Robust Estimation of Probit Models with Endogeneity

Andrea A. Naghi[1]
Máté Váradi[1]
Mikhail Zhelonkin[1]

[1] Erasmus University Rotterdam

Robust Estimation of Probit Models with Endogeneity

Andrea A. Naghi[*], Máté Váradi[†], Mikhail Zhelonkin[‡]

December 2020

Abstract

Probit models with endogenous regressors are commonly used models in economics and other social sciences. Yet, the robustness properties of parametric estimators in these models have not been formally studied. In this paper, we derive the influence functions of the endogenous probit model's classical estimators (the maximum likelihood and the two-step estimator) and prove their non-robustness to small but harmful deviations from distributional assumptions. We propose a procedure to obtain a robust alternative estimator, prove its asymptotic normality and provide its asymptotic variance. A simple robust test for endogeneity is also constructed. We compare the performance of the robust and classical estimators in Monte Carlo simulations with different types of contamination scenarios. The use of our estimator is illustrated in several empirical applications.

Keywords: Binary outcomes, Probit model, Endogenous variable, Instrumental variable, Robust Estimation

1 Introduction

Probit models are widely used binary outcome specifications in economics and social science studies (see for example Cameron and Trivedi (2005), Wooldridge (2010) for a standard textbook treatment). To account for common empirical challenges driven by endogeneity issues, most empirical applications resort to probit models that accommodate endogenous regressors.

An endogenous probit model can be represented by the regression system

$$y_{1i} = x_{1i}^\top \gamma_1 + x_{2i}^\top \gamma_2 + \epsilon_{1i}, \tag{1}$$

$$y_{2i} = \mathcal{I}(x_{1i}^\top \beta + \alpha y_{1i} + \epsilon_{2i} \geq 0), \tag{2}$$

where x_{1i} is a $p_1 \times 1$ vector of exogenous regressors, x_{2i} is a $p_2 \times 1$ vector of instrumental variables (IV) and y_{1i} is an endogenous continuous regressor. The error terms ϵ_{1i} and ϵ_{2i} follow a bivariate normal distribution with variances σ_1^2, $\sigma_2^2 = 1$ and correlation ρ, $\mathcal{I}(A)$ denotes an indicator function which is equal to one if A is true and zero otherwise, and $i = 1, \ldots, N$. For simplicity of exposition we consider the case of one endogenous variable (Wooldridge 2010, Section 15.7.2). The intercepts are included in γ_1 in equation (1) and in β in equation (2). Note that for identification we need $p_2 \geq 1$. Also, x_2 needs to be a significant predictor of y_1 and needs to be uncorrelated with the error terms.

Equation (1) is a reduced form equation which explains the endogenous variable in terms of strictly exogenous variables including instruments. The IV's are excluded from (2), which is the equation of interest. The model (1)-(2) is typically estimated by the Maximum Likelihood Estimator (MLE) or by two-step procedures. A well used two-step procedure, the estimator of Rivers and Vuong (1988), employs the Ordinary Least Squares (OLS) estimator for (1), then the probit MLE is fitted for (2). In the MLE step, the first step residuals $\hat{\epsilon}_1$ are used as an additional regressor, i.e., the linear predictor in (2) becomes $\alpha y_{1i} + x_{1i}^\top \beta + \hat{\epsilon}_{1i} \lambda$. An important virtue of this approach is that it leads to a simple test of exogeneity of y_1, by testing the

null $H_0 : \lambda = 0$. The above mentioned parametric estimators heavily rely on distributional assumptions. Unfortunately, even when the normality assumption is approximately correct, the estimators and tests can become unstable. If there are outliers in the data, the classical parametric estimators can become arbitrarily biased (see for example Ronchetti and Trojani (2001) for the context of GMM estimation, Freue et al. (2013) for the instrumental variable setting, and Zhelonkin et al. (2016) for the sample selection model).

In this paper we study the robustness properties (to small data contamination) of parametric estimators of the endogenous probit model: the MLE and the two-step estimator of Rivers and Vuong (1988). As it turns out that these estimators are not robust, we propose robust alternatives. Throughout this paper we assume that the parametric model holds approximately, i.e., a certain (small) proportion of data can come from an unknown contaminating model. The simplest example of this type of contamination is contamination by outliers. More specifically, we assume that the true data generating process is the Huber (1964) gross-error model $F_\varepsilon = (1 - \varepsilon)F + \varepsilon G$, where ε denotes a (small) contamination proportion. Then, we are interested in the estimation of the model F, which is the model (1)-(2), although the data can come from F_ε. We treat G as an unknown/arbitrary contamination. In a recent paper, Bonhomme and Weidner (2018) refer to the gross-error model from above as nonparametric contamination.

Note that if a practitioner concludes that the model F is not applicable even approximately, then the alternative would be to use a semi- or nonparametric approach, see Blundell and Powell (2004). The semi- or nonparametric approaches directly estimate F_ε. Although these methods are more flexible, the flexibility comes at a cost of identification, efficiency and computational difficulties. Our robust approach is parametric in nature and hence it identifies exactly the same parameters as a fully parametric case. The price to pay is a small loss of efficiency if the model F holds exactly. However, in the presence of contamination, as we show in the simulation

study below, the classical estimators become less efficient than the robust alternative.

By means of the influence function approach (Hampel et al. 1986) we prove the non-robustness of classical estimators of the endogenous probit model in Section 2. In Section 3 we propose a robust two-step estimator and, as a by-product, a robust test for endogeneity. Section 4 analyzes the performance of the classical and robust estimators under different types of contamination. Three empirical applications in Section 5 illustrate the use of the proposed estimator. Finally, Section 6 concludes.

2 Robustness Properties

In this section we investigate the robustness properties of classical endogenous probit estimators. To this end, we start with deriving the influence function (IF) of the two-step estimator, then we briefly discuss the properties of the IF of the maximum likelihood estimator.

We consider a parametric probit model with endogeneity (1)-(2), denoted by F_θ, where $\theta = (\alpha, \beta, \gamma_1, \gamma_2, \rho, \sigma_1)$ lies in Θ, a compact subset of $\mathbb{R} \times \mathbb{R}^{2p_1} \times \mathbb{R}^{p_2} \times [0,1] \times \mathbb{R}^+$. Let F_N be the empirical distribution function with mass $1/N$ at each observation $z_i = (x_{1i}, x_{2i}, y_{1i}, y_{2i})$, with $i = 1, \ldots, N$. Let F be the distribution function of z_i. To make the notation more concise, let us denote $z_{1i} = (x_{1i}, x_{2i}, y_{1i})$ and $z_{2i} = (x_{1i}, y_{1i}, y_{2i})$.

