Controlling for Outliers in Efficiency Analysis: A Contaminated Normal-Half Normal Stochastic Frontier Model

Phill Wheat, Alexander D. Stead and William H. Greene

Abstract

Whilst the key innovation of stochastic frontier analysis over deterministic frontier analysis was controlling for the presence of normally distributed noise when estimating production and cost frontiers, in many practical applications, we find a disproportionate number of observations appear within the tails of the efficiency distribution, resulting in implausibly low efficiency predictions. The presence of outliers can be conceptualised in terms of heteroskedasticity, in which for a minority of observations identified by a latent indicator variable v is drawn from a distribution with higher variance. This leads us to propose a robust stochastic frontier model in which v is drawn from a mixture of two normal distributions, often referred to as a contaminated normal distribution. The model handles outlying observations in a different way than the standard model, since the change in efficiency prediction in the low noise regime is offset by an increasing probability that the observation belongs to the high noise regime, which can result in a non-monotonic relationship between the residuals and the resulting efficiency predictions. We apply the model to real datasets, and consider how to test down to the standard model.

P. Wheat • A.D. Stead
Institute for Transport Studies, University of Leeds, Leeds, LS2 9LJ, UK
e-mail: a.d.stead@leeds.ac.uk

W.H. Greene
Department of Economics, Stern School of Business, New York University, New York City, New York, 10012, USA

1. Introduction

This paper is concerned with the problem of how to deal appropriately with noise in the stochastic frontier model, when it is suspected that there are observations which have substantial measurement error in the dependent variable. This often occurs when the dependent variable is data is collected from multiple decision making units, and particularly when the dependent variables relates to an accounting measure, such as cost, revenue or profit. Multiple decision making units could be a set of regulated companies or from a set of regional government departments. In all cases, the analyst is relying on agents within the companies to provide data in a means which is comparable to that collected from others.

From a statistical perspective this can be thought of as a heteroskedasticity problem, with the variance of the noise term differing across decision making units. However there are no *a priori* variables which are correlated with the variance of the noise error, and so the usual approach of modelling heteroskedasticity as a function of a set of covariates cannot be undertaken.

To account for this issue, we adopt a mixture approach in which the noise variance is a function of a latent indicator variable, so that observations fall into one of two classes or regimes, one with low noise variance and another with high noise variance, with some unobserved probability. The model formulation is similar to the zero inefficiency stochastic frontier model of Kumbhakar et al. (2013) with the key difference that the variance of the noise term, rather than that of the inefficiency term, differs between regimes. All other parameters are restricted to be the same across the two regimes. Another way of conceptualising the model is that we are assuming a heavy tailed noise distribution, in common with recent proposals by Stead et al. (2017 ; 2018), since a mixture of normal distributions – commonly referred to as a contaminated normal distribution – in which both normal components have identical means and differing variances has excess kurtosis.

The structure of this paper is as follows. Section 2 provides a brief literature review on outliers, focusing on the context of SFA, Section 3 describes the contaminated normal-half normal model and considers efficiency prediction and testing against the standard stochastic frontier, Section 3 provides applications to the datasets of Christensen and Greene (1976) electricity costs and the Caves et al. (1984) on airline costs. Finally, we give our conclusions in Section 5.

2. Literature Review

The standard stochastic frontier model applied to cross sectional data (Aigner et al., 1977 ; Meeusen and van Den Broeck, 1977) can be written as:

$$y_i = \boldsymbol{x_i}'\boldsymbol{\beta} + \varepsilon_i \quad i = 1, \dots, N$$
$$\varepsilon_i = su_i + v_i \tag{1}$$

Where y is the dependent variable, $\boldsymbol{x}$ is a vector of covariates and $\boldsymbol{\beta}$ the associated vector of parameters, $u_i \sim |N(0, \sigma_u^2)|$ represents inefficiency, $v_i \sim N(0, \sigma_v^2)$ represents statistical noise, and for a production frontier $s = 1$, while for a cost frontier $s = -1$. N is the sample size. The standard model therefore explicitly controls for noise, and predictions for u based on the distribution of $u_i|\varepsilon_i$, as proposed by Jondrow et al. (1982) allows the analyst to decompose the estimated residual ε into noise and inefficiency. However, given the assumption of normality of v_i, the model does not allow for heavy tails in the noise distribution, i.e. outliers in the data.

An extensive literature review of methods of handling outliers in the context of stochastic frontier analysis (SFA) is provided by Stead et al. (2017b), which possibilities such as the detection and removal of outliers, the use of alternative efficiency predictors or efficiency distributions, heteroskedastic stochastic frontier (SF) models, and alternative noise distributions. The authors argue that the latter method is the best way of explicitly allowing for the presence of outliers in the data and alleviating their effects, i.e. inflated inefficiency variance and insufficient shrinkage of efficiency predictions at the tails.

Previous proposals for alternative noise distributions in SFA include the Student's *t* distribution (Tancredi, 2002), the Cauchy distribution (Nguyen, 2010 ; Gupta and Nguyen, 2010) and the Laplace distribution (Nguyen, 2010 ; Horrace and Parmeter, 2015). In all of the aforementioned studies, the proposed noise distribution is paired with its left truncation, and the convolution of the noise and inefficiency distribution has a closed form. In contrast, the convolutions of these noise distributions with some of the more commonly used inefficiency distributions in SFA, e.g. the half normal or exponential distributions, do not have closed forms. Stead et al. (2017 ; 2018) therefore propose the use of maximum simulated likelihood methods – see Train (2009), or Greene (2003) for an explanation in the context of SFA – to pair the canonical inefficiency distributions with the logistic and Student's *t* noise distributions, respectively, allowing for more direct comparison to the standard SF model; the advantages of the Student's *t* distribution in particular is its flexibility as its kurtosis may vary with the degrees of freedom parameter, and its ability to nest the normal distribution as this parameter approaches infinity.

By assuming heavy tailed noise distributions in this way, the SF model is made more robust to outliers, and in particular the efficiency predictions for outlying observations in either direction can be substantially different from those obtained from the standard SF model. As alluded to in the previous section, an alternative way of accounting for outliers is in terms of heteroskedasticity in the noise term. A number of studies have explored the issue of heteroskedasticity in the context of SF models, mainly introducing heteroskedasticity in the inefficiency term, the variance of which is modelled as a function of a vector of covariates, as in Reifschneider and Stevenson (1991), Caudill and Ford (1993) and Caudill et al. (1995). Hadri (1999) introduces a doubly heteroskedastic SF model, in which the noise variance is likewise modelled as a function of a set of covariates.

