GTL Regression: A Linear Model with Skewed and Thick-Tailed Disturbances

Wim Vijverberg*

Ph.D Program in Economics, City University of New York Graduate Center; and IZA, Bonn

Takuya Hasebe†

Faculty of Liberal Arts, Sophia University, Tokyo

19 June 2017

Abstract

If the disturbances of a linear regression model are skewed and/or thick-tailed, a maximum likelihood estimator is efficient relative to the customary Ordinary Least Squares (OLS) estimator. In this paper, we specify a highly flexible Generalized Tukey Lambda (GTL) distribution to model skewed and thick-tailed disturbances. The Maximum-Likelihood-based GTL-regression estimator is consistent and asymptotically normal. We demonstrate the potential gains of the GTL estimator over the OLS estimator in a Monte Carlo study and in two applications that are typical of applied economics research problems: a study of trade creation and trade diversion that result from preferential trade agreements, and an analysis of speeding tickets.

Key words: Linear Regression, Robust Estimation, Generalized Tukey Lambda Distribution

JEL classification: C16, C21

*Address: Wim Vijverberg, PhD Program in Economics, CUNY Graduate Center, 365 5th Avenue, New York, NY 10016-4309, tel: 212-817-8262, fax: 212-817-1514, email: wvijverberg@gc.cuny.edu

†Address: Takuya Hasebe, Faculty of Liberal Arts, Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo 102-8554, tel: 03-3238-4051, email: thasebe@sophia.ac.jp

1 Introduction

In estimating models of economic behavior, researchers pay more attention to the specification of the systematic component of the regression model than to the disturbances. After all, the researcher has full control over the manner in which the systematic component is built up from observable factors. The disturbance combines all unobservables in a single composite term that is both a nuisance because it prevents a perfect explanation of the outcome variable and a convenience because it permits a parsimonious representation of ignorance.

A researcher may resort to classical assumptions about the disturbances (independently and identically distributed with a zero mean and a constant finite variance), or he may choose to describe this aggregate disturbance factor with familiar tools such as heteroskedasticity, serial correlation, and ARCH and GARCH modeling. Such tools address patterns in the behavior among the disturbances. But consider a situation where such patterns do not exist: what options exist then? What is actually the nature of the distribution of the disturbances? The classical assumptions imply that the Ordinary Least Squares (OLS) estimator is efficient within the class of linear unbiased estimators. Moreover, if the disturbances are normally distributed, the OLS estimator coincides with the Maximum Likelihood (ML) estimator, which has optimality properties itself by virtue of the Cramer-Rao theorem. However, normality is not guaranteed by economic theory; in fact, economic theory rarely has anything to say about the distribution of the disturbances. Rather, the researcher is implicitly relying on the Central Limit Theorem, assuming that many unobservables play a role and none is dominant, thus yielding an approximately normally distributed aggregate disturbance. But that is an untested assumption.

This paper addresses the situation where disturbances are independently and identically distributed but are not "standard": they may be skewed; they may have long or short tails; and their moments may not even be defined. The Ordinary Least Squares estimator may not work well under these circumstances. We offer an ML estimator based on a highly flexible Generalized Tukey Lambda (GTL) distribution, that approximately nests the normal distribution but can also handle thick, skewed tails and nonexisting moments. We show that this flexible description of the distribution yields a more efficient estimator than OLS.

The literature offers several other approaches. First, tail data may be dealt with by means of data trimming or winsorizing (Chen and Dixon, 1972; Yale and Forsythe, 1976; Chen et al., 2001). These techniques view tail observations as data contaminations arising from a different data generating process; the interior of the scatterplot contains the information that is relevant for the process under focus. However, for thick-tailed processes, while the interior of the scatterplot is indeed relevant for the location parameters, the tails are not considered contaminations but rather provide information about scale and shape (and also location) parameters. Moreover, trimming and winsorizing lead to biased parameter estimates if the disturbances are generated with a skewed distribution.

Second, M-estimation maximizes an objective function that may or may not coincide with the log-likelihood function but, more importantly, may be selected so as to downplay extreme values in order to reduce their impact on the slope (or other) estimates (Huber, 1964; Huber and Ronchetti, 2009). The least squares estimator that follows the trimming or winsorizing of the data may be seen as a special case of the M-estimator. Ultimately, the objective of M-estimation is to estimate particular characteristics of the data generating process, such as location (intercept and slope), dispersion, or skewness; the objective is not to describe the distribution of the disturbances *per se*. As a result, the information content of tail observations is not (optimally) utilized. The same may be said about L-estimation, which formulates estimators from order statistics (or sample quantiles) and of which the Least Absolute Deviation estimator is a special case (Koenker and Bassett, 1978; Koenker and Portnoy, 1987; Koenker, 2005).

Third, by design, the semi-parametric estimation approach allows the distribution of the disturbances to be more arbitrary. For example, the various semiparametric estimators of the single index model only require the existence of several moments of the dependent variable; at the very least, the variance of the disturbance must be finite.[1] Heavy-tailed disturbances, especially those for which higher-order moments do not exist, are likely to cause estimation problems for location parameters, and estimates do not provide direct evidence about the nature of the distribution of the disturbances.

The fourth approach is more direct, explicitly specifying the distribution of the disturbances. Our paper fits within this approach. The current literature offers several alternatives. The first to come to mind is the Student's $t(\nu)$ distribution where variation in the degrees of freedom ν yields varying degrees of tail thickness (Zellner, 1976; Lange et al., 1989; Geweke, 1993; Breusch et al., 1997).[2] Other examples are the skewed-t, skewed generalized error, normal inverse gaussian, generalized hyperbolic, asymmetric Laplace, and asymmetric power distributions as well as mixtures of various distributions.[3] A series of recent studies rely on the family of so-called stable distributions, members of which are typically heavy-tailed and potentially skewed. Blattberg and Sargent (1971) and Samorodnitsky et al. (2007) develop a linear estimator for a model with a single explanatory variable in which both the disturbance and the explanatory variable are drawn from a stable distribution. Nolan and Ojeda-Revah (2013) develop an ML estimator of a multivariate linear (and nonlinear) model, and Hallin et al. (2011, 2013) use rank estimation as the strategy to deal with heavy-tailed stable disturbances. The family of stable distributions nests normality, but all other members do not have a second or higher moment and some even lack a first moment. All of these studies demonstrate the relative inefficiency of OLS estimators. Stable distributions are sometimes proposed as a way to model heavy-tailed disturbances in regression models especially in finance

[1]For example, see Powell et al. (1989), Ichimura (1993), Carroll et al. (1997), Horowitz (1998), and Li and Racine (2007).

[2]Fernancez and Steel (1999) show that the likelihood function of a multivariate $t(\nu)$-based regression model is unbounded for ν approaching 0, raising questions about models based on the multivariate t distribution.

[3]E.g., McDonald and Butler (1987); Barndorff-Nielsen (1997); Fernandez and Steel (1998); Azzalini and Capitanio (2003); Ferreira and Steel (2006); Komunjer (2007); Hansen et al. (2007); Wichitaksorn and Tsurumi (2013); Harvey and Sucarrat (2014); Wraith and Forbes (2015). Author (yeara, Section 5) compares many of these distributions with the GTL distribution, both analytically and through simulation.

applications, but the fact that the variance of the disturbance does not exist for all but the normal-distribution member of the family could be seen as a disadvantage.

A separate literature examines the tails of the OLS estimator under heavy-tailed disturbances; e.g., see He et al. (1990), Jureckova et al. (2001), Mikosch and de Vries (2013). The GTL distribution can have tails of a kind that conforms to the type of distributions analyzed in this literature. One of the conclusions is that the OLS estimator behaves poorly if the disturbance is non-Gaussian. Moreover, the OLS estimator is adversely affected if the explanatory variables are heavy-tailed, possibly even when the disturbances are normally distributed.

This study specifies the GTL distribution as the source of variation in the disturbances. This distribution is discussed in Section 2 and is entered into the linear regression model in Section 3, where we prove the consistency and asymptotic normality of the ML-based GTL estimator. We also provide an LM test for non-normality that can be applied to OLS residuals and points towards the GTL regression model if non-normality is discovered. Section 4 presents results of a Monte Carlo study that compares the performance of the OLS and GTL estimators when data are generated with GTL disturbances and possibly also with heavy-tailed GTL-distributed explanatory variables. Section 5 presents two examples where OLS and GTL estimates are contrasted: an analysis of the effect of preferential trade agreements on bilateral trade, and a study of the degree of police officer discretion in setting fines for speeding. In both cases, normality of the disturbances is resoundingly rejected. In a first example we find an efficiency gain of nearly 10% and essential differences in many of the main policy effects, and in the second example the OLS estimates are completely misleading and the GTL standard errors are less than one tenth of the OLS standard errors. Section 6 concludes.

2 The GTL Distribution

Pregibon (1980) and Freimer et al. (1988) developed a generalization of the Tukey lambda distribution that permits not only varying levels of tail thickness (as the Tukey lambda distribution does) but also skewness and therefore is highly flexible. We refer to this distribution as the Generalized Tukey Lambda (GTL) distribution; in the literature, it is also known as the GLD-FMKL distribution.[4] In its canonical form, the GTL(α, δ) distribution is described by its link function $G(u)$ with $u \in [0, 1]$:[5]

$$\epsilon = G(u) = \frac{u^{\alpha-\delta} - 1}{\alpha - \delta} - \frac{(1-u)^{\alpha+\delta} - 1}{\alpha + \delta}. \tag{1}$$

α and δ are real-valued parameters potentially anywhere between $-\infty$ and $+\infty$. For $\alpha - \delta \to 0$, the first term converges to $\ln u$; for $\alpha + \delta \to 0$, the second term converges to $\ln(1 - u)$. Broadly speaking, α is related to tail thickness, and δ determines skewness. Because α and δ often appear in pairs, we define $\lambda_1 = \alpha - \delta$ and $\lambda_2 = \alpha + \delta$

A link function is the inverse of a cumulative distribution function (cdf). For general values of λ_1 and λ_2, this inverse does not have an analytical closed-form solution but may be determined accurately with numerical algorithms; we use a combination of the bisection and Newton-Raphson algorithms (Press et al., 1986). The probability density function of ϵ is given by

$$f(\epsilon) = \frac{1}{u^{\lambda_1 - 1} + (1-u)^{\lambda_2 - 1}} \equiv \frac{1}{G'(u)} \qquad \text{with} \qquad u = G^{-1}(\epsilon). \tag{2}$$

The support of ϵ is not necessarily infinite: the lower bound equals $-\infty$ if $\lambda_1 \leq 0$ or $-1/\lambda_1$ if $\lambda_1 > 0$, and the upper bound equals ∞ if $\lambda_2 \leq 0$ or $1/\lambda_2$ if $\lambda_2 > 0$. The density $f(\epsilon)$ is regularly varying in the left (right) tail with index $-1/\lambda_1$ $(-1/\lambda_2)$ if $\lambda_1 < 0$ $(\lambda_2 < 0)$.

[4]Earlier, a generalization of the Tukey lambda distribution first appeared in the statistics literature in a paper by Ramberg and Schmeiser (1974) and has become known as the Generalized Lambda distribution (GLD), sometimes also referred to as GLD-RS. The parameter space of the GLD has gaps: as demonstrated in Karian et al. (1996), the GLD is not defined in the following regions of the shape parameters λ_3 and λ_4: (*i*) $\lambda_3 \leq 0$ and $0 \leq \lambda_4 \leq 1$; (*ii*) $-1 \leq \lambda_3 \leq 0$ and $\lambda_4 > 1$ and $(1-\lambda_3)^{1-\lambda_3}(\lambda_4 - 1)^{\lambda_4 - 1}(\lambda_4 - \lambda_3)^{\lambda_3 - \lambda_4}\lambda_4 \geq -\lambda_3$; and (*iii*) symmetric regions relative to (*i*) and (*ii*) obtained by interchanging λ_3 and λ_4. As a result, the feasible parameter space for (λ_3, λ_4) consists of four non-contiguous areas. The parameter space of the GTL distribution is free of such gaps and is therefore more amenable to maximum likelihood estimation.

[5]Referring to (1) as "the canonical form" does not imply that the distribution of ϵ is standardized with mean 0 and variance 1. A few other well-known distributions that are commonly stated in their canonical form are the Student's t, logistic, and χ^2 distributions.

For such values of λ_1 and λ_2, the GTL density has algebraic tails and thus is said to be heavy-tailed (He et al., 1990). Because of skewness, this heavy-tail property may apply in only one tail; with a suitable selection of α and δ, the other tail may even be truncated.

The k^{th} moment of ϵ exists only if $\min(\lambda_1, \lambda_2) > -1/k$.[6] In other words, the mean of ϵ does not exist if $\lambda_1 \leq -1$ or $\lambda_2 \leq -1$; if it does exist, $E[\epsilon] = -2\delta/((\lambda_1 - 1)(\lambda_2 - 1))$. Similarly, $\text{Var}(\epsilon)$ exists only if $\lambda_1 > -1/2$ and $\lambda_2 > -1/2$.[7] We shall denote $E[\epsilon]$ and $\text{Var}(\epsilon)$ of a canonical GTL(α, δ)-distributed ϵ as μ_ϵ and σ_ϵ^2.

With an appropriate selection of (α, δ), the GTL distribution closely approximates many well-known distributions (Freimer et al., 1988; Author, yearb).[8] For example, GLT$(0.1436, 0)$ is nearly indistinguishable from the normal distribution. Mathematically, the L_1-norm of the difference between the two densities equals 0.0031, such that the gap between the densities is only 0.31% of the total area underneath the densities. Statistically, the difference is negligible as well: simulations show that, in a large sample ($n = 5000$) of normally distributed random values, the power of an LM test (see Section 3.4) of GTL$(0.1436, 0)$ is only 8.1%. This LM test does have power against even minor formal deviations from GTL$(0.1436, 0)$: if disturbances are generated with slightly thicker tails $(\alpha, \delta) = (0.10, 0)$ (which closely approximates a Student's $t(30)$ distribution), power equals 86.9%; and similarly for a slightly thinner tail $(\alpha, \delta) = (0.20, 0)$ or slight skewness $(0.1436, 0.02)$, power is 99.8% and 77.7% respectively. We do note that under GTL$(0.1436, 0)$, ϵ ranges from -6.96 to 6.96 and not from $-\infty$ to ∞ as under normality.

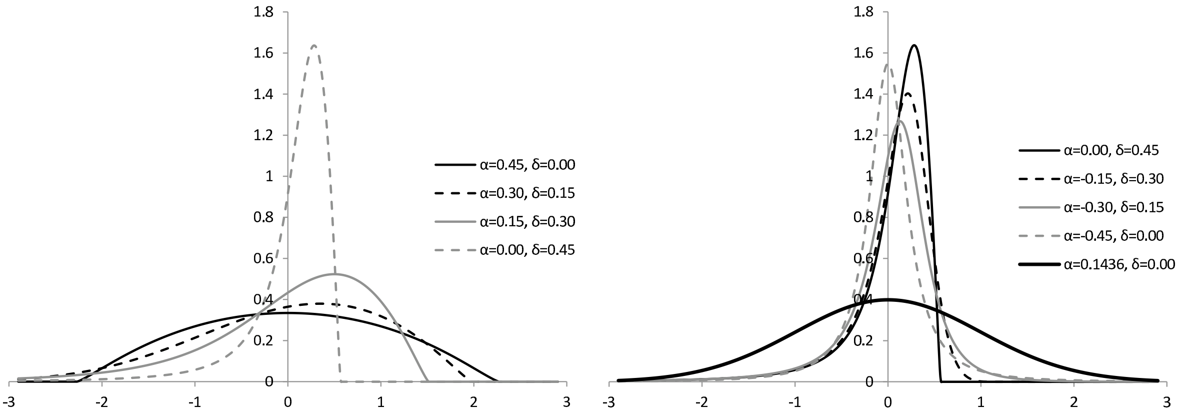
The GTL distribution can mimic other distributions as well. With GTL$(0, 0)$, the GTL family nests the logistic distribution, and the GTL distribution simplifies to a uniform distribution for four different combinations of (α, δ): $(1, 0)$, $(2, 0)$, $(\alpha, \alpha-1)$ with $\alpha \to \infty$, and $(\alpha, 1 - \alpha)$ with $\alpha \to \infty$. For $-0.8416 \leq \alpha \leq 0.1436$, GTL$(\alpha, 0)$ closely approxi-

[6] Equivalently, the k^{th} moment exists only if $-\alpha - \frac{1}{k} < \delta < \alpha + \frac{1}{k}$.

[7] $\text{Var}(\epsilon) = E[\epsilon^2] - E[\epsilon]^2$, where $E[\epsilon^2] = \frac{2}{(\lambda_1+1)(2\lambda_1+1)} - \frac{1}{\lambda_1\lambda_2}(B(\lambda_1 + 1, \lambda_2 + 1) - \frac{1}{\lambda_1+1} - \frac{1}{\lambda_2+1} + 1) + \frac{2}{(\lambda_2+1)(2\lambda_2+1)}$ and $B(\cdot, \cdot)$ is the Beta function.

[8] We determine such approximations by minimizing the L_1-norm of the difference between densities. Other methods for selecting (α, δ) include matching third and fourth moments, a least squares approach, matching quantiles, and more (Ramberg et al., 1979; Öztürk and Dale, 1985; Karian and Dudewicz, 1999; Su, 2011).

Figure 1: Standardized GTL densities for selected values of α and δ



(a) Thin-tailed left-skewed densities (b) Thick-tailed left-skewed densities

mates the $t(\nu)$ distribution with $1 \leq \nu \leq \infty$ degrees of freedom. GTL$(0.1422, -0.2290)$ approximates the Gumbel distribution. Similarly, many other distributions may be closely approximated by minimizing the difference between the density functions with the L_1 norm. The GTL distribution is quite flexible, indeed.[9]

Figure 1 shows seven examples of the GTL density. The distribution is symmetric when $\delta = 0$ or $\lambda_1 = \lambda_2$, right-skewed if $\delta < 0$ or $\lambda_1 > \lambda_2$, and left-skewed if $\delta > 0$ or $\lambda_1 < \lambda_2$.[10] Tails are longer and thicker if α is more negative.

3 The GTL Regression Model

3.1 Specification

The GTL regression model is specified as follows:

$$y_i = x_i'\beta + \sigma\epsilon_i \qquad \text{where} \qquad \epsilon_i \sim \text{iid GTL}(\alpha, \delta). \tag{3}$$

where $i = 1, \ldots, n$ denotes observations (individuals, states, time periods, etc.). x_i and β are $k \times 1$ vectors. ϵ has a canonical GTL(α, δ) distribution and is assumed independent

[9] See Appendix F for a more detailed discussion of the approximations.

[10] As Freimer et al. (1988) show, the direction of the skew actually reverses for large values of α but then the size of the skew is small.

of x, such that $E[\epsilon|x] = E[\epsilon]$ if indeed ϵ has a first moment. The existence of this first moment cannot be taken for granted if ϵ is generated by a GTL distribution with unknown parameters. In particular, if $\min(\lambda_1, \lambda_2) \leq -1$, $E[\epsilon]$ does not exist. For this reason also, we do not impose the restriction that $E[\epsilon] = 0$. σ is merely a scaling parameter. The variance of the disturbance $(\sigma\epsilon)$ equals $\sigma^2\sigma_\epsilon^2$ whenever σ_ϵ^2 exists. However, σ_ϵ^2 does not exist if $\min(\lambda_1, \lambda_2) \leq -1/2$.

If $E[\epsilon] \equiv \mu_\epsilon$ exists, it is a function of (α, δ) and thus not generally equal to 0. Thus, in that case, $E[y|x] = x'\beta + \sigma\mu_\epsilon$. If we denote the intercept of the model with β_1, it follows that the magnitude of β_1 is sensitive to the mean of ϵ. For the sake of comparability with other estimators such as OLS that assume $E[\epsilon|x] = 0$, we may compute an adjusted estimate of the intercept as $\hat{\beta}_1^* = \hat{\beta}_1 + \hat{\sigma}\hat{\mu}_\epsilon$. Its standard error is straightforwardly derived with the delta method. We note also that, assuming exogeneity of x, regardless of whether $E[\epsilon]$ exists, β_j represents the marginal effect of x_{ji} on y_i, holding all other observed and unobservable factors constant.

Given the linear regression equation (3), let $\theta = (\beta', \ \sigma, \ \alpha \ , \delta)'$, and define $u_i = G^{-1}(\frac{1}{\sigma}(y_i - x_i'\beta))$. The GTL estimator (GTLE) maximizes the following likelihood function:

$$L(y, x, \theta) = -n \ln \sigma - \sum_{i=1}^{n} \ln G'(u_i) = -n \ln \sigma - \sum_{i=1}^{n} \ln \left(u_i^{\lambda_1 - 1} + (1 - u_i)^{\lambda_2 - 1}\right). \quad (4)$$

3.2 OLS estimation, GTL estimation, M-estimation: a matter of weighting observations

Equation (3) is a simple linear regression model: why not estimate it with Ordinary Least Squares (OLS)? Provided that ϵ has first and second moments (and subject to other assumptions),[11] the OLS estimator $\hat{\beta}_{OLS}$ is efficient in the class of linear unbiased estimators, regardless of the distribution of ϵ. If ϵ is normally distributed, $\hat{\beta}_{OLS}$ is identical to the maximum likelihood estimator (MLE), and if ϵ has some other distribution, $\hat{\beta}_{OLS}$

[11] See White (1982, 1984) and Gourieroux et al. (1984) for a technical discussion.

is equivalent to the quasi-ML estimator that uses a gaussian criterion function (QMLg). Therefore, if ϵ is non-normal with a finite second moment, $\hat{\beta}_{OLS}$ is asymptotically normally distributed with a well-defined variance (Gourieroux et al., 1984, Theorem 3); if ϵ merely has a constant finite first moment but no finite second moment, the slope elements of $\hat{\beta}_{OLS}$ are consistent (Gourieroux et al., 1984, Theorem 1) and unbiased by the usual argument, but it remains unclear whether they are asymptotically normal. Indeed, our Monte Carlo analysis in Section 4 below provides an example where $\hat{\beta}_{OLS}$ is *not* asymptotically normal. If ϵ does not even have a finite first moment, $\hat{\beta}_{OLS}$ may be computed numerically but its value lacks meaning.

If ϵ has a GTL distribution, the OLS estimator generally differs from the GTLE. As proven below, the GTLE of β is consistent and asymptotically normal, even when ϵ does not possess first and second moments. And even when these do exist, OLS is no longer efficient relative to GTLE if the disturbance's distribution departs from normality.

The difference between the OLS and GTL estimators follows from the first order conditions that are implied by maximization of the respective criterion functions. Provided that σ_ϵ exists, the first-order condition of OLS (or QMLg) implies:

$$\sum_{i=1}^{n} \tilde{\epsilon}_i x_i = 0 \tag{5}$$

where $\tilde{\epsilon}_i = (y_i - x_i'\beta)/(\sigma\sigma_\epsilon)$ is the standardized disturbance. The first-order condition of GTLE yields:

$$\sum_{i=1}^{n} \frac{1}{\sigma} \frac{G''(u_i)}{(G'(u_i))^2}\, x_i \equiv \sum_{i=1}^{n} w_{GTL}(\tilde{\epsilon}_i)\tilde{\epsilon}_i x_i = 0 \tag{6}$$

where $u_i = G^{-1}(\sigma_\epsilon\tilde{\epsilon}_i + \mu_\epsilon)$. The term $w_{GTL}(\tilde{\epsilon}_i) = \frac{1}{\sigma}\frac{G''(u_i)}{(G'(u_i))^2\tilde{\epsilon}_i}$ is the weight given to $\tilde{\epsilon}_i x_i$ for observation i. This weight distinguishes OLS from GTLE: under OLS, the weight that is implied by equation (5) is $w_{OLS}(\tilde{\epsilon}_i) = 1$ for all $\tilde{\epsilon}_i$, whereas $w_{GTL}(\tilde{\epsilon}_i)$ varies with $\tilde{\epsilon}_i$. As shown in Web Appendix B, when tails are thick, $w_{GTL}(\tilde{\epsilon}_i)$ is greater than 1 for smaller values of $\tilde{\epsilon}_i$ and declines to 0 for values of $\tilde{\epsilon}_i$ in the tails. (Skewed distributions yield more complicated patterns in $w_{GTL}(\tilde{\epsilon}_i)$.) By varying the weight according to the distribution of $\tilde{\epsilon}$ that is detected in the data, the GTLE avoids the disruptive effect of tail values in the data and thus exploit the tail information more effectively than OLS.

M-estimation makes use of an objective function $L = \sum_{i=1}^{n} \rho(\tilde{\epsilon}_i)$, which is minimized to find $\hat{\beta}_M$. The first order conditions may be stated as:

$$\sum_{i=1}^{n} \frac{1}{\sigma} \frac{d\rho(\tilde{\epsilon}_i)}{d\tilde{\epsilon}_i} x_i \equiv \sum_{i=1}^{n} w_M(\tilde{\epsilon}_i)\tilde{\epsilon}_i x_i = 0 \tag{7}$$

with $w_M(\tilde{\epsilon}) = \frac{1}{\sigma\tilde{\epsilon}} \frac{d\rho(\tilde{\epsilon})}{d\tilde{\epsilon}}$. Thus, M-estimation differs from GTL regression in that M-estimation chooses $\rho(\tilde{\epsilon})$ a priori, and thus also $w_M(\tilde{\epsilon})$, without reference to the sample data at hand.

Non-normality may well be common in real-life data. For example, financial data are often reported to have thick tails and sometimes exhibit skewness as well; (e.g., Harvey and Siddique, 1999; Grigoletto and Lisi, 2009). Mikosch and de Vries (2013) show that statistical inference based on OLS can be misleading if the disturbance has thick tails. GTL regression can gain efficiency by properly accounting for tail thickness and by properly favouring the information in one tail relative to the other as a result of asymmetry in the distribution. We examine this further in a Monte Carlo study in Section 4.

3.3 Asymptotic Properties of the GTL estimator

Recall that $\theta = (\beta', \ \sigma, \ \alpha \ , \delta)'$ is the complete parameter vector of the GTL regression model. Let θ_0 denote the true parameter vector. Let Θ be the parameter space: Θ is a subset of $\mathbb{R}^{k+3}$ subject to restrictions indicated below.

Assumption A.1 *(i) $x \in \mathcal{X} \subset \mathbb{R}^k$. (ii) x is weakly exogenous. (iii) x has finite moments up to the fourth order. (iv) $E[xx']$ has full rank.*

Assumption A.2 *(i) The sequence $\{\epsilon_i\}$ is independently and identically GTL(α, δ)-distributed, with the GTL distribution defined by the link function in equation (1) and the probability density function in equation (2). (ii) ϵ is independent of x.*

Assumption A.3 Θ *is compact with* $\sigma > 0$, $\alpha + \delta \leq 1$, *and* $\alpha - \delta \leq 1$.

Assumption A.4 Θ *is compact with* $\sigma > 0$, $\alpha + \delta < \frac{1}{2}$, *and* $\alpha - \delta < \frac{1}{2}$.

Assumption A.5 $\theta_0 \in int(\Theta)$.

The following theorems establish the consistency and asymptotic normality of the GTL estimator $\hat{\theta}$ of the GTL regression model:

Theorem 1 *Given Assumptions A.1-A.3,* $\hat{\theta}$ *is a consistent estimator of* θ_0.

