Goodness-of-Fit Tests for Logistic Regression with Complex Survey Data

Amang Sukasih
Donsig Jang
Haixia Xu

2007 Joint Statistical Meeting
Salt Lake City, UT, July 31, 2007

MATHEMATICA
Policy Research, Inc.

Overview

- Background on weighting adjustment for nonresponse using response propensity scores
- Methodology on goodness-of-fit test for response propensity model
- Simulation study to investigate model misspecification
- Conclusion

Background

- Weighting adjustment

$$\hat{\theta}_A = \sum_{i \in R} \frac{w_i}{a_i} Y_i$$

Calculating adjustment factor:

$$a_{i(c)} = \sum_{i \in R_c} w_{i(c)} \Big/ \sum_{i \in S_c} w_{i(c)}$$

$$a_i = \hat{\pi}_i, \quad \log\left[\pi_i / \left(1 - \pi_i\right)\right] = X_i \beta$$

Background

- Weighting adjustment for nonresponse using response propensity model
  - individual estimate of propensity score as adjustment factor
  - weighting cells based on estimated propensity scores
- Both techniques depend on estimates of propensity scores
- Estimated propensity scores depend on goodness-of-fit of model

Background

- Goodness-of-fit test:
  - Hosmer-Lemeshow (H-L) test for simple random samples, available in SAS (unweighted)
  - for complex samples, available in SUDAAN and STATA (design-based, different in rejection regions)
- Effect of model misspecification
  - goodness-of-fit test
  - distribution of propensity scores
  - weighting cells

Methodology

- Goodness-of-fit test

  H_0: model fit is good
  H_1: model fit is not good

- Test for simple random sample

$$HL = \sum_{c=1}^{G} \frac{\left(O_c - N_c \bar{\pi}_c\right)^2}{N_c \bar{\pi}_c \left(1 - \bar{\pi}_c\right)} \sim \chi^2_{(G-2)}$$

Methodology

- Test for complex sample

$$\hat{\theta} = (O_1 - E_1, \cdots, O_G - E_G)$$

$$O_c - E_c = \sum_{i \in S_c} w_i (R_i - \hat{\pi}_i)$$

(SUDAAN)

$$O_c - E_c = \sum_{i \in S_c} w_i (R_i - \hat{\pi}_i) \Big/ \sum_{i \in S_c} w_i$$

(STATA)

$$Q = \hat{\theta}' \left[\text{var}(\hat{\theta}) \right]^{-1} \hat{\theta}$$

Methodology

- Rejection regions
  - SUDAAN Chi-square $Q \sim \chi^2_{(G-2)}$
  - SUDAAN Satterthwaite Adjusted *F*

$$\frac{G}{\bar{\lambda}(1+a^2)^2} Q^* \sim F^{G/(1+a^2)}_{K-G+2}, \quad \text{where } K = \#\text{PSUs} - \#\text{strata}$$

  - SUDAAN Wald *F* $\frac{Q}{G} \sim F^{G-1}_{K}$
  - STATA Wald *F* $\frac{K-G+2}{GK} Q \sim F^{G}_{K-G+2}$

Simulation Study

- Goals
  - to understand available GOF tests for complex survey data
  - to understand effect of model misspecification on GOF test
  - to understand effect of model misspecification on weighting adjustment

Simulation Study

- Data generation
  - 5 covariates (categorical): *X, Y, Z, U, V*
  - *X, Y, U, V* are independent; *Z* is highly correlated with *Y* but not with others

Covariate	Value	Probability
X	1, 2, 3	.5, .3, .2
Y	1, 2, 3, 4	.25, .25, .25, .25
Z	1, 2, 3, 4, 5	*Y* + Bernoulli(.5)
U	1, 2	.6, .4
V	1, 2, 3	.65, .25, .1

