Semiparametric Smooth-coefficient Stochastic Frontier Model

Kai Sun*[1] and Subal C. Kumbhakar[2]

[1]Aston Business School, Aston University, Birmingham, B4 7ET, UK
[2]Department of Economics, State University of New York at Binghamton, NY 13902, USA

Abstract

This paper proposes a semiparametric smooth-coefficient stochastic production frontier model where regression coefficients are unknown smooth functions of environmental factors, which shift the production frontier non-neutrally. Technical inefficiency enters into the model in the form of a parametric scaling function which also depends on the environmental factors. A residual-based bootstrap test of the relevance of the environmental factors is suggested. Results show that the semiparametric model captures parameter heterogeneity and yields comparable estimates of technical efficiency.

Keywords: Semiparametric smooth-coefficient Model; Stochastic Frontier Model; Environmental Factor

1 Introduction

Following the seminal work of Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977), the literature on the estimation of technical inefficiency using stochastic frontier framework (see Kumbhakar and Lovell (2000) for references) has been growing exponentially. More recently, attention was paid to the modeling of environmental factors (hereafter, Z variables) affecting inefficiency (u). They are the exogenous factors, such as education, age, experience, R&D, etc., in addition to traditional input(s) and output(s) in frontier models. There are many different ways by which the Z variables can explain inefficiency. For example, Kumbhakar (1990) and Battese and Coelli (1992) proposed a multiplicative decomposition of technical inefficiency in a panel model, where $u_{it} = g(t)u_i$, $g(t)$ is a deterministic function of time, and u_i is one-sided random variable assumed to be normally distributed and truncated at zero from below. Alvarez, Amsler, Orea and Schmidt (2006) called this formulation the scaling property of technical inefficiency. The idea of this property was proposed earlier by Simar, Lovell and van den Eeckaut (1994) and further studied by Wang and Schmidt (2002), among others. Alvarez et al. (2006) interpreted the standard truncated normal

*Corresponding author. Address: Aston Business School, Aston University, Aston Triangle, Birmingham, B4 7ET, UK. Email: k.sun@aston.ac.uk. Tel: +44 121 204 3162.

random variable as "the firms' base efficiency level which captures things like the manager's natural skills", but "how well these natural skills are exploited to manage the firm efficiently depends on ... measures of the environment in which the firm operates."[1]

The novelty of this paper lies in the fact that we not only consider the impact of Z variables on the technical inefficiency part, but we also introduce the Z variables into the frontier part in a semiparametric fashion. Specifically, in a production framework, we express the intercept and slope coefficients as unknown functions of the Z variables. This allows the environmental factors to shift the frontier non-neutrally. The advantage of the semiparametric approach over its parametric counterpart is that the regression coefficients are fully flexible and no prior knowledge about the functional forms are required. Meanwhile, Z is still allowed to affect technical inefficiency as some parametric stochastic frontier models can do. This allows one to compare the technologies, including technical efficiencies, for different firms which are linked by the Z variable, say, R&D. In this regard, our proposed model has several advantages over Battese, Prasada Rao and O'Donnell (2004) and O'Donnell, Prasada Rao and Battese (2008) (hereafter, B&O) who suggested a metafrontier framework for the comparison of firms under different technologies: (1) B&O's model is more liable to sample misclassification due to potentially different grouping criteria whereas grouping is not required in our model; (2) B&O's model only yields group-specific estimates while ours is individual-specific; (3) our model yields comparable estimates linked by the Z variables and there is no need to estimate a common metafrontier.

To give more credibility of the inclusion of the Z variables into the model, a residual-based wild bootstrap testing procedure, borrowed from Li and Racine (2010), for the relevance of the environmental factors is proposed. We show that the model under the null of irrelevance of Z is the same as a standard parametric stochastic frontier model without environmental factors. We then apply our proposed methodology in the Norwegian forestry, with a cross-section of 3249 active forest owners. Both standard and semiparametric frontier models are estimated and results are compared.

