Erasmus University Rotterdam

Erasmus School of Economics

Bachelor Thesis Bsc^2 Econometrics and Economics

Using outliers to identify mismeasured models

Willem Castelijns (504070)

Supervisor:	J. Klooster
Second assessor:	A.A. Naghi
Date final version:	2nd July 2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Contents

Abstract

In this thesis we develop two methods to identify simultaneous equation models, in particular systems such as mismeasured models. These methods relate to the method proposed by Lewbel (2012), who uses heteroskedasticity to identify such models. Therefore, we combine the method proposed by Lewbel (2012) with results from the field of robust statistics. By distinguishing between different types of outliers, we have developed methods using vertical outliers as defined by Rousseeuw and Van Zomeren (1990) for identification. We show that by classifying these vertical outliers using Huber's criterion (Huber, 1964) or using a binary classification, we can construct variables that can be employed in the method of Lewbel (2012). These methods work well in the simulated examples but show mixed results in a real data study. We also replicate the results found in the appendix of Lewbel (2012) and we extend the simulation study in Lewbel (2012) for set identification.

1 Introduction

Identification is an important topic in econometrics, it occurs in the setting of Simultaneous Equation Models (SEM) where not all parameters can be estimated. These models were introduced in the beginning of the 20th century due to economists' interests to analyze supply and demand settings (Angrist, Graddy & Imbens, 2000). However, beside supply and demand settings, these models occur also in other situations such as in the presence of measurement errors and endogenous variables. This often reduces the SEM to a triangular model such as in equations 1 and 2:

$$y_1 = X'\beta_{10} + y_2\gamma_{10} + \epsilon_1, \tag{1}$$

$$y_2 = X'\beta_{20} + \epsilon_2, \tag{2}$$

where y_1 is the $(n \times 1)$ dependent variable, y_2 the $(n \times 1)$ endogenous variable, X the $(n \times m)$ matrix containing the exogenous variables and ϵ_i for $i \in \{1, 2\}$ are $(n \times 1)$ the residuals. The most important approach to identify these models is by use of instrumental variables. The effect of an instrumental variable goes via y_2 to y_1, such that γ_{10} can be determined when we know the relation between the instrumental variable and y_2. However, finding adequate instruments is difficult and not all instruments are sufficiently powerful. Therefore, recent developments by Lewbel (2012) offer a new interesting perspective on instrumental variables. In the method proposed by Lewbel (2012), instruments are constructed using the heteroskedasticity. We built on this method introduced by Lewbel (2012) to explore whether the method can be extended to provide identification in the triangular SEM model using outliers.

Thus, in this thesis we explore the identification of SEMs by use of outliers. Therefore, we combine the method proposed by Lewbel (2012) and results from the field of robust statistics. This field is specialized in coping with deviations from the standard assumptions regarding statistical analyses (Huber, 1981). Especially distributional robustness, which concerns outliers, is an important topic in robust statistics (Huber, 1981). Outliers can be the result of a different generating distribution, where either location, scale or distributional shape is different from the 'good' observations. We show how differences in the variance and more general in the distributions generating outliers, can be exploited to obtain identification.

In order to use outliers for identification, we first set-out to determine what outliers are for our identification context. In this regard, we follow a classification of outliers as described in Rousseeuw and Van Zomeren (1990). This leads to a tripartite classification of outliers into *vertical outliers*, *good leverage points* and *bad leverage points*. Using this classification, we show that vertical outliers can be used for identification. For vertical outliers where distribution and scale differ from the good observations, we show that we can construct variables by means of techniques from robust statistics which can then be employed as z variable in the method of Lewbel (2012). We show the use of these techniques in a number of simulations and we apply the method to a real data example. We find that using the Huber criterion (Huber, 1964) to assign weights to observations on their 'outlyingness' is an efficient variable for the method proposed by Lewbel (2012).

Another part of this thesis will focus more on the results in Lewbel (2012). We reproduce the results of the Monte Carlo simulation in the appendix of Lewbel (2012). In addition we extend the simulation study by simulations on the topic of set identification as discussed in Lewbel (2012). We find that the bounds of the sets cannot always be computed and that this leads to slightly biased set bounds. In this simulation we also include a model where the assumption of $\text{cov}(Z, \epsilon_1\epsilon_2) = 0$ is actually relaxed. We show how this impacts the bounds of the set identification by estimating the bounds and computing the theoretical values.

Considering the previous discussion, we formulate a main research question and several subquestions to answer:

Can we use outliers in combination with the method proposed by Lewbel (2012) to obtain identification in simultaneous equation models?

To answer to main research question, we pose the following sub-questions:

1. *Can we reproduce the results presented in Lewbel (2012)? In particular can we replicate the Monte Carlo simulation in the appendix and extend this with a simulation where* $\text{cov}(Z, \epsilon_1\epsilon_2) \neq 0$ *to verify the set bound results*?

2. *Can we obtain identification in simultaneous equation models with an adapted version of the method proposed by Lewbel (2012) using vertical outliers?*

3. *Can we obtain identification in simultaneous equations models with an adapted version of the method proposed by Lewbel (2012) using leverage points?*

4. *Can we obtain identification in an example with real data using outliers?*

These questions contribute to the existing scientific literature by extending the possibilities of constructing instrumental variables. We show that there are differences between heteroskedasticity as employed by Lewbel (2012) and the methods proposed in this thesis. Beside the contribution to the scientific literature, the methods proposed in this thesis will extend the econometrician's toolbox. Instrumental variables are notoriously hard to find and a broad arsenal of methods to find instrumental variables is therefore useful for applied research encountering models with endogeneity or mismeasured variables.

The remainder of this paper has the following structure, we first discuss the literature concerning identification and robust statistics in section 2. Then, in section 3, we discuss the methodological framework of our methods and the techniques on which they are based. In section 4, we present the results of the simulation studies and real application as described in the methodology. Finally we present our conclusion and a discussion of the results in section 5. The appendices contain a proof, discussion of the programming code used for obtaining the results and additional tables to support the results.

2 Literature Review

2.1 Identification

Here we discuss the literature concerning identification and instrumental variables. We first give a general overview of the issue and techniques employed to solve it. We especially focus on the problem of finding adequate instruments, as instruments are the method to obtain identification. Then we focus on the findings in Lewbel (2012) and shortly discuss his method as it will be used further in this paper as a starting point.

2.1.1 Identification General

Simultaneous equation models were developed, among others, by Wright (1928), Tinbergen (1930) and Haavelmo (1943) often in the context of supply and demand equations (Angrist et al., 2000). In the case of supply and demand, we consider the fully simultaneous model as presented in equations 3 and 4:

$$y_1 = X'\beta_{10} + y_2\gamma_{10} + \epsilon_1, \tag{3}$$

$$y_2 = X'\beta_{20} + y_1\gamma_{20} + \epsilon_2. \tag{4}$$

In general with G equations, there are $G \times (G-1)$ unknown parameters γ and $G \times M$ unknown parameters β (Hausman, 1983). With the reduced form we can get $G \times M$ parameters identified, thus in order to identify all structural parameters we need $G \times (G-1)$ instrumental variables (Hausman, 1983). Three types of instrumental variables are common, the exclusion restriction, linear or nonlinear coefficient restrictions and covariance restrictions. With exclusion restrictions, a variable y_2 in equation 3 could be identified when one of the parameters in β_{10} in equation 3 is set to be 0. For a linear or nonlinear coefficient restrictions, the system can be identified when a combination of parameters needs to adhere to a prescribed constraint (Hausman, 1983). Finally, when the residuals are uncorrelated, we can identify the variables using the residuals as instruments. Because without correlation between the residuals, the effect of ϵ_2 goes exclusively through y_2 into y_1. In the methods proposed later it is important to check whether the methods also work in the case that the residuals are correlated, to ensure that the methods proposed do not actually exploit the covariance restriction type of identification that arises with uncorrelated residuals.

The interest in estimating structural form equations instead of reduced form equations lies in the economical modelling behind the studied phenomena. According to Nachtigall, Kroehne, Funke and Steyer (2003), structural equations allow researchers to include latent variables in the models. Structural equations represent the connections between latent variables and reduced form equations represent the connection between latent and observable variables (Nachtigall et al., 2003). Modelling classical measurement errors is one example of using a SEM for modelling latent variables, as the true value is a latent variable and the mismeasured variable the observed variable. Therefore, mismeasured models can be estimated by means of a SEM. Although SEMs are adequate for modelling effects between latent variables, in general SEMs do not measure causal effects (Angrist, Imbens & Rubin, 1996). An acceptable fit does however indicate that the modeled dependencies are supported by the data. As Nachtigall et al. (2003) points out, we can conclude from a fitting SEM that the model is not rejected but we cannot conclude causal relationships. Causal effects can be estimated in a SEM with additional assumptions as discussed in Angrist et al. (1996). For more detail about causal inferences and how it fits in a SEM, we refer to literature on causality such as Angrist et al. (1996).

We now discuss the literature concerning identification by use of outliers. The causal transmission mechanism described in Bazinas and Nielsen (2022) uses catalysts to identify structural equations via causal transmission. Catalysts are similar to instrumental variables which can be found as natural experiment but can also be found searching for outliers in the data whilst also implying causal relationships (Bazinas & Nielsen, 2022). In an empirical example, Bazinas and Nielsen (2022) use oil shocks and fiscal expansions as catalysts which correspond to outlying observations. In this case we can consider the outlier observations to be dependent on additional exogenous variables, that did not influence the good observations. The causal transmission mechanism as described in Bazinas and Nielsen (2022) is therefore adequate when variables can be found that explain the distributional differences between the outliers and good observations. In our analysis we distinguish from the method proposed by Bazinas and Nielsen (2022) by focusing on outliers generated by differences in the variance. Thus, we focus on cases where no variables can be found to explain the outlier observations but where it is clear that the observations have be generated via a different process than the good observations.

2.1.2 Identification using heteroskedasticity

We now discuss the method proposed by Lewbel (2012) to identify SEMs using heteroskedasticity. Lewbel (2012) constructs instruments by means of a variable z (we denote the random variable corresponding to the variable z as Z), this variable needs to fulfill two important conditions as given in equation 5:

$$\operatorname{cov}(Z, \epsilon_1 \epsilon_2) = 0, \quad \operatorname{cov}\big(Z, \epsilon_2^2\big) \neq 0. \tag{5}$$

For a triangular model, such as in equations 1 and 2, the method uses instruments constructed as given in equation 6 (Lewbel, 2012):

$$IV = (z - \bar{z})\epsilon_2, \tag{6}$$

where IV is the instrumental variable, z the variable responsible for the heteroskedasticity and ϵ_2 the first stage residual. For an instrument, it is important that the effect goes to y_1 only via y_2 and that it is correlated sufficiently with y_2. We discuss the instrument presented in equation 6 and argue intuitively why it is a valid instrument. The heteroskedasticity in ϵ_2 indicates that the variance of y_2 is higher/lower for specific values of the variable z, this is guaranteed by the condition $\mathrm{cov}\big(Z, \epsilon_2^2\big) \neq 0$. The condition $\mathrm{cov}(Z, \epsilon_1\epsilon_2) = 0$ ensures that ϵ_1 has different heteroskedasticity than ϵ_2. Thus, conditional on z, y_2 has unique additional or reduced variance. Thus using ϵ_2, which makes the value of y_2 more extreme, and z, which is related to the more extreme values of ϵ_2, we can create a variable that explains y_2 (namely when y_2 is relatively more extreme) but not y_1.

So the constructed variable as presented in equation 6, follows the previous intuitive description. As we are interested in relative values of z (the average of z gets incorporated in the mean value of the variance), we have the term $(z - \bar{z})$, when z is included as explanatory variable as well, which is possible in the method, these relative differences are necessary to distinguish from the levels associated with the location of y_2. The product with ϵ_2 is necessary to establish the relation between the residual and the variable z responsible for the heteroskedasticity.

