# Modeling Semicontinuous Longitudinal Expenditures: A Practical Guide

VALERIE A. SMITH, DRPH
MAREN K. OLSEN, PHD

HERC CYBERSEMINAR

JUNE 20, 2018

# Agenda

- Review and compare strategies for analyzing “semicontinuous” health care expenditures collected longitudinally over multiple time points

- Discuss model specification, software available, advantages and disadvantages

- Based on manuscript:

  Smith, Valerie A., Matthew L. Maciejewski, and Maren K. Olsen. "Modeling Semicontinuous Longitudinal Expenditures: A Practical Guide." *Health Services Research* (2018).

# Goals for this CyberSeminar

- Review well-known strategies for analyzing expenditure data
- Introduce less well-known strategies that may be useful
- Provide a “roadmap” for when to use which approach
- Review frequently encountered modeling complications and how to overcome them

# Poll Question #1: What is your experience with analyzing longitudinal healthcare expenditures?

- None, or planned for upcoming work
- Some (< 1 year)
- Moderate (1-5 years)
- Considerable (>5 years)

# Health Care Expenditures: An Example of Semicontinuous Data

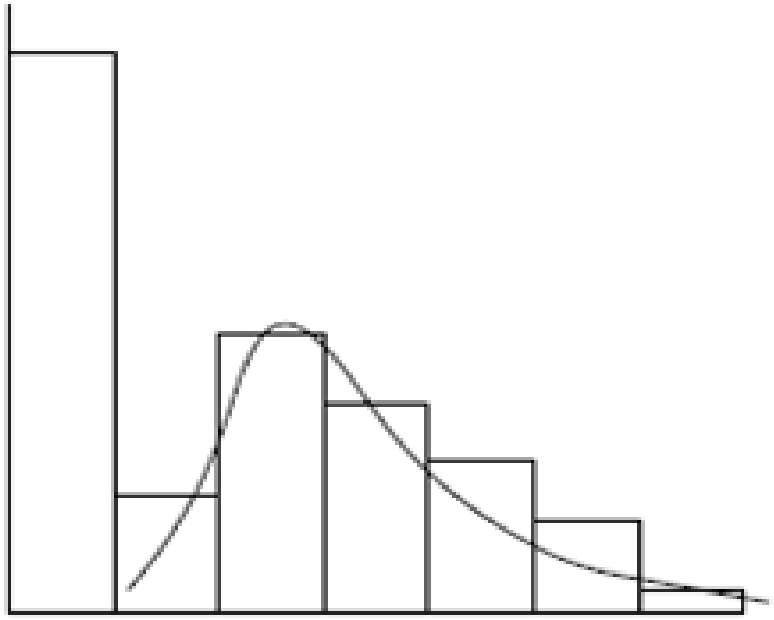
Data characterized by two components:

1) A clumping of zero values, representing a subgroup of “non-users”

- no utilization → zero healthcare costs

2) Paired with a continuous distribution of positive values among users

- some varying amount of utilization → varying levels of costs



# Health Care Expenditures: An Example of Semicontinuous Data

Data characterized by two components:

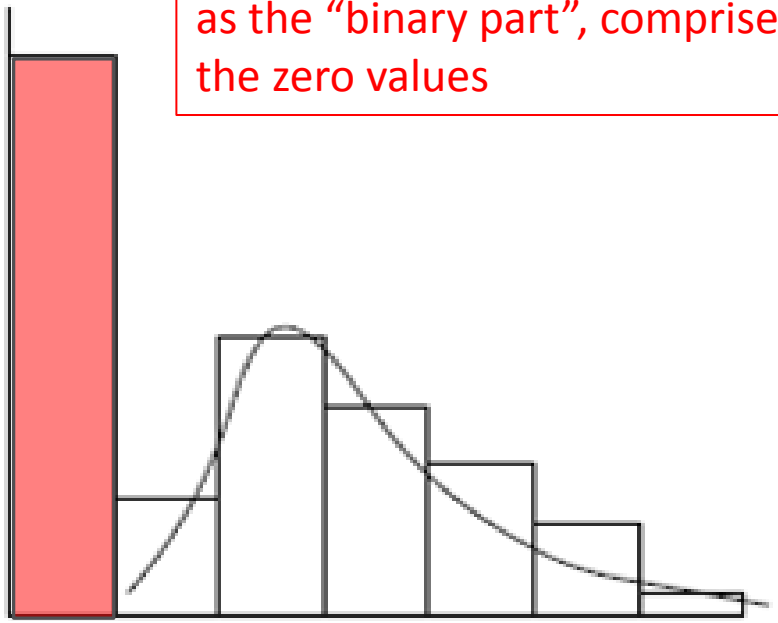
1) A clumping of zero values, representing a subgroup of "non-users"

