We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter 11

Bayesian Two-Stage Robust Causal Modeling with Instrumental Variables using Student's t Distributions

Dingjing Shi and Xin Tong

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.70393

Abstract

In causal inference research, the issue of the treatment endogeneity is commonly addressed using the two-stage least squares (2SLS) modeling with instrumental variables (IVs), where the local average treatment effect (LATE) is the causal effect of interest. Because practical data are usually heavy tailed or contain outliers, using traditional 2SLS modeling based on normality assumptions may result in inefficient or even biased LATE estimate. This study proposes four types of Bayesian two-stage robust causal models with IVs to model normal and nonnormal data, and evaluates the performance of the four types of models with IVs. The Monte Carlo simulation results show that the Bayesian two-stage robust causal modeling produces reliable parameter estimates and model fits. Particularly, in different types of the two-stage robust models with IVs, the models that take outliers into consideration and use Student's t distributions in the second stage to model heavy-tailed data or data containing outliers provide more accurate and efficient LATE estimates and better model fits than other distribution models when data are contaminated. The preferred models are recommended to be adopted in general in the two-stage causal modeling with IVs.

Keywords: Bayesian methods, two-stage causal modeling with instrumental variables, nonnormal data, robust method using Student's t distributions

1. Introduction

Causal inference and experimental researchers are often interested in the average treatment effect (ATE), measured by the outcome difference between participants who are assigned to the treatment and those being assigned to the control. The estimation of ATE for the whole population is neither reliable nor feasible when certain conditions are not achieved or assumptions are violated [6, 9]. The treatment effects for only a subset of participants is instead estimated, which is called the local average treatment effect (LATE) [2, 13]. Different studies may have different

LATEs, depending on the subgroup of interest. Often the subgroup of interest is those who have been assigned to the treatment and have actually received the treatment [3]. One way to estimate the LATE is to incorporate instrumental variables (IVs), which are correlated with both the endogenous regressors and error terms when the linearity assumption of the traditional linear models is violated and the endogenous regressors are correlated with the errors. Instrumental variables are incorporated in the analysis to estimate the LATE, or a part of the treatment effect whose estimation is not contaminated by the violation of the linearity assumption.

Two-stage least squares (2SLS) modeling [1] is widely used to estimate the LATE with IVs. In the first stage, IVs are used to predict the partial treatment effect that can be explained by the variations of IVs, and in the second stage, the fitted treatment values are used to predict the experimental outcome, and to estimate the LATE. In estimating the LATE in traditional 2SLS modeling with IVs, it is typically assumed that the measurement errors at both stages are normally distributed. However, practical data in social and behavioral research usually violate the normality assumption and often have heavy tails or contain outliers [25]. Failure to take the nonnormal data into consideration but instead treating the heavy-tailed data or data containing outliers as if they were normally distributed may result in unreliable parameter estimates and inflated type I error rates [35, 38–40], which will eventually lead to misleading statistical inference.

Routine methods to accommodate heavy-tailed data or data with outliers include data transformation and data truncation. However, transformed data are often difficult to interpret especially when the raw scores have meaningful scales [17], and the exclusion of outliers may lead to underestimated standard errors and reduced efficiency [14, 32]. Alternatively, different robust procedures have been developed to provide reliable parameter estimates, the associated standard errors, and statistical tests. The rationale of most robust procedures is to weigh each observation according to its distance from the center of the majority of the data, so that outliers that are far from the center of the data are downweighted [10, 11, 37]. In recent research, more and more robust methods have been used to estimate complex models, such as linear and generalized linear mixed-effects models [19, 26], structural equation models [15, 31], and hierarchical linear and nonlinear models [20, 29].

Over the past decades, robust procedures based on Student's t distributions have been developed and advanced to model heavy-tailed data or data containing outliers [14, 33]. For example, Student's t distributions have been applied under the structural equation modeling framework and were found to produce reliable parameter estimates and inferences [15, 16]; in robust mixture models, Wang et al. [30] used the multivariate t distribution to fit heavy-tailed data and data with missing information, Shoham [24] implemented a robust clustering algorithm in mixture models by modeling data that are contaminated by outliers using multivariate t distributions, Seltzer et al. [21] and Seltzer and Choi [22] conducted sensitivity analysis employing Student's t distributions in robust multilevel models and downweighted outliers in level two (the between-subject level), and Tong and Zhang [28] and Zhang et al. [36] advanced the Student's t distributions to robust growth curve models and provided online software to

carry out the analysis. Although robust methods based on Student's t distributions have been used in different modeling frameworks, few have been adopted in the causal modeling, where heavy-tailed data or data containing outliers are not uncommon [18].