The method proposed in Lewbel (2012) is presented for triangular models as in equations 1 and 2 and for fully simultaneous models as in equations 3 and 4. The focus is also on latent variable modelling such as measurement errors or latent factor modelling. Similar to the discussion about causal interpretation in SEMs in the forgoing section 2.1.1, also Baum and Lewbel (2019) stress that with this identification method no average local treatment effects are measured, i.e. it is not generally possible to perform causal inference with the method proposed in Lewbel (2012). However, this is inherent to the SEMs in general and not a particular disadvantage of the method proposed by (Lewbel, 2012). Lewbel (2012) also includes estimates for parameter sets when the condition $\mathrm{cov}(Z, \epsilon_1\epsilon_2) = 0$ is violated. Small violations of this condition are no problem but will make the instrument weaker as there is less unique influence of the instrument going via y_2. Finally, Lewbel (2012) also provides a way to use the method in non linear settings by piece wise linear approximations.

2.2 Outliers

We add a new method for identification to the existing literature. This method is based on the method introduced by Lewbel (2012) with adaptations such that identification is obtained using outliers. Therefore, we discuss the relevant literature concerning outliers in this section. We consider how to define an outlier, which types of outliers can be distinguished and how outliers can be detected. Most of the discussed literature comes from the field of robust statistics, Huber (1981, p. 1) states that: '*robustness signifies insensitivity to small deviations from the assumptions*'. In his work on robust statistics Huber (1981) focuses on distributional robustness, which is when the shape of the underlying distribution and the shape of the assumed distribution differ. We show how this connects to outliers, but first we define what an outlier is.

A definition of an outlier is given in Heij et al. (2004, p. 379) as: 'the value of the dependent variable y_i differs substantially from what would be expected from the general pattern of the other observations'. However, observations can deviate from the general pattern in multiple ways. Many attempts have been made in the literature to make distinctions, Aguinis, Gottfredson and Joo (2013) found 14 outlier definitions in a literature review of 46 sources on outlier methodology. We follow the outlier types distinguished in Rousseeuw and Van Zomeren (1990), i.e. we consider *vertical outliers*, *good leverage points* and *bad leverage points*. Before elaborating on these types of outliers, we go over the definition of a leverage point. However, first we consider two effects that could occur in an outlier context, *masking* and *swamping* (Rousseeuw & Hubert, 2011). The *masking* effect is the influence of outliers on a classical regression such that the outlier data points are no longer detectable. When this leads to detection methods labeling 'good' data points as outliers, it is called *swamping*.

Following the definition of an outlier, when for some data points the explanatory variables in X differ substantially from the values of the explanatory variables of other data points, these points are considered to be *leverage* points (Everitt & Skrondal, 2010). Often the hat-matrix $H = X(X'X)^{-1}X'$ is used to diagnose leverage points and Rousseeuw and Van Zomeren (1990) note that many authors even define leverage as the diagonal elements of the hat-matrix. However, according to Rousseeuw and Van Zomeren (1990) this is an improper practice as the diagonal elements h_{ii} of the hat-matrix do not necessarily detect leverage points due to masking and are therefore only (non-robust) diagnostics.

A *good leverage point* is a leverage point i such that the observation's dependent variable y_i fits the linear trend of the non-outlier points. Such a data point is regarded an outlier as it differs substantially from the other data points but fits in the linear pattern and is therefore informative and improving the estimation. A *bad leverage point* is a leverage point i such that the observation's dependent variable y_i does not fit in the linear trend of the non-outlier points, therefore there is probably something else happening at this observation. The point could for example be generated by a different data generating process (DGP), meaning that the parameters of the DGP could differ in this instance, an omitted variable could have significant impact on this particular observation or the randomness is different either in magnitude or distribution. Such a bad leverage point can influence the estimation, especially as it has a certain amount of leverage, such that a researcher should handle with such an observation. Finally we discuss the *vertical outliers*, these points are outliers but non-leverage points, i.e. the explanatory variables do not differ substantially from the non-outlier observations' explanatory variables but the dependent variable does differ substantially. The three types of outliers are visually depicted in figure 1 (Rousseeuw & Van Zomeren, 1990).

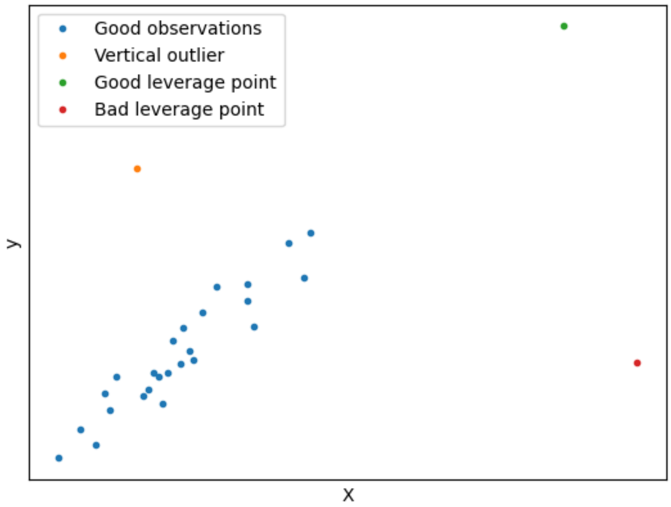

Figure 1: The three types of outliers and good observations

We now look how these three types of outliers fit into the distributional robustness definition stated by Huber (1981). For robust statistics the differences between the actual and assumed distributions should not be too great. In an outlier context, we can define this as follows. Let there be n total observations of which h 'good' observations, these observations follow the assumed distribution and let there be $n - h$ 'bad' observations, these observations have a different distribution. Then, the distribution of a random y_i is composed of the distribution of the 'good' observations and the distributions of the 'bad' observations. For sufficiently large h, this will mean that the distribution of a general y_i is close to the assumed distribution but slightly different. This shows how vertical outliers can be seen in the framework of distributional robustness. This also corresponds to the definition of robust statistics provided by Rousseeuw and Hubert (2011). They define robust statistics as the fit that would have been found without outliers. Without outliers we only have the h 'good' observations and thus indeed the assumed and actual distribution are equal. Thus, vertical outliers we can also define as observations with a different underlying distribution in the generating process. This gives a definition that is based on the generating process instead of the definition found in Heij et al. (2004) where the focus is on realized data points. For good leverage points this definition is not applicable as they fit the general pattern of the other observations, but with a substantially different set of exogenous variables. Bad leverage points could also be seen as a combination of a vertical outlier and a good leverage point.

We discuss a number of techniques used in robust statistics. These techniques are relevant for our research, as we identify outliers and construct variables for identification based on methods in robust statistics. One robust measure for location in the normal distribution is the median (Rousseeuw & Hubert, 2011). A quantitative measure of robustness, the breakdown point, was introduced by Hampel (1971). According to Huber (1981) the breakdown point gives the fraction of outliers an estimator can handle. For the common arithmetic mean, the breakdown point $\epsilon^* = 1/n$, meaning that in the limit the mean can handle 0 bad outliers (where bad outliers are either vertical outliers or bad leverage points), i.e. by changing one observation the mean can change unlimited. The breakdown point of the median is $\epsilon^* = \frac{1}{2}$, indicating that it can handle much more bad outliers. The scale estimate of the normal distribution can similarly be estimated robustly, as the standard estimate $s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$ also has a breakdown point of 0% in the limit (Rousseeuw & Hubert, 2011). Therefore, the Median of all absolute deviations from the median (MAD) can be used:

$$MAD_i = 1.483 \underset{i=1,\dots,n}{\text{median}} |e_i - \underset{j=1,\dots,n}{\text{median}}(e_j)|, \tag{7}$$

where the factor 1.483 is a correction to make the MAD unbiased for the normal distribution. The median and MAD come at the cost of efficiency (Rousseeuw & Hubert, 2011). To find a balance between bias and efficiency maximum-likelihood-estimators or M-estimators have been proposed (Huber, 1964), this class of functions is characterized as follows for a linear model $y_i = x_i'\beta + \epsilon_i$:

$$\min \sum_{i=1}^{n} \rho(y_i - x_i'\hat{\beta}) = \min \sum_{i=1}^{n} \rho(e_i), \tag{8}$$

as described in Heij et al. (2004). When choosing $\rho(e_i) = -\text{log}\ (f(e_i))$ with f the probability distribution of the error term, it is clear that the maximum likelihood estimator is a specific instance of a M-estimator, also for $\rho(e_i) = e_i^2$ we get the ordinary least squares (OLS) estimator that equals the arithmetic mean and for $\rho = |e_i|$ we get an estimate that equals the median estimate of location (Huber, 1964). To make the estimator scale invariant, an implicitly defined form of the M-estimator can be denoted as:

$$\sum_{i=1}^{n} \psi\left(\frac{e_i}{s}\right) x_i = 0, \tag{9}$$

where $\psi = \frac{\partial}{\partial e_i}\rho(e_i)$ and s is an estimate of the standard deviation (Huber, 1981). To estimate the M-estimators, both the standard deviation and the parameters must be estimated. To do this an alternating procedure is described in Huber (1981). This procedure also requires an initial guess, this could be provided by a robust method such as the least trimmed squares estimator (LTS), introduced by Rousseeuw (1984), and for an initial guess of the standard deviation we could use the median of all absolute deviation from the MAD as described in equation 7. Estimating the standard deviation with the estimate $s^{(m)}$, in step m as in Huber (1981) we get:

$$(s^{(m+1)})^2 = \frac{1}{na} \sum_{i=1}^{n} \chi_H \left(\frac{e_i}{s^{(m)}}\right)^2 \left(s^{(m)}\right)^2, \tag{10}$$

with $\chi(e_i) = e_i\psi(e_i) + \rho(e_i)$ and a a bias correcting factor. Then, the parameters can be estimated as in Huber (1981) by determining:

$$\arg \min_{\tau} \sum_{i=1}^{n} \left(e_i^* - \boldsymbol{x}_i'\tau\right), \tag{11}$$

with $e_i^* = \psi_H\left(\frac{r_i}{s^{(m)}}\right) s^{(m)}$ and then obtaining the new parameter estimate with an arbitrary relaxation factor $0 < q < 2$ by $\beta^{(m+1)} = \beta^{(m)} + q\hat{\tau}$. The interesting fact is that the parameter estimation can also be rewritten in terms of weights, essentially giving more weight to 'good' observations and less weight to outliers. The weights for estimation are defined in Huber (1981) as:

$$w_i = \frac{\psi\left(e_i/s^{(m)}\right)}{e_i/s^{(m)}}, \tag{12}$$

using the weights we can also determine the parameter estimation by:

$$\boldsymbol{\tau} = (X^T W X)^{-1} X^T W \boldsymbol{e}, \tag{13}$$

where W is a diagonal matrix with elements w_i such that $\beta^{(m+1)} = \beta^{(m)} + \tau$.

We now explore the Huber criterion as described in Huber (1981). The Huber criterion is a particular function designed by Huber (1964) for M-estimators (see equations 8 and 9) that combines the efficiency of OLS, defined as $\rho(e_i) = e_i^2$, and the robustness of the median, defined as $\rho(e_i) = |e_i|$. The criterion is defined as follows in Huber (1981):

$$\rho_H(e_i) = \begin{cases} \frac{1}{2}e_i^2 + \frac{1}{2}\gamma & \text{for } |e_i| < c \\ c|e_i| - \frac{1}{2}c^2 + \frac{1}{2}\gamma & \text{for } |e_i| \geq c, \end{cases} \tag{14}$$

where $\gamma = \gamma(c)$ with $\gamma(c) = \int \min(c^2, e_i^2)\ \Phi(de_i)$ corrects for the bias in a normal distribution. Then, it follows that:

$$\psi_H(e_i) = \max[-c, \min(c, e_i)], \tag{15}$$

where $\psi_H(e_i) = \frac{d}{de_i}\rho_H$ see equation 9. By obtaining the weight factors for every observation we have a variable that indicates the 'outlyingness' of an observation and thereby can be employed as the z variable in the method proposed by Lewbel (2012).

3 Methodology

3.1 Replication Lewbel

In this section we set out how we reproduce the results found in Lewbel (2012). We replicate the results in the appendix of Lewbel (2012) and we extend section 5 in Lewbel (2012) with a simulation study to show the theoretical findings obtained there.

