On the use of predicted values in nutritional epidemiology: Example from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL)

Pamela Shaw

Department of Biostatistics, Epidemiology and Informatics

University of Pennsylvania

shawp@upenn.edu

SER Annual Meeting

June 23, 2021

DEPARTMENT of BIOSTATISTICS EPIDEMIOLOGY & INFORMATICS

Perelman School of Medicine University of Pennsylvania

Acknowledgments

This is joint work with members of STRATOS TG4: Measurement Error and Misclassification Topic Group and other collaborators

- Laurence Freedman (Gertner Institute, STRATOS TG4 chair)
- Victor Kipnis (NCI, STRATOS TG4 chair)
- Paul Gustafson (UBC, STRATOS TG4 member)
- Daniela Sotres-Alvarez (UNC-Chapel Hill)
- Jenny Shen (UPenn)

Introduction

- In epidemiology, there are many measurements that are difficult to obtain directly:
  - Expensive (Resting Energy Expenditure)
  - Burdensome (24 hour urinary sodium)
  - Impossible (Usual energy intake)
- One strategy is to use prediction equations to measure them indirectly
- Many analyses proceed with predicted values as if they were observed data
- Using predicted values instead of observed data in study analyses can corrupt study results if the (Berkson) prediction error is not handled appropriately

Berkson vs Classical measurement error (Keogh et al 2021)

- **Classical error adds random noise to the true value X**

X* = X + error

Example: A single measure of blood pressure X* can fluctuate randomly around an innate true average value X

- **Observations with Berkson error are less variable than true value X**

X = X* + error

Example: A predicted value from a regression equation has less variability than the original outcome, due to unexplained variance

Example from the Hispanic Community Health Study

(Lavange et al 2010)

Question of interest: Does sodium intake vary by Hispanic ethnicity?

HCHS main cohort: n = 16,415

Male: 40%

Age: mean 43y; range: 18-74y

Main dietary assessment: two 24 hour recalls, known to be subject to bias

SOLNAS: Calibration sub-study: n = 477

Biomarker: 24 hour urinary sodium was obtained to create calibration equations that correct for the measurement error/bias in self-reported sodium (Mossavar-Rahmani et al 2017)

Calibration equations as prediction equations

If a biomarker Y has classical error one can estimate true intake (Y)**

by regressing Y** on self-reported Y* and other covariates (age, BMI, gender, language preference, restaurant score, fast food intake)

Step 1: use Y** to Fit Model:

$Y = b_0 + b_1$ Y* + b_2 X2 + b_3 X3 + …. b_k Xk + epsilon

Step 2: Use fitted regression equation to derive predicted (mean) intake for a give sent of covariates.

- $\hat{Y} = \hat{b}_0 + \hat{b}_1$ Y* + $\hat{b}_2$ X2 + $\hat{b}_3$ X3 + …. $\hat{b}_k$ Xk

- The unexplained variance from the calibration equation results in the Berkson error in measure $\hat{Y}$
  - Y = $\hat{Y}$ + e

A simple fix for Berkson error

- **The fundamental problem of predicted values is their Berkson error makes them less variable than they should be.**

- **When the distribution is of interest, a simple fix is to add back the missing variance to the calibrated value.**
  - This can be accomplished from simulating error e ~ N(0,σ^2), where $\sigma^2 = \sigma^2_{Resid}$- σ^2_{within}
  - Y_{imp} = $\hat{Y}$ + e
  - A multiple imputation approach can be used to estimate the quantities of interest

- If the error-prone variable Y* only had classical measurement error, the NCI method (SAS macro) could be used to estimate the distribution (Tooze et al 2006)

Numerical Study

Here consider a computer simulation study

- $X_i = \gamma_0 + \gamma_1 X_i^* + \gamma_2 Z_i + \varepsilon_i$ (relationship between truth and self-report)
- $X_i^{**} = X_i + u_i.$ (relationship between truth and biomarker)

