A Big Simulation Study[1]

STA2053 Fall 2022

[1] See last slide for copyright information.

Design

A big simulation study(Brunner and Austin, 2009) with six factors:

- Sample size: $n = 50$, 100, 250, 500, 1000
- $Corr(X_1, X_2)$: $\phi_{12} = 0.00$, 0.25, 0.75, 0.80, 0.90
- Proportion of variance in Y explained by X_1: 0.25, 0.50, 0.75
- Reliability of W_1: 0.50, 0.75, 0.80, 0.90, 0.95
- Reliability of W_2: 0.50, 0.75, 0.80, 0.90, 0.95
- Distribution of latent variables and error terms: Normal, Uniform, t, Pareto.

There were $5 \times 5 \times 3 \times 5 \times 5 \times 4 = 7{,}500$ treatment combinations.

Simulation study procedure

Within each of the $5 \times 5 \times 3 \times 5 \times 5 \times 4 = 7{,}500$ treatment combinations,

- 10,000 random data sets were generated
- For a total of 75 million data sets
- All generated according to the true model, with $\beta_2 = 0$.
- Fit naive model, test $H_0 : \beta_2 = 0$ at $\alpha = 0.05$.
- Proportion of times H_0 is rejected is a Monte Carlo estimate of the Type I Error Probability.
- It should be around 0.05.

Look at a small part of the results

- Both reliabilities = 0.90
- Everything is normally distributed
- $\beta_0 = 1$, $\beta_1 = 1$ and of course $\beta_2 = 0$.

Table 1 of Brunner and Austin (2009, p.39)

Canadian Journal of Statistics, Vol. 37, Pages 33-46, Used without permission

TABLE 1: Estimated Type I error rates when independent variables and measurement errors are all normal, and reliability of W_1 and W_2 both equal 0.90.

	Correlation between X_1 and X_2				
N	0.0	0.2	0.4	0.6	0.8
25% of variance in Y is explained by X_1					
50	0.0476†	0.0505†	0.0636	0.0715	0.0913
100	0.0504†	0.0521†	0.0834	0.0940	0.1294
250	0.0467†	0.0533†	0.1402	0.1624	0.2544
500	0.0468†	0.0595†	0.2300	0.2892	0.4649
1,000	0.0505†	0.0734	0.4094	0.5057	0.7431
50% of variance in Y is explained by X_1					
50	0.0460†	0.0520†	0.0963	0.1106	0.1633
100	0.0535†	0.0569†	0.1461	0.1857	0.2837
250	0.0483†	0.0625	0.3068	0.3731	0.5864
500	0.0515†	0.0780	0.5323	0.6488	0.8837
1,000	0.0481†	0.1185	0.8273	0.9088	0.9907
75% of variance in Y is explained by X_1					
50	0.0485†	0.0579†	0.1727	0.2089	0.3442
100	0.0541†	0.0679	0.3101	0.3785	0.6031
250	0.0479†	0.0856	0.6450	0.7523	0.9434
500	0.0445†	0.1323	0.9109	0.9635	0.9992
1,000	0.0522†	0.2179	0.9959	0.9998	1.00000

†Not significantly different from 0.05, Bonferroni corrected for 7,500 tests.

Marginal Mean Type I Error Probabilities

Base Distribution

normal	Pareto	t Distr	uniform
0.38692448	0.36903077	0.38312245	0.38752571

Explained Variance

0.25	0.50	0.75
0.27330660	0.38473364	0.48691232

Correlation between Latent Independent Variables

0.00	0.25	0.75	0.80	0.90
0.05004853	0.16604247	0.51544093	0.55050700	0.62621533

Sample Size n

50	100	250	500	1000
0.19081740	0.27437227	0.39457933	0.48335707	0.56512820

Reliability of W_1

0.50	0.75	0.80	0.90	0.95
0.60637233	0.46983147	0.42065313	0.26685820	0.14453913

Reliability of W_2

0.50	0.75	0.80	0.90	0.95
0.30807933	0.37506733	0.38752793	0.41254800	0.42503167

Poison

- The poison combination is measurement error in the variable for which you are "controlling," and correlation between latent explanatory variables.
- As the sample size increases, the problem gets worse
- For a large enough sample size, no amount of measurement error in the explanatory variables is safe, assuming that the latent explanatory variables are correlated.

Other kinds of regression, other kinds of measurement error

- Logistic regression
- Proportional hazards regression in survival analysis
- Log-linear models: Test of conditional independence in the presence of classification error
- Median splits
- Even converting X_1 to ranks inflates Type I Error probability.

Moral of the story

Use models that allow for measurement error in the explanatory variables.

Copyright Information