Analysis of National Health Interview Survey Data

Chris Moriarity

National Conference on Health Statistics

August 18, 2010

cdm7@cdc.gov

Presentation outline

National Health Interview Survey (NHIS) overview

NHIS estimates and variance estimates

Analysis methods for pooled (combined annual samples) NHIS data – need to account for year-to-year correlation

Analysis of multiply imputed income data

The National Health Interview Survey (NHIS)

Conducted continuously nationwide since July 1957

Personal visit interview protocol, collecting data on a broad range of health topics

NHIS home page URL:

www.cdc.gov/nchs/nhis.htm

Estimates from NHIS data

NHIS has a complex sample design, including higher sampling rates of certain groups (black, Hispanic, Asian persons) - sampling weights should be used to make estimates from NHIS data

Variance estimation procedure must take account of complex sample design in order to be valid

Software for NHIS variance estimation

Reference: excellent Web page maintained by Alan Zaslavsky

http://www.hcp.med.harvard.edu/statistics/survey-soft/

Software list, comparative summaries, review articles

Software package list at Alan's website

AM Software	free	American Inst. for Research
Bascula	$	Statistics Netherlands
CENVAR	free	U.S. Bureau of the Census
CLUSTERS	free	University of Essex
Epi Info	free	Centers for Disease Control
GES	$	Statistics Canada
IVEware	free	University of Michigan
PCCARP	$	Iowa State University
R survey	free	www.r-project.org
SAS/STAT	$	SAS Institute
SPSS	$	SPSS
Stata	$	Stata Corporation
SUDAAN	$	Research Triangle Institute
VPLX	free	U.S. Bureau of the Census
WesVar	$	Westat, Inc.

Variance estimation guidance at NHIS methods page - 1963 to 2009

www.cdc.gov/nchs/nhis/methods.htm

SUDAAN, Stata, R survey, SAS survey procedures, SPSS, VPLX: Sample code provided for use with NHIS data

SAS, SPSS: Guidance provided to avoid problems with missing DOMAIN/SUBPOP variables in analyses of NHIS data

NHIS year-to-year correlation: why?

The U.S. counties (PSUs) selected at the beginning of a sample design period remain the same for the entire sample design period

Consecutive annual sample cases tend to be close together geographically - they tend to have similar characteristics

Year-to-year correlation over a ~10 year sample design period

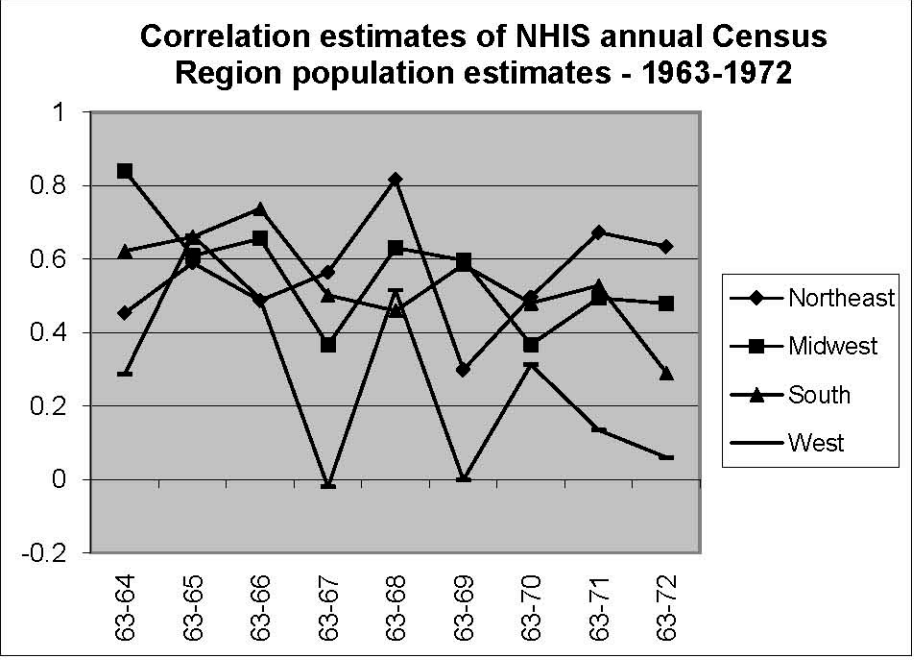
Correlation is present during the entire sample period