The two-step estimator can be expressed as a solution of the empirical counterpart of the following system of M-estimators

$$\int \Psi_1\{z_1; S(F)\}dF = 0, \tag{3}$$

$$\int \Psi_2[z_2; h\{z_1; S(F)\}, T(F)]dF = 0, \tag{4}$$

where Ψ_1 and Ψ_2 denote the score functions of the first and second step estimators, respectively. In the case of the Rivers and Vuong (1988) estimator, the score functions are

$$\Psi_1\{z_1; S(F)\} = (y_1 - x_1^\top \gamma_1 - x_2^\top \gamma_2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \tag{5}$$

4

$$\Psi_2[z_2; h\{z_1; S(F)\}, T(F)] = \{y_2 - \Phi(\tilde{x}^\top \delta)\} \frac{\phi(\tilde{x}^\top \delta)}{\Phi(\tilde{x}^\top \delta)\{1 - \Phi(\tilde{x}^\top \delta)\}} \tilde{x}, \tag{6}$$

where $\tilde{x} = (x_1^\top, y_1, h)^\top$, $\delta = (\beta^\top, \alpha, \lambda)^\top$, and h is a function that connects two estimation steps, i.e., $h\{z_1; S(F)\} = y_1 - x_1^\top \gamma_1 - x_2^\top \gamma_2$, which in our case expresses the residuals in the first step estimator. The density and the cumulative distribution function of the standard normal distribution are denoted by ϕ and Φ, respectively.

2.1 Influence Function

For a statistical functional $T(F)$ the influence function (IF) is defined (Hampel 1974) as

$$\text{IF}(z; T, F) = \lim_{\varepsilon \to 0}[T\{(1 - \varepsilon)F + \varepsilon\Delta_z\} - T(F)]/\varepsilon,$$

where Δ_z is a point mass at z. The IF characterizes the standardized asymptotic bias of the estimator due to contamination ε. If the IF is unbounded then the worst possible bias in the ϵ neighborhood of F can be infinite. Hence, for an estimator to be (locally) robust, a bounded IF is required.

Let us start with the IF of the first step. The IF for the OLS estimator is known (Hampel et al. 1986) and is given by

$$\text{IF}(z; S, F) = M_1 \Psi_1\{z_1; S(F)\},$$

where M_1 is a constant matrix given by

$$M_1 = -\int \begin{pmatrix} x_1 x_1^\top & x_1 x_2^\top \\ x_2 x_1^\top & x_2 x_2^\top \end{pmatrix} dF.$$

The IF is clearly unbounded in both dependent variable y_1 and covariate space (x_1, x_2).

Now, the estimator of the second step depends on an outcome of the first step. Thus, the IF of the estimator of the second step depends on the IF of the first step. The connection is formulated in the following proposition. The derivations of Proposition 1 are given in Appendix A.

5

Proposition 1. *For model (1)-(2), the IF of the Rivers-Vuong two-step estimator is*

$$IF(z; T, F) = M_2^{-1} \left(\Psi_2[(x_1, y_1, y_2); h\{(x_1, x_2, y_1); S(F)\}, T(F)] + \int AdF \cdot IF(z; S, F) \right), \quad (7)$$

where M_2 is a probit Hessian matrix and $\int AdF$ is a constant matrix given in the Appendix.

Note that the score function Ψ_2 is unbounded. In addition, the second step estimator also inherits unboundedness of the IF from the first step, since the IF of the second step linearly depends on the IF of the first step. Hence, the entire two-step estimator is not robust.

Remark 1: Notice that the IF of the full information MLE is also unbounded. For completeness, the score functions are given in Appendix B. In this case the estimation of the parameters of the two equations (1)-(2) is not separable. Thus, the robustification of this estimator is much more complex than that of the two-step estimator as it requires complicated consistency corrections which can lead to numerical instability. Hence, in this paper, we proceed with the robust version of the two-step estimator.

3 Robust Estimation and Test for Endogeneity

In this section we propose a robust estimator for the endogenous probit model and construct a simple robust test for endogeneity.

3.1 Robust Two-Step Estimator

The expression of the IF of the two-step estimator derived in Proposition 1 formalizes the non-robustness problem and paves the way to construct a robust estimator. The IF in (7) contains two sources of unboundedness. The first source is the IF of the estimator of the first step, while the second is the unbounded score function of the probit MLE. Thus, it is clear that as the first step we need to have a bounded-influence estimator for the linear regression. The topic of robustness in linear regressions is well-studied (Hampel et al. 1986, Maronna et al.

2006) and there are a large number of robust alternatives to OLS. One can use for example the MM-estimator (Yohai 1987) or an M-estimator of Mallows type. With the M-estimator of Mallows type the deviations in the dependent variable y_1 and covariates (x_1, x_2) are bounded separately. The score function is as follows

$$\Psi_1^r\{z_1; S(F)\} = \psi_c(y_1 - x_1^\top \gamma_1 - x_2^\top \gamma_2)\omega_1(x_1, x_2), \tag{8}$$

where ψ_c is the Huber function defined by

$$\psi_c(e) = \begin{cases} e, & |e| \leq c, \\ c \cdot \mathrm{sgn}(e), & |e| > c, \end{cases} \tag{9}$$

and where the tuning constant c can be chosen to ensure a certain level of asymptotic efficiency (we used $c = 1.345$, as recommended in Hampel et al. (1981)). The weight function ω_1 is based on the robust Mahalanobis distance $d(x_1, x_2)$:

$$\omega_1(x_1, x_2) = \begin{cases} (x_1, x_2)^\top, & \text{if } d(x_1, x_2) \leq \tilde{c}, \\ (x_1, x_2)^\top \tilde{c}/d(x_1, x_2), & \text{if } d(x_1, x_2) > \tilde{c}. \end{cases} \tag{10}$$

Given that the squared Mahalanobis distance follows a χ^2-distribution, we use the 5% critical level for the choice of $\tilde{c}$. The expectation and covariance matrix for the robust Mahalanobis distance are estimated by the Minimum Covariance Determinant (MCD), see Rousseeuw and Van Driessen (1999).

In the second estimation step we need to use the estimated error terms from step one. Since the first step can be estimated robustly, we can then compute the error $\hat{h} = y_1 - x_1^\top \hat{\gamma}_1 - x_2^\top \hat{\gamma}_2$ without modifications. The idea is that if there are large errors due to outliers or misspecification, then they will be transferred to the second stage, where a robust estimator needs to be used, as the score function is unbounded. For the second step we propose to replace the probit MLE by the robust version of the quasi-likelihood estimator, which is also a Mallows-type M-estimator. We use the procedure proposed by Cantoni and Ronchetti (2001) for generalized linear models. The score function is

$$\Psi_2^r\{z; h, T(F)\} = \nu(z_2; \mu)\omega_2(\tilde{x})\mu' - a(\delta), \tag{11}$$

7

where $a(\delta) = (1/N) \sum_{i=1}^{N} E\{\nu(z_{2i}; \mu_i)\}\omega_2(\tilde{x}_i)\mu'$ ensures Fisher consistency with the expectation taken with respect to the conditional distribution of $y_2|\tilde{x}$. The weight functions $\nu(\cdot, \cdot)$ and $\omega_2(\cdot)$ are defined below, $\mu_i = \Phi(\tilde{x}_i^\top \delta)$ and $\mu_i' = \partial\mu_i/\partial\delta$. The weight function