- no utilization → zero healthcare costs

2) Paired with a continuous distribution of positive values among users

- some varying amount of utilization → varying levels of costs

First component, often referred to as the "binary part", comprised of the zero values



# Health Care Expenditures: An Example of Semicontinuous Data

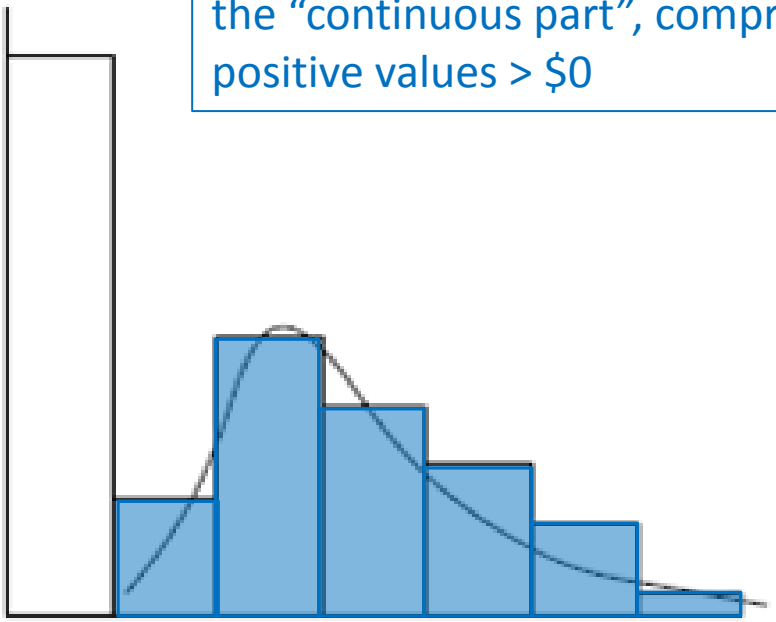
Data characterized by two components:

1) A clumping of zero values, representing a subgroup of "non-users"

- no utilization → zero healthcare costs

2) Paired with a continuous distribution of positive values among users

- some varying amount of utilization → varying levels of costs



Second component, often referred to as the "continuous part", comprised of the positive values > $0

# Modeling issues with semicontinuous data

- Highly right-skewed distribution precludes direct linear modeling
  - Use link function (e.g., log link) in a generalized linear model (GLM)
  - Transform the data (e.g., log transform) prior to using a linear model

- Decide how to accommodate zero values
  - If log-transform is used with standard model, must add small constant first
  - With a GLM, could choose to not treat zero values as any different than others
  - Utilize two-part models to separately account for the two components

# Incorporating Longitudinal Data

- Need to account for correlation among repeated measurements on same individual over time
  - Include random effects for each individual (mixed models)
  - Working correlation structure via generalized estimating equations (GEEs)
- Issues specific to semicontinuous longitudinal data:
  - Also need to account for correlation across the two components over time
    - (i.e., having a zero vs. positive expenditure at one time point may be correlated with level of expenditures incurred at another time point)
  - The distribution and proportion of positive values depends upon time frame under consideration (e.g., more people will incur costs in a year than in a month)