Simulation Study

- Data generation (continued)
  - Regression coefficients

$$\log\left[\pi/(1-\pi)\right] = \beta_0 + \beta_1 X + \beta_2 Y + \beta_3 Z + \beta_4 U + \beta_5 V$$

$$\beta_0 = -6,\ \beta_1 = 2.5,\ \beta_2 = 0.05,\ \beta_3 = 0,\ \beta_4 = 1.5,\ \beta_5 = -1.5$$

  - Response propensity and respondent indicator

$$\pi = \text{response propensity} = \frac{e^{X\beta}}{1+e^{X\beta}}$$

$$R = \text{respondent indicator} \sim Bernoulli(\pi)$$

Simulation Study

- Data generation (continued)
  - Sampling strata (*S*): Corr(*S*,*X*) is high
  - Sampling weight (*W*) depends on *X*

X	*W*
1	66
2	79
3	406

  - Sample size = 50,000; replicates = 1,000

Simulation Study

- Model misspecification
  - true model: $$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X + \beta_2 Y + \beta_4 U + \beta_5 V$$
  - overspecified model: $$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X + \beta_2 Y + \beta_3 Z + \beta_4 U + \beta_5 V$$
  - underspecified model: $$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_2 Y + \beta_4 U + \beta_5 V$$

Simulation Study

- Evaluation:
  - estimates of regression coefficients
  - goodness-of-fit test's *p*-values
  - response propensity score distribution and weighting cells construction
- Software: SAS, SUDAAN, STATA

