National Cancer Institute

Joinpoint Regression Methods for Complex Survey Data

Benmei Liu, National Cancer Institute (NCI)

Coauthors:

Hyune-Ju Kim (Syracuse University)

Joe Zou (IMS)

Eric J. Feuer (NCI, Retired)

Barry I Graubard (NCI, Retired)

Presented at the SEER*Stat Seminar

May 11, 2026

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

Outline

Introduction & research goals

Aggregate-level modeling

Empirical studies

Unit-level modeling

Simulation studies

Empirical studies

Discussion

Introduction

Time trend analyses of complex survey data are widely used by U.S. government agencies to monitor changes in health, environmental, and economic outcomes over time.

Agencies use methods such as pairwise tests, trend contrasts, regression models, and joinpoint regression to assess trends and detect change-points.

NCI's Joinpoint software, widely used for trend analysis of population-based disease surveillance data, has also been considered by several U.S. government agencies for analyzing complex survey data.

Before version 4.9, the software was limited because it relied on aggregated time-specific estimates and did not account for more general correlation across time points beyond the autoregressive AR(1) option.

As a workaround, analysts sometimes used an ad hoc two-step approach, but this could be problematic because joinpoints from aggregated estimates may differ from those based on data of sampled-units.

Research Goals

- Extend joinpoint methods for aggregate outcomes to incorporate the general variance-covariance matrix among time-specific estimates in trend analysis using complex survey data.
- Develop unit-level models that account for both this correlation structure and the design-based degrees of freedom required for valid inference.

Aggregate-level modeling

Aggregate-level Joinpoint Regression Methods

- Let $(t_1, y_1), \dots, (t_m, y_m)$ be the estimates at m time points t for an outcome of interest y. Assuming there are k joinpoints (k is unknown).

  The joinpoint regression model is:

$$g(y_i) = \beta_0 + \beta_1 t_i + \delta_1 (t_i - \tau_1) I_{\tau_1}(t_i) + \cdots + \delta_k (t_i - \tau_k) I_{\tau_k}(t_i) + \boldsymbol{\varepsilon_i^{(k)}} \quad \textbf{(1)}$$

where $g(\cdot)$ is a function, the τ_k's are the unknown joinpoints and $I_{\tau_k}(t_i) = 1$ if $t_i > \tau_k$, $\beta_0, \beta_1, \delta_1, \dots, \delta_k$ are model parameters, $\varepsilon_i^{(k)}$ is the error term, $i = 1, \dots, m$.

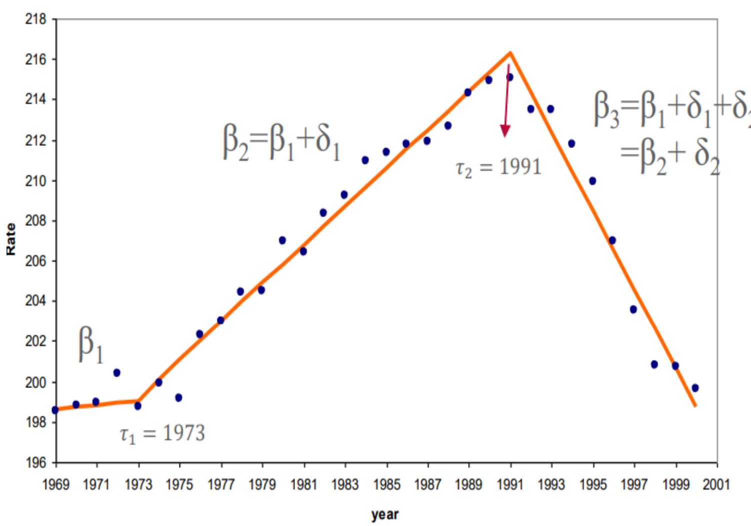
✓ When log-scale $g(\cdot) = log(\cdot)$ is used, the slope of the sth segment corresponding to annual percent change (APC):

$$APC_s = 100 \times \exp(\beta_s - 1) \approx 100\beta_s$$

✓ The APC difference:

$$APC_2 - APC_1 = 100\left(e^{\delta_1 + \beta_1} - e^{\beta_1}\right) \approx 100\delta_1$$

➢ For registry data (e.g., SEER), the error terms are typically assumed uncorrelated or follow first order autocorrelation (AR1).



Joinpoint Regression Model Fitting

- The NCI's Joinpoint software fits the model by least squares or weighted least squares (WLS) up to a specified number of joinpoints
- Top two methods to determine the number and location of joinpoints
  - Permutation test
  - Weighted Bayesian Information Criterion (BIC)

Joinpoint Under Complex Survey Framework

- In complex survey data, time-specific aggregate estimates $y_{1,} \ldots, y_n$ (e.g., totals, means, proportions) may be correlated because of features like panel survey designs or the reuse of the same sampled PSUs across multiple years.

- To account for this dependence, a (full) variance-covariance matrix can be computed across the set of time-specific estimates.

Extension of Joinpoint Regression Method for Correlated Data

Given joinpoints $\tau_1, \dots, \tau_k$. When $\varepsilon_i^{(k)}$ are correlated:

- Estimate the variance-covariance matrix $V = (v_{ij})$, where $v_{ii} = \widehat{Var}(y_i)$ and $v_{ij} = \widehat{Cov}(y_i, y_j\ )$ for $i = 1, .., n; j = 1, \dots, m;\ i \neq j$.

- Estimate $\theta \equiv [\beta_0, \beta_1, \delta_1, \dots, \delta_k]$:

  - Fit the joinpoint model using WLS;

  - The estimator of θ is $\hat{\theta} = (T'WT)^{-1}T'Wy$, $W = V^{-1}$ and T is a $m \times (k+2)$ matrix with ith row equal to $[1, x_i, (x_i - \tau_1)^+, \dots, (x_i - \tau_k)^+]$.

- When we model $\log(y)$, the estimate of variance-covariance matrix $\tilde{V} = (\tilde{v}_{ij})$ can be derived using Taylor series linearization method:

  - $\tilde{v}_{ii} = \frac{v_{ii}}{y_i^2}$; and
  - $\widehat{Cov}\big(\log(y_i)\,, \log\,(y_j)\big) = \frac{v_{ij}}{y_i y_j} = \frac{\widehat{Cov}(y_i, y_j)}{y_i y_j}$,

  for $i = 1, .., m; j = 1, \dots, m;\ i \neq j$. where Cov denotes covariance and v_{ii} is the estimated variance of y_i.

✓ https://surveillance.cancer.gov/help/joinpoint/setting-parameters/input-file-tab/heteroscedastic-errors-option/variance-covariance-matrix

Incorporate Variance-covariance Matrix in Joinpoint

- The full variance-covariance matrix has been incorporated as an option in the NCI's Joinpoint software (version 4.9 and later) for joinpoint model fitting.

  - Standard survey software, such as SAS PROC SURVEY procedures and SUDAAN can be used to produce both the time-specific aggregate estimates and their variance-covariance matrix.

  - The aggregate estimates ($y_{i,}\ i = 1, \dots, n$) and their variance-covariance matrix (V) are then used to fit the joinpoint regression model.

Empirical Studies

Panel Study of Income Dynamics (PSID)

- Panel design directed by University of Michigan
- Data extracted from 1997-2017 with cross-sectional sample weights for each year
- **Outcomes of interest**: unemployment status for age 16+ (Yes/No) and physical health status for age<55 (good/bad)

The National Health Interview Survey (NHIS)

- Cross-sectional household survey conducted by National Center for Health Statistics (NCHS)
- Data extracted from 1991-2016
- **Outcomes of interest**: Body Mass Index (BMI) and obesity status for age 18+ (Yes/No)

The National (Nationwide) Inpatient Sample (NIS)

- Cross-sectional establishment survey administered by Agency for Healthcare Research and Quality (AHRQ)
- Data extracted for 1993-2015 for gastric cancer patients
- **Outcomes of interest**: in-hospital mortality status, total charges ($), total gastrectomy status, and partial gastrectomy status (Yes/No)

Statistical Analysis

Annual estimates for the eight measures in the three surveys, along with the associated standard errors and the variance-covariance matrix of the estimates, obtained using SUDAAN version 11.0.

The point estimates, standard errors, or full variance-covariance matrix for eight outcomes were input into Joinpoint to fit joinpoint models.

Both the permutation method and the weighted BIC method were used to determine the number of joinpoints.

Models were fit both with and without accounting for the covariances between the yearly data.