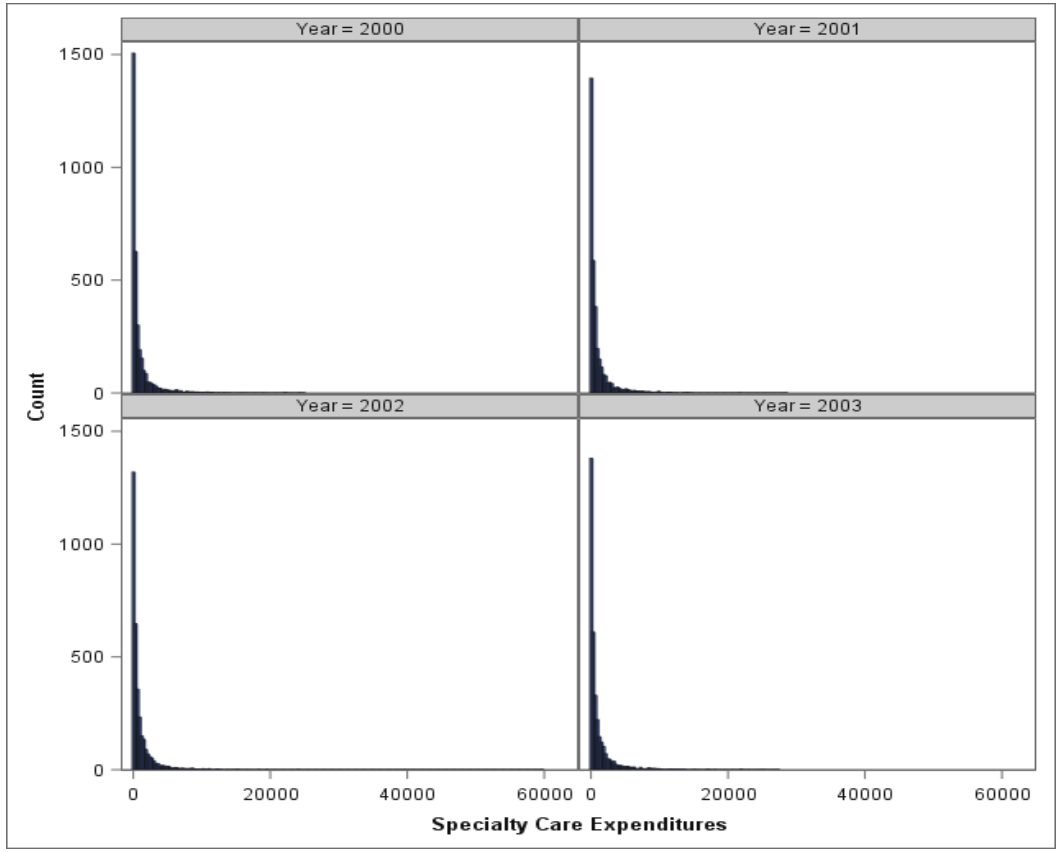
# Example: VA Specialty Care Expenditures

- December 2001: VA increased specialty care visit copayments from $15 to $50 → natural experiment to examine outpatient expenditure changes due to copayments
- Specialty care expenditures determined for each year 2000-2003, 2 years prior to & 2 years following copayment change
- Control group: Veterans exempt from copayments due to low income or service-connected disability because they did not experience the copayment increase
- 1 to 1 propensity score matching resulted in a sample of 1,693 veterans exempt from copayments & 1,693 veterans required to pay copayments, who experienced the increase

*Maciejewski, Matthew L., et al. "How price responsive is the demand for specialty care?." Health economics 21.8 (2012): 902-912.*

# Example: VA Specialty Care Expenditures

| Year | % Zeros | Mean (SD) | Median (Q1-Q3) | Maximum |
|---|---|---|---|---|
| 2000 | 29% | $904 (1923) | $225 (0-902) | $24,800 |
| 2001 | 26% | $963 (1854) | $277 (0-1008) | $28,619 |
| 2002 | 23% | $953 (2060) | $286 (21-1019) | $59,565 |
| 2003 | 27% | $999 (2024) | $271 (0-1047) | $27,320 |



# Strategies for Analysis of Longitudinal Health Care Expenditures

- "One-part" generalized linear models
  - Treats zero and positive values the same – i.e., as coming from one distribution
- "Two-part" uncorrelated generalized linear models
  - Separates the two components but requires strong, often unrealistic assumptions to be valid
- Correlated conditional two-part models
  - Separates the two components with the second component being conditional on having a positive expenditure
- Correlated marginalized two-part models
  - Explicitly incorporates zero-values but target of inference is mean of both components