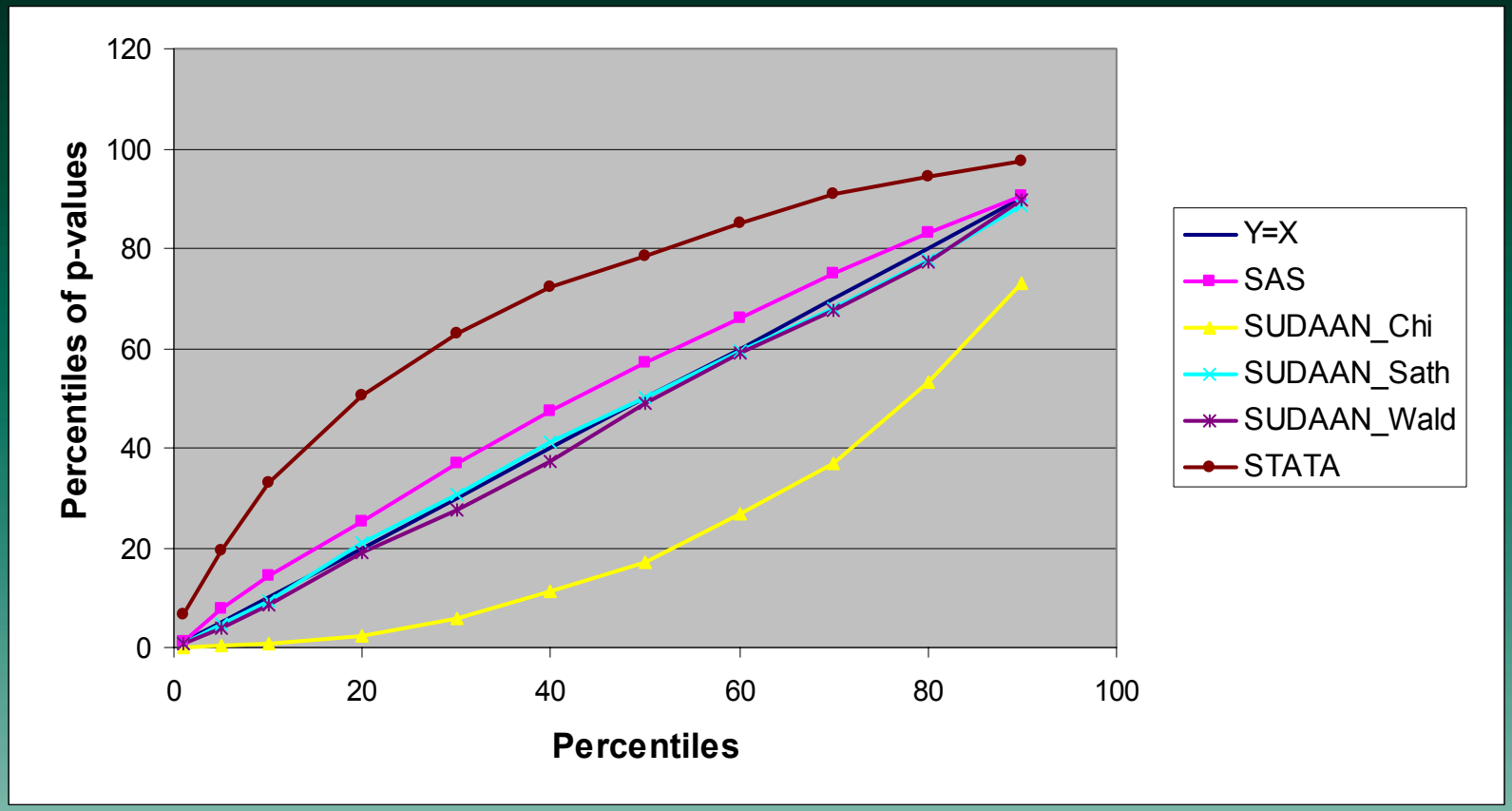
Simulation Study

Means of Regression Coefficients

BETA	True coefficient	True model			Overspecified model			Underspecified model		
		SAS	SUDAAN	STATA	SAS	SUDAAN	STATA	SAS	SUDAAN	STATA
Beta_0	-3.5	-3.50163	-3.50112	-3.50101	-3.50178	-3.50137	-3.50137	-2.5162	-1.90706	-1.90706
X = 2	1	0.9995	0.99938	0.99938	0.99964	0.99963	0.99963			
X = 3	2	2.00074	2.00052	2.00052	2.00097	2.00095	2.00095			
Y = 2	0.5	0.50164	0.50003	0.50003	0.50749	0.51145	0.51145	0.44943	0.44756	0.44756
Y = 3	1	1.00052	0.99913	0.99913	1.00899	1.01524	1.01524	0.90045	0.8982	0.8982
Y = 4	1.5	1.49982	1.49933	1.49933	1.50556	1.51638	1.51638	1.35128	1.34807	1.34807
Z = 2	0				-0.00591	-0.01161	-0.01161			
Z = 3	0				-0.00877	-0.01639	-0.01639			
Z = 4	0				-0.00563	-0.01697	-0.01697			
Z = 5	0				-0.00963	-0.02295	-0.02295			
U = 2	2.5	2.50145	2.50128	2.50128	2.50173	2.50185	2.50185	2.23122	2.22006	2.22006
V = 2	-1.5	-1.50051	-1.49863	-1.49863	-1.50068	-1.49895	-1.49895	-1.34454	-1.33848	-1.33848
V = 3	-3	-3.00051	-2.9975	-2.99746	-3.00082	-2.99822	-2.99822	-2.70838	-2.70305	-2.70305