Fit calibration equation and compare distributions of

True X

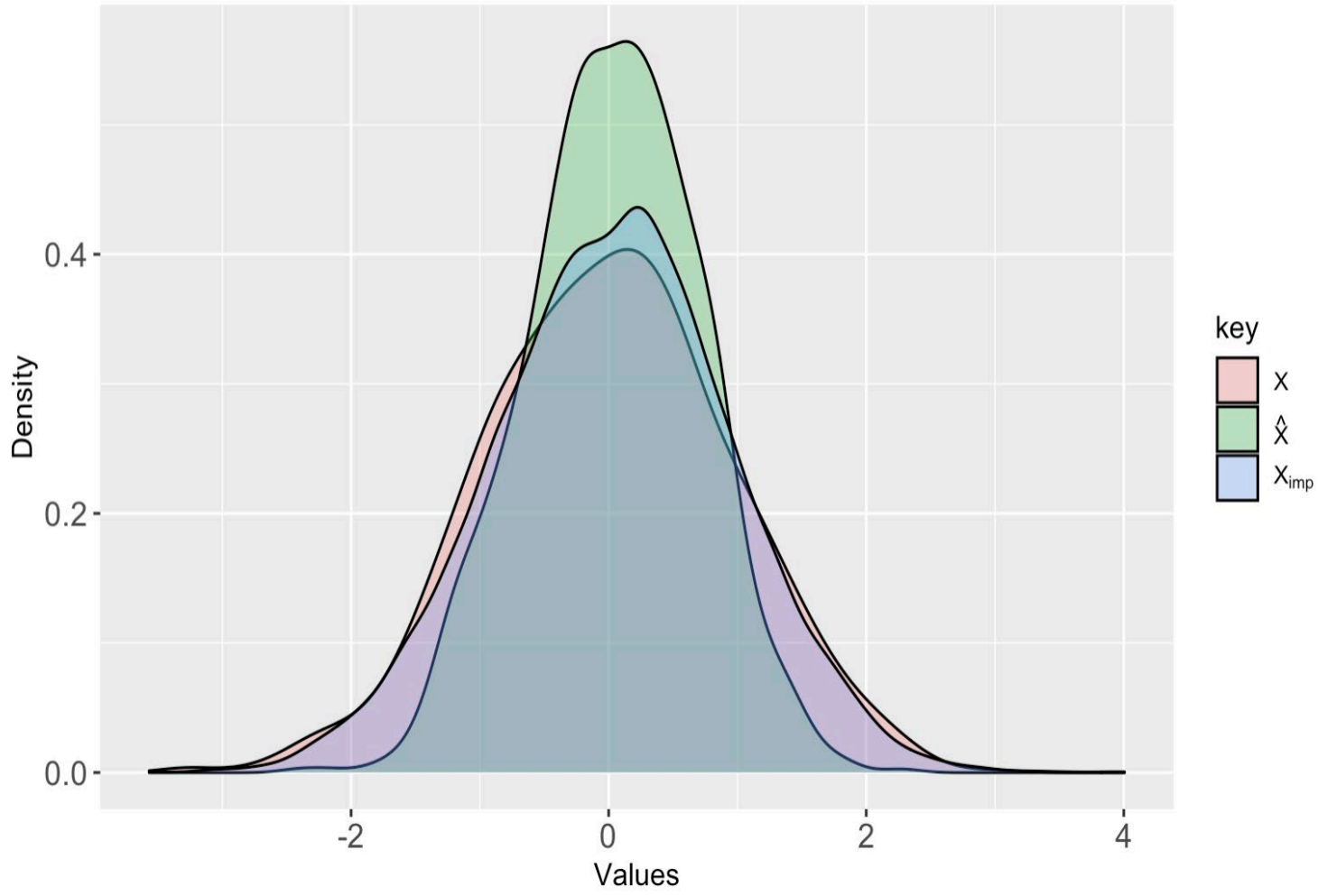
Predicted values $\widehat{\boldsymbol{X}}$: From calibration equation regressing X_i^{**} on (X,Z)

Imputed X_{imp}: $X_{imp,i}^{(m)} = \widehat{\boldsymbol{X}}_i^{(m)} + e_i^m$

Simulation settings: normal distributions (simplistic); classical error in X*; var(ε)=var(X); 1000 simulations; 1000 bootstraps (CI for Ximp quantiles)

R software (version 4.04)

