Modeling errors in physical activity data

Sarah Nusser

Department of Statistics and

Center for Survey Statistics and Methodology

Iowa State University

Collaborators: N. Beyler, A. Carriquiry, W. Fuller, G. Welk

Presented at

Measurement of Active and Sedentary Behaviors: Closing the Gaps in Self-Report Methods

National Institutes of Health, Bethesda, MD • July 21, 2010

A brief tour through ideas

- How do we frame the modeling problem to address public health questions?
- Why is it important to use measurement error models?
- How do we think about constructing statistical models that account for errors?
- What do these models look like? How do you design a study to use them?
  - Unbiased activity measure
  - Biased activity measure

What is the relevant concept?

- Many health outcomes are related to the long-term behavior of individuals, not what an individual did yesterday or last week
- Long-term behavior is characterized by the usual activity for an individual

Usual Activity

An individual's activity per unit time averaged over a long period of time

Inference about whom?

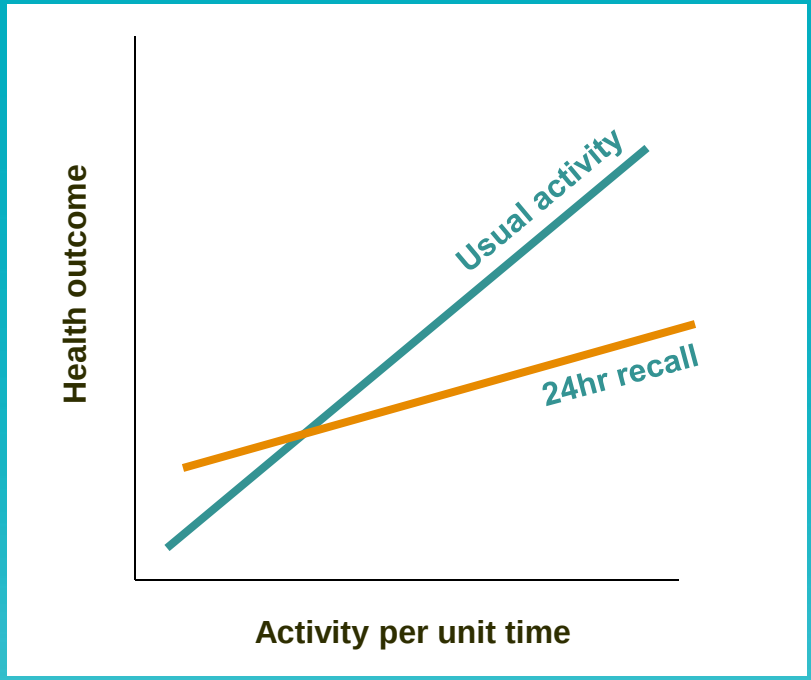
- Although usual activity is an individual-level metric, our focus is on public health questions
- To address these questions, we need to characterize usual activity patterns for a group or **population of individuals**
  - Requires individual measurements on a (probability) sample of individuals from the pop.
- From this, we can estimate pop. characteristics
  - % of population with inadequate activity
  - Relation between health outcome & usual activity

Basic approach

- Use self-report activity data on a sample of individuals to make inferences about usual activity in the population
- Because self-report activity data measure usual activity with error, we'll need measurement error models to obtain inferences about usual activity in the population that are adjusted for error