Recently, Shi and Tong [23] implemented a robust Bayesian estimation method using Student's t distributions to the two-stage causal modeling with IVs to fit data that contain outliers or are normally distributed concurrently at both stages. However, in the two-stage causal models with IVs, data at either stage are equally likely having outliers or are nonnormally distributed. Previous studies have noticed such a situation. For example, Pinheiro et al. [19] used a robust estimation to the linear mixed-effects model and applied the multivariate t distribution to both the random effects and intraindividual errors simultaneously. Tong and Zhang [28] conducted a robust estimation to growth curve modeling and modeled the measurement errors and random effects separately with t distributions or normal distributions rather than the same distribution for the two effects. Therefore, this article extends the study of Shi and Tong [23] and proposes four possible types of two-stage causal models with IVs to the data. The study evaluates the performance of the robust method in four types of models. In the following section, the robust method based on Student's t distributions is reviewed. Then, the two-stage causal models with IVs, the associated LATE, and the corresponding four types of models are introduced. Next, a Monte Carlo simulation study is conducted to evaluate the performance of the robust method in four possible types of two-stage causal models with IVs. In the end, conclusions are summarized and discussions are provided.

2. Robust methods based on Student's t distributions

As a robust procedure, the fundamental idea of using Student's t distributions to model heavy-tailed data or data containing outliers is to assign a weight to each case and properly downweight cases that are far from the center of the majority of the data [10, 11, 37]. Suppose a population of k random variables, $\mathbf{y}$, follow a multivariate t distribution, with mean vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Psi}$, and degrees of freedom ν, denoted by t ($\boldsymbol{\mu}$, $\boldsymbol{\Psi}$, ν). The probability density function of $\mathbf{y}$ can be expressed as:

$$p(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Psi}, \nu) = \frac{|\boldsymbol{\Psi}|^{-\frac{1}{2}}\Gamma\left(\frac{\nu+k}{2}\right)}{\left(\Gamma\left(\frac{1}{2}\right)^{k}\Gamma\left(\frac{\nu}{2}\right)\nu^{\frac{k}{2}}\right)} \times \left(1 + \frac{(\boldsymbol{y}-\boldsymbol{\mu})^{T}\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{(\nu+k)}{2}}. \quad (1)$$

The maximum likelihood estimates of model parameters under the model with t distribution assumptions satisfy

$$\Sigma_{i=1}^{n} w_i \boldsymbol{A}_i \boldsymbol{\Psi}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}) = 0, \quad (2)$$

where n is the total sample size, $\boldsymbol{y_i}$ is a sample from $\boldsymbol{y}$, $\boldsymbol{A_i}$ is the partial derivatives of $\boldsymbol{\mu}$, and

$$w_i = \frac{\nu + \tau_i}{\nu + \sigma_i^2} \tag{3}$$

is the weight assigned to case i. In the equation for w_i, τ_i is the dimension of the parameter for each i and σ_i^2 is the squared Mahalanobis distances $\sigma_i^2 = \left(\boldsymbol{y}_i - \boldsymbol{\mu}\right)^T \boldsymbol{\Psi}^{-1} \left(\boldsymbol{y}_i - \boldsymbol{\mu}\right)$. Note that $(\boldsymbol{y}_i - \boldsymbol{\mu})$ is the distance between each observation and the population mean, and a large $(\boldsymbol{y}_i - \boldsymbol{\mu})$ indicates a potential outlier as well as a large squared Mahalanobis distance σ_i^2. The outliers are downweighted in the analysis because the weight w_i decreases with increasing squared Mahalanobis distances σ_i^2, given fixed degrees of freedom ν, and dimensions τ_i [14].

The shape of a t distribution is controlled by its degrees of freedom ν, and ν can be set a priori or estimated in the analysis. Under certain conditions, the degrees of freedom have been recommended setting a priori. Lange et al. [14] and Zhang et al. [36] suggested fixing the value for the degrees of freedom of Student's t distributions when sample size is small, as small sample sizes could lead to biased degrees of freedom estimate. Moreover, Tong and Zhang [28] argued that by fixing the degrees of freedom, more accurate parameter estimates and credible intervals can be obtained when model specification is built on solid substantive theories. In contrast, estimating the degrees of freedom can make the model more flexible. When the degrees of freedom ν are freely estimated, Student's t distributions have an additional parameter ν, compared with normal distributions. As the degrees of freedom ν increase, the Student's t distribution approaches a normal distribution.

There are several advantages in using Student's t distributions for robust data analysis [28]. First, unlike the nonparametric robust analysis, Student's t distributions have parametric forms, and inferences based on them can be carried out relatively easily through maximum likelihood estimation or Bayesian estimation methods. Second, the degrees of freedom of Student's t distributions control the weight of outliers and can flexibly set a priori or be estimated. Third, when data have heavy tails or contain outliers, considering Student's t distribution as a natural extension of the normal distribution is rather intuitive.