# One-part Generalized Linear Models (GLMs)

- Observed expenditures (*Y*) are from a single distribution
- Link function (often log link) accommodates non-normally distributed data:

$$\log\left(E\left(Y_{ij}\right)\right) = \beta_0 + \beta_1 x_{1ij} + \ \ldots + \beta_p x_{pij}$$

$i =$ individual, $j =$ time point

- Fit using generalized estimating equations (GEEs) coupled with empirical "sandwich" variance estimation (often referred to as "robust" standard errors)

# One-part GLMs: Advantages

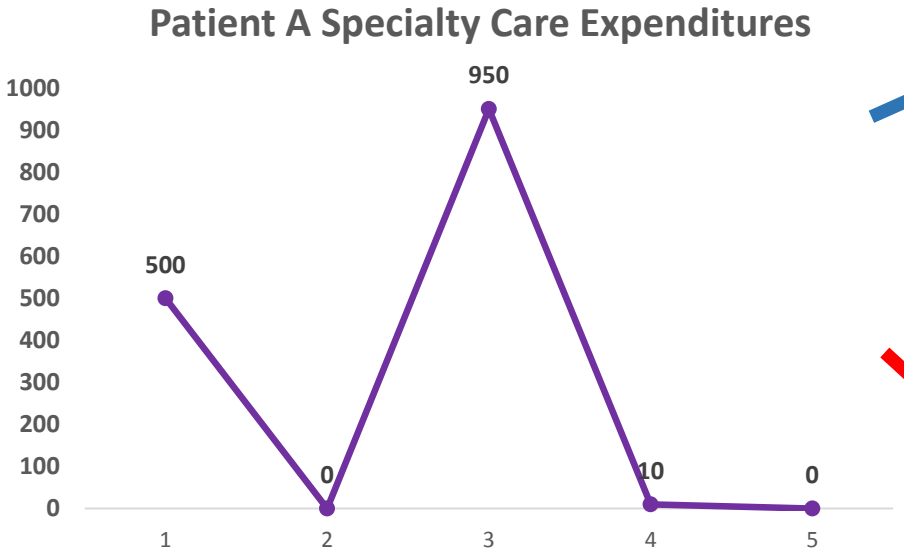
- Easily implemented in most statistical software
  - SAS's PROC GENMOD using REPEATED statement or PROC GEE
  - Stata's xtgee
  - R's geepack
- Single component with link function allows simple estimation of population-average effects on original (i.e., dollar) scale
- When a log link is used, $\exp(\beta_k)$ represents the multiplicative effect on the overall mean associated with a one-unit increase in $x_{kij}$
- Results are typically easy to communicate with research team or policy makers

