Cautionary notes for regression analyses that use predicted values as an outcome or exposure

Pamela Shaw
Division of Biostatistics
Kaiser Permanente Washington Health Research Institute
Pamela.A.Shaw@kp.org

IBC, Riga
July 14, 2022

KAISER PERMANENTE®

Acknowledgments

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

- Laurence Freedman, Paul Gustafson, Victor Kipnis, Douglas Midthune, Pamela Shaw (STRATOS TG4)
- Lillian Boe, Eunyung Park (UPenn)
- Daniela Sotres-Alvarez (UNC-Chapel Hill)

- We would like to thank the investigators of the Hispanic Community Health Study/ Study of Latinos (HCHS/SOL) for the use of HCHS/SOL data.
- This work was supported in part by NIH grant R01-AI131771 (PS, LB) and by NHLBI contract 75N92019D00010 (DSA).

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 $\hat{X}$ from a regression equation has less variability than the original outcome, due to unexplained variance

Example from the Hispanic Community Health Study/Study of Latinos

(Lavange et al 2010)

Questions of interest: Is potassium intake associated with hypertension? Does potassium intake vary by level of acculturation or Hispanic ethnicity?

HCHS/SOL main cohort: N = 16,415, recruited 2008-2011 from the Bronx, Chicago, Miami and San Diego

Male: 40%

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

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

SOLNAS: Calibration sub-study: n = 477

Biomarker: 24 hour urinary potassium was obtained to create calibration equations that correct for the measurement error/bias in self-reported sodium. A subset had repeated measures of biomarker (Mossavar-Rahmani et al 2017)

Regression Calibration

- **Popular method for addressing covariate measurement error**

Suppose:

$$\begin{aligned} Y &= \beta_0 + \beta_x X + \beta_z Z + \varepsilon \quad \text{and} \\ X^* &= \alpha_0 + \alpha_x X + \alpha_z Z + U \end{aligned}$$

Then

$$\begin{aligned} \mathrm{E}[Y|X^*,Z] &= \mathrm{E}_{X|X^*,Z}[\mathrm{E}(Y|X^*,Z)|X] = \mathrm{E}_{X|X^*,Z}[\mathrm{E}(Y|Z,X)] \\ &= \mathrm{E}_{X|X^*,Z}[\beta_0 + \beta_x X + \beta_z Z] \\ &= \beta_0 + \beta_x \mathrm{E}[X|X^*,Z] + \beta_z Z \end{aligned}$$

Conclusion: regress Y on $\mathrm{E}[X|X^*,Z]$ and Z to get right β coefficients.
$\mathrm{E}[X|X^*,Z]$ is referred to as the calibrated exposure

Calibration equations as prediction equations

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

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

Step 1: use X** to Fit Model:

X= b_0 + b_1 X* + b_2 Z1 + b_3 Z2 + …. b_{k+1} Zk + epsilon

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

- $\hat{X}$ = $\hat{b}_0$ + $\hat{b}_1$ X* + $\hat{b}_2$ Z1 + $\hat{b}_3$ Z2 + …. $\hat{b}_{k+1}$ Zk
- The unexplained variance from the calibration equation results in the Berkson error in measure $\hat{X}$
  - X = $\hat{X} + \mathrm{e}$

Predicted values as covariates in a regression

- Regression calibration: Replace unobserved X with predicted value $\hat{X}=E(X|X^*,Z)$ in the outcome model
- Berkson error in a covariate will not bias a linear regression coefficient (so long as prediction equation correct, independent error)
- Approximation if non linear outcome model

Many common and underappreciated pitfalls when applying regression calibration

- Standard errors still need to be adjusted to account for extra uncertainty
- Prediction model needs all covariates in outcome model to avoid bias
- Extra covariates can be included in prediction model if not correlated with outcome given the truth
- Special considerations when calibration model covariate is a mediator

Simulation Study: Regression Calibration in logistic regression

Variables	Parameter values
$X^* = a_0 + a_1X + a_2Z + a_3V + e$	$a_0 = 0.4$, $a_1 = 0.5$, $a_2 = 0.5$, $a_3 = 0.2$; $\sigma_e^2 = 0.49$
$X^{**} = X + d$	$\sigma_d^2 = 0.7$
(X,Z,V)	Multivariate normal $\mu_X = \mu_Z = \mu_V = 0$ $\sigma_X^2 = \sigma_Z^2 = \sigma_V^2 = 1$ cor(X,Z)= cor(X,V)=cor(Z,V) =0.5
$\text{Logit}(Y) = b_0 + b_1X + b_2Z + b_3V$	$b_0 = -1.0$, $b_1 = \log(1.5)$, $b_2 = -\log(1.3)$, $b_3 = \log(1.75)$