3. Bayesian two-stage robust causal modeling with IVs

In causal Ordinary Least Squares (OLS) regression, when the error terms are related to some regressors, the estimated ATE is biased due to the violation of the linearity assumption. Variables that are related to both endogenous regressors and errors are used as instruments to differentiate the correlations between endogenous regressors and errors, leaving only a part of the treatment effects that have not been contaminated by the violation of the linearity assumption to be estimated, and such variables are called instrumental variables (IVs). The ATE of interest becomes the LATE of interest. For example, Currie and Yelowitz [8] studied the effect of public housing voucher program of having a larger housing unit on housing quality and educational attainment. Based on the fact that some families in voucher program tradeoff physical housing amenities and reductions in rental payments that are bad and have negative effects for the housing quality and their children, some regressors are correlated with errors and become endogenous. Previous theory supports that a household having an extra number of kids is

entitled to a larger housing unit, whether there are extra kids in the household and the sex decomposition of the extra kids are chosen as the IVs, to study the voucher program effect to participants who have one girl and one boy (i.e., having sex decomposition) in the household. It was found that the voucher program participants who have the sex decomposition in the household are more likely to have better housing quality and educational attainment. The example shows that when IVs are introduced, the external validity is traded for the improvement of the internal validity, and the ATE (i.e., all of the voucher program participants) becomes the LATE (i.e., program participants who have extra kids and who have the sex decomposition).

One commonly used framework to estimate LATE is the 2SLS modeling with IVs. Let d_i and y_i be the treatment and the outcome for individual i, respectively, and $\boldsymbol{Z}_i = (z_{i1}, \ldots, z_{ij})'$ be a vector of instrumental variables for individual i ($i = 1, \ldots, N$). Here, N is the sample size and J is the total number of instrumental variables. In the first stage of the 2SLS model, the IVs **Z** are used to predict the treatment **d**. In other words, the portion of variations in the treatment **d** is identified and estimated by the IVs **Z**; and then the second stage relies on the estimated exogenous portion of treatment variations in the form of the predicted treatment values to estimate the treatment effect on the outcome **y**. A typical form of the 2SLS model with IVs can be expressed as:

$$d_i = \pi_{10} + \boldsymbol{\pi}_{11}\boldsymbol{Z}_i + e_{1i}, \tag{4}$$

$$y_i = \pi_{20} + \pi_{21}\widehat{d}_i + e_{2i}, \tag{5}$$

where π_{10} and $\boldsymbol{\pi_{11}} = (\pi_{11}, \ldots, \pi_{1J})'$ are the intercept and regression coefficients for the linear model where the treatment **d** is regressed on the IVs **Z**, respectively; and π_{20} and π_{21} are the intercept and slope for the linear model where the outcome y is regressed on the predicted treatment values of $\widehat{\boldsymbol{d}}$, respectively. The IVs help estimate the treatment effects in which the causal effect of IVs on the treatment is first estimated in Eq. (4), and the causal effect of this estimated partial treatment effect on the outcome is then estimated in Eq. (5). From the model, $\boldsymbol{\pi}_{11}$ is the causal effect of the IVs **Z** on the treatment Z, and $\boldsymbol{\pi}_{21}$ is the treatment effect on the outcome **y** for a subset of participants whose treatment effect has been partialled out and explained by the IVs **Z**. π_{21} is the causal effect of interest and is called LATE. There are several advantages in using 2SLS modeling to estimate LATE. First, unlike method of point estimate such as Wald estimator [4], 2SLS modeling also provides standard error estimate and confidence intervals of the LATE, making statistical inferences more efficient. Second, when 2SLS models are used, covariates could be controlled simultaneously at both stages of the 2SLS model when the effect of **Z** on **d** and the effect of $\widehat{d}$ on **y** are estimated. Mathematically, the estimated LATE $\widehat{\pi}_{21}$ in 2SLS can be derived as:

$$\widehat{\pi}_{21} = \frac{cov\left(y_i, \widehat{d}_i\right)}{var\left(\widehat{d}_i\right)} = \frac{cov\left(y_i, \widehat{\pi}_{10} + \widehat{\pi}_{11}z_{i1} + \cdots + \widehat{\pi}_{1J}z_{iJ}\right)}{var\left(\widehat{\pi}_{10} + \widehat{\pi}_{11}z_{i1} + \cdots + \widehat{\pi}_{1J}z_{iJ}\right)} = \frac{\widehat{\pi}_{11}cov\left(y_i, z_{i1}\right) + \cdots + \widehat{\pi}_{1J}cov\left(y_i, z_{iJ}\right)}{\widehat{\pi}^2{}_{11}var(z_{i1}) + \cdots + \widehat{\pi}^2{}_{11}var(z_{i1})}. \tag{6}$$