$$\nu(z_{2i}; \mu_i) = \psi_c(r_i)\frac{1}{V^{1/2}(\mu_i)}, \tag{12}$$

where $r_i = (y_{2i} - \mu_i)/V^{1/2}(\mu_i)$ are Pearson residuals, $V(\mu_i) = \Phi(\tilde{x}_i^\top \delta)\{1 - \Phi(\tilde{x}_i^\top \delta)\}$, and ψ_c is a Huber function defined in (9) with a possibly different tuning constant. As advocated by Cantoni and Ronchetti (2001), the suitable value for c is 1.345. The weights on the covariate space $\omega_2(\cdot)$ can be computed in the same way as in the first step (10), i.e., based on the robust Mahalanobis distance. A second option is $\omega_2(\tilde{x}_i) = \sqrt{1 - H_{ii}}$, where H_{ii} is the i'th diagonal element of the hat matrix $H = \tilde{X}(\tilde{X}^\top \tilde{X})^{-1}\tilde{X}^\top$. Notice that the weights based on the hat matrix were a first option suggested by Cantoni and Ronchetti (2001). In the case of the endogenous probit model, our simulation study suggests that the weights based on the robust Mahalanobis distances provide a more robust performance with respect to data contamination. However, applications in economics and social science often have covariate spaces that contain a lot of dummy variables. This can make the MCD's work computationally unstable. Hence, the hat matrix-based weights can still be viable alternatives. Finally, the expectation $E\{\psi_c(r_i)\}$ in the $a(\delta)$ term in (11) can be computed explicitly and equals to

$$E\left[\psi_c\left\{\frac{y_{2i} - \mu_i}{V^{1/2}(\mu_i)}\right\}\right] = \psi_c\left\{\frac{-\mu_i}{V^{1/2}(\mu_i)}\right\}\{1 - \Phi(\tilde{x}_i^\top \delta)\} + \psi_c\left\{\frac{1 - \mu_i}{V^{1/2}(\mu_i)}\right\}\Phi(\tilde{x}_i^\top \delta). \tag{13}$$

The properties of the proposed estimator are formalized in the following proposition.

Proposition 2. *Under the regularity condition stated in Appendix C, the two-step estimator defined by the score functions (8) and (11) is robust, consistent and asymptotically normal, with asymptotic variance defined by*

$$V(T, F) = M\left(\Psi_2^r\right)^{-1}\int \{\Upsilon_1(z)\Upsilon_1(z)^\top + \Upsilon_2(z)\Upsilon_2(z)^\top\}dFM\left(\Psi_2^r\right)^{-1}, \tag{14}$$

where

$$M(\Psi_2^r) = -\int \frac{\partial}{\partial \delta} \Psi_2^r(z; h, T) dF,$$

$\Upsilon_2(z) = \Psi_2^r(z; h, T)$, *and*

$$\Upsilon_1(z) = \int \frac{\partial}{\partial \zeta} \Psi_2^r \{z_2; \zeta, S(F)\} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} dF \left\{ \int \frac{\partial}{\partial \gamma} \Psi_1^r(z_1; \gamma) dF \right\}^{-1} \Psi_1^r(z_1; S).$$

The proof of Proposition 2 is given in Appendix C.

3.2 Testing for Endogeneity

When using the two-step estimator, testing for the endogeneity of y_1 is straightforward. It is a standard t-test testing if the coefficient λ is equal to zero. Denote the test statistic by $\tau = \sqrt{N} T(F)/V(T, F)^{1/2}$, where $V(T, F)$ denotes the asymptotic variance of the estimator at the model.

We study the behavior of the test statistic under contamination. We make a von Mises (1947) expansion of the test statistic:

$$\sqrt{N} \frac{T(F_\varepsilon)}{V(T, F_\varepsilon)^{1/2}} = \sqrt{N} \frac{T(F)}{V(T, F)^{1/2}} + \varepsilon \left\{ \sqrt{N} \frac{IF(z; T, F)}{V(T, F)^{1/2}} + \frac{\sqrt{N}}{2} T(F) \frac{CVF(z; T, F)}{V(T, F)^{3/2}} \right\} + o(\varepsilon),$$

(15)

where $CVF(z; T, F)$ is the change-of-variance function (Hampel et al. 1981). The change-of-variance function is analogous to the IF in the sense that it quantifies the influence of a small amount of contamination on the variance of an estimator. The von Mises expansion in (15) provides an approximation for the bias of the test statistic at the contaminated model. The behavior of the t-test statistic was studied in different contexts (Avella-Medina 2020, Zhelonkin et al. 2016). It is clear that the bias of the test statistic depends on the IF and CVF, however, the term containing the CVF is of higher order. Hence, from a practical perspective, for a stable performance of the test, the estimator with a bounded IF is required, while the boundedness of the CVF is desirable.

9

Remark 2: In the interest of brevity, we do not make a complete derivation of the CVF for our problem. The derivation of the CVF for general two-step M-estimators can be found in Zhelonkin (2013). In our case, the CVF depends on the score functions of both estimators and on the second derivatives of the score from the second step. Thus, it follows that the Rivers and Vuong (1988) estimator has an unbounded CVF. Our proposed robust estimator has bounded score functions, and hence possesses a more stable variance estimator.

As a simple robust alternative to test endogeneity, i.e., $H_0 : \lambda = 0$ against $H_A : \lambda \neq 0$, we propose to use a t-test based on the robust estimator of λ. The standard error can be obtained by estimation of the asymptotic variance in (14). The natural estimator is the sandwich estimator (Eicker 1967, Huber 1967, White 1980). The second term in (14) corresponds to the asymptotic variance of the one-step probit robust quasi-likelihood estimator and is estimated by

$$\hat{M}\left(\Psi_2^r\right)^{-1} \frac{1}{N} \sum_{i=1}^{N} \hat{\Upsilon}_2(z_i)\hat{\Upsilon}_2(z_i)^\top \hat{M}\left(\Psi_2^r\right)^{-1},$$

where $\hat{M}\left(\Psi_2^r\right)$ and $\hat{\Upsilon}_2(z)$ denote the sample versions of $M\left(\Psi_2^r\right)$ and $\Upsilon_2(z)$, respectively. The first term in (14) contains the asymptotic variance of the robust linear regression estimator given by (8), premultiplied and postmultiplied by the constant matrix as follows

$$
\begin{aligned}
\hat{M}\left(\Psi_2^r\right)^{-1} \frac{1}{N} \sum_{i=1}^{N} \hat{\Upsilon}_1(z_i)\hat{\Upsilon}_1(z_i)^\top \hat{M}\left(\Psi_2^r\right)^{-1} \;=\; & \hat{M}\left(\Psi_2^r\right)^{-1} \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \Psi_{2i}^r}{\partial \gamma} \\
& \times \widehat{\mathrm{Var}}(S, F) \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\partial \Psi_{2i}^r}{\partial \gamma} \right)^\top \hat{M}\left(\Psi_2^r\right)^{-1},
\end{aligned}
$$

where

$$\frac{\partial \Psi_{2i}^r}{\partial \gamma} = \frac{\partial \Psi_2^r \{z_{2i}; h, T(F)\}}{\partial h} \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix}.$$

4 Simulation Study

In this simulation study we illustrate the performance of the proposed robust estimator relative to the the non-robust alternatives: the MLE and the classical two-step estimator. We focus on

a setup with no contamination and then on two simple contamination scenarios: contamination by outliers and contamination of the distribution of the errors.

First we generate the uncontaminated data. To this end, we generate one exogenous variable x_1 and one instrument x_2 independently from each other following a standard normal distribution. The error terms follow a bivariate normal distribution with variances equal to 1 and correlation 0.5. The parameters are $\gamma_1 = (0, 1)^\top$, $\gamma_2 = 0.5$, $\alpha = 1$, and $\beta = (0, 1)^\top$. The sample size is $N = 2000$ and we repeat the study 500 times.