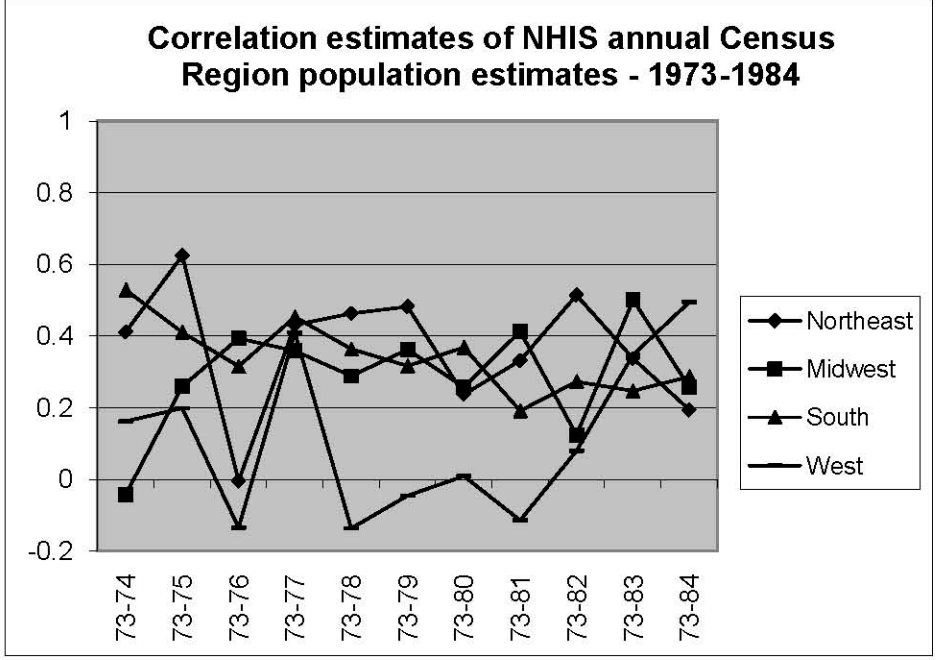
Correlation may be less for annual samples years apart than for annual samples closer together

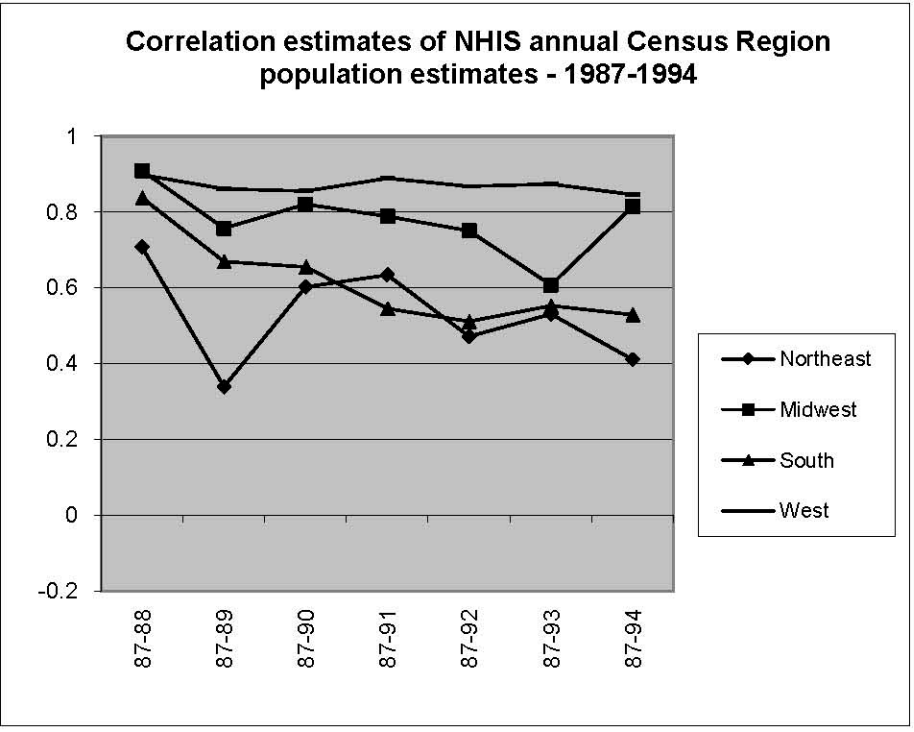
Year-to-year correlation example: Census Region population totals (4)