Traditional causal 2SLS models with IVs are commonly estimated using OLS methods or maximum likelihood estimation from the frequentist approach. The measurement errors at both stages, e_{1i} and e_{2i}, are assumed to be normally distributed as $e_{1i} \sim N(0, \sigma^2_{e_1})$ and

$e_{2i} \sim N(0, \sigma^2_{e_2})$. Because practical data usually violate the normality assumption, it was proposed from a Bayesian approach that the normal distributions can be replaced by Student's t distributions for heavy-tailed data or data containing outliers [23, 28, 36]. In the two-stage causal model with IVs, data at either stage are equally likely to be nonnormal or containing outliers. Therefore, we propose four possible types of Bayesian two-stage causal models to data with (a) normal measurement errors at both stages, denoted as *Bayesian normal model*, (b) t measurement errors in the first stage and normal measurement errors in the second stage, denoted as *Bayesian nonnormal-s1 model*, (c) normal measurement errors in the first stage and t measurement errors in the second stage, denoted as *Bayesian nonnormal-s2 model*, and (d) t measurement errors at both stages, denoted as *Bayesian nonnormal-both model*. The four types of Bayesian two-stage causal models have the same mathematical model expressions as those from the frequentist approach. Namely, for the Bayesian normal model, measurement errors are assumed to be distributed as $e_{1i} \sim N(0, \sigma^2_{e_1})$ and $e_{2i} \sim N(0, \sigma^2_{e_2})$; for the Bayesian nonnormal-s1 model, the measurement errors are assumed to be distributed as $e_{1i} \sim t(0, \sigma^2_{e_1}, \nu_1)$ and $e_{2i} \sim N(0, \sigma^2_{e_2})$; for the Bayesian nonnormal-s2 model, the measuremenet errors are assumed to be distributed as $e_{1i} \sim N(0, \sigma^2_{e_1})$ and $e_{2i} \sim t(0, \sigma^2_{e_2}, \nu_2)$; finally, for the Bayesian nonnormal-both model, the measurement errors are assumed to be distributed as $e_{1i} \sim t(0, \sigma^2_{e_1}, \nu_1)$ and $e_{2i} \sim t(0, \sigma^2_{e_2}, \nu_2)$. All four types of models are estimated using Bayesian methods.

In the Bayesian approach, we obtain the joint posterior distributions of the parameters based on the prior distributions of the parameters and the likelihood of the data information. Making statistical inferences directly from the joint posterior distributions is usually difficult. Gibbs sampling, a Markov chain Monte Carlo (MCMC) method is a widely used algorithm to draw a sequence of samples from the joint posterior distribution of two or more random variables, given that the conditional posterior distributions of the model parameters can be obtained [7]. In specific, Gibbs sampling alternately samples parameters one at a time from their conditional posterior distribution on the current values of other parameters, which are treated as known. After a sufficient number of iterations, the sequence of samples constitutes a Markov chain that converges to a stationary distribution. This stationary distribution is the sought-after joint posterior distribution of the parameters [12].

The Gibbs sampling algorithm is used to obtain the LATE estimate for the two-stage causal model with IVs. Because the t distribution can be viewed as a normal distribution with variance weighted by a Gamma distribution, the data augmentation method is used here to simplify the posterior distribution. Specifically, a Gamma random variable ω is augmented with a normal random variable because if $\omega_i \sim G\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$, and $\boldsymbol{y}_i | \omega_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Psi}/\omega_i)$, then $\boldsymbol{y}_i \sim t(\boldsymbol{\mu}, \boldsymbol{\Psi}, \nu)$. The detailed steps of the Gibbs sampling algorithm for the Bayesian nonnormal-s2 model are given below. The Gibbs sampling procedures for the other models are similar.

1. Start with initial values $\boldsymbol{\pi}_1^{(0)}$, $\boldsymbol{\pi}_2^{(0)}$, $\sigma_{e1}^{2\,(0)}$, $\sigma_{e2}^{2\,(0)}$, $\nu^{(0)}$, $\omega_i^{(0)}$, where $\boldsymbol{\pi}_1^{(0)} = \left(\pi_{10}^{(0)}, \boldsymbol{\pi}_{10}^{(0)}\right)'$ and $\boldsymbol{\pi}_2^{(0)} = \left(\pi_{20}^{(0)}, \pi_{21}^{(0)}\right)'$.