The results of the estimation without contamination are reported in Table 1, columns 2 and 3 and in Figure 1. Both the non-robust estimators (MLE and two-step) and the robust estimator (as described in Section 3.1) perform well. The biases are close to zero. As expected from the theory, the robust estimator is less efficient - the efficiency loss is approximately 15% relative to the two-step estimator. We have also tried the MM-estimator instead of the Mallows-type for the first step regression and the performance is similar (not reported here).

In order to study the robustness of the estimators we now add contamination. With probability $\varepsilon = 0.01$ we replace the original observations by outliers. We consider two types of outliers: first when the point mass is at $(x_1, x_2, y_1, y_2) = (2, 2, -2, 1)$; second when the point mass is at $(x_1, x_2, y_1, y_2) = (-2, -2, 2, 0)$. Although the setup is simple, the detection of contamination is not straightforward. The leverage outliers in the covariate space of the reduced form equation are two standard deviations away from the mean, which is not straightforward to detect with standard exploratory analysis. The endogenous variable y_1 has a standard deviation of approximately 1.5. Moreover, x_1 in the equation of interest becomes a good leverage point, while it was a bad leverage point in the reduced form equation.

The results of the estimation with this first contamination are summarized in Table 1, columns 4-7 and Figure 2 and Figure 3. It is clear that both non-robust estimators are biased, while the robust alternative is nearly unbiased. As expected from the theory, the robust

estimator allows for a small bias, but it is controlled and it looks negligible. The efficiency of the non-robust estimators is also affected. In the estimation of the reduced form equation, the non-robust estimators become less efficient than the robust estimator.

The second contamination scenario is the contamination of the distribution of the error terms. The setup is the same as above, but with probability 10% we generate error terms from a bivariate $t-$distribution with 2 degrees of freedom. We present the results in Figure 4 and Table 2. This contamination is less severe than the one by outliers. However, it is clear that our robust estimator outperforms the MLE and the two-step estimator. Notice that the non-robust estimators become less efficient (for many parameters) than our robust estimator.

5   Empirical Applications

In this section we illustrate the use of the robust estimator in three different settings. The first two examples revisit empirical studies about the causal effect of education on the body mass index (BMI), and about the impact of immigration on US natives' decision to major in science and technology. The third example uses a data set employed in Stata to illustrate the estimation of probit models with a continuous endogenous regressor. We present the estimation results with both the proposed robust estimator and the classical two-step estimator. In the first application, the results given by the classical and robust estimator are the same; in the second application we observe mild differences between the two sets of results; in the final example the robust estimates are clearly different from the classical estimates.

5.1   The Causal Effect of Education on the Body Mass Index

Our first empirical application revisits the paper by Brunello et al. (2013) on the causal impact of years of schooling on adult BMI. The authors argue that the main empirical challenges in estimating the causal effect of education on health (for which the BMI constitutes a proxy)

is reverse causality (improved health, as a result of a better BMI, can reduce dropout rates and improve cognitive skills); or the existence of some unobservable variables that affect both education and health (such as parental background or attitude towards risk). To overcome these challenges, the causal impact is identified by the exogenous variation induced by the number of compulsory years of education in a given country. Control variables include: age and its square, dummy variables for year-of-birth, survey year and country, and second-order polynomials in age interacted with country dummies – to capture smooth changes over time in education and BMI.

Brunello et al. (2013) focus on European countries and pool together data from the European Community Household Panel (ECHP), the Survey of Health, Ageing, and Retirement in Europe (SHARE), the English Longitudinal Survey of Ageing (ELSA), the German Socioeconomic Panel (SOEP), and the British Household Panel Survey. The authors start with estimating ordinary least squares (OLS) and two-stage least squares (2SLS) regressions for males and females separately and find that education has a protective role on the BMI for females, but no significant effect is found for males. Next, they estimate the probability of being overweight and obese using probit models where they treat years of schooling as both exogenous and endogenous. To this end, they construct a dependent variable which takes value one when the BMI is greater than 25 for the overweight category, and greater than 30 for the obese category. The results indicate that education reduces the probability of being overweight, for women. Finally, they delve into the potential mechanism that could explain the protective effect of education and the gender differences.

We focus here on their endogenous probit specifications and revisit these with the two-step estimator of Rivers and Vuong (1988)) and our proposed robust estimator. We report the results on the main variable of interest: number of years of schooling. In this empirical example we find that the classical and robust estimates (as well as their standard errors) are

close to each other for both females and males and for both overweight and obese groups. The robust estimates confirm the results of the original paper and the significance of the coefficients remains the same. Thus, for this first empirical application, we can conclude that there are no violations of the distributional assumptions and the classical estimator is reliable.

5.2 The Impact of Immigration on US Natives' Decision to Major in Science and Engineering

As a second example, we revisit the paper by Orrenius and Zavodny (2015) which examines whether immigration affects US natives' decision to major in STEM (science, technology, engineering, and mathematics) fields. The effect can go both ways: immigrants may crowd out US natives from STEM fields or they can have positive spillovers (such as positive peer effects, as foreign students tend to be among the best in their home countries) on US natives and thus attract or retain them in those fields. The paper uses US data from the American Community Survey (ACS) that records information on college majors, demographics and labor market outcomes, as well as Census data to compute immigrant share measures.

To answer the question of how the probability of majoring in a STEM field is related to immigrant shares, the authors use linear probability models (LPM) where the dependent variable is equal to one if the individual majored in a STEM field. Two measures of immigrant share variables are used, separately: immigrant share while in high school and immigrant share while in college. In the original paper, the immigrant share measures are merged with the ACS data by individuals' cohort and their state of birth. The control variables include: age and its square, dummy variables for race/ethnicity, state and cohort fixed effects, measures of the relative labor market attractiveness of STEM jobs (the fraction of STEM employed college graduates to non-STEM employed graduates, the change in this fraction in the past 10 years, the ratio of log average STEM annual earning to non-STEM earning, and the change in this ratio in the past 10 years). Additionally, some specifications also include state-specific linear

time trends to control for smooth changes within states that could affect the probability of US natives to major in STEM fields.

An important concern in the study is that the variable immigrant shares could be endogenous. Factors that determine immigrants to choose to live in a state, may also affect natives' choice to major in a STEM field (for example, the Internet boom in the Silicon Valley in the 1990s). If the controls from above do not fully account for such factors, the OLS estimates will be biased. Thus, the authors present results based on two instruments: the 1960 distribution of immigrant across states and the lagged immigrant composition within states.[1] The main results indicate that women are less likely to major in STEM fields, when the immigrants share is higher in their college cohort. In contrast, the effect on men is not significant.

We revisit with our robust estimation approach the regression specifications for females with state and cohort fixed effects, and with labour market controls. We focus on the specifications that use the lagged immigrant composition within states as instrument. Instead of IV regressions of linear probability models, we work here with endogenous probit specifications. Table 4 in the appendix summarizes the results. Two points are worth to be noted. First, our classical two-stage probit estimates are negative and thus support the linear probability results of the original paper. Second, the robust estimates change sign in most cases, indicating a positive (although not significant) effect of immigrant shares on the probability of women to major in a STEM field. These differences between the classical and robust estimates might cast doubt on the distributional assumptions of the model.