Theorem 2 *Given the linear regression model of equation (3) and Assumptions A.1, A.2, A.4 and A.5,* $\hat{\theta}$ *is asymptotically normally distributed* $N(\theta_0, V_0)$, *where* V_0 *is estimated as*

$$V(\hat{\theta}) = -\Big(\sum_{i=1}^{n} \nabla_{\theta\theta}\ell_i(\hat{\theta})\Big)^{-1} \tag{8}$$

where $\ell_i(\theta) = \ln g_y(y_i|x_i;\theta)$ *and* $g_y(y|x_i;\theta)$ *is the conditional density of* y.

The proofs of Theorems 1 and 2 are provided in Web Appendices C.1 and C.2, respectively.

Assumption A.3 restricts the range of (α, δ) values. For both $(\alpha, \delta) = (1, 0)$ and $(\alpha, \delta) = (2, 0)$, GTL becomes a uniform distribution. Thus, to satisfy identification requirements, both sets of values cannot be part of Θ at the same time. Assumption A.3 could be phrased less restrictively (e.g., requiring $\alpha < 2$ only); as stated, permissible densities have tails that go down to 0 except for when (α, δ) is on the boundary and one or both tails are high. But this is still not restrictive enough for Theorem 2, which requires $\alpha - \delta < \frac{1}{2}$ and $\alpha + \delta < \frac{1}{2}$, as stated in Assumption A.4, which permits the interchange of differentiation and integration (Lemma 2 in Web Appendix C.2).[12] Thus, for all (α, δ)

[12]In general, GTL densities may take on many forms and could even be U-shaped with high tails. Assumption A.4 implies a unimodal density with $f(\epsilon) \downarrow 0$ as ϵ approaches the lower or upper bound of its support and $f'(\epsilon) = 0$ for ϵ equal to a finite lower or upper bound. That is, the density is tangent to the horizontal axis at the finite lower and upper bound and approaches the axis at the infinite lower and upper bound. Practically speaking, this implies that draws from feasible GTL distributions near the endpoint are rare; there is no probability mass at the endpoint.

in the parameter space for which the GTL estimator is asymptotically normal, the GTL estimator is also $\sqrt{n}$-consistent (which is stronger than Theorem 1 states for a broader range of (α, δ) values). Moreover, because of Lemma 2, the information matrix equality holds and the GTL estimator is asymptotically efficient.

Note that Assumption A.2 is essential for these results. If ϵ has a non-GTL distribution, $\hat{\theta}$ is no longer consistent by Gourieroux et al. (1984, Theorem 2) because the assumed GTL-based likelihood function does not belong to the linear exponential family. We will return to this issue in Section 4 with examples of simulation of $\hat{\theta}$ in the presence of non-GTL disturbances.

3.4 Testing for normality

As mentioned, the normal distribution is closely similar to the $\mathrm{GTL}(0.1436, 0)$ distribution. Thus, on the basis of the GTL regression model, several tests for normality present themselves: a Wald test that uses $(\hat{\alpha}, \hat{\delta})$ and $V(\hat{\theta})$; a likelihood ratio (LR) test; a Lagrange multiplier (LM) test; and a Vuong test that examines the observations' contributions to the likelihood function under $\mathrm{GTL}(\hat{\alpha}, \hat{\delta})$ and under normality. Tests of other distributions (e.g., t) proceed in similar ways. Strictly speaking, the Null hypothesis of the Wald, LM and LR tests is the $\mathrm{GTL}(0.1436, 0)$ regression model.[13] However, a model with $\mathrm{GTL}(0.1436, 0)$ disturbances deviates trivially little from a model with true normal distributed disturbances (Web Appendix D): LM could be used as a post-estimation diagnostic test of the OLS model, pointing in the direction of GTL regression if non-normality is discovered.

4 Monte Carlo evidence: the effect of GTL disturbances

By the Gauss-Markov Theorem, skewed and unusually-tailed data do not bias the OLS estimator or render it inefficient relative to other linear unbiased estimators, provided

[13]The Vuong test compares $\mathrm{GTL}(\hat{\alpha}, \hat{\delta})$ with strict normality and thus involves non-nested models. A Jarque-Bera test of normality of OLS residuals examines only the third and fourth moments of the residuals and thus does not point specifically to GLT disturbances if normality is rejected.

that the disturbance has first and second moments. Yet, non-linear estimators may exist that are efficient relative to OLS. The GTL estimator is such an estimator. In this section, we report the results of a Monte Carlo analysis where data are generated with GTL disturbances and, in a few cases, with non-GTL disturbances.

The data generating process is as follows: $y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \sigma\epsilon_i$ for $i = 1, \ldots, n$ with $n = 250$ or 5000, where x_{2i} is a standard normal random variate, x_{3i} is a $\chi^2(5)$ random variate that is standardized to have mean 0 and variance 1 and is independent of x_{2i}, and ϵ_i is a canonical GTL(α, δ) variate for various combinations of α and δ. The values of β_1, β_2 and β_3 are all equal to 1. The skewness of x_3 inserts a bit of skewness into y. α varies from 0.33 (GTL with truncated tail) to -1.00 (GTL is so thick-tailed that the $E[\epsilon]$ does not even exist); δ is varied such that the data generating process includes right-skewed ($\delta < 0$), symmetric ($\delta = 0$), and left-skewed ($\delta < 0$) experiments. Because of the range of (α, δ) combinations we cannot standardize the GTL distribution. Instead, we select the scaling parameter σ such that the distance between the 0.1% quantile and the 99.9% quantile of the GTL distribution is the same as that of a normal $N(0, 2)$ distribution. This keeps the R^2 of the OLS regression in the neighborhood of 0.50 as long as the standard deviation of ϵ is defined, but the R^2 diminishes as the tails of the GTL distribution become longer. We replicate each design 1000 times, drawing new explanatory variables each time and using the same sets of draws across the different experiments.[14]

Table 1 presents the root mean squared error (RMSE) of the OLS and GTL estimators of the slopes and intercept for different values of (α, δ).[15] In Panel A, data are generated with GTL(0.1436,0) and normal $N(0, 2)$ disturbances. Since a GTL(0.1436,0) distribution closely approximates a normal distribution, it is not surprising that the OLS and GTL

[14]One might question the wisdom of a Monte Carlo design that specifies thick-tailed unobservables but regular-tailed observable determinants. In fact, regression designs that contain thick-tailed explanatory variables may be problematic (e.g. Huber, 1981; He et al., 1990; Jureckova et al., 2001), since the OLS estimator may lose its consistency property. Web Appendix E.4 provides Monte Carlo results for a design with thick-tailed observables and thin- or thick-tailed unobservables.

[15]The OLS estimate of the intercept is not adjusted for the fact that it is a priori biased when δ is nonzero.

Table 1: RMSE of OLS and GTL estimators of slopes and intercept for GTL-generated data

DGP			RMSE of OLS			RMSE of GTLE		
α	δ	σ	β_2	β_3	β_1	β_2	β_3	β_1
A: Comparing GTL(0.1436,0) with $N(0,2)$; $N=250$								
0.1436	0.00	1.188	0.1072	0.1097	0.1101	0.1085	0.1115	0.1256
$N(0,2)$	n.a.	n.a.	0.1072	0.1097	0.1101	0.1084	0.1114	0.1253
B: Various GTL distributions, small sample, $N=250$								
0.33	-0.10	1.477	0.106	0.108	0.199	0.089	0.094	0.132
0.33	0.00	1.508	0.107	0.109	0.109	0.097	0.101	0.133
0.33	0.10	1.477	0.106	0.108	0.200	0.091	0.093	0.132
-0.33	-0.10	0.454	0.133	0.144	0.255	0.066	0.068	0.066
-0.33	0.00	0.482	0.118	0.122	0.124	0.071	0.073	0.071
-0.33	0.10	0.454	0.127	0.126	0.239	0.067	0.069	0.067
-0.67	-0.25	0.136	1.737	1.667	1.932	0.025	0.026	0.025
-0.67	0.00	0.202	0.336	0.337	0.341	0.038	0.040	0.037
-0.67	0.25	0.136	1.046	0.823	1.126	0.025	0.026	0.025
-1.00	-0.50	0.024	112.585	89.227	103.243	0.005	0.005	0.005
-1.00	0.00	0.079	2.503	2.217	2.408	0.018	0.019	0.018
-1.00	0.50	0.024	39.742	27.930	35.099	0.005	0.005	0.005
C: Various GTL distributions, large sample, $N=5000$								
0.33	-0.10	1.477	0.023	0.024	0.169	0.018	0.019	0.032
0.33	0.00	1.508	0.023	0.024	0.025	0.021	0.021	0.030
0.33	0.10	1.477	0.023	0.024	0.171	0.019	0.019	0.033
-0.33	-0.10	0.454	0.034	0.032	0.209	0.015	0.014	0.015
-0.33	0.00	0.482	0.027	0.028	0.029	0.016	0.015	0.016
-0.33	0.10	0.454	0.031	0.032	0.210	0.014	0.014	0.015
-0.67	-0.25	0.136	6.075	3.868	4.933	0.005	0.005	0.006
-0.67	0.00	0.202	0.233	0.180	0.220	0.008	0.008	0.008
-0.67	0.25	0.136	1.903	1.456	2.512	0.005	0.005	0.006
-1.00	-0.50	0.024	9916.534	5631.352	7274.680	0.001	0.001	0.001
-1.00	0.00	0.079	12.606	7.908	10.359	0.004	0.004	0.004
-1.00	0.50	0.024	1997.902	832.310	2381.450	0.001	0.001	0.001

Note: Bias and standard errors are reported separately in Table E.1 of the Web Appendix.

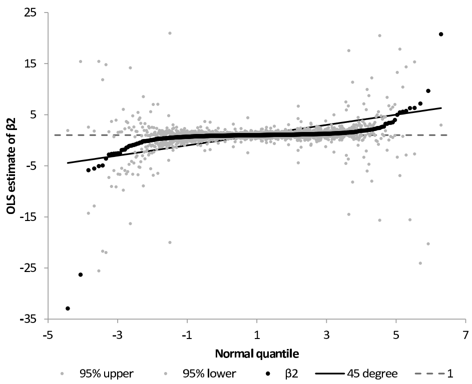
estimators perform nearly identically. In Panel B with $n = 250$ (a quite small sample) and Panel C with $n = 5000$ (a large sample), the GTL distribution varies from having truncated tails (for $\alpha = 0.33$) to having very long tails (for $\alpha = -1$), with right skew if δ is negative and left skew if δ is positive. The direction of the skewness matters somewhat for the RMSEs because of the skewness inherent in one of the explanatory variables (x_3), but the overriding concern is the length of the tails. The more negative α is, the worse OLS performs: sample moments blow up because theoretical moments of the GTL distribution (and thus of the OLS estimator) no longer exist. The GTL estimator is hardly affected by the thickness or skewness of the tails; in fact, the RMSEs diminish as the GTL distribution becomes more heavy-tailed. For a larger negative α, the GTL estimator outperforms the OLS estimator by a vast margin.

Detailed inspection of simulation results shows that the RMSEs of the slope estimators represent mostly the standard deviation of the simulated slope estimates. Given the exogeneity of the explanatory variables, the OLS estimator is unbiased as long as the mean of ϵ exists, but even when (α, δ) is such that the first moment of ϵ no longer exists and the OLS estimator loses its theoretical moments also, the average of the simulated values is still close to 1. The GTL estimator of the intercept shows a bias as expected (Section 3.1), but it is minor compared to the standard deviation of the simulated values.

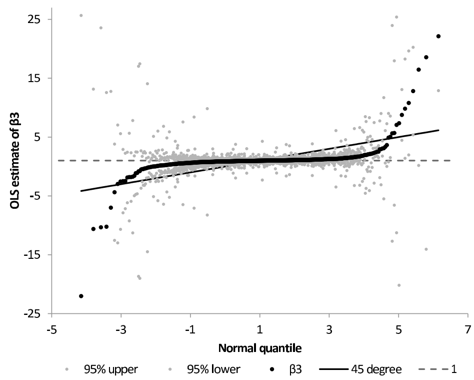
Separate from concerns about bias and precision is the question whether the small-sample distribution of the estimators is approximately normal. One way to examine the small-sample properties is to compute coverage ratios, i.e., the rate at which the confidence intervals contain the population parameter values. Table 2 illustrates our findings with three experiments; the detailed results for all experiments are provided in Web Appendix E.3. Surprisingly, in Panel A of Table 2, the coverage rates of the OLS slope estimators are essentially 95% in all experiments; the intercept's estimator is biased when the GTL distribution of ϵ is skewed as the GTL distribution is utilized in its canonical form. This would suggest that perhaps the sampling distribution of the OLS estimators is nearly normal after all. However, the Jarque-Bera test in Panel B squashes that conclusion: in the presence of GTL disturbances, the sampling distribution of the OLS estimator is not normal even for $n = 5000$.

How can the findings of Panels A and B be reconciled? Figure 2 presents QQ plots of both OLS and GTLE slope estimators, together with the 95% confidence intervals for each replication and a dashed reference line at 1. In this figure, (α, δ) equals $(-0.67, 0.25)$, such that the variance of the GTL-distributed disturbance is not even defined. Clearly, the QQ plots of the OLS estimators indicate strong deviations from normality: the systematic

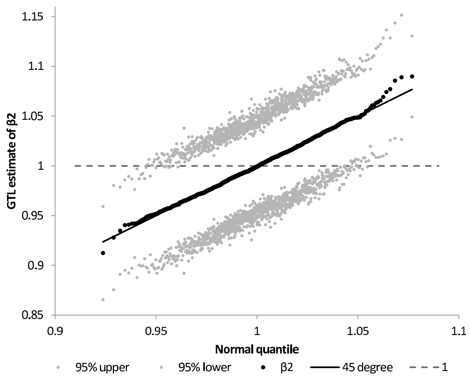
Figure 2: QQ plots of OLS and GTLE estimators of β_2 and β_3 for $(\alpha, \delta) = (-0.67, -0.25)$ and $n = 250$



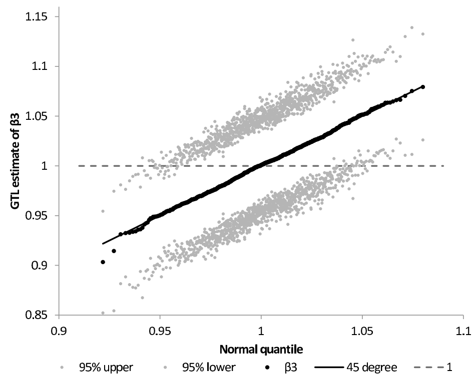
(a) OLS: $\hat{\beta}_2$



(b) OLS $\hat{\beta}_3$



(c) GTLE $\hat{\beta}_2$



(d) GTLE $\hat{\beta}_3$

Table 2: Diagnostics of the distribution of OLS and GTL estimators: baseline design

DGP			OLS			GTLE					
α	δ	σ	β_2	β_3	β_1	β_2	β_3	β_1	σ	α	δ
A: Coverage rate at a 95% confidence level											
Small sample, $N = 250$											
0.1436	0.00	1.188	0.947	0.948	0.954	0.944	0.937	0.943	0.941	0.918	0.924
-0.33	-0.10	0.454	0.958	0.948	0.655	0.952	0.936	0.951	0.943	0.946	0.950
-0.67	-0.25	0.136	0.955	0.959	0.337	0.955	0.933	0.955	0.951	0.948	0.943
Large sample, $N = 5000$											
0.1436	0.00	1.188	0.960	0.948	0.939	0.960	0.946	0.950	0.949	0.932	0.946
-0.33	-0.10	0.454	0.956	0.944	0.000	0.960	0.956	0.951	0.941	0.948	0.944
-0.67	-0.25	0.136	0.952	0.961	0.093	0.953	0.956	0.949	0.938	0.949	0.944
B: p-values of Jarque Bera tests for normality of the sampling distribution of the estimator											
Small sample, $N = 250$											
0.1436	0.00	1.188	0.91	0.10	0.62	0.46	0.22	0.55	0.00	0.00	0.01
-0.33	-0.10	0.454	0.00	0.00	0.00	0.23	0.21	0.86	0.00	0.10	0.73
-0.67	-0.25	0.136	0.00	0.00	0.00	0.07	0.70	0.80	0.00	0.06	0.55
Large sample, $N = 5000$											
0.1436	0.00	1.188	0.14	0.88	0.14	0.21	0.94	0.96	0.15	0.78	0.01
-0.33	-0.10	0.454	0.00	0.00	0.00	0.51	0.50	0.99	0.30	0.44	0.06
-0.67	-0.25	0.136	0.00	0.00	0.00	1.00	0.39	0.98	0.71	0.42	0.11
C: Ratio of average estimated variance to Monte Carlo variance											
Small sample, $N = 250$											
0.1436	0.00	1.188	1.045	1.000	0.979	1.010	0.961	0.988	0.905	0.878	0.856
-0.33	-0.10	0.454	1.073	0.926	0.915	1.018	0.956	1.031	0.970	0.992	0.929
-0.67	-0.25	0.136	0.875	1.010	0.866	1.005	0.945	1.051	1.007	1.019	0.933
Large sample, $N = 5000$											
0.1436	0.00	1.188	1.051	1.005	0.935	1.048	1.004	0.978	0.951	0.935	0.990
-0.33	-0.10	0.454	1.050	1.046	0.901	1.027	1.036	0.992	0.912	0.908	0.951
-0.67	-0.25	0.136	1.839	1.897	0.998	1.006	1.058	0.996	0.918	0.932	0.940

"sling" around the 45 ° line indicates a sampling distribution with fat tails.[16] Now, for the given (α, δ), $\text{Var}(\hat{\beta}_{OLS})$ is not defined, and therefore the 95% confidence intervals are not trustworthy. The coverage rate reflects how many of the supposed "confidence intervals" bracket the dashed line. Thus, the fact that the coverage rate happens to be 95% in all cases is purely coincidental and not at all evidence that the usual statistical inference methods apply.

[16]For $(\alpha, \delta) = (0.1436, 0)$, the QQ plot of the OLS estimator is very similar to that shown in Panels 2c and 2d; for $(\alpha, \delta) = (-0.33, 0)$, some of this same sling becomes apparent. Thus, the more ϵ deviates from normality, the more the small sampling distribution of the OLS estimators deviates from normality as well.

In contrast, the GTLE results indicate a good coverage rate (Table 2), and Panels 2c and 2d in Figure 2 shows highly satisfactory QQ plots for $(\alpha, \delta) = (-0.67, -0.25)$, the experiment that gives OLS so much trouble. Other experiments yield similar results.

Panel C of Table 2 contributes one more facet about the small-sample behavior of the estimators. The variance of the estimator, OLS or GTLE, is itself estimated as well. Each replication yields another estimate of the variance, as well as an estimate of θ: the average of the estimated variances ought to be close to the variance of the replicated $\hat{\theta}$. As the ratios of these two quantities in Panel C of Table 2 shows, the estimated OLS variance becomes problematic as the GTL distribution differs more greatly from normality, and the ratio of the GTLE variances stays close to 1.

In the simulations reported so far, the disturbances were generated with the GTL distribution. This gives GTL regression an advantage over OLS. To show that this advantage also exists when disturbances follow other distributions, we give five examples: normal, $t(5)$, $\chi^2(8)$, Gumbel, and Skewed-$t(5, 0.67)$. These distributions differ substantially from each other, in shape, in skewness and in kurtosis or heavy-tailedness. They are illustrated in detail in Appendix F, which also shows how closely the GTL distribution approximates each of these when (α, δ) is estimated by the (quasi-)maximum likelihood approach. Table 3 focuses on the two slope parameters of the simulation model. Bias is not an issue, neither with OLS (as expected) nor with GTL regression. The standard deviations are computed from the 1000 replication in each simulation. When the true distribution is normal, GTL regression lags slightly behind OLS regression, which ought to be efficient since it is equivalent to maximum likelihood estimation with a correctly specified disturbance distribution. In each other case, GTL regression outperforms OLS estimation, even though the distribution is technically misspecified—but still is closely approximated because of the flexibility of the GTL distribution. Thus, the advantage of GTL regression is not restricted merely to cases where disturbances are GTL-distributed.

Table 3: Bias, efficiency and RMSE of OLS and GTL estimators of slopes for non-GTL-generated data

Disturbance	n	Estimator	bias($\hat{\beta}_1$)	bias($\hat{\beta}_2$)	s.d.($\hat{\beta}_1$)	s.d.($\hat{\beta}_2$)	RMSE($\hat{\beta}_1$)	RMSE($\hat{\beta}_2$)
Normal	250	OLS	0.0042	0.0006	0.1072	0.1097	0.1072	0.1097
		GTL	0.0037	0.0012	0.1084	0.1114	0.1084	0.1114
	5000	OLS	0.0009	-0.0002	0.0237	0.0242	0.0237	0.0242
		GTL	0.0010	-0.0002	0.0237	0.0242	0.0237	0.0242
$t(5)$	250	OLS	0.0029	0.0004	0.0818	0.0842	0.0819	0.0842
		GTL	0.0023	0.0015	0.0750	0.0770	0.0751	0.0770
	5000	OLS	0.0007	0.0000	0.0183	0.0187	0.0183	0.0187
		GTL	0.0005	0.0001	0.0166	0.0170	0.0166	0.0170
$\chi^2(8)$	250	OLS	0.0036	-0.0007	0.0878	0.0896	0.0879	0.0896
		GTL	0.0020	0.0048	0.0641	0.0683	0.0642	0.0685
	5000	OLS	0.0008	-0.0002	0.0193	0.0198	0.0193	0.0198
		GTL	0.0006	0.0008	0.0150	0.0151	0.0150	0.0152
Gumbel	250	OLS	0.0038	-0.0008	0.0951	0.0971	0.0951	0.0971
		GTL	0.0022	0.0048	0.0742	0.0789	0.0742	0.0790
	5000	OLS	0.0008	-0.0002	0.0209	0.0215	0.0209	0.0215
		GTL	0.0007	0.0002	0.0169	0.0171	0.0169	0.0171
Sk-$t(5, 0.67)$	250	OLS	0.0031	-0.0011	0.0810	0.0832	0.0810	0.0832
		GTL	0.0009	0.0013	0.0439	0.0455	0.0439	0.0455
	5000	OLS	0.0007	-0.0001	0.0180	0.0185	0.0180	0.0185
		GTL	0.0003	0.0000	0.0095	0.0096	0.0096	0.0096

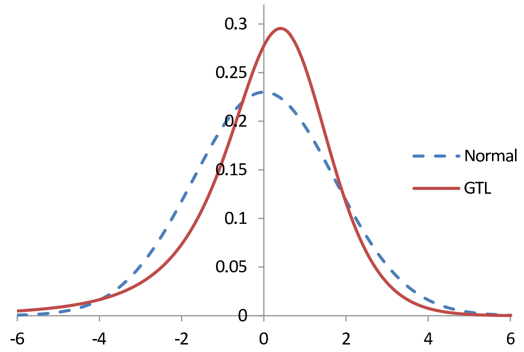
5 Applications

We apply the GTL regression model to two examples: the question whether preferential trade agreements are trade-creating or trade-diverting, and a study of the determinants of fines for speeding. These examples illustrate the difference that can result when more attention is paid to the characteristics of the disturbances of the model. In both cases, we find some skewness and substantial kurtosis, and we highlight the changes in the slopes estimates and the gain in the precision of these estimates. Web Appendix A provides the data sources, variable definitions, and descriptive statistics.

5.1 Trade creation and diversion

Many studies have examined the effect of preferential trade agreements (PTAs) on trade. Two trading partners signing on to the same PTA may trade more with each other than before and less with other partners who are not part of the PTA because of changes

Figure 3: Estimated distributions of ϵ for the trade creation and diversion equation



in the relative cost of trading. Thus, following work by Ghosh and Yamarik (2004) and Subramanian and Wei (2007), Eicher et al. (2012) estimate a regression model that relates bilateral imports from country j to country i to (a) trade creation dummy variables that indicate whether countries i and j are both members of various PTAs, and (b) trade diversion dummy variables that indicate whether only one of the (i, j) country pair belongs to a given PTA, augmented with (c) various control variables such as the national income of each country, the distance between them, the colonial history between them, the similarity in language, differences in human capital and income level, indices of remoteness, trade policy, exchange rate volatility, and so forth. See Table A.1 in the Web Appendix for more details.[17] The data describe trade between 97 countries arranged in 4069 importer-exporter country pairs over five-year intervals from 1960 to 2000 with a total of 37983 observations in an unbalanced panel.

The estimation results in Table 4 correspond with specification 2 in Table III of Eicher et al. (2012). Table 4 highlights the estimates of the trade creation and trade diversion effects, whereas estimated slopes of control variables may be found in Table G.1 in the Web Appendix. The normality assumption that is consistent with the OLS estimation method is rejected convincingly by every indicator; the distribution of the disturbances is left-skewed (as $\hat{\delta} > 0$, with $\kappa_3 = -2.34$) and quite thick-tailed (as $\hat{\alpha} < 0$, with $\kappa_4 = 40.17$). Figure 3 illustrates the density functions.

[17]More precise definitions of the PTAs may be found in Ghosh and Yamarik (2004), Subramanian and Wei (2007), and Eicher et al. (2012).

Table 4: Trade creation and diversion, 1960-2005

	OLS		GTLE			
	Estimate	Stan.Err.	Estimate	Stan.Err.	$\frac{\hat{\beta}_{OLS}-\hat{\beta}_{GTL}}{\hat{\beta}_{GTL}}$	$\frac{SE_{GTL}}{SE_{OLS}}$
Trade creation dummy variables						
tc.nafta	0.371	0.252	0.500	0.220	-0.257	0.874
tc.eu	0.427	0.115	0.196	0.098	1.182	0.853
tc.efta	0.685	0.129	0.518	0.116	0.323	0.896
tc.eea	0.179	0.095	0.261	0.079	-0.312	0.837
tc.caricom	2.823	0.513	2.417	0.459	0.168	0.894
tc.ap	0.828	0.186	0.804	0.159	0.030	0.852
tc.mercosur	1.086	0.306	1.035	0.315	0.049	1.030
tc.asean	0.467	0.216	0.492	0.185	-0.051	0.858
tc.anzcerta	0.969	0.141	0.748	0.130	0.295	0.920
tc.apec	1.599	0.095	1.291	0.085	0.238	0.889
tc.laia	-0.133	0.141	-0.432	0.134	-0.691	0.950
tc.cacm	2.314	0.150	1.931	0.139	0.198	0.927
tc.bilateralPTA	0.110	0.128	0.098	0.117	0.122	0.916
Trade diversion dummy variables						
td.nafta	0.151	0.073	0.081	0.061	0.875	0.841
td.eu	0.651	0.051	0.434	0.047	0.500	0.908
td.efta	0.376	0.059	0.202	0.054	0.866	0.921
td.eea	-0.142	0.048	-0.101	0.043	0.409	0.894
td.caricom	-0.577	0.100	-0.539	0.097	0.071	0.972
td.ap	0.105	0.074	0.115	0.068	-0.088	0.925
td.mercosur	0.030	0.073	-0.019	0.063	-2.554	0.869
td.asean	0.474	0.070	0.395	0.061	0.198	0.869
td.anzcerta	-0.759	0.098	-0.657	0.086	0.156	0.879
td.apec	0.439	0.049	0.341	0.042	0.288	0.871
td.laia	-0.561	0.060	-0.533	0.054	0.051	0.913
td.cacm	-0.174	0.078	-0.120	0.074	0.451	0.945
td.bilateralPTA	-0.292	0.054	-0.275	0.045	0.064	0.832
σ			0.809	0.013		
α			-0.078	0.008		
δ			0.139	0.005		
log Likelihood	-74773.2		-72585.9			
(Absolute) Average					0.294	0.911

Dependent variable: Log of bilateral imports. The model also includes control variables (reported in Table G.1 in the Web Appendix) and time dummy variables (not reported). Number of observations = 37983. Skewness and kurtosis of OLS residuals equal -0.71 and 4.90; the Jarque-Bera test of normality of the OLS residuals has a p-value of less than 0.001. The Wald test of the GTL estimates of (α, δ) equals 1326.8, rejecting normality with a p-value of less than 0.001. The LM test equals 1299 with a p-value of less than 0.001 and with 24 support violations (all in the left tail). The Vuong test that compares OLS and GTLE equals -24.96 in favor of the GTL model with a p-value of less than 0.001.