3.1.1 Replication Appendix Simulation

In this section, we first discuss the replication of the results presented in the supplemental appendix of Lewbel (2012). In the supplemental appendix of Lewbel (2012), a Monte Carlo simulation is performed with the estimators introduced in the article. Data were generated as follows (we present the reduced form of the data generating process):

$$Y_1 = \frac{1}{1 - \gamma_1\gamma_2}\left(\beta_{11} + \beta_{21}\gamma_1 + X\beta_{12} + X\beta_{22}\gamma_1 + \epsilon_2\gamma_1 + \epsilon_1\right), \tag{16}$$

$$Y_2 = \frac{1}{1-\gamma_1\gamma_2}\left(\beta_{21} + \beta_{11}\gamma_2 + X\beta_{22} + X\beta_{12}\gamma_2 + \epsilon_1\gamma_2 + \epsilon_2\right), \tag{17}$$

where the residuals were generated according to $\epsilon_1 = U + e^{X}S_1$ and $\epsilon_2 = U + e^{-X}S_2$ with X, U, S_1 and S_2 independent distributed standard normal scalars. The parameters are assigned the following values to generate the data, $\beta_{11} = \beta_{12} = \beta_{21} = \beta_{22} = \gamma_1 = 1$, for a triangular design the parameter $\gamma_2 = 0$ and $z = X$. For a fully simultaneous design $\gamma_2 = -0.5$ and $z = \left(X, X^2\right)$. We replicated the Monte Carlo presented in Lewbel (2012) with a simulation in Python 3 using the Numpy and Scipy libraries, setting Numpy random seed to 0 for generating the data. The data were generated (for both the triangular design and the fully simultaneous design) as described in equations 16 and 17, with the residuals generated as $\epsilon_1 = U + e^{X}S_1$ and $\epsilon_2 = U + e^{-X}S_2$. To estimate the parameters in the triangular design we implemented the two stage least squares estimator discussed in Lewbel (2012) and in accordance with his simulation we also perform 10,000 simulations with 500 observations each.

For the fully simultaneous model we use the moment conditions in section 3 of Lewbel (2012) in combination with the standard Hansen (1982) estimator for the Generalized Method of Moments (GMM) where the weighting matrix Ω_n is an estimate of $\mathrm{E}\left(Q\left(\theta_0, S\right)\left(\theta_0, S\right)'\right)$. To solve the minimization problem in this set-up we use the python Scipy function optimize.minimize. The 'Powell' method was used as it was the only method to always converge. As this algorithm is rather slow, we first apply the 'TNC' algorithm (Truncated Newton Algorithm) and for the non-converging simulations we afterwards apply the 'Powell' algorithm to guarantee that all simulations converge. Again as in accordance with the simulation in Lewbel (2012), we perform 10,000 simulation of 500 observations each.

3.1.2 Set Identification Simulation

In section 5 of Lewbel (2012), the assumption of $\mathrm{cov}(Z, \epsilon_1\epsilon_2) = 0$ is relaxed. With certain conditions on the ratio between the covariance of the residuals and the heteroskedasticity of the residuals, parameters can be set identified. For set identification, Lewbel (2012) creates a bounded set Γ_1 such that $\gamma_1 \in \Gamma_1$. In Lewbel (2012), this set identification is discussed in the setting of the data generating process presented in equations 16 and 17. This data generating process fulfills all assumptions, so the assumption $\mathrm{cov}(z, \epsilon_1\epsilon_2) = 0$ is satisfied as well. However, a researcher who is not aware of the data generating process may in this case still assume that the $|\mathrm{corr}\left(z, \epsilon_1\epsilon_2\right)| \leq \tau|\mathrm{corr}\left(z, \epsilon_2^2\right)|$ with a certain value of τ. In his paper, Lewbel (2012) works out the analytical expression for the set bounds in the model of equations 16 and 17. For estimation, Lewbel (2012) indicates that replacing the reduced form errors W_j with sample estimates and using these estimates in the equation that defines the set. The equation is given as:

$$\frac{\mathrm{cov}(W_1W_2, Z)^2}{\mathrm{cov}\left(W_2^2, Z\right)^2} - \frac{\mathrm{var}(W_1W_2)}{\mathrm{var}\left(W_2^2\right)}\tau^2 + 2\left(\frac{\mathrm{cov}\left(W_1W_2, W_2^2\right)}{\mathrm{var}\left(W_2^2\right)}\tau^2 - \frac{\mathrm{cov}(W_1W_2, Z)}{\mathrm{cov}\left(W_2^2, Z\right)}\right)\gamma_1 + \left(1-\tau^2\right)\gamma_1^2 = 0. \tag{18}$$

we perform a simulation study in the model of equations 16 and 17 to verify whether these are in accordance with the analytical expression. The parameter values and residuals are the same as in the triangular model in section 3.1 and we again perform 10,000 simulations with 500 observations each. We also present an adjusted model that consists of equations 16 and 17 with the following new residuals:

$$\epsilon_1 = U + e^{aX}S, \tag{19}$$

$$\epsilon_2 = U + e^{-X}S, \tag{20}$$

with U and S standard normal distributed and a an adjustable parameter. Then, we have $\mathrm{E}\left(X\epsilon_1\right) = 0$, $\mathrm{E}\left(X\epsilon_2\right) = 0$, $\mathrm{cov}\big(Z, \epsilon_2^2\big) = \mathrm{E}\left(Xe^{-2X}\right) = -2e^2$ and finally $\mathrm{cov}(Z, \epsilon_1\epsilon_2) = \mathrm{E}\left(Xe^{(a-1)X}\right) = (a-1)e^{\frac{a^2-2a+1}{2}}$ where $Z = X$, so that $\mathrm{cov}(Z, \epsilon_1\epsilon_2)$ is no longer zero (with $a = 1$ the expression is 0). This can be used to generate data for which the ratio τ is known and we can develop the analytical bounds for this particular data generating process. Thus, we generate the data such that when we calculate bounds with $\tau = c$, $\frac{|\mathrm{corr}(Z,\epsilon_1\epsilon 2)|}{|\mathrm{corr}\big(Z,\epsilon_2^2\big)|} = c$, we can do this by picking the right value for a in equation 19. We can then see how the set identification performs when the assumption is actually violated. Apart from the different residuals, the parameters are the same as in the triangular model in section 3.1 and we again perform 10,000 simulations with 500 observations each.

3.2 Extension

Here we discuss our extension on the method presented by Lewbel (2012). We show how outliers can be regarded in the context of heteroskedasticity and how this leads to identification.

3.2.1 Vertical Outliers with Huber's Criterion

As described in the section 2.2, vertical outliers are data points that differ substantially from the value of the dependent variable of the other data points, while the explanatory variables fit in with those of the other data points. We show that this type of outlier can be used for identification and we specify the conditions that have to be met to ensure that identification is successful. The approach we take combines the identification method proposed in Lewbel (2012) with techniques from robust statistics, especially using the criterion proposed by Huber (1964) for M-estimators.

We first discuss the conditions in which the identification method works. We assume that there is no common heteroskedasticity between observations, i.e. the heteroskedasticity in the outliers does no come from a model driven by a variable but we assume that the distribution of each outlier differs randomly. Thus, a model as used in the original simulation described in Lewbel (2012) with residuals generated as $\epsilon_1 = U + e^X S_1$ and $\epsilon_2 = U + e^{-X}S_2$, would therefore not qualify as the heteroskedasticity is driven by the observable variable x. This leaves two options to identify the model using the outlier, the first option is that we find a variable (e.g. corresponding to a rare event) that can explain the extra vertical distance. Such variable can be used for identification, similar to the causal transmission mechanism described in Bazinas and Nielsen (2022). The second option is that there is no (observed) variable responsible for the extra vertical distance. In this case, we can model the outlier as being drawn from a different distribution than the good observations either because the distribution's shape differs or because the parameters defining the distribution differ. Then again, a non-zero location for such distribution could for example correspond to a (group of) unobserved variables that define a causal transmission (Bazinas & Nielsen, 2022). In this second case, the approach as described in Lewbel (2012) is no longer applicable as it will be impossible to find a variable z that explains the heteroskedasticity. With techniques from robust statistics, we can construct a variable that indicates whether a data point is an outlier. For these outlier points, we can regard their 'outlyingness' with respect to the robust fit as the result of unique heteroskedasticity only present at that observation. This variable can be constructed based on the Huber criterion (Huber, 1964), basically giving an estimate to each outlier of how much its distribution differs from the distribution of the 'good' observations. It is also possible to use a binary classification such as used in the Least Trimmed Squared (LTS) estimator (Rousseeuw, 1984). We show that this can then be used in the method of Lewbel (2012) to obtain identification. Therefore, we first define outliers for our identification set-up in a data generating process. After all, the definitions presented in section 2 where constructed to classify outliers in observed samples and it is important to consider how these outlier observations are created.

Here we describe a model for generating outliers and explain how to distinguish it from the heteroskedasticity as employed in the method of Lewbel (2012). We first consider Huber's contamination model or gross error model, as defined by Huber (1964) and discussed in Huber (1981). Later we turn to the outlier generating model as presented in Berenguer-Rico, Johansen and Nielsen (2021). Huber's contamination model is an outlier model on which many concepts in robust statistics, such as e.g. maximum bias and breakdown points are based (Mu & Xiong, 2023) and plays an important role in robust statistics. The model assumes the following generating process for any observation (Huber, 1981):

$$f_i = (1-\eta)g + \eta h, \tag{21}$$

where $\eta \in [0, \frac{1}{2}]$, f_i the distribution of observation i, g the distribution of the good observation and h the distribution of the η fraction of outlier observations. This model provides a broad interpretation of what drives the generation of an outlier. The fact that the outlier is generated by the alternative distribution h is a versatile definition. As described in the previous paragraph, the distribution could differ in multiple aspects. The outlier location could differ due to an (unobserved) variable such as with the causal transmission as described in Bazinas and Nielsen (2022). Another reason for an alternate location is that an outlier is not affected by one of the explanatory variables in the DGP of the good observations (non-constant parameters). Then, there could be differences in the variance, maybe outlier observations are generated with very skewed or much larger variances (if they would be generated with smaller variances than the good observations we would probably not be able to observe them in the data). Finally, there could be higher moments on which the outlier generation depends. This could be captured by allowing the outlier generation to be the result of another distributional function. For the remainder of this section we focus on outliers that have been generated due to differences in the variance or distributional function. As set out before, we also will assume that every outlier is generated different, to distinct it from heteroskedasticity. In this context we mean that the contamination distribution h differs for each outlier, either in parameters or shape. Therefore, we replace the distribution h with h_i:

$$f_i = (1 - \eta)g + \eta h_i \tag{22}$$

We now describe this in more detail, where we follow the notation as presented in Berenguer-Rico et al. (2021). Let there be n observations, we say that the elements $i \in \zeta$ with $\zeta \subseteq \{1, ..., n\}$ and $|\zeta| = h$ such that $h \leq n$ are the 'good' (non-outlier) observations and that the $n-h$ observations $j \notin \zeta$ are the outliers. For the linear model $y_i =_i' \beta + \epsilon_i$, with standard assumptions, we know that the distribution of y_i equals the distribution of ϵ_i with location $\mu_i = x_i'\beta$. We assume that the outliers $j \notin \zeta$ are drawn from a different distribution, such that each outlier observations j has its own distribution $h_j(y)$ different to the distribution $g(y)$ of the good observations as in equation 22. Further we stress that each outlier $j \notin \zeta$ can have an unique distribution such that for each $j, k \notin \zeta$ we have in general that $h_j \neq h_k$. We also make a distinction between outliers that have a distribution $s_j(y)$ with a location parameter $\mu_{y_j} = x_i'\beta$ and with a location parameter $\mu_{y_j} \neq x_i'\beta$. The difference is important, as outliers with $\mu_{s_j} = x_i'\beta$ can be regarded as observations where the outlier only has 'pure' heteroskedasticity because only the variance is affected. For the other type of outliers, there is more to it than only heteroskedasticity as the location of the outlier observation differs from that of the good observations. Non-zero location distributions for outliers will lead to bias in first stage regressions, which has to be accounted for (due the fact that non-zero locations in the residuals are absorbed in the constant). When these location parameters differ due to an observable event, identification could be obtained using the causal transmission mechanism described in (Bazinas & Nielsen, 2022).