2. Assume at the jth iteration, we have $\boldsymbol{\pi}_1^{(j)}$, $\boldsymbol{\pi}_2^{(j)}$, $\sigma_{e1}^{2\,(j)}$, $\sigma_{e2}^{2\,(j)}$, $\nu^{(j)}$, $\omega_i^{(j)}$, where $\boldsymbol{\pi}_1^{(j)} = \left(\pi_{10}^{(j)}, \boldsymbol{\pi}_{10}^{(j)}\right)'$ and $\boldsymbol{\pi}_2^{(j)} = \left(\pi_{20}^{(j)}, \pi_{21}^{(j)}\right)'$.
At the (j+1)th iteration,

3. Step 3
3.1 Sample $\boldsymbol{\pi}_1^{(j+1)}$ from $p\left(\boldsymbol{\pi}_1|\sigma_{e1}^{2\,(j)}, d_i, \boldsymbol{Z}_i, i = 1, \ldots, N\right)$;
3.2 Sample $\sigma_{e1}^{2\,(j+1)}$ from $p\left(\sigma_{e1}^2|\boldsymbol{\pi}_1^{(j+1)}, d_i, \boldsymbol{Z}_i, i = 1, \ldots, N\right)$;
3.3 Sample $\sigma_{e2}^{2\,(j+1)}$ from $p\left(\sigma_{e2}^2|\boldsymbol{\pi}_2^{(j)}, \widehat{d}_{i'}, y_{i'}, \omega_i^{(j)}, i = 1, \ldots, N\right)$;
3.4 Sample $\nu^{(j\,+\,1)}$ from $p\left(\nu|\omega_i^{(j)}, i = 1, \ldots, N\right)$;
3.5 Sample $\omega_i^{(j+1)}, i = 1, \ldots N$, from $p\left(\omega_i|\nu^{(j+1)}, \sigma_{e2}^{2\,(j+1)}, \widehat{d}_{i'}, y_{i'}, \boldsymbol{\pi}_2^{(j)}, i = 1, \ldots, N\right)$;
3.6 Sample $\boldsymbol{\pi}_2^{(j+1)}$ from $p\left(\boldsymbol{\pi}_2|\omega_i^{(j+1)}, \sigma_{e2}^{2\,(j)}, \widehat{d}_{i'}, y_{i'}, i = 1, \ldots, N\right)$.

4. Repeat Step 3.

4. Evaluation of four types of distributional 2SLS models

In this section, the performance of the four types of two-stage robust causal models is evaluated through a Monte Carlo simulation study. Data are generated from a general causal inference model as presented in Eq. (7). Full Bayesian methods are used for the estimation of all four types of two-stage causal models. In specific, noninformative priors are applied to all model parameters, conditional posterior distributions of all model parameters are obtained and Markov chains are generated through Gibbs sampling algorithm, convergence tests are conducted and finally statistical inferences for the model parameters are made. Free software (R Development Core Team, 2011) R [41] and OpenBUGS [42] (Thomas, O'Hara, Ligges, & Sturtz, 2006) were used for the implementation of MCMC algorithms and model estimation. A total of 20,000 iterations was conducted for each simulation condition, with the first 10,000 iterations as the burn-in period.

4.1. Study design

Data are generated from a general causal inference model

$$y_i = 3 + 0.5x_i + e_i, \tag{7}$$

where y_i is the causal outcome, x_i is the causal treatment, and e_i is the measurement error. Three potential influential factors are considered. First, sample size (N) is either 200 or 600. Second, correlation between $\mathbf{x}$ and $\mathbf{e}(\Phi)$ is manipulated to be either 0.3 or 0.7, reflecting relatively weak or strong linear relationship between the treatment and the measurement

error. Third, a proportion of observations that contains outliers is manipulated. The proportion of outliers (OP) is considered to be 0, 5, or 10%. When the OP is 0%, data contain no outliers and measurement errors e_i are normally distributed. When the OP is above zero, data contain outliers. For outliers, the measurement errors are generated from a different normal distribution with the same standard deviation, but a larger mean (eight times of the standard deviation). An IV is also generated from a normal distribution and correlated with **x** with the correlation coefficient being 0.6.

If we fit a linear regression to the generated data, we will immediately notice that the residuals and the regressors are not independent. Therefore, we adopt the two-stage causal model with IVs. The four types of two-stage models (normal model, nonnormal-s1 model, nonnormal-s2 model, and nonnormal-both model) are used to fit the data. In the first stage, the IV is used to predict the endogenous treatment, and the estimated treatment is then used in the second stage to estimate the LATE. Based on Eq. (6), the theoretical LATE is 5/6.