More importantly, the OLS and GTLE slopes differ by an average of nearly 30 percent, and the standard errors of the GTLE slopes are 8.9 percent smaller on average. The degree of trade creation varies substantially by estimation method. For example, the European Union (EU) effect drops from 42.7 percent to 19.6 percent; NAFTA rises from

37.1 percent to 50 percent; the effect of the ANZCERTA agreement between Australia and New Zealand is nearly one third smaller than the OLS suggests (though still large); and the *negative* effect of LAIA (the Latin America Free Trade Association/Latin America Integration Agreement) is tripled and now also statistically significant. As for the trade diversion effect of PTAs, the GTL estimates that correspond with the five positive and statistically significant OLS estimates are all substantially smaller, whereas all of the negative OLS slopes estimates are matched by similar GTL estimates. Thus, the evidence in favor of a PTA trade diversion effect is actually stronger than the OLS estimates suggest, although the GTL estimates still present a mixed picture.[18]

5.2 Speeding tickets, Massachusetts

Makowsky and Stratmann (2009) examine the determinants of traffic citations and fines for speeding, using a database that contains all speeding-related stops in Massachusetts from April 1, 2001 through May 31, 2001.

A traffic stop results in either a ticket or a warning. When a ticket is issued, a driver has to pay a fine. Whether a police officer issues a ticket or gives a warning is at the officer's discretion; in this database, about 46% of the 68,357 stops resulted in a speeding ticket. If a ticket is issued, state law provides a formula for the amount of the fine: \$50 + \$10 × (speed − (speed limit + 10)). Makowsky and Stratmann (2009, p.513) discuss the political economy hypothesis and the opportunity-cost hypothesis of officer behavior. The former relates the officers' decision to "the fiscal condition of the government that employs them and to whether the driver is a potential voter in local elections," and the latter predicts that "officers have a higher likelihood of issuing a ticket and issuing a larger fine amount when the opportunity cost for contesting the ticket is higher for drivers."

For the purpose of this paper, we ignore the selection issue whether a ticket is issued and concentrate on the amount of fine, which in this database averages \$122. The regression model expresses this outcome variable in logarithmic form. The explanatory

[18]Web Appendix G explores another model of Eicher et al. (2012) with similar results.

Table 5: Speeding tickets, Massachusetts, 2005

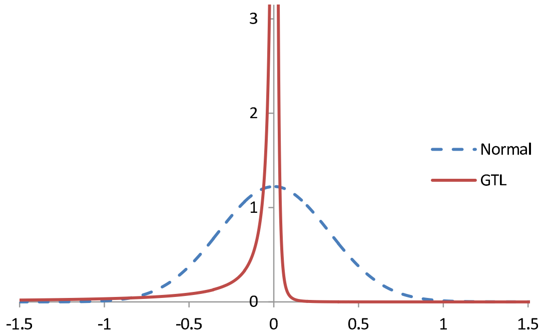
	OLS		GTLE			
	Estimate	Stan.Err.	Estimate	Stan.Err.	$\frac{\hat{\beta}_{OLS}-\hat{\beta}_{GTL}}{\hat{\beta}_{GTL}}$	$\frac{SE_{GTL}}{SE_{OLS}}$
ln(Mph over)	0.8649	0.0144	1.2332	0.0057	-0.299	0.396
OutTown	0.0071	0.0107	0.0001	0.0006	119.524	0.054
OutState	0.0353	0.0070	0.0007	0.0004	52.026	0.062
Afr.American	-0.0207	0.0095	-0.0010	0.0004	20.402	0.045
Hispanic	0.0216	0.0101	-0.0003	0.0007	-85.838	0.064
Female	-0.0564	0.0375	0.0003	0.0017	-204.497	0.046
ln(Age)	0.0002	0.0073	0.0015	0.0007	-0.878	0.092
Female $\times$ ln(Age)	0.0092	0.0105	0.0000	0.0005	946.064	0.049
ln(CourtDist)	0.0158	0.0035	-0.0002	0.0003	-81.445	0.074
ln(Pvalue.pc)	-0.0043	0.0218	-0.0007	0.0011	5.209	0.050
OR	0.0202	0.0700	0.0058	0.0020	2.502	0.028
OR $\times$ OutTown	0.0083	0.0630	-0.0082	0.0019	-2.008	0.031
OR $\times$ ln(CourtDist)	0.0019	0.0089	0.0003	0.0004	5.139	0.042
SP	0.0299	0.2949	-0.0016	0.0155	-19.662	0.053
SP $\times$ OutTown	0.0258	0.0187	0.0000	0.0008	-4569.156	0.041
SP $\times$ ln(CourtDist)	0.0059	0.0040	0.0007	0.0004	7.853	0.101
SP $\times$ ln(Pvalue.pc)	-0.0018	0.0265	-0.0002	0.0014	9.790	0.051
SP $\times$ OR	-0.0164	0.0319	-0.0018	0.0017	8.270	0.053
Intercept	2.2716	0.2551	1.4817	0.0190		
σ			0.0030	0.0005		
α			-1.7612	0.0833		
δ			1.3083	0.0483		
log Likelihood	-9440.55		25066.89			
(Absolute) Average					360.723	0.075

Dependent variable: Log of amount of fine. Number of observations = 31674. Skewness and kurtosis of OLS residuals equal -1.20 and 5.07; the Jarque-Bera test of normality of the OLS residuals has a p-value of less than 0.001. The Wald test of the GTL estimates of (α, δ) equals 791.1, rejecting normality with a p-value of less than 0.001. The LM test equals 5072 with a p-value of less than 0.001 and with 8 support violations. The Vuong test that compares OLS and GTLE equals -52.67 in favor of the GTL model with a p-value of less than 0.001.

variables include the excess speed of the driver ("Mph over", in log form), driver characteristics (residence, race, ethnicity, gender, age, and the distance to court), and measures of the fiscal condition of a municipality (a dummy "OR" whether a municipality rejected a tax increase via an override referendum applicable to the operating budget of the 2001 fiscal year; property value per capita; and a dummy "SP" whether the traffic stop was made by a state police officer, who may have different incentives than a local police officer). The regression model includes several interactions as well.

The difference in the estimates is striking (Table 5). The first sign of the relative inadequacy of the OLS estimator is seen in the skewness (-1.20) and kurtosis (5.07) of the

Figure 4: Estimated distributions of ϵ for the speeding ticket equation



residuals. As for the GTLE, the estimates $(\hat{\alpha}, \hat{\delta}) = (-1.761, 1.308)$ are far from $(0.1436, 0)$ that represents normality; none of the moments of the estimated GTL distribution even exists. The disturbances are sharply peaked and have thick tails, with the left tail thicker than the right; see Figure 4 where the GTL density that peaks at 33.23 at $\epsilon = 0.0075$ is actually top-truncated to show the tails better. The two densities are drawn such that the median of both falls at 0. They intersect at -0.8468, -0.0611, 0.0362, and 1.5974, dividing the real axis in five segments. For normality, the probability of these segments equals 0.0047, 0.4210, 0.1185, 0.4558, and 4.9E-07; for GTL, the probability equals 0.1102, 0.1413, 0.7275, 0.0211, and 5.5E-06. This highlights the thick tails, especially on the left, and the high pile-up of the GTL disturbances in the interval from -0.0611 to 0.0362. That means that the actual log-fine generally deviates little from the predicted log-fine—but if it deviates, it can deviate much.

About the determinants of the log-fine, both estimates show that excess speed (ln(Mph over)) is the major determinant, as it should be, but the two estimates have different implications. The OLS slope estimate is less than 1, indicating that the fine is inelastic with respect to the severity of the speeding violation (miles over the speed limit). On the other hand, the GTL estimate indicates an elasticity greater than 1: the fine is elastic. This is more intuitive: a more serious speeding violation draws an increasingly severe penalty; this also corresponds with the prescription in state law.[19]

[19]Simple algebra with the formula for the amount of fine reveals that the elasticity exceeds 1 as long as the driver exceeded the speed limit by 5 miles, and the elasticity is not constant as it is in the estimated double-log model. In the entire sample, only 1% of the stopped drivers were going less than five miles

The other slopes are also much different. The OLS estimates suggest that African-Americans (and perhaps young females) pay lower fines, and that Hispanic and out-of-state drivers as well as those living farther away from the courthouse pay more. Regardless of whether these effects are intuitively plausible, none of the GTL-estimated slopes of these variables is economically significant any longer, even if a few of them are still statistically significant. In other words, the amount of fine is not varying at the discretion of the police officer in response to observable factors, in contrast with the findings by Makowsky and Stratmann (2009), but unobservable factors can occasionally cause major deviations from the fine that state law prescribes.

6 Concluding remarks

Because of the well-known Gauss-Markov theorem, OLS is the workhorse estimation approach to linear regression models, to be discarded only if its basic assumptions are clearly violated, such as endogenous regressors or serial correlation. Skewed and thick-tailed disturbances do not constitute a violation of those assumptions, and as long as the first and second moments of the disturbance exist, OLS is still the best estimator within the class of linear unbiased estimators. But even in this case, nonlinear estimators that fall outside this class may be efficient relative to OLS. Moreover, when the distribution of the disturbances is strongly skewed or has very thick tails, the second (and even the first) moment may not even exist, which invalidates statistical inference with the OLS estimator. Plainly stated, with skewed and thick-tailed disturbances, a nonlinear estimator may have a smaller variance than OLS.

This paper develops the GTL estimator, which is the maximum likelihood estimator of a linear regression model with GTL-distributed disturbances. Statistically, it has good properties (consistent and asymptotically normal). Monte Carlo comparisons demonstrate its dominance when disturbances are non-normally distributed. Two applications highlight the practical relevance of the GTL regression approach. First, they illustrate the

over speed limit; one fifth of them received a ticket.

fact that disturbances are often not normally distributed but rather exhibit skewness and a higher degree of kurtosis than normality. Second, they illustrate that in some research problems the location parameters (the slopes and intercept of the regression model) are highly sensitive to variations in the distributional assumption and in other research problems these parameters are very robust. This difference in the effect of non-normality is not foreseeable; the only way to find out is to test for normality of the residuals of OLS and, if normality is rejected, to estimate the regression model with a GTL estimator. Third, the applications illustrate that, robust or not, the location parameters can be estimated with greater precision with a GTL estimator. Efficiency gains vary between research applications and generally are larger when disturbances are more distinctly non-normal.

Disturbances need not be GTL-distributed. The assumption of a GTL distribution is one of several alternatives the applied econometrician can choose from when confronted with non-normal disturbances. But, among the class of unimodal distributions, the GTL distribution has much to commend it: it is parsimonious, highly flexible, able to accommodate both thin (truncated) tails and thick tails as well any shape between symmetry and extreme degrees of skewness, and able to model data generated by a distribution that does not have finite moments. This flexibility also makes the GTL distribution appealing for, e.g., discrete choice (Author, yearb) and GARCH modeling (Author, yeara), where, just as in linear regression models, applied econometricians may benefit from a wider range of modeling tools.

References

Author, yeara. Blinded title. Blinded Journal .

Author, yearb. Blinded title. Blinded Journal .

Azzalini, A., Capitanio, A., 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 65(2), 367–389.

Barndorff-Nielsen, O. E., 1997. Normal inverse gaussian distributions and stochastic volatility modelling. Scandinavian Journal of Statistics 24(1), 1–13.

Blattberg, R., Sargent, T., 1971. Regression with non-gaussian stable disturbances: some sampling results. Econometrica 39(3), 501–510.

Breusch, T. S., Robertson, J. C., Welsh, A. H., 1997. The emperor's new clothes: a critique of the multivariate t regression model. Statistica Neerlandica 51(3), 269–286.

Carroll, R., Fan, J., Gijbels, I., Wand, M., 1997. Generalized partially linear single-index models. Journal of the American Statistical Association 92(438), 477–489.

Chen, E. H., Dixon, W. J., 1972. Estimators for the linear regression model based on Winsorized observations. Journal of the American Statistical Association 67(339), 664–671.

Chen, L.-A., Welsh, A. H., Chan, W., 2001. Estimators for the linear regression model based on Winsorized observations. Statistica Sinica 11(1), 147–172.

Eicher, T. S., Henn, C., Papageorgiou, C., 2012. Trade creation and diversion revisited: accounting for model uncertainty and natural trading partner effects. Journal of Applied Econometrics 27(2), 296–321.

Fernancez, C., Steel, M. F., 1999. Multivariate Student-t regression models: Pitfalls and inference. Biometrika 86(1), 153–167.

Fernandez, C., Steel, M. F. J., 1998. On bayesian modeling of fat tails and skewness. Journal of the American Statistical Association 93(441), 359–371.

Ferreira, J., Steel, M., 2006. A constructive representation of univariate skewed distributions. Journal of the American Statistical Association 101(474), 823–829.

Freimer, M., Mudholkar, G. S., Kollia, G., Lin, C. T., 1988. A study of the generalized tukey lambda family. Communications in Statistics: Theory and Methods 17, 3547–3567.

Geweke, J., 1993. Bayesian treatment of the independent student-t linear model. Journal of Applied Econometrics 8(Suppl.), S19–S40.

Ghosh, S., Yamarik, S., 2004. Are regional trading arrangement trade creating? An application of extreme bounds analysis. Journal of International Economics 63(2), 369–395.

Gourieroux, C., Monfort, A., Trognon, A., 1984. Pseudo maximum likelihood methods: Theory. Econometrica 52(3), 681–700.

Grigoletto, M., Lisi, F., 2009. Looking for skewness in financial time series. Econometrics Journal 12(2), 310–323.

Hallin, M., Swan, Y., Verdebout, T., Veredas, D., 2011. Rank-based testing in linear models with stable errors. Journal of Nonparametric Statistics 23(2), 305–320.

Hallin, M., Swan, Y., Verdebout, T., Veredas, D., 2013. One-step R-estimation in linear models with stable errors. Journal of Econometrics 172(2), 195–204.

Hansen, C. B., McDonald, J. B., Theodossiou, P., 2007. Some flexible parametric models for partially adaptive estimators of econometric models. Economics: The Open-Access, Open-Assessment E-Journal 1(2007-7).

Harvey, A. C., Sucarrat, G., 2014. Evaluating density forecasts with applications to financial risk management. Computational Statistics and Data Analysis 76(SI), 320–338.

Harvey, C. R., Siddique, A., 1999. Autoregressive conditional skewness. Journal of Financial and Quantitative Analysis 34, 465–487.

He, X., Jureckova, J., Koenker, R., Portnoy, S., 1990. Tail behavior of regression estimators and their breakdown points. Econometrica 58(5), 1195–1214.

Horowitz, J. L., 1998. Semiparametric Methods in Econometrics. Springer-Verlag, New York.

Huber, P. J., 1964. Robust estimation of a location parameter. Annals of Mathematical Statistics 35, 73—101.

Huber, P. J., 1981. Robust Statistics. Wiley, New York.

Huber, P. J., Ronchetti, E. M., 2009. Robust Statistics. Wiley, New York, 2nd edition.

Ichimura, H., 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single index models. Journal of Econometrics 38(1-2), 195–204.

Jureckova, J., Koenker, R., Portnoy, S., 2001. Tail behavior of the least-squares estimator. Statistics and Probability Letters 55(4), 377–384.

Karian, Z. A., Dudewicz, E. J., 1999. Fitting the Generalized Lambda Distribution to data: A method based on percentiles. Communications in Statistics: Simulation & Computation 28(3), 793–819.

Karian, Z. A., Dudewicz, E. J., McDonald, P., 1996. The extended Generalized Lambda Distribution system for fitting distributions to data: History, completion of theory, tables, applications, the 'final word' on moment fits. Communications in Statistics–Simulation 25(3), 611–642.

Koenker, R., 2005. Quantile Regression. Cambridge University Press, New York.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46(1), 33–50.

Koenker, R., Portnoy, S., 1987. L-Estimation for linear models. Journal of the American Statistical Association 82(399), 851–857.

Komunjer, I., 2007. Asymmetric power distribution: theory and applications to risk measurement. Journal of Applied Econometrics 22(5), 891–921.

Lange, K. L., Little, R. J. A., Taylor, J. M. G., 1989. Robust statistical modeling using the t distribution. Journal of the American Statistical Association 84(408), 881–896.

Li, Q., Racine, J. S., 2007. Nonparametric Econometrics. Princeton University Press, Princeton.

Makowsky, M. D., Stratmann, T., 2009. Political economy at any speed: What determines traffic citations? American Economic Review 99(1), 509–527.

McDonald, J. B., Butler, R. J., 1987. Generalized mixture distributions with an application to unemployment duration. Review of Economics and Statistics 69(2), 232–240.

Mikosch, T., de Vries, C. G., 2013. Heavy tails of OLS. Journal of Econometrics 172(2), 205–221.

Nolan, J. P., Ojeda-Revah, D., 2013. Linear and nonlinear regression with stable errors. Journal of Econometrics 172, 186–194.

Öztürk, A., Dale, R. F., 1985. Least squares estimation of the parameters of the Generalized Lambda distribution. Technometrics 27(1), 81–84.

Powell, J., Stock, J., Stoker, T. M., 1989. Semiparametric estimation of index coefficients. Econometrica 51(2), 1403–1430.

Pregibon, D., 1980. Goodness of link tests for generalized linear models. Journal of the Royal Statistical Society, Series C (Applied Statistics) 29(1), 15–23.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., 1986. Numerical Recipes: The Art of Scientific Computing. Cambridge University Press, Cambridge.

Ramberg, J., Schmeiser, B., 1974. An approximate method for generating asymmetric random variables. Communications ACM 17, 78–82.

Ramberg, J. S., Tadikamalla, P. R., Dudewicz, E. J., Mykytka, E. F., 1979. A probability distribution and its uses in fitting data. Technometrics 21(2), 201–214.

Samorodnitsky, G., Rachev, S. T., Kurz-Kim, J.-R., Stoyanov, S. V., 2007. Asymptotic distribution of unbiased linear estimators in the presence of heavy-tailed stochastic regressors and residuals. Probability and Mathematical Statistics 27(2), 275–302.

Su, S., 2011. Fitting GLD to data via quantile matching method. In: Karian, Z., Dudewicz, E. (Eds.), Handbook on Fitting Statistical Distributions with R, chapter 14, pp. 557–583, Boca Raton.

Subramanian, A., Wei, S.-J., 2007. The WTO promotes trade, strongly but unevenly. Journal of International Economics 72(1), 151–175.

White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50(1), 1–25.

White, H., 1984. Asymptotic Theory for Econometrians. Academic Press, San Diego, California.

Wichitaksorn, N., Tsurumi, H., 2013. Comparison of mcmc algorithms for the estimation of tobit model with non-normal error: The case of asymmetric laplace distribution. Computational Statistics and Data Analysis 67, 226–235.

Wraith, D., Forbes, F., 2015. Location and scale mixtures of gaussians with flexible tail behaviour: Properties, inference and application to multivariate clustering. Computational Statistics and Data Analysis 90, 61–73.

Yale, C., Forsythe, A. B., 1976. Winsorized regression. Technometrics 18(3), 291–300.

Zellner, A., 1976. Bayesian and non-bayesian analysis of the regression model with multivariate Student t error terms. Journal of the American Statistical Association 71(354), 400–405.

Web Appendices

A Variable Definitions and Descriptive Statistics

The data used in the trade creation and trade diversion application are derived from the dataset used in Eicher et al. (2012);[20] variable names have been slightly changed. "Remoteness" is defined as the sum of the log of the average distance, weighed by relative GDP, of each country from all trading partners. The data are available online at `http://qed.econ.queensu.ca/jae/2012-v27.2 /eicher-henn-papageorgiou/`. Definition of the variables and descriptive statistics are provided in Table A.1.

The data used in the speeding tickets application are derived from Makovsky and Stratmann (2009);[21] variable names have been slightly changed. The data are available online at `http://www.aeaweb.org/ issue.php?journal=AER&volume=99&issue=1`. Definition of the variables and descriptive statistics are provided in Table A.2.

[20] T.S. Eicher, C. Henn, and C. Papageorgiou, "Trade creation and diversion revisited: accounting for model uncertainty and natural trading partner effects." *Journal of Applied Econometrics*, 2012, 27(2):296-321.

[21] M.D. Makowsky and T. Stratmann, "Political Economy at Any Speed: What Determines Traffic Citations?" 2009, 99(1):509-527.

Table A.1: Trade creation and trade diversion, 1960-2000: Descriptive statistics

Variable	Definition	Mean	St.Dev	Min	Max
lnimportij	log of bilateral imports from j to i (current US dollars)	11.14	3.27	-3.54	21.01
Trade creation dummy variables					
tc.nafta	1 = both countries are NAFTA members	0.000	0.019	0	1
tc.eu	1 = both countries are EU members	0.016	0.124	0	1
tc.efta	1 = both countries are EFTA members	0.006	0.077	0	1
mx.eea	1 = both countries are EEA members	0.012	0.108	0	1
tc.caricom	1 = both countries are CARICOM members	0.001	0.030	0	1
tc.ap	1 = both countries are Andean Pact members	0.004	0.061	0	1
tc.mercosur	1 = both countries are MERCOSUR members	0.001	0.025	0	1
tc.asean	1 = both countries are ASEAN FTA members	0.001	0.027	0	1
tc.certain	1 = both countries are ANZCERTA members	0.000	0.015	0	1
tc.apec	1 = both countries are APEC members	0.013	0.114	0	1
tc.laia	1 = both countries are LAIA members	0.021	0.143	0	1
tc.cacm	1 = both countries are CACM members	0.004	0.064	0	1
tc.bilateralPTA	1 = both countries are in a joint BPTA	0.005	0.070	0	1
Trade diversion dummy variables					
td.nafta	1 = only one country is a NAFTA member	0.033	0.178	0	1
td.eu	1 = only one country is an EU member	0.262	0.440	0	1
td.efta	1 = only one country is an EFTA member	0.160	0.367	0	1
td.eea	1 = only one country is an EEA member	0.107	0.310	0	1
td.caricom	1 = only one country is a CARICOM member	0.047	0.211	0	1
td.ap	1 = only one country is an Andean Pact member	0.091	0.287	0	1
td.mercosur	1 = only one country is a MERCOSUR member	0.026	0.159	0	1
td.asean	1 = only one country is an ASEAN FTA member	0.033	0.178	0	1
td.anzcerta	1 = only one country is an ANZCERTA member	0.032	0.176	0	1
td.apec	1 = only one country is an APEC member	0.134	0.340	0	1
td.laia	1 = only one country is a LAIA member	0.200	0.400	0	1
td.cacm	1 = only one country is a CACM member	0.083	0.276	0	1
td.bilateralPTA	1 = only one country is a member of a BPTA	0.170	0.376	0	1
Non-PTA control variables					
lpgdpij	Sum over (i,j) of log nominal GDP	6.58	3.06	-4.29	17.65
lpgdppcij	Sum over (i,j) of log real GDP per capita	17.38	1.41	12.24	20.85
ldist	Log of bilateral distance	8.23	0.81	4.40	9.42
sachsij	Sum over (i,j) of the Sachs-Warner trade policy index	1.22	0.70	0	2
vola3	St.dev. of the volatility in the bilateral exchange rate	4.83	7.32	0	97.61
floatij	Sum over (i,j) of dummy: 1 if floating exchange	0.71	0.73	0	2
cu	1 = a common currency union	0.009	0.097	0	1
adifsecschool25	Abs. log difference in years of secondary schooling	1.13	0.91	0.00	5.86
adifdensity	Abs. log difference in population density	1.70	1.32	0.00	8.24
adifgdppc	Abs. log difference of real GDP per capita	1.25	0.91	0.00	4.07
border	1 = common land border	0.033	0.180	0	1
islandij	Sum over (i,j) of dummy: 1 if island	0.272	0.487	0	2
landlockij	Sum over (i,j) of dummy: 1 if landlocked	0.228	0.448	0	2
lpareaij	Sum over (i,j) of log surface area	24.69	2.89	11.82	32.08
lremoteij	Remoteness	17.94	0.35	16.97	19.03
colony	1 = one country was a former colony of the other	0.028	0.165	0	1
comcol	1 = common colonizer	0.064	0.244	0	1
comlang	1 = common language	0.235	0.424	0	1
N of obs		37983			

Table A.2: Speeding tickets, Massachusetts, 1987: Descriptive statistics

Variable	Definition	Mean	St.Dev	Min	Max
Amount	Fine amount (in $)	122.03	56.25	3.00	725.00
ln(Amount)	log of Amount	4.707	0.438	1.099	6.586
Mph over	Miles per hour over the speed limit	17.08	5.79	1	75
ln(Mph over)	log of mph over	2.783	0.333	0	4.317
Afr.American	1 if the driver is African American	0.051	0.219	0	1
Hispanic	1 if the driver is Hispanic	0.047	0.211	0	1
Female	1 if the driver is female	0.332	0.471	0	1
ln(Age)	log of age (in years)	3.442	0.366	2.485	4.585
OutTown	1 if out of town driver	0.847	0.360	0	1
OutState	1 if out of state driver	0.221	0.415	0	1
ln(CourtDist)	log of distance to court (in miles)	2.886	1.298	1.609	8.529
ln(Pvalue.pc)	log property value per capita	11.165	0.499	9.828	13.580
OR	1 if a tax increase rejected via override refe	0.026	0.160	0	1
SP	1 if the officer is state police	0.445	0.497	0	1
N of obs		31674			

B Contrasting OLS estimation, GTL estimation, and M-estimation

Recall the regression model that is the subject of this paper:

$$y_i = x_i'\beta + \sigma\epsilon_i \qquad \text{where} \qquad \epsilon_i \sim \text{iid GTL}(\alpha, \delta). \tag{B.1}$$

where $i = 1, \ldots, n$ denotes observations (individuals, states, time periods, etc.). x_i and β are $k \times 1$ vectors. ϵ has a canonical GTL(α, δ) distribution and is assumed independent of x, such that $E[\epsilon|x] = E[\epsilon]$ if indeed ϵ has a first moment.