We point out that Huber's contamination model as in equation 21 and the form in equation 22, are not necessarily detectable as outliers as defined in section 2.2. A vertical outlier as defined in section 2.2, requires that the observation differs substantially from the good observations given that the exogenous variables are similar. Of course, intuitively it makes sense to classify undetectable observations with different distributions as outliers as well. However, because we cannot distinct these observations in the sample, we pose certain restrictions on the distribution h_i in equation 22 to ensure that all generated outliers are detectable as outliers. Therefore, we adapt the model for generating outliers as presented in Berenguer-Rico et al. (2021). Their model ensures that outliers in the generation stage can later be classified as outliers as outliers will always have larger residuals than the good observations. So let the outlier observations $j \notin \zeta$ have residuals according to the following generating process:

$$\epsilon_j = (\max_{i \in \zeta} \epsilon_i + \xi_j)1_{(\xi_j > 0)} + (\min_{i \in \zeta} \epsilon_i + \xi_j)1_{(\xi_j < 0)}, \tag{23}$$

where ξ_j is drawn from a distribution s_j which can have many forms. We can regard this model as the contamination model in equation 22 where the distribution h_j is replaced by the distribution of $y_j = x_j\beta + \epsilon_j$. This model generates for some observations an unexplained additional randomness and as shown in Berenguer-Rico et al. (2021) the LTS estimator is the maximum likelihood for this model. Although this model unites the definition of a vertical outlier as defined in section 2.2 with the DGP for outliers, there might be some objections about the realism of the model. For many 'real world' economic processes it is questionable whether data are generated according to this process. An operation such as $\max_{i\in\zeta}\epsilon_i$ seems unrealistic for 'real' data as it defines a strict sequence between generating the good and outlier observations. However, we deem the process to be realistic enough to continue our analysis with this model, especially as it allows for a clear distinction between outlier and good observations.

Now we discuss the identification based on the method discussed in the foregoing paragraph. Central in the procedure are the weights obtained from the M-estimator and in this specific case from the Huber-criterion. These weights w_i are influenced by the choice of c in the criterion and are in the Huber criterion constructed such that 'good' observations have weights 1, these are all the observation for which the error $e_i < |c|$ other observations get a declining weight. This means that the largest outliers get a weight that is small, $w_j << 1$ if observation j could be considered a vertical outlier. We use the complement of the weights, thus $p_i = 1 - w_i$ such that the variable has the interpretation such that larger values correspond to vertical outliers. We state that the constructed measure based on Huber's criterion fulfills the conditions set out by Lewbel (2012) in proposition 1.

We show that this works by demonstrating that weights obtained with the Huber criterion fulfill the conditions specified in Lewbel (2012). To use the method proposed by Lewbel (2012) in a triangular model as described in equations 1 and 2, we need to ensure that $\mathrm{cov}(Z,\epsilon_1\epsilon_2) = 0$ and $\mathrm{cov}(Z,\epsilon_2^2) \neq 0$. With $Z = P$ where P is the random variable $P = 1 - \frac{\psi_H(\epsilon_2/\sigma)}{\epsilon_2/\sigma}$ corresponding to the realisations p_i. This leads to proposition 1.

Proposition 1 *Assuming the outlier generation model from Berenguer-Rico et al. (2021) as in equation 23. The 'outlyingness' variable* $P = 1 - \frac{\psi_H(\epsilon_2/\sigma)}{\epsilon_2/\sigma}$ *based on Huber's criterion, fulfills the conditions* $\mathrm{cov}(P,\epsilon_1\epsilon_2) = 0$ *and* $\mathrm{cov}(P,\epsilon_2^2) \neq 0$ *required for identification with the method as presented in Lewbel (2012) in the model as shown in equations 1 and 2.*

The proof of proposition 1 can be found in the appendix A.

3.2.2 Vertical Outliers with Binary Classification

In this section we describe a procedure similar to the one discussed in section 3.2.1, but with a binary classification variable rather than the constructed weights. This is a simpler approach and when outliers share a common generating process it might even be more efficient. Another advantage is that the the classification variable does not have to be constructed algorithmically but can also be constructed on e.g. visual inspection. The idea is to label outliers and use this labeling variable as the z variable in the method proposed by Lewbel (2012). Although the classification does not necessarily have to be performed algorithmically, we can label the outliers with help of the (fast) Least Trimmed Squares (LTS) estimator (Rousseeuw & Van Driessen, 2000). This estimator is basically an OLS estimator but only considering h 'good' (non-outlier observations). When providing the number of h good observations, the algorithm sorts the observations into a group of $n - h$ outliers and h good observations. This method allows for a data driven binary classification of the outliers. This classification variable can then be used to be employed as the z variable in the method proposed by Lewbel (2012). The (fast) LTS algorithm is implemented as described in Rousseeuw and Van Driessen (2000) but without the enhancement for samples with more than 600 observations as all our simulation are for 500 observations. This enhancement is also only to increase computational efficiency.

3.2.3 Leverage points

The previous section shows how vertical outliers can be used for identification. In this section we discuss why leverage points are not adequate for identification when using the method proposed by Lewbel (2012). We focus on good leverage points because we regard bad leverage points as untrustworthy points due to the fact that both that dependent variable and the explanatory variables differ substantially. Bad leverage points could be considered as vertical outliers on leverage points but there are many other reasons why these observations could be so different. The leverage point itself is not useful, as the both the first stage equation and the second stage equation will both have the same leverage points. The reason is that the leverage points are defined by the exogenous variables that occur both in the first and second stage regression (X variable in equations 1 and 2), both equations have leverage points at the same place. Therefore, when we would construct an instrument based on the leverages in the first stage regression, the effect will not only go through the endogenous variable y_2 but also directly through the exogenous variables on to y_1. This excludes leverage points to be used for identification as for an instrument the effect must on y_1 must go via the endogenous variable y_2. Also when constructing an instrument as performed in Lewbel (2012), we still need to ensure that the constructed instrument only influences y_1 via y_2. The heteroskedasticity used in the method of Lewbel (2012) must therefore differ between y_1 and y_2 and as discussed leverage points do not differ between the equations. Therefore, the leverage does not a provide an opportunity on lower moment conditions as in usual instruments but also in higher moment conditions, as employed in the method of Lewbel (2012), we cannot regard the usage of leverages as a new technique. This is because leverage is defined by the extremes in the exogenous variables, using leverages in higher moment conditions, like in the variance, would boil down exactly to the method described in Lewbel (2012). Because when observations with extreme exogenous variable locations have different variances, this can be regarded as heteroskedasticity and this heteroskedasticity can then be modeled be use of the exogenous variables. Therefore, when encountering good leverage points, a good idea is to consider the method described in Lewbel (2012). For bad leverage points, due to the nature of their deviation from the good observations their seems not to be a standard procedure to use these points for identification.

3.2.4 Simulation study

In this section we discuss an implementation of the method described in section 3.2.1. We implement the method using the Huber criterion (Huber, 1964) and labeling the outliers binary as classified by the Least Trimmed Square Estimator (LTS).

First we discuss the simulation set-up for identification with outliers using the Huber criterion (Huber, 1964). Therefore, we generate data with the model as presented in the appendix of Lewbel (2012) but with residuals generated as described by Berenguer-Rico et al. (2021) and found in equation 23. We choose three different implementations of the variable ξ_j, namely (1) $\xi_j \sim N(0,1)$, (2) $\xi_j \sim N(1,1)$ and (3) $\xi_j \sim N(0,A)$ with $A \sim U(0,10)$ with $U(0,10)$ the uniform distribution from 0 till 10. These choices are for the following reasons, with model (1) we can test the general idea whether we can obtain identification with outliers. With model (2) we demonstrate the issues when the distribution has a non-zero location as in the first stage the constant will not be estimated correctly as the location of the residual is absorbed in the constant. In the model (3), we show that if every outlier has its own distribution the identification method still works, i.e. the method works also if all distributions of ξ_j differ. We also control the number of good observations h and let it differ between 90% and 99%. An important choice that we have to make is the value of c in equations 14 and 15. This parameter controls which observations are classified as outliers. In the model of Berenguer-Rico et al. (2021), we would ideally have $c = \max\left(\max_{i\in\zeta} \epsilon_i, |\min_{i\in\zeta} \epsilon_i|\right)$, such that only outliers are considered to be assigned weights. We run 10,000 simulations with 500 observations each. If 80% is good, this means that 400 observations are distributed with a standard normal distribution. We would like to choose c such that only the outlier generated observations get a non-zero p_i value. However, we can better include too many observations in the outlier pool, as the actual correct observations will get small p_i values. Therefore, we choose $c = 2s$ with s the estimated standard deviation. It is well known that for a normal distribution roughly two third of the observations are within two standard deviations. This means that we expect that one third is included with the outliers.

We also perform a simulation where we classify the outliers binary, so an observations is either an outlier or not. The data will be generated the same way as with the Huber criterion. We expect that this approach will work more efficient for scenarios (1) and (2) as all ξ_j variables have the same distribution and there is no need differentiate between the 'outlyingness' of observations. However, for scenario (3) we expect the approach with Huber's criterion to work better, as it allows to give weights to the different distributions, so that more extremely distributed ξ_j variables get higher weights p_i. Classifying the data as outlier or good observations will be done as described in section 3.2.2 using the fast LTS as described by Rousseeuw and Van Driessen (2000). Again we generate 10,000 simulations with 500 observations each.

3.2.5 Applied study

In this section we set out the framework for how to apply the method discussed in section 3.2.1, using vertical outliers with Huber's criterion to a real data example. Therefore, we investigate whether our method can provide identification in the analysis performed by Alesina and Zhuravskaya (2011). In their research Alesina and Zhuravskaya (2011) measured the effect of segregation on government quality by regressing segregation with a set on control variables on the World Bank's Governance indicators. To overcome endogeneity, Alesina and Zhuravskaya (2011) use an instrumental variable to obtain causal inference about the influence of segregation on the quality of government. The instrument uses a relation between the major groups in neighbouring countries with the spatial distribution of population groups within a country (Alesina & Zhuravskaya, 2011). The system of equations can therefore be represented as in equations 24 and 25:

$$Q_i = \alpha_Q + \beta_Q S_i + \gamma_Q F_i + \delta_Q' X_i + \epsilon_i^Q, \tag{24}$$

$$S_i = \alpha_S + \beta_S S_i^p + \gamma_S F_i + \delta_S' X_i + \epsilon_i^S, \tag{25}$$

where S_i^p is the instrument, a variable for the predicted segregation, S_i is a segregation index, F_i an (ethnic) fractionalization index and X is a vector of additional covariates.

In their paper they conclude that influential observations, which we regard as outliers, do not drive their results. However, the online appendix does discuss the influence of certain observations on the first and second stage regressions. Chile and Zimbabwe turn out to be outlier observations that enhance the results in favor of the hypothesis, excluding these observations for linguistic segregation, makes the instrument weak in the second stage. Without Chile and Zimbabwe, the USA are an outlier observation in the first stage regression, by excluding the USA as well the instrument is sufficiently strong enough for identification. We are interested in the model without Chile and Zimbabwe for linguistic segregation and we apply the method proposed in section 3.2.1, using vertical outliers with the Huber criterion, to obtain identification. We explore whether we can find significant effects on the government quality indicators 'voice and accountability' and 'political stability' as these were identified without the USA. In the full sample there were 4 outlier observations, Chile, Zimbabwe, Bulgaria and Russia. Therefore, we also apply the method to the full sample to determine whether the outlier constructed instruments can replace the instrumented used in Alesina and Zhuravskaya (2011) as it is clear from the discussion in the online appendix that these four observations can be considered outliers.

4 Results

4.1 Extension

Here we present the results of the simulation study and applied study using the method proposed in sections 3.2.1 and 3.2.2. For the simulation study, every simulation was performed at with 10% and 1% outliers. Only one 10% table is included here, the other tables are in the appendix.

4.1.1 Simulation study

First we present the results for the simulation using the Huber criterion in section 3.2.1 for identification. We present the results for model (1), i.e. with $\xi_j \sim N(0,1)$ in table 1 and table 2. We note that the method indeed seems to work and using the outliers we can identify the parameter γ_1. As we can see, more outliers make the method more efficient in the second stage but less efficient in the first stage. This is as we would expect, as with more outliers the first stage is of lower quality. This means that the Huber criterion, interpreted in weights, has a lower total weight distribution and thus lower efficiency. The fact that the second stage becomes more efficient with more outliers is also as we would expect, as we have created a stronger instrument.