However, we do not actually observe the extent of misreporting, or have any other variable capturing the extent to which an observation is outlying. This can be seen as a latent variable problem, in which the variable determining the noise variance is not observed. A simple way to model this would be to imagine that we have two classes, or regimes, to which observations may belong: one with a lower variance, and another with a higher variance, membership being denoted by a latent dummy variable. Outside of the context of SFA, Huber (1964) introduced the contamination model, in which a proportion of observations are drawn from a contaminating distribution, as a robust alternative to OLS; in our context the contaminating distribution is simply a higher variance distribution, such that the distribution of v is

$$f_v(v) = p\frac{1}{\sigma_{v1}}\phi\left(\frac{v}{\sigma_{v1}}\right) + (1-p)\frac{1}{\sigma_{v2}}\phi\left(\frac{v}{\sigma_{v2}}\right) \quad (2)$$

Where σ_{v1} and σ_{v2} are the standard deviations of the normal components, and p is the mixing proportion, or the probability that an observation belongs to the first regime. In other words, v is drawn from a mixture of normal distributions, also known as a contaminated normal distribution, or more specifically in this case the scale contaminated normal distribution. This distribution, providing that the variances of the two normal components differ, always has excess kurtosis, and we can therefore also see this model as a further addition to the literature on introducing heavy tailed distributions for v.

As discussed above, recent developments in this literature have been the pairing of heavy tailed noise distributions with any given inefficiency distribution via simulation (Stead et al., 2018), and the proposal of a Student's t distribution for v, which allows the degree of kurtosis to vary and nests the normal distribution (Stead et al., 2017). The model proposed in this study, in which v follows a contaminated normal distribution, retains these advantages: since the contaminated normal distribution nests the normal distribution when $\sigma_{v1} = \sigma_{v2}$ or p is equal to zero or one, the model nests the standard SF model, and kurtosis is a function of all three parameters. In addition, since it is clear that if the convolution of a normal distribution and a given inefficiency distribution has a closed form then the convolution of a contaminated normal distribution with that inefficiency distribution will also have a closed form, the contaminated normal distribution has the advantage that its pairing with many of the proposed distributions for u does not require simulation; the derivation of likelihood functions, efficiency predictors, etc. is also straightforward.

3. Model

The model we propose is similar to the standard model in. Given that u is drawn from a one sided distribution the probability density function for ε, suppressing the i subscript, is given by the convolution

$$f_\varepsilon(\varepsilon) = \int_0^\infty f_v(\varepsilon + su) f_u(u) du \tag{3}$$

Where f_u is the probability density function for u. Where f_v is as given in (2) it is clear that this becomes

$$f_\varepsilon(\varepsilon) = p \int_0^\infty \frac{1}{\sigma_{v1}} \phi\left(\frac{\varepsilon + su}{\sigma_{v1}}\right) f_u(u) du + (1-p) \int_0^\infty \frac{1}{\sigma_{v2}} \phi\left(\frac{\varepsilon + su}{\sigma_{v2}}\right) f_u(u) du \tag{4}$$

That is, the weighted sum of the convolutions of normally distributed random variables with u. Therefore, for any given distribution of u, the integrals simply give the probability density function for ε when v is normally distributed. In the normal-half normal case, this is given by Aigner et al. (1977), and substituting this in gives

$$f_\varepsilon(\varepsilon) = \sum_{j=1}^{2} p_j \frac{2}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}} \phi\left(\frac{\varepsilon}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}}\right) \Phi\left(-\frac{s\varepsilon\, \sigma_u / \sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}}\right) \tag{5}$$

Where $p_1 = p$ and $p_2 = 1 - p$. Note that, due to the additive nature of the likelihood function, the log-likelihood function is not linearised in contrast to those of other SF models, and therefore the first order conditions are somewhat cumbersome: these are given in the appendix. While we focus on the contaminated normal-half normal case, it is straightforward to derive the log-likelihood function when u follows some other distribution in the same way, by simply substituting in the probability density function for ε when v is normally distributed.

3.1. Testing Against the Standard SF Model

If $\sigma_{v1} = \sigma_{v2}$ or $p_j = 0\ \forall\ j = 1{,}2$, then the model reduces to the normal-half normal model. Clearly, it is desirable to test either of these null hypotheses in order to assess the performance of the model against the standard SF model. However, under either null hypothesis the parameter of interest in on a boundary, and there is also an unidentified nuisance parameter. The distribution of the likelihood ratio statistic in these cases is therefore uncertain.

This is a familiar issue in latent class modelling generally, and therefore testing to determine the appropriate number of classes in this way is usually done using some information theoretic or bootstrapping approach instead. Nylund et al. (2007) present evidence using simulated data on the relative effectiveness of each of the commonly used approaches, finding that a bootstrap likelihood ratio test performs well in general, and that the Bayesian Information Criterion (BIC) (Schwarz, 1978) outperformed other information criteria. In this paper, we use the Akaike Information Criteria (AIC) and BIC to inform model selection.

3.2. Conditional probability and efficiency predictors

The most commonly used efficiency predictor in the SF literature is the mean of the distribution of $u|\varepsilon$ as proposed by Jondrow et al. (1982). This is derived by solving the integral

$$E(u|\varepsilon) = \frac{\int_0^\infty u f_v(\varepsilon + su) f_u(u) du}{f_\varepsilon(\varepsilon)} \tag{6}$$

And similar to the likelihood function, given f_v this is simply a weighted sum of means as if v was normally distributed, as given in Jondrow et al. (1982). This gives

$$E(u|\varepsilon) = \sum_{j=1}^{2} \frac{p_j \sigma_u \sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}} \left[\frac{\phi\left(\frac{s\varepsilon\,\sigma_u/\sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}}\right)}{\Phi\left(-\frac{s\varepsilon\,\sigma_u/\sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}}\right)} - \frac{s\varepsilon\,\sigma_u/\sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}} \right] \tag{7}$$