Equation (B.1) is a simple linear regression model that may be estimated with Ordinary Least Squares (OLS): provided that ϵ has first and second moments (and subject to other assumptions),[22] the OLS estimator $\hat{\beta}_{OLS}$ is efficient in the class of linear unbiased estimators, regardless of the distribution of ϵ. If ϵ is normally distributed, $\hat{\beta}_{OLS}$ is identical to the maximum likelihood estimator (MLE), and if ϵ has some other distribution, $\hat{\beta}_{OLS}$ is equivalent to the quasi-ML estimator that uses a gaussian criterion function (QMLg). Var$(\hat{\beta}_{OLS})$ exists if ϵ has a finite second moment, and the slope elements of $\hat{\beta}_{OLS}$ are unbiased if ϵ has a constant finite first moment. If these moments do not exist, $\hat{\beta}_{OLS}$ may be computed numerically, but the usual statistical inference is invalid.[23]

If ϵ has a GTL distribution, the OLS estimator generally differs from the GTLE and is no longer efficient relative to GTLE if the GTL distribution deviates from normality. As proven in Appendix C.2, the GTLE of β is consistent and asymptotically normal, even when ϵ does not possess first and second moments.

[22] For a technical discussion, see H. White, *Asymptotic Theory for Econometricians*, Academic Press, San Diego, California, 1984.

[23] Under the theory of quasi (pseudo)-maximum likelihood, violation of the normality assumption by itself does not necessarily lead to inconsistency or asymptotic non-normality of the OLS estimator. The OLS estimator is a quasi maximum likelihood estimator with a gaussian criterion function (QMLg). Since this criterion function belongs to the class of linear exponential distributions, the OLS estimator of β is consistent and asymptotically normal when disturbances have finite first and second moments; and the OLS estimator of β is consistent but not necessarily asymptotically normal when disturbances only have a finite first moment. If neither moment is finite, even the consistency property of OLS is lost. See H. White, "Maximum likelihood estimation of misspecified models." *Econometrica*, 1982, 50(1):1-25, and C. Gourieroux, A. Monfort, and A. Trognon, "Pseudo maximum likelihood method: theory." *Econometrica*, 1984, 52(3):681-700 (in particular, Theorems 1, 2 and 3).

The difference between the OLS and GTL estimators follows from the first order conditions that are implied by maximization of the respective criterion functions. Provided that σ_ϵ exists, the first-order condition of OLS (or QMLg) implies:

$$\sum_{i=1}^{n} \tilde{\epsilon}_i x_i = 0 \tag{B.2}$$

where $\tilde{\epsilon}_i = (y_i - x_i'\beta)/(\sigma\sigma_\epsilon)$ is the standardized disturbance. The first-order condition of GTLE yields:

$$\sum_{i=1}^{n} \frac{1}{\sigma} \frac{G''(u_i)}{(G'(u_i))^2} x_i \equiv \sum_{i=1}^{n} w_{GTL}(\tilde{\epsilon}_i)\tilde{\epsilon}_i x_i = 0 \tag{B.3}$$

where $u_i = G^{-1}(\sigma_\epsilon \tilde{\epsilon}_i + \mu_\epsilon)$. The term $w_{GTL}(\tilde{\epsilon}_i) = \frac{1}{\sigma} \frac{G''(u_i)}{(G'(u_i))^2 \tilde{\epsilon}_i}$ is the weight given to $\tilde{\epsilon}_i x_i$ for observation i. This weight distinguishes OLS from GTLE: under OLS, the weight that is implied by equation (5) is $w_{OLS}(\tilde{\epsilon}_i) = 1$ for all $\tilde{\epsilon}_i$, whereas $w_{GTL}(\tilde{\epsilon}_i)$ varies with $\tilde{\epsilon}_i$. By varying the weight according to the distribution of $\tilde{\epsilon}$ that is detected in the data, the GTLE is able to avoid the disruptive effect of tail values in the data and thus exploit the tail information more effectively than OLS.

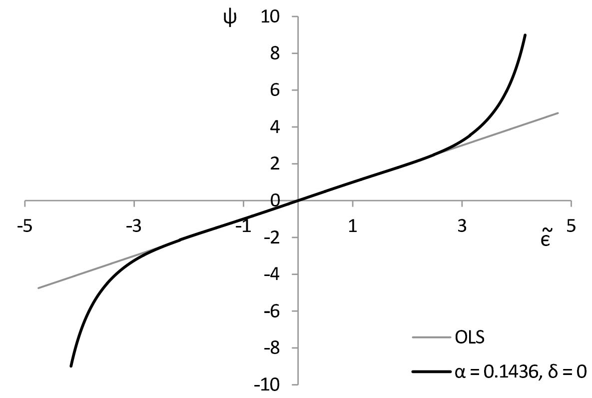
M-estimation makes use of an objective function $L = \sum_{i=1}^{n} \rho(\tilde{\epsilon}_i)$, which is minimized to find $\hat{\beta}_M$. The first order conditions may be stated as:

$$\sum_{i=1}^{n} \frac{1}{\sigma} \frac{d\rho(\tilde{\epsilon}_i)}{d\tilde{\epsilon}_i} x_i \equiv \sum_{i=1}^{n} w_M(\tilde{\epsilon}_i)\tilde{\epsilon}_i x_i = 0 \tag{B.4}$$

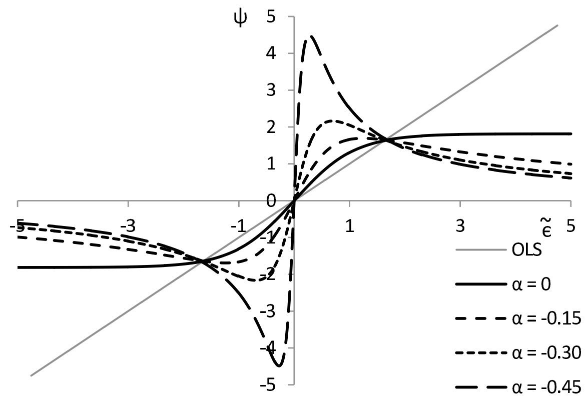
with $w_M(\tilde{\epsilon}) = \frac{1}{\sigma\tilde{\epsilon}} \frac{d\rho(\tilde{\epsilon})}{d\tilde{\epsilon}}$. Thus, M-estimation differs from GTL regression in that M-estimation chooses $\rho(\tilde{\epsilon})$ a priori, and thus also $w_M(\tilde{\epsilon})$, without reference to the sample data at hand.

For greater clarity, for $E = \{OLS, M, GTL\}$, consider the influence functions $\psi_E(\tilde{\epsilon}) = w_E(\tilde{\epsilon})\tilde{\epsilon}$, such that $\psi_{OLS}(\tilde{\epsilon}) = \tilde{\epsilon}$; $\psi_M(\tilde{\epsilon}) = \frac{d\rho(\tilde{\epsilon})}{d\tilde{\epsilon}}$; and $\psi_{GTL}(\tilde{\epsilon}) = \frac{1}{\sigma} \frac{G''(u)}{(G'(u))^2}$ with $u = G^{-1}(\sigma_\epsilon\tilde{\epsilon} + \mu_\epsilon)$. Figure B.1 contrasts $\psi_{GTL}(\tilde{\epsilon})$ for various (α, δ) combinations with $\psi_{OLS}(\tilde{\epsilon}_i)$, which is a simple 45° line. Figure B.1a compares ψ_{OLS} with ψ_{GTL} for a GTL(0.1436,0) distribution. Over the interval $[-3, 3]$, which contains 99.5% of the distribution, the two lines are nearly

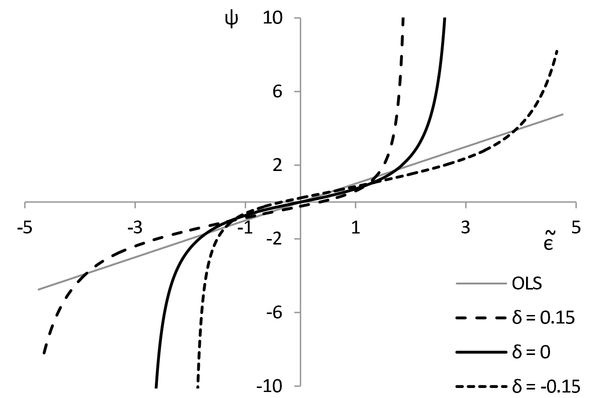
Figure B.1: Comparing $\psi_{OLS}(\tilde{\epsilon})$ with $\psi_{GTL}(\tilde{\epsilon})$ for selected values of α and δ



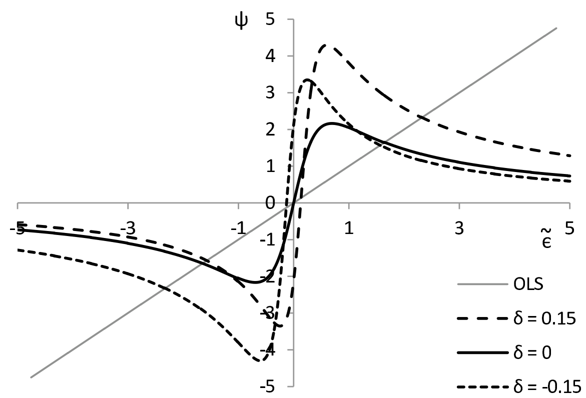
(a) Comparing $N(0,1)$ with GTL$(0.1436, 0)$



(b) Increasingly thick tails: $\delta = 0$



(c) Thin tails with skewness: $\alpha = 0.30$



(d) Thick tails with skewness: $\alpha = -0.30$

identical; tail values outside this interval have more influence on the GTLE estimator than OLS because for $(\alpha, \delta) = (0.1436, 0)$ the feasible support of $\tilde{\epsilon}$ is finite. In contrast, when tails become thicker (Figure B.1b), observations with large $\tilde{\epsilon}$ are increasingly discounted.[24] Thus, a large $\tilde{\epsilon}$ is informative in a distribution with thin tails and a finite support because it more clearly defines the set of plausible values of $\hat{\beta}$, whereas a large $\tilde{\epsilon}$ in the context of a distribution with thick tails is rather uninformative. Figures B.1c and B.1d show that skewness in the distribution generates asymmetry in $\psi_{GTL}(\tilde{\epsilon})$ around $\tilde{\epsilon} = 0$. When δ is negative and the GTL distribution is right-skewed, the right tail is longer than the left and thus observations with large positive disturbances are weighted less heavily than observations with large negative disturbances.

It should be noted that graphs of $w_{GTL}(\tilde{\epsilon})$ may be deduced from Figure B.1 since $w_{GTL}(\tilde{\epsilon}) = \psi_{GTL}(\tilde{\epsilon})/\psi_{OLS}(\tilde{\epsilon})$. Moreover, recall that $w_{OLS}(\tilde{\epsilon}) = 1$. Thus, for configurations of Figure B.1b, the typical w_{GTL}-curve exceeds 1 in the middle around the origin and trails off to 0 near the ends of the support. However, for skewed configurations, ψ_{GTL} is negative for some values where $\tilde{\epsilon}$ is positive or vice versa, giving rise to locally negative values of w_{GTL} near the origin and vertical asymptotes at the origin. Because of this, graphs of ψ_{GTL} provide better intuition.

$\psi_M(\tilde{\epsilon})$ is the influence function of M-estimation, for which the literature shows many examples. This influence function depends on the choice of $\rho(\tilde{\epsilon})$, which is perhaps strategic but also *a priori* and not data-driven. In contrast, $\psi_{GTL}(\tilde{\epsilon})$ reflects the distribution detected in the data, which the GTL distribution approximates in a flexible way. In this way, GTL regression gives proper weight to outliers—provided, of course, that the outliers are generated by the same probability law that generates the typical ("non-

[24] For a similar pattern in the context of a stable distribution that permits thick-tailed and skewed disturbances, see J.P. Nolan and D. Ojeda-Revah, "Linear and nonlinear regression with stable errors." *Journal of Econometrics*, 2013, 172:186-194. The distribution is characterized in part by an index of stability denoted as α (which has a different meaning than the parameter α in the GTL distribution), with $0 < \alpha \leq 2$. The comparision of the GTL and stable distributions is left as a future research topic, but suffice it to note that stable distributions have no moments of order α or higher—and thus no variance unless $\alpha = 2$, which is the special case that corresponds with the normal distribution; see G. Samorodnitsky, S.T. Rachev, J.-R. Kurz-Kim, and S.V. Stoyanov, "Asymptotic distribution of unbiased linear estimators in the presence of heavy-tailed stochastic regressors and residuals." *Probability and Mathematical Statistics*, 2007, 27(2):275-302.

outlier") members of the sample. Thus, the flexibility of the GTL distribution gives GTL regression an advantage over M-estimation.

C Proofs

C.1 Proof of Theorem 1

The proof of Theorem 1 relies on Theorem 2.5 of Newey and McFadden (1994)[25], which consists of five components: (i) ϵ_i must be i.i.d.; (ii) θ_0 must be in a compact set; (iii) the log-likelihood function must be continuous at every θ with probability 1; (iv) θ must be identified; and (v) $E[\sup_{\theta\in\Theta} |\ln g(z|\theta)|] < \infty$.

Conditions (i) and (ii) are satisfied by virtue of Assumptions A.2(i) and A.3. Regarding condition (iii), continuity of the likelihood function is straightforwardly verified when $\alpha \neq \pm\delta$. When $\alpha = \delta$ and therefore $\lambda_1 = 0$, we have

$$\lim_{\lambda_1\to 0} \frac{u^{\lambda_1} - 1}{\lambda_1} = \lim_{\lambda_1\to 0} \frac{u^{\lambda_1}\ln u}{1} = \ln u.$$

by L'Hôpital's Rule. A similar argument applies to the case where $\alpha = -\delta$. Thus, the log likelihood function is continuous at every θ.

Regarding condition (iv), Lemma 2.2 of Newey and McFadden (op. cit.) provides conditions for identification:

> If θ_0 is identified and $E[|\ln g(z|\theta)|] < \infty$ for all θ, then $Q_0(\theta) = E[\ln g(z|\theta)]$ has a unique maximum at θ_0.

We shall first check whether $E[|\ln g(z|\theta)|] < \infty$ for all θ.

The regression model is given in equation (3). Define the support of ϵ as $\mathcal{E}(\theta)$. Define $z = (y, x)$ with $x \in \mathcal{X}$ and $y \in \mathcal{Y}(x, \theta)$, where $\mathcal{Y}(x, \theta)$ is $\mathcal{E}(\theta)$ shifted by $x'\beta$. Given the independence of x and ϵ, $g(z) = g_y(y|x, \theta)g_x(x)$ with $g_y(y|x, \theta) = \frac{1}{\sigma} f(\frac{y-x'\beta}{\sigma}|\theta)$, where $f(\cdot|\theta)$ is the GTL density. Successive transformation of variables ($\epsilon = (y - x'\beta)/\sigma$ and

[25] Whitney K. Newey and Daniel McFadden, "Large Sample Estimation and Hypothesis Testing." In R.F. Engle and D.L. McFadden, eds., *Handbook of Econometrics*, Vol. IV, Ch. 36, pp. 2111-2245, 1994, New York: Elsevier Science B.V.

$u = G^{-1}(\epsilon)$) yields

$$E[|\ln g(z|\theta)|] = \int_{\mathcal{X}} \int_{\mathcal{Y}(x,\theta)} |\ln g(y,x|\theta)| g(y,x|\theta) dy dx \tag{C.1}$$

$$= \int_{\mathcal{X}} \int_{\mathcal{E}(\theta)} |\ln\{\frac{1}{\sigma} f(\epsilon|\theta) g_x(x)\}| f(\epsilon|\theta) g_x(x) d\epsilon dx \tag{C.2}$$

$$= \int_{\mathcal{X}} \Big(\int_0^1 |-\ln\sigma - \ln G'(u) + \ln g_x(x)| du \Big) g_x(x) dx. \tag{C.3}$$

since the density of u is uniform on $[0,1]$. With $\sigma > 0$ and x being properly distributed as a vector of explanatory variables, the focus of equation (C.3) is on $G'(u)$. Thus, write the term in the large parentheses as

$$\int_0^1 |A - \ln G'(u)| du = \int_0^1 |A - \ln(u^{\lambda_1 - 1} + (1-u)^{\lambda_2 - 1})| du. \tag{C.4}$$

$G'(u)$ is well-behaved for any $u \in [u_\eta, 1-u_\eta]$ for a small u_η but may have asymptotes at $u = 0$ and/or $u = 1$. Thus, we must check for the behavior of $G'(u)$ in the intervals $[0, u_\eta]$ and $[1-u_\eta, 1]$. For $\lambda_1 > 1$ and $u \downarrow 0$, $G'(u) \to 1$ since $u^{\lambda_1 - 1} \downarrow 0$ and $(1-u)^{\lambda_2 - 1} \to 1$. For $\lambda_1 = 1$, $u^{\lambda_1 - 1} = 1$, so as $u \downarrow 0$, $G'(u) \to 2$ since $(1-u)^{\lambda_2 - 1} \to 1$. For $\lambda_1 < 1$, consider that $G'(u) \to \infty$ as $u \downarrow 0$. Thus, choose u_η such that $\ln G'(u_\eta) \geq A$. Rewrite $-\ln G'(u) = -\ln(u^{\lambda_1 - 1}(1 + u^{1-\lambda_1}(1-u)^{\lambda_2 - 1})) \equiv -((\lambda_1 - 1)\ln u + \ln B(u))$, where $B(0) = 1$ and $B(u) > 1$ for $u \in (0, u_\eta)$ for a small enough u_η. Thus, since $A - \ln G'(u) \leq 0$ for all $u \in [0, u_\eta]$, we have

$$\int_0^{u_\eta} |A - \ln G'(u)| du = \int_0^{u_\eta} ((\lambda_1 - 1)\ln u + \ln B(u) - A) du < \int_0^{u_\eta} ((\lambda_1 - 1)\ln u - A) du + u_\eta \ln B(u_\eta) < \infty,$$

since $\int_0^{u_\eta} \ln u = [u \ln u - u]_0^{u_\eta} = u_\eta \ln u_\eta - u_\eta$ is finite. Thus, $G'(u)$ is sufficiently well-behaved for any λ_1 at the left bound of the interval $[0,1]$. By a similar argument, the right bound yields no problems for any λ_2 either. Thus, $E[|\ln g(z|\theta)|] < \infty$ for all θ.

The other part of Lemma 2.2 of Newey and McFadden (op. cit.) refers to identification of θ_0. As indicated in Section 2, the GTL(α, δ) distribution turns into a uniform distribution for four parameter pairs: $(\alpha, \delta) = (1, 0)$, $(2, 0)$, $(\alpha, 1-\alpha)$ for $\alpha \to \infty$, and $(\alpha, \alpha-1)$ for $\alpha \to \infty$. Given Assumption A.3, only the first of these four parameter pairs is part of Θ and the latter three parameter pairs are immediately ruled out. Thus, θ_0 is always identified. Note that Theorem 2 implies that $(\alpha, \delta) = (1, 0)$ is ruled out both by Assumption A.4: the MLE estimator is not asymptotically normally distributed when $\alpha \geq \delta + \frac{1}{2}$ and/or $\alpha \geq -\delta + \frac{1}{2}$. and in particular for $(\alpha, \delta) = (1, 0)$.

Regarding condition (v), take equations (C.3) and (C.4) one step further by inserting the sup condition:

$$E[\sup_{\theta \in \Theta} |\ln g(z|\theta)|] = \int_{\mathcal{X}} \Big(\int_0^1 \sup_{\theta \in \Theta} |A - \ln G'(u)| du \Big) g_x(x) dx \tag{C.5}$$

with $A = -\ln \sigma + \ln g_x(x)$ and $G'(u) = u^{\lambda_1 - 1} + (1-u)^{\lambda_2 - 1}$. First, note that compactness of Θ implies that A is finite for all values of σ in Θ. Next, we distinguish two situations. First, for values of $u \in (0, 1)$ for which $A - \ln G'(u) > 0$, consider that

$$\begin{aligned} A - \ln G'(u) &= A - \ln(u^{\lambda_1 - 1} + (1-u)^{\lambda_2 - 1}) \\ &< A - \frac{1}{2} \ln u^{\lambda_1 - 1} - \frac{1}{2} \ln(1-u)^{\lambda_2 - 1} \\ &= A + \frac{1}{2}(\lambda_1 - 1)(-\ln u) + \frac{1}{2}(\lambda_2 - 1)(-\ln(1-u)) \end{aligned}$$

since for $u \in (0, 1)$ we have $u^{\lambda_1 - 1} > 0$ and $(1-u)^{\lambda_2 - 1} > 0$, and we know that $-\ln(a+b) < -\ln a$ if $b > 0$. Since $-\ln u > 0$ and $-\ln(1-u) > 0$, it is clear that $\sup_{\theta \in \Theta}(A - \ln G'(u)) = \sup_{\sigma \in \Theta} A$ because the expression obtains its largest value for $\lambda_1 = \lambda_2 = 1$. At the endpoints of the unit interval of u, we have $G'(u) \to \infty$ and thus $-\ln G'(u) \to -\infty$ as $u \downarrow 0$ or $u \uparrow 1$, such that it is impossible that $A - \ln G'(u) > 0$ any longer. Thus, for values of u for which $A - \ln G'(u) > 0$, we have $\sup_{\theta \in \Theta}(A - \ln G'(u))$ is at most equal to $\sup_{\theta \in \Theta} A$, which is finite, and so is the integral of this expression over the appropriate interval of u.

Second, for values of $u \in (0,1)$ for which $A - \ln G'(u) < 0$, we must evaluate $\sup_{\theta\in\Theta}(-A + \ln G'(u))$. Consider the functions $v_1(u) = u^{\lambda_1 - 1}$ and $v_2(u) = (1-u)^{\lambda_2 - 1}$ for $u \in [0,1]$. $v_1(u)$ decreases from ∞ to 1 as u goes from 0 to 1, and $v_2(u)$ rises from 1 to ∞ as u goes from 0 to 1. Define u^* as the value of u where these curves cross each other: $v_1(u^*) = v_2(u^*)$. Note that $u^* = \frac{1}{2}$ if $\lambda_1 = \lambda_2$. Then, for $u < u^*$, it is clear that $v_2(u) < v_1(u))$ and therefore

$$-A + \ln G'(u) < -A + \ln(2v_1(u)) = -A + \ln 2 + \ln u^{\lambda_1 - 1} = -A + \ln 2 + (\lambda_1 - 1)\ln u.$$

Since $\ln u < 0$, the supremum is no larger than the largest value that this expression can take, which is attained for the smallest (most negative) value of λ_1 in Θ, and this largest value approaches ∞ as $u \downarrow 0$. Similarly, for $u > u^*$, we have $v_2(u) > v_1(u))$ and

$$-A + \ln G'(u) < -A + \ln(2v_2(u)) = -A + \ln 2 + \ln(1-u)^{\lambda_2 - 1} = -A + \ln 2 + (\lambda_2 - 1)\ln(1-u),$$

such that the supremum is no larger than the largest value that this expression can take, which is attained for the smallest (most negative) value of λ_2 in Θ, approaching ∞ as $u \uparrow 1$. Since $\int_0^{u^*} \ln u \, du$ and $\int_{u^*}^1 \ln(1-u) du$ are finite (as argued already above in the discussion of condition (iv)), the compactness of Θ guarantees that the integral of $\sup_{\theta\in\Theta}(-A + \ln G'(u))$ is finite.[26]

C.2 Proof of Theorem 2

The proof of the asymptotic normality of the MLE estimator $\hat{\theta}$ relies on verification of the conditions of Theorem 13.2 of Wooldridge (2002, p.395)[27] in relation to, in particular, the

[26] The fact that u^* depends on λ_1 and λ_2 is irrelevant, and if the smallest values of λ_1 and λ_2 in Θ are equal, u^* equals $\frac{1}{2}$. Furthermore, it is possible that $A - \ln G'(u)$ is positive for some range of values of $u \in [0,1]$ and negative for other values (certainly those near the endpoints of the interval). The discussion of each situation implies that the integral of the supremum over each range, and thus the integral over the entire interval $[0,1]$, is finite.

[27] Jeffrey M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, MIT Press, 2002, Cambridge, Massachusetts.

values of the parameters α and δ. These conditions are verified in the following lemmas, the proofs of which are provided in separate sections below.

Lemma 1 *For each* $(x, y) \in \mathcal{X} \times \mathcal{Y}(x, \theta)$, $L(y, x, \cdot)$ *is twice continuously differentiable on int*(Θ) *for all* $\theta \in \Theta$ *except where* $\lambda_1, \lambda_2 = 1, 2, 3$.

Proof Derivatives are provided and examined in Appendix C.2.1. ∎

The dependence of $\mathcal{Y}(x, \theta)$ on θ is usually problematic, because proofs of asymptotic normality rely on the interchange of differentiation and integration, and the dependence of the bounds of y on θ may prevent that. The following lemma addresses this issue:

Lemma 2 *Define* $g_y(y|x_i; \theta)$ *as the conditional density of* y. *Define* $\ell_i(\theta) = \ln g_y(y|x_i; \theta)$. *Define* $s_i(\theta) = \nabla_\theta \ell_i(\theta)$. *Then, for all* $x_i \in \mathcal{X}$ *and* $\theta \in int(\Theta)$ *with* $\lambda_1 < \frac{1}{2}$ *and* $\lambda_2 < \frac{1}{2}$, *we have:*

(i) $\nabla_\theta \Big(\int\limits_{\mathcal{Y}(x,\theta)} g_y(y|x_i; \theta) dy \Big) = \int\limits_{\mathcal{Y}(x,\theta)} \nabla_\theta g_y(y|x_i; \theta) dy$, *and*

(ii) $\nabla_\theta \Big(\int\limits_{\mathcal{Y}(x,\theta)} s_i(\theta) g_y(y|x_i; \theta) dy \Big) = \int\limits_{\mathcal{Y}(x,\theta)} \nabla_\theta s_i(\theta) g_y(y|x_i; \theta) dy$.

Proof See Appendix C.2.2. ∎

Lemma 3 *The elements of* $\nabla^2_\theta L(y, x, \theta)$ *are bounded in absolute value by a function* $b(y, x)$ *with finite expectation for all* $\theta \in \Theta$ *where* (i) $\lambda_1 < \frac{1}{2}$ *and* $\lambda_2 < \frac{1}{2}$, *or* (ii) $\lambda_1 > 2$ *and* $\lambda_2 > 2$ *with* $\lambda_1, \lambda_2 \neq 3$.

Proof See Appendix C.2.1. ∎

Lemma 4 *Define* $A_0 = -E[H(\theta_0)]$ *where* $H(\theta_0) = \nabla_{\theta\theta} \ell(\theta_0)$. *Then* A_0 *is positive definite.*

Proof See Appendix C.2.3. ∎

In the light of these results, the asymptotic normality of the MLE estimator $\hat{\theta}$ of the GTL regression model follows:

$$\sqrt{n}(\hat{\theta} - \theta_0) \stackrel{d}{\to} N(0, A_0^{-1}) \tag{C.6}$$

which implies the statement in Theorem 2:

$$\hat{\theta} \stackrel{a}{\sim} N(\theta_0, V_0) \tag{C.7}$$

where V_0 is estimated as $V(\hat{\theta}) = -\Big(\sum_{i=1}^{n} \nabla_{\theta\theta}\ell_i(\hat{\theta})\Big)^{-1}$

Note that, relative to the assumptions adopted for Theorem 1, Assumption A.5 is added because the population parameters must not lie on the boundary of Θ, and Assumption A.3 is replaced with the more restrictive Assumption A.4 because of Lemma 2.

C.2.1 Proof of Lemmas 1 and 3

In this appendix, we first examine derivatives of the function $G(u)$ and then list the derivatives of the log-likelihood function. This leads directly to the proof of Lemma 1 and sets up the examination of the second order derivatives for Lemma 3.