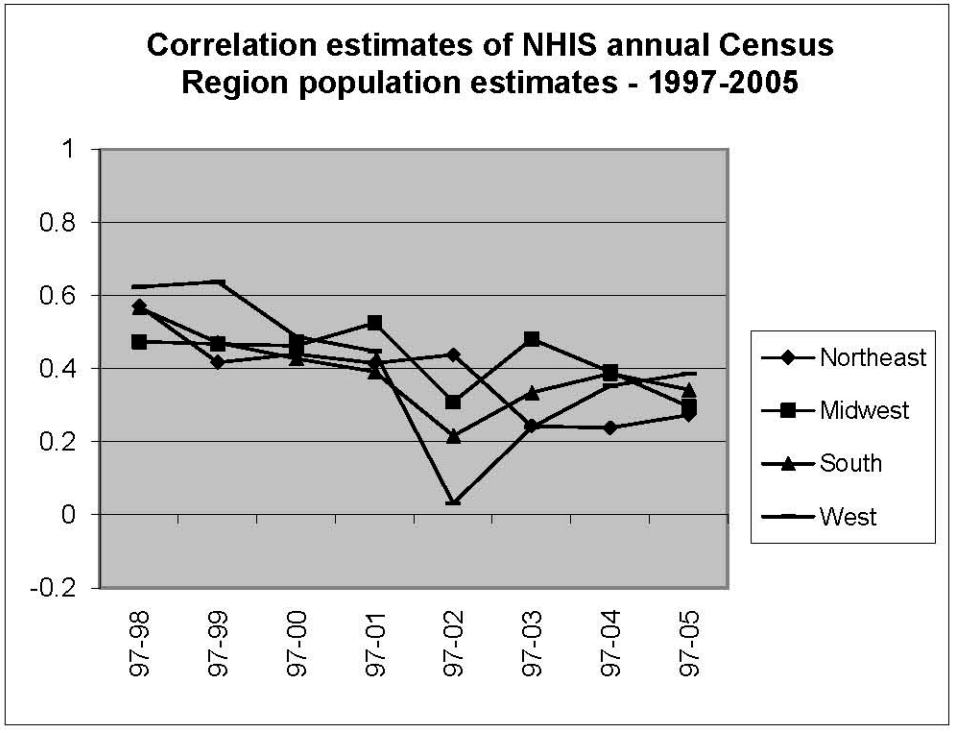
Available for all years NHIS microdata are available; Census Region consistently defined