5.3 Modelling Women's Decision to Work

In our final application we work with the data set employed in Stata which exemplifies the estimation of probit models with continuous endogenous regressors (command **ivprobit**). This

---

[1]See Orrenius and Zavodny (2015) for the motivation of these instruments and a detailed description on how they are constructed.

is a hypothetical data set on 500 two-parent households with the aim of fitting a labor-supply model, i.e., to model whether the women is employed. The dependent variable is equal to one if the woman is employed and zero otherwise. The decision to work is assumed to be a function of the number of children at home (*kids*), the number of years of schooling that the women completed (*fem_educ*), and other household income (*other_inc*). The variable *other_inc* is treated as endogenous, as unobserved shocks affecting the household's other income might also affect the woman's decision to work. Thus, the number of years of schooling completed by the man (*male_educ*) is considered as an instrument.

For this empirical example, we perform a sensitivity analysis of the estimators. We start by fitting the model and estimating it with the two-step estimator of Rivers and Vuong (1988) and our proposed robust estimator on the clean data set. Then, we contaminate the data and study the behavior of the two estimators. For the robust versions, we employ both the MM-estimator of (Yohai 1987) and the M-estimator of Mallows type. For both these estimators, we try two different estimation options: one with weights using MCD, and one without weights on the covariate space. The results on the clean data set are reported in Table 5, where *step1.res* denotes the first step residuals. We notice that the estimates given by the classical and the robust estimators are close to each other, have the same sign, and the significance level is preserved. The standard errors are again similar in magnitude.

We now contaminate the data by changing the employment status of 10 employed women who have no kids and have the highest number of years of education (16 years in this data set). The estimates are reported in Table 6. We notice that the estimates for *fem_educ* and *kids* are underestimated if we use the classical two-step estimator. In addition, the significance levels for *kids* changes. Notice also that the coefficient on the first stage residuals changes significance level indicating endogeneity only at 10% level, while the significance level was 1% in the clean data. The robust estimators, however, perform better: the estimates are closer

to the values obtained in the cleaned data, while the coefficient on the first stage residuals is significant at 5% level.

6 Conclusion

The paper introduced robust estimators for the endogenous probit model which provide reliable estimates in the presence of small but harmful deviations from the model's assumptions. Monte Carlo simulations show a good performance of the robust estimator relative to the classical estimators in different contamination scenarios.

Our robust procedure can serve the applied researcher in several ways. First, as illustrated by our empirical example on the effect of education on BMI, if the classical and robust estimates are close, it is an indication that the distributional assumptions hold and the classical estimators can be trusted. In this case, we view our estimator as a useful addition to the applied researcher's toolkit for various robustness checks typically performed in applied work. Second, if the researcher already has doubts about the modelling assumptions in the preliminary data analysis (for example notices atypical but valid observations), our proposed approach can serve as a stand-alone method. Of course, if the distributional assumption do not hold even approximately, the applied researcher should resort instead to semi- and nonparametric methods.

References

Avella-Medina, M. (2020), "Privacy-Preserving Parametric Inference: A Case for Robust Statistics," *Journal of the American Statistical Association*, to Appear.

Blundell, R. W. and Powell, J. L. (2004), "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71, 655–679.

Bonhomme, S. and Weidner, M. (2018), "Minimizing Sensitivity to Model Misspecification," *Working Paper*.

Brunello, G., Fabbri, D., and Fort, M. (2013), "The causal effect of education on body mass: Evidence from Europe," *Journal of Labor Economics*, 31, 195–223.

Cameron, A. C. and Trivedi, P. K. (2005), *Microeconometrics: methods and applications*, Cambridge university press.

Cantoni, E. and Ronchetti, E. (2001), "Robust Inference for Generalized Linear Models," *Journal of the American Statistical Association*, 96, 1022–1030.

Duncan, G. M. (1987), "A Simplified Approach to M-Estimation with Application to Two-Stage Estimators," *Journal of Econometrics*, 34, 373–389.

Eicker, F. (1967), "Limit Theorems for Regression with Unequal and Dependent Errors," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L.M. LeCam, J. Neyman (Eds.), Berkeley: University of California Press, 59–82.

Freue, G. V. C., Ortiz-Molina, H., and Zamar, R. H. (2013), "A natural robustification of the ordinary instrumental variables estimator," *Biometrics*, 69, 641–650.

Hampel, F. (1974), "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383–393.

Hampel, F., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley and Sons.

Hampel, F., Rousseeuw, P. J., and Ronchetti, E. (1981), "The Change-of-Variance Curve and Optimal Redescending M-Estimators," *Journal of the American Statistical Association*, 76, 643–648.

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35, 73–101.

— (1967), "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L.M. LeCam, J. Neyman (Eds.), Berkeley: University of California Press, 221–233.

Maronna, R. A., Martin, G. R., and Yohai, V. J. (2006), *Robust Statistics: Theory and Methods*, Chichester: John Wiley and Sons.

Orrenius, P. M. and Zavodny, M. (2015), "Does immigration affect whether US natives major in science and engineering?" *Journal of Labor Economics*, 33, S79–S108.

Rivers, D. and Vuong, Q. H. (1988), "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics*, 39, 347–366.

Ronchetti, E. and Trojani, F. (2001), "Robust Inference with GMM Estimators," *Journal of Econometrics*, 101, 37–69.

Rousseeuw, P. J. and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212–223.

von Mises, R. (1947), "On the Asymptotic Distribution of Differentiable Statistical Functions," *The Annals of Mathematical Statistics*, 18, 309–348.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: The MIT Press, 2nd ed.

Yohai, V. J. (1987), "High Breakdown-Point and High Efficiency Robust Estimates for Regression," *The Annals of Statistics*, 15, 642–656.

Zhelonkin, M. (2013), "Robustness in Sample Selection Models," *PhD Thesis, University of Geneva, Switzerland.*

Zhelonkin, M., Genton, M. G., and Ronchetti, E. (2012), "On the Robustness of Two-Stage Estimators," *Statistics & Probability Letters*, 82, 726–732.

— (2016), "Robust Inference in Sample Selection Models," *Journal of the Royal Statistical Society B*, 78, 805–827.

A Derivation of Proposition 1

We derive the IF of the Rivers-Vuong two-step estimator. The first step is OLS for which the IF is known. Using our notation we have

$$\text{IF}\{(x_1, x_2, y_1); S, F\} = -\int \begin{pmatrix} x_1 x_1^\top & x_1 x_2^\top \\ x_2 x_1^\top & x_2 x_2^\top \end{pmatrix} dF \cdot (y_1 - x_1^\top \gamma_1 - x_2^\top \gamma_2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \tag{16}$$

Then, we use the result for the two-step M-estimators from Zhelonkin et al. (2012).