We have $G(u) = \frac{1}{\lambda_1}(u^{\lambda_1} - 1) - \frac{1}{\lambda_2}((1-u)^{\lambda_2} - 1)$, where $\lambda_1 = \alpha - \delta$ and $\lambda_2 = \alpha + \delta$. In the following, apostrophes refer to derivatives with respect to u, and subscripts indicate derivatives with respect to parameters.

$$G'(u) = u^{\lambda_1 - 1} + (1-u)^{\lambda_2 - 1} \tag{C.8}$$

$$G''(u) = (\lambda_1 - 1)u^{\lambda_1 - 2} - (\lambda_2 - 1)(1-u)^{\lambda_2 - 2} \tag{C.9}$$

$$G'''(u) = (\lambda_1 - 1)(\lambda_1 - 2)u^{\lambda_1 - 3} + (\lambda_2 - 1)(\lambda_2 - 2)(1-u)^{\lambda_2 - 3} \tag{C.10}$$

$$G_\alpha(u) = \frac{1}{\lambda_1}u^{\lambda_1}\ln u - \frac{1}{\lambda_1^2}(u^{\lambda_1} - 1) - \frac{1}{\lambda_2}(1-u)^{\lambda_1}\ln(1-u) + \frac{1}{\lambda_2^2}((1-u)^{\lambda_2} - 1) \tag{C.11}$$

$$G'_\alpha(u) = u^{\lambda_1 - 1}\ln u + (1-u)^{\lambda_2 - 1}\ln(1-u) \tag{C.12}$$

$$
\begin{aligned}
G''_{\alpha}(u) &= u^{\lambda_1-2}(1+(\lambda_1-1)\ln u)-(1-u)^{\lambda_2-2}(1+(\lambda_2-1)\ln(1-u)) && \text{(C.13)}\\
G_{\alpha\alpha}(u) &= \frac{1}{\lambda_1}u^{\lambda_1}\ln u\big(\ln u-\frac{2}{\lambda_1}\big)+\frac{2}{\lambda_1^3}(u^{\lambda_1}-1)\\
&\quad -\frac{1}{\lambda_2}(1-u)^{\lambda_2}\ln(1-u)\big(\ln(1-u)-\frac{2}{\lambda_2}\big)-\frac{2}{\lambda_2^3}((1-u)^{\lambda_2}-1) && \text{(C.14)}\\
G'_{\alpha\alpha}(u) &= u^{\lambda_1-1}(\ln u)^2+(1-u)^{\lambda_2-1}(\ln(1-u))^2 && \text{(C.15)}\\
G_{\delta}(u) &= -\frac{1}{\lambda_1}u^{\lambda_1}\ln u+\frac{1}{\lambda_1^2}(u^{\lambda_1}-1)-\frac{1}{\lambda_2}(1-u)^{\lambda_1}\ln(1-u)+\frac{1}{\lambda_2^2}((1-u)^{\lambda_2}-1) && \text{(C.16)}\\
G'_{\delta}(u) &= -u^{\lambda_1-1}\ln u+(1-u)^{\lambda_2-1}\ln(1-u) && \text{(C.17)}\\
G''_{\delta}(u) &= -u^{\lambda_1-2}(1+(\lambda_1-1)\ln u)-(1-u)^{\lambda_2-2}(1+(\lambda_2-1)\ln(1-u)) && \text{(C.18)}\\
G_{\delta\delta}(u) &= \frac{1}{\lambda_1}u^{\lambda_1}\ln u\big(\ln u-\frac{2}{\lambda_1}\big)+\frac{2}{\lambda_1^3}(u^{\lambda_1}-1)\\
&\quad -\frac{1}{\lambda_2}(1-u)^{\lambda_2}\ln(1-u)\big(\ln(1-u)-\frac{2}{\lambda_2}\big)-\frac{2}{\lambda_2^3}((1-u)^{\lambda_2}-1) && \text{(C.19)}\\
G'_{\delta\delta}(u) &= u^{\lambda_1-1}(\ln u)^2+(1-u)^{\lambda_2-1}(\ln(1-u))^2 && \text{(C.20)}\\
G_{\alpha\delta}(u) &= -\frac{1}{\lambda_1}u^{\lambda_1}\ln u\big(\ln u-\frac{2}{\lambda_1}\big)-\frac{2}{\lambda_1^3}(u^{\lambda_1}-1)\\
&\quad -\frac{1}{\lambda_2}(1-u)^{\lambda_2}\ln(1-u)\big(\ln(1-u)-\frac{2}{\lambda_2}\big)-\frac{2}{\lambda_2^3}((1-u)^{\lambda_2}-1) && \text{(C.21)}\\
G'_{\alpha\delta}(u) &= -u^{\lambda_1-1}(\ln u)^2+(1-u)^{\lambda_2-1}(\ln(1-u))^2 && \text{(C.22)}
\end{aligned}
$$

λ_1 and λ_2 appear in the denominator in G, G_α, G_δ, $G_{\alpha\alpha}$, $G_{\alpha\delta}$ and $G_{\delta\delta}$. Nevertheless, these functions are well-defined when $\lambda_1 \to 0$ or $\lambda_2 \to 0$. Because of symmetry, it is necessary to show this only for $\lambda_1 \to 0$. In $G(u)$, we have

$$\lim_{\lambda_1\to 0}\frac{u^{\lambda_1}-1}{\lambda_1}=\lim_{\lambda_1\to 0}\frac{u^{\lambda_1}\ln u}{1}=\ln u. \tag{C.23}$$

by L'Hôpital's Rule. Similarly, in $G_\alpha(u)$ and $G_\delta(u)$, we find

$$\lim_{\lambda_1\to 0}\frac{\lambda_1 u^{\lambda_1}\ln u-(u^{\lambda_1}-1)}{\lambda_1^2}=\lim_{\lambda_1\to 0}\frac{u^{\lambda_1}\ln u+\lambda_1 u^{\lambda_1}\ln u-u^{\lambda_1}\ln u}{2\lambda_1}=\frac{1}{2}(\ln u)^2, \tag{C.24}$$

and in regard to $G_{\alpha\alpha}(u)$, $G_{\alpha\delta}(u)$ and $G_{\delta\delta}(u)$, we find

$$\lim_{\lambda_1\to 0}\frac{\lambda_1^2 u^{\lambda_1}(\ln u)^2-2\lambda_1 u^{\lambda_1}\ln u+2(u^{\lambda_1}-1)}{\lambda_1^3}=\lim_{\lambda_1\to 0}\frac{\lambda_1^2 u^{\lambda_1}(\ln u)^3}{3\lambda_1^2}=\frac{1}{3}(\ln u)^3. \tag{C.25}$$

A more serious issue exists at $\lambda_1 = 1$. For example, as $u \downarrow 0$, $G'(u) \to 1$ when $\lambda_1 > 1$; $G'(u) \to 2$ when $\lambda_1 = 1$; and $G'(u) \to \infty$ when $\lambda_1 < 1$. Thus, $G'(0)$ is not a continuous function of λ_1. Other derivatives (G'_α and others) are similarly impacted at $\lambda_1 = 1$. Moreover, $G''(0)$ is discontinuous in λ_1 at $\lambda_1 = 2$ and $G'''(0)$ is discontinuous in λ_1 at $\lambda_1 = 3$. Similar discontinuities exist at $u = 1$ for $\lambda_2 = 1, 2, 3$.

Next, we present the first and second order derivatives of the log-likelihood function: $L = \sum_{i=1}^{n} \ell_i(\theta)$. For ease of notation, we drop the argument u_i from the G-function and its derivatives. The first order derivatives of $\ell_i(\theta)$ are:

$$\nabla_\beta \ell_i = \frac{1}{\sigma}\frac{G''}{(G')^2}x_i \tag{C.26}$$

$$\nabla_\sigma \ell_i = -\frac{1}{\sigma} + \frac{1}{\sigma^2}\frac{G''}{(G')^2}(y_i - x_i\beta) \tag{C.27}$$

$$\nabla_\alpha \ell_i = \frac{G''G_\alpha}{(G')^2} - \frac{G'_\alpha}{G'} \tag{C.28}$$

$$\nabla_\delta \ell_i = \frac{G''G_\delta}{(G')^2} - \frac{G'_\delta}{G'} \tag{C.29}$$

The second order derivatives are:

$$\nabla_{\beta\beta} \ell_i = -\frac{1}{\sigma^2}\frac{G'G''' - 2(G'')^2}{(G')^4}x_i x_i' \tag{C.30}$$

$$\nabla_{\beta\sigma} \ell_i = -\frac{1}{\sigma}\nabla_\beta \ell_i - \frac{1}{\sigma^3}\frac{G'G''' - 2(G'')^2}{(G')^4}(y_i - x_i\beta)x_i \tag{C.31}$$

$$\nabla_{\beta\alpha} \ell_i = -\frac{1}{\sigma}\frac{G'G''' - 2(G'')^2}{(G')^4}G_\alpha x_i + \frac{G'G''_\alpha - 2G''G'_\alpha}{(G')^3}x_i \tag{C.32}$$

$$\nabla_{\beta\delta} \ell_i = -\frac{1}{\sigma}\frac{G'G''' - 2(G'')^2}{(G')^4}G_\delta x_i + \frac{G'G''_\delta - 2G''G'_\delta}{(G')^3}x_i \tag{C.33}$$

$$\nabla_{\sigma\sigma} \ell_i = -\frac{2}{\sigma}\nabla_\sigma \ell_i - \frac{1}{\sigma^2} - \frac{1}{\sigma^2}\frac{G'G''' - 2(G'')^2}{(G')^4}(y_i - x_i\beta)^2 \tag{C.34}$$

$$\nabla_{\sigma\alpha} \ell_i = -\frac{1}{\sigma^2}\frac{G'G''' - 2(G'')^2}{(G')^4}G_\alpha(y_i - x_i\beta) + \frac{1}{\sigma^2}\frac{G'G''_\alpha - 2G''G'_\alpha}{(G')^3}(y_i - x_i\beta) \tag{C.35}$$

$$\nabla_{\sigma\delta} \ell_i = -\frac{1}{\sigma^2}\frac{G'G''' - 2(G'')^2}{(G')^4}G_\alpha(y_i - x_i\beta) + \frac{1}{\sigma^2}\frac{G'G''_\delta - 2G''G'_\delta}{(G')^3}(y_i - x_i\beta) \tag{C.36}$$

$$\nabla_{\alpha\alpha} \ell_i = -\frac{G'G''' - 2(G'')^2}{(G')^4}G_\alpha^2 + 2\frac{G'G''_\alpha - 2G''G'_\alpha}{(G')^3}G_\alpha + \frac{G''G_{\alpha\alpha} - G'G'_{\alpha\alpha} + (G'_\alpha)^2}{(G')^2} \tag{C.37}$$

(C.38)

$$
\begin{aligned}
\nabla_{\alpha\delta}\ell_i &= -\frac{G'G''' - 2(G'')^2}{(G')^4}G_\alpha G_\delta + \frac{G'G''_\alpha - 2G''G'_\alpha}{(G')^3}G_\delta \\
&\quad + \frac{G'G''_\delta - 2G''G'_\delta}{(G')^3}G_\alpha + \frac{G''G_{\alpha\delta} - G'G'_{\alpha\delta} + G'_\alpha G'_\delta}{(G')^2} \qquad \text{(C.39)} \\
\nabla_{\delta\delta}\ell_i &= -\frac{G'G''' - 2(G'')^2}{(G')^4}G_\delta^2 + 2\frac{G'G''_\delta - 2G''G'_\delta}{(G')^3}G_\delta + \frac{G''G_{\delta\delta} - G'G'_{\delta\delta} + (G'_\delta)^2}{(G')^2} \qquad \text{(C.40)}
\end{aligned}
$$

As for Lemma 1, the properties of G and all its derivatives imply continuity of all second order derivatives in θ except at $\lambda_1, \lambda_2 = 1, 2, 3$.

As for Lemma 3, we replace $(y_i - x_i'\beta)/\sigma$ by $G(u_i)$ in the first and second order derivatives. This yields expressions in u_i, multiplied in some cases with x_i or $x_i x_i'$. Thus, relative to x, Lemma 3 requires that $\mathrm{E}[x]$ and $\mathrm{E}[xx']$ are finite. Taking expectations with respect to y_i turns into an integration over u_i, which has a uniform density, which simplifies the analysis considerably. Thus, we examine whether $\nabla_{\theta\theta}$ is integrable for $u \in [0,1]$. Since G and its derivatives are well-behaved for any $u \in [u_\eta, 1-u_\eta]$ for a small u_η, we must examine the behavior of $\nabla_{\theta\theta}$ over the interval $[0, u_\eta]$; the argument for the interval $[1-u_\eta, 1]$ is similar.

In $\nabla_{\theta\theta}$, G' appears in the denominator frequently. Since $G' \to \infty$ for $u \downarrow 0$ when $\lambda_1 < 1$, we consider the cases of $\lambda_1 < 1$, $\lambda_1 = 1$ and $\lambda_1 > 1$ separately. Furthermore, note that as $u \downarrow 0$, terms with $1-u$ in equations (C.26) to (C.40) contribute at most a constant to the limit of the expression for any finite λ_2.

The case of $\lambda_1 < 1$.

We start with $\nabla_{\beta\beta}\ell_i$, dividing the ratio in equation (C.30) into two parts:

$$
\begin{aligned}
\frac{G'''}{(G')^3} &\to \frac{(\lambda_1 - 1)(\lambda - 2)u^{\lambda_1 - 3} + (\lambda_2 - 1)(\lambda_2 - 2)}{(u^{\lambda_1 - 1} + 1)^3} \\
&= \;= \frac{u^{\lambda_1 - 3}\big((\lambda_1 - 1)(\lambda - 2) + (\lambda_2 - 1)(\lambda_2 - 2)u^{3-\lambda_1}\big)}{u^{3\lambda_1 - 3}(1 + u^{1-\lambda_1})^3} \\
&\to \frac{u^{\lambda_1 - 3}(\lambda_1 - 1)(\lambda - 2)}{u^{3\lambda_1 - 3}} = u^{-2\lambda_1}(\lambda_1 - 1)(\lambda - 2) \qquad \text{(C.41)}
\end{aligned}
$$

$$
\begin{aligned}
\frac{-2(G'')^2}{(G')^4} &\to -2\frac{((\lambda_1-1)u^{\lambda_1-2}-(\lambda_2-1))^2}{(u^{\lambda_1-1}+1)^4} \\
&= -2\frac{u^{2\lambda_1-4}((\lambda_1-1)-(\lambda_2-1)u^{2-\lambda_1})^2}{u^{4\lambda_1-4}(1+u^{1-\lambda_1})^4} \\
&\to -2\frac{(\lambda_1-1)^2u^{2\lambda_1-4}}{u^{4\lambda_1-4}} = -2(\lambda_1-1)^2u^{-2\lambda_1} \qquad \text{(C.42)}
\end{aligned}
$$

since $u^{m-\lambda_1} \to 0$ for $m = 1, 2, 3$. Combining these two terms yields

$$
\frac{G'G''' - 2(G'')^2}{(G')^4} \to -\lambda_1(\lambda_1-1)u^{-2\lambda_1} \qquad \text{(C.43)}
$$

which has a finite integral on $[0, u_\eta]$ if $-2\lambda_1 \geq -1$ or $\lambda_1 \leq 1/2$.[28]

For the other terms, we follow the same strategy. For $\nabla_{\beta\sigma}\ell_i$, we have:

$$
\begin{aligned}
\frac{G''}{G'} &\to \frac{(\lambda_1-1)u^{\lambda_1-2}-(\lambda_2-1)}{(u^{\lambda_1-1}+1)^2} \to (\lambda_1)u^{-\lambda_1} \qquad \text{(C.44)} \\
\frac{G'G'''-2(G'')^2}{(G')^4}G &\to -\lambda_1(\lambda_1-1)u^{-2\lambda_1}\frac{u^{\lambda_1}-1}{\lambda_1} = -(\lambda_1-1)(u^{-\lambda_1}-u^{-2\lambda_1}) \qquad \text{(C.45)}
\end{aligned}
$$

which, altogether, has a finite integral if $\lambda_1 \leq 1/2$.

For $\nabla_{\beta\alpha}\ell_i$, we have:

$$
\begin{aligned}
\frac{G'G'''-2(G'')^2}{(G')^4}G_\alpha &\to -\lambda_1(\lambda_1-1)u^{-2\lambda_1}\big(\frac{1}{\lambda_1}u^{\lambda_1}\ln u - \frac{1}{\lambda_1^2}(u^{\lambda_1}-1)\big) \\
&= (\lambda_1-1)u^{-\lambda_1}\ln u - \frac{\lambda_1-1}{\lambda_1}(u^{-\lambda_1}-u^{-2\lambda_1}) \qquad \text{(C.46)} \\
\frac{G'G''_\alpha - 2G''G'_\alpha}{(G')^3} &\to u^{-\lambda_1} - u^{-2\lambda_1+2} - (\lambda_1-1)u^{-\lambda_1}\ln u \qquad \text{(C.47)}
\end{aligned}
$$

The first term in (C.46) cancels against the last term of (C.46). $u^{-\lambda_1}$, $u^{-2\lambda_1}$, and $u^{2-2\lambda_1}$ all have a finite integral if $\lambda_1 < 1/2$. The result for $\nabla_{\beta\delta}\ell_i$ is parallel to this case.

[28] Lemma 3 requires $\nabla_{\beta\beta}\ell_i(\theta)$ to be bounded in absolute value by a function $b(y, x)$. As equations (C.41) and (C.42), the expression in equation (C.43) omits multiplicative functions that converge to 1 as $u \downarrow 0$. Thus, $b(y, x)$ is found by multiplying (C.43) with the largest value that such functions take over the interval $[0, u_\eta]$, which is finite.

For $\nabla_{\sigma\sigma}\ell_i$, we have:

$$\begin{aligned}\frac{G''}{(G')^2}G &\to \Big(\frac{(\lambda_1-1)u^{-\lambda_1-2}-(\lambda_2-1)}{(u^{\lambda_1-1}+1)^2}\Big)\Big(\frac{u^{\lambda_1}-1}{\lambda_1}\Big)\\ &\to \frac{1}{\lambda_1}\big((\lambda_1-1)(1-u^{-\lambda_1})-(\lambda_2-1)(u^{2-\lambda_1}-u^{2-2\lambda_1})\big) \qquad \text{(C.48)}\end{aligned}$$

$$\begin{aligned}\frac{G'G'''-2(G'')^2}{(G')^4}G^2 &\to -\lambda_1(\lambda_1-1)u^{-2\lambda_1}\Big(\frac{u^{\lambda_1}-1}{\lambda_1}\Big)^2\\ &= -\frac{(\lambda_1-1)}{\lambda_1}(1-2u^{-\lambda_1}+u^{-2\lambda_1}) \qquad \text{(C.49)}\end{aligned}$$

In (C.48) and (C.49), $u^{-\lambda_1}$, $u^{2-\lambda_1}$, $u^{2-2\lambda_1}$ and $u^{-2\lambda_1}$ all have a finite integral if $\lambda_1 < 1/2$.

For $\nabla_{\sigma\alpha}\ell_i$, we have with the aid of (C.43) and (C.47):

$$\nabla_{\sigma\alpha}\ell_i \to \frac{1}{\lambda_1^2}\big(1-\lambda_1(u^{2-\lambda_1}-u^{2-2\lambda_1})+(\lambda_1-2)u^{-\lambda_1}-(\lambda_1-1)u^{-2\lambda_1}\big) \qquad \text{(C.50)}$$

Once again, finite integrability depends on $u^{-2\lambda_1}$, such that λ_1 must be less than 1/2. The result for $\nabla_{\sigma\delta}\ell_i$ is parallel to this case.

With respect to $\nabla_{\alpha\alpha}\ell_i$, we have three components:

$$\begin{aligned}-\frac{G'G'''-2(G'')^2}{(G')^4}G_\alpha^2 &\to \frac{\lambda_1-1}{\lambda_1}(\ln u)^2-\frac{2(\lambda_1-1)}{\lambda_1^2}\ln u+\frac{2(\lambda_1-1)}{\lambda_1^2}u^{-\lambda_1}\ln u\\ &\quad +\frac{\lambda_1-1}{\lambda_1^3}(1-2u^{-\lambda_1}+u^{-2\lambda_1}) \qquad \text{(C.51)}\end{aligned}$$

$$\begin{aligned}2\frac{G'G''_\alpha-2G''G'_\alpha}{(G')^3}G_\alpha &\to -2\frac{\lambda_1-1}{\lambda_1}(\ln u)^2+\frac{2(\lambda_1-1)}{\lambda_1^2}\ln u-\frac{2(\lambda_1-1)}{\lambda_1^2}u^{-\lambda_1}\ln u\\ &\quad +\frac{2}{\lambda_1^2}\big(\lambda_1\ln u-\lambda_1 u^{2-\lambda_1}\ln u\big)\\ &\quad +\frac{2}{\lambda_1^2}\big(u^{-\lambda_1}-1+u^{2-\lambda_1}-u^{2-2\lambda_1}\big) \qquad \text{(C.52)}\\ \frac{G''G_{\alpha\alpha}-G'G'_{\alpha\alpha}+(G'_\alpha)^2}{(G')^2} &\to \frac{\lambda_1-1}{\lambda_1}(\ln u)^2-2\frac{\lambda_1-1}{\lambda_1^2}\ln u+2\frac{\lambda_1-1}{\lambda_1^3}(1-u^{-\lambda_1}) \qquad \text{(C.53)}\end{aligned}$$

The first line of (C.51) and (C.52) and the first term of (C.53) cancel against each other. On the second line of (C.52), $u^{2-\lambda_1}\ln u$ vanishes more rapidly as $u \downarrow 0$ than $\ln u$, which itself has a finite integral. The remainder has a finite integral if $\lambda_1 \leq 1/2$. The results for $\nabla_{\alpha\delta}\ell_i$ and $\nabla_{\delta\delta}\ell_i$ are parallel to this case.

The case of $\lambda_1 \geq 1$.

From Lemma 1, it is clear that $\nabla_{\theta\theta}\ell_i$ has discontinuities at $\lambda_1 = 1, 2, 3$. With methods similar to above, we have in regard to $\nabla_{\beta\beta}\ell_i$, as $u \downarrow 0$:

$$1 < \lambda_1 < 2 \quad : \quad \frac{G'G''' - 2(G'')^2}{(G')^4} \to (\lambda_1 - 1)(\lambda_1 - 2)u^{\lambda_1 - 3} - 2(\lambda_1 - 1)^2 u^{2\lambda_1 - 4} \tag{C.54}$$

$$2 < \lambda_1 < 3 \quad : \quad \frac{G'G''' - 2(G'')^2}{(G')^4} \to (\lambda_1 - 1)(\lambda_1 - 2)u^{\lambda_1 - 3} - 2(\lambda_2 - 1)^2 \tag{C.55}$$

$$3 < \lambda_1 < \infty \quad : \quad \frac{G'G''' - 2(G'')^2}{(G')^4} \to -\lambda_2(\lambda_2 - 1) \tag{C.56}$$

In (C.55), the expression has a finite integral if $\lambda_1 - 3 > -1$ or $\lambda_1 > 2$, which is compatible with the range of λ_1 considered in the derivation of (C.55). In (C.54) however, for the expression to have a finite integral over $[0, u_\eta]$, λ_1 must exceed 2, which is beyond the range considered. With similar derivations, it can be shown that every component of $\nabla_{\theta\theta}\ell_i$ has a finite integral only if $\lambda_1 > 2$ and $\lambda_2 > 2$ with a discontinuity at $\lambda_1 = \lambda_2 = 3$.

C.2.2 Proof of Lemma 2

For $\lambda_1 > 0$, the lower bound of ϵ is $\underline{\epsilon} = \frac{-1}{\lambda_1}$, and for $\lambda_2 > 0$, the upper bound of ϵ is $\bar{\epsilon} = \frac{1}{\lambda_2}$. Thus the bounds on y are $\underline{y} = x'\beta + \sigma\underline{\epsilon}$ and $\bar{y} = x'\beta + \sigma\bar{\epsilon}$, respectively. To prove Lemma 2, we abbreviate notation slightly and examine

$$\nabla_\theta\Big(\int_{\underline{y}}^{\bar{y}} g_y(y|\theta)dy\Big) = \int_{\underline{y}}^{\bar{y}} \nabla_\theta g_y(y|\theta) - g_y(\underline{y}|\theta)\underline{y}_\theta + g_y(\bar{y}|\theta)\bar{y}_\theta. \tag{C.57}$$

Thus, interchanging differentiation and integration is permissible as long as $g_y(\underline{y}|\theta) = g_y(\bar{y}|\theta) = 0$. We have $g_y(y|\theta) = [G'(u)]^{-1}/\sigma$ with $u = G^{-1}(\epsilon)$ with $\epsilon = \frac{1}{\sigma}(y - x'\beta)$.

Therefore, as $y \to \underline{y}$, we have $\epsilon \to \underline{\epsilon}$ and $u \to 0$, and thus for $\lambda_1 < 1$, $G'(u) \to \infty$ and $g_y(\underline{y}|\theta) \to 0$. Similarly, for $\lambda_2 < 1$, $g_y(\bar{y}|\theta) \to 0$.

Since $\int s(\theta) g_y(y|\theta) dy = \int \nabla_\theta g_y(y|\theta) dy$, part (*ii*) of Lemma 2 concerns the derivative of the first term in equation (C.57):

$$\nabla_\theta \Big(\int_{\underline{y}}^{\bar{y}} \nabla_\theta g_y(y|\theta) dy \Big) = \int_{\underline{y}}^{\bar{y}} \nabla_{\theta\theta} g_y(y|\theta) - \nabla_\theta g_y(\underline{y}|\theta) \underline{y}_\theta + \nabla_\theta g_y(\bar{y}|\theta) \bar{y}_\theta. \tag{C.58}$$

$\nabla_\theta g_y(y|\theta)$ approaches the following functions as $u \downarrow 1$:

$$\nabla_\beta g_y = \frac{G''(u) x}{\sigma^2 (G'(u))^3} \to \big((\lambda_1 - 1) u^{1-2\lambda_1} - (\lambda_2 - 1) u^{3-3\lambda_1} \big) \tag{C.59}$$

$$\begin{aligned} \nabla_\sigma g_y &= -\frac{1}{\sigma^2 G'(u)} - \frac{G''(u)(y - x'\beta)}{\sigma^3 (G'(u))^3} \\ &\to -\sigma^{-2} \big(u^{1-\lambda_1} - \lambda_1^{-1} ((\lambda_1 - 1) u^{1-2\lambda_1} - (\lambda_2 - 1) u^{3-3\lambda_1}) \big) \end{aligned} \tag{C.60}$$

$$\begin{aligned} \nabla_\alpha g_y &= \frac{G''(u) G_\alpha}{\sigma (G'(u))^3} \\ &\to \frac{\lambda_1 - 1}{\lambda_1} u^{1-\lambda_1} \ln u - \frac{\lambda_2 - 1}{\lambda_1} u^{3-2\lambda_1} \ln u \\ &\quad - \frac{\lambda_1 - 1}{\lambda_1^2} (u^{1-\lambda_1} - u^{1-2\lambda_1}) + \frac{\lambda_2 - 1}{\lambda_1^2} (u^{3-2\lambda_1} - u^{3-3\lambda_1}) \end{aligned} \tag{C.61}$$

$$\begin{aligned} \nabla_\delta g_y &= \frac{G''(u) G_\delta}{\sigma (G'(u))^3} \\ &\to -\frac{\lambda_1 - 1}{\lambda_1} u^{1-\lambda_1} \ln u + \frac{\lambda_2 - 1}{\lambda_1} u^{3-2\lambda_1} \ln u \\ &\quad + \frac{\lambda_1 - 1}{\lambda_1^2} (u^{1-\lambda_1} - u^{1-2\lambda_1}) - \frac{\lambda_2 - 1}{\lambda_1^2} (u^{3-2\lambda_1} - u^{3-3\lambda_1}) \end{aligned} \tag{C.62}$$

As $u \downarrow 1$, each of these functions goes to 0 if and only if $\lambda < 1/2$. By a similar argument, $\lambda_2 < 1/2$ is necessary for $\nabla_\theta g_y(y|\theta)$ to go to 0 as $u \to 1$.[29]

[29] The derivative of the second and third terms of (C.57) equals 0 as well when $\lambda_1 < 1/2$ and $\lambda_2 < 1/2$. That is, for both the first and second order derivatives of $\int g_y(y|\theta) dy$, differentiation and integration is interchangeable.