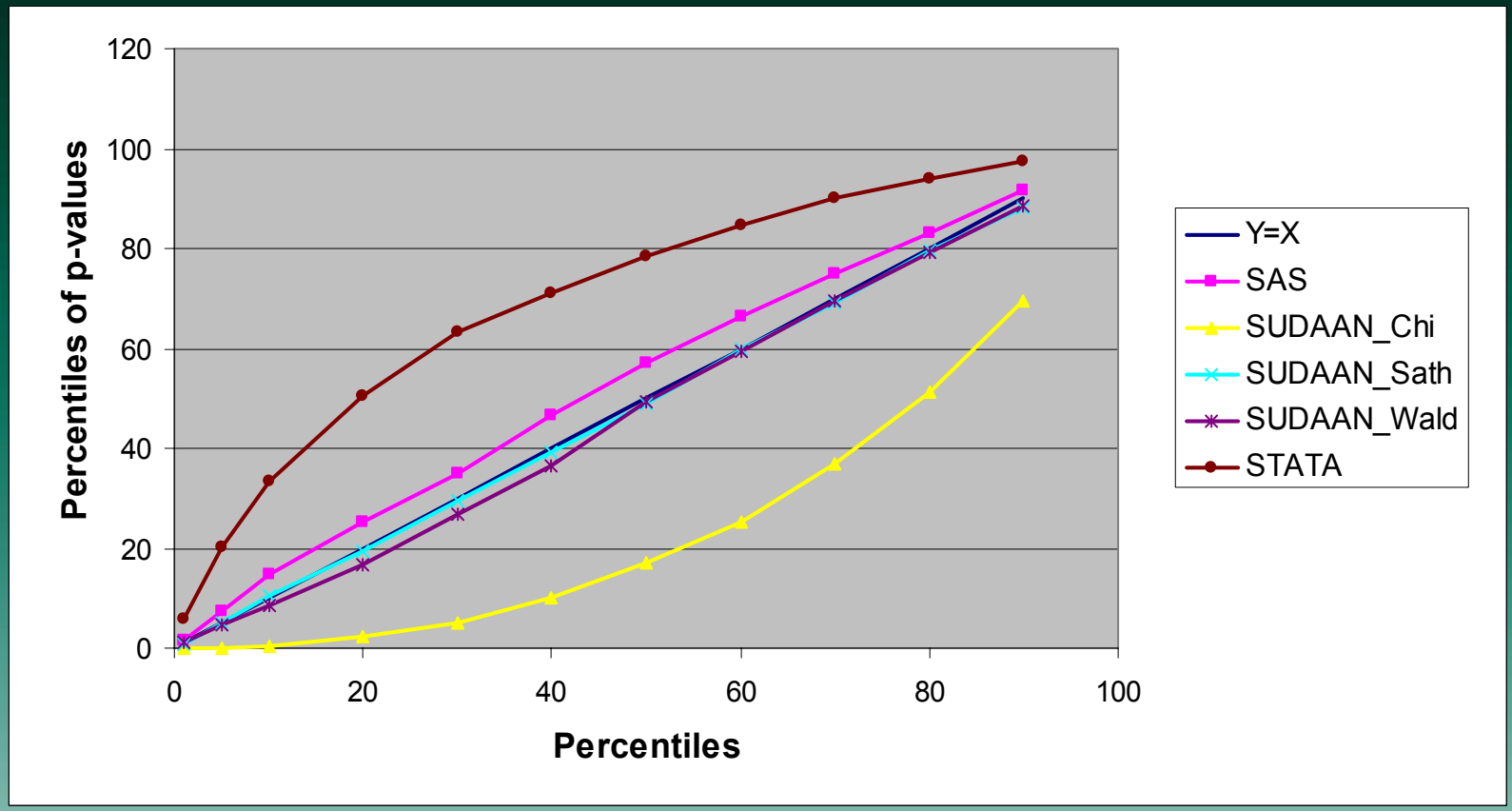
Simulation Study

Percentiles for *p*-values of True Model



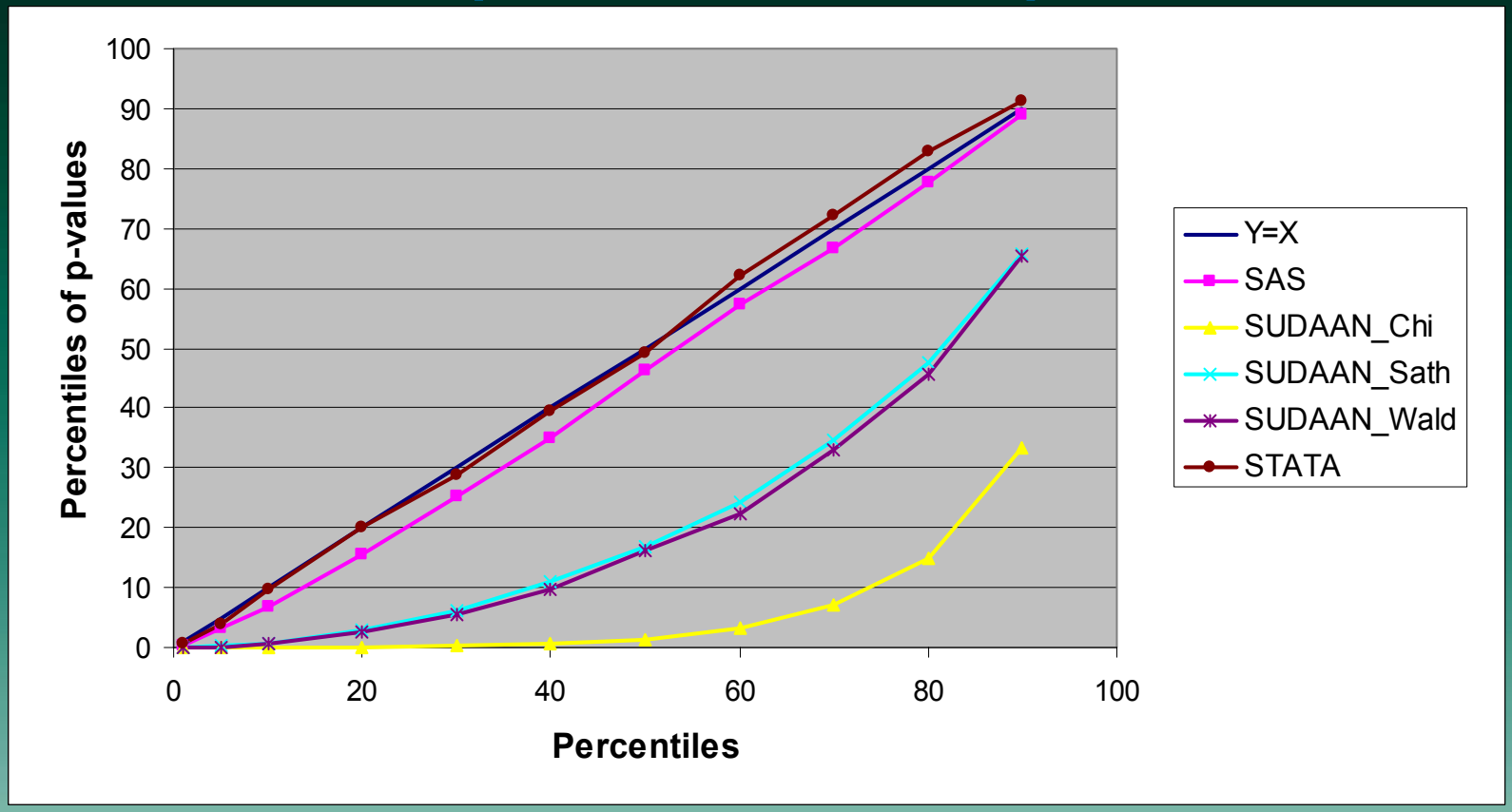
Simulation Study

Percentiles for *p*-values of Overspecified Model



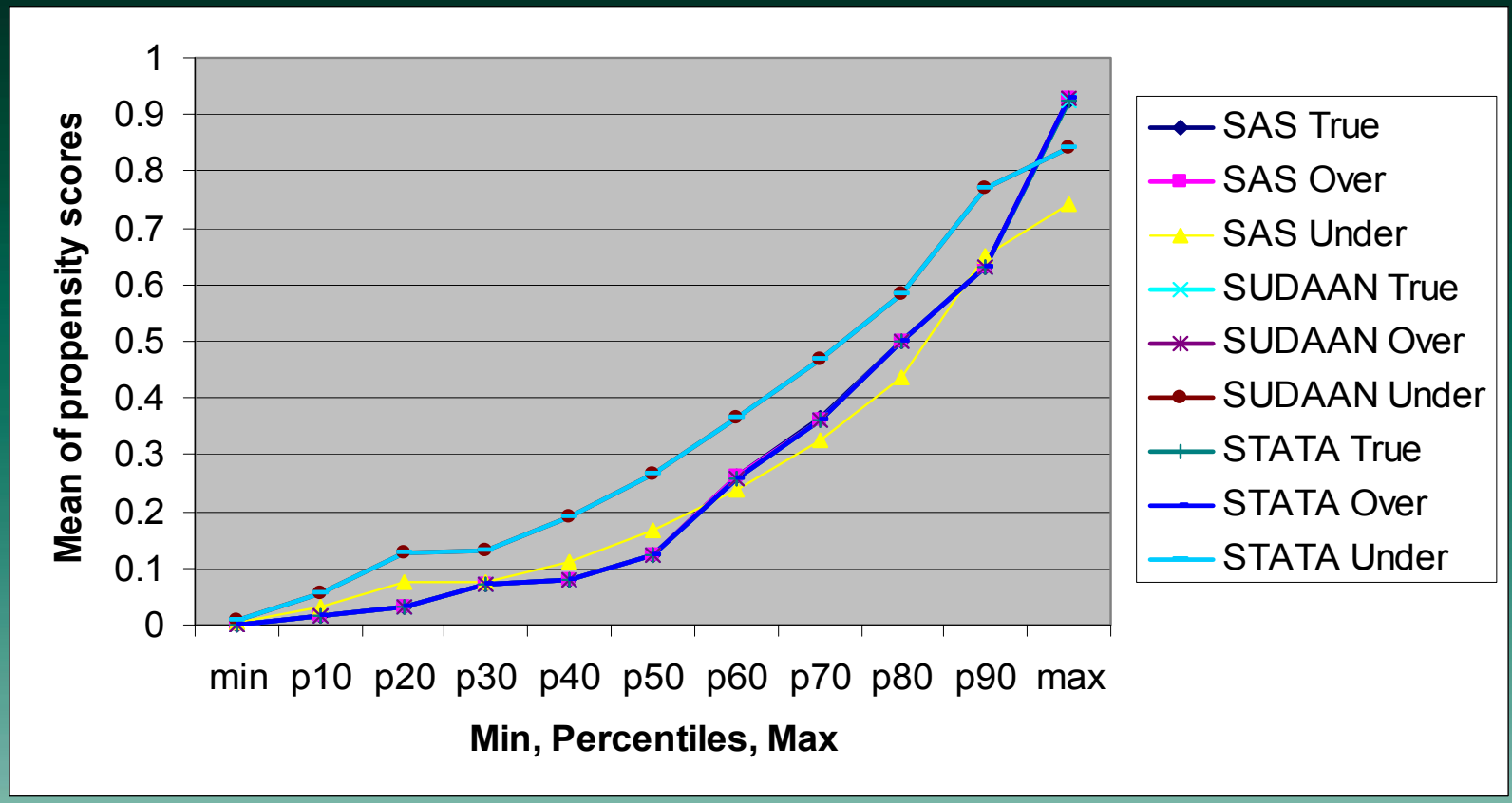
Simulation Study

Percentiles for *p*-values of Underspecified Model



Simulation Study

Means of Propensity Score Percentiles



Conclusion

- Simple random sample H-L and STATA GOF tests overstate the true *p*-value
- Chi-square rejection region works poorly for GOF test for complex survey data
- Underspecified model can be detected by low *p*-value in the GOF test
- Minor model overspecification does not affect distribution of propensity scores, nor the weighting adjustment factors