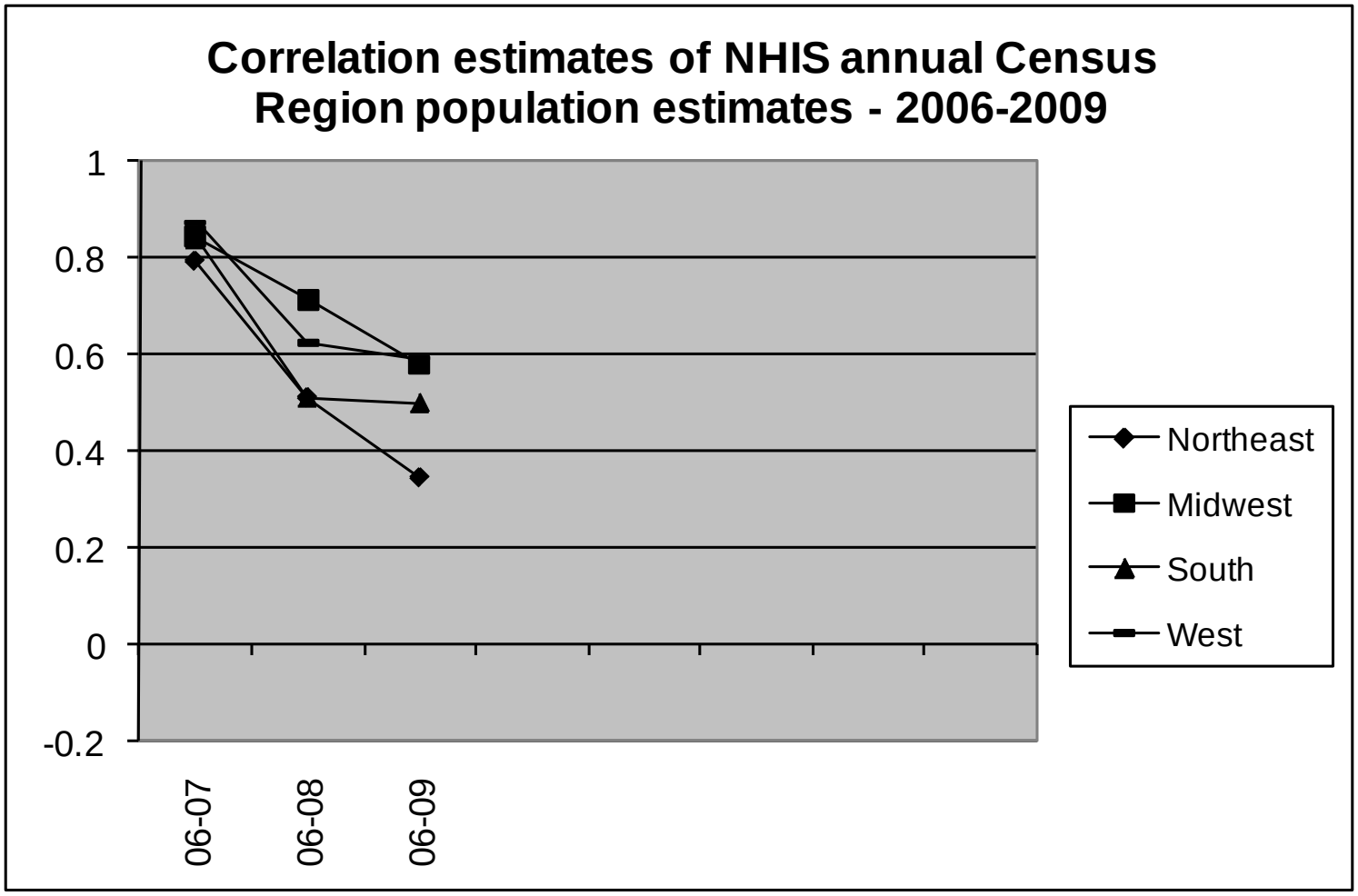
Reasonable to expect high level of correlation for adjacent years, perhaps a decline over time


Correlation estimates of NHIS annual Census Region population estimates - 1963-1972
1
0.8
0.6
0.4
0.2
0
-0.2
63-64
63-65
63-66
63-67
63-68
63-69
63-70
63-71
63-72
Northeast
Midwest
South
West


Correlation estimates of NHIS annual Census Region population estimates - 1973-1984
1
0.8
0.6
0.4
0.2
0
-0.2
73-74
73-75
73-76
73-77
73-78
73-79
73-80
73-81
73-82
73-83
73-84
Northeast
Midwest
South
West


Correlation estimates of NHIS annual Census Region population estimates - 1987-1994
1
0.8
0.6
0.4
0.2
0
-0.2
87-88
87-89
87-90
87-91
87-92
87-93
87-94
Northeast
Midwest
South
West


Correlation estimates of NHIS annual Census Region population estimates - 1997-2005
1
0.8
0.6
0.4
0.2
0
-0.2
97-98
97-99
97-00
97-01
97-02
97-03
97-04
97-05
Northeast
Midwest
South
West


Correlation estimates of NHIS annual Census Region population estimates - 2006-2009
1
0.8
0.6
0.4
0.2
0
-0.2
06-07
06-08
06-09
Northeast
Midwest
South
West

Variance estimation guidance for combined (pooled) analyses

Documentation for public use files available online at NHIS methods page: www.cdc.gov/nchs/nhis/methods.htm