Example SUDAAN Codes to Compute the Variance-Covariance Matrix

```
proc DESCRIPT data=ind97_17;
NEST ER31996 ER31997;
weight weight;
var unemployed;
class year;
*PRINT nsum mean semean COVMEAN
     /meanfmt=F10.3 semeanfmt=F10.6 COVmeanfmt=F10.6;
OUTPUT nsum mean semean /meanfmt=F10.4 semeanfmt=F10.9 FILENAME =unemployed_Mean9717 FILETYPE = SAS REPLACE;
OUTPUT COVMEAN /COVmeanfmt=F10.9 FILENAME =unemployed_MeanCOV9717 FILETYPE = SAS REPLACE;
run;
```

Stratum & PSU

	VARIABLE	ROWNUM	DDF	B001	B002	B003	B004	B005	B006	B007	B008	B009	B010	B011	B012
1	1	1	70.000	0.000012292	0.000008343	0.000006255	0.000010896	0.000011645	0.000013280	0.000012799	0.000013642	0.000015488	0.000017158	0.000011415	0.000011508
2	1	2	70.000	0.000008343	0.000014623	0.000005983	0.000007675	0.000008249	0.000005786	0.000007540	0.000008202	0.000009477	0.000010910	0.000006873	0.000006117
3	1	3	70.000	0.000006255	0.000005983	0.000016700	0.000006108	0.000003136	0.000006613	0.000005332	0.000004643	0.000005517	0.000006493	0.000003809	0.000005789
4	1	4	70.000	0.000010896	0.000007675	0.000006108	0.000018154	0.000011977	0.000012404	0.000012446	0.000008792	0.000012867	0.000011606	0.000006863	0.000009903
5	1	5	70.000	0.000011645	0.000008249	0.000003136	0.000011977	0.000017677	0.000012186	0.000011545	0.000012686	0.000013770	0.000013954	0.000010972	0.000008801
6	1	6	70.000	0.000013280	0.000005786	0.000006613	0.000012404	0.000012186	0.000021430	0.000013585	0.000013030	0.000018275	0.000016949	0.000010695	0.000011175
7	1	7	70.000	0.000012799	0.000007540	0.000005332	0.000012446	0.000011545	0.000013585	0.000019872	0.000013442	0.000016166	0.000018369	0.000009339	0.000010822
8	1	8	70.000	0.000013642	0.000008202	0.000004643	0.000008792	0.000012686	0.000013030	0.000013442	0.000024860	0.000016532	0.000020568	0.000011994	0.000011588
9	1	9	70.000	0.000015488	0.000009477	0.000005517	0.000012867	0.000013770	0.000018275	0.000016166	0.000016532	0.000024749	0.000021914	0.000013701	0.000013082
10	1	10	70.000	0.000017158	0.000010910	0.000006493	0.000011606	0.000013954	0.000016949	0.000018369	0.000020568	0.000021914	0.000030937	0.000017963	0.000014917
11	1	11	70.000	0.000011415	0.000006873	0.000003809	0.000006863	0.000010972	0.000010695	0.000009339	0.000011994	0.000013701	0.000017963	0.000021692	0.000008282
12	1	12	70.000	0.000011508	0.000006117	0.000005789	0.000009903	0.000008801	0.000011175	0.000010822	0.000011588	0.000013082	0.000014917	0.000008282	0.000023009
13	1	1	70.000	0.079299419	0.061274672	0.058401298	0.061017803	0.074815378	0.059245388	0.067200480	0.107897873	0.110570267	0.096884882	0.084713257	0.079120582
14	1	1	70.000	0.000000579	.	.	.	.	.	.	.	.	.	.	.
15	1	2	70.000	.	0.000006218	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
16	1	3	70.000	.	0.000000000	0.000005586	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
17	1	4	70.000	.	0.000000000	0.000000000	0.000005456	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
18	1	5	70.000	.	0.000000000	0.000000000	0.000000000	0.000006163	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
19	1	6	70.000	.	0.000000000	0.000000000	0.000000000	0.000000000	0.000004846	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
20	1	7	70.000	.	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000005305	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
21	1	8	70.000	.	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000007906	0.000000000	0.000000000	0.000000000	0.000000000
22	1	9	70.000	.	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000008037	0.000000000	0.000000000	0.000000000
23	1	10	70.000	.	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000007089	0.000000000	0.000000000
24	1	11	70.000	.	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000006347	0.000000000
25	1	12	70.000	.	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000005580
26	1	1	70.000	126185.000000000	9251.000000000	9846.000000000	10502.000000000	11233.000000000	11502.000000000	11818.000000000	12176.000000000	12237.000000000	12343.000000000	12218.000000000	13059.000000000
27	1	1	70.000	1632753734.000000	125049482.0000000	131012105.0000000	134955646.0000000	143830055.0000000	150074535.0000000	153970016.0000000	155678797.0000000	156909579.0000000	157357564.0000000	162420695.0000000	161495260.0000000

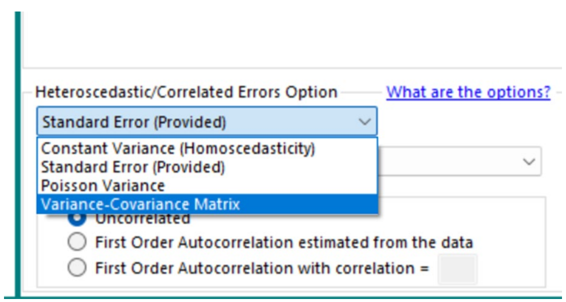
Example SAS Codes to Compute the Variance-Covariance Matrix

```
proc surveymeans data=ind97_17;
strata ER31996;
cluster ER31997;
weight weight;
var unemployed;
domain year/ cov;
ods output domain=MyStat_unemployed DomainMeanCov=DomainMeanCov_Unemployed;
run;
```

	Levels	DomainLabel	year	VarName	Cov1	Cov2	Cov3	Cov4	Cov5	Cov6	Cov7	Cov8	Cov9	Cov10	Cov11
1	Cov1	year	1997	unemployed	0.000015	5.983E-6	7.675E-6	8.249E-6	5.786E-6	7.54E-6	8.202E-6	9.477E-6	0.000011	6.873E-6	6.117E-6
2	Cov2	year	1999	unemployed	5.983E-6	0.000017	6.108E-6	3.136E-6	6.613E-6	5.332E-6	4.643E-6	5.517E-6	6.493E-6	3.809E-6	5.789E-6
3	Cov3	year	2001	unemployed	7.675E-6	6.108E-6	0.000018	0.000012	0.000012	0.000012	8.792E-6	0.000013	0.000012	6.863E-6	9.903E-6
4	Cov4	year	2003	unemployed	8.249E-6	3.136E-6	0.000012	0.000018	0.000012	0.000012	0.000013	0.000014	0.000014	0.000011	8.801E-6
5	Cov5	year	2005	unemployed	5.786E-6	6.613E-6	0.000012	0.000012	0.000021	0.000014	0.000013	0.000018	0.000017	0.000011	0.000011
6	Cov6	year	2007	unemployed	7.54E-6	5.332E-6	0.000012	0.000012	0.000014	0.000020	0.000013	0.000016	0.000018	9.339E-6	0.000011
7	Cov7	year	2009	unemployed	8.202E-6	4.643E-6	8.792E-6	0.000013	0.000013	0.000013	0.000025	0.000017	0.000021	0.000012	0.000012
8	Cov8	year	2011	unemployed	9.477E-6	5.517E-6	0.000013	0.000014	0.000018	0.000016	0.000017	0.000025	0.000022	0.000014	0.000013
9	Cov9	year	2013	unemployed	0.000011	6.493E-6	0.000012	0.000014	0.000017	0.000018	0.000021	0.000022	0.000031	0.000018	0.000015
10	Cov10	year	2015	unemployed	6.873E-6	3.809E-6	6.863E-6	0.000011	0.000011	9.339E-6	0.000012	0.000014	0.000018	0.000022	8.282E-6
11	Cov11	year	2017	unemployed	6.117E-6	5.789E-6	9.903E-6	8.801E-6	0.000011	0.000011	0.000012	0.000013	0.000015	8.282E-6	0.000023

!! There may be insufficient memory issue if the dimensions of the matrix are too large. Switch to SUDAAN if this issue occurs.