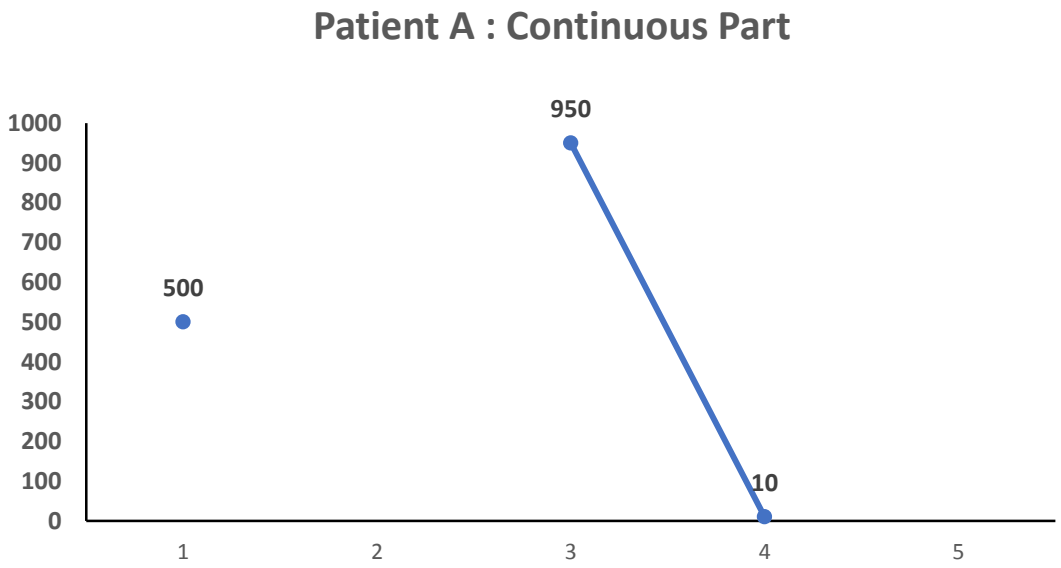
# One-part GLMs: Limitations

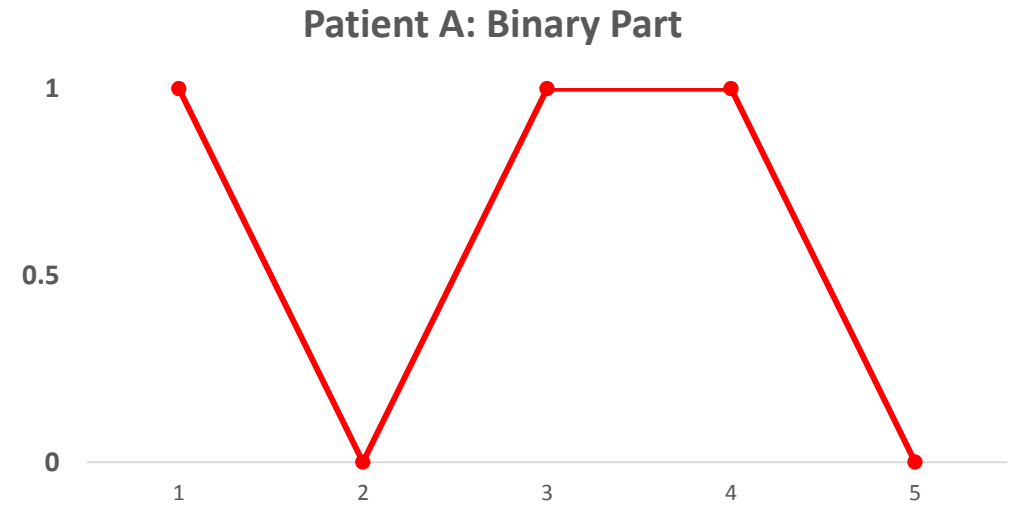
- Not suitable for data containing a significant proportion of zeros at any given time point, particularly for smaller sample sizes

- Simulation studies have shown significantly biased covariate effect estimates and high type I error rates when fit to data containing 20% or more zeros (Smith et al. 2017)
  - While GEE methodology is asymptotically unbiased, sample size needs to be very large (> 50,000)
  - “Robust” standard errors utilized with GEEs do not overcome this

# Recoding Longitudinal Expenditures for a Two-part Model



# Uncorrelated Two-Part GLMs via GEEs

Two GLM models: one for "binary part" and one for "continuous part"

$$\text{logit}\left(\Pr\left(Y_{ij} > 0\right)\right) = \alpha_0 + \alpha_1 x_{1ij} + \ \ldots + \alpha_p x_{pij}$$

$$\log\left(E\left(Y_{ij} \middle| Y_{ij} > 0\right)\right) = \gamma_0 + \gamma_1 x_{1ij} + \ \ldots + \gamma_p x_{pij}$$

- Each part separately accounts for correlation among repeated measures over time
- Each part is fit via GEEs similarly to one-part GLMs

# Uncorrelated Two-Part GLMs: Limitations

- Challenging interpretation
  - Binary component: population-average estimates of probability of incurring positive expenditures *for the entire sample at all time points*
  - Continuous component: estimates of mean level of expenditures *among the subset of individuals who incurred expenses at each time point* → target population changes depending on time point
- Potentially biased estimates
  - Two components are typically correlated over time (e.g., probability of incurring any expense at one time point correlated with level of expense at another time point)
  - Failure to account for the correlation between the two components leads to biased results (Su, Tom, Farewell 2009)

**Due to simple implementation, this provides a tempting option but there are critical problems with this approach! We do not recommend this as an alternative.**

# Correlated Conditional Two-Part (CTP) Models

Two mixed-effects models: one for “binary component” and one for “continuous component”

$$\text{logit}\left(\Pr\left(Y_{ij} > 0\right)\right) = \alpha_0 + \alpha_1 x_{1ij} + \ \ldots + \alpha_p x_{pij} \ + b_{1i}$$
$$E\left(\log\left(Y_{ij} \middle| Y_{ij} > 0\right)\right) = \delta_0 + \delta_1 x_{1ij} + \cdots \delta_p x_{pij} + b_{2i}$$

