 1990 Census counts as an approximation to N_{id} in equations (7) and (8). We used SAS IML (1995) for the calculations and 1200 iterations of the Gibbs sampler. The values shown were obtained for the prior distribution with $k=10^4$. We used sensitivity analysis to insure that our prior was overdispersed.

We compare the design and model-based estimates in Figure 1 for one of the six race/gender categories. The results of figure 1 are typical of the other categories. The model-based estimates tracked the design-based estimates well for all categories and all ages. We take the results as verification that our model-based estimates may match the design-based estimates at a national level.

5.2 State Estimates

We computed the overweight prevalence estimate by state and show the results in figure 2. The figure shows the following:

- •Relatively small range (.32 to .40)
- •North/South difference (reflecting difference in minority population)

Figure 3 shows the estimated coefficient of variation (CV) by state plotted against the square root of the number of adults in sample. All the CVs meet the NCHS publication standard of 30%.

Since the proportions do not exhibit much variation, the CVs are approximately proportional to the state standard deviations. If state data is preferentially used for state estimation, one would expect an inverse relationship. This relationship appears to be approximately correct. The only outlier is Alaska, which appears to have a larger variance. This may be because of its small population and because it was treated as a single county -- due to data limitations.

Acknowledgments

We want to thank Dr. Kurt Maurer for helpful comments and direction.

References

1. Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd Ed.), Springer-Verlag, New York.
2. Best, N., Cowles, M. K. and Vines, K. (1995), CODA: Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output, Version 0.30, MRC Biostatistics Unit, Cambridge.
3. Breslow, N. E. and Clayton, D.G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9-25.
4. Chib, S. and Greenberg E. (1995), "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*,

[1]The accuracy of the design-based estimates may not meet NCHS publication standards since we used 5-year age categories.

49, 327-335.
5. Folsom, R. E. and Liu, J. (1994), "Small Area Estimation for the National Household Survey of Drug Abuse," *Proceedings of the American Association Survey Methodology Section*, 565-569.
6. Ghosh, M., and Rao J.N.K. (1994), "Small Area Estimation: An Appraisal," *Statistical Science*, 9, 55-93.
7. Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97-109.
8. Hinkins, S., Oh, H.L. and Scheuren, F. (1994), "Inverse Sampling Design Algorithms," *1994 Proceedings of the American Statistical Association Survey Methodology Section*, 626-631.
9. Krieger, A. and Pfeffermann, D. (1992) "Maximum Likelihood Estimation from Complex Sample Surveys", *Survey Methodology*, 18, No 2, 225-239.
10. Kuczmarski, R. J., Flegal, K. M., Campbell, S. M. and Johnson, C. L. (1994), "Increasing Prevalence of Overweight Among US Adults: The National Health and Examination Surveys, 1960 to 1991," *Journal of the American Medical Association*, 272 No. 3, 205-211.
11. Malec, D., Sedransk, J. and Tompkins, L. (1993), "Bayesian Predictive Inference for Small Areas for Binary Variables in the National Health Interview Survey," In *Case Studies in Bayesian Statistics*, editors: Gatsonis, C., Hodges, J. S., Kass, R. E. and Singpurwalla, N. D., Springer-Verlag, New York, pp. 377-389.
12. Mohadjer, L., Montaquila, J., Waksberg, J., Bell, B., James, P., Flores-Cervantes, I. and Montes, M. (1996), "National Health and Nutrition Examination Survey III: Weighting and Estimation Methodology," Westat Inc., Rockville, MD.
13. *SAS IML Software: Usage and Reference Manual* (1995), Ver 6, SAS Institute, Cary, N.C.
14. Scott, A.J. (1977) "On the Problem of Randomization in Survey Sampling", *Sankhya*, 39, Series C, Pt.1, 1-9.
15. Troiano, R. P., Frongillo, E. A. Jr., Sobal, J. and Levitsky, D. A. (1996) "The relationship between body weight and mortality: a quantitative analysis combining information from existing studies, *International Journal of Obesity*, 20, 63-75.
16. U.S. Department of Health and Human Services (1989), "The Area Resource File (ARF) System", ODAM Report No. 7-89.