Illustration of the Variance-Covariance Matrix Option in Joinpoint (version 4.9+)



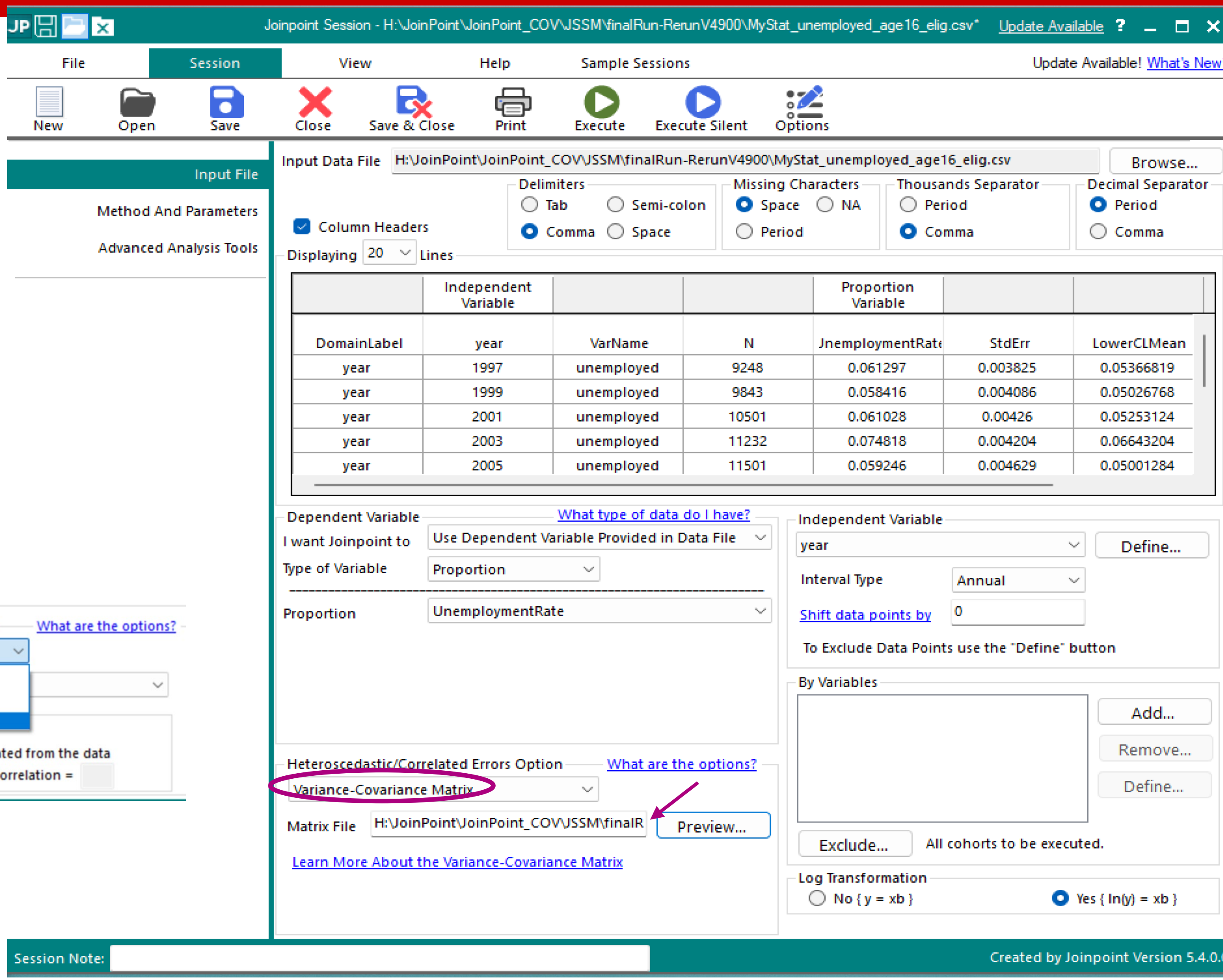


Illustration of the Joinpoint Variance-Covariance Matrix Option

Variance-Covariance Matrix Preview

Matrix File: H:\JoinPoint\JoinPoint_COV\JSSM\finalRun-RerunV4900\DomainMeanCov_Unemployed_age16_eligMatrix.csv

Browse...

Delimiters: Tab / Semi-colon / Comma (selected) / Space

Number of Observations: 11
Number of Rows: 11
Number of Columns: 11

	1997	1999	2001	2003	2005	2007	2009	2011	2013	2015	2017
1997	0.000015	5.98E-06	7.67E-06	8.24E-06	5.77E-06	7.53E-06	8.19E-06	9.47E-06	0.000011	6.87E-06	6.11E-06
1999	5.98E-06	0.000017	6.10E-06	3.14E-06	6.61E-06	5.33E-06	4.66E-06	5.52E-06	6.49E-06	3.81E-06	5.78E-06
2001	7.67E-06	6.10E-06	0.000018	0.000012	0.000012	0.000012	8.79E-06	0.000013	0.000012	6.86E-06	9.91E-06
2003	8.24E-06	3.14E-06	0.000012	0.000018	0.000012	0.000012	0.000013	0.000014	0.000014	0.000011	8.81E-06
2005	5.77E-06	6.61E-06	0.000012	0.000012	0.000021	0.000014	0.000013	0.000018	0.000017	0.000011	0.000011
2007	7.53E-06	5.33E-06	0.000012	0.000012	0.000014	0.00002	0.000013	0.000016	0.000018	9.34E-06	0.000011
2009	8.19E-06	4.66E-06	8.79E-06	0.000013	0.000013	0.000013	0.000025	0.000017	0.000021	0.000012	0.000012
2011	9.47E-06	5.52E-06	0.000013	0.000014	0.000018	0.000016	0.000017	0.000025	0.000022	0.000014	0.000013
2013	0.000011	6.49E-06	0.000012	0.000014	0.000017	0.000018	0.000021	0.000022	0.000031	0.000018	0.000015
2015	6.87E-06	3.81E-06	6.86E-06	0.000011	0.000011	9.34E-06	0.000012	0.000014	0.000018	0.000022	8.28E-06
2017	6.11E-06	5.78E-06	9.91E-06	8.81E-06	0.000011	0.000011	0.000012	0.000013	0.000015	8.28E-06	0.000023

OK | Cancel | Help

‼ No row and column names are allowed in the input matrix file.