Which is the weighted sum of predictions from the two regimes. Each observation is weighted by the same probabilities, since *a priori* observation-specific probabilities are not known. However in practice, the larger $|\varepsilon_i|$, the greater the probability will be that the observation belongs to the regime with the largest noise variance parameter. An alternative is therefore to weight each regime by conditional probabilities which are observation-specific and reflect this fact. We denote the probability that a specific observation i belongs to the regime j, conditional on ε, as p_i, which is given by:

$$p_{ij} = \frac{p_j \frac{2}{\sqrt{\sigma_{v1}^2 + \sigma_u^2}} \phi\left(\frac{\varepsilon_i}{\sqrt{\sigma_{v1}^2 + \sigma_u^2}}\right) \Phi\left(-\frac{s\varepsilon_i\,\sigma_u/\sigma_{v1}}{\sqrt{\sigma_{v1}^2 + \sigma_u^2}}\right)}{\sum_{j=1}^{2} p_j \frac{2}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}} \phi\left(\frac{\varepsilon_i}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}}\right) \Phi\left(-\frac{s\varepsilon_i\,\sigma_u/\sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}}\right)} \quad (8)$$

In terms of predictors of inefficiency, in line with the literature, particularly, Kumbhakar et al. (2013) we can consider two candidates. The first candidate is to assign each observation a regime post estimation based on whether $p_i > 0.5$ and use the Jondrow et al. (1982) predictor for inefficiency in that regime. The second candidate is to compute a weighted predictor based on the Jondrow et al. (1982) predictors for each of the regimes weighted by the conditional probabilities. Of the two, we favour the latter for two reasons. Firstly, this makes most use of the available information from estimation rather than adopting a threshold approach for which regime the observation belongs. The weighted approached is equivalent to the expectation of the inefficiency conditional on the realised error applied to both the probability of belonging to each regime and the inefficiency distribution within each regime. This is in the spirit of the original proposal of using the conditional expectation as a predictor of inefficiency by Jondrow et al (1982). Secondly, the motivation for the model is not to pigeonhole individual DMUs into one regime or another, instead it is to allow the probability that an observation belongs to the noisy regime—i.e. that it is an outlier—to vary. Thus our preferred inefficiency predictor is:

$$E(u_i|\varepsilon_i) = \sum_{j=1}^{2} \frac{p_{ij}\sigma_u\sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}} \left[\frac{\phi\left(\frac{s\varepsilon_i\,\sigma_u/\sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}}\right)}{\Phi\left(-\frac{s\varepsilon_i\,\sigma_u/\sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}}\right)} - \frac{s\varepsilon_i\,\sigma_u/\sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}} \right] \quad (9)$$

One important implication of the inefficiency predictor is that inefficiency in this model is not a monotonic function of the estimated residual, in contrast to the ALS model. In particular large positive and negative residuals imply greater weight to the noisy regime and as such the predictor of inefficiency is closer to the un-conditional mean of the inefficiency distribution due to the relatively higher noise in that regime. This issue is returned to in the empirical examples.

Wheat et al. (2014) show that minimum width prediction intervals are derived by solving the integral

$$\min_{U,L} U - L + \lambda\left[F_{u|\varepsilon}(L) + 1 - F_{u|\varepsilon}(U) - \alpha\right] \tag{10}$$

And that there are two possible solutions: either $L = 0$, in which case U is given by the quantile function for the distribution of $u|\varepsilon$ evaluated at $1 - \alpha$, or

$$\exists f_{u|\varepsilon}(L) = f_{u|\varepsilon}(U): \int_L^U f_{u|\varepsilon}(u|\varepsilon)du = 1 - \alpha, \qquad L, U > 0 \tag{11}$$

In the former case, it is obvious that the quantile function for the distribution of $u|\varepsilon$ is the weighted sum of that when v is normally distributed: Horrace and Schmidt (1996) give the quantile function for $u|\varepsilon$ for the normal-half normal model, and therefore for the contaminated normal-half normal model, we arrive at

$$U = \sum_{j=1}^{2} \frac{p_j \varepsilon \sigma_u^2}{\sigma_{vj}^2 + \sigma_u^2} + \sum_{j=1}^{2} \frac{p_j \sigma_u \sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}} \Phi^{-1}\left[\alpha \Phi\left(-\frac{\varepsilon\, \sigma_u / \sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}}\right)\right] \tag{12}$$

In the latter case, it is equally clear that the integral in (11) is the weighted sum of the same integrals when $v \sim N(0, \sigma_{v1})$ and $v \sim N(0, \sigma_{v2})$, and therefore the formulae derived by Wheat et al. (2014) become

$$U = \sum_{j=1}^{2} \frac{p_j \varepsilon \sigma_u^2}{\sigma_{vj}^2 + \sigma_u^2} + \sum_{j=1}^{2} \frac{p_j \sigma_u \sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}} \Phi^{-1}\left[\frac{2 - \alpha}{2} \Phi\left(-\frac{\varepsilon\, \sigma_u / \sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}}\right)\right] \tag{13}$$

$$L = \sum_{j=1}^{2} \frac{p_j \varepsilon \sigma_u^2}{\sigma_{vj}^2 + \sigma_u^2} + \sum_{j=1}^{2} \frac{p_j \sigma_u \sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}} \Phi^{-1}\left[\frac{\alpha}{2} \Phi\left(-\frac{\varepsilon\, \sigma_u / \sigma_{vj}}{\sqrt{\sigma_{vj}^2 + \sigma_u^2}}\right)\right]$$

In the contaminated normal-half normal case. Again, in practice we will prefer to replace with the posterior probability p_{ij}.

4. Empirical examples

We now apply the model to two commonly used datasets. The purpose of the use of three (as opposed to one) datasets is to illustrate the features of the model and to demonstrate that the model can be applied in multiple settings. For each dataset we provide the parameter estimates, an efficiency against residual plot and descriptive statistics for the efficiency scores predicted from the normal half normal and latent class model.

Before proceeding with a description of the results for each dataset we offer our experience of estimation of this model using real data. As previously stated we use the latent class estimation routines in LIMDEP v10 to estimate this model and we have replicated the results in Stata[1]. This estimation routine convergence in all empirical cases and all estimations took less than one second on an Intel i7 machine (2.6GHz). However particular note needs to be given to starting values. As a general rule we adopted the parameter estimates for the ALS frontier model as starting values for the frontier parameters ($\boldsymbol{\beta}$) and the estimate of σ_u^2. For $(\sigma_v^2|q = 1,2)$ we adopt $0.95.\hat{\sigma}_v^2$ and $1.05.\hat{\sigma}_v^2$ where $\hat{\sigma}_v^2$ is the estimate of σ_v^2 from the ALS frontier model. For the electricity and airlines data, after trying various other starting values for the variance components, we concluded that these starting values do provide the basis to reach a global maxima.