The rest of the paper is organized as follows. Section 2 presents the estimation procedure of a semiparametric stochastic production frontier model with environmental factors. Section 3 proposes a test for the relevance of the environmental factors. Section 4 applies the method to the Norwegian forestry. Section 5 concludes.

[1] Alternatively, Wang (2002) specified $u_{it} \sim N^+(\mu(Z_{it}), \sigma_u^2(Z_{it})), u_{it} \geq 0$. See also Kumbhakar, Ghosh and McGuckin (1991), Battese and Coelli (1995) and Huang and Liu (1994).

2 Technical Inefficiency in Semiparametric Models

Consider a stochastic production frontier model with the following specification:

$$y_i = \alpha(Z_i) + X_i'\beta(Z_i) + v_i - u(Z_i), \tag{1}$$

where y_i is the log of output, $X_i' = [x_{1i}, \ldots, x_{ki}]$ is a vector of the log of k-inputs, Z_i is a p-vector of environmental factors (e.g., time, R&D, among others), $\alpha(\cdot)$ is the intercept and $\beta(\cdot)$ is a $k \times 1$ parameter vector. Both of them are expressed as unknown functions of Z_i. $v_i \sim iidN(0, \sigma_v^2)$ is the noise term, and $u(Z_i) = \sigma_u(Z_i)\eta_i$, where $\eta_i \sim iidN^+(0,1)$ and $\sigma_u(Z_i) > 0$. We parameterize $\sigma_u(Z_i)$ such that $\sigma_u(Z_i) = \exp(\delta_0 + \delta_1' Z_i)$, to guarantee its positivity. Furthermore, η and v are assumed to be independent of each other and independent of X and Z. These assumptions indicate $E(u(Z_i)|Z_i) = \sigma_u(Z_i)E(\eta_i|Z_i) = \sqrt{2/\pi}\sigma_u(Z_i) = \sqrt{2/\pi}\exp(\delta_0 + \delta_1' Z_i)$.

For estimation we rewrite (1) as:

$$\begin{aligned} y_i &= \alpha(Z_i) + X_i'\beta(Z_i) + v_i - (u(Z_i) - E(u(Z_i)|Z_i)) - E(u(Z_i)|Z_i) \\ &= \theta(Z_i) + X_i'\beta(Z_i) + \varepsilon_i \end{aligned} \tag{2}$$

where $\theta(Z_i) = \alpha(Z_i) - E(u(Z_i)|Z_i)$, and $\varepsilon_i = v_i - (u(Z_i) - E(u(Z_i)|Z_i))$. The model in (2) can then be consistently estimated as a semiparametric smooth coefficient model (Li, Huang, Li and Fu 2002).

Define $\rho(Z_i) = [\theta(Z_i),\ \beta'(Z_i)]$, and $W_i' = [1,\ X_i']$, and (2) becomes $y_i = W_i'\rho(Z_i) + \varepsilon_i$. Using the population moment condition $E(W_i\varepsilon_i|Z_i) = 0$[2] gives:

$$\rho(Z_i) = \left[E(W_i W_i'|Z_i)\right]^{-1} E(W_i y_i|Z_i). \tag{3}$$

Then, use the Nadaraya-Watson estimator (Li and Racine 2007) for the conditional expectations, viz., $E(W_i W_i'|Z_i)$ and $E(W_i y_i|Z_i)$, and the smooth coefficient estimator can be written as:

$$\hat{\rho}(Z_i) = \left[\sum_{j=1}^{n} W_j W_j' K\left(\frac{Z_j - Z_i}{h}\right)\right]^{-1} \sum_{j=1}^{n} W_j y_j K\left(\frac{Z_j - Z_i}{h}\right), \tag{4}$$

where n is sample size, $K(\cdot)$ is product kernel function, and h is a p-vector of bandwidth, which can be selected via least-squares cross-validation method (Li and Racine 2010). We use the consistent estimators of

[2]This is because $E(\varepsilon|X,Z) = E(v - u + E(u|Z)|X,Z) = E(v|X,Z) - E(u|X,Z) + E(E(u|Z)|X,Z) = 0 - E(u|Z) + E(u|Z) = 0$.