Snapshot of the variance-covariance matrix for mean BMI from NHIS 1991-2016

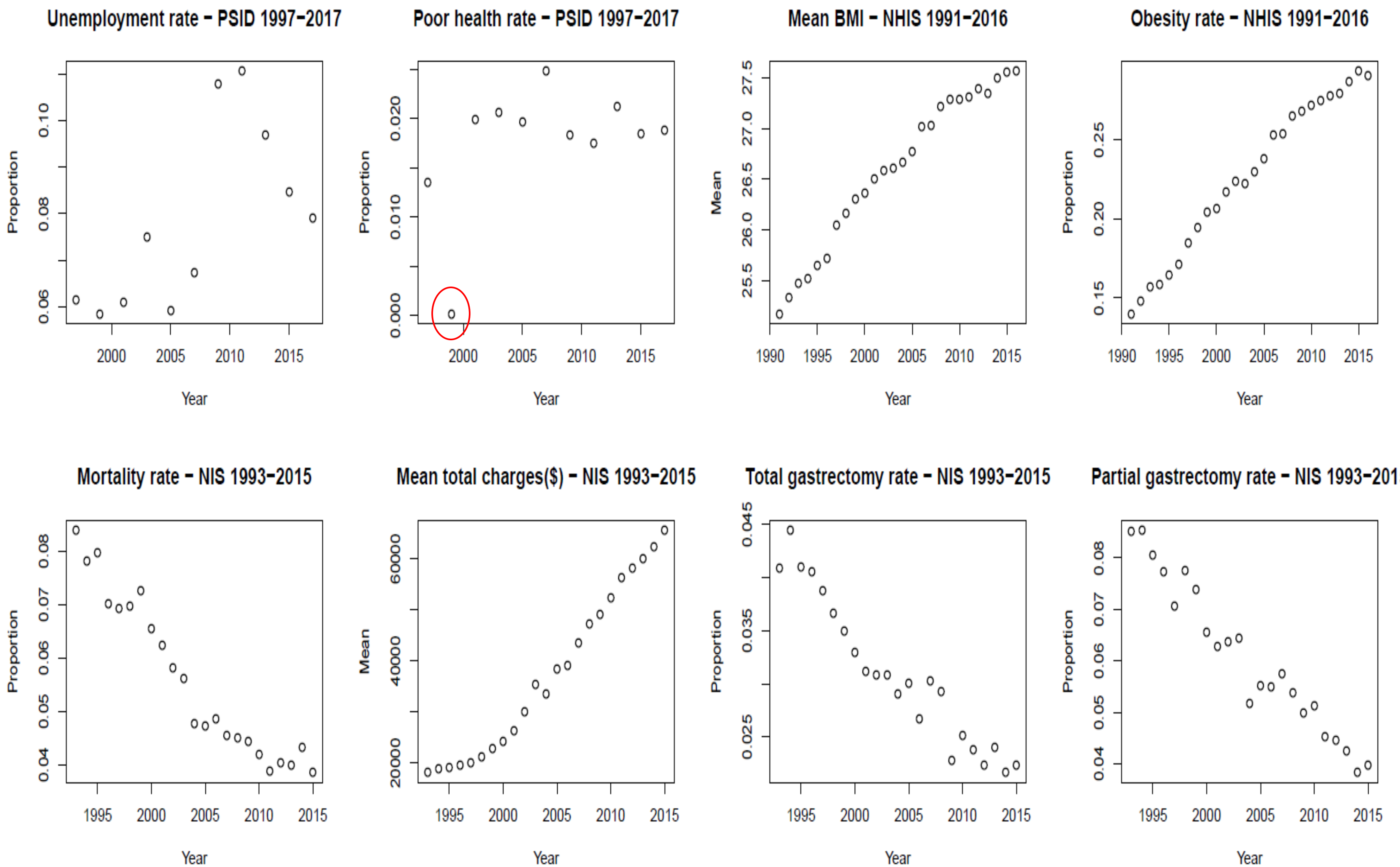
	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
1991	0.00068	0.00029	0.00028	0.00032	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1992	0.00029	0.00068	0.00026	0.00034	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1993	0.00028	0.00026	0.00072	0.00031	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1994	0.00032	0.00034	0.00031	0.00082	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1995	0	0	0	0	0.00077	0.00021	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1996	0	0	0	0	0.00021	0.00165	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1997	0	0	0	0	0	0	0.0011	0.00014	0.0001	9.9E-05	9.9E-05	0.00023	0.00018	0.00021	-2.7E-05	0	0	0	0	0	0	0	0	0	0	0
1998	0	0	0	0	0	0	0.00014	0.00098	0.00017	9.6E-06	9.4E-05	0.0001	0.00025	0.00014	0.00012	0	0	0	0	0	0	0	0	0	0	0
1999	0	0	0	0	0	0	0.0001	0.00017	0.00132	0.00018	0.00011	5.8E-05	5.3E-05	0.00011	0.00014	0	0	0	0	0	0	0	0	0	0	0
2000	0	0	0	0	0	0	9.9E-05	9.6E-06	0.00018	0.00111	0.00013	0.00011	9.6E-05	-3.9E-05	2.7E-05	0	0	0	0	0	0	0	0	0	0	0
2001	0	0	0	0	0	0	9.9E-05	9.4E-05	0.00011	0.00013	0.00118	7E-05	0.00025	7.8E-05	0.0001	0	0	0	0	0	0	0	0	0	0	0
2002	0	0	0	0	0	0	0.00023	0.0001	5.8E-05	0.00011	7E-05	0.00151	0.00027	0.00015	0.00021	0	0	0	0	0	0	0	0	0	0	0
2003	0	0	0	0	0	0	0.00018	0.00025	5.3E-05	9.6E-05	0.00025	0.00027	0.00161	0.00026	0.00025	0	0	0	0	0	0	0	0	0	0	0
2004	0	0	0	0	0	0	0.00021	0.00014	0.00011	-3.9E-05	7.8E-05	0.00015	0.00026	0.00146	5.8E-05	0	0	0	0	0	0	0	0	0	0	0
2005	0	0	0	0	0	0	-2.7E-05	0.00012	0.00014	2.7E-05	0.0001	0.00021	0.00025	5.8E-05	0.00158	0	0	0	0	0	0	0	0	0	0	0
2006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.00247	0.00047	0.00038	0.00015	0.00025	0.00033	0.00017	0.00016	0.0001	0.0002	0
2007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.00047	0.00197	0.00013	0.00024	0.0003	0.00019	0.00017	0.00012	5E-05	0.00022	0
2008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.00038	0.00013	0.00234	0.00043	0.00012	0.00015	5.5E-05	6.3E-05	0.00034	0.00016	0
2009	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.00015	0.00024	0.00043	0.00217	0.00021	0.00037	0.00028	-4.2E-05	0.00028	0.00018	0
2010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.00025	0.0003	0.00012	0.00021	0.00179	0.00019	0.00037	-5.8E-05	0.0001	0.00025	0
2011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.00033	0.00019	0.00015	0.00037	0.00019	0.00178	0.00029	0.00022	0.00034	0.00037	0
2012	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.00017	0.00017	5.5E-05	0.00028	0.00037	0.00029	0.00177	0.00021	0.00015	0.00017	0
2013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.00016	0.00012	6.3E-05	-4.2E-05	-5.8E-05	0.00022	0.00021	0.00197	0.00019	0.00029	0
2014	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0001	5E-05	0.00034	0.00028	0.0001	0.00034	0.00015	0.00019	0.00222	0.00031	0
2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0002	0.00022	0.00016	0.00018	0.00025	0.00037	0.00017	0.00029	0.00031	0.00197	0
2016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0024

Table 1: The mean, median and range of the Correlation Coefficients* between annual cross-sectional estimates for eight outcome variables

Variables (data source)	mean	min	median	max
Unemployment rate age 16+ (PSID 1997-2017)	0.504	0.180	0.510	0.790
Poor health rate age <55 (PSID 1997-2017)	0.185	-0.172	0.152	0.747
Mean BMI age 18+ (NHIS 1991-2016)	0.033	-0.031	0.000	0.461
Obesity rate age 18+ (NHIS 1991-2016)	0.024	-0.087	0.000	0.390
In-hospital mortality rate (NIS 1993-2015)	0.021	-0.018	0.006	0.292
Mean total charges $ (NIS 1993-2015)	0.068	-0.04	0.011	0.681
Total gastrectomy rate (NIS 1993-2015)	0.043	-0.027	0.002	0.531
Partial gastrectomy rate (NIS 1993-2015)	0.034	-0.03	0.002	0.789

$$* \rho_{i,j} = \frac{cov(y_i, y_j)}{\sqrt{var(y_i)var(y_j)}}$$

Scatter plots of the annual point estimates of eight outcomes from PSID, NHIS and NIS