Numerical Study

- Results from 1000 simulations of regression calibration:
- Cohort N=2500; calibration substudy n=250

Method	Mean	% Bias	Empirical standard error	Average estimated standard error	Coverage probability
Model-based	0.407	0.3	0.136	0.113	0.915
Bootstrap-based				0.140	0.954

Underappreciated bias when models not aligned

Method	Mean	Empirical Standard Error of Mean	% Bias
Naïve regression	0.201	0.057	-50.3
Correct RC model	0.407	0.136	0.3
RC, Non-aligned outcome model	0.912	0.194	125.0
RC, Non-aligned calibration model	0.366	0.115	-9.7

Returning to HCHS Example

- Is potassium associated with lower odds of hypertension?
  - For the outcome model: also adjust for potential confounders: age, sex, Hispanic/Latino background, education, income, current smoking, body mass index (BMI)
  - Supplement intake is a useful covariate for the calibration model
  - Recommended approach: include supplement intake into both the outcome and calibration models.

HCHS Analysis: Results

Method of Estimation	OR	95% CI*
Including supplement use in outcome model	0.76	0.60 – 0.96
Omitting supplement use from outcome model	0.90	0.75 – 1.07

Regression with a predicted outcome

Measurement Error Model: Y = $\hat{Y}$ + e

Outcome model of interest: Y = β_xX + β_zZ + u

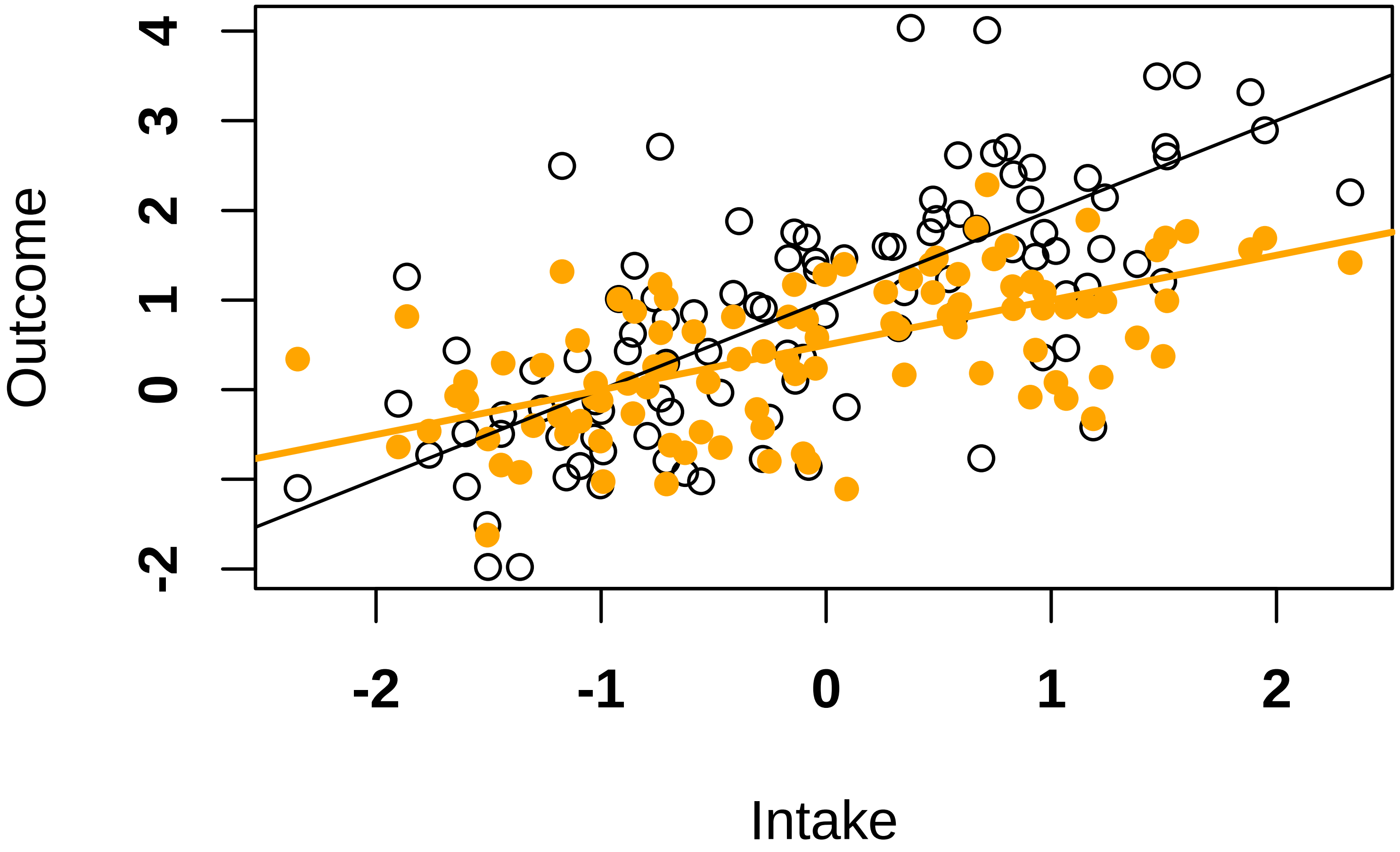
Common Setting: Want to relationship between Y and X, but Y hard to measure. Prediction equation developed in previous study.