$$\begin{aligned}
\text{IF}(z_2; T, F) &= M_2^{-1} \Big(\Psi_2[(x_1, y_1, y_2); h\{(x_1, x_2, y_1), S(F)\}, T(F)] \\
&\quad + \int \frac{\partial}{\partial \zeta} \Psi_2\{z_2; \zeta, S(F)\} \frac{\partial h\{(x_1, x_2, y_1); \eta\}}{\partial \eta} dF \cdot \text{IF}(z_1; S, F) \Big),
\end{aligned} \tag{17}$$

where M_2 is the probit Hessian matrix given by

$$M_2 = \int \frac{\phi(\tilde{x}^\top \delta)^2}{\Phi(\tilde{x}^\top \delta)\{1 - \Phi(\tilde{x}^\top \delta)\}} \tilde{x} \tilde{x}^\top dF, \tag{18}$$

$$\frac{\partial h\{(x_1, x_2, y_1); \eta\}}{\partial \eta} = \frac{\partial \{y_1 - (x_1^\top, x_2^\top)\gamma\}}{\partial \gamma} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \tag{19}$$

where $\gamma = (\gamma_1^\top, \gamma_2^\top)$. Finally, the derivative $\partial \Psi_2 / \partial \zeta$ is defined as follows

$$\frac{\partial \Psi_2\{z_2; \zeta, S(F)\}}{\partial \zeta} = \frac{\partial}{\partial \zeta} \left([y_2 - \Phi\{(x_1^\top, y_1, \zeta)\delta\}] \frac{\phi\{(x_1^\top, y_1, \zeta)\delta\}}{\Phi\{(x_1^\top, y_1, \zeta)\delta\}[1 - \Phi\{(x_1^\top, y_1, \zeta)\delta\}]} \delta \right). \tag{20}$$

B IF of the joint MLE

The log-likelihood function for an observation i is given by

$$l_i = y_{2i} \log \Phi(u_i) + (1 - y_{2i}) \log\{1 - \Phi(u_i)\} + \log \phi \left(\frac{y_{1i} - x_{1i}^\top \gamma_1 - x_{2i}^\top \gamma_2}{\sigma_1} \right) - \log \sigma_1,$$

where

$$u_i = \frac{x_{1i}^\top \beta + \alpha y_{1i} + \rho(y_{1i} - x_{1i}^\top \gamma_1 - x_{2i}^\top \gamma_2)/\sigma_1}{(1 - \rho^2)^{1/2}}.$$

Since MLE is the particular case of an M-estimator, its IF is known to be proportional to the score function (Hampel et al., 1986). Hence the IF is (un)bounded if the score function is (un)bounded. The score functions for different parameters are as follows:

$$\frac{\partial l}{\partial \gamma_j} = -y_2 \frac{\phi(u)}{\Phi(u)} \frac{\rho x_j / \sigma_1}{(1 - \rho^2)^{1/2}} + (1 - y_2) \frac{\phi(u)}{\{1 - \Phi(u)\}} \frac{\rho x_j / \sigma_1}{(1 - \rho^2)^{1/2}} + \frac{1}{\sigma_1^2}(y_1 - x_{1i}^\top \gamma_1 - x_{2i}^\top \gamma_2) x_j, \tag{21}$$

for $j = 1, 2$.

$$\frac{\partial l}{\partial \beta} = \phi(u) \left[\frac{y_2 - \Phi(u)}{\Phi(u)\{1 - \Phi(u)\}} \right] \frac{x_1}{(1 - \rho^2)^{1/2}}, \quad (22)$$

$$\frac{\partial l}{\partial \alpha} = \phi(u) \left[\frac{y_2 - \Phi(u)}{\Phi(u)\{1 - \Phi(u)\}} \right] \frac{y_1}{(1 - \rho^2)^{1/2}}. \quad (23)$$

The score functions in (22) and (23) are structurally the same as (6). It is clear that all the score functions in (21)-(23) are unbounded.

C   Assumptions and Proof of Proposition 2

Denote $\Psi^r(z; \theta) = \left\{ \Psi_1^r(z_1; \gamma)^\top, \Psi_2^r(z; \gamma, \delta)^\top \right\}$. Assume that the following regularity conditions (adapted from Duncan 1987) hold:

1. z_i, for $i = 1, \ldots, N$, is a sequence of independent identically distributed random vectors with distribution F defined on a space $\mathcal{Z}$.

2. The parameter space Θ is a compact subset of $\mathbb{R} \times \mathbb{R}^{2p_1} \times \mathbb{R}^{p_2} \times [0, 1] \times \mathbb{R}^+$.

3. $\int \Psi^r(z; \theta) dF = 0$ has a unique solution θ_0 in the interior of Θ.

4. $\Psi^r(z; \theta)$ and $\frac{\partial}{\partial \theta} \Psi^r(z; \theta)$ are measurable for each θ in Θ, continuous for each z in $\mathcal{Z}$, and there exist F-integrable functions ξ_1 and ξ_2 such that for all $\theta \in \Theta$ and $z \in \mathcal{Z}$ $|\Psi^r(z; \theta)\Psi^r(z; \theta)^T| \leq \xi_1$ and $|\frac{\partial}{\partial \theta} \Psi^r(z; \theta)| \leq \xi_2$.

5. $\int \Psi^r(z; \theta)\Psi^r(z; \theta)^\top dF$ is non-singular for each $\theta \in \Theta$.

6. $\int \partial \Psi^r(z; \theta_0)/\partial \theta dF$ is finite and non-singular.

Proof of Proposition 2. Consistency and asymptotic normality follow from Theorems 1-4 in Duncan (1987). The asymptotic variance reduces to two terms. This is because the error terms ϵ_1 in the first step and $y_2 - \tilde{x}^\top \delta$ in the second step are independent by construction and thus $\Upsilon_1(z)\Upsilon_2(z)^\top$ and $\Upsilon_2(z)\Upsilon_1(z)^\top$ vanish after integration. $\qquad \square$

Table 1: Bias and Standard Deviation of the MLE, two-step and robust estimators for the parameters of the outcome equation (2), at the model, and under two types of contamination

$N = 1000$	Not contaminated		$y_2 = 1$		$y_2 = 0$	
	Bias	Std.Dev.	Bias	Std.Dev.	Bias	Std.Dev.
MLE						
γ_{10}	0.000	0.022	-0.046	0.024	0.045	0.025
γ_{11}	0.001	0.022	-0.093	0.029	-0.092	0.030
γ_2	0.000	0.022	-0.092	0.029	-0.091	0.029
β_0	-0.002	0.043	0.050	0.045	-0.054	0.044
β_1	-0.001	0.071	-0.056	0.109	-0.057	0.110
α	0.003	0.134	0.260	0.127	0.258	0.130
ρ	0.006	0.072	-0.454	0.114	-0.450	0.124
Two-step						
γ_{10}	0.000	0.022	-0.046	0.024	0.045	0.025
γ_{11}	0.001	0.022	-0.093	0.029	-0.092	0.030
γ_2	0.000	0.022	-0.092	0.029	-0.091	0.029
β_0	-0.003	0.050	0.051	0.045	-0.054	0.045
β_1	0.009	0.119	-0.203	0.114	0.114	0.116
α	0.006	0.112	0.114	0.124	0.567	0.126
ρ	0.009	0.112	-0.530	0.116	-0.526	0.126
Robust						
γ_{10}	0.000	0.023	-0.014	0.023	0.013	0.024
γ_{11}	0.001	0.023	-0.030	0.024	-0.030	0.025
γ_2	0.000	0.023	-0.029	0.024	0.028	0.024
β_0	-0.003	0.055	-0.009	0.055	-0.004	0.056
β_1	0.008	0.142	0.014	0.146	0.012	0.146
α	0.015	0.135	-0.019	0.141	-0.018	0.140
ρ	0.004	0.136	0.017	0.152	0.016	0.152

Table 2: Bias and Standard Deviation of the MLE, two-step and robust estimators, when the errors are contaminated by the $t-$distribution with 2 degrees of freedom