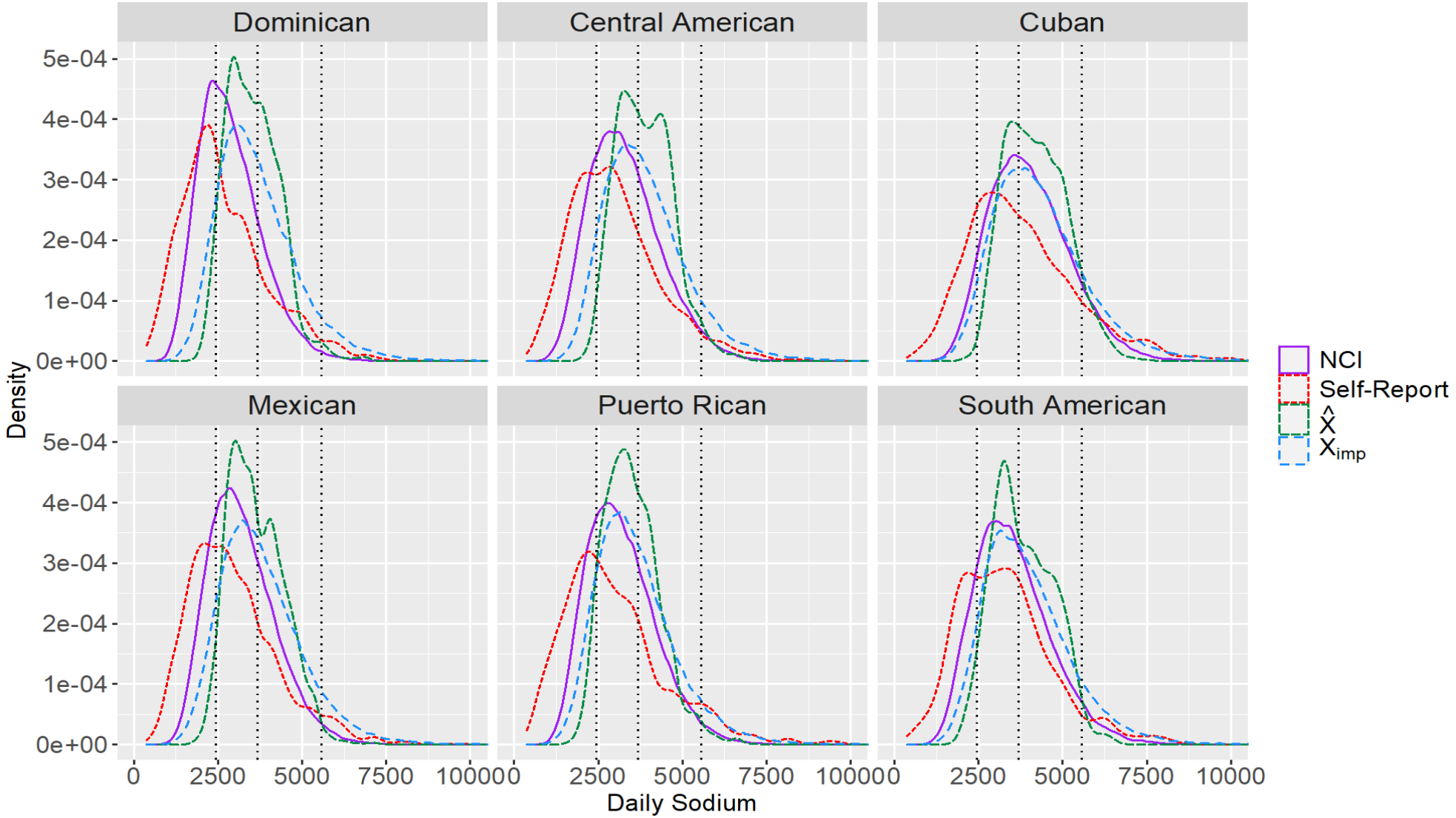
Simulation Study Results



Berkson error biases quantiles and standard errors

	X			$\hat{X}$			X_{imp}		
%-tile	Mean	ESE	CP	Mean	ESE	CP	Mean	ESE	CP
10th	-1.277	0.055	94.7	-0.949	0.105	1.3	-1.289	0.121	97.1
25th	-0.672	0.043	94.8	-0.501	0.079	8.0	-0.679	0.089	96.6
50th	-0.001	0.039	96.1	-0.002	0.067	6.0	-0.002	0.074	97.4
75th	0.674	0.043	94.5	0.498	0.078	8.3	0.675	0.087	96.5
90th	1.276	0.053	95.8	0.947	0.103	0.8	1.284	0.119	96.9

Similar results seen in HCHS/SOL



Marked differences in percentiles and SE (low R-squared)

Percentile	Self-report X*	$\hat{X}$ Treated as observed	X_{imp}
10th	1571 (66)	2711 (37)	2401 (164)
25th	2091 (63)	3017(33)	2869 (125)
50th	2853 (80)	3514 (56)	3547 (102)
75th	3920 (108)	4153 (77)	4422 (168)
90th	5026 (132)	4837 (78)	5371 (323)