Refer also to appendix "Merging Data Files and Combining Years of Data in the NHIS" in the annual NHIS survey description document, part of annual NHIS public use file data release

Variance estimation for pooled annual samples

Annual samples within a sample design period are not statistically independent

Annual samples in different sample design periods are (essentially) statistically independent

Variance estimation within a sample design period (dependent)

- Treat pooled annual samples like one big annual sample for variance estimation

- No recoding of variance estimation variables required

Variance estimation across sample design periods (independent)

Need to recode variance estimation stratum variables in different sample design periods to make sure they are different

Variance estimation stratum variable values always are <1000; use this fact when recoding

Variance estimation across sample design periods - recodes

Construct a new variance estimation stratum variable from existing variables by adding 1000 in one design period, 2000 in the next design period, etc.

This guarantees the values will be distinct in different design periods

Variance estimation for both "within" and "across"

Example: a 2004-2008 pooled analysis

Conceptually, the "within" step comes first: 2004-2005 in one sample design period, 2006-2008 in a different sample design period

Variance estimation for both "within" and "across" (continued)

Conceptually, the "across" step follows the "within" step: do recoding of variance estimation strata variables across the sample design periods (2004-2005 versus 2006-2008) while combining the five annual datasets into one pooled dataset

Recommended weight adjustment for all pooled analyses

Divide weights by the number of years being pooled - simple and defensible

Example: 2004-2008 pooled analysis (5 years): divide weights by 5

More sophisticated weight adjustment for pooled analyses

A user focusing on a particular pooled estimate may prefer a weight adjustment designed to minimize the estimate's variance

If sample sizes stable: both methods (simple, sophisticated) usually give similar weights

Before doing a pooled analysis - need to check data are similar

Analyses of pooled data are meaningful only when the data being pooled are similar

Question wording the same?
Answer categories the same?
Same target population?

1968: a special case for pooled analyses

There are 1968 calendar year and 1968 fiscal year (July 1967-early July 1968) data files; overlap of 67,608 persons

The overlap (January-early July 1968) should be removed for a pooled analysis that includes both fiscal and calendar 1968 data

Imputed NHIS income data

High item nonresponse to income questions

1990-6: hot deck single imputation

1997-present: multiple imputation (5 imputations)

1990-6 imputed data

Imputed items have allocation flags which allow identification of imputed data

No simple method available to estimate uncertainty from imputation process

1997-present imputed data

Imputed items have allocation flags which allow identification of imputed data

Can use Rubin's method to estimate uncertainty from imputation process

New 1997-present imputed data

New files contain multiply-imputed values, not just ranges, for family income and personal earnings

Top ~5% of values are top-coded

Already released for 2008, releases for 1997-2007 and 2009 are coming soon

Correct analysis of multiply imputed data

Carry out analysis for each imputation

Combine results of analyses to obtain final result

Incorrect analyses of multiply imputed data

Pick just 1 imputation and do 1 analysis

Take the average of the imputations and do 1 analysis

Combining results of analyses

Can do manually, e.g., by writing a SAS macro program

Can do with software such as SAS PROC MIANALYZE, mitools R package

Can do analysis and combination automatically with software such as mi estimate in Stata, mi_files, mi_count in SUDAAN, etc.

Example: 2006 family income

Pick just 1 imputation (incorrect): $55,583, s.e. $601

Take the average of the imputations and do 1 analysis (incorrect): $55,376, s.e. $599

Correct: $55,376, s.e. $642

Summary

Weights should be used in analyses of NHIS data

Variance estimation requires care, particularly for subdomains

Annual NHIS samples are correlated within a sample design period; not correlated across sample design periods; pooled analyses need to account for correlation/lack of correlation

Analyses of multiply imputed data should follow the standard protocol in order to obtain appropriate estimates and uncertainty estimates

Year-to-year Correlation Reference

Moriarity, C. and Parsons, V.: Year-to-Year Correlation in National Health Interview Survey Estimates, Presented at the 2008 Joint Statistical Meetings

Available online at:

http://www.amstat.org/Sections/Srms/Proceedings/y2008/Files/301235.pdf