Fundamental Question: If fit model $\hat{Y}$ = β_x^*X + β_z^* Z+ v will β_x^*= β_x?

Answer: No

Intuition: $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2$

Impact of Berkson error in the Y



Addressing bias from Berkson error in outcome

$\beta_x^* = \text{var}(Y_{pred})/\text{var}(Y)\ \beta_x$, if non-differential measurement error in $\hat{Y}$, that is if $f(\hat{Y}\ |Y,X,Z) = f(\hat{Y}\ |Y,Z)$

- If non-differential error can apply Buonaccorsi (1991) adjustment:

  $Y_{adj} = (\hat{Y} - \alpha_0)/\ \alpha_1$

  $\alpha_1 = \text{var}(\hat{Y})/\ \text{var}(Y)$ and $\alpha_0 = \mu_{\hat{Y}} - \alpha_1\ \mu_Y$

- Differential error can occur if X or other confounders should have been in prediction model for $\hat{Y}$.
  - For linear regression example can test: $\hat{Y} \perp\!\!\!\perp X\ |\ Y$,Z
  - The challenge: May not have (X,Y,Z) all in same dataset
  - In data example, the objective biomarker M=Y+error can be used to estimate Buonaccorsi coefficients and test this condition
  - For Buonaccorsi adjustment, estimate: var(Y) = var(M) - var(M1-M2)/2

Does sodium intake vary with acculturation (English preference)?

- Calibration (prediction) equation obtained from regression of ln (M) on log 24h recall sodium (24hrK), age, BMI, gender

$$\widehat{\ln(\mathrm{NA})} = 7.268 + 0.136 \ln(24hrK) + 0.001\ age + 0.017\ BMI - 0.274\ I(Female)$$

- Can use M to test non-differentiality condition, equivalent to $\hat{Y} \coprod X \mid Y$,Z in example
  - Check regression of $\hat{Y}$ on X,Y, Z.
  - Since M = Y + error, can perform second regression calibration. Replace Y with E(Y|M,X,Z)= E(M2|M1,X,Z)

Three ways of regressing sodium intake on acculturation (English preference)

1. Regress $\hat{Y}$ on acculturation (biased)
2. Buonaccorsi's correction (unbiased if error is non-differential)
3. Regress M on acculturation (unbiased)

$$E(M|X,Z) = E(Y+e|X,Z) = E(Y|X,Z)$$

- Methods 1 and 2 can be done in full HCHS or SOLNAS
- Method 3 can be done only in SOLNAS

- Method 3 is usually not available – but here the SOLNAS substudy makes it possible.

HCHS/SOL Results
Regression coefficient and SE

	$\hat{Y}$	**Buonaccorsi Adjustment**	**Unbiased**
SOLNAS	0.064 (0.026)	0.166 (0.085)	-0.056 (0.056)
HCHS	0.029 (0.017)	0.075 (0.051)	-----

$\hat{Y}$ method appears biased as expected, and Buonaccorsi adjustment appears to increase the bias!

- Differential error check: regress $\hat{Y}$ on X,Z, E(Y|X,Z)
- Regression Coefficient for X = 0.235; 95%CI = (0.095, 0.711)

Other examples of differential error

Intake	Regression coefficient for English preference	95% CI
Potassium	0.125	(0.016, 0.531)
Protein	0.101	(0.043, 0.174)
Total energy	0.038	(-0.007, 0.083)

Discussion

- **There is increasing use of prediction/calibration equations in medicine**
- **Naïve analyses with predicted outcomes are subject to multiple biases**
  - Regressions reliant on predicted outcomes will have biased coefficients
  - Regressions reliant on predicted values need SE adjustment
  - Distributional summaries are biased, quantiles appear less extreme
- **Prediction model needs to be correct or all bets are off**
    - This includes alignment of outcome and prediction model covariates
- **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

- Forthcoming on ArXiv: Boe L , Shaw PA, Midthune D, Gustafson P, Kipnis V, Park E, Sotres-Alvarez D, Freedman L. Issues in Implementing Regression Calibration Analyses.
- 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.