C.2.3 Proof of Lemma 4

$E[H(\theta)]$ may be written as integrals of expressions found in equations (C.30)-(C.40) (multiplied by $g_x(x)$) over u and x. As functions of u, integration of equations (C.30)-(C.40) has no analytical solution. Thus, to provide evidence that $A_0 = -E[H(\theta)]$ is positive definite, we resort to numerical integration by simulation. This evidence is tied to a model that must therefore be specified in a generic fashion.

We specify x as a vector of three elements: $x_1 = 1$ to allow for an intercept, x_2 is standard normal, and x_3 is a standardized $\chi^2(5)$ variable that introduces some skewness into the explanatory variables. β is set to $(1,1,1)'$. Note that $x'\beta$ generates a location shift only. λ_1 and λ_2 vary from -3 to 0.49, the former value out of concern for numerical over- and underflow in the computations and the latter value in deference to the upper bound of the feasible parameter space. σ cannot really be held constant because variations in α and δ generate ϵs of a different scale such that models with a high λ_1 and λ_2 have a much higher signal-to-noise ratio. Since $x'\beta$ has a variance of 2 by design, we choose $\sigma = (2IQR_{N(0,1)}/IQR_{GTL(\alpha,\delta)})^{0.5}$. Thus, if the GTL density is close to the standard normal density, the variance of ϵ is close to 2 and the signal-to-noise ratio is about 1. σ declines as λ_1 and λ_2 fall and the interquartile range of ϵ rises.[30] For other values of λ_1 and λ_2, the signal-to-noise ratio probably differs from 1 but should be in the neighborhood of it.

For the given range of λ_1 and λ_2, Table C.3 reports the smallest eigenvalues of the simulated matrix $\hat{A}_0$, which itself is computed on the basis of 10,000 replications. The smallest eigenvalue is always solidly positive: at least for the model that is examined here, A_0 appears to be positive definite.

[30] Recall that $\mathrm{Var}(\epsilon)$ does not exist if $\min(\lambda_1, \lambda_2)$ is less than $-1/2$.

Table C.3: Simulated values of the smallest eigenvalue of $\hat{A}_0$

	λ_2 for $N = 100$					λ_2 for $N = 1000$				
λ_1	-3	-2	-1	0	0.49	-3	-2	-1	0	0.49
-3	6.56	8.01	9.46	11.14	11.93	65.29	79.81	94.12	110.77	117.87
-2	8.04	10.69	13.10	16.27	17.91	79.82	106.30	130.31	161.48	175.66
-1	9.48	13.13	15.17	17.16	18.07	94.15	130.36	152.74	170.78	178.06
0	11.16	16.30	17.17	24.22	23.79	110.84	161.62	170.99	242.28	234.88
0.49	11.94	17.92	18.09	23.69	37.17	118.08	176.12	178.49	235.53	352.26

D The LM test for normality

As discussed in Section 2 of the paper, the normal distribution is closely similar to the GTL$(0.1436, 0)$ distribution. Thus, on the basis of the GTL regression model, an assumption of normal-distributed disturbances may be tested with a Wald test that uses $(\hat{\alpha}, \hat{\delta})$ and $V(\hat{\theta})$; a likelihood ratio (LR) test; a Lagrange multiplier (LM) test; and a Vuong test that examines the observations' contributions to the likelihood function under GTL$(\hat{\alpha}, \hat{\delta})$ and under normality. Tests of other distributions (e.g., t) proceed in similar ways. Strictly speaking, the Null hypothesis of the Wald, LM and LR tests is the GTL$(0.1436, 0)$ regression model.[31]

However, a model with GTL$(0.1436, 0)$ disturbances deviates trivially little from a model with true normal distributed disturbances. In the context of the simulations with the baseline design discussed in the Section 4, results show the power of the LM test to be only 7.4% in a large sample ($n = 5000$). However, the LM test does have power against even minor formal deviations from GTL$(0.1436, 0)$: if disturbances are generated with slightly thicker tails $(\alpha, \delta) = (0.10, 0)$ (which closely approximates a Student's $t(30)$ distribution), power equals 84.9%; and similarly for a slightly thinner tail $(\alpha, \delta) = (0.20, 0)$ or slight skewness $(0.1436, 0.02)$, power is 99.5% and 77.0% respectively. Thus, in a regression model, GTL$(0.1436, 0)$ disturbances are virtually indistinguishable from strictly normal disturbances.

The LM test requires estimation only of the restricted model. Thus, could the LM test be considered as a post-estimation diagnostic test after OLS regression for the existence of GTL disturbances? In this light, let us examine the LM test. Let the restricted estimator of θ be denoted as $\breve{\theta} = (\breve{\beta}', \breve{\sigma}, 0.1436, 0)'$; let $s(\breve{\theta})$ be the score vector of the log-likelihood function (see Lemma 2 in Web Appendix C.2); and let $V(\breve{\theta})$ be the negative of the inverted hessian evaluated at $\breve{\theta}$ (see equation (8)). Then approximate normality may

[31]The Vuong test compares GTL$(\hat{\alpha}, \hat{\delta})$ with strict normality and thus involves non-nested models. A Jarque-Bera test of normality of OLS residuals is also easily implemented but examines only the third and fourth moments of the residuals and thus does not point specifically to GLT disturbances if normality is rejected.

be tested with:

$$LM = s(\breve{\theta})' \, V(\breve{\theta})^{-1} \, s(\breve{\theta}) \;\overset{a}{\sim}\; \chi^2(2) \tag{D.1}$$

Conceptually, the computation of LM makes use of the residuals of the GTL$(0.1436, 0)$ regression model. Suppose OLS residuals are used instead: estimate the linear regression model with OLS, deduce the proper estimate of σ from the root mean squared error of the residuals, evaluate $s(\breve{\theta})$ and $V(\breve{\theta})$, and compute LM. However, it might happen that the largest OLS residuals fall outside the support of ϵ, which for $(\alpha, \delta) = (0.1436, 0)$ equals $(-6.694, 6.694)$ if ϵ is GTL in canonical form or $(-4.819, 4.819)$ if ϵ is a standardized GTL disturbance (i.e., $\tilde{\epsilon}$). Since s and V are derived in terms of $u_i = G^{-1}(\epsilon_i) \in [0, 1]$ (see Appendix C.2.1), one might set $u_i = 0$ if $\tilde{\epsilon}_i < -4.819$ and $u_i = 1$ if $\tilde{\epsilon}_i > 4.819$ and proceed with the computation, which then typically yields a large LM value. Nevertheless, if ϵ is truly normally distributed, the chance that ϵ exceeds the GTL support is approximately 1.44×10^{-6}. Thus, the encounter of any support violations may well indicate that ϵ is not normally distributed anyway.[32] Thus, the LM test may be implemented as a post-estimation diagnostic test of the OLS model, as a way to suggest more efficient alternatives of estimating the regression slopes.

[32]For example, the chance of any violations in a sample of 10000 observations is 0.0143.

E Monte Carlo analysis: Further results

E.1 Monte Carlo results: Bias and standard errors of OLS and GTL Estimators of β

The first three columns of Table E.1 shows the bias of the OLS and GTL estimators of $beta_2$ and β_3, which are slope parameters, and β_1 which is the intercept of the regression equation. Columns 4 through 6 report the Monte Carlo standard deviation of these estimators, and columns 7 through 9 report the average asymptotic standard errors based on the standard formula for OLS estimators or the inverse Hessian for the GTL estimator.

The table offers several conclusions. (*i*) As long as the second moment of the GTL disturbances is finite (i.e., in our experiments, for $\alpha \geq -0.33$), the asymptotic OLS standard errors accurately measure the Monte Carlo standard deviation, but when the GTL disturbances lack a finite standard deviation (for $\alpha \leq -0.67$), this correspondence breaks down. (*ii*) OLS bias of $\hat{\beta}_2$ and $\hat{\beta}_3$ is minimal as long as GTL disturbances have finite moments, but the OLS estimator becomes wild when the GTL disturbances lack a finite standard deviation. (*iii*) As expected, the OLS estimator of the intercept is biased when GTL disturbances are skewed, since the GTL disturbance is not centered at its (constant but possibly non-zero) mean. (*iv*) The GTL estimator has minimal bias, even in small samples. (*v*) In both small and large samples, the analytical (asymptotic) standard errors closely approximate the Monte Carlo standard deviation.

E.2 Monte Carlo results: Estimators of α, δ, and σ

For the various experiments that are discussed in Section 4 of the paper, Table E.2 examines the biases and standard deviations of the GTL estimator of the parameters of the GTL distribution that generates the disturbance, namely the shape parameters α and δ as well as the scaling parameter σ that is determined by linking the length of the quantile range of the normal and GTL distributions as described above. σ varies with α and δ as indicated in the table. Table E.2 indicates a small bias for $\alpha = 0.33$,

Table E.1: Bias and precision of $\hat{\beta}_{OLS}$ and $\hat{\beta}_{GTL}$ for GTL-generated data

DGP		Bias			Monte Carlo standard deviation			Analytical standard error		
α	δ	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1
A: Small Sample: $n = 250$										
A1: OLS										
0.1436	0	0.004	0.001	0.000	0.107	0.110	0.110	0.110	0.110	0.109
N(0,2)	n.a.	0.004	0.001	0.000	0.107	0.110	0.110	0.110	0.110	0.109
0.33	-0.1	0.004	0.000	0.167	0.106	0.108	0.109	0.108	0.108	0.107
0.33	0	0.004	0.001	-0.001	0.107	0.109	0.109	0.108	0.109	0.108
0.33	0.1	0.004	0.001	-0.169	0.106	0.108	0.108	0.108	0.108	0.107
-0.33	-0.1	0.001	-0.001	0.210	0.133	0.144	0.144	0.138	0.139	0.138
-0.33	0	0.003	0.001	0.003	0.118	0.122	0.124	0.122	0.123	0.121
-0.33	0.1	0.003	0.003	-0.202	0.127	0.126	0.128	0.129	0.130	0.129
-0.67	-0.25	-0.072	-0.001	0.841	1.736	1.667	1.739	1.623	1.675	1.618
-0.67	0	-0.010	0.004	0.026	0.336	0.337	0.340	0.329	0.336	0.328
-0.67	0.25	-0.017	0.022	-0.670	1.046	0.823	0.905	0.932	0.962	0.923
-1	-0.5	-4.006	-0.974	15.248	112.514	89.221	102.110	94.165	99.165	93.824
-1	0	-0.113	0.020	0.207	2.501	2.217	2.399	2.279	2.359	2.268
-1	0.5	-1.163	0.192	-6.572	39.725	27.930	34.479	35.005	36.486	34.414
A2: GTL										
0.1436	0	0.004	0.001	-0.001	0.108	0.111	0.126	0.109	0.109	0.125
N(0,2)	n.a.	0.004	0.001	-0.001	0.108	0.111	0.125	0.109	0.109	0.124
0.33	-0.1	0.003	0.004	0.020	0.089	0.094	0.130	0.094	0.095	0.128
0.33	0	0.004	0.001	-0.001	0.097	0.101	0.133	0.100	0.101	0.132
0.33	0.1	0.004	-0.003	-0.023	0.091	0.093	0.130	0.094	0.095	0.128
-0.33	-0.1	0.002	0.002	-0.001	0.066	0.068	0.066	0.067	0.067	0.067
-0.33	0	0.002	0.002	-0.001	0.071	0.073	0.071	0.071	0.071	0.072
-0.33	0.1	0.002	0.001	-0.001	0.067	0.069	0.067	0.067	0.067	0.067
-0.67	-0.25	0.000	0.001	0.000	0.025	0.026	0.025	0.025	0.025	0.025
-0.67	0	0.001	0.001	0.000	0.038	0.040	0.037	0.038	0.038	0.038
-0.67	0.25	0.001	0.000	0.000	0.025	0.026	0.025	0.025	0.025	0.025
-1	-0.5	0.000	0.000	0.000	0.005	0.005	0.005	0.005	0.005	0.005
-1	0	0.000	0.001	0.000	0.018	0.019	0.018	0.018	0.018	0.018
-1	0.5	0.000	0.000	0.000	0.005	0.005	0.005	0.005	0.005	0.005
B: Large Sample: $n = 5000$.										
B1: OLS										
0.1436	0	0.001	0.000	-0.001	0.024	0.024	0.025	0.024	0.024	0.024
0.33	-0.1	0.001	0.000	0.167	0.023	0.024	0.025	0.024	0.024	0.024
0.33	0	0.001	0.000	-0.001	0.023	0.024	0.025	0.024	0.024	0.024
0.33	0.1	0.001	0.000	-0.169	0.023	0.024	0.025	0.024	0.024	0.024
-0.33	-0.1	0.001	0.000	0.206	0.034	0.032	0.035	0.034	0.034	0.034
-0.33	0	0.001	0.000	-0.001	0.027	0.028	0.029	0.028	0.028	0.028
-0.33	0.1	0.001	0.000	-0.207	0.030	0.032	0.033	0.032	0.032	0.032
-0.67	-0.25	-0.138	0.023	1.148	6.074	3.868	4.797	4.891	4.931	4.903
-0.67	0	-0.005	-0.001	0.008	0.233	0.180	0.220	0.219	0.220	0.219
-0.67	0.25	-0.054	-0.048	-0.938	1.902	1.455	2.331	2.352	2.358	2.354
-1	-0.5	-251.913	78.897	365.907	9913.334	5630.799	7265.472	7422.937	7496.666	7449.832
-1	0	-0.386	-0.033	0.394	12.600	7.908	10.351	10.511	10.592	10.537
-1	0.5	-61.597	-34.904	-115.755	1996.952	831.578	2378.635	2395.306	2404.220	2398.811
B2: GTL										
0.1436	0.000	0.001	0.000	-0.001	0.024	0.024	0.028	0.024	0.024	0.027
0.33	-0.100	0.001	0.000	0.014	0.018	0.019	0.029	0.020	0.020	0.028
0.33	0.000	0.001	0.000	-0.001	0.021	0.021	0.030	0.022	0.022	0.029
0.33	0.1	0.001	0.000	-0.016	0.019	0.019	0.029	0.020	0.020	0.028
-0.33	-0.1	0.001	0.000	0.000	0.015	0.014	0.015	0.015	0.015	0.015
-0.33	0	0.001	0.000	0.000	0.016	0.015	0.016	0.016	0.016	0.016
-0.33	0.1	0.001	0.000	0.000	0.014	0.014	0.015	0.015	0.015	0.015
-0.67	-0.25	0.000	0.000	0.000	0.005	0.005	0.006	0.005	0.005	0.006
-0.67	0	0.000	0.000	0.000	0.008	0.008	0.008	0.008	0.008	0.008
-0.67	0.25	0.000	0.000	0.000	0.005	0.005	0.006	0.005	0.005	0.006
-1	-0.5	0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.001
-1	0	0.000	0.000	0.000	0.004	0.004	0.004	0.004	0.004	0.004
-1	0.5	0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.001

regardless of δ; the reason is that the maximum likelihood function is augmented with a small regularity penalty function[33] to keep $(\hat{\alpha}, \hat{\delta})$ within the feasible parameter area where the GTL estimator is consistent and asymptotically normal (Theorem 2). The penalty function is no longer needed when the values of α and δ of the data generating process are solidly inside this feasible area. The Monte Carlo results indicate that even

[33] See Web Appendix E.3 for details.

Table E.2: GTLE bias and precision of GTL parameters for GTL-generated data

DGP			Bias			Standard deviation		
α	δ	σ	σ	α	δ	σ	α	δ
A: GTL as an approximation of the standard normal distribution ($N = 250$)								
0.1436	0.00	1.188	0.0277	0.0219	-0.0007	0.1277	0.0638	0.0334
$N(0,2)^a$	n.a.	n.a.	0.0168	0.0158	-0.0007	0.1269	0.0650	0.0343
B: Various GTL distributions, small sample ($N = 250$)								
0.33	-0.10	1.477	-0.090	-0.032	0.005	0.101	0.041	0.028
0.33	0.00	1.508	-0.060	-0.017	-0.001	0.112	0.046	0.028
0.33	0.10	1.477	-0.090	-0.032	-0.006	0.097	0.040	0.029
-0.33	-0.10	0.454	0.001	0.006	-0.001	0.056	0.081	0.052
-0.33	0.00	0.482	0.001	0.006	-0.001	0.059	0.081	0.052
-0.33	0.10	0.454	0.001	0.006	0.000	0.056	0.081	0.052
-0.67	-0.25	0.136	0.000	0.003	-0.002	0.019	0.100	0.069
-0.67	0.00	0.202	-0.001	0.002	-0.001	0.028	0.099	0.068
-0.67	0.25	0.136	0.000	0.003	0.000	0.019	0.100	0.069
-1.00	-0.50	0.024	0.000	0.002	-0.002	0.004	0.120	0.088
-1.00	0.00	0.079	0.000	-0.001	-0.001	0.012	0.118	0.085
-1.00	0.50	0.024	0.000	0.001	0.000	0.004	0.119	0.086
C: Various GTL distributions, large sample ($N = 5000$)								
0.33	-0.10	1.477	-0.085	-0.037	0.004	0.022	0.009	0.006
0.33	0.00	1.508	-0.066	-0.028	0.000	0.025	0.009	0.005
0.33	0.10	1.477	-0.085	-0.037	-0.004	0.022	0.009	0.005
-0.33	-0.10	0.454	0.000	0.001	0.000	0.012	0.018	0.011
-0.33	0.00	0.482	0.000	0.001	0.000	0.013	0.018	0.011
-0.33	0.10	0.454	0.000	0.001	0.000	0.012	0.018	0.011
-0.67	-0.25	0.136	0.000	0.001	0.001	0.004	0.023	0.015
-0.67	0.00	0.202	0.000	0.000	0.001	0.006	0.023	0.015
-0.67	0.25	0.136	0.000	0.000	0.000	0.004	0.023	0.015
-1.00	-0.50	0.024	0.000	0.001	0.001	0.001	0.027	0.019
-1.00	0.00	0.079	0.000	0.000	0.001	0.003	0.027	0.019
-1.00	0.50	0.024	0.000	0.000	0.001	0.001	0.028	0.020

Note: [a] Bias and RMSE relative to equivalent GTL parameters.

for relatively small samples, bias in the GTL estimator of σ, α and δ is no issue.

E.3 Monte Carlo results: Diagnostics of small-sample distributions

Tables E.3 - E.5 present diagnostic results of the small-sample distributions of the OLS and MLE estimators of the various experiments that are discussed in Section 4 of the paper. We refer to these experiments as the "baseline design" since we consider an alternative design in Web Appendix E.4 below.

Prior to examining the diagnostic results, some remarks are in order. Theoretically, since the GTL regression estimator is a maximum likelihood estimator, the variance (denoted as V_h) of the estimator may be calculated as the inverse of the negative Hessian. However, when a penalty function[34] is added to the log-likelihood function in order to keep the estimator of (α, δ) in bounds, V_h is impacted by the curvature in this penalty function and thus may well be biased. More properly, the log-likelihood-with-penalty function may be seen as a criterion function within the context of quasi-likelihood estimation (White, 1982; Gourieroux et al., 1984),[35] in which case the familiar sandwich estimator (denoted as V_{sw}) ought to be used to compute the variance. In our set of experiments, we added a penalty function only for the experiments with $\alpha = 0.33$ and $\delta = -0.10, 0.00, 0.10$. For other experiments, the sandwich estimator should not be needed. However, as a check, we compute the sandwich estimator also when $\alpha = -0.33$.

Table E.5 shows that for $n = 5000$ and $\alpha = 0.33$, V_{sw} yields an estimated variance that is much closer to the variance of the Monte Carlo draws than V_h. But its use is questionable when $\alpha = -0.33$: the variance ratio moves away from 1 for four of the six parameters, increasing it for $\hat{\beta}_1$ and $\hat{\beta}_2$ and decreasing it for $\hat{\alpha}$ and $\hat{\delta}$. For small samples ($n = 250$), when $\alpha = 0.33$, V_h tends to be too large; V_{sw} tends to overcorrect and ends up to be often too small but is usually closer to the Monte Carlo variance. For $\alpha = -0.33$, the sandwich estimator yields a better value of the variance for $\hat{\beta}_2$ but worsens the variance for $\hat{\beta}_3$, $\hat{\beta}_1$, $\hat{\alpha}$, and $\hat{\delta}$.

As for coverage ratios (the proportion of confidence intervals that include the true population parameter), Table E.4 considers a significance level of 0.05. Ideally, therefore, the coverage ratios should equal 0.95. Use of the sandwich estimator when $\alpha = -0.33$

[34] The penalty function takes the following form. Let $\lambda_1 = \alpha - \delta$ and $\lambda_2 = \alpha + \delta$. Let the lower and upper limit on λ_j be denoted as λ_{jL} and λ_{jU}, respectively; in our Monte Carlo analysis, we set $\lambda_{jL} = -3$ and $\lambda_{jU} = 0.5$. Let $\lambda_{jR} = (\lambda_{jU} - \lambda_{jL})/2$ measure half of the feasible range of λ_j. As always, n denotes the number of observations. Then the penalty function is written as $P(\lambda_1, \lambda_2, n) = 0.005\ n\ \big(p(\lambda_1) + p(\lambda_2)\big)$

$$p(\lambda_j) = \ln\{(\lambda_j - \lambda_{jL})/\lambda_{jR}\} + \ln\{(\lambda_{jL} - \lambda_j)/\lambda_{jR}\}$$

This function p takes on a value of 0 at the midpoint -1.25 of the feasible range of λ_j, is symmetric around this midpoint, and for $\lambda_j = 0.20,\ 0.40,\ 0.45,\ 0.49,\ 0.499$ equals -0.00580, -0.01099, -0.01438, -0.02237, -0.03387 respectively. It has an asymptote of $-\infty$ at the endpoints of the range.

[35] H. White, "Maximum likelihood estimation of misspecified models." *Econometrica*, 1982, 50(1):1-25; C. Gourieroux, A. Monfort, and A. Trognon, "Pseudo maximum likelihood method: theory." *Econometrica*, 1984, 52(3):681-700.

makes little difference when $n = 5000$ and worsens coverage when $n = 250$. When $\alpha = 0.33$, the coverage ratio improves for $\hat{\beta}_2$ and $\hat{\beta}_3$; the other parameters suffer from bias especially when $\delta \neq 0$ and thus exhibit poor coverage performance.[36]

Based on these comparisons, the use of a sandwich estimator is recommended when a penalty function is added to the log-likelihood function but may do more harm than good when the regression model is estimated without a penalty function.

[36]The penalty function keeps $(\hat{\alpha}, \hat{\delta})$ within the feasible area but also has the effect of driving it away from the boundary of the feasible area. This produces a small but significant bias in $(\hat{\alpha}, \hat{\delta})$ in experiments with $\alpha = 0.33$. Experimentation with different penalty function specifications indicated that functions that become active only near the boundary (and thus are less likely to cause bias) typically cause non-convergence during some replications.

Table E.3: p-values of Jarque-Bera tests for normality of OLS and GTL estimators: baseline design

DGP			OLS			GTL					
α	δ	σ	β_2	β_3	β_1	β_2	β_3	β_1	σ	α	δ
A: GTL as an approximation of the standard normal distribution, $N = 250$											
0.1436	0.00	1.188	0.91	0.10	0.62	0.46	0.22	0.55	0.00	0.00	0.01
B: Various GTL distributions, small sample, $N = 250$											
0.33	-0.10	1.477	0.80	0.07	0.48	0.36	0.07	0.30	0.29	0.00	0.03
0.33	0.00	1.508	0.90	0.11	0.51	0.47	0.51	0.16	0.08	0.00	0.43
0.33	0.10	1.477	0.93	0.17	0.51	0.17	0.46	0.14	0.31	0.00	0.10
-0.33	-0.10	0.454	0.00	0.00	0.00	0.23	0.21	0.86	0.00	0.10	0.73
-0.33	0.00	0.482	0.00	0.00	0.02	0.11	0.10	0.72	0.00	0.12	0.59
-0.33	0.10	0.454	0.00	0.00	0.00	0.07	0.07	0.54	0.00	0.18	0.29
-0.67	-0.25	0.136	0.00	0.00	0.00	0.07	0.70	0.80	0.00	0.06	0.55
-0.67	0.00	0.202	0.00	0.00	0.00	0.00	0.43	0.78	0.00	0.04	0.63
-0.67	0.25	0.136	0.00	0.00	0.00	0.04	0.51	0.41	0.00	0.08	0.21
-1.00	-0.50	0.024	0.00	0.00	0.00	0.53	0.93	0.64	0.00	0.09	0.33
-1.00	0.00	0.079	0.00	0.00	0.00	0.00	0.57	0.81	0.00	0.02	0.57
-1.00	0.50	0.024	0.00	0.00	0.00	0.40	0.53	0.23	0.00	0.04	0.10
C: Various GTL distributions, large sample, $N = 5000$											
0.33	-0.10	1.477	0.09	0.99	0.07	0.68	0.23	0.78	0.71	0.78	0.09
0.33	0.00	1.508	0.10	0.96	0.16	0.65	0.55	0.70	0.48	0.69	0.10
0.33	0.10	1.477	0.14	0.89	0.33	0.87	0.67	0.61	0.59	0.56	0.26
-0.33	-0.10	0.454	0.00	0.00	0.00	0.51	0.50	0.99	0.30	0.44	0.06
-0.33	0.00	0.482	0.83	0.07	0.21	0.35	0.39	0.98	0.23	0.44	0.07
-0.33	0.10	0.454	0.00	0.00	0.00	0.19	0.41	0.97	0.15	0.42	0.08
-0.67	-0.25	0.136	0.00	0.00	0.00	1.00	0.39	0.98	0.71	0.42	0.11
-0.67	0.00	0.202	0.00	0.00	0.00	0.91	0.24	0.87	0.29	0.61	0.20
-0.67	0.25	0.136	0.00	0.00	0.00	0.40	0.40	0.73	0.13	0.50	0.30
-1.00	-0.50	0.024	0.00	0.00	0.00	0.75	0.49	0.93	0.98	0.42	0.07
-1.00	0.00	0.079	0.00	0.00	0.00	0.83	0.21	0.76	0.30	0.86	0.30
-1.00	0.50	0.024	0.00	0.00	0.00	0.18	0.45	0.47	0.04	0.74	0.53

Note: These test results pertain to simulations reported in Tables 1 and E.2.