4.1. U.S. Electricity Data

In this section we apply the contaminated normal-half normal SF model to the Christensen and Greene (1976) dataset on US electricity generation costs; this consists of a cross section of 123 utilities from 1970, and includes data on costs, factor prices and factor shares, and electricity generated. We estimate an augmented Cobb-Douglas cost model including a squared output term:

$$\ln(c_i/e_i) = \beta_0 + \beta_1 \ln q_i + \beta_2(\ln q_i)^2 + \beta_3 \ln(w_i/e_i) + \beta_4 \ln(r_i/e_i) + \varepsilon_i \qquad (14)$$

[11] Code available at http://www.its.leeds.ac.uk/bear

Where c is total cost, q is electricity generated, and w, r and e are prices for labour, capital and energy inputs, respectively. We compare the results from this model to those from the standard normal-half normal SF model in Table 1.

Table 1: Outputs from the contaminated normal-half normal model (electricity data)

	Contaminated Normal-Half Normal			Normal-Half Normal		
	Estimate	s.e.	Sig	Estimate	s.e.	Sig
β_0	3.744	0.011	***	3.735	0.035	***
β_1 ($\ln q_i$)	0.960	0.002	***	0.966	0.013	***
β_2 ($(\ln q_i)^2$)	0.027	0.056	***	0.030	0.003	***
β_3 ($\ln(w_i/e_i)$)	0.327	0.050	***	0.261	0.066	***
β_4 ($\ln(r_i/e_i)$)	0.026	0.036		0.055	0.062	
σ_u	0.138	0.049	-	0.149	0.049	-
σ_{v1}	0.067	0.022	-	0.109	0.023	-
σ_{v2}	0.262	0.079	-	-	-	-
p	0.845	0.096	-	1.000	-	-
Log Likelihood	71.345	-	-	66.865	-	-

Statistical significance at the: * 10% level, ** 5% level, *** 1% level

Table 1 shows that the contaminated normal-half normal and normal-half normal SF models yield very similar estimates of the frontier parameters. The difference, as expected, is in the estimated variance parameters. We estimate a standard deviation for v of 0.067 for the first regime and 0.262 for the second regime, with a mixing proportion of 0.845; compared to an estimated standard deviation for v of 0.109 from the normal-half normal model, this means that most observations are drawn from a lower noise variance regime, with a small proportion drawn from a higher variance regime. We also see that the estimated $VAR(u)$ is smaller than that from the normal-half normal model. The variance of u is given by

$$VAR(u) = \frac{\pi - 2}{\pi}\sigma_u^2 \qquad (15)$$

While the variance of v is given by

$$VAR(v) = \sigma_v^2 \qquad (16)$$

$$VAR(v) = p\sigma_{v1}^2 + (1-p)\sigma_{v2}^2 \qquad (17)$$

In the contaminated normal and normal cases, respectively. The variances of the error components, and of the composed error, is given in Table 2. We can see that, as expected, we obtain a lower estimated $VAR(u)$ and a higher estimated $VAR(v)$ from the contaminated normal-half normal model.

Table 2: Estimated error variances (electricity data)

	Contaminated Normal-Half Normal	Normal-half normal
$\mathrm{VAR}(u)$	0.007	0.008
$\mathrm{VAR}(v)$	0.014	0.012
$\mathrm{VAR}(\varepsilon)$	0.021	0.020

In line with the discussion in Section 2, we would expect given these results to obtain a narrower range of efficiency predictions. We calculate efficiency predictions according to the formula given in (9), and compare these to those obtained from the standard model. Figure 1 shows the relationship between the residuals and efficiency scores in both cases.

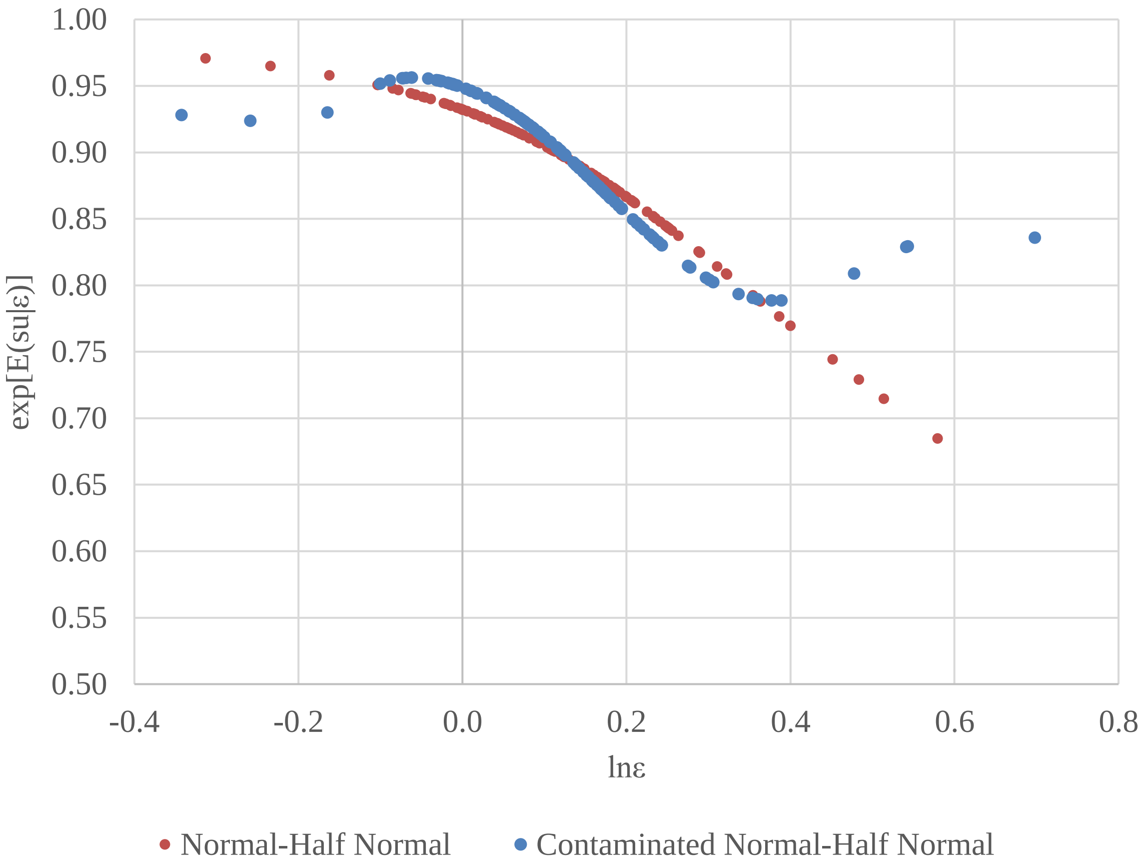

Figure 1: Scatterplot of cost efficiency scores vs residuals (electricity data)

We can see that, as expected, the efficiency predictions from the contaminated normal-half normal model have a considerably narrower range; the mean efficiency prediction was also higher. Specifically, the efficiency predictions from the contaminated normal-half normal model range from 0.789 to 0.956, with a mean of 0.897, while predictions from the normal-half normal model range from 0.685 to 0.971, with a mean of 0.890.