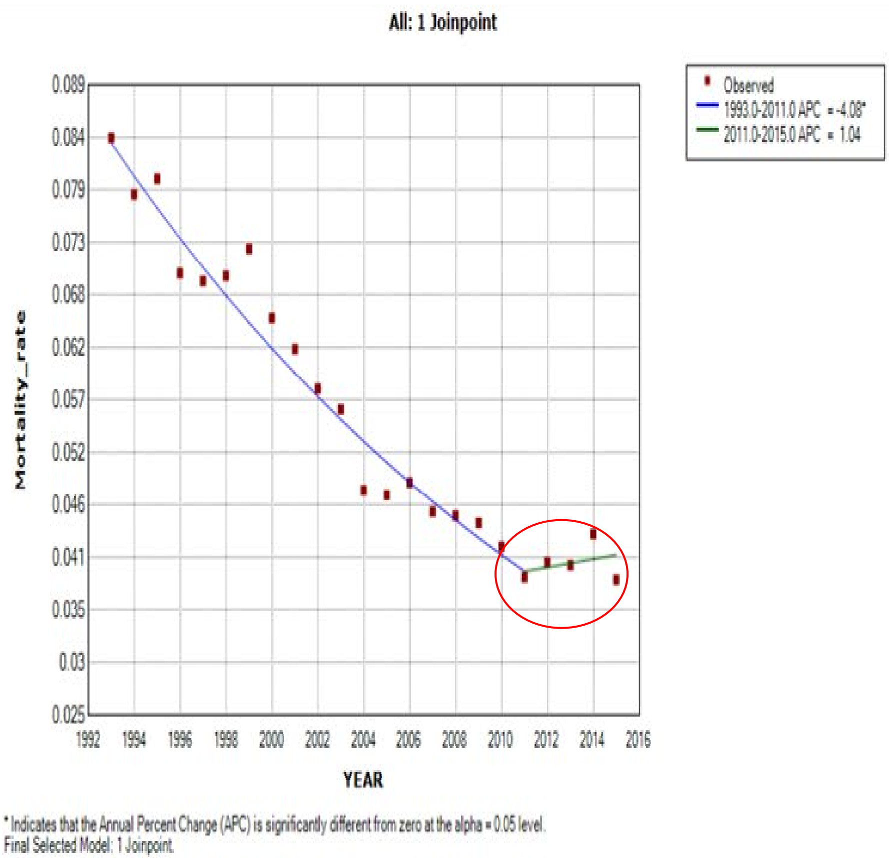


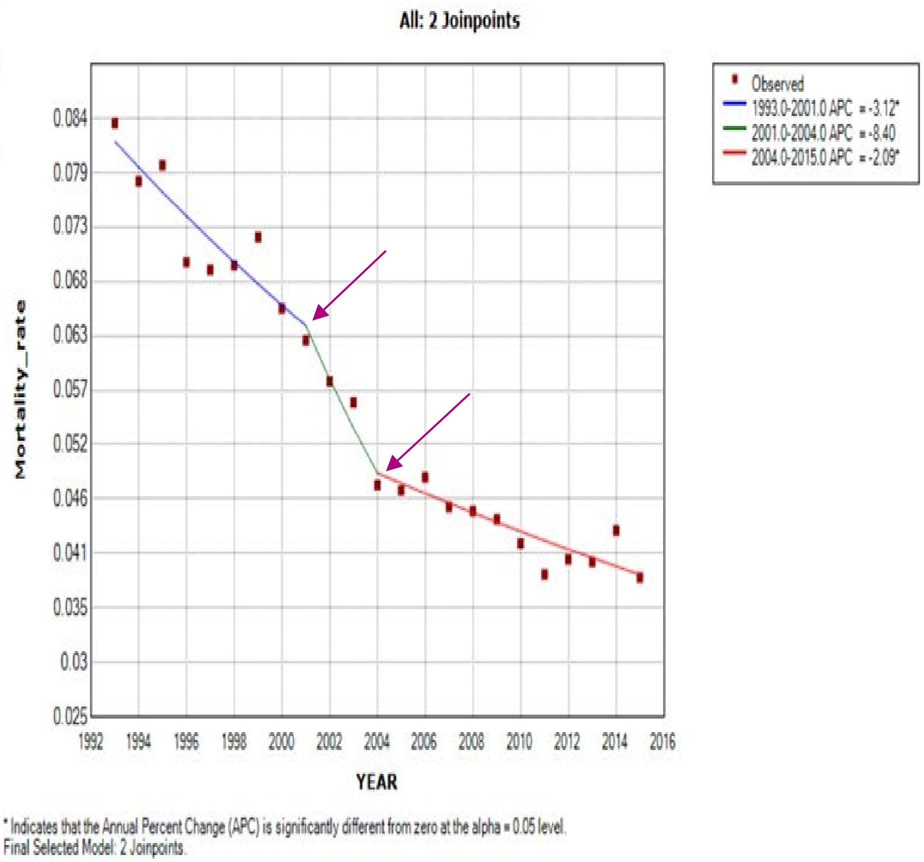
Number of joinpoints found

Variables (data source)	Permutation		Weighted BIC	
	diagonal standard errors only	**with variance-covariance matrix**	diagonal standard errors only	**with variance-covariance matrix**
Unemployment rate aged 16+ (PSID 1997-2017)	1 joinpoint (2011)	1 joinpoint (2011)	1 joinpoint (2011)	1 joinpoint (2011)
Poor health rate age <55 (PSID 1997-2017)	0 joinpoint	0 joinpoint	0 joinpoint	1 joinpoint (2007)
Poor health rate age <55(PSID 1997-2017, excluding data point 1999)	0 joinpoint	0 joinpoint	1 joinpoint (2003)	1 joinpoint (2007)
Mean BMI age 18+ (NHIS 1991-2016)	2 joinpoints (1999, 2009)	2 joinpoints (1999, 2009)	2 joinpoints (1999, 2009)	2 joinpoints (1999, 2009)
Obesity rate age 18+ (NHIS 1991-2016)	2 joinpoints (1999, 2008)	2 joinpoints (1999, 2008)	2 joinpoints (1999, 2008)	2 joinpoints (1999, 2008)
In-hospital mortality rate (NIS 1993-2015)	1 joinpoint (2011)	1 joinpoint (2011)	1 joinpoint (2011)	2 joinpoint (2011, 2004)
Mean total charges $ (NIS 1993-2015)	2 joinpoints (1998, 2008)	3 joinpoints (1999, 2003, 2011)	2 joinpoints (1998, 2008)	3 joinpoints (1999, 2003, 2011)
Total gastrectomy rate (NIS 1993-2015)	0 joinpoint	0 joinpoint	0 joinpoint	0 joinpoint
Partial gastrectomy rate (NIS 1993-2015)	0 joinpoint	0 joinpoint	0 joinpoint	0 joinpoint

Trends of in-hospital mortality rates among gastric cancer patients (NIS 1993-2015)

Weighted BIC method: left panel without and right panel with variance-covariance matrix





DISCUSSION

The empirical studies demonstrate the necessity of accounting for the correlated errors, through incorporating the full variance-covariance matrix of time-specific estimates into the joinpoint regression model fitting.

The models considered are at aggregate level, the results may differ from results computed by fitting models at the unit level.

For technical details, see Liu et al (2022):

Liu B, Kim H-J, Feuer E.J., Graubard B.I. (2022). Joinpoint regression methods of aggregate outcomes for complex survey data. *Journal of Survey Statistics and Methodology*, 11(4), 967-989. https://doi.org/10.1093/jssam/smac014.

Unit-level modeling

Motivations

When modeling aggregated data, each time point has only a single value, so the degrees of freedom for variance estimation in joinpoint analysis are based on the number of time points rather than the typically larger number of sampled PSUs used in individual-level survey analyses.

The smaller degrees of freedom (based on time points) can limit the statistical power.

Unit-level Joinpoint Regression Methods

- We consider the following model with k (≥ 0) joinpoints:

  - Log-normal joinpoint regression model for continuous y:
  $$log\left(y_{ijl}^{(t)}\right) = \beta_0 + \beta_1 t + \delta_1(t-\tau_1)I_{\tau_1}(t) + \cdots + \delta_k(t-\tau_k)I_{\tau_k}(t) + \varepsilon_{ijl}^{(t)} \quad \textbf{(2)}$$

  - Logistic joinpoint regression model for binary y:
  $$logit\left(p\left(y_{ijl}^{(t)} = 1\right)\right) = \beta_0 + \beta_1 t + \delta_1(t-\tau_1)I_{\tau_1}(t) + \cdots + \delta_k(t-\tau_k)I_{\tau_k}(t), \quad \textbf{(3)}$$

  where $y_{ijl}^{(t)}$ and $\varepsilon_{ijl}^{(t)}$ denote the response and error term for unit l in stratum i and PSU j at time point t (t=$1, \ldots, T$), $\tau_k{}'s$ are the unknown joinpoints and $I_{\tau_k}(t) = 1$ if $t > \tau_k$.

- **We are interested in:**
  - How many change-points there are?
  - Where is/are the change-point(s)?
  - How variable is the estimate?

- Challenge: Need to find an appropriate model selection method to determine k.

Review: Lumley & Scott (2015) modified AIC

For Statisticians

- Lumley & Scott (2015) proposed the following modification of the classical Akaike information criterion (AIC) for regression with complex survey data:

$$dAIC = -2n\hat{l}(\hat{\theta}) + 2p\hat{\hat{\delta}},$$

where $\hat{l}(\hat{\theta}) = \frac{1}{N}\sum_{i\in S} w_i l_i(\hat{\theta}) = \frac{1}{N}\sum_{i\in S} w_i log f_{\hat{\theta}}(y_i|x_i)$, n is the sample size, N is the population size, $\sum_{i=1}^{n} w_i = N, p$ is the # of covariates, $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)'$ and

$$\hat{\hat{\delta}} = trace\{\hat{J}(\hat{\theta})^{-1}\hat{V}(\hat{\theta})\},$$ sample design effect

$\hat{J}(\hat{\theta})$ and $\hat{V}(\hat{\theta})$ are the variance-covariance matrix associated with $\hat{\theta}$ assuming SRS and the complex design, respectively. The first row and column of the matrices corresponding to $\hat{\beta}_0$ removed in this calculation.

Grid search to find the best jp$k$ model for given k (>0)

- Do a grid search to find the best jp$k$ model across models with different joinpoint locations $\tau_1, \ldots, \tau_k$ for each given k ($k = 1,2, \ldots, K$).
- Follow the default rule:
  - Minimum three time points from a joinpoint to either end of the data (including the joinpoint);
  - Minimum four time points between two joinpoints (including the join points)
- Use appropriate survey procedure to fit each model and obtain
  - adjusted R-square (for continuous data) or
  - $-2logL$ (for binary data).
- Select the best jp$k$ model with the largest adjusted R-square or smallest $-2logL$ rule.

➢ **Question:** Which model (jp0, jp1, jp2,…, jp$K$) fits the best?