$\theta(Z_i)$ and $\beta'(Z_i)$ to estimate $\alpha(Z_i)$ and $u(Z_i)$ in the second step in which we make use of the distributional assumptions on v_i and η_i.

In the second step we use the residuals from (2). Recall that $\varepsilon_i = v_i - u(Z_i) + E(u(Z_i)|Z_i)$ where $E(u(Z_i)|Z_i) = \sqrt{2/\pi}\sigma_u(Z_i)$, and therefore, the estimating equation for the second step of the estimation is:

$$\begin{aligned} \varepsilon_i &= \sqrt{2/\pi}\sigma_u(Z_i) + v_i - \sigma_u(Z_i)\eta_i \\ &= \sqrt{2/\pi}\exp(\delta_0 + \delta_1' Z_i) + v_i - \exp(\delta_0 + \delta_1' Z_i)\eta_i \end{aligned} \tag{5}$$

The standard stochastic frontier estimation technique (maximum likelihood) can be applied in this step. Define $\varepsilon_i^* = v_i - \exp(\delta_0 + \delta_1' Z_i)\eta_i = \varepsilon_i - \sqrt{2/\pi}\sigma_u(Z_i)$, the log-likelihood function can be written as:

$$\ln L = Constant - \frac{1}{2}\sum_i \ln\left[\sigma_u^2(Z_i) + \sigma_v^2\right] + \sum_i \ln\Phi\left(-\frac{\varepsilon_i^*\lambda_i}{\sigma_i}\right) - \frac{1}{2}\sum_i \frac{\varepsilon_i^{*2}}{\sigma_i^2}, \tag{6}$$

where $\sigma_u^2(Z_i) = \exp[2(\delta_0 + \delta_1' Z_i)]$, $\sigma_i^2 = \sigma_v^2 + \sigma_u^2(Z_i) = \sigma_v^2 + \exp[2(\delta_0 + \delta_1' Z_i)]$, and $\lambda_i = \sigma_u(Z_i)/\sigma_v = \exp(\delta_0 + \delta_1' Z_i)/\sigma_v$. Maximization of the above log-likelihood[3] will give estimates of δ_0, δ_1, and σ_v^2, which can be used to obtain $\sigma_u^2(Z_i)$ and therefore $E(u_i|Z_i) = \sqrt{2/\pi}\sigma_u(Z_i)$. We use this to estimate the intercept in (1) as $\alpha(Z_i) = \theta(Z_i) + E(u(Z_i)|Z_i)$.

Finally, we use the Battese and Coelli's (1988) technique to estimate technical efficiency, viz,

$$TE_i = E[\exp(-u(Z_i))|\varepsilon_i^*] = \frac{\Phi(\mu_{*i}/\sigma_{*i} - \sigma_{*i})}{\Phi(\mu_{*i}/\sigma_{*i})} \cdot \exp(-\mu_{*i} + 0.5\sigma_{*i}^2), \tag{7}$$

where $\mu_{*i} = -\varepsilon_i^*\sigma_u^2(Z_i)/\sigma_i^2$, $\sigma_{*i}^2 = \sigma_u^2(Z_i)\sigma_v^2/\sigma_i^2$.

3 Testing for the Relevance of Environmental Factors

The environmental factors, Z_i, shift the production frontier (both intercept and slopes) as well as technical inefficiency. One naturally wants to test if Z_i matters, that is, to test if (1) can be estimated as a standard stochastic frontier model:

$$y_i = \alpha + X_i'\beta + \nu_i - \mu_i, \tag{8}$$

where ν_i is the normal noise term, and μ_i is the half-normal technical inefficiency term. ν and μ are independent of each other and of X. In this model, neither the coefficients nor the technical inefficiency

[3]Since ε_i is not observed we follow the standard practice and use the residuals from (2) in (6).