We now discuss the results of model (2), which can be found in tables 3 and 15. Again we see that with less outliers the first stage is more efficient and the second stage is less efficient. We also see that the fact that the location of the outliers is non-zero can be found in the first stage constant. This is what we would expect, but of course the effect is stronger with more outliers as we can see in the tables. A solution is to re-estimate the first stage regression with robust statistics methods, which was not done in the current method. This would solve the problem seen here, but ensures that the second stage can still be identified with the outliers.

Table 1: Simulation Results Huber model (1) with 10% outliers

parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
β_{11}	1.00	1.00	0.060	0.960	1.00	1.04	0.060	0.048	0.040
β_{12}	1.00	1.00	0.060	0.959	1.00	1.04	0.060	0.048	0.041
γ_1	1.00	1.00	0.040	0.973	1.00	1.03	0.040	0.032	0.027
β_{21}	1.00	1.00	0.077	0.949	1.00	1.05	0.485	0.130	0.091
β_{22}	1.00	1.00	0.059	0.954	1.00	1.05	0.069	0.055	0.046

Note: MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

Table 2: Simulation Results Huber model (1) with 1% outliers

parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
β_{11}	1.00	1.00	0.104	0.932	1.00	1.07	0.104	0.082	0.068
β_{12}	1.00	1.00	0.104	0.933	1.00	1.07	0.104	0.082	0.069
γ_1	1.00	1.00	0.093	0.937	1.00	1.06	0.093	0.074	0.062
β_{21}	1.00	1.00	0.048	0.967	1.00	1.03	0.048	0.038	0.033
β_{22}	1.00	1.00	0.048	0.968	1.00	1.03	0.048	0.038	0.032

Note: MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

Table 3: Simulation Results Huber model (2) with 1% outliers

parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
β_{11}	1.00	1.00	0.102	0.932	1.00	1.07	0.102	0.082	0.069
β_{12}	1.00	1.00	0.100	0.934	1.00	1.07	0.100	0.079	0.067
γ_1	1.00	1.00	0.090	0.940	1.00	1.06	0.090	0.071	0.060
β_{21}	1.00	1.03	0.047	0.999	1.03	1.06	0.056	0.045	0.038
β_{22}	1.00	1.00	0.049	0.986	1.00	1.03	0.049	0.039	0.033

Note: MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

We present now the results for model (3) in tables 4 and 16, here we compare the results with the findings for the simulations with the Huber criterion. As we can see table 4 has more efficient results in the second stage than the simulations corresponding to tables 2 and 3. The results in models (1) and (2) had very similar efficiencies, with slightly better uncertainties in model (2) where the outliers were larger. In model (3) however, the outliers are considerably larger than those in models (2) and (3). This shows that the the size of the outliers affects the efficiency, as we would expect.

Table 4: Simulation Results Huber model (3) with 1% outliers

parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
β_{11}	1.00	1.00	0.081	0.949	1.00	1.05	0.081	0.063	0.052
β_{12}	1.00	1.00	0.080	0.949	1.00	1.05	0.080	0.062	0.051
γ_1	1.00	1.00	0.067	0.961	1.00	1.04	0.067	0.050	0.039
β_{21}	1.00	1.00	0.059	0.961	1.00	1.04	0.059	0.047	0.039
β_{22}	1.00	1.00	0.058	0.962	1.00	1.04	0.058	0.046	0.039

Note: MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

We now present the results of the simulation models (1), (2) and (3) but with the binary classification model. This classification was based on (fast) LTS estimation. We first present the binary classification identification for model (1) in tables 5 and 17. We see that the results are close to those with the Huber criterion in table 2 but slightly worse in efficiency. We had not necessarily expected this, because all ξ_j were generated with the standard normal distribution and therefore assigning different weights, as done in the Huber criterion, seemed redundant and therefore only wasting efficiency. However, as we can see the Huber criterion is actually more efficient. Perhaps, the fast LTS algorithm does assign some good observations to the outliers. This would then definitely reduce the efficiency.

Table 5: Simulation Results Binary (1) with 1% outliers

parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
β_{11}	1.00	1.00	0.133	0.913	1.00	1.09	0.133	0.106	0.088
β_{12}	1.00	1.00	0.133	0.912	1.00	1.09	0.133	0.106	0.089
γ_1	1.00	1.00	0.126	0.916	1.00	1.08	0.126	0.100	0.083
β_{21}	1.00	1.00	0.048	0.967	1.00	1.03	0.048	0.038	0.033
β_{22}	1.00	1.00	0.048	0.968	1.00	1.03	0.048	0.038	0.032

Note: MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

We now present the results for model (2) using the binary classification method. Again we see that the results are slightly worse (less efficient) than those with the Huber criterion in tables 6 and 15. In this instance as well, we actually expected that the binary classification method would work at least as good as the Huber criterion. But the reason for the fact that actual results are worse could again be caused by the fact that the (fast) LTS method makes some classification mistakes which would have significant impact on the efficiency.

Finally, we present the binary classification method for model (3) in tables 7 and 19. For this

Table 6: Simulation Results Binary (2) with 1% outliers

parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
β_{11}	1.00	1.00	0.126	0.917	1.00	1.09	0.126	0.100	0.084
β_{12}	1.00	1.00	0.123	0.922	1.00	1.08	0.123	0.097	0.081
γ_1	1.00	1.00	0.115	0.922	1.00	1.07	0.115	0.091	0.076
β_{21}	1.00	1.03	0.047	0.999	1.03	1.06	0.056	0.045	0.038
β_{22}	1.00	1.00	0.049	0.986	1.00	1.03	0.049	0.039	0.033

Note: MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

Table 7: Simulation Results Binary (3) with 1% outliers

parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
β_{11}	1.00	1.00	0.093	0.944	1.00	1.06	0.093	0.071	0.056
β_{12}	1.00	1.00	0.092	0.946	1.00	1.05	0.092	0.070	0.055
γ_1	1.00	1.00	0.081	0.958	1.00	1.04	0.081	0.059	0.043
β_{21}	1.00	1.00	0.059	0.961	1.00	1.04	0.059	0.047	0.039
β_{22}	1.00	1.00	0.058	0.962	1.00	1.04	0.058	0.046	0.039

Note: MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

model, we had expected that the results would be worse than those using the Huber criterion as in table 4. We see that this is indeed the fact, however as we noticed for models (1) and (2) this was also the case. When we look at how much better the Huber criterion performed in model (3) compared to models (1) and (2), we see that the Huber criterion actually seems to perform better in models (1) and (2). We cannot spot an immediate striking effect of a joint worse LTS performance and an advantage for the Huber criterion in taking the amount of 'outlyingness' into account. The results of the binary classification generally show that the Huber criterion is performing better in all these three scenarios. We can consider the Huber criterion method as a more sophisticated method, as both do basically the same but with the Huber criterion we are not restricted to a binary classification.

4.1.2 Applied study

We estimated the regressions as described in section 3.2.5. We present the results in table 8, where we present the results of four models. The model IV-Original, is the regression using equations 24 and 25, with the instrument constructed in Alesina and Zhuravskaya (2011). The OLS model presents the results as presented in Alesina and Zhuravskaya (2011) as well, for equation 24 without use of the instrument. Then, we measure two models with the Huber Criterion using vertical outliers.

The first model, IV-Outliers, is applied on the full sample and allows us to compare the instrument constructed in Alesina and Zhuravskaya (2011) with the instrument constructed in section 3.2.1. We see that for the two most interesting dependent variables, 'voice and accountability' and 'political stability', the estimates obtained in Alesina and Zhuravskaya (2011) are quite different from the estimates we obtain with our instrument. However, we also notice

Table 8: Linguistic segregation and the quality of government for different models

Measure	Model	Voice	Political stability	Government effectiveness	Regulatory quality	Rule of law	Control of corruption
S	IV-Original	-2.65^{***}	-2.92^{***}	-1.54^{*}	-1.95	-1.80^{**}	-1.29
	IV-Outliers	-1.90^{***}	-2.03^{***}	-1.44^{*}	-2.07	-1.84^{**}	-1.53
	OLS	-1.38^{***}	-1.53^{***}	-0.57	-0.69	-1.15^{**}	-0.80
	IV-Outliers-Red	-1.22^{**}	-0.91^{**}	-1.17^{*}	-1.90	-1.22^{**}	-1.03
F	IV-Original	0.44^{*}	0.24	0.44^{*}	0.48	0.31	0.13
	IV-Outliers	0.37^{*}	0.12	0.45^{*}	0.50	0.32	0.16
	OLS	0.26	0.05	0.31	0.3	0.22	0.06
	IV-Outliers-Red	0.23	-0.06	0.41^{*}	0.18	0.24	0.11

Note: S is the segregation and F the fractionalization as in equation 24. All models use all control variables as described in Alesina and Zhuravskaya (2011), three models use the full sample but the IV-Outliers-Red excludes Chile and Zimbabwe. *** denotes a p-value of 1 percent, ** denotes a p-value of 5 percent and * denotes a p-value of 10 percent.

that using our instrument the estimates are considerably different from the results with only OLS. We see that using our instrument, estimates are closer to the estimates with the instrument used by Alesina and Zhuravskaya (2011) compared to OLS. For 'Rule of law' the estimates are even very close and for the other non-significant variables the estimates with our instrument are close to the estimates with the instrument used by (Alesina & Zhuravskaya, 2011) as well. We therefore conclude that in the full sample our instrument seems to have limited power but does also show that it could work.

Without Chile and Zimbabwe, the USA turned out to be an outlier in the first stage regression. Therefore, in the model IV-Outliers-Red, we estimated the model without Chile and Zimbabwe with the knowledge that there is at least one outlier, the USA. We see that the results are again quite different for the dependent variables 'voice and accountability' and 'political stability', only for regulatory quality is the estimate reasonably similar to the estimate obtained with the instrument in Alesina and Zhuravskaya (2011). In this context the method proposed in 3.2.1 does not seem to work well. Perhaps having only the USA as outlier is not enough for identification or maybe the first stage and second stage outliers are correlated. This would reduce the strength of the instrument using outliers.

We see that the applied study presents a mixed view on the application of the instrument using Huber's criterion. For the full sample, it did not work as well as the instrument constructed by Alesina and Zhuravskaya (2011) but their instrument is dedicated to this problem whereas the instrument we presented is a more generally applied method. Therefore, it is not remarkable that our instrument gives some different results, although for most estimates the values are quite close. For the model without Chile and Zimbabwe, we see that our method fails, it is often quite close to the OLS estimates. Therefore, it is important to consider whether all assumptions necessary for the method to work are fulfilled.

4.2 Replication Results Lewbel

In this section we present the results from the replication of the Monte Carlo simulation in the appendix of Lewbel (2012) and we present a simulation study of the findings in section 5 of

Lewbel (2012). We find that the results presented in Lewbel (2012) are very well reproducible and that differences can be explained.

4.2.1 Replication Results Appendix Simulation

We first present the results of the simulation of the triangular model in table 9. As we can see the results are similar and very close to the findings in Lewbel (2012).

Table 9: Simulation Results Triangular Model Two Stage Least Squares

parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
β_{11}	1.00	1.00	0.133	0.913	1.00	1.09	0.133	0.104	0.087
β_{12}	1.00	1.00	0.272	0.835	1.00	1.17	0.272	0.206	0.166
γ_1	1.00	1.00	0.034	0.981	1.00	1.02	0.034	0.025	0.019
β_{21}	1.00	1.00	0.128	0.915	1.00	1.08	0.128	0.101	0.085
β_{22}	1.00	1.00	0.267	0.835	1.01	1.17	0.267	0.205	0.167

Note: MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

The results for the simultaneous model can be found in table 10. As we can see the results are in some aspects different to the results in Lewbel (2012). First, the results for the quantile statistics (LQ, MED, UQ and MDAE) are nearly the same, indicating that the distributions are very close. Also the mean statistic is very close to the results in Lewbel (2012), although the mean of γ_2 is a bit worse in our results. Especially the results for the standard deviation (SD), the root mean squared error (RMSE) and the mean absolute error (MAE) are different. As the bias in our results and in Lewbel (2012) is very small compared to the standard deviation, the root mean squared is for all cases very close to the standard deviation. We explain the different results due to occurrence of a small group of extreme estimates, as Lewbel (2012) also mentions when presenting his results. The quantile statistics are only very limited influenced by these large outliers and therefore these statistics correspond very good to the results in Lewbel (2012). However, in Lewbel (2012) the optimization seems to end more often at extreme estimates (although this effect is probably limited otherwise the quantile statistics would be influenced more) or at more extreme estimates for the parameter estimates. This explains the much higher standard deviation and root mean squared and also the fact that the mean absolute error is consistently higher in Lewbel (2012). As the standard deviation and root mean squared pay more weight to the extreme observations than the mean absolute error we also see that the SD and RMSE statistics are indeed more different than the MAE. So we conclude that the differences are very likely to descend from differences in the extreme estimates from the optimization.