Proportion with Sodium Intake <2300 mg/day

Ethnic Background	Self-Report (X*)	$\widehat{X}$ (treated as observed)	X_{imp} (account for Berkson error)
Dominican (n= 1451)	0.456 (0.018)	0.025 (0.005)	0.104 (0.055)
Central American (n=1708)	0.334 (0.018)	0.004 (0.002)	0.054 (0.034)
Cuban (n=2329)	0.174 (0.009)	0.001 (0.001)	0.034 (0.023)
Mexican (n=6405)	0.347 (0.010)	0.018 (0.003)	0.080 (0.044)
Puerto Rican (n=2677)	0.367 (0.014)	0.033 (0.004)	0.111 (0.045)
South American (n=1061)	0.272 (0.018)	0.020 (0.007)	0.067 (0.036)
All (N=15825)	0.320 (0.007)	0.016 (0.001)	0.075 (0.004)

Comments regarding other analyses

- **Predicted values as covariates in a regression (classic regression calibration)**
  - Berkson error in a covariate will not bias regression coefficient (so long as prediction equation correct)
  - Standard errors still need to be adjusted to account for uncertainty in predict model coefficients
- **Predicted values as the outcome in a regression (classic regression calibration)**
  - Need Berkson error to be independent of the covariates in the regression model
  - Coefficients will be biased
  - Buonaccorsi method (1991) can be used to address bias, so long as non-differential error in predicted value

Discussion

- **There is increasing use of prediction and calibration equations in medicine**
- **Naïve analyses with predicted outcomes are subject to multiple biases**
  - Distributional summaries are biased, quantiles appear less extreme
  - Regressions reliant on predicted outcomes will have biased coefficients
  - Regressions reliant on predicted exposures need SE adjustment
- **Presented methods do not address when prediction error is differential**
  - Deficiencies in the prediction model leads to correlation between prediction error and other analysis variables
  - Recent work (Haber et al ; Ogburn et al 2021) has outlined bias related to misspecified prediction models
- **Awareness of the effects of Berkson error and methods to adjust for it need more attention**

References

- Buonaccorsi J. Measurement errors, linear calibration and inferences for means. *Comp stat and Data Analysis,*1991;11(3):239-57.
- Haber G, Sampson J, Graubard B. Bias due to Berkson error: issues when using predicted values in place of observed covariates. Biostatistics. 2020 Feb 10.
- Haber G, Sampson J, Flegal KM, Graubard B. The perils of using predicted values in place of observed covariates: an example of predicted values of body composition and mortality risk. The American Journal of Clinical Nutrition. 2021 Apr 8.
- Lavange L et al. Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Ann Epidemiol.* 2010;20(8):642-649.
- Mossavar-Rahmani Y, Sotres-Alvarez D, Wong W, Loria C, Gellman M, Van Horn L, Alderman M, Beasley J, Lora C, Siega-Riz AM, Kaplan R, Shaw PA. Applying recovery biomarkers to calibrate self-report measures of sodium and potassium in the Hispanic Community Health Study/Study of Latinos *Journal of Human Hypertension*, 2017; 31(7): 462-473, Jul 2017.
- Ogburn EL, Rudolph KE, Morello-Frosch R, Khan A, Casey JA. A Warning About Using Predicted Values From Regression Models for Epidemiologic Inquiry. American Journal of Epidemiology, In Press
- Keogh RH, Shaw PA, Gustafson P, Carroll RJ, Deffner V, Dodd KW, Küchenhoff H, Tooze JA, Wallace MP, Kipnis V, Freedman LS. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part I – basic theory, validation studies and simple methods of adjustment. *Statistics in Medicine* 2020 Jul 20;39(16):2197-2231.
- Shaw PA, Gustafson P, Carroll RJ, Deffner V, Dodd KW, Keogh RH, Kipnis V, Tooze JA, Wallace MP, Küchenhoff H, Freedman LS. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part II –more complex methods of adjustment and advanced topics. *Statistics in Medicine* 2020 Jul 20;39(16):2232-2263.
- Tooze JA, Kipnis V, Buckman DW, Carroll RJ, Freedman LS, Guenther PM, Krebs-Smith SM, Subar AF, Dodd KW. A mixed-effects model approach for estimating the distribution of usual intake of nutrients: the NCI method. Statistics in medicine. 2010 Nov 30;29(27):2857-68.