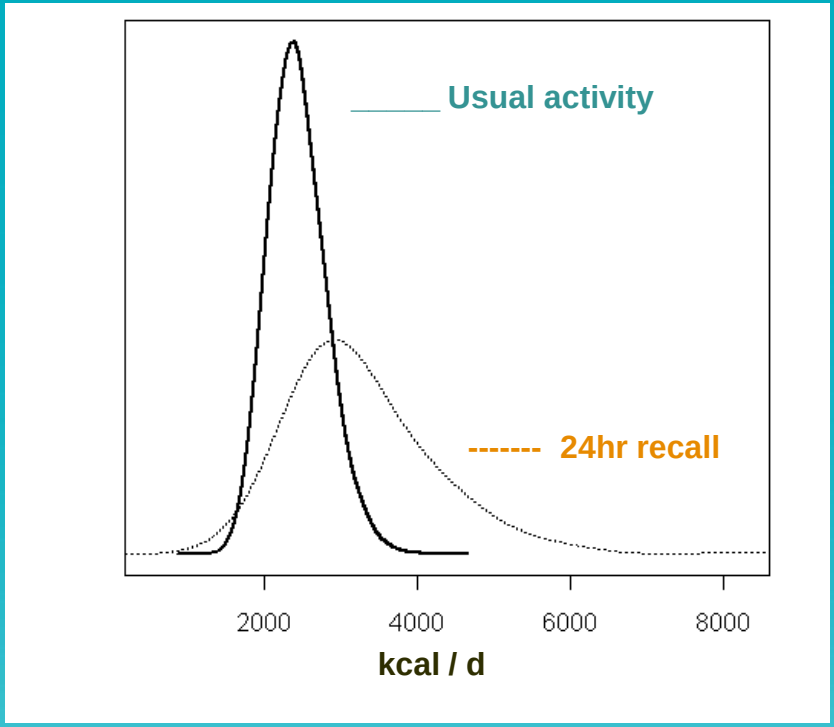
What if we don't adjust for error in self-report activity?

- Biased estimates of correlation and regression coefficients
- For correlation or a regression that uses self-reported activity as the only covariate, slope is attenuated
- Bias not predictable in regression with more covariates than self-reported activity



What if we don't adjust for error in self-report activity?

- Want to estimate average activity levels in the population?
  - If self-reported data are biased, estimated mean usual activity is biased
- Want to estimate the percent of population with poor activity levels?
  - Extra variation in self-reported data cause the spread of distribution to be too wide, percentages are biased



Addressing errors in variables requires attention to design and analysis methods

Prevention
Avoid error

- "Right" measurement: data collected are related to the concept of interest (if possible)
- "Good" measurement: protocol or question to generate accurate and precise measurements or responses

Adjustment
Reduce impact of error

- Statistical models that include terms to describe systematic and random errors
- Statistical designs that generate the kind of data required to estimate these parameters

Sources of errors

- Error framework used in statistical surveys to describe errors
  - Sampling error
  - Coverage (frame) error
  - Nonresponse error
  - Specification error
  - Measurement error
  - Processing error

Sample design and response process

Measurement and data generation process

Specification error

- We are generally interested in collecting data on a specific concept (construct), even if we can't always directly observe it

Specification Error
Error that arises when our question or measurement method does not match the target concept

- Example
  - Concept: usual daily activity
  - Measurement: activity yesterday (1 day)

Measurement error

Measurement error
Error that arises in obtaining a response or measurement during data collection

- Examples
  - Respondent errors in recalling prior activities
  - Deliberate or subconscious over-reporting of time spent in vigorous activity
  - Mistakes in recording data

Processing error

Processing error
Error that is induced in manipulating the raw data to create analysis variables

- Examples
  - Assigning METs to self-reported activities
  - Criteria for classifying a behavior as MVPA or as sedentary
  - Proprietary processing algorithms applied to data from an activity monitor

Modeling errors in activity data

- To illustrate concepts, we will consider some examples of models
  - Self-report: 24 hr recall
  - Activity metric: total energy expenditure
  - Target concept: usual daily energy expenditure

Target variable: usual daily activity (U) (… and some notation)

- Usual activity of individual k = U_k
- We'll focus on the distribution of usual daily activity across individuals in population
- Parameters for usual activity distribution
  - Mean usual activity for the population = μ_U
  - Person-to-person variation in usual activity = σ_U^2

Modeling 24 hr recall data

- 24 hr recall for person k on day j = R_{kj}
- Short-term recall data do not measure usual activity directly, but they are related to usual activity (specification error)
- Self-report instruments also have some form of measurement (and processing) error
  - Some errors are random, varying from person to person or day to day
  - Other errors are systematic throughout the population, e.g., over-reporting activity

Simple model for 24 hr recall (R) and usual daily activity (U)