Further the bias in the mean of γ_2 is striking, as it is considerably higher than the bias in Lewbel (2012). As the quantile distribution for this parameter is very close to the results in Lewbel (2012), we expect that this stems from extreme estimates as well. A closer look at the parameter estimates learns that there are more extreme estimates below the real value of γ_2 than above the real value of γ_2. We think that this is also due to the optimization, the algorithm might be more inclined to make extreme estimates below the true value of γ_2 due to the shape of the objective function. The fact that the result in Lewbel (2012) is more accurate in the mean is therefore probably due to the choice of optimization algorithm but we can also see that the results in Lewbel (2012) are slightly biased in the same direction and that the quantiles are more skewed towards higher values of γ_2 (and thus lower value of $-\gamma_2$).

Table 10: Simulation Results Simultaneous System GMM

parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
β_{11}	1.00	1.00	0.154	0.916	1.00	1.09	0.154	0.104	0.085
β_{12}	1.00	1.00	0.280	0.839	1.01	1.17	0.280	0.208	0.167
γ_1	1.00	1.01	0.234	0.973	1.00	1.02	0.234	0.044	0.026
β_{21}	1.00	1.02	0.484	0.912	1.00	1.09	0.485	0.130	0.091
β_{22}	1.00	1.02	0.567	0.833	1.00	1.18	0.567	0.232	0.172
$-\gamma_2$	0.500	0.516	0.368	0.527	0.501	0.477	0.386	0.048	0.025

Note: MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

Finally, we discuss the differences and similarities between the results in the triangular design and the simultaneous design. We can regard the triangular design as a nested model of the simultaneous design with $\gamma_2 = 0$. So the simultaneous design estimates one additional parameter in comparison with the triangular design. Therefore, we may expect that the bias and standard deviations are not too different but slightly larger in the simultaneous design as we estimate an additional parameter. We can see indeed that biases are slightly worse in the simultaneous design for the parameters β_{11}, β_{12}, β_{21}, β_{22}, and γ_1. Also the standard deviations for these parameters are higher.

We expected that the estimation of equation 16 and 17 would not differ in precision and efficiency as the set-up is ‘symmetric’ (the only difference is the value of the parameter γ_2 and the generation of the residuals). However, standard deviations in equation 17 are considerably higher, something that Lewbel (2012) finds as well. The strength of the instruments using $Z = \big(X, X^2\big)$ is as strong with the residuals of equation 16 as equation 17 ($\text{cov}\big(X, \epsilon_1^2\big) = 2e^2$, $\text{cov}\big(X, \epsilon_2^2\big) = -2e^2$, $\text{cov}\big(X^2, \epsilon_1^2\big) = 1 + 5e^2$, $\text{cov}\big(X^2, \epsilon_2^2\big) = 1 + 5e^2$). Therefore, it is unlikely that the generation of the residuals is of influence. The difference in the value of γ_2 must therefore be the cause of the difference in efficiency. We conclude that the optimization algorithm has more difficulties with optimizing the parameters β_{21}, β_{22} and γ_2 due to the value of γ_2 and its impact on the optimization.

4.2.2 Set Identification Simulation Results

We first present the results for the data generated with the model in equations 16 and 17 and the original residuals $\epsilon_1 = U + e^X S_1$ and $\epsilon_2 = U + e^{-X} S_2$. For each simulation we obtained the lower bound and the upper bound of the set Γ_1 (the set that contains the parameter γ_1), the results are shown in table 11. The TRUE parameters are calculated using the analytical expression provided in Lewbel (2012) for this particular model. As we can see, the estimates for the bounds lead to larger sets Γ_1 than those provided by the analytical expression, as both the lower bound is underestimated and the upper bound is overestimated. This is partially explainable by the fact that roughly 13% (14% for $\tau = 0.1$, 14% for $\tau = 0.5$ and 12% for

$\tau = 0.9$) of the estimates results in functions that do not have roots (always > 0). The estimates without roots would be the estimates with bounds lying close together, thus high lower bound and low upper bounds. This results in poor estimates for the statistics as we use only 90% and throw away an important 10%. However, the characteristics shown here could also be present in applied econometric research, where especially for small sample sizes it will not be guaranteed that the function presented in Lewbel (2012) has roots.

We also observe that estimates become worse for larger values of τ. The increase in bias is probably because the estimates' standard deviation increases, meaning more extremer points while still a fraction of the estimates is thrown away. With the 13% of discarded estimates being those that should counter balance the statistics' estimates.

Table 11: Simulation Results Triangular Model Set Identification

τ	parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
0.1	$\Gamma_{1_{LB}}$	0.995	0.985	0.038	0.965	0.988	1.008	0.040	0.029	0.022
	$\Gamma_{1_{UB}}$	1.005	1.012	0.036	0.991	1.010	1.032	0.037	0.027	0.020
0.5	$\Gamma_{1_{LB}}$	0.973	0.915	0.069	0.883	0.928	0.961	0.090	0.064	0.046
	$\Gamma_{1_{UB}}$	1.023	1.069	0.055	1.033	1.061	1.095	0.070	0.051	0.038
0.9	$\Gamma_{1_{LB}}$	0.892	0.612	0.313	0.483	0.681	0.822	0.421	0.296	0.212
	$\Gamma_{1_{UB}}$	1.084	1.224	0.194	1.105	1.186	1.294	0.240	0.159	0.106

Note: $\Gamma_{1_{LB}}$ denotes the lower bound of the identifying set Γ_1 and $\Gamma_{1_{UB}}$ denotes the upper bound. MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

We now present the results from the simulation with the model of equations 16 and 17 with residuals generated as 19 and 20. This is the model with an actual non-zero covariance between Z and $\epsilon_1\epsilon_2$. The results for this adjusted model can be found in table 12.

Table 12: Simulation Results Adjusted Triangular Model Set Identification

τ	parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
0.1	$\Gamma_{1_{LB}}$	0.990	0.974	0.048	0.951	0.979	1.002	0.050	0.035	0.025
	$\Gamma_{1_{UB}}$	1.000	1.006	0.043	0.983	1.005	1.029	0.043	0.031	0.023
0.5	$\Gamma_{1_{LB}}$	0.885	0.726	0.289	0.658	0.789	0.878	0.330	0.182	0.103
	$\Gamma_{1_{UB}}$	1.016	1.119	0.220	1.024	1.072	1.153	0.243	0.118	0.063
0.9	$\Gamma_{1_{LB}}$	0.511	-0.223	1.369	-0.538	0.097	0.502	1.554	0.832	0.432
	$\Gamma_{1_{UB}}$	1.081	1.484	0.925	1.137	1.270	1.523	1.009	0.417	0.189

Note: $\Gamma_{1_{LB}}$ denotes the lower bound of the identifying set Γ_1 and $\Gamma_{1_{UB}}$ denotes the upper bound. MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

Again the same problem arises as in the simulation that was discussed in table 11, a fraction of the estimates (11% for $\tau = 0.1$, 3% for $\tau = 0.5$ and 2% for $\tau = 0.9$), cannot be used as the expression to determine the set bounds does not have roots. This means again that the estimates presented in table 12 are biased towards too large sets. We do also see that in this model with actual violation of the assumption that $\text{cov}(Z, \epsilon_1\epsilon_2) = 0$, both the analytical bounds and the estimated bounds are larger than in the model used in Lewbel (2012). For $\tau = 0.9$ we see that the estimates indicate a very large set (although this is still influenced by the fact that counter balancing estimates could not be used). But also the analytical bounds of the set are much larger in this instance than those presented in Lewbel (2012), such that in models where the violation is actually violated for large values of τ (close to 1) the set might become rather large after all. Also, we see that in this particular model for $\tau = 0.1$ the true parameter $\gamma_1 = 1$ is only included at the very boundary of the set, actually for every value of τ we see that the true value is closer to the upper bound. We assume this is a property of this particular set-up but it does show that for large set intervals it is really uncertain what the true parameter is as the value could be anywhere in the interval.

We also present the following tables to support our claim that the bias is largely due to the fact that a certain fraction of estimates is unusable. We re-estimated the models above with only 100 simulations and 500,000 observations in the original model used for the triangular design in Lewbel (2012) and 100 simulations and 50,000 observations in the adjusted model with residuals as in equations 19 and 20. The results for the triangular model as described in Lewbel (2012) can be found in table 13, with this number of observations 0% of estimates is thrown away. We see indeed that biases are much smaller, but of course in practice data sets with 500,000 observations are not the standard and 100 simulations are rather few to make strong statements about the distribution.

Table 13: Simulation Results Triangular Model Set Identification With Increased Observations

τ	parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
0.1	$\Gamma_{1_{LB}}$	0.995	0.995	0.001	0.994	0.995	0.996	0.001	0.001	0.001
	$\Gamma_{1_{UB}}$	1.005	1.005	0.001	1.004	1.005	1.006	0.001	0.001	0.001
0.5	$\Gamma_{1_{LB}}$	0.973	0.970	0.006	0.966	0.970	0.973	0.007	0.005	0.005
	$\Gamma_{1_{UB}}$	1.026	1.028	0.006	1.025	1.029	1.032	0.006	0.005	0.004
0.9	$\Gamma_{1_{LB}}$	0.892	0.884	0.025	0.867	0.882	0.900	0.027	0.021	0.020
	$\Gamma_{1_{UB}}$	1.083	1.093	0.019	1.082	1.096	1.105	0.021	0.017	0.016

Note: $\Gamma_{1_{LB}}$ denotes the lower bound of the identifying set Γ_1 and $\Gamma_{1_{UB}}$ denotes the upper bound. MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

Table 14: Simulation Results Adjusted Triangular Model Set Identification With Increased Observations

τ	parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
0.1	$\Gamma_{1_{LB}}$	0.990	0.988	0.004	0.986	0.988	0.991	0.004	0.004	0.003
	$\Gamma_{1_{UB}}$	1.000	1.002	0.004	0.999	1.002	1.005	0.004	0.003	0.003
0.5	$\Gamma_{1_{LB}}$	0.885	0.863	0.030	0.851	0.865	0.878	0.037	0.029	0.024
	$\Gamma_{1_{UB}}$	1.016	1.035	0.023	1.020	1.032	1.047	0.030	0.022	0.017
0.9	$\Gamma_{1_{LB}}$	0.511	0.426	0.125	0.371	0.430	0.488	0.151	0.116	0.098
	$\Gamma_{1_{UB}}$	1.081	1.139	0.070	1.094	1.132	1.172	0.091	0.065	0.053

Note: $\Gamma_{1_{LB}}$ denotes the lower bound of the identifying set Γ_1 and $\Gamma_{1_{UB}}$ denotes the upper bound. MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

In table 14 we present the results for the adjusted model with 100 simulations and 50,000 observations, again all estimates could be used. We see again that biases are much smaller.

Overall we can conclude that, at least with this model set-up, only with large number of observations roots can with great certainty be found. But the results in tables 13 and 14 should be taken interpreted with cause as the number of simulations was very low.

5 Conclusion

In order to answer our main research question we will first treat the sub-questions posed in the introduction. Our first sub-question was: ‘Can we reproduce the results presented in Lewbel (2012)? In particular can we replicate the Monte Carlo simulation in the appendix and extend this with a simulation where $\text{cov}(Z, \epsilon_1\epsilon_2) \neq 0$ to verify the set bound results?’. We conclude that the results in Lewbel (2012) can be replicated. There are some differences, especially in the simultaneous design but these differences can be explained. We also extended the simulation in Lewbel (2012) with simulation concerning the set identification theory. We find that indeed set identification is a strong technique to cope with violated assumptions but that bounds for the sets cannot always be found. In addition we also found that the true parameter values can be at the very border of the sets, indicating that when sets are very large their usage is only limited.