As discussed previously, Bayesian methods using Gibbs sampling algorithm are used to obtain the LATE estimates in four types of two-stage causal models. The bias and standard error (SE) of the LATE estimate for each of the four distributional models are assessed. In addition, the deviance information criterion (DIC) [27] for each condition is examined to study the model fit. A lower value of DIC indicates a better model fit.

4.2. Results

The bias and SEs of the LATE estimates from four types of models when $\varphi = 0.3$ are presented in **Table 1**.

In almost all cases, models that use normal distributions to model the normal data and that use Student's t distributions to model the data with outliers provide the best estimates with smaller bias and SEs among other types of two-stage causal models. For example, when N = 200, the normal model provides smaller bias and SE for normal data; similarly, nonnormal-s2 and nonnormal-both models lead to the smaller bias and SEs when they are used to fit data containing outliers. This shows that using Student's t distributions to model data containing

			Normal model		**Nonnormal-s1 model**		**Nonnormal-s2 model**		**Nonnormal-both model**	
N	**Data**	**OP**	**Bias**	**SE**	**Bias**	**SE**	**Bias**	**SE**	**Bias**	**SE**
200	Normal	0%	0.001	0.154	−0.004	0.154	−0.004	0.155	−0.003	0.155
	Nonnormal	5%	0.210	0.283	0.200	0.281	−0.022	0.177	−0.020	0.171
		10%	0.342	0.379	0.341	0.378	−0.060	0.157	−0.050	0.155
600	Normal	0%	−0.021	0.076	−0.023	0.076	−0.023	0.077	−0.023	0.076
	Nonnormal	5%	0.180	0.168	0.170	0.167	−0.064	0.099	−0.060	0.096
		10%	0.390	0.230	0.380	0.210	−0.077	0.099	0.070	0.098

Table 1. Bias and SEs of the LATE estimates for all the conditions when $\Phi = 0.3$.

outliers is an effective way to accommodate heavy-tailed data or data containing outliers, and this finding is consistent with the previous research [34, 36]. In causal inference study, because practical data at either stage are equally likely having outliers or are normally distributed in the two-stage causal model with IVs, we fit all four types of distributional models and try to decide which one is the best-fitted model. From the results, modeling heavy-tailed data or data containing outliers with nonnormal-both model provides more reliable parameter estimates than traditional methods that ignore the data distributions and model all data exclusively with normal distributions.

Although it is always a good choice to model normal data with normal distributions and heavy-tailed data or data containing outliers with Student's t distributions, in practice, researchers may not know whether the first stage or the second stage of the model should account for the nonnormality. The simulation results show that when data contain outliers, the nonnormal-s2 model and nonnormal-both model that use t distributions in the second stage produce the smallest bias and SEs of the LATE estimates. This is probably because the causal effect of interest, LATE, is housed in the second stage, and using Student's t distribution to model outliers in that stage is effective in capturing the LATE. On the contrary, in the normal model or the nonnormal-s1 model, the normal distribution is being used to model the second stage data that are heavy tailed or contain outliers. For example, for all the nonnormal data that contain outliers (i.e., OP = 5 or 10%), the nonnormal-s2 model and the nonnormal-both model, both of which use t distributions to model data in the second stage, outperform other models, providing smaller bias and SEs of the LATE estimates regardless of sample size (N) and proportion of outliers (OP). Comparing between nonnormal-s2 and nonnormal-both models, the nonnormal-both models perform slightly better than the nonnormal-s2 model does. Take N = 600 and OP = 10% as an example, the bias and SEs for the nonnormal-s2 model are −0.077 and 0.099, whereas those for the nonnormal-both model are slightly smaller to be 0.070 and 0.098, showing that fitting the nonnormal data with Student's t distributions at both stages has the best performance in terms of accuracy and efficiency of the LATE estimate.

Table 2 presents the results for DICs for the four types of two-stage causal models when $\Phi = 0.3$.

In practice, DIC can be used as a model selection criteria. To select the best-fitted parsimonious model, we first fit all four types of models to the data, and then select the model with the

N	Data	OP	Normal model	Nonnormal-s1 model	Nonnormal-s2 model	Nonnormal-both model
200	Normal	0%	1145.09	1145.83	1145.82	1146.54
	Nonnormal	5%	1380.18	1380.82	1241.48	1242.04
		10%	1488.71	1489.43	1315.18	1315.87
600	Normal	0%	3418.20	3419.25	3419.23	3420.53
	Nonnormal	5%	4126.86	4128.00	3705.62	3706.93
		10%	4448.88	4450.07	3922.32	3923.73

Table 2. DICs of all the distributional models when $\Phi = 0.3$.

smallest DIC. Notice that for normal data, all four types of models have similar DIC values. When data contain outliers, nonnormal-s2 and nonnormal-both models provides the smallest DIC, indicating that these types of models fit the data better. In all data conditions in the study, the DICs of the nonnormal-s2 model and the nonnormal-both model are very similar, and either model can be adopted.