Modified dAIC to select the best joinpoint model

- Let $\hat{\tau} = (\hat{\tau}_1, \dots, \hat{\tau}_k)$ denote the k estimated joinpoints for a given k ($k = 1,2,\dots,K$) from the grid search. To find which jp$k$ ($k \geq 0$) model fits the best, we use the following **modified $\boldsymbol{dAIC}$**:
  - For the log-normal joinpoint model (2):

$$m.dAIC = n * MSE + 2(2k+1)\hat{\hat{\delta}},$$

  - For the logistic jointpoint model (3):

$$m.dAIC = \frac{n}{N} * (-2logL) + 2(2k+1)\hat{\hat{\delta}},$$

where $\hat{\hat{\delta}} = trace\{\hat{J}(\hat{\theta})^{-1}\hat{V}(\hat{\theta})\}, \hat{\theta} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\delta}_1, \dots, \hat{\delta}_k\right)'$.

- Estimation of MSE and $-2logL$
  - Run SAS PROC SURVEYREG to get square root of MSE for the log-normal model; Run PROC SURVEYLOGISTIC to get $-2logL$.
  - Use the R Survey package to compute both
- Estimation of $\hat{\hat{\delta}}$:
  - Constrained model approach: Assume the mean functions are continuous at $\hat{\tau}_S$.
  - Unconstrained model approach: Not assume that the mean functions are continuous at $\hat{\tau}_S$.

Computation of Annual Percent Change (APC)

- For a continuous outcome under the log-normal joinpoint model (2), the model for the sth segment, $s = 1, \ldots, k + 1$, is assumed as:

$$\log\left(y_{ijl}^{(t)}\right) = \beta_{s,0} + \beta_{s,1}t + \varepsilon_{ijl}^{(t)},$$

where $\beta_{s,0}$ and $\beta_{s,1}$ denote the intercept and slope of the segment s, respectively.

- APC in segment *S* (APC_s) measures the annual percent change in the sample-weighted **mean outcome** (denoted as $\bar{y}$) from year (t) to year (t+1):

$$APC_s = 100\left[\hat{\bar{y}}^{(t+1)} - \hat{\bar{y}}^{(t)}\right]/\hat{\bar{y}}^{(t)} = 100\left(e^{\hat{\beta}_{s,1}} - 1\right),$$

where $\hat{\beta}_{s,1}$ is the estimate of the slope ($\beta_{s,1}$) in segment s.

- AAPC is defined as average APC over a given time period.

Computation of Odds Annual Percent Change (OAPC)

- For a binary outcome under the logistic joinpoint model (3), let $Odds_{y_{ijl}^{(t)}=1} = prob\left(y_{ijl}^{(t)}=1\right)/\left(1-prob\left(y_{ijl}^{(t)}=1\right)\right)$, the model for the sth segment, $s=1,\ldots,k+1$, is assumed as:

$$\log\left(Odds_{y_{ijl}^{(t)}=1}\right)=\beta_{s,0}+\beta_{s,1}t,$$

where $\beta_{s,0}$ and $\beta_{s,1}$ denote the intercept and slope of the segment s, respectively.

- OAPC in segment *S* ($OAPC_s$) measures the annual percent change in the sample-weighted **mean predicted odds** from year (t) to year (t+1):

$$OAPC_s = 100\,\frac{\sum_i\sum_j\sum_l w_{ijl}^{(t+1)}\,\widehat{Odds}_{y_{ijl}^{(t+1)}=1}/\sum_i\sum_j\sum_l w_{ijl}^{(t+1)} - \sum_i\sum_j\sum_l w_{ijl}^{(t)}\,\widehat{Odds}_{y_{ijl}^{(t)}=1}/\sum_i\sum_j\sum_l w_{ijl}^{(t)}}{\sum_i\sum_j\sum_l w_{ijl}^{(t)}\,\widehat{Odds}_{y_{ijl}^{(t)}=1}/\sum_i\sum_j\sum_l w_{ijl}^{(t)}}$$

$$=100\left(e^{\widehat{\beta}_{s,1}}-1\right),$$

where $\hat{\beta}_{s,1}$ is the estimate of the slope ($\beta_{s,1}$) in segment s.

- OAAPC is defined as average OAPC over a given time period.

Simulation Study: Data Generation

- **Objective:** Generate 20 years of data (t=1997-2016) with:
  - Total Population per year: 1 million;
  - # of clusters per year: 1000;
  - # of observations per cluster per year: 1000
  - # of sampled clusters per year: 80
  - Sampling rate within each cluster (**5%, 20%, 50%**)
  - # of simulations: 100
- **Data generation steps:**
  - Generate yearly population means/proportions using:

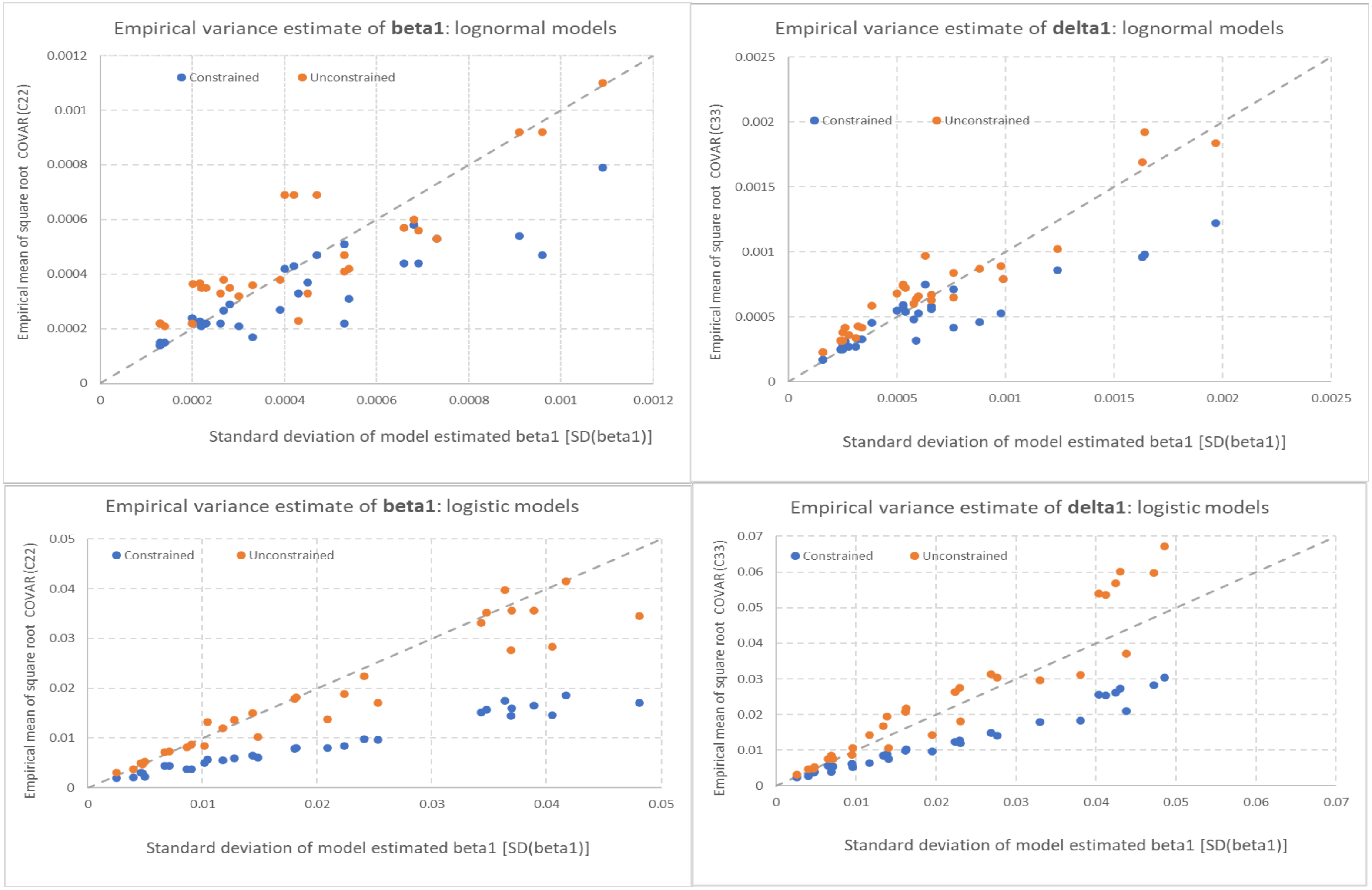
$f(\mu_t) = \beta_0 + \beta_1 t + \delta_1 (t - \tau_1) I_{\tau_1}(t)$, where t=1997-2016, $\tau_1 = 2003$, $f(\cdot) = \log(\cdot)$ or $\text{logit}(\cdot)$, $\beta_0 = -6.57705$; $\beta_1 = 0.00492$, $\delta_1 = (\mathbf{0, -0.0041, -0.041})$.