Our second sub-question was: ‘can we obtain identification in simultaneous equation models with an adapted version of the method proposed by Lewbel (2012) using vertical outliers?’. Therefore, we proposed two methods to obtain identification using vertical outliers. The first method used Huber’s criterion (Huber, 1964) to construct weights determining the ‘outlyingness’ of observations. This method is fully data-driven and assigns higher weights to more aberrant observations. The other method was a binary classification, that could either be used in a set-up where outliers are identified by the researcher or by a data-driven process based on the fast-LTS algorithm (Rousseeuw & Van Driessen, 2000). The methods are distinct from those in Lewbel (2012) by introducing a data-driven variable for explaining the heteroskedasticity and by not explicitly assuming a common variable responsible for common heteroskedasticity. The simulations showed that in the outlier model as presented by Berenguer-Rico et al. (2021), the methods are adequate in obtaining identification. The method using Huber’s criterion (Huber, 1964) seems to be the most efficient method. We expect that this method is more efficient because it can assign larger weights to more outlying observations.

Our third sub-question was: ‘Can we obtain identification in simultaneous equations models with an adapted version of the method proposed by Lewbel (2012) using leverage points?’. We concluded that we think it is unlikely to construct instruments based on leverage for identification. At least not in a set-up similar to the method as presented in Lewbel (2012). This is because the exogenous variables show up in both the first and second stage regression. Leverage points are therefore by their definition present in both the first and second stage. Because the leverage points are also present at the same locations (the exogenous variables are the same), there is no unique effect going through y_2 to y_1. However, we do not exclude the possibility that leverages could be used in another way to obtain identification. Therefore, it could be interesting for future research to investigate whether leverage points can be used for identification.

Our final sub-question was: ‘Can we obtain identification in an example with real data using outliers?’. We applied our method to the data in the research by Alesina and Zhuravskaya (2011) and we found mixed results. We had expected that by excluding Chile and Zimbabwe, we could perhaps obtain identification as there was an outlier in the form of the USA. However, in this regression the estimation method based on Huber's criterion performed rather bad. However, in the full sample without the instrument used by Alesina and Zhuravskaya (2011) the method seemed to outperform OLS and was often close to the estimates found in Alesina and Zhuravskaya (2011). Only for the two most significant results, the estimates were more off. However, although Chile and Zimbabwe were outliers in the full sample, they were not explicitly outliers in the first stage. Therefore, the results obtained with our method should be seen in a more positive setting. It would be interesting for future research to use the method in a data set with a SEM analysis where outlier are clearly present in the first stage. This would learn whether the method is useful in applied settings.

Now we will answer the main research question posed: 'Can we use outliers in combination with the method proposed by Lewbel (2012) to obtain identification in simultaneous equation models?'. We conclude that the simulations show that the methods proposed using Huber's criterion and the binary classification are adequate in particular outlier generating settings. However, it remains the question how 'realistic' these outliers model are in the real world. The applied setting could not yet give conclusive evidence on this matter but did show some promising results. Therefore, we think it is very dependent on the situation whether the method can be useful. When outliers have a strong presence in the first stage and a SEM model is used, the methods proposed in this paper could be used. However, to give a conclusive advice, we think future research should first find stronger evidence in real data application before using the methods described in this paper. We also focused predominantly on triangular models, therefore we cannot generalize the results yet to larger systems of SEM variants. However, the triangular model is frequently present, e.g. in mismeasured models. Therefore, the methods described here should be considered for such triangular models and future research could see how the methods fit in a more general simultaneous set-up.

References

Aguinis, H., Gottfredson, R. K. & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*(2), 270–301. doi: DOI:10.1177/1094428112470848

Alesina, A. & Zhuravskaya, E. (2011). Segregation and the quality of government in a cross section of countries. *American Economic Review*, *101*(5), 1872–1911. doi: 10.1257/aer.101.5.1872

Angrist, J. D., Graddy, K. & Imbens, G. W. (2000, 07). The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish. *The Review of Economic Studies*, *67*(3), 499-527. Retrieved from `https://doi.org/10.1111/1467-937X.00141` doi: 10.1111/1467-937X.00141

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, *91*(434), 444–455.

Baum, C. F. & Lewbel, A. (2019). Advice on using heteroskedasticity-based identification. *The Stata Journal*, *19*(4), 757–767. doi: 10.1177/1536867X19893614

Bazinas, V. & Nielsen, B. (2022). Causal transmission in reduced-form models. *Econometrics*, *10*(2), 14. doi: 10.3390/econometrics10020014

Berenguer-Rico, V., Johansen, S. & Nielsen, B. (2021). A model where the least trimmed squares estimator is maximum likelihood. In *Book of abstracts* (p. 19).

Everitt, B. S. & Skrondal, A. (2010). *The Cambridge Dictionary of Statistics.* Cambridge University Press.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, 1–12.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The annals of mathematical statistics*, *42*(6), 1887–1896. doi: 10.1214/aoms/1177693054

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*(4), 1029–1054. Retrieved 2023-05-16, from `http://www.jstor.org/stable/1912775` doi: 10.2307/1912775

Harter, H. L. (1961). Expected values of normal order statistics. *Biometrika*, *48*(1/2), 151–165. Retrieved 2023-06-28, from `http://www.jstor.org/stable/2333139`

Hausman, J. A. (1983). Specification and estimation of simultaneous equation models. *Handbook of econometrics*, *1*, 391–448.

Heij, C., Heij, C., de Boer, P., Franses, P. H., Kloek, T., van Dijk, H. K. et al. (2004). *Econometric methods with applications in business and economics.* Oxford University Press.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101. Retrieved from `http://www.jstor.org/stable/2238020`

Huber, P. J. (1981). *Robust statistics.* John Wiley & Sons.

Lewbel, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics*, *30*(1), 67–80. doi: 10.1080/07350015.2012.643126

Mu, W. & Xiong, S. (2023). On huber's contaminated model. *Journal of Complexity*, *77*, 101745. doi: https://doi.org/10.1016/j.jco.2023.101745

Nachtigall, C., Kroehne, U., Funke, F. & Steyer, R. (2003). Pros and cons of structural equation modeling. *Methods Psychological Research Online*, *8*(2), 1–22.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association*, *79*(388), 871–880.

Rousseeuw, P. J. & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, *1*(1), 73–79. doi: 10.1002/widm.2

Rousseeuw, P. J. & Van Driessen, K. (2000). An algorithm for positive-breakdown regression based on concentration steps. *Data analysis: Scientific modeling and practical application*, 335–346. doi: 10.1007/978-3-642-58250-9_27

Rousseeuw, P. J. & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, *85*(411), 633–639. doi: 10.1080/01621459.1990.10474920

Tinbergen, J. (1930). Determination and interpretation of supply curves: an example. *Zeitschrift fur Nationalokonomie*, *1*(5), 669–679.

Wright, P. G. (1928). Tariff on animal and vegetable oils. *New York: Macmillan*.

A Proofs

In this section we demonstrate the claims made in the thesis.

A.1 Proof of proposition 1

We set P is the random variable $P = 1 - \frac{\psi_H(\epsilon_2/\sigma)}{\epsilon_2/\sigma}$ corresponding to the realisations p_i, we show that conditions stated in the proposition are fulfilled. We start with the first condition that $\text{cov}(P, \epsilon_1\epsilon_2) = 0$:

$$\begin{aligned}\text{cov}(P, \epsilon_1\epsilon_2) &= \mathrm{E}(P\epsilon_1\epsilon_2) - \mathrm{E}(P)\,\mathrm{E}(\epsilon_1\epsilon_2)\\ &\overset{indep.}{=} \mathrm{E}(\epsilon_1)\,\mathrm{E}(P\epsilon_2) - \mathrm{E}(P)\,\mathrm{E}(\epsilon_1)\,\mathrm{E}(\epsilon_2)\\ &= 0 - 0 = 0,\end{aligned}$$

and for the other condition that $\text{cov}(P, \epsilon_2^2) \neq 0$:

$$\begin{aligned}\text{cov}(P, \epsilon_2^2) &= \mathrm{E}(P\epsilon_2^2) - \mathrm{E}(P)\,\mathrm{E}(\epsilon_2^2)\\ &= \mathrm{E}\left(\left(1 - \frac{\psi_H(\epsilon_2/\sigma)}{\epsilon_2/\sigma}\right)\epsilon_2^2\right) - \mathrm{E}\left(1 - \frac{\psi_H(\epsilon_2/\sigma)}{\epsilon_2/\sigma}\right)\mathrm{E}(\epsilon_2^2),\end{aligned}$$

from here we show it first for ϵ_i with $i \in \zeta$ (these are standard normal distributed) and then for ϵ_j with $j \notin \zeta$ and let $\Phi(z)$ be the cumulative distribution function of the normal distribution:

$$\begin{aligned}\text{cov}(P, \epsilon_2^2) &= \mathrm{E}\left(\left(1 - \frac{\psi_H(\epsilon_2/\sigma)}{\epsilon_2/\sigma}\right)\epsilon_2^2\right) - \mathrm{E}\left(1 - \frac{\psi_H(\epsilon_2/\sigma)}{\epsilon_2/\sigma}\right)\mathrm{E}(\epsilon_2^2)\\ &= (1 - 2\Phi(-c/\sigma))\,\mathrm{E}\,(0) + \Phi(-c/\sigma)\,\mathrm{E}\left(\left(1 + \frac{c}{\epsilon_2}\right)\epsilon_2^2\right) +\\ &(1 - \Phi(c/\sigma))\,\mathrm{E}\left(\left(1 - \frac{c}{\epsilon_2}\right)\epsilon_2^2\right) - \left(\Phi(-c/\sigma)\,\mathrm{E}\left(1 + \frac{c}{\epsilon_2}\right)\right.\\ &\left. + (1 - \Phi(c/\sigma))\,\mathrm{E}\left(1 - \frac{c}{\epsilon_2}\right)\right)\mathrm{E}(\epsilon_2^2)\\ &= \Phi(-c/\sigma)\left(\mathrm{E}(\epsilon_2^2) + c\,\mathrm{E}\,(\epsilon_2)\right) + \Phi(-c/\sigma)\left(\mathrm{E}(\epsilon_2^2) - c\,\mathrm{E}\,(\epsilon_2)\right)\\ &- \left(\Phi(-c/\sigma)\left(1 + \mathrm{E}\left(\frac{c}{\epsilon_2}\right)\right) + \Phi(-c/\sigma)\left(1 - \mathrm{E}\left(\frac{c}{\epsilon_2}\right)\right)\right)\mathrm{E}\left(\epsilon_2^2\right)\\ &= 2\Phi(-c/\sigma)\,\mathrm{E}\left(\epsilon_2^2\right) - 2\Phi(-c/\sigma)\,\mathrm{E}\left(\epsilon_2^2\right)\\ &= 0.\end{aligned}$$