	γ_{10}	γ_{11}	γ_2	β_0	β_1	α	ρ
MLE							
Bias	0.000	-0.002	0.000	-0.001	-0.066	-0.065	-0.103
Std.Dev.	0.033	0.033	0.032	0.043	0.101	0.142	0.143
Two-step							
Bias	0.000	-0.002	0.000	-0.001	-0.090	-0.096	-0.093
Std.Dev.	0.033	0.033	0.032	0.049	0.135	0.120	0.132
Robust							
Bias	0.002	-0.002	0.000	0.001	0.003	-0.010	-0.008
Std.Dev.	0.023	0.024	0.024	0.055	0.149	0.133	0.136

Table 3: Classical two-step and robust estimates of the causal effect of education on BMI

Panel A:
Males

	Overweight		Obese	
	Two-Step	Robust	Two-Step	Robust
Years of	-0.009	-0.021	-0.011	-0.012
Schooling	(0.068)	(0.073)	(0.08)	(0.094)

Panel B:
Females

	Overweight		Obese	
	Two-Step	Robust	Two-Step	Robust
Years of	-0.107 **	-0.119 **	-0.077	-0.101
Schooling	(0.042)	(0.056)	(0.054)	(0.082)

Notes: Significance at 1% '***', significance at 5% '**', significance at 10% '*'.

Table 4: Classical two-step and robust estimates of the causal effect of immigrant shares on US natives' (females) decision to major in a STEM field

Panel A: College

		State and Cohort Fixed Effects		Labor market controls	
		Two-step	Robust	Two-step	Robust
	Immigrant Share	-0.026	-0.026	-0.011	0.005
	While in College	(0.059)	(0.069)	(0.053)	(0.074)

Panel B: High School

		State and Cohort Fixed Effects		Labor market controls	
		Two-step	Robust	Two-step	Robust
	Immigrant Share	-0.021	0.009	-0.016	0.024
	While in High School	(0.027)	(0.043)	(0.025)	(0.044)

Table 5: Classical two-step and robust estimates - Stata Example on Clean Data

	Two-Step	MM - weights	MM - no weights	M - weights	M - no weights
Intercept	0.396	0.607	0.492	0.625	0.497
	(0.477)	(0.518)	(0.485)	(0.52)	(0.487)
fem_educ	0.227 ***	0.214 ***	0.217 ***	0.213 ***	0.217 ***
	(0.027)	(0.03)	(0.028)	(0.03)	(0.028)
kids	-0.196 ***	-0.179 ***	-0.194 ***	-0.177 ***	-0.192 ***
	(0.048)	(0.053)	(0.05)	(0.053)	(0.05)
other_inc	-0.058 ***	-0.06 ***	-0.058 ***	-0.06 ***	-0.058 ***
	(0.009)	(0.01)	(0.009)	(0.01)	(0.009)
step1.res	0.024 ***	0.027 ***	0.024 **	0.027 ***	0.024 **
	(0.009)	(0.01)	(0.01)	(0.01)	(0.01)

Notes: Significance at 1% '***', significance at 5% '**', significance at 10% '*'.

Table 6: Classical two-step and robust estimates - Stata Example on Contaminated Data

	Two-Step	MM - weights	MM - no weights	M - weights	M - no weights
Intercept	0.339	0.653	0.476	0.638	0.48
	(0.463)	(0.512)	(0.473)	(0.513)	(0.474)
fem_educ	0.164 ***	0.19 ***	0.176 ***	0.19 ***	0.176 ***
	(0.025)	(0.029)	(0.026)	(0.029)	(0.026)
kids	-0.105 **	-0.14 ***	-0.132 ***	-0.14 ***	-0.13 ***
	(0.046)	(0.052)	(0.048)	(0.052)	(0.048)
other_inc	-0.047 ***	-0.057 ***	-0.051 ***	-0.057 ***	-0.051 ***
	(0.008)	(0.01)	(0.009)	(0.01)	(0.009)
step1.res	0.017 *	0.026 **	0.019 **	0.025 **	0.019 **
	(0.009)	(0.01)	(0.009)	(0.01)	(0.009)

Notes: Significance at 1% '***', significance at 5% '**', significance at 10% '*'.
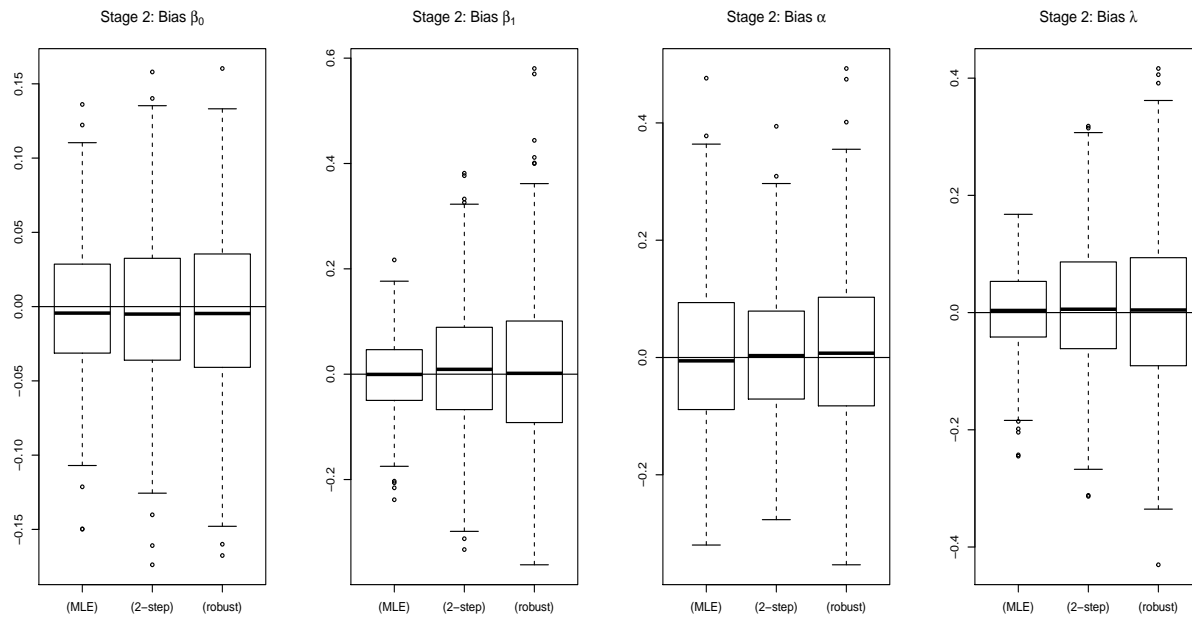