Random intercepts, $b_{1i}$ and $b_{2i}$, assumed to jointly follow a multivariate normal distribution to both account for correlation among repeated measures over time & correlation among the two components of the model

# Correlated Conditional Two-Part (CTP) Models: Advantages

- Fully parametric model fit with maximum likelihood estimation
  - Missing at random assumption for missing data
- Variance components of random effects quantifies how two parts are related:
  - Positive estimate of covariance between the binary component's random intercept and the continuous component's random slope → probability of any expenditure is positively related to the amount of expenditures over time.
- Models with random intercepts can be fit in standard software packages
  - SAS PROC NLMIXED, Mplus

# Correlated Conditional Two-Part (CTP) Models: Limitations

- Models with more complex random effects specification are computationally challenging
- Conditional model does not allow for model estimates to convert predictions from the log-$ scale back to the $ scale
- Subject-specific interpretation may not be of primary interest to researchers and policy makers

# Correlated Marginalized Two-Part (MTP) Models

- Blends marginal interpretations of the one-part GLMs and the structure of the correlated CTP model

- First part models the probability of a positive expenditure, but second part incorporates both the zero and positive values:

$$\text{logit}\left(\Pr\left(Y_{ij} > 0\right)\right) = \alpha_0 + \alpha_1 x_{1ij} + \ \ldots + \alpha_p x_{pij} + b_{1i}$$

$$\log\left(E\left(Y_{ij}\right)\right) = \beta_0 + \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_{2i}$$

- Random intercepts, $b_{1i}$ and $b_{2i}$, assumed to jointly follow a multivariate normal distribution to both account for correlation among repeated measures over time & correlation among the two components of the model

# Correlated MTP Models: Advantages

- Allows estimation of both probability of use component and covariate effects on the overall mean expenditures, which may of interest to policy-makers or investigators
  - $\exp(\alpha_k)$ represents subject-specific odds ratio for incurring positive expenditures associated with one-unit increase in $k^{th}$ covariate
  - $\exp(\beta_k)$ represents multiplicative effect on overall mean expenditures of entire population associated with a one-unit increase in $k^{th}$ covariate. This has a dual population average & subject specific interpretation, assuming the $k^{th}$ covariate is not also a random effect itself.

- When implemented in SAS PROC MCMC, any value calculated from the parameters can be easily obtained with corresponding credible intervals or highest posterior density intervals (Bayesian analog to confidence intervals)

# Correlated MTP Models: Limitations

- Computationally challenging
  - It is difficult computationally to allow the cross-part correlation, which can lead to long run times and convergence issues

- Not as easily implemented in standard software
  - We provide SAS PROC MCMC code, and can provide SAS PROC NLMIXED code, but it is not incorporated as standard procedures in any software

# Poll Question #2: What is your experience using any the 4 methods presented?

- One Part GLM
- Two Part GLM
- Correlated Conditional Two-Part Model
- Correlated Marginalized Two-Part Model

# How do I choose which method to use?

First steps:

- Use descriptive statistics and plots to understand distribution at EACH time point
  - What is the % of 0's at each time point?
  - Degree of skewness and extreme values at each time point
- What is the overall sample size at each time point?
- What are my research questions of interest?

# How do I choose which method to use?

| | N at each time point | % of 0's at each time point |
|---|---|---|
| Scenario 1 | 200 – 1,000 | ≤ 10% |
| Scenario 2 | 1,000 – 50,000 | 10% - 15% |
| Scenario 3 | > 50,000 | 15% - 30% |

→ **One-part GLM**

Examples of research questions that can be answered with this method:

1. What is the effect of being required to pay a copayment on **overall mean specialty care expenditures** in each year?

2. What are the estimated or predicted **overall mean expenditures** for those with and without a copayment requirement in each year?