Now we present the derivation for the outlier observations. Let the distribution of the $\max \epsilon_{2_i}$ for $i \in \zeta$ be Φ^h and similarly for $\min \epsilon_{2_i}$ let it be distributed as $1 - (1 - \Phi)^h$ with $h = |\zeta|$ as these are order statistics. Then the CDF of $\max \epsilon_{2_i} + \xi_j$ for $j \notin \zeta$ is given by $h_+ = \Phi^h * s_j$ (s_j is the PDF of ξ_j) and similarly the CDF for $\min \epsilon_{2i} + \xi_j$ is given by $h_- = (1 - (1 - \Phi)^h) * s_j$ with H denoting the CDF. Beside this, we define S_j to be the CDF of the probability density function s_j. Also we define $\epsilon_2 > 0$ as ϵ_2^+ and $\epsilon_2 \leq 0$ as ϵ_2^-, then:

$$
\begin{aligned}
\operatorname{cov}(P,\epsilon_2^2) &= \mathrm{E}\Bigg(\bigg(1-\frac{\psi_H(\epsilon_2/\sigma)}{\epsilon_2/\sigma}\bigg)\epsilon_2^2\Bigg) - \mathrm{E}\bigg(1-\frac{\psi_H(\epsilon_2/\sigma)}{\epsilon_2/\sigma}\bigg)\mathrm{E}(\epsilon_2^2)\\
&= S_j\,(0)\,\Bigg(\big(1-H_-(c/\sigma)\big)\Bigg(\mathrm{E}\Bigg(\bigg(1-\frac{c}{\epsilon_2^-}\bigg)\epsilon_2^{-2}\Bigg)-\mathrm{E}\bigg(1-\frac{c}{\epsilon_2^-}\bigg)\mathrm{E}\Big(\epsilon_2^{-2}\Big)\Bigg)+\\
&\big(H_-(-c/\sigma)\big)\Bigg(\mathrm{E}\Bigg(\bigg(1+\frac{c}{\epsilon_2^-}\bigg)\epsilon_2^{-2}\Bigg)-\mathrm{E}\bigg(1+\frac{c}{\epsilon_2^-}\bigg)\mathrm{E}\Big(\epsilon_2^{-2}\Big)\Bigg)+\\
&\big(1-(1-H_-(c/\sigma))-H_-(-c/\sigma)\big)\Big(\mathrm{E}\Big(0\epsilon_2^{-2}\Big)-\mathrm{E}\,((0))\,\mathrm{E}\Big(\epsilon_2^{-2}\Big)\Big)+\\
&\big(1-S_j(0)\big)\Bigg(\big(1-H_+(c/\sigma)\big)\Bigg(\mathrm{E}\Bigg(\bigg(1-\frac{c}{\epsilon_2^+}\bigg)\epsilon_2^{+2}\Bigg)-\mathrm{E}\bigg(1-\frac{c}{\epsilon_2^+}\bigg)\mathrm{E}\Big(\epsilon_2^{+2}\Big)\Bigg)+\\
&\big(H_+(-c/\sigma)\big)\Bigg(\mathrm{E}\Bigg(\bigg(1+\frac{c}{\epsilon_2^+}\bigg)\epsilon_2^{+2}\Bigg)-\mathrm{E}\bigg(1+\frac{c}{\epsilon_2^+}\bigg)\mathrm{E}\Big(\epsilon_2^{+2}\Big)\Bigg)+\\
&\big(1-(1-H_+(c/\sigma))-H_+(-c/\sigma)\big)\Big(\mathrm{E}\Big((0)\epsilon_2^{+2}\Big)-\mathrm{E}\,(0)\,\mathrm{E}\Big(\epsilon_2^{+2}\Big)\Big)\\
&= S_j(0)\,\Bigg(\big(1-H_-(c/\sigma)\big)\Bigg(\mathrm{E}\Big(\epsilon_2^{-2}-c\epsilon_2^-\Big)-\bigg(1-\mathrm{E}\bigg(\frac{c}{\epsilon_2^-}\bigg)\bigg)\mathrm{E}\Big(\epsilon_2^{-2}\Big)\Bigg)+\\
&\big(H_-(-c/\sigma)\big)\Bigg(\mathrm{E}\Big(\epsilon_2^{-2}+c\epsilon_2^-\Big)-\bigg(1+\mathrm{E}\bigg(\frac{c}{\epsilon_2^-}\bigg)\bigg)\mathrm{E}\Big(\epsilon_2^{-2}\Big)\Bigg)\\
&+\big(1-S_j(0)\big)\Bigg(\big(1-H_+(c/\sigma)\big)\Bigg(\mathrm{E}\Big(\epsilon_2^{+2}-c\epsilon_2^+\Big)-\bigg(1-\mathrm{E}\bigg(\frac{c}{\epsilon_2^+}\bigg)\bigg)\mathrm{E}\Big(\epsilon_2^{+2}\Big)\Bigg)+\\
&\big(H_+(-c/\sigma)\big)\Bigg(\mathrm{E}\Big(\epsilon_2^{+2}+c\epsilon_2^+\Big)-\bigg(1+\mathrm{E}\bigg(\frac{c}{\epsilon_2^+}\bigg)\bigg)\mathrm{E}\Big(\epsilon_2^{+2}\Big)\Bigg)\\
&= S_j(0)\,\Bigg(\big(1-H_-(c/\sigma)-H_-(-c/\sigma)\big)\Bigg(-c\,\mathrm{E}\left(\epsilon_2^-\right)+c\,\mathrm{E}\bigg(\frac{1}{\epsilon_2^-}\bigg)\mathrm{E}\Big(\epsilon_2^{-2}\Big)\Bigg)\Bigg)+\\
&\big(1-S_j(0)\big)\Bigg(\big(1-H_+(c/\sigma)-H_+(-c/\sigma)\big)\Bigg(-c\,\mathrm{E}\left(\epsilon_2^+\right)+c\,\mathrm{E}\bigg(\frac{1}{\epsilon_2^+}\bigg)\mathrm{E}\Big(\epsilon_2^{+2}\Big)\Bigg)\Bigg)
\end{aligned}
$$

We now assess whether this expression will not equal zero. The term $S_j(0)$ is bounded by $0 < S_j < 1$ or else equals either $S_j = 0$ or $S_j = 1$. Such that always one of the two terms in the last expression is non-zero. More important are the terms $(1 - H_+(c/\sigma) - H_+(-c/\sigma))$ and $(1 - H_-(c/\sigma) - H_-(-c/\sigma))$, in symmetric cases these expressions will equal zero. However, we know that H_- and H_+ must generally be non-zero. This is due to the set-up of the H distribution. For H_-, the CDF of $\min_{i\in\zeta}\epsilon_i + \xi_j$, we know that it is composed of the first order statistic of the normal distribution and another unknown distribution. At least for the first order statistic of the normal distribution (and similarly for the largest order statistic), we know the distribution is non-symmetric. So only in a specific case is the sum of this order statistic with ξ_j symmetric. Therefore, in general H_- and H_+ are non-symmetric. We should impose a condition on the s_j distributions to prevent the result of a symmetric distribution. Thus, therefore we impose that the distribution of s_j must be such that $(1-(1-\Phi)^h) * s_j$ is non-symmetric. Then, we can conclude that $(1 - H_+(c/\sigma) - H_+(-c/\sigma))$ and $(1 - H_-(c/\sigma) - H_-(-c/\sigma))$ are non-zero. Finally, we address the terms $\left(-c\,\mathrm{E}\left(\epsilon_2^-\right) + c\,\mathrm{E}\left(\frac{1}{\epsilon_2^-}\right)\mathrm{E}\left(\epsilon_2^{-2}\right)\right)$ and $\left(-c\,\mathrm{E}\left(\epsilon_2^+\right) + c\,\mathrm{E}\left(\frac{1}{\epsilon_2^+}\right)\mathrm{E}\left(\epsilon_2^{+2}\right)\right)$, the expectation of $\mathrm{E}\left(\epsilon_2^-\right)$ and $\mathrm{E}\left(\epsilon_2^+\right)$ is again generally non-zero. This due to the fact of the order statistics again. Without ξ_j (thus distribution s_j) exactly counterbalancing the expected value of the first/largest normal order statistic the expectation is non-zero, see also (Harter, 1961). So, we should also assume that distribution s_j is such that $\mu_{(1-(1-\Phi)^h)} + \mu_{s_j} \neq 0$. The last terms $c\,\mathrm{E}\left(\frac{1}{\epsilon_2^-}\right)\mathrm{E}\left({\epsilon_2^-}^2\right)$ and $c\,\mathrm{E}\left(\frac{1}{\epsilon_2^+}\right)\mathrm{E}\left({\epsilon_2^+}^2\right)$ are more complex to analyze. In general they will not be zero, but the main concern is that they exactly cancel out with $-c\,\mathrm{E}\left(\epsilon_2^-\right)$ and $-c\,\mathrm{E}\left(\epsilon_2^+\right)$ respectively. However, it is safe to state that in general $\mathrm{E}\left(\frac{1}{\epsilon_2^+}\right)\mathrm{E}\left({\epsilon_2^+}^2\right) \neq \mathrm{E}\left(\epsilon_2^+\right)$ and $\mathrm{E}\left(\frac{1}{\epsilon_2^-}\right)\mathrm{E}\left({\epsilon_2^-}^2\right) \neq \mathrm{E}\left(\epsilon_2^-\right)$. So that $\mathrm{cov}(P, \epsilon_2^2) \neq 0$, at least in general (in some specific cases it might turn out to be 0).

B Programming code

The programming code used for the thesis was coded in Python 3, using the numpy and scipy libraries as well as the pandas library for the applied study and matplotlib for figure 1. The code is provided in a Jupyter Notebook and by running the cells sequentially, all results can be obtained. For the replication, all code until the markdown block 'extension' can be run and this will supply all the results presented in the thesis. For the extension, the same applies for the simulation study. But to run the different models, you should run everything from 'extension' till applied study again with another uncommented model in the block 'Outlier generation code'. This will be clear in the Jupyter Notebook. For the applied study, run everything after 'applied study' until 'outlier picture'. The results are obtained in the last cell before 'outlier picture', by commenting the exclusion of Chile and Zimbabwe the full model can be run. To run the model for the different dependent variables, change the dependent variable in the line indicated by the comment.

C Other result tables

Here we present tables with other results that were not relevant for the main text. We start with the results of the Huber and Binary simulation studies and their 10% outlier simulations.

Table 15: Simulation Results Huber model (2) with 10% outliers

parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
β_{11}	1.00	1.00	0.067	0.957	1.00	1.05	0.067	0.053	0.044
β_{12}	1.00	1.00	0.059	0.961	1.00	1.04	0.059	0.047	0.040
γ_1	1.00	1.00	0.038	0.974	1.00	1.03	0.038	0.030	0.025
β_{21}	1.00	1.31	0.070	1.258	1.30	1.35	0.313	0.305	0.305
β_{22}	1.00	1.00	0.072	0.952	1.00	1.05	0.072	0.058	0.049

Note: MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

Table 16: Simulation Results Huber model (3) with 10% outliers

parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
β_{11}	1.00	1.00	0.052	0.966	1.00	1.03	0.052	0.041	0.034
β_{12}	1.00	1.00	0.052	0.966	1.00	1.04	0.052	0.041	0.035
γ_1	1.00	1.00	0.025	0.986	1.00	1.01	0.025	0.018	0.013
β_{21}	1.00	1.00	0.126	0.925	1.00	1.08	0.126	0.097	0.077
β_{22}	1.00	1.00	0.125	0.924	1.00	1.07	0.125	0.095	0.074

Note: MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

Table 17: Simulation Results Binary (1) with 10% outliers

parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
β_{11}	1.00	1.00	0.060	0.960	1.00	1.04	0.060	0.048	0.040
β_{12}	1.00	1.00	0.060	0.959	1.00	1.04	0.060	0.048	0.041
γ_1	1.00	1.00	0.040	0.973	1.00	1.03	0.040	0.032	0.027
β_{21}	1.00	1.00	0.077	0.949	1.00	1.05	0.077	0.061	0.051
β_{22}	1.00	1.00	0.069	0.954	1.00	1.05	0.069	0.055	0.046

Note: MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

Table 18: Simulation Results Binary (2) with 10% outliers

parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
β_{11}	1.00	1.00	0.066	0.956	1.00	1.04	0.066	0.053	0.044
β_{12}	1.00	1.00	0.059	0.961	1.00	1.04	0.059	0.047	0.039
γ_1	1.00	1.00	0.038	0.974	1.00	1.02	0.038	0.030	0.025
β_{21}	1.00	1.31	0.070	1.258	1.30	1.35	0.313	0.305	0.305
β_{22}	1.00	1.00	0.072	0.952	1.00	1.05	0.072	0.058	0.049

Note: MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

Table 19: Simulation Results Binary (3) with 10% outliers

parameter	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
β_{11}	1.00	1.00	0.052	0.966	1.00	1.04	0.052	0.041	0.034
β_{12}	1.00	1.00	0.052	0.966	1.00	1.03	0.052	0.041	0.034
γ_1	1.00	1.00	0.025	0.986	1.00	1.01	0.025	0.018	0.013
β_{21}	1.00	1.00	0.126	0.925	1.00	1.08	0.126	0.097	0.077
β_{22}	1.00	1.00	0.125	0.924	1.00	1.07	0.125	0.095	0.074

Note: MEAN and SD are the mean and standard deviation of the estimates across simulation. LQ, MED and UQ are the 25%, 50% and 75% quantiles. RMSE, MAE and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.