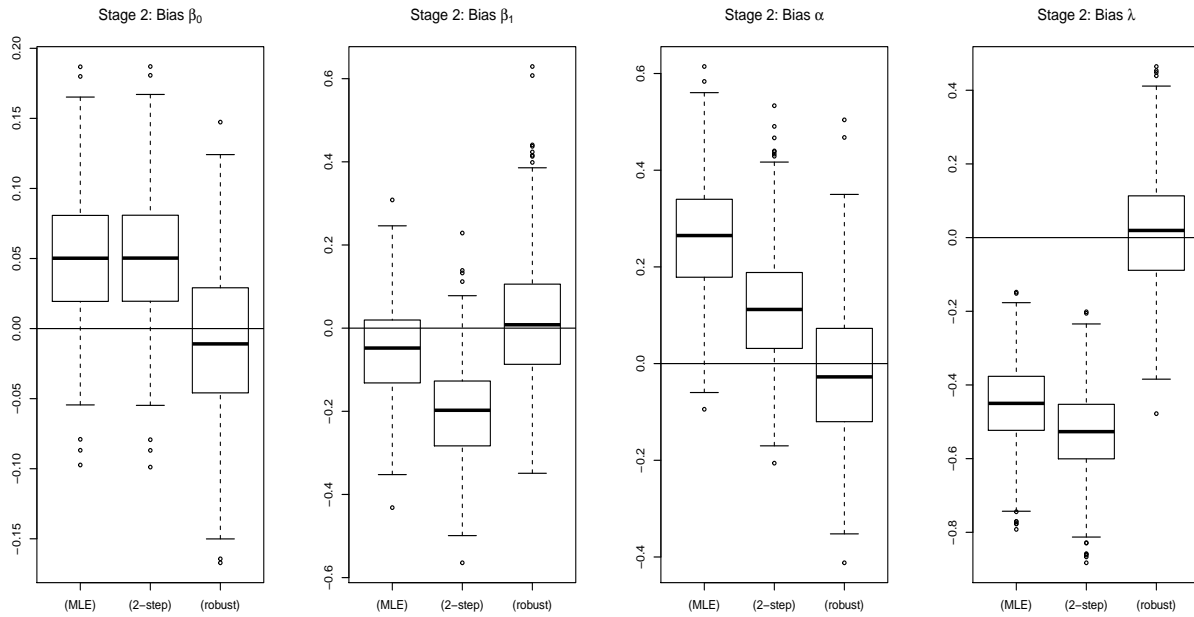Figure 1: Biases of parameter estimates of the equation of interest at the model

Figure 2: Biases of parameter estimates of the equation of interest under contamination of Type 1, i.e., the outlier is at $(x_2, x_2, y_1 y_2) = (2, 2, -2, 1)$
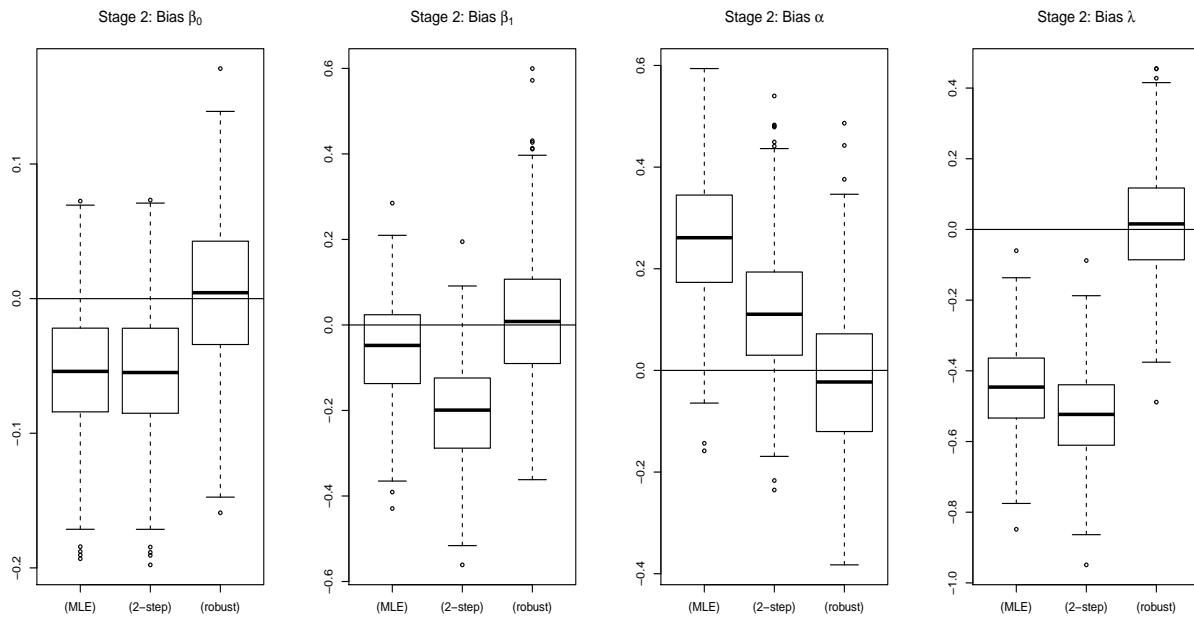

Figure 3: Biases of parameter estimates of the equation of interest under contamination of Type 1, i.e., the outlier is at $(x_2, x_2, y_1 y_2) = (-2, -2, 2, 0)$
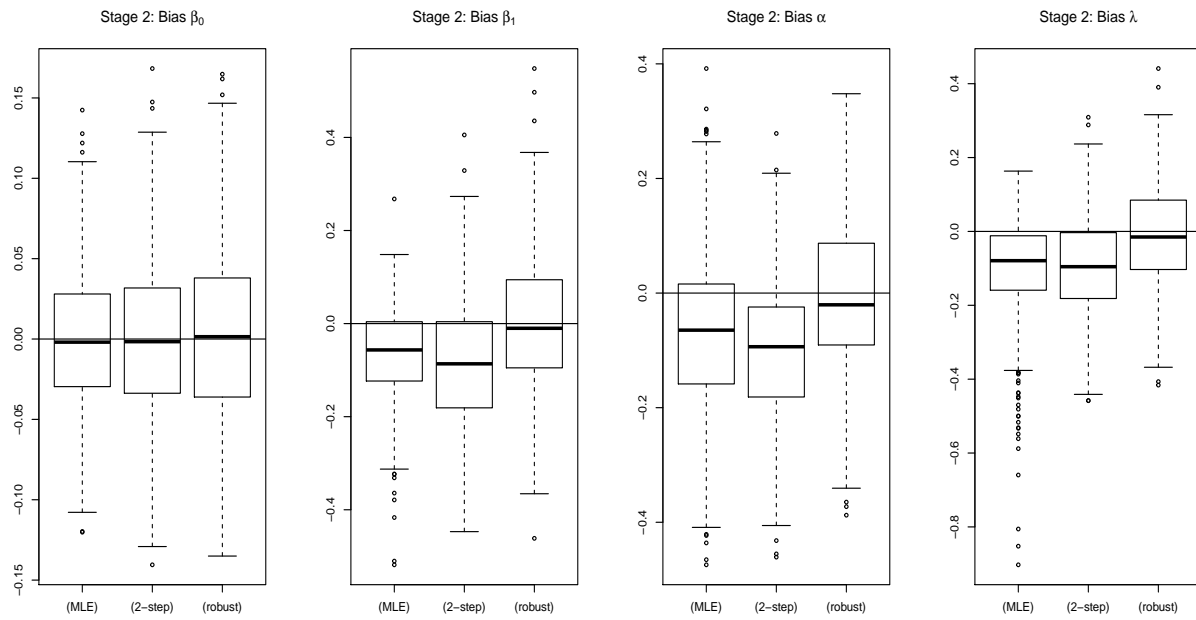
Figure 4: Biases of parameter estimates of the equation of interest when the errors are contaminated by the $t-$distribution with 2 degrees of freedom