vary with Z_i. This is the same as testing whether the parameters in (9) are constants, viz.,

$$y_i = \theta + X_i'\beta + \epsilon_i = W_i'\rho + \epsilon_i, \tag{9}$$

where $\theta = \alpha - E(\mu_i)$, $\rho' = [\theta,\ \beta']$, and $\epsilon_i = \nu_i - (\mu_i - E(\mu_i))$. The null hypothesis can be stated as $H_0 : \rho(Z_i) = \rho$.[4] Following Li and Racine (2010), the consistent model specification test statistic is constructed as:

$$\hat{I}_n = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} W_i'W_j\hat{\epsilon}_i\hat{\epsilon}_j K\left(\frac{Z_i - Z_j}{h}\right) \tag{10}$$

where $K(\cdot)$ is the product kernel function, $\hat{\epsilon}_i = y_i - \hat{\theta} - X_i'\hat{\beta}$ is obtained from the parametric model (9) via OLS. We follow Li and Racine's (2010) residual-based wild bootstrap method to determine whether to reject the null hypothesis or not:

Step 1: Estimate (9), obtain $\hat{\rho}$ and $\hat{\epsilon}_i$, and generate wild bootstrap disturbance $\epsilon_i^\star$;

Step 2: From $\epsilon_i^\star$, generate $y_i^\star = W_i'\hat{\rho} + \epsilon_i^\star$;

Step 3: Use $\{y_i^\star, W_i\}_{i=1}^n$ to estimate the parametric model (9), and obtain $\hat{\rho}^\star$, and $\hat{\epsilon}_i^\star = y_i^\star - W_i'\hat{\rho}^\star$;

Step 4: The bootstrap statistic $\hat{I}_n^\star$ is obtained from (10), replacing $\hat{\epsilon}_i\hat{\epsilon}_j$ by $\hat{\epsilon}_i^\star\hat{\epsilon}_j^\star$.

Step 5: Repeat Steps 1-4 a large number of times, say $B = 399$ times, and calculate the p-value: $p = \frac{1}{B}\sum_{b=1}^{B}\mathbf{I}(I_n^* > I_n)$, where $\mathbf{I}(\cdot)$ is the indicator function with a value of 1 if the statement in the parenthesis is true.

Note that $y_i^\star$ is generated under the null hypothesis, and therefore, the p-value is the size of the test. The null hypothesis can be rejected if the p-value is less than the level of significance, say 0.05.

4 An Empirical Application

In this section, we consider estimation of stochastic production frontier in the Norwegian forestry. The data, compiled by *Statistics Norway*, were drawn from a cross-section of 3249 active forest owners. All data are for the year 2003. The output variable consists of annual timber sales from the forest, measured in cubic meters. The labor input variable is the sum of hours worked by contractors and hours worked by the owner, his family or hired labor in 2003. The land input variable measures forest area cut in hectares, which is the area of various types of final fellings in 2003. The capital input variable is the value of timber stock that can be cut without affecting future harvesting. Our choices of the environmental factors are: (1) income

[4]Constant ρ implies constant θ and β, and constant θ implies constant α and $E(\mu_i)$, assuming $\alpha \neq E(\mu_i)$.

from outfield-related productions (i.e., recreational services), (2) income from agriculture, (3) wage income, (4) a binary variable with a value of 1 indicating there is a management plan, and 0 otherwise, (5) a binary variable with a value of 1 indicating the forest owner has an education level of Bachelor or higher, and 0 otherwise, (6) a binary variable with a value of 1 indicating its properties are located in central municipalities, and 0 otherwise. Lien, Størdal and Baardsen (2007) used this data to assess technical inefficiency of these Norwegian forests. Table 1 presents summary statistics in the sample. Further details on the source of the data and definitions of the variables were provided in their study.