- A 24 hr recall R is related, but not equal to usual activity U

$$R_{kj} = U_k + D_{kj}$$

- Some model assumptions

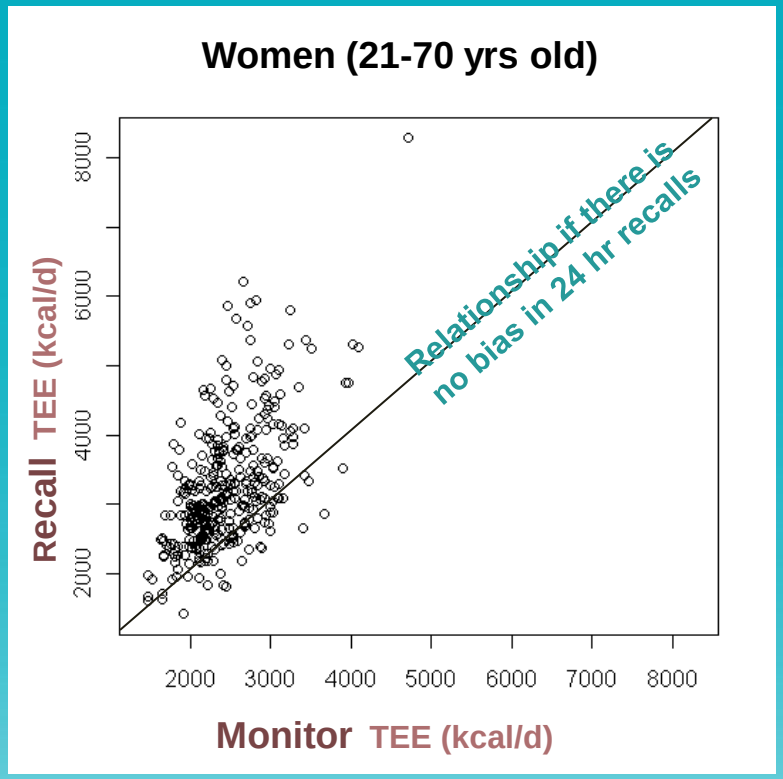
	Variable	Mean	Variance	
Usual activty for person k	U_k	μ_U	σ_U^2	Among person variance (variation from person to person)
Deviation of today's activity from usual act. on day j for person k	D_{kj}	0	σ_D^2	Within person variance (day to day variation for a person)
24 hr recall on day j for person k	R_{kj}	μ_U	$\sigma_U^2 + \sigma_D^2$	Recall data include both types of variance

Simple model for 24 hr recall (R) and usual daily activity (U)

- The among person variance for U can be estimated (i.e., separated from the within-person variance for D) if our design includes
  - 2+ measurement days for subsample or all of the study sample to provide info on within-person variation
  - Model assumptions typically require independent days (e.g., separated by a few days or more)

Self-report data may be biased

- Example using early data from Physical Activity Measurement Survey (PAMS) (NIH grant)
- Sample of Iowa adults
- Concurrent observations on each sampled adult
  - 24 hr recall of all activities
  - 24 hr activity data from multi-sensor monitor



Accounting for population-level bias in self-reports

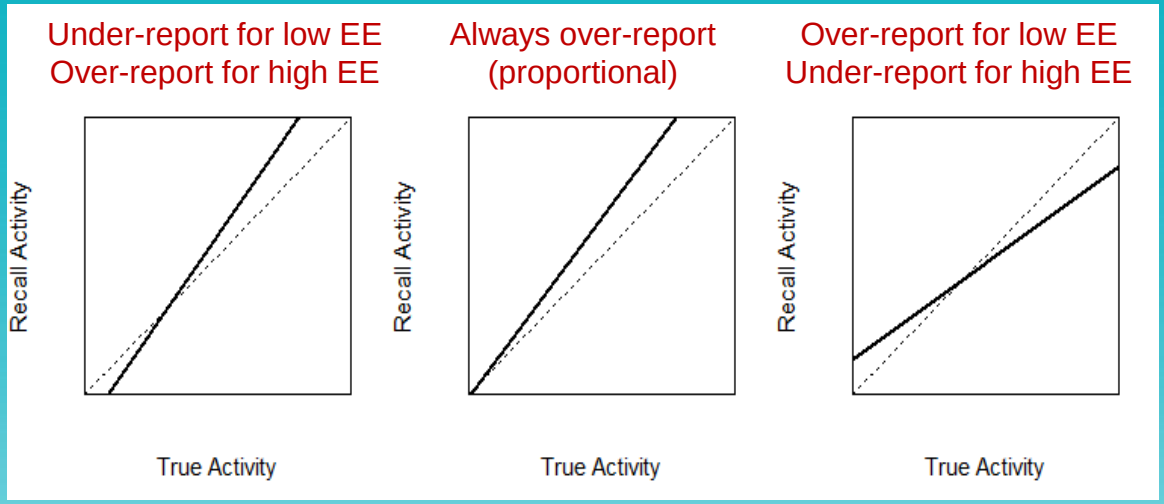
- We can add an intercept and slope to express systematic reporting (& processing) bias in pop.