# How do I choose which method to use?

| | N at each time point | % of 0's at each time point |
|---|---|---|
| Scenario 4 | 200 – 1,000 | > 10% |
| Scenario 5 | 1,000 – 50,000 | > 15% |
| Scenario 6 | > 50,000 | > 30% |

→ **Correlated MTP or CTP**

Choice depends upon:

- Research questions and estimated quantities of interest
- Computational challenges of model estimation

# Correlated CTP vs MTP: Example Research Questions

| | Conditional Two-Part Model | Marginalized Two-Part Model |
|---|---|---|
| Binary part | In each year, what is probability of incurring specialty care expenditures for an individual required to pay a copayment as compared to if that individual was not required to pay a copayment? | |
| Continuous part | Conditional upon having a specialty care visit, is there a difference in specialty log-expenditures in each year for an individual required to pay a copayment as compared to an individual not required to pay? | What are the estimated overall mean specialty care expenditures for those with and without a copayment requirement in each year? |
| Variance components of random effects | How is probability of any expenditure related to log-positive expenditures over time? | How is probability of any expenditure related to overall mean expenditures over time? |

# Distribution Choices

- For either the CTP or MTP models, three popular distributional choices are:
  - Log-normal: Assumes log of positive values follows a normal distribution
  - Log-skew-normal: Allows additional skewness on the log-scale & takes the log-normal as a special case
    - Chai & Bailey, 2008; Smith et al., 2017b
  - Generalized gamma: Takes Weibull, gamma, inverse gamma, and log-normal distributions as special cases
    - Liu et al., 2010

# Correlated CTP and MTP Computational Challenges

- Correlated random effects make it difficult for these models to converge
  - Maximum likelihood solution (e.g., proc nlmixed) can take many hours
  - Both CTP and MTP may be fit with Bayesian methods
    - Smith et al (2018) provides example SAS proc mcmc code for MTP
    - Cooper et al (2007) provides example WinBugs code for CTP
- Extreme skewness also impacts model fit and convergence

# Correlated CTP and MTP Computational Challenges: Our tips

- Start with estimating uncorrelated model
  - Make sure this simplified model converges
  - Use final estimates as starting values for correlated model estimation
- Try different software options --- underlying maximization and estimation algorithms differ slightly
  - See *Tips and Strategies for Mixed Modeling with SAS/STAT Procedures (2012)*
- Try a different distribution choice
- Consider simple modifications to your data or research question
  - increasing time period of observation → decreases % of 0's
  - reducing study design to pre-post, rather than repeated longitudinal

# Conclusion

- Model selection driven by both data structure and research goals
  - Two-part models are often necessary when data have > 10-15% zeros
  - Two-part models accommodate multiple research question forms

- Correctly modeling longitudinal expenditures with many zeros can be computationally complex, but code is provided in many papers

- Statistical methods still evolving in this area

# References

Smith, Valerie A., Matthew L. Maciejewski, and Maren K. Olsen. "Modeling Semicontinuous Longitudinal Expenditures: A Practical Guide." *Health Services Research* (2018)

Maciejewski, Matthew L., et al. "How price responsive is the demand for specialty care?" *Health Economics* 21.8 (2012): 902-912.

Smith, Valerie A., et al. "Two parts are better than one: modeling marginal means of semicontinuous data." *Health Services and Outcomes Research Methodology* 17.3-4 (2017): 198-218.

Cooper, Nicola J., et al. "Predicting costs over time using Bayesian Markov chain Monte Carlo methods: an application to early inflammatory polyarthritis." *Health economics* 16.1 (2007): 37-56.

# References

Chai, High Seng, and Kent R. Bailey. "Use of log-skew-normal distribution in analysis of continuous data with a discrete component at zero." *Statistics in Medicine* 27.18 (2008): 3643-3655.

Smith, Valerie A., et al. "A marginalized two-part model for longitudinal semicontinuous data." *Statistical methods in medical research* 26.4 (2017b): 1949-1968.

Liu, Lei, et al. "A flexible two-part random effects model for correlated medical costs." *Journal of Health Economics* 29.1 (2010): 110-123.

Kiernan, Kathleen, Jill Tao, and Phil Gibbs. “Tips and Strategies for Mixed Modeling with SAS/STAT® Procedures.” *SAS Global Forum 2012.* Available at http://support.sas.com/resources/papers/proceedings12/332-2012.pdf

# Questions?

Feel free to contact:

*Valerie Smith*

valerie.smith9@va.gov

*Maren Olsen*

maren.olsen@va.gov