Figure 1. Design and Model-Based Estimates of Overweight Percentage for Black (non-Hispanic) Females using NHANES III

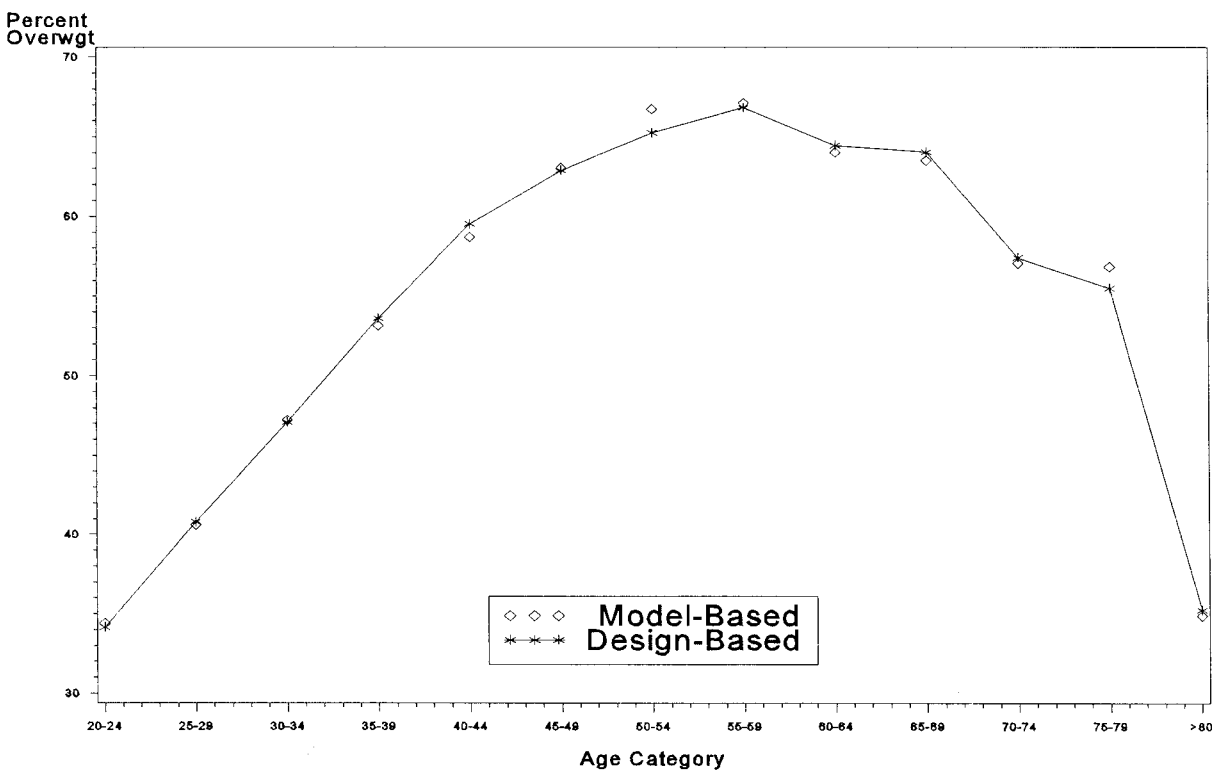

Figure 2: Proportion Overweight

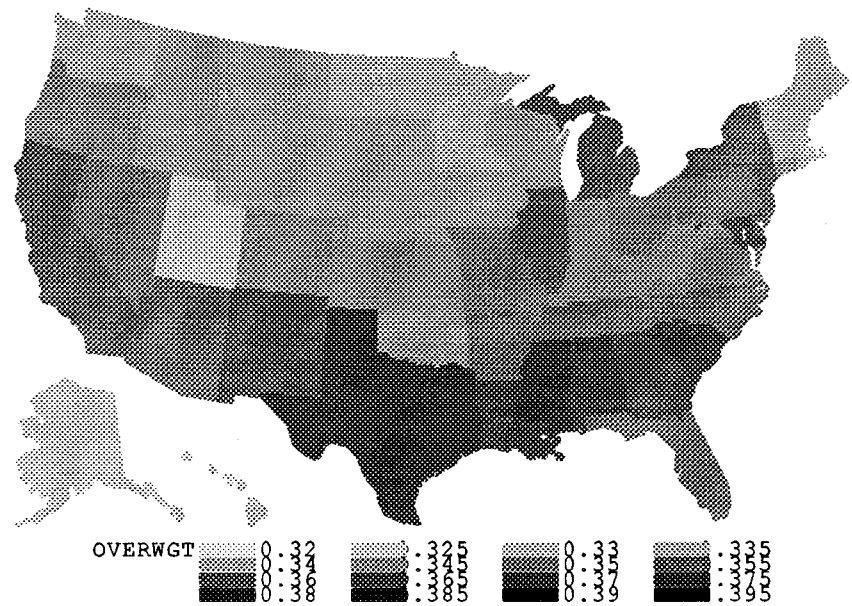

Figure 3. Adult Overweight Prevalence Coefficient of Variation vs. Square Root of State Sample Size for NHANES III