Table 1: **Summary Statistics of the Variables**

Symbol	Variable Name	Mean	Sd.	Min.	Max.	Bandwidth[1]
y	Log of output (Harvesting level)	5.6680	1.665462	0.6931	10.74	-
x_1	Log of labor (Working hours)	2.882	1.637169	-2.072	7.876	-
x_2	Log of land (Forest area cut)	0.6692	1.522922	-4.4240	5.434	-
x_3	Log of capital (Value of timber stock)	11.780	1.184578	7.297	16.6	-
Z_1	Income from outfield related productions (1000NOK)	70.98	467.0222	0.00	11810	27.94376593
Z_2	Income from agricullture (1000NOK)	54.21	125.9468	0.00	2488	9.94951468
Z_3	Wage income (1000NOK)	240.3	269.1531	0.00	2183	122.03785955
Z_4	Management plan (0/1)	0.6898	0.4626668	0.00	1.00	0.21638380
Z_5	Education, Bachelor or higher (0/1)	0.2416	0.4281267	0.00	1.00	0.45613206
Z_6	Centrality (0/1)	0.3764	0.4845628	0.00	1.00	0.01324032

1. The bandwidths are selected via least-squares cross-validation.

We consider three specifications for the stochastic production frontier: (1) the semiparametric smooth-coefficient stochastic frontier model as described in (1) (i.e., with environmental factors which enter the coefficients and inefficiency), (2) the standard parametric stochastic production frontier model as described in (8) (i.e., without any environmental factors), and (3) a parametric stochastic production frontier model with environmental factors affecting technical inefficiency only. Technical efficiencies are calculated from all these models using $TE_i = E[\exp(-u_i)|\varepsilon_i^*]$ (see Kumbhakar and Lovell (2000), p. 78 for the exact formula).

The average estimated technical efficiency is 0.97 for the semiparametric model, 0.86 for the standard parametric model without Z, and 0.98 for the parametric model with Z. These results are comparable to Lien et al. (2007) who found the average technical efficiency to be 0.90 using a different model. The semiparametric (parametric without Z) model shows that about 4% (6%) of the forest owners have an efficiency estimate of less than 0.75. To get an overall picture, the histograms of the estimated technical

efficiencies for the three models are reported in Figure 1, and those of the estimated functional parameters are reported in Figure 2. Figure 1 shows that, most of the forest owners are fully technically efficient under the semiparametric model and the standard parametric model with Z variables, with the mode of technical efficiency around one. However, under the standard parametric model without Z variables, the mode occurs about 0.9, and much fewer forest owners are estimated to be fully efficient. This may have some implication on the impact of model specification and the inclusion of Z variables on the estimated technical efficiency.

While all the three models yield observation-specific technical efficiency estimates, only the semiparametric model can generate observation-specific parameters. The distributions of the regression coefficients in Figure 2 show that the semiparametric model better captures parameter heterogeneity while its standard parametric counterparts only yield estimates that are degenerate. More specifically, the labor, land, and capital productivity (represented by $\hat{\beta}_1(Z_i)$, $\hat{\beta}_2(Z_i)$, and $\hat{\beta}_3(Z_i)$, respectively) estimates under the standard parametric models only approximate the means of those estimates under the semiparametric model. With a micro-level data set, it is generally more interesting and informative to investigate each forest owner as opposed to an average forest owner.

With all these differences in results between the semiparametric and its parametric counterpart, one would naturally perform specification test of one model against another. We test the standard parametric model without Z against the semiparametric model by testing the relevance of the environmental factors using the testing procedure described in section 3, because the semiparametric model without the environmental factors becomes the standard parametric model. The zero bootstrapped p-value suggests that these factors are relevant; and therefore the semiparametric model is preferred. This testing result is not very surprising based on the estimation results.

5 Conclusion

This paper proposes a semiparametric smooth-coefficient stochastic production frontier model, where all the coefficients, including intercept and slopes, along with the inefficiency term, are expressed as functions of a set of environmental factors. Thus, these factors affect the production frontier non-neutrally, as opposed to traditional inputs which only affect the frontier neutrally. Using micro-level