Another important feature of the contaminated normal-half normal model is that the relationship between the efficiency predictions is non-monotonic. For most of the range, efficiency decreases as the residual becomes large, however, for outlying observations this relationship is reversed. Intuitively this is because the larger $|\varepsilon|$ becomes, the more likely it is that an observation is simply an outlier, and the uncertainty around predictions of u increases. Figure 2 shows the relationship between the efficiency predictions and residuals in the first and second (high and low noise) regimes, along with the conditional probability that an observation belongs to the first, and the weighted prediction.

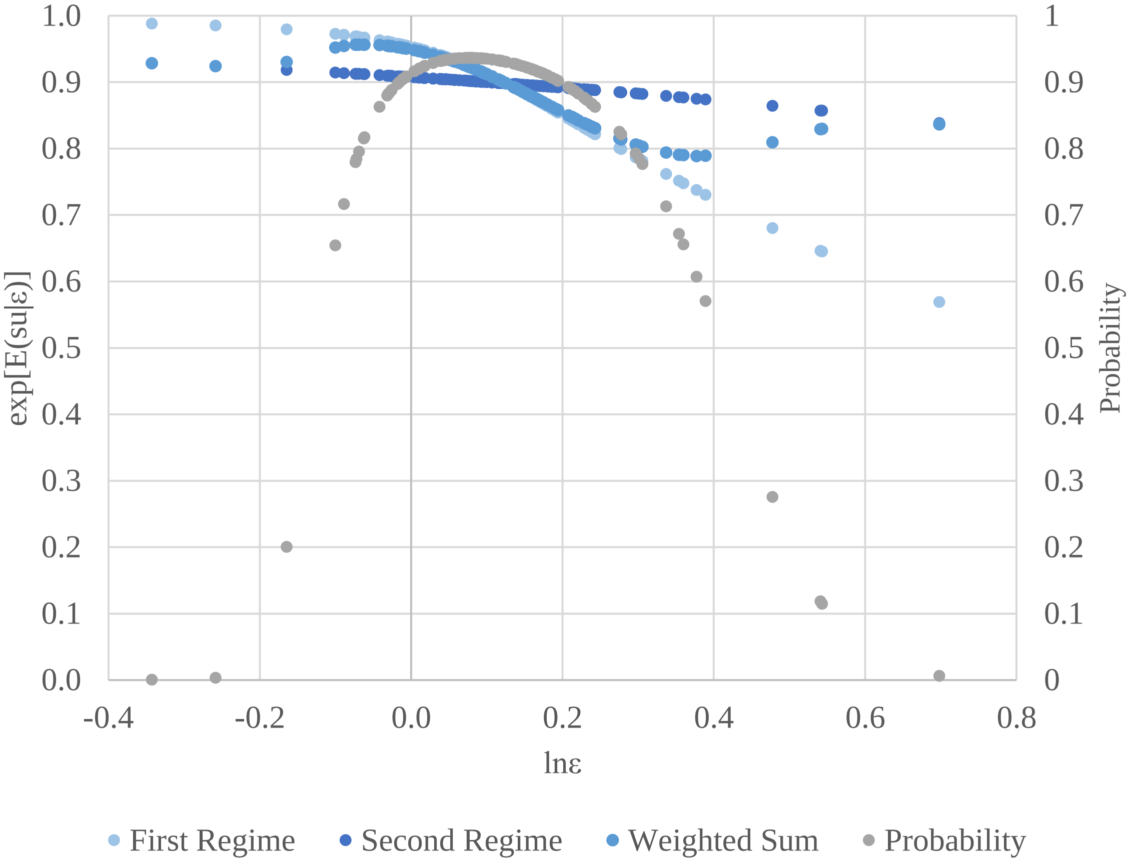

Figure 2: Efficiency scores and probabilities vs residuals (electricity dataset)

From above, we can see that non-monotonicity arises, despite the fact that the relationship in each regime is monotonic, because the probability of belonging to the second, noisy, regime increases with $|\varepsilon|$. The resulting non-monotonic relationship is similar to that found by Stead et al. (2017b) in the Student's *t*-half normal case, except that for the very extreme outliers, we can see that efficiency begins to increase (decrease) again as $s\varepsilon$ increases (decreases); this is because, as the probability of belonging to the second regime approaches one, the relationship continues in its usual direction.

We now consider model selection on the basis of information criteria. Specifically, we use the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), both of which are based on log-likelihood values and including penalties based on the number of parameters in the model. The AIC value is given by

$$AIC = 2k - 2\ln L \quad (18)$$

While the BIC value is given by

$$AIC = k\ln n - 2\ln L \quad (19)$$

Where $\ln L$ is the log likelihood value, k is the number of parameters to be estimated, and n is the number of observations. Therefore we obtain AIC and BIC values of -124.690 and -99.380 from the contaminated normal-half normal model, and -119.922 and -97.424 from the normal half normal. The former model is therefore preferred in terms of both criteria.

4.2. U.S. Airlines Data

The second of the 'benchmark' dataset is the U.S. Airlines dataset which can be traced back to Caves et al (1980) and Trethaway and Windle (1983) and has been used in similar comparison exercises such as Greene (2008) and Horrace and Parmeter (2014), Whilst the dataset is a panel comprising 256 observations (unbalanced panel of 15 years and 25 airlines), we treat this as a pooled cross section in line with uses of this data in the literature, e.g. Horrace and Parmeter (2015). We adopt the 'augmented' Cobb Douglas specification (squared term in output only) and include stage length, number of points served and load factor directly in the frontier following Greene (2008). The model we estimate is:

$$\ln(c_i/p_i) = \beta_0 + \beta_1 \ln q_i + \beta_2 (\ln q_i)^2 + \beta_3 \ln(m_i/p_i) + \beta_4 \ln(f_i/p_i) + \beta_5 \ln(e_i/p_i) + \beta_6 \ln(l_i/p_i) + \beta_7 \ln stages_i + \beta_7 \ln points_i + \beta_7 loadfctr_i + \varepsilon_i \quad (20)$$

Where c is total cost, q is an index of outputs, $stages$ is stage length, $points$ is the number of points served, $loadfctr$ is the load factor, and m, f, e, l and p are input prices for materials, fuel, equipment, labour and property, respectively. The estimated parameters and standard errors are shown in Table 3.