$$R_{kj} = \beta_0 + \beta_1 (U_k + D_{kj})$$

intercept slope

Dashed line ---------
No systematic bias
$\beta_0 = 0$ and $\beta_1 = 1$

Solid line _____
Systematic bias present
$\beta_0 \neq 0$ and/or $\beta_1 \neq 1$



Some individuals are more accurate reporters than others

- A person may have their own tendency to over- or under-report that is different from the population tendency
- This is expressed as random variation across individuals in the pop. via the term S_k

$$R_{kj} = \beta_0 + \beta_1 (U_k + D_{kj}) + S_k$$

Random measurement errors

- Finally, there are generally some additional unexplained deviations in a person's reported activity on day j relative to their true activity
- Random measurement error that varies from day to day for a person is expressed via E_{kj}

$$R_{kj} = \beta_0 + \beta_1 (U_k + D_{kj}) + S_k + E_{kj}$$

Some model assumptions

$$R_{kj} = \beta_0 + \beta_1 (U_k + D_{kj}) + S_k + E_{kj}$$

	Variable	Mean	Variance	
Usual activity for person k	U_k	μ_U	σ_U^2	Among person variance (variation from person to person)
Deviation of today's activity from usual act. on day j for person k	D_{kj}	0	σ_D^2	Within person variance (day to day variation for a person)
Individual deviations in reporting bias for pers. k	S_k	0	σ_S^2	Person-to-person variance in reporting bias
Measurement error for person k on day j	E_{kj}	0	σ_E^2	Variation due to random measuremt & processing error
24 hr recall on day j for person k	R_{kj}	$\beta_0 + \beta_1 \mu_U$	$(\beta_1^2 \sigma_U^2) + (\beta_1^2 \sigma_D^2) + \sigma_S^2 + \sigma_E^2$	Recall data biased and include all types of variances

What does this model tell us?

- The mean of the recall data may not be equal to the mean of the usual activity metric

Without adjustment, mean of recall data may over or under estimate usual activity mean

- Random variation in the recall data contains more than person-to-person variation in usual activity levels

Without adjustment, variance of the recall data overstates true variation in usual activity

Design considerations

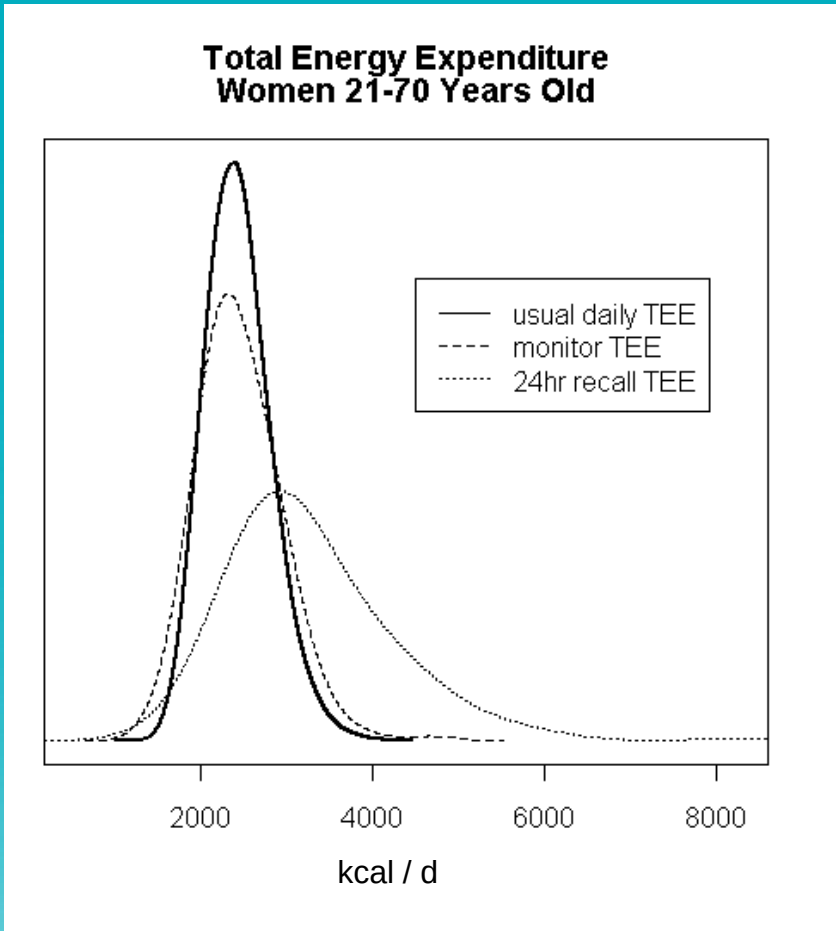
- To estimate person-to-person usual activity variance, need replicate observations on at least subset of sampled individuals
- To estimate systematic bias in self-reports, need a second unbiased measurement (M) (e.g., multisensor monitor) as a benchmark
- Measurement error models for monitor M & recall R