Table E.4: Coverage rates of OLS and GTL estimators under a 95% confidence level: baseline design

DGP			OLS			GTL					
α	δ	σ	β_2	β_3	β_1	β_2	β_3	β_1	σ	α	δ
A: GTL as an approximation of the standard normal distribution, $N = 250$											
0.1436	0.00	1.188	0.947	0.948	0.954	0.944	0.937	0.943	0.941	0.918	0.924
B: Various GTL distributions, small sample, $N = 250$											
0.33	-0.10	1.477	0.948	0.951	0.658	0.959	0.947	0.942	0.905	0.970	0.965
						0.935	*0.915*	*0.949*	*0.853*	*0.899*	*0.913*
0.33	0.00	1.508	0.943	0.950	0.951	0.951	0.944	0.944	0.935	0.976	0.964
						0.938	*0.912*	*0.948*	*0.909*	*0.909*	*0.909*
0.33	0.10	1.477	0.944	0.950	0.651	0.959	0.948	0.946	0.908	0.976	0.963
						0.941	*0.914*	*0.948*	*0.862*	*0.893*	*0.915*
-0.33	-0.10	0.454	0.958	0.948	0.655	0.952	0.936	0.951	0.943	0.946	0.950
-0.33	0.00	0.482	0.952	0.946	0.951	0.956	0.944	0.946	0.944	0.945	0.947
-0.33	0.10	0.454	0.956	0.945	0.632	0.950	0.939	0.948	0.939	0.936	0.936
-0.67	-0.25	0.136	0.955	0.959	0.337	0.955	0.933	0.955	0.951	0.948	0.943
-0.67	0.00	0.202	0.954	0.947	0.962	0.955	0.936	0.955	0.944	0.955	0.937
-0.67	0.25	0.136	0.952	0.946	0.326	0.950	0.935	0.952	0.938	0.943	0.935
-1.00	-0.50	0.024	0.951	0.957	0.571	0.949	0.926	0.953	0.951	0.955	0.940
-1.00	0.00	0.079	0.951	0.946	0.972	0.954	0.929	0.956	0.940	0.949	0.934
-1.00	0.50	0.024	0.947	0.956	0.558	0.949	0.939	0.951	0.939	0.944	0.932
C: Various GTL distributions, large sample, $N = 5000$											
0.33	-0.10	1.477	0.964	0.950	0.000	0.966	0.960	0.913	0.052	0.024	0.932
						0.955	*0.948*	*0.913*	*0.026*	*0.007*	*0.868*
0.33	0.00	1.508	0.963	0.949	0.938	0.961	0.950	0.942	0.290	0.285	0.977
						0.955	*0.941*	*0.942*	*0.214*	*0.142*	*0.950*
0.33	0.10	1.477	0.963	0.951	0.000	0.965	0.954	0.910	0.065	0.032	0.946
						0.953	*0.936*	*0.911*	*0.027*	*0.007*	*0.889*
-0.33	-0.10	0.454	0.956	0.944	0.000	0.960	0.956	0.951	0.941	0.948	0.944
-0.33	0.00	0.482	0.959	0.944	0.944	0.954	0.951	0.948	0.942	0.942	0.943
-0.33	0.10	0.454	0.957	0.947	0.001	0.956	0.952	0.947	0.941	0.940	0.945
-0.67	-0.25	0.136	0.952	0.961	0.093	0.953	0.956	0.949	0.938	0.949	0.944
-0.67	0.00	0.202	0.959	0.952	0.967	0.956	0.952	0.946	0.942	0.948	0.941
-0.67	0.25	0.136	0.954	0.954	0.100	0.947	0.956	0.944	0.949	0.937	0.944
-1.00	-0.50	0.024	0.951	0.954	0.486	0.952	0.954	0.949	0.939	0.952	0.948
-1.00	0.00	0.079	0.950	0.964	0.978	0.953	0.958	0.948	0.947	0.949	0.943
-1.00	0.50	0.024	0.958	0.956	0.481	0.953	0.943	0.945	0.947	0.934	0.939

Notes: These test results pertain to simulations reported in Tables 1 and E.2. For single-row sets of results, the coverage ratio is based on a variance that is computed from the information matrix (inverse negative Hessian). For double-row sets of results, the first row uses a variance that is computed from the information matrix; the second row (in italics) uses a sandwich estimator to compute the variance.

Table E.5: Ratio of average estimated variance to Monte Carlo variance: baseline design

DGP			OLS			GTL					
α	δ	σ	β_2	β_3	β_1	β_2	β_3	β_1	σ	α	δ
A: GTL as an approximation of the standard normal distribution, $N = 250$											
0.1436	0.00	1.188	1.045	1.000	0.979	1.010	0.961	0.988	0.905	0.878	0.856
B: Various GTL distributions, small sample, $N = 250$											
0.33	-0.10	1.477	1.033	0.995	0.972	1.131	1.028	0.959	1.280	1.606	1.226
						0.932	*0.827*	*0.994*	*0.974*	*0.865*	*0.812*
0.33	0.00	1.508	1.037	0.994	0.984	1.064	1.006	0.979	1.255	1.460	1.157
						0.932	*0.858*	*1.001*	*1.017*	*0.869*	*0.765*
0.33	0.10	1.477	1.042	0.995	0.996	1.080	1.055	0.960	1.392	1.691	1.167
						0.888	*0.851*	*0.994*	*1.064*	*0.916*	*0.773*
-0.33	-0.10	0.454	1.073	0.926	0.915	1.018	0.956	1.031	0.970	0.992	0.929
-0.33	0.00	0.482	1.073	1.004	0.960	1.005	0.945	1.024	0.966	0.996	0.929
-0.33	0.10	0.454	1.037	1.071	1.018	0.989	0.933	1.016	0.960	0.994	0.928
-0.67	-0.25	0.136	0.875	1.010	0.866	1.005	0.945	1.051	1.007	1.019	0.933
-0.67	0.00	0.202	0.960	0.994	0.931	0.997	0.922	1.045	0.994	1.026	0.937
-0.67	0.25	0.136	0.794	1.368	1.042	0.987	0.912	1.032	0.977	1.020	0.946
-1.00	-0.50	0.024	0.700	1.235	0.844	0.991	0.940	1.052	1.039	1.019	0.933
-1.00	0.00	0.079	0.830	1.132	0.893	0.978	0.896	1.054	1.008	1.032	0.939
-1.00	0.50	0.024	0.777	1.707	0.996	0.991	0.903	1.033	0.993	1.030	0.961
C: Various GTL distributions, large sample, $N = 5000$											
0.33	-0.10	1.477	1.053	1.006	0.952	1.177	1.137	0.958	1.342	1.609	1.338
						1.031	*0.992*	*0.967*	*0.992*	*0.931*	*0.974*
0.33	0.00	1.508	1.051	1.008	0.942	1.111	1.058	0.965	1.201	1.420	1.346
						1.038	*0.986*	*0.963*	*0.954*	*0.913*	*0.989*
0.33	0.10	1.477	1.049	1.010	0.931	1.177	1.110	0.957	1.262	1.583	1.415
						1.031	*0.970*	*0.965*	*0.932*	*0.914*	*1.029*
-0.33	-0.10	0.454	1.050	1.046	0.901	1.027	1.036	0.992	0.912	0.908	0.951
-0.33	0.00	0.482	1.035	1.006	0.897	1.034	1.038	0.997	0.907	0.905	0.949
-0.33	0.10	0.454	1.076	0.999	0.921	1.042	1.035	1.003	0.907	0.906	0.948
-0.67	-0.25	0.136	1.839	1.897	0.998	1.006	1.058	0.996	0.918	0.932	0.940
-0.67	0.00	0.202	1.307	1.446	0.980	1.019	1.066	1.005	0.908	0.921	0.931
-0.67	0.25	0.136	1.529	2.627	1.020	1.035	1.034	1.011	0.920	0.918	0.925
-1.00	-0.50	0.024	1.676	0.981	1.011	0.995	1.046	1.002	0.948	0.967	0.945
-1.00	0.00	0.079	2.727	0.966	0.998	1.008	1.080	1.017	0.923	0.943	0.924
-1.00	0.50	0.024	20.456	0.382	1.005	1.024	0.998	1.014	0.944	0.922	0.909

Notes: These test results pertain to simulations reported in Tables 1 and E.2. For single-row sets of results, the numerator of the ratio is based on a variance that is computed from the information matrix (inverse negative Hessian). For double-row sets of results, the first row uses a variance that is computed from the information matrix; the second row (in italics) uses a sandwich estimator to compute the variance in the numerator.

E.4 A second design: thick-tailed observables

One might question the wisdom of a Monte Carlo design that specifies thick-tailed unobservables but regular-tailed observable determinants. Regression designs that contain thick-tailed explanatory variables may be problematic:[37] the OLS estimator may lose its consistency property when the data generating process creates a few observations with outlying values of x that have an outsized influence on the position of the regression line. However, such problems appear to be alleviated when the disturbance is normally distributed (Jureckova, Koenker, and Portnoy (op.cit.)).

Table E.6 contains a few cases with thick-tailed disturbances and both regular- and thick-tailed x-variables, generated as GTL variates. Specifically, X_j is distributed GTL$(\alpha_{xj}, \delta_{xj})$ for $j = 2, 3$ with $\alpha_{xj} = -0.67$ and $\delta_{xj} = -0.25, 0.00, 0.25$. In this design, if X denotes the matrix of explanatory variables including a column of ones for the intercept, $plim\ X'X/n$ is not defined. As before, the GTL distribution of these GTL variates is scaled such that the distance between the 0.1% quantile and the 99.9% quantile is the same as that of a standard normal distribution. It is well-known that the OLS estimator is no longer consistent; the GTL-regression estimator is likely similarly impacted.

The disturbances in Panel A are distributed GTL$(-0.67, 0.25)$ and thus are thick-tailed and skewed. The explanatory variables change from two standard normal ones to two thick-tailed skewed ones: the performance of the OLS estimator actually improves, as does the GTL estimator. Panel B shows that this also happens when the disturbances are standard normal (or, closely similar, GTL(0.1436,0)). These results seem at odds with the research of Huber (op. cit.) that the OLS estimator is inconsistent with heavy-tailed explanatory variables. However, Panels A and B of Table E.6 do not address the matter of consistency as sample size is unchanged between these experiments. The reason for the improved performance is that despite the scaling of the GTL distribution the variation

[37] See, e.g., P.J. Huber, *Robust Statistics*. Wiley, New York, 1981; X. He, J. Jureckova, R. Koenker, and S. Portnoy, "Tail behavior of regression estimators and their breakdown points." *Econometrica*, 1990, 58(5):1195-1214; and J. Jureckova, R. Koenker, and S. Portnoy, "Tail behavior of the least-squares estimator." *Statistics and Probability Letters*, 2001, 55(4):377-384.

Table E.6: RMSE of OLS and GTL estimators of β with thick-tailed determinants

DGP for x_2		DGP for x_3		OLS			GTLE		
α_{x2}	δ_{x2}	α_{x3}	δ_{x3}	β_2	β_3	β_1	β_2	β_3	β_1
A: GTL-Disturbances are generated with $(\alpha, \delta) = (-0.67, 0.25)$ for $n = 250$									
0.1436	0.00	0.1436	0.00	1.0466	0.9638	1.1273	0.0250	0.0259	0.0248
0.1436	0.00	-0.33	0.00	1.0464	0.5981	1.1254	0.0249	0.0152	0.0249
0.1436	0.00	-0.33	0.10	1.0486	0.6754	1.1130	0.0249	0.0156	0.0250
0.1436	0.00	-0.67	0.00	1.0456	0.5650	1.1226	0.0249	0.0134	0.0249
0.1436	0.00	-0.67	0.25	1.0484	0.7007	1.1037	0.0249	0.0151	0.0252
-0.67	0.00	-0.67	0.00	0.4929	0.5712	1.1231	0.0135	0.0134	0.0248
-0.67	0.25	-0.67	0.25	0.4370	0.7064	1.1105	0.0155	0.0151	0.0255
-0.67	0.25	-0.67	-0.25	0.4355	0.3487	1.1504	0.0155	0.0144	0.0253
-0.67	-0.25	-0.67	-0.25	0.3201	0.3622	1.1519	0.0145	0.0146	0.0253
B: GTL-Disturbances are generated with $(\alpha, \delta) = (0.1436, 0)$ for $n = 250$									
0.1436	0.00	0.1436	0.00	0.1073	0.1108	0.1101	0.1085	0.1126	0.1257
-0.67	0.00	-0.67	0.00	0.0488	0.0494	0.1105	0.0491	0.0498	0.1252
-0.67	0.25	-0.67	0.25	0.0512	0.0515	0.1134	0.0516	0.0518	0.1291
-0.67	0.25	-0.67	-0.25	0.0512	0.0507	0.1133	0.0517	0.0513	0.1284
-0.67	-0.25	-0.67	-0.25	0.0517	0.0506	0.1132	0.0522	0.0512	0.1281
C: GTL-Disturbances are generated with $(\alpha, \delta) = (-0.67, 0.25)$ for $n = 5000$									
0.1436	0.00	0.1436	0.00	1.8968	1.5547	2.5085	0.0054	0.0054	0.0056
0.1436	0.00	-0.33	0.00	1.9060	0.4835	2.5109	0.0054	0.0028	0.0056
0.1436	0.00	-0.33	0.10	1.9043	0.4290	2.5738	0.0054	0.0026	0.0056
0.1436	0.00	-0.67	0.00	1.9124	0.0693	2.5149	0.0054	0.0015	0.0056
0.1436	0.00	-0.67	0.25	1.9111	0.0313	2.5252	0.0054	0.0009	0.0056
-0.67	0.00	-0.67	0.00	0.2560	0.0697	2.5241	0.0015	0.0015	0.0055
-0.67	0.25	-0.67	0.25	0.0940	0.0328	2.5778	0.0009	0.0009	0.0056
-0.67	0.25	-0.67	-0.25	0.0940	0.0303	2.5676	0.0009	0.0010	0.0056
-0.67	-0.25	-0.67	-0.25	0.0784	0.0303	2.5334	0.0010	0.0010	0.0056

in the explanatory variables increases dramatically as the tails thicken. The R^2 rises for many of these OLS regressions from around 0.33 to nearly 1. Across replications, the largest diagonal element of the hat-matrix (i.e., $X(X'X)^{-1}X'$) averages 0.0523 when both explanatory variables are standard normal and averages 0.7315 (ranging from 0.1223 to 0.9999) when both explanatory variables are drawn from a GTL(-0.67,0.25) distribution.

This leaves the question what happens when n increases. In Panel C, we raise n to 5000. If the estimator is $n^{1/2}$-consistent, the standard errors should diminish by a factor of $20^{1/2} = 4.47$, and the RMSE should follow suit since bias is not much of an issue (except for in the intercept). However, as x-variables are drawn from $\text{GTL}(-0.33, \delta)$ and $\text{GTL}(-0.67, \delta)$ for several values of δ, they are so thick-tailed that the fourth moment of x no longer exists and Assumption A.1 is violated. Whenever x_2 is standard normal, the

RMSE of $\hat{\beta}_{2OLS}$ rises with n, but when x_2 is GTL$(-0.67, \delta)$ for any δ, it falls by a factor of approximately 4.5. As for the RMSE of $\hat{\beta}_{3OLS}$, it rises when x_3 is standard normal; it falls slightly when x_3 is GTL$(-0.33, \delta)$ for any δ; and when x_3 is GTL$(-0.67, \delta)$ for any δ, it falls by a factor of up to 22. The distinction between these cases is that the second moment of a GTL distribution is finite for GTL$(-0.33, \delta)$ but is not defined for GTL$(-0.67, \delta)$. It appears to be the relative thickness of the tails of the explanatory variables and the disturbances that determines the large-sample behavior of the OLS estimator.

As for the GTL estimator, the RMSE of $\hat{\beta}_{jGTL}$ decreases by a factor of approximately 4.5 when x_j is standard normal or GTL$(-0.33, \delta)$, which indicates $\sqrt{n}$-consistency, and it decreases by a factor of 15 to 17.5 when x_j is GTL$(-0.67, \delta)$, which might indicate a type of super consistency.

As with the baseline, we also examine the proximity to normality of the small-sample distribution of the estimators. Table E.7 illustrates that normality of the OLS slope estimators is destroyed by thick tails in the disturbances or explanatory variables. The GTL-regression estimator of the slopes displays normality for large n in the estimators of α, δ, σ, and β_1. If X_2 is well-behaved and X_3 is thick-tailed, $\hat{\beta}_2$ is normally distributed but $\hat{\beta}_3$ is not. Table E.8 shows that the inverted negative hessian underestimates the Monte Carlo variance of the GTL-regression estimator if determinants are thick-tailed. However, the measured deviation is not as large as it is for the OLS estimator. (Note that this “deviation” cannot be called a “bias” since the mean of the variance estimator is not defined.) Table E.9 shows that these poor estimates impact the coverage ratio as well: for OLS, unlike in the baseline design, thick-tailed determinants lead to a excessive coverage ratio (too large confidence intervals), and for the GTL-regression estimator, the coverage ratios are too low (too small confidence intervals).

Table E.7: p-values of Jarque-Bera tests for normality of OLS and GTL estimators: design with thick-tailed determinants

DGP for x_1		DGP for x_2		OLS			MLE					
α_{x2}	δ_{x2}	α_{x3}	δ_{x3}	β_2	β_3	β_1	β_2	β_3	β_1	σ	α	δ
A: GTL-Disturbances are generated with $(\alpha, \delta) = (-0.67, 0.25)$ for $n = 250$												
0.1436	0.00	0.1436	0.00	0.00	0.00	0.00	0.03	0.62	0.43	0.00	0.11	0.21
0.1436	0.00	-0.33	0.00	0.00	0.00	0.00	0.05	0.02	0.32	0.00	0.15	0.15
0.1436	0.00	-0.33	0.10	0.00	0.00	0.00	0.03	0.00	0.39	0.00	0.16	0.19
0.1436	0.00	-0.67	0.00	0.00	0.00	0.00	0.09	0.00	0.39	0.00	0.12	0.12
0.1436	0.00	-0.67	0.25	0.00	0.00	0.00	0.03	0.00	0.48	0.00	0.13	0.20
-0.67	0.00	-0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.54	0.00	0.07	0.12
-0.67	0.25	-0.67	0.25	0.00	0.00	0.00	0.00	0.00	0.55	0.00	0.09	0.24
-0.67	0.25	-0.67	-0.25	0.00	0.00	0.00	0.00	0.00	0.36	0.00	0.10	0.06
-0.67	-0.25	-0.67	-0.25	0.00	0.00	0.00	0.00	0.00	0.35	0.00	0.07	0.04
B: GTL-Disturbances are generated with $(\alpha, \delta) = (0.1436, 0)$ for $n = 250$												
0.1436	0.00	0.1436	0.00	0.94	0.27	0.64	0.48	0.48	0.54	0.00	0.00	0.00
-0.67	0.00	-0.67	0.00	0.00	0.00	0.61	0.00	0.00	0.41	0.00	0.08	0.14
-0.67	0.25	-0.67	0.25	0.00	0.00	0.37	0.00	0.00	0.66	0.00	0.07	0.05
-0.67	0.25	-0.67	-0.25	0.00	0.00	0.55	0.00	0.00	0.75	0.00	0.06	0.04
-0.67	-0.25	-0.67	-0.25	0.00	0.00	0.47	0.00	0.00	0.31	0.00	0.07	0.17
C: GTL-Disturbances are generated with $(\alpha, \delta) = (-0.67, 0.25)$ for $n = 5000$												
0.1436	0.00	0.1436	0.00	0.00	0.00	0.00	0.38	0.24	0.72	0.14	0.50	0.29
0.1436	0.00	-0.33	0.00	0.00	0.00	0.00	0.39	0.37	0.72	0.14	0.50	0.30
0.1436	0.00	-0.33	0.10	0.00	0.00	0.00	0.39	0.06	0.92	0.14	0.51	0.29
0.1436	0.00	-0.67	0.00	0.00	0.00	0.00	0.41	0.00	0.76	0.13	0.51	0.31
0.1436	0.00	-0.67	0.25	0.00	0.00	0.00	0.39	0.00	0.88	0.14	0.52	0.30
-0.67	0.00	-0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.76	0.13	0.51	0.30
-0.67	0.25	-0.67	0.25	0.00	0.00	0.00	0.00	0.00	0.87	0.13	0.54	0.31
-0.67	0.25	-0.67	-0.25	0.00	0.00	0.00	0.00	0.00	0.61	0.12	0.51	0.32
-0.67	-0.25	-0.67	-0.25	0.00	0.00	0.00	0.00	0.00	0.50	0.12	0.51	0.31

Note: These test results pertain to simulations reported in Table E.6.

Table E.8: Ratio of average estimated variance to Monte Carlo variance: design with thick-tailed determinants

DGP for x_1		DGP for x_2		OLS			MLE					
α_{x2}	δ_{x2}	α_{x3}	δ_{x3}	β_2	β_3	β_1	β_2	β_3	β_1	σ	α	δ
A: GTL-Disturbances are generated with $(\alpha, \delta) = (-0.67, 0.25)$ for $n = 250$												
0.1436	0.00	0.1436	0.00	0.792	0.958	1.038	0.987	0.920	1.031	0.977	1.020	0.946
0.1436	0.00	-0.33	0.00	0.792	0.883	1.043	0.992	0.860	1.026	0.974	1.018	0.945
0.1436	0.00	-0.33	0.10	0.788	0.667	1.079	0.993	0.830	1.028	0.976	1.018	0.945
0.1436	0.00	-0.67	0.00	0.794	0.792	1.050	0.995	0.743	1.025	0.982	1.022	0.943
0.1436	0.00	-0.67	0.25	0.788	0.568	1.109	0.996	0.630	1.013	0.985	1.027	0.941
-0.67	0.00	-0.67	0.00	0.833	0.773	1.050	0.785	0.707	1.029	0.976	1.026	0.938
-0.67	0.25	-0.67	0.25	0.718	0.558	1.105	0.559	0.664	1.007	0.986	1.033	0.935
-0.67	0.25	-0.67	-0.25	0.724	2.408	1.018	0.564	0.588	1.019	0.975	1.025	0.931
-0.67	-0.25	-0.67	-0.25	2.115	2.285	1.028	0.548	0.608	1.014	0.970	1.023	0.939
B: GTL-Disturbances are generated with $(\alpha, \delta) = (0.1436, 0)$ for $n = 250$												
0.1436	0.00	0.1436	0.00	1.045	0.973	0.979	1.010	0.933	0.986	0.903	0.876	0.852
-0.67	0.00	-0.67	0.00	1.051	1.000	0.972	1.037	0.983	0.989	0.921	0.903	0.880
-0.67	0.25	-0.67	0.25	1.038	1.007	0.960	1.024	0.992	0.965	0.922	0.906	0.870
-0.67	0.25	-0.67	-0.25	1.037	1.002	0.961	1.018	0.980	0.975	0.909	0.895	0.878
-0.67	-0.25	-0.67	-0.25	1.006	1.011	0.963	0.993	0.991	0.980	0.919	0.904	0.884
C: GTL-Disturbances are generated with $(\alpha, \delta) = (-0.67, 0.25)$ for $n = 5000$												
0.1436	0.00	0.1436	0.00	2.707	0.589	0.989	1.034	1.039	1.010	0.918	0.917	0.924
0.1436	0.00	-0.33	0.00	2.714	1.299	0.986	1.033	1.029	1.011	0.918	0.918	0.924
0.1436	0.00	-0.33	0.10	2.710	1.914	0.935	1.037	1.019	1.005	0.918	0.917	0.924
0.1436	0.00	-0.67	0.00	2.710	7.029	0.979	1.036	0.842	1.013	0.919	0.918	0.925
0.1436	0.00	-0.67	0.25	1.516	85.169	1.013	1.039	0.753	1.008	0.918	0.918	0.926
-0.67	0.00	-0.67	0.00	1.836	6.924	0.980	0.836	0.840	1.015	0.917	0.918	0.925
-0.67	0.25	-0.67	0.25	28.386	77.614	0.978	0.685	0.758	1.007	0.917	0.919	0.926
-0.67	0.25	-0.67	-0.25	14.215	12.190	0.983	0.683	0.623	1.010	0.919	0.919	0.924
-0.67	-0.25	-0.67	-0.25	3.485	12.194	0.982	0.555	0.621	1.020	0.918	0.919	0.926

Note: These test results pertain to simulations reported in Table E.6.

Table E.9: Coverage rates of OLS and GTL estimators under a 95% confidence level: design with thick-tailed determinants

DGP for x_1		DGP for x_2		OLS			MLE					
α_{x2}	δ_{x2}	α_{x3}	δ_{x3}	β_2	β_3	β_1	β_2	β_3	β_1	σ	α	δ
A: GTL-Disturbances are generated with $(\alpha, \delta) = (-0.67, 0.25)$ for $n = 250$												
0.1436	0.00	0.1436	0.00	0.952	0.947	0.323	0.951	0.944	0.951	0.939	0.945	0.935
0.1436	0.00	-0.33	0.00	0.953	0.950	0.325	0.953	0.915	0.952	0.937	0.944	0.935
0.1436	0.00	-0.33	0.10	0.951	0.951	0.328	0.952	0.909	0.950	0.937	0.944	0.935
0.1436	0.00	-0.67	0.00	0.953	0.957	0.329	0.951	0.838	0.951	0.938	0.946	0.936
0.1436	0.00	-0.67	0.25	0.951	0.966	0.334	0.954	0.793	0.944	0.941	0.944	0.934
-0.67	0.00	-0.67	0.00	0.954	0.957	0.328	0.839	0.831	0.953	0.938	0.942	0.935
-0.67	0.25	-0.67	0.25	0.952	0.965	0.358	0.780	0.791	0.950	0.941	0.945	0.934
-0.67	0.25	-0.67	-0.25	0.951	0.964	0.358	0.790	0.788	0.950	0.940	0.946	0.932
-0.67	-0.25	-0.67	-0.25	0.964	0.963	0.342	0.779	0.790	0.949	0.936	0.943	0.930
B: GTL-Disturbances are generated with $(\alpha, \delta) = (0.1436, 0)$ for $n = 250$												
0.1436	0.00	0.1436	0.00	0.948	0.947	0.954	0.946	0.934	0.943	0.939	0.917	0.925
-0.67	0.00	-0.67	0.00	0.945	0.943	0.953	0.945	0.938	0.945	0.946	0.920	0.931
-0.67	0.25	-0.67	0.25	0.952	0.947	0.947	0.946	0.943	0.939	0.948	0.930	0.927
-0.67	0.25	-0.67	-0.25	0.951	0.961	0.951	0.945	0.955	0.948	0.946	0.924	0.929
-0.67	-0.25	-0.67	-0.25	0.948	0.961	0.951	0.946	0.954	0.948	0.947	0.922	0.929
C: GTL-Disturbances are generated with $(\alpha, \delta) = (-0.67, 0.25)$ for $n = 5000$												
0.1436	0.00	0.1436	0.00	0.954	0.947	0.098	0.945	0.954	0.945	0.950	0.935	0.945
0.1436	0.00	-0.33	0.00	0.955	0.948	0.099	0.946	0.950	0.945	0.949	0.936	0.943
0.1436	0.00	-0.33	0.10	0.954	0.952	0.102	0.946	0.931	0.947	0.949	0.935	0.943
0.1436	0.00	-0.67	0.00	0.954	0.972	0.098	0.948	0.854	0.948	0.947	0.936	0.943
0.1436	0.00	-0.67	0.25	0.955	0.982	0.099	0.947	0.816	0.948	0.949	0.937	0.942
-0.67	0.00	-0.67	0.00	0.980	0.972	0.098	0.863	0.857	0.947	0.947	0.936	0.945
-0.67	0.25	-0.67	0.25	0.988	0.982	0.098	0.782	0.816	0.950	0.948	0.937	0.945
-0.67	0.25	-0.67	-0.25	0.988	0.987	0.101	0.783	0.792	0.946	0.950	0.936	0.944
-0.67	-0.25	-0.67	-0.25	0.985	0.987	0.101	0.789	0.793	0.943	0.948	0.936	0.944

Note: These test results pertain to simulations reported in Table E.6.