  - Generate clusters with $ICC = (\mathbf{0, 0.01, 0.075})$
  - Generate individual records using normal distribution for continuous data and Beta-Bernoulli for binary data.
  - Sample clusters: Use the same set of sampled clusters for years 1997-2006 and another set of sampled clusters for years 2008-2016.
  - Sampling within sampled clusters: n=(4000, 16,000, 40,000) per year.

Simulation Study: Analytic goals

- Apply the unit-level models and the aggregate-level models with & without variance-covariance matrix to compute the following statistics for each scenario:
  - Compare how many times # of joinpoints were identified using unit-level model with & without constrains for dAIC and those from the aggregate-level models
  - Check whether the $\widehat{\boldsymbol{V}}(\hat{\theta})$ from the **constrained** model or the $\widehat{\widehat{\boldsymbol{V}}}$ from the **unconstrained** model is a better estimator of the true covariance matrix

Figure 1: Variance estimation of beta1 & delta1 from the individual-level models



X-axis: Empirical square root simulation error;

Y-axis: Empirical mean square root of variances for the constrained and unconstrained models corresponding to the diagonal terms in the variance-covariance matrices

Simulation study: Results

Table 1: Percent (%) of samples correctly identified true # of joinpoint(s) for a continuous outcome: log-normal model

sampling rate	APC difference	ICC=0				ICC=0.01				ICC=0.075			
		Unit-level model		Aggregate-level model		Unit-level model		Aggregate-level model		Unit-level model		Aggregate-level model	
		constrained	unconstrained	without COV	with COV	constrained	unconstrained	without COV	with COV	constrained	unconstrained	without COV	with COV
5% (n=4000)	0.00%	90	76	***93***	89	***97***	30	94	88	***100***	0	83	89
	-0.41%	90	67	***94***	***94***	***95***	36	91	***95***	***95***	28	85	93
	-4.10%	93	78	***97***	96	96	51	94	***97***	***98***	21	83	96
20%	0.00%	86	77	***91***	88	***100***	0	86	90	***100***	0	53	87
	-0.41%	89	59	***94***	93	***98***	22	84	95	***99***	22	58	94
	-4.10%	90	71	95	***96***	***99***	22	91	96	***100***	14	61	96
50%	0.00%	87	78	***91***	89	***100***	0	72	88	***100***	0	40	88
	-0.41%	87	69	***94***	***94***	***97***	20	71	91	***99***	14	51	92
	-4.10%	89	72	***95***	***95***	***99***	23	73	92	***98***	10	73	92
unif 5%~15%	-4.10%	91	90	94	***100***	96	45	91	***100***	97	16	75	***100***

Simulation study: Results

NATIONAL CANCER INSTITUTE

Table 2: Percent (%) of replicates correctly identified true # of joinpoint(s) for a binary outcome: logistic joinpoint model

sampling rate	APC difference	ICC=0				ICC=0.01				ICC=0.075			
		Unit-level model		Aggregate-level model		Unit-level model		Aggregate-level model		Unit-level model		Aggregate-level model	
		constrained	unconstrained	without COV	with COV	constrained	unconstrained	without COV	with COV	constrained	unconstrained	without COV	with COV
5% (n=4000)	0	69	***73***	72	71	***74***	***74***	71	71	***76***	74	*75*	*75*
	-0.485%	38	35	18	21	36	***38***	26	24	32	35	32	31
	-4.85%	***88***	87	72	69	***86***	82	73	74	***85***	83	83	83
20%	0	31	24	74	75	***64***	60	82	84	71	70	75	***77***
	-0.485%	***86***	76	42	38	60	***62***	35	34	37	39	33	37
	-4.85%	***90***	89	74	73	***85***	82	76	73	***93***	88	92	87
50%	0	3	7	73	76	55	53	***84***	***84***	72	70	**73**	69
	-0.485%	***91***	84	57	58	***70***	***70***	34	33	37	**38**	34	36
	-4.85%	***91***	90	76	74	91	***92***	84	78	**92**	88	91	88
uniform 5%~15%	-4.85%	93	***100***	74	73	92	***100***	72	78	86	80	81	***87***

Empirical Study

- Data source:

  The National Health Interview Survey 1991-2016

- Outcomes of interest:
  - Body mass Index (BMI)
  - Obesity (=1 if BMI>=30; =0 otherwise)

- Analyses conducted
  - Fit unit-level joinpoint models (using both constrained and unconstrained models)
  - Fit aggregate-level joinpoint models (with and without incorporating the variance-covariance matrix of the parameters

Empirical study: Results

Table 3: The final joinpoints identified and the APC (%) and AAPC (%) results from the log-normal models for mean BMI

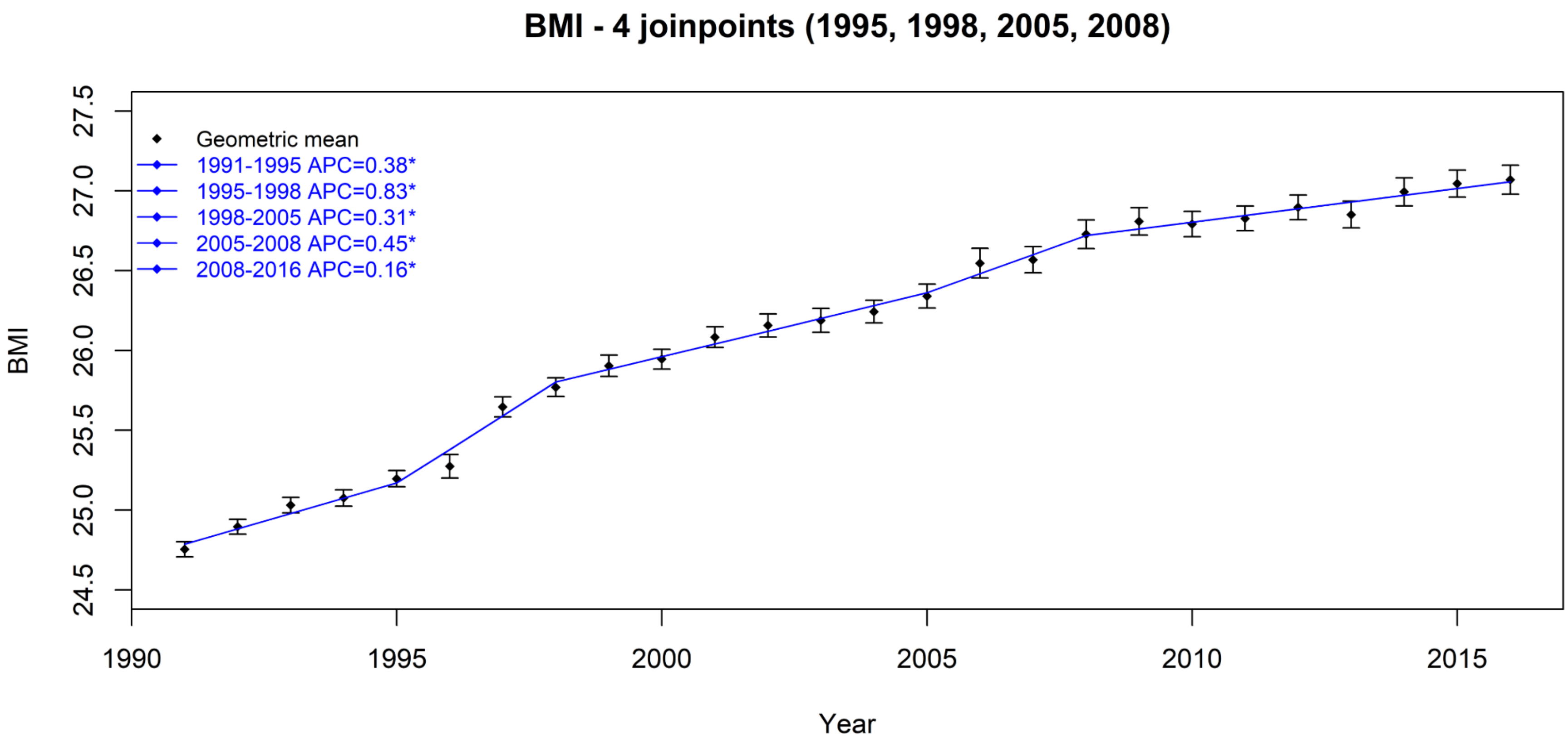
	Aggregate-level model				Individual-level model*	
	without COV		with COV			
Joinpoints	1999, 2009		1999, 2009		1995, 1998, 2005, 2008	
Segment	Range	APC (95% CI)	Range	APC (95% CI)	Range	APC (95% CI)
1	1991-1999	0.54 (0.48,0.60)	1991-1999	0.54 (0.48,0.59)	1991-1995	0.38 (0.32, 0.45)
2	1999-2009	0.37 (0.30,0.43)	1999-2009	0.36 (0.29,0.42)	1995-1998	0.83 (0.45, 1.21)
3	2009-2016	0.17 (0.07, 0.27)	2009-2016	0.18 (0.08, 0.28)	1998-2005	0.31 (0.24, 0.37)
4					2005-2008	0.45 (0.03, 0.87)
5					2008-2016	0.16 (0.11, 0.21)
AAPC						
whole range	1991-2016	0.37 (0.33, 0.41)	1991-2016	0.37 (0.33, 0.41)	1991-2016	0.35 (0.28, 0.42)
Range corresponding to the aggregate-level model					1991-1999	0.54 (0.39, 0.69)
					1999-2009	0.33 (0.20, 0.47)
					2009-2016	0.16 (0.11, 0.21)