Table 3: Outputs from the contaminated normal-half normal model (airline data)

	Contaminated Normal-Half Normal			Normal-Half Normal		
	Estimate	s.e.	Sig	Estimate	s.e.	Sig
β_0	2.237	0.237	***	2.242	0.275	***
β_1 ($\ln q_i$)	0.980	0.019	***	0.961	0.018	***
β_2 ($(\ln q_i)^2$)	0.031	0.006	***	0.025	0.005	***
β_3 ($\ln(m_i/p_i)$)	0.804	0.098	***	0.774	0.090	***
β_4 ($\ln(f/p_i)$)	-0.009	0.017		-0.022	0.018	
β_5 ($\ln(e_i/p_i)$)	0.086	0.064		0.060	0.064	
β_6 ($\ln(l_i/p_i)$)	0.088	0.093		0.147	0.080	*
β_7 ($\ln stages_i$)	-0.146	0.026	***	-0.160	0.026	***
β_8 ($\ln points_i$)	0.151	0.024	***	0.150	0.026	***
β_9 ($loadfctr_i$)	-1.186	0.190	***	-1.052	0.174	***
σ_u	0.082	0.080	-	0.088	0.031	-
σ_{v1}	0.058	0.045	-	0.088	0.011	-
σ_{v2}	0.138	0.045	-	-	-	-
p	0.680	0.208	-	1.000	-	-
Log Likelihood	221.326	-	-	218.523	-	-

Statistical significance at the: * 10% level, ** 5% level, *** 1% level

The frontier parameter estimates are again similar between the ALS and latent class model. Indeed, the correlation coefficient between the residuals from each model is 0.992. Again, as shown in Table 4 we find that the contaminated normal-half normal model yields a lower estimated $VAR(u)$ and a higher estimated $VAR(v)$ than the normal-half normal case. Note, however, that in this case the estimated total variance of the composed error is greater in the standard model.

Table 4: Estimated error variances (airlines data)

	Contaminated Normal-Half Normal	Normal-half normal
VAR(u)	0.0299	0.0321
VAR(v)	0.0084	0.0078
VAR(ε)	0.0383	0.0399

Again, following from this we expect to see a narrower range of efficiency predictions, and a non-monotonic relationship between the residuals and efficiency predictions observed with the electricity dataset. Figure 3 shows that the range of efficiency predictions is indeed narrower in the contaminated normal-half normal model; specifically, the efficiency predictions range from 0.881 to 0.964, with a mean of 0.937, compared to a range between 0.804 to 0.976, with a mean of 0.932 from the normal-half normal model.

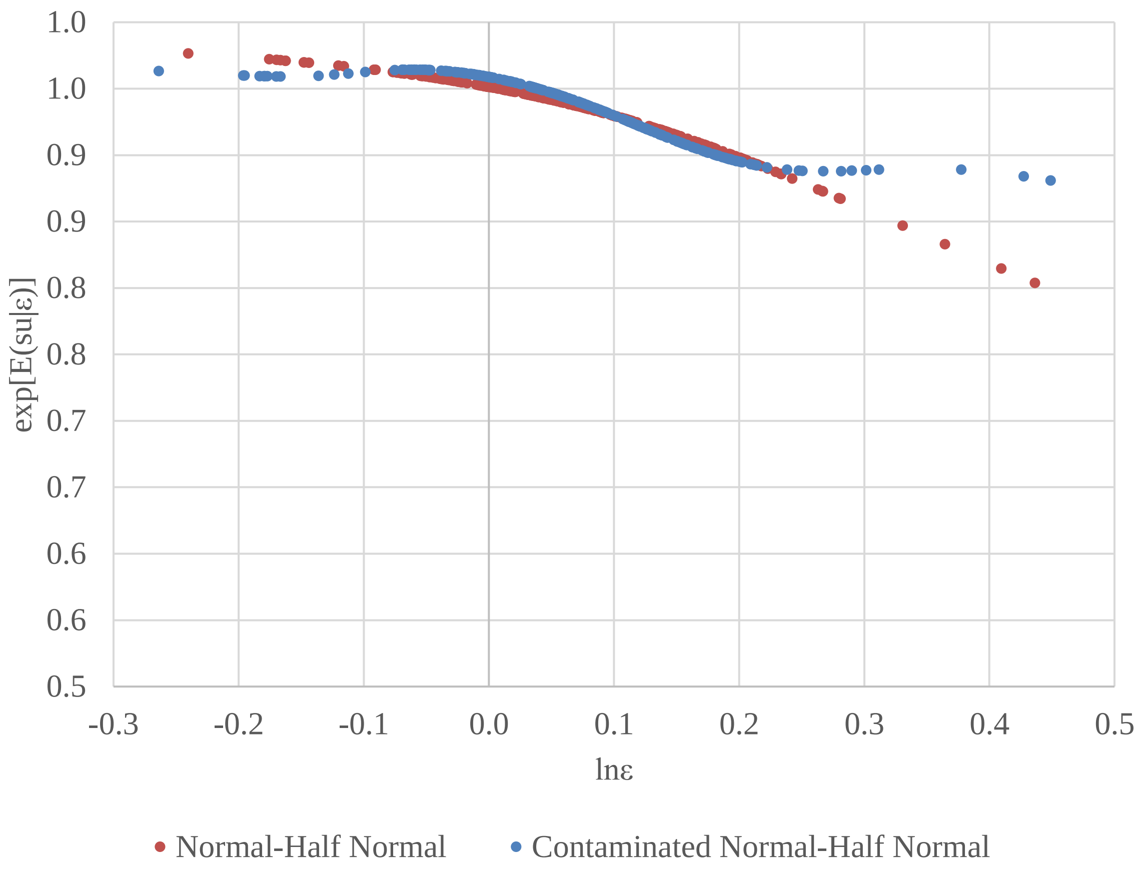

Figure 3: Scatterplot of cost efficiency scores vs residuals (airlines data)

Also as expected, we see a non-monotonic relationship between the residuals and the efficiency predictions as with the electricity data. Similarly Figure 4 shows how this results from the increasing probability that an observation belongs to the noisy regime as $|\varepsilon|$ becomes large.