The proportions of outliers contained in the data have effect on the performance of the nonnormal-s2 model and the nonnormal-both model. Specifically, the larger the proportions of outliers, the more salient the advantages of the nonnormal-s2 and nonnormal-both models. For example, for the nonnormal data with N = 200 and OP = 5%, the bias from the normal model, the nonnormal-s2 model and the nonnormal-both model is 0.210, −0.022, and −0.020, respectively; when OP becomes 10%, the bias from the normal model jumps to 0.342, whereas the bias from the nonnormal-s2 model changes slightly to −0.060 and that from the nonnormal-both model is −0.050. Similarly, the preferred models provide less biased LATE estimates when sample size is small, and the advantage of the preferred models is more apparent under small sample conditions (e.g., [23]).

When Φ = 0.7, consistent with the results from previous conditions when Φ = 0.3, when data have outliers, using Student's t distributions to model the data provides more accurate and efficient LATE estimates and better model fits than using normal distribution to model the data. The advantage of using t distributions is more obvious when sample size is small and the proportion of outliers is large.

5. Discussion

In causal inference research, the issue of the treatment endogeneity is commonly addressed in the 2SLS model with IVs, where the LATE is the causal effect of interest. Because practical data usually violate the normality assumption, using normal distributions to model heavy-tailed data or data containing outliers may result in inefficient or even biased LATE estimate. In the 2SLS model with IVs, data at either stage are equally likely having outliers or are normally distributed. To address this problem, this study proposes four possible types of Bayesian two-stage robust causal models with IVs to the data, and evaluates the performance of the robust method using Student's t distributions in the causal modeling. The Monte Carlo simulation results show that modeling normal data with normal distributions and normal or heavy-tailed data or data containing outliers with Student's t distributions gives good performance in terms of accuracy, efficiency, and model fit. When data are normally distributed, the methods that either use normal distributions or the Student's t distributions perform equally well as they provide similar bias, SEs and DICs. In the presence of outliers, the nonnormal-s2 and the nonnormal-both models that take outliers into consideration and use Student's t distributions in the second stage to model heavy-tailed data or data containing outliers outperform other distribution models that use normal distributions to model either exclusively all the data or the second stage data in two-stage causal models with IVs with smaller bias and higher efficiency. In addition, the nonnormal-s2 model and the nonnormal-both model have smaller DICs than the other two models,

suggesting evidence of better model fit. The nonnormal-s2 and nonnormal-both models are especially preferred when sample size is small and the proportion of outliers is large as they produce more accurate and efficient LATE estimates.

Note that fitting the nonnormal-both model to data may require longer Markov chains as degrees of freedom for t distributions at both stages need to be estimated. We also want to be cautious to simply use Student's t distributions to model all the data as this method is numerically not optimal all the time and computationally time consuming [28]. Additionally, Student's t distributions are sensitive to the skewness, so some nonnormally distributed data may not be modeled by them. If data are highly skewed, alternative robust method, such as robust methods based on skewed-t distributions may be considered [5].

Author details

Dingjing Shi and Xin Tong*

*Address all correspondence to: xtong@virginia.edu

Department of Psychology, University of Virginia, Charlottesville, Virginia, USA

References

[1] Angrist JD, Imbens G. Two stage least squares estimation of average causal effects in models with variable treatment intensity. Journal of the American Statistical Association. 1995;**90**:431-442

[2] Angrist JD, Imbens G, Rubin D. Identification of causal effects using instrumental variables. Journal of the American Statistical Association. 1996;**91**:444-455

[3] Angrist JD, Pischke J. Mastering Metrics: The Path from Cause to Effect. Princeton, NJ: Princeton University Press; 2014

[4] Angrist JD, Pischke J. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton, NJ: Princeton University Press; 2008

[5] Azzalini A, Genton MG. Robust likelihood methods based on the skew-t and related distributions. International Statistical Review. 2008;**76**:106-129

[6] Baiocchi M, Cheng J, Small D. Instrumental variable methods for causal inference. Statistics in Medicine. 2014;**33**:2297-2340

[7] Casella G, George EI. Explaining the Gibbs sampler. The American Stastician. 1992;**46**: 167-174

[8] Currie J, Yelowitz A. Are public housing projects good for kids? Journal of Public Economics. 2000;**75**:99-124

[9] Gerber AS, Green DP. Field Experiments: Design, Analysis and Interpretation. New York, NY: W.W.Norton & Company; 2011

[10] Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. Robust Statistics: The Approach Based on Influence Functions. New York: John Wiley & Sons, Inc; 1986