F Comparison of GTL distribution with other distributions

The GTL distribution is highly flexible. In this appendix, by means of example, the GTL distribution is shown to closely approximate various other distributions: normal, $t(5)$, $\chi^2(8)$, Gumbel, and a highly skewed t distribution.

Pertaining to the GTL distribution or to its cousin, the Generalized Lambda Distribution of Ramberg and Schmeiser (1974)[38] (see also Karian and Dudewicz (2000,2011)[39]), many approximation methods are suggested in the literature of statistical data analysis. Ramberg et al. (1979)[40] propose to estimate the distribution's parameters by the method of moments using the first four moments. Öztürk and Dale (1985)[41] suggest a least squares estimation method. King and MacGillivray (1999)[42] fit the GTL distribution to the data by the "starship" method, which is a computationally intensive grid-search, while Karian and Dudewicz (1999)[43] and Su (2011)[44] derive parameter estimates by quantile matching, and Karvanen and Nuutinen (2008)[45] do the same from L-moments. Su (2007)[46] proposes an algorithm that combines a random grid search with maximum likelihood estimation and examines the bias and variance of this estimator by simulation.

To illustrate how closely the GTL distribution can approximate any other distribution, we use the L_1-norm of the absolute difference between the GTL density $f(.)$ and another

[38] J.S. Ramberg and B.W. Schmeiser. "An approximate method for generating asymmetric random variables." *Communications of the ACM*, 1974, 17(2):78-82.

[39] Z.A. Karian and E.J. Dudewicz. *Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods.* 2000, CRC Press, Boca Raton, FL. Z.A. Karian and E.J. Dudewicz. *Handbook of Fitting Statistical Distributions with R.* 2011 CRC Press., Boca Raton, FL.

[40] J.S. Ramberg, P.R. Tadikamalla, E.J. Dudewicz, and E.F. Mykytka. "A probability distribution and its uses in fittingdata." *Technometrics*, 1979, 21(2):201-214.

[41] A. Öztürk and R.F. Dale. "Least squares estimation of the parameters of the Generalized Lambda distribution." *Technometrics*, 1985, 27(1):81-84.

[42] R.A.R. King and H.L. MacGillivray. "A starship estimation method for the Generalized λ Distributions." *Australian & New Zealand Journal of Statistics*, 1999, 41(3):353-374.

[43] Z.A. Karian and E.J. Dudewicz. "Fitting the Generalized Lambda Distribution to data: A method based on percentiles." *Communications in Statistics: Simulation & Computation*, 1999, 28(3):793-819.

[44] S. Su. "Fitting GLD to data via quantile matching method." In: Z.A. Karian and E.J. Dudewicz, (Eds.), *Handbook on Fitting Statistical Distributions with R.* 2011, Boca Raton, FL., Ch. 14, pp. 557-583.

[45] J. Karvanen and A. Nuutinen. "Characterizing the Generalized Lambda distribution by L-moments." *Computational Statistics & Data Analysis*, 2008, 52(4):1971-1983.

[46] S. Su. "Numerical maximum log likelihood estimation for Generalized Lambda distributions." *Computational Statistics & Data Analysis*, 2007, 51(8): 3983-3998.

density $g(.)$:

$$L_1(f,g) = \int |f(\epsilon) - g(\epsilon))| d\epsilon. \tag{F.1}$$

In principle, the approximation is obtained by choosing (α, δ) such that $L_1(f,g)$ is minimized. If g is phrased in a standardized form with mean 0 and variance 1, f is similarly standardized; if g is phrased in a canonical form, f is stated in a general form and $L_1(f,g)$ is minimized with respect to not only (α, δ) but also a location and scaling parameter. In a regression context such as is implemented in this paper, these four parameters are estimated jointly with the slope parameters by means of the maximum likelihood approach. As will be illustrated below, this can impact the fit of the estimated GTL distribution with the distribution that underlies the data generating process.

The standard normal density function is given by:

$$g(\epsilon) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\epsilon^2\right) \tag{F.2}$$

As is well-known, the normal distribution has a skewness value of 0, and its kurtosis equals 3.

In its canonical form, the density of the $t(\nu)$ distribution is given by:

$$g(\epsilon) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{\epsilon}{\nu}\right)^{-(\nu+1)/2} \tag{F.3}$$

The mean of this distribution equals 0 and the variance equals $\frac{1}{\nu-2}$ for $\nu > 2$. Since this distribution is symmetric, skewness equals 0. Kurtosis equals $6/(\nu-4)$ for $\nu > 4$.

In its canonical form, the density of the $\chi^2(\nu)$ distribution is given by:

$$g(\epsilon) = \frac{1}{2^{\nu/2}\Gamma(\frac{\nu}{2})} \epsilon^{((\nu/2)-1)} e^{-\epsilon/2}. \tag{F.4}$$

The mean of this distribution equals ν and the variance equals 2ν. Skewness and kurtosis are given by $\sqrt{(8/\nu)}$ and $12/\nu$, respectively.

In its canonical form, the density of the Gumbel distribution is given by:

$$g(\epsilon) = e^{-e^{-\epsilon}} e^{-\epsilon}. \tag{F.5}$$

The mean of this distribution equals approximately 0.5772 and the standard deviation equals $\pi/\sqrt{(6)}$. Skewness and kurtosis are equal to 1.14 (approximately) and 5.40, respectively.

The literature offers several versions of a skewed-t distribution; we use one offered by Fernández and Steel (1998).[47] In its canonical form, the density of this distribution is given by:

$$g(\epsilon) = \frac{2}{\gamma + \frac{1}{\gamma}} \left(h\left(\frac{\epsilon}{\gamma}\right) I_{\epsilon<0} + h(\gamma\epsilon) I_{\epsilon\geq 0} \right), \tag{F.6}$$

where $h(.)$ is the density function of a $t(\nu)$ distribution. Moments are found as follows. Define

$$N_j = \int_0^\infty 2\epsilon^j h(\epsilon) d\epsilon$$

where $h(\epsilon)$ is given in equation (F.3). By means of Gradshteyn and Ryzhik (1988,p.295, eq.3.251.11),[48] this solves to

$$N_j = \frac{\nu^{j/2}\Gamma(\frac{j+1}{2})\Gamma(\frac{\nu-j}{2})}{\sqrt{\pi}\Gamma(\frac{\nu}{2})}.$$

The j^{th} moment around 0 is then defined as

$$M_j = N_j \frac{(-1)^j \gamma^j + \frac{1}{\gamma^j}}{\gamma + \frac{1}{\gamma}}.$$

Thus, the mean of the Skewed-$t(\nu.\gamma)$ distribution is given by M_1, the variance by $M_2 - M_1^2$, the third central moment by $M_3 - 3M_2M_1^2 + 2M_1^3$, and the fourth central moment by $M_4 - 4M_1M_3 + 6M_1^2M_2 - 3M_1^4$. Skewness and kurtosis are then straightforwardly derived.

[47] C. Fernández and M.F. Steel. "On bayesian modeling of fat tails and skewness." *Journal of the American Statistical Association*, 1998, 93(441):359-371.

[48] I.S. Gradshteyn and I.M. Ryshik, *Table of Integrals, Series and Products*, 4th ed., 1980, Academic

Table F.1: GTL fit to alternative distributions, by L_1-norm and by simulation with maximum likelihood

	Skewness	Kurtosis	Fitted with L_1-norm α	δ	L_1	Average simulated value α	δ	L_1
Normal	0.00	3.00	0.1436	0.0000	0.0031	0.1356	0.0000	0.0063
$t(5)$	0.00	9.00	-0.0710	0.0000	0.0128	-0.0713	-0.0005	0.0128
$\chi^2(8)$	1.00	4.50	0.2072	-0.2381	0.0261	0.1773	-0.2127	0.0371
Gumbel	1.14	5.40	0.1422	-0.2290	0.0307	0.0736	-0.1992	0.0573
Skewed-$t(5, 0.67)$	-0.83	13.34	0.0576	-0.2059	0.0603	-0.3724	-0.2245	0.1641

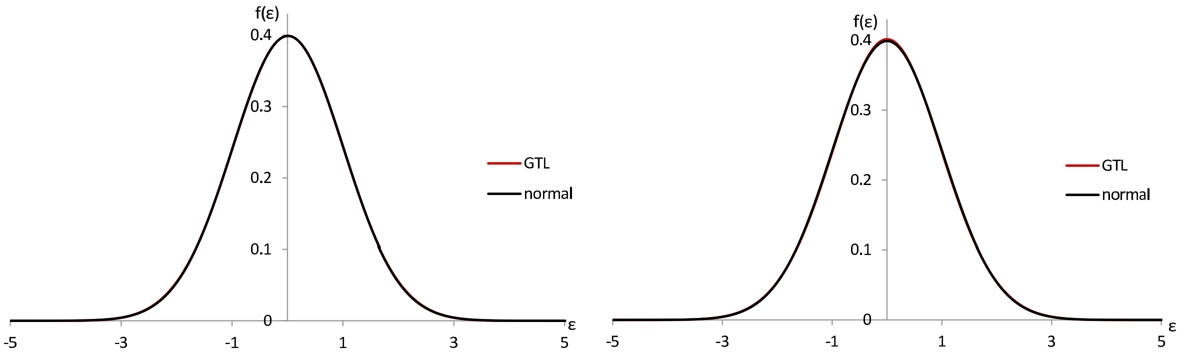
Table F.1 lists the values of (α, δ) of the GTL distribution that is fitted to the distribution listed in each row. The distributions vary from symmetric (normal and $t(5)$) to right-skewed ($\chi^2(8)$, Gumbel and Skewed-$t(5, 0.67)$), and the four non-normal distributions all have higher kurtosis (and thus longer tails) than the normal distribution.

With one exception that will be discussed below, each distribution is specified in its standardized form, with mean 0 and standard deviation 1. The value of the L_1-norm is provided as well. For example, the normal density is fitted with a GTL$(0.1436, 0)$ density with $L_1 = 0.0031$, which means that the total area between the two density curves equals 0.0031, which is only 0.31% of the entire area under the density function. Panel (a) in Figures F.2 to F.6 shows the fit visually. Each panel shows the GTL density in red and the alternative distribution in black; where the red curve is not visible, it is because the black curve overlays it completely. The latter happens for the normal and $t(5)$ densities, for which the fit is very close indeed. The $\chi^2(5)$ density has a truncated left tail, and the fitted GTL density mimics this truncation closely. The Gumbel and Skewed-$t(5, 0.67)$ densities have infinite tails on both sides, but their right skewness is best approximated in the fitted GTL density with a truncated tail on the thin side of the alternative distribution.

Table F.1 also shows the average values of (α, δ) in simulations of the base design with $n = 5000$ and disturbances generated not with a GTL distribution but with each of these alternative distributions. Recall that estimation is done by the maximum likelihood approach, which is fundamentally different from the minimization of the L_1-norm. As

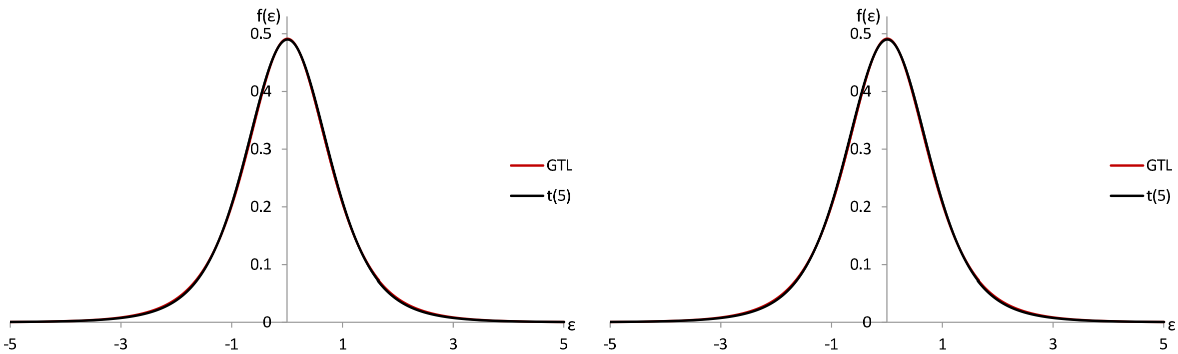
Press, New York.

Figure F.2: Standardized GTL and normal densities



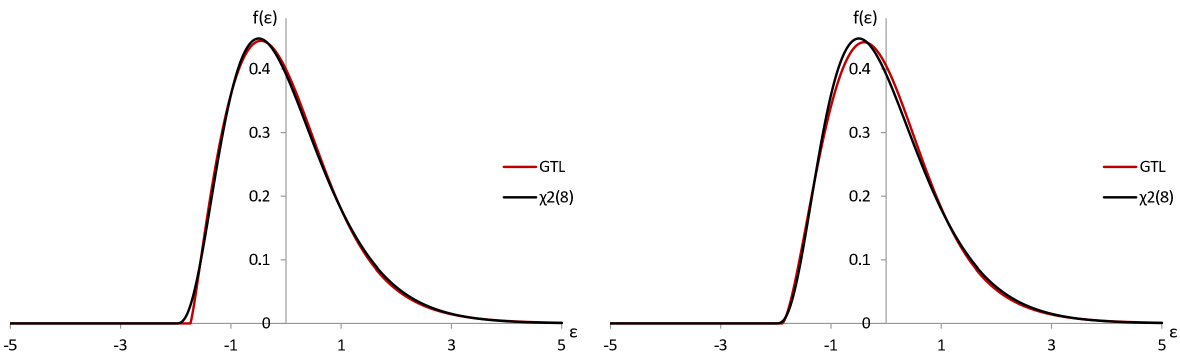
(a) GTL density fitted with L_1 norm (b) GTL density at average simulated (α, δ)

Figure F.3: Standardized GTL and $t(5)$ densities



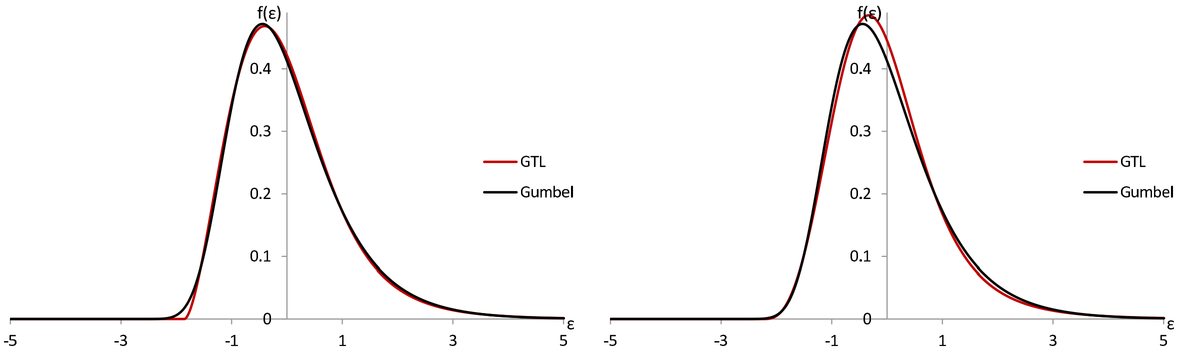
(a) GTL density fitted with L_1 norm (b) GTL density at average simulated (α, δ)

Figure F.4: Standardized GTL and $\chi^2(8)$ densities



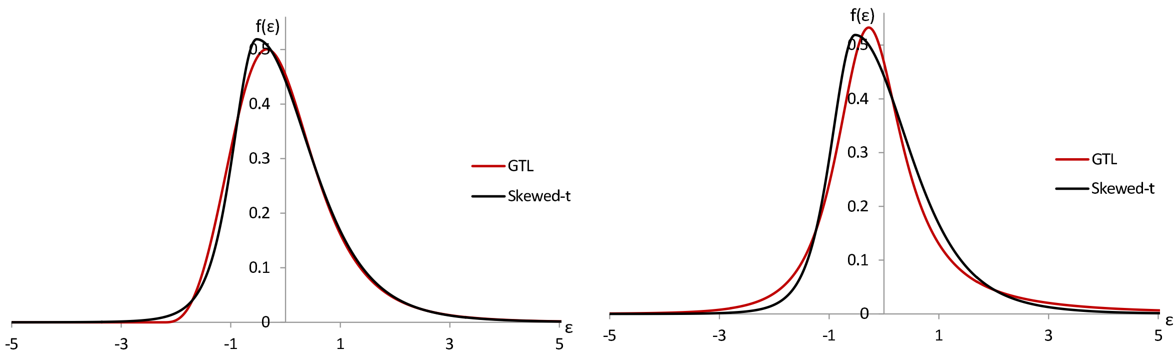
(a) GTL density fitted with L_1 norm (b) GTL density at average simulated (α, δ)

Figure F.5: Standardized GTL and Gumbel densities



(a) GTL density fitted with L_1 norm (b) GTL density at average simulated (α, δ)

Figure F.6: Standardized GTL and Skewed-$t(5, 0.67)$ densities



(a) GTL density fitted with L_1 norm (b) GTL density at average simulated (α, δ)

Figures F.5a and F.6a show, the fit with the L_1-norm allows the fitted GTL density to equal 0 for some value of ϵ where the Gumbel or Skewed-t density is positive. In contrast, with maximum likelihood estimation, each observed value of ϵ must be associated with a positive density mass around that value. In a large sample, tail values are likely to occur, which means that estimated GTL densities are likely to have longer tails than L_1-fitted GTL densities.

With normally or $t(5)$-distributed disturbances, the average (α, δ) hardly differ from those obtained by minimizing the L_1-norm. The truncation point of the $\chi^2(8)$ density is closely mimicked by the estimated GTL$(0.1773, -0.2127)$ density (Figure F.4b). When disturbances are Gumbel-distributed, the estimated GTL$(0.0736, -0.1992)$ distribution has longer tails than the L_1-fitted GTL$(0.1422, -0.2290)$ distribution (Figure F.5).

As mentioned above, one of the ten figures pertains to a case where the GTL is not stated in its standardized form. In Figure F.6b, the Skewed-$t(5, 0.67)$ is displayed in standardized form but the estimated GTL$(-0.3724, -0.2245)$ does not have a finite variance (nor skewness or kurtosis) and thus cannot be represented in standardized form. Rather, the GTL distribution is displayed here with location and scale parameters that, conditionally on the estimated $(\alpha, \delta) = (-0.3724, -0.2245)$ are fitted by minimizing the L_1-norm of the difference between the resulting unstandardized GTL density and the standardized Skewed-$t(5, 0.67)$ density. This estimated density has infinite tails, unlike the L_1-fitted GTL$(0.0576, -0.2079)$ density, in line with the nature of the Skewed-t distribution. As Table F.1 and Figure F.6 shows, the area between the two densities widens slightly but the fit is still highly satisfactory.

In GTL-regression, the objective is usually not so much to obtain a good approximation of the shape of the distribution but rather to obtain precisely estimated parameters of the mean (or location) equation. Evidence of this is provided in Table 3 of Section 4.

G Trade Creation and Trade Diversion: Additional Estimation Results

Table G.1: Trade creation and diversion, 1960-2000: Slopes of control variables

	OLS		GTL			
	Estimate	Stan.Err.	Estimate	Stan.Err.	$\frac{\hat{\beta}_{OLS}-\hat{\beta}_{GTL}}{\hat{\beta}_{GTL}}$	$\frac{SE_{GTL}}{SE_{OLS}}$
lpgdpij	0.942	0.012	0.904	0.012	0.043	0.954
lpgdppcij	0.277	0.027	0.268	0.026	0.034	0.948
ldist	-1.078	0.034	-1.014	0.031	0.063	0.915
sachsij	0.214	0.029	0.148	0.027	0.449	0.936
vola3	0.000	0.002	0.000	0.001	-1.280	0.882
floatij	0.097	0.020	0.050	0.017	0.923	0.849
cu	1.212	0.197	1.198	0.199	0.012	1.011
adifsecschool25	0.038	0.025	0.026	0.023	0.447	0.930
adifdensity	0.126	0.014	0.104	0.013	0.205	0.905
adifgdppc	0.060	0.028	0.054	0.026	0.119	0.960
border	0.404	0.152	0.386	0.130	0.047	0.855
islandij	-0.222	0.050	-0.214	0.047	0.039	0.931
landlockij	-0.271	0.046	-0.263	0.044	0.032	0.951
lpareaij	-0.083	0.010	-0.072	0.009	0.147	0.955
lremoteij	1.327	0.086	1.366	0.080	-0.029	0.936
colony	1.123	0.124	1.089	0.104	0.031	0.841
comcol	0.556	0.095	0.561	0.090	-0.007	0.939
comlang	0.277	0.051	0.258	0.046	0.074	0.904

Notes: See Table 4.

Eicher et al. (2012) estimate more elaborate models than the one presented in Table 4. Their preferred model appears to be one with import-export-pair fixed effects in order to account for unobserved time-invariant country-pair heterogeneity (their specification 4 in Table IV). This introduces 7342 dummy variables, which in OLS estimation are dealt with by computing within-pair deviations. GTL estimates are obtained by maximum likelihood, which does not permit within-pair deviations as an estimation shortcut. This implies that, because of the short panel, the inconsistency of the fixed effects estimator spills over to the estimator of all other slopes.[49] Practically, however, estimating another 7342 slopes is not even feasible. As a substitute, we explore a specification with 36 continent import-export dummies, which are likely to pick up at least some of the unobserved time-invariant country-pair heterogeneity. These results are presented in Table G.2 in the Web Appendix, which yields similar conclusions as Table 4, though the magnitude of the estimated trade effects differs.

Table G.2: Trade creation and diversion, 1960-2000: Adding continent import-export pair dummies

	OLS		GTL			
	Estimate	Stan.Err.	Estimate	Stan.Err.	$\frac{\hat\beta_{OLS}-\hat\beta_{GTL}}{\hat\beta_{GTL}}$	$\frac{SE_{GTL}}{SE_{OLS}}$
Trade creation dummy variables						
tc.nafta	-0.353	0.350	-0.011	0.328	32.493	0.937
tc.eu	0.985	0.125	0.783	0.108	0.257	0.861
tc.efta	1.332	0.140	1.161	0.125	0.147	0.893
tc.eea	0.292	0.088	0.350	0.075	-0.168	0.848
tc.caricom	1.950	0.484	1.791	0.424	0.089	0.876
tc.ap	0.867	0.190	0.831	0.173	0.043	0.912
tc.mercosur	1.009	0.305	1.010	0.324	-0.001	1.063
tc.asean	0.435	0.216	0.503	0.183	-0.135	0.843
tc.anzcerta	-0.694	0.313	-0.435	0.259	0.598	0.828
tc.apec	1.417	0.100	1.136	0.090	0.247	0.899
tc.laia	-0.389	0.179	-0.664	0.171	-0.414	0.956
tc.cacm	1.636	0.193	1.541	0.175	0.062	0.905
tc.bilateralPTA	0.158	0.097	0.126	0.092	0.249	0.950

(continued)

[49] A country pair may be in the sample for a maximum of nine period. The average number of times a country pair appears is 5.17.

Table G.2 continued

	OLS		GTL			
	Estimate	Stan.Err.	Estimate	Stan.Err.	$\frac{\hat{\beta}_{OLS}-\hat{\beta}_{GTL}}{\hat{\beta}_{GTL}}$	$\frac{SE_{GTL}}{SE_{OLS}}$
Trade diversion dummy variables						
td.nafta	0.154	0.076	0.121	0.065	0.279	0.845
td.eu	0.790	0.057	0.625	0.052	0.265	0.903
td.efta	0.510	0.063	0.402	0.057	0.268	0.900
td.eea	-0.249	0.047	-0.149	0.042	0.674	0.897
td.caricom	-0.752	0.107	-0.692	0.102	0.088	0.951
td.ap	0.061	0.072	0.057	0.064	0.062	0.886
td.mercosur	-0.013	0.072	-0.042	0.065	-0.687	0.896
td.asean	0.435	0.070	0.395	0.061	0.100	0.878
td.anzcerta	-0.394	0.094	-0.299	0.084	0.315	0.893
td.apec	0.383	0.050	0.271	0.044	0.414	0.888
td.laia	-0.870	0.085	-0.852	0.082	0.022	0.967
td.cacm	-0.323	0.090	-0.233	0.086	0.386	0.958
td.bilateralPTA	-0.237	0.051	-0.219	0.043	0.080	0.847
Non-PTA control variables						
lpgdpij	0.944	0.014	0.911	0.013	0.036	0.938
lpgdppcij	0.321	0.033	0.318	0.031	0.010	0.939
ldist	-0.990	0.051	-0.909	0.044	0.090	0.869
sachsij	0.242	0.029	0.193	0.027	0.256	0.915
vola3	0.000	0.002	0.000	0.001	-2.109	0.877
floatij	0.092	0.020	0.035	0.017	1.635	0.841
cu	1.052	0.217	0.951	0.245	0.105	1.129
adifsecschool25	0.027	0.026	0.003	0.024	7.690	0.910
adifdensity	0.087	0.014	0.066	0.013	0.316	0.915
adifgdppc	0.091	0.029	0.084	0.028	0.080	0.954
border	0.454	0.145	0.380	0.123	0.194	0.850
islandij	-0.161	0.051	-0.145	0.047	0.109	0.927
landlockij	-0.274	0.045	-0.275	0.041	-0.004	0.929
lpareaij	-0.078	0.012	-0.076	0.011	0.025	0.941
lpremoteij	1.309	0.118	1.193	0.111	0.097	0.936
colony	1.097	0.118	1.093	0.102	0.004	0.862
comcol	0.511	0.099	0.442	0.093	0.156	0.933
comlang	0.262	0.052	0.230	0.048	0.138	0.915
σ			0.777	0.014		
α			-0.093	0.009		
δ			0.136	0.005		
log Likelihood	-74136.2		-71867.5			
(Absolute) Average					1.006	0.915

Dependent variable: Log of bilateral imports. The model also includes control variables (reported in Table G.1 in the Web Appendix) and time dummy variables (not reported). Number of observations = 37983. Skewness and kurtosis of OLS residuals equal -0.71 and 4.90; the Jarque-Bera test of normality of the OLS residuals has a p-value of less than 0.001. The Wald test of the GTL estimates of (α, δ) equals 1482.7, rejecting normality with a p-value of less than 0.001. The Vuong test that compares OLS and GTL equals -25.03 in favor of the GTL model with a p-value of less than 0.001. The GTL estimates imply $\kappa_3 = -2.56$ and $\kappa_4 = 61.72$.