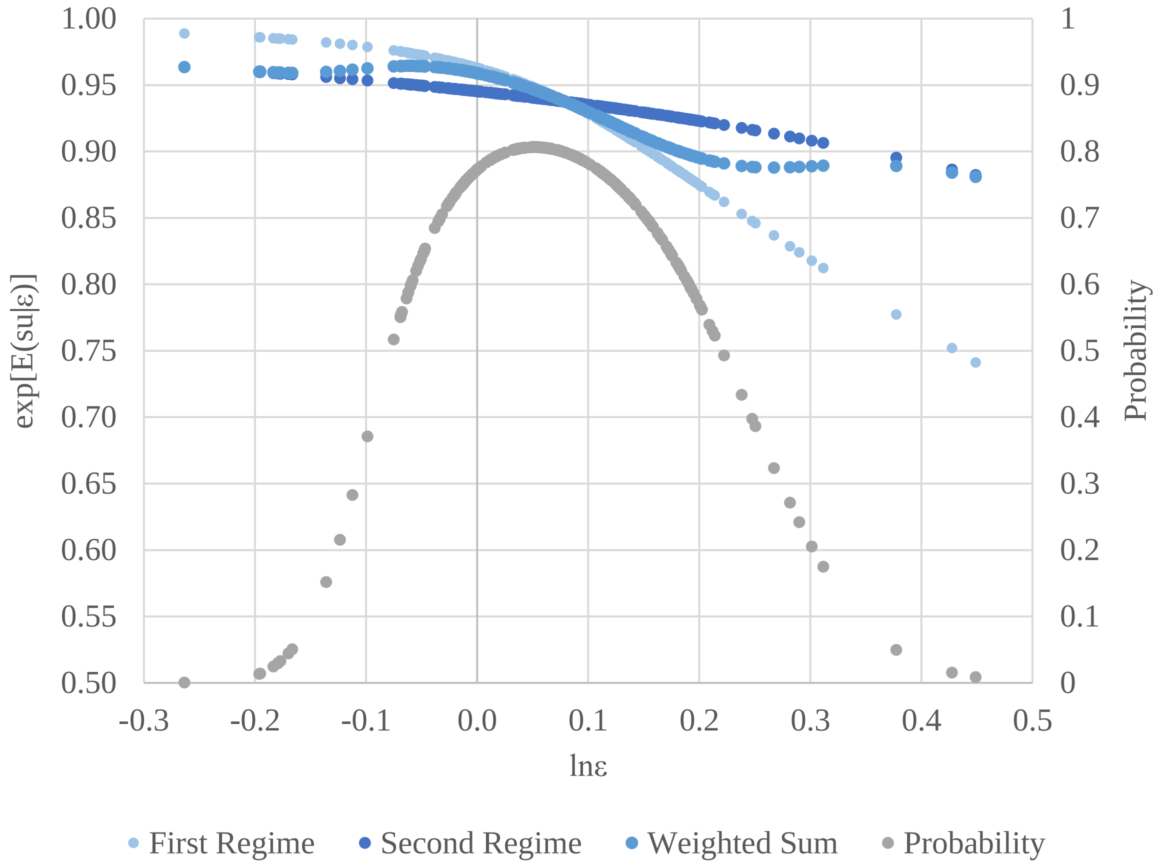

Figure 4: Efficiency Predictions and probabilities vs residuals (airlines data)

Looking again at model selection, using the airlines data we obtain AIC and BIC values of -365.020 and -414.653 from the contaminated normal-half normal model and -370.504 and -413.046 from the normal-half normal model. Therefore the former model is again preferred on the basis of the AIC, although not in this case on the basis of the BIC.

4.3. Synthesis of empirical examples

The two empirical examples demonstrate several important features of our proposed modelling approach. Firstly, the model is implementable using two commercial software packages. Such a finding is important from the perspective demonstrating that this model has potential for wide uptake within the practitioner community.

Secondly, in both examples both the parameter estimates and the majority of the efficiency predictions are similar between the Contaminated Normal-Half Normal SF model and the Normal-Half Normal SF model. The difference lies in the way that observations in which $|\varepsilon|$ is large are handled by the models: in the standard SF model, these are treated as observations in which efficiency is very high or low, while in the contaminated normal-half normal model they are treated as outliers with large $|v|$. That is, those observations with the large residuals are likely to be drawn from a noise distribution with large variance. Hence the 'best' predictor of efficiency is closer to the unconditional mean than would be had the noise been drawn from a tighter distribution. As a result, the Contaminated Normal-Half Normal model yields a more restricted set of efficiency predictions.

Thirdly, efficiency predictions from the contaminated normal-half normal model based on the conditional expectation are no longer monotonic with respect to residuals. This is in contrast to the normal-half normal model. Practically this result means that the contaminated normal-half normal considers that large outliers are primarily generated by noise and as such the conditional expectation of efficiency gravitates toward the unconditional mean of efficiency. That is the best that can be cleaned from this (unreliable data). This potentially has useful incentive properties for overcoming informational asymmetries in economic regulation, particularly in terms of discouraging firms to try and game the process by submitting too favourable data into the process e.g. under reporting costs in the case of cost benchmarking.

Fourthly, we have shown that in both datasets considered, at least one of the model selection criteria support the use of the contaminated normal/half normal model over the simpler normal/half normal model. Thus there are statistical fit reasons for preferring our proposed model in these cases. However, in general we would caution on sole reliance of statistical criteria to select between the two models. Given the similarity in model parameter estimates and the similarity in the majority of the efficiency predictions, we suggest that an important consideration is whether the analyst or key stakeholders are concerned about the extremes of the efficiency distribution. In economic regulation applications, this is often the case since the extreme observations may represent a company subject to the regulatory regime. As such the proposed model, which yields a less extreme distribution of efficiency predictions, may still be an important candidate model for regulatory consideration even if model fit criteria do not indicate superiority. What this is highlighting is that statistical fit criteria relate to all observations, however we are highlighting that there are some observations that appear to have odd (potentially implausible) efficiency predictions in the simple normal/half normal model, which the contaminated rectifies. This issue may or may not be picked up prominently by model fit criteria, it depends on the extent of the extreme observations, in terms of their frequency relative to the sample size and the magnitude of anomalies.