[11] Huber PJ. Robust Statistics. New York: John Wiley & Sons, Inc; 1981

[12] Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1984;**6**:721-741

[13] Imbens G, Angrist JD. Identification and estimation of local average treatment effects. Econometrica. 1994;**62**:467-475

[14] Lange KL, Little RJ, Taylor JM. Robust statistical modeling using the t distribution. Journal of the American Statistical Association. 1989;**84**:881-896

[15] Lee SY, Xia YM. Maximum likelihood methods in treating outliers and symmetrically heavy-tailed distributions for nonlinear structural equations. Psychometrika. 2006;**71**:565-585

[16] Lee SY, Xia YM. A robust Bayesian approach for structural equation models with missing data. Psychometrika. 2008;**73**:343-364

[17] Osbourne, J. W. (2002). Notes on the Use of Data Transformation. Practical Assessment, Research & Evaluation, **8**(6), n6.

[18] Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). Practical assessment, research & evaluation, **9**(6), 1-12.

[19] Pinheiro JC, Liu C, Wu Y. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. Journal of Computational and Graphical Statistics. 2001;**10**:249-276

[20] Rachman-Moore D, Wolfe RG. Robust analysis of a nonlinear model for multilevel educational survey data. Journal of Educational Statistics. 1984;**9**:277-293

[21] Seltzer M, Novak J, Choi K, Lim N. Sensitivity analysis for hierarchical models employing t level-1 assumptions. Journal of Educational and Behavioral Statistics. 2002;**27**:181-222

[22] Seltzer M, Choi K. Sensitivity analysis for hierarchical models: Downweighting and identifying extreme cases using the t distribution. Multilevel Modeling: Methodological Advances, Issues, and Applications. 2003;**1**:25-52

[23] Shi D, Tong X. Robust Bayesian estimation in causal two-stage least squares modeling with instrumental variables. In: van der Ark LA, Culpepper S, Douglas JA, Wang W-C, Wiberg M, editors. Quantitative Psychology Research. Springer: New York; 2017

[24] Shoham S. Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions. Pattern Recognition. 2002;**35**:1127-1142

[25] Simmons J, Nelson L, Simon S. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science. 2011;**22**:1359-1366

[26] Song P, Zhang P, Qu A. Maximum likelihood inference in robust linear mixed-effects models using multivariate t distribution. Statistica Sinica. 2007;**17**:929-943

[27] Spiegelhalter D, Best N, Carlin B, van der Linder A. Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society, Series B. 2002;**64**:583-639

[28] Tong X, Zhang Z. Diagnostics of robust growth curve Modeling using Student's t distribution. Multivariate Behavioral Research. 2012;**47**:493-518

[29] Wang J, Lu Z, Cohen AS. The sensitivity analysis of two-level hierarchical linear models to outliers. Quantitative Psychology Research. New York: Springer; 2015. 307-320

[30] Wang H, Zhang Q, Luo B, Wei S. Robust mixture modelling using multivariate t-distribution with missing information. Pattern Recognition Letters. 2004;**25**:701-710

[31] Yuan K-H, Bentler PM. Structural equation modeling with robust covariances. Sociological Methodology. 1998;**28**:363-396

[32] Yuan K-H, Bentler PM. On normal theory based inference for multilevel models with distributional violations. Pschometrika. 2002;**67**:539-561

[33] Yuan K-H, Zhang Z. Structural equation modeling diagnostics using R package semdiag and EQS. Structural Equation Modeling. 2012;**19**:683-702

[34] Yuan K-H, Bentler PM, Chan W. Structural equation modeling with heavy tailed distributions. Psychometrika. 2004;**69**:421-436

[35] Yuan K-H, Lambert PL, Fouladi RT. Mardia's multivariate kurtosis with missing data. Multivariate Behavioral Research. 2004;**39**:413-437

[36] Zhang Z, Lai K, Lu Z, Tong X. Bayesian inference and application of robust growth curve models using Student's t distribution. Structural Equation Modeling. 2013;**20**:47-78

[37] Zhong X, Yuan K-H. Weights. In: Salkind NJ, editors. Encyclopedia of Research Design. Thousand Oaks, CA: Sage; 2010. pp. 1617-1620

[38] Zimmerman D. A note on the influence of outliers on parametric and nonparametric tests. Journal of General Psychology. 1994;**121**:391-401

[39] Zimmerman D. Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. Journal of Experimental Education. 1998;**67**:55-68

[40] Zu J, Yuan K-H. Local influence and robust procedures for mediation analysis. Multivariate Behavioral Research. 2010;**45**:1-44

[41] R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2011

[42] Thomas A, O'Hara B, Ligges U, Sturtz S. Making BUGS open. R News. 2006:**6**;12-17

IntechOpen