$$M_{kj} = (U_k + D_{kj}) + F_{kj}$$

$$R_{kj} = \beta_0 + \beta_1 (U_k + D_{kj}) + S_k + E_{kj}$$

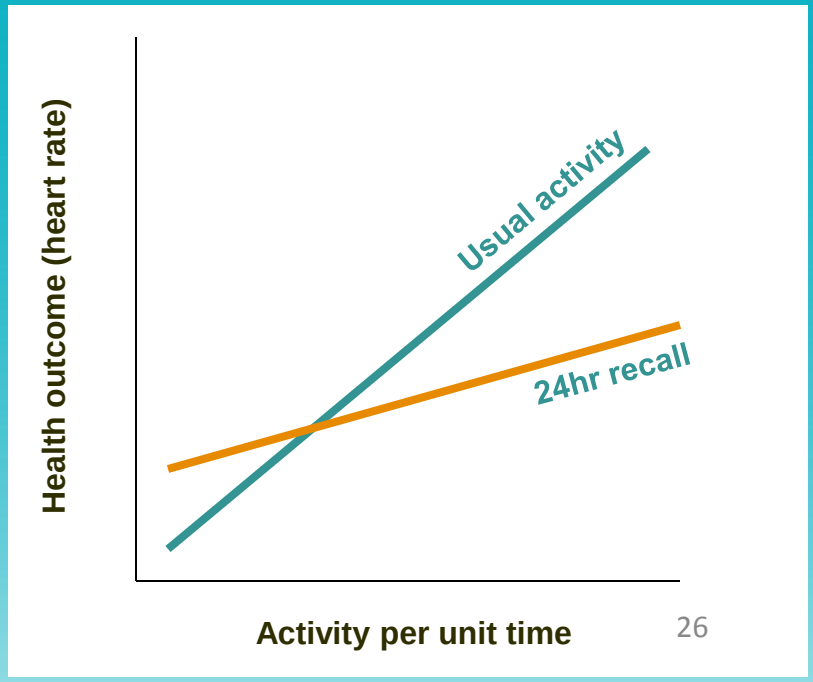
Estimated total energy expenditure distributions (Early PAMS data)

- 24 hr recall
  - Over-reporting bias
  - Over-dispersion
- 24 hr multi-sensor monitor
  - Benchmark to adjust for over-reporting
  - Still has extra variation
- Usual activity
  - Unbiased estimate of mean
  - Proper spread of distribution



Back to the relationship between health outcomes and activity

- With this model, we can estimate attenuation factor associated with recall data for adjusting correlation
- We can also generate predicted usual activity values for individuals to use in regressions (these are poor estimates of individual usual activity, but have appropriate mean and variance at population level)



Summary

- A few examples of measurement models and study designs to adjust self-reports for errors
  - Use understanding of the errors to build the model
  - Develop study design to support model estimation
- With model, can obtain estimates that reflect properties of usual activity in the population
  - Usual activity distribution: appropriate mean and variance (bias removed)
  - Regression: obtain unbiased estimates of relationship between activity and health outcome