5. Conclusions

This paper explores the literature on handling outliers in stochastic frontier analysis and proposes an alternative SF model. This is a really important issue particularly when using n accountancy measure for the dependent variable such as cost, revenue or profit. Given differing accounting practices, and that accounting data can often be collected from multiple sources for different firms, there can heteroskedasticity associated with noise across observations. We build on recent work by Stead et al. (2017a; 2017b) which uses simulation to combine a heavy tailed distributions for v with commonly-used distributions for u, but propose the contaminated normal distribution for v: this has the advantage that if the convolution of the normal distribution and any given distribution for u has a closed form, then so will the convolution of the contaminated normal distribution and the distribution of u, removing the need for simulation while retaining the advantages of the aforementioned studies, such as allowing for heavy tails in v – the scale contaminated normal distribution always has excess kurtosis – and allowing for various different inefficiency distributions. Furthermore, as with the SF model with Student's t distributed v, the model nests the standard SF model and allows for varying degrees of kurtosis depending upon the estimated scale and mixing parameters of the contaminated normal error component.

We discuss derivation of the model and approaches to obtaining point predictors for efficiency, which are straightforward, and apply the model to estimate cost frontier using two different datasets: the Christensen and Greene (1976) dataset on electricity generation, and the Caves et al. (1984) dataset on airlines. In both applications, we show that the contaminated normal-half normal model yields similar estimates of the frontier parameters, but higher estimated $VAR(v)$ and lower estimated $VAR(u)$ than the standard SF model, and that as a result, the range of estimated efficiency predictions is considerably smaller. We also find, as in the Student's t-half normal model, that there is a non-monotonic relationship between the residuals and efficiency predictions from the contaminated normal-half normal model, as the probability of an observation being an outlier increases with $|\varepsilon|$.

In terms of model selection, we find that the contaminated normal-half normal model is preferred to the standard model on the basis of the AIC and BIC in our electricity application, whereas in our application to the airlines data, we find that the model is preferred to the standard model on the basis of the AIC but not the BIC. Overall, this emphasises the importance of controlling for outliers in SFA.

We envisage that our model is attractive on a regulatory setting where the data refers to many regulated firms under the same regime. As such a credible distribution of efficiency is essential for the model to be used to support robust and transparent incentive setting. We have shown that the model can easily be estimated within two commonly available econometric software packages (and this has been demonstrated here for two datasets) and this coupled with the frontier parameter estimates being found to be comparable with the simpler normal/half normal model, should provide confidence that the model is a potential candidate model for use by practitioners. Further, we highlight that the non-monotonicity of efficiency predictions to residuals may actually have attractive incentive properties in regulatory settings in terms of eliciting truly comparable dependent variable data (e.g. cost, revenue, profit) from regulated firms.

6. References

Aigner, D., Lovell, C.A.K. and Schmidt, P. 1977. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics.* **6**(1), pp.21-37.

Caudill, S.B. and Ford, J.M. 1993. Biases in frontier estimation due to heteroscedasticity. *Economics Letters.* **41**(1), pp.17-20.

Caudill, S.B., Ford, J.M. and Gropper, D.M. 1995. Frontier Estimation and Firm-Specific Inefficiency Measures in the Presence of Heteroscedasticity. *Journal of Business & Economic Statistics.* **13**(1), pp.105-111.

Caves, D.W., Christensen, L.R. and Tretheway, M.W. 1984. Economies of Density versus Economies of Scale: Why Trunk and Local Service Airline Costs Differ. *The RAND Journal of Economics.* **15**(4), pp.471-489.

Christensen, L.R. and Greene, W.H. 1976. Economies of Scale in U.S. Electric Power Generation. *Journal of Political Economy.* **84**(4), pp.655-676.

Greene, W.H. 2003. Simulated Likelihood Estimation of the Normal-Gamma Stochastic Frontier Function. *Journal of Productivity Analysis.* **19**(2), pp.179-190.

Gupta, A.K. and Nguyen, N. 2010. Stochastic frontier analysis with fat-tailed error models applied to WHO health data. *International Journal of Innovative Management, Information & Production.* **1**(1), pp.43-48.

Hadri, K. 1999. Estimation of a Doubly Heteroscedastic Stochastic Frontier Cost Function. *Journal of Business & Economic Statistics.* **17**(3), pp.359-363.

Horrace, W.C. and Parmeter, C.F. 2015. A Laplace Stochastic Frontier Model. *Econometric Reviews.* pp.1-27.

Horrace, W.C. and Schmidt, P. 1996. Confidence statements for efficiency estimates from stochastic frontier models. *Journal of Productivity Analysis.* **7**(2), pp.257-282.

Huber, P.J. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics.* **35**(1), pp.73-101.

Jondrow, J., Knox Lovell, C.A., Materov, I.S. and Schmidt, P. 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics.* **19**(2), pp.233-238.

Kumbhakar, S.C., Parmeter, C.F. and Tsionas, E.G. 2013. A zero inefficiency stochastic frontier model. *Journal of Econometrics.* **172**(1), pp.66-76.

Meeusen, W. and van Den Broeck, J. 1977. Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error. *International Economic Review.* **18**(2), pp.435-444.

Nguyen, N. 2010. *Estimation of Technical Efficiency in Stochastic Frontier Analysis*. PhD thesis, Bowling Green State University.

Nylund, K.L., Asparouhov, T. and Muthén, B.O. 2007. Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling: A Multidisciplinary Journal.* **14**(4), pp.535-569.

Reifschneider, D. and Stevenson, R. 1991. Systematic Departures from the Frontier: A Framework for the Analysis of Firm Inefficiency. *International Economic Review.* **32**(3), pp.715-723.

Schwarz, G. 1978. Estimating the Dimension of a Model. pp.461-464.

Stead, A.D., Wheat, P. and Greene, W.H. 2017. *Robust Stochastic Frontier Analysis: A Student's t-Half Normal Model with Application to Highways Maintenance Costs in England*. Working Paper. University of Leeds, Insitute for Transport Studies.

Stead, A.D., Wheat, P. and Greene, W.H. 2018. Estimating efficiency in the presence of extreme outliers: A logistic-half normal stochastic frontier model with application to highway maintenance costs in England. In: Greene, W.H., Khalaf, L., Makdissi, P., Sickles, R.C., Veall, M. and Voia, M. eds. *Productivity and Inequality*. Springer, pp.1-19.

Tancredi, A. 2002. *Accounting for heavy tails in stochastic frontier models*. Working Paper No. 2002.16. Department of Statistical Sciences, University of Padua.

Train, K.E. 2009. *Discrete Choice Methods with Simulation.* Second ed. Cambridge University Press.

Wheat, P., Greene, W. and Smith, A. 2014. Understanding prediction intervals for firm specific inefficiency scores from parametric stochastic frontier models. *Journal of Productivity Analysis.* **42**(1), pp.55-65.