Empirical study: Results (cont'd)

Table 4: The final joinpoints identified and the OAPC (%) and OAAPC (%) results for obesity: aggregate-level log-normal models and individual-level logistic models

Joinpoints	**Aggregate level**				**Individual-level***	
	without COV		with COV			
	1999, 2008		1999, 2008		1999, 2008	
Segment	Range	APC (95% CI)	Range	APC (95% CI)	Range	OAPC (95% CI)
1	1991-1999	4.62 (4.13,5.11)	1991-1999	4.65 (4.17, 5.14)	1991-1999	5.63 (5.02, 6.20)
2	1999-2008	3.03 (2.49, 3.58)	1999-2008	3.00 (2.45, 3.45)	1999-2008	3.99 (3.35, 4.64)
3	2008-2016	1.35 (0.85, 1.85)	2008-2016	1.37 (0.87, 1.88)	2008-2016	1.86 (1.28, 2.45)
	AAPC				**OAAPC**	
Whole range	1991-2016	2.99 (2.72, 3.27)	1991-2016	3.00 (2.72, 3.28)	1991-2016	3.82 (3.47, 4.17)

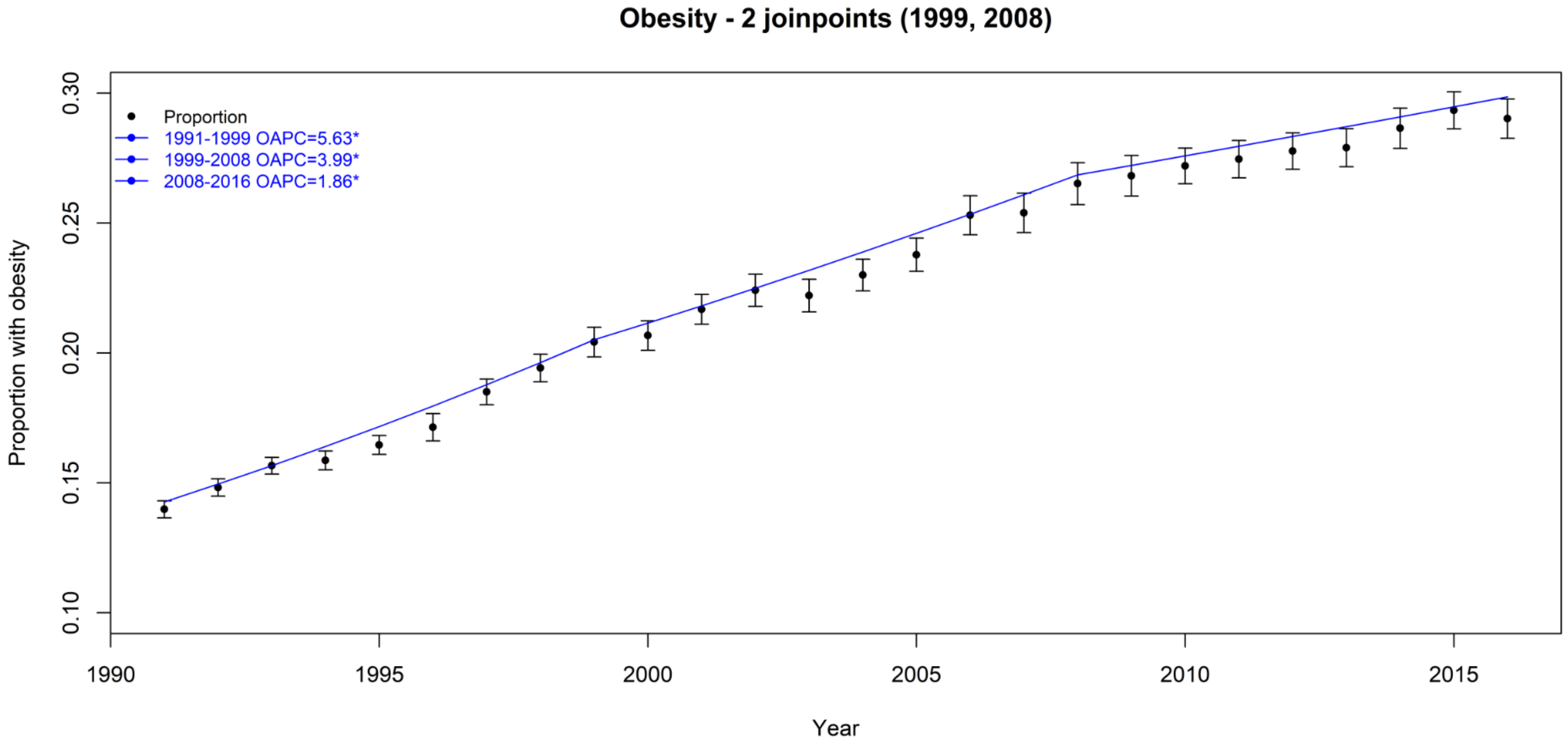
Figure 2: Plot of the annual geometric mean[1] along with the 95% confidence intervals and the fitted joinpoint line from individual-level log-normal model for BMI chosen from the constrained approach



*APC is statistically significant, p-value<0.05;

[1] $$GM^{(t)} = \exp\left[\frac{1}{\sum_i \sum_j \sum_l w_{ijl}^{(t)}} \sum_i \sum_j \sum_l w_{ijl}^{(t)} log\left(y_{ijl}^{(t)}\right)\right], t=1, \ldots, T.$$

Figure 3: Plot of the annual obesity prevalence (and 95% confidence interval) along with the fitted joinpoint line estimated from the individual- level logistic model for probability of obesity chosen from the constrained approach



The OAPC is for the annual percent change of the odds of obesity;
* OAPC is statistically significant, p-value<0.05.
The fitted line is for predicted probability of obesity.

The R Program: Unit_Level_Joinpoint_CompSurv.R

- An R program has been released at the NCI Joinpoint website:

  https://surveillance.cancer.gov/joinpoint/Unit_Level_Jointpoint_CompSurv.R

  - **The program can be used to analyze trend using unit-level joinpoint regression models to complex survey data.**
    - Created by Joe Zoe (11/21/2025)
    - Provided by The National Cancer Institute & The IMS inc.
    - Program was developed from the project described in Liu et al. (2026).
  - **The program defines two functions**:

    1) unit_level_n_jp: This function finds the best joinpoint model with given number of joinpoint (=n_jp) for complex survey data;

    2) unit_level: This function calls for function 'unit_level_n_jp' and finds the overall best joinpoint model.
  - **Users just need read in their data, define the required variables, call the 'unit-level' function, and run for results.**

    An example of applying this program to trend analysis using BMI data from the NHIS(1991–2016) is provided at the end of the program.

Summary & Discussion

Developed and evaluated individual-level joinpoint methods for complex survey data within a finite population framework, using simulations and NHIS empirical data, and compared them with traditional aggregate-level methods across varying ICCs, APC/OAPC differences, and sample sizes.

Individual-level constrained models performed best when ICC > 0 and design effects were reasonable, particularly for correctly identifying the number of joinpoints; **aggregate-level models were more appropriate when ICC was low or zero**.

Recommended workflow: use the **constrained individual-level model** to determine the number and location of joinpoints, then estimate **APC/OAPC and AAPC/OAAPC** from the constrained model, while deriving **standard errors from the corresponding unconstrained model**.

Practical implications and future directions: individual- and aggregate-level models produced generally similar results for NHIS data. Avoiding detection of very small changes may require a **minimum detectable APC/OAPC threshold**. Future work includes **covariate adjustment through a weighting approach**.

Liu B, Kim H-J, Zou J, Feuer E.J., Graubard B.I. (2026). Extended joinpoint regression methodology for complex survey data. *Statistics in Medicine*, 45(1-2):e70374. https://doi.org/10.1002/sim.70374.

Thank you!

For questions, feel free to contact:

Benmei Liu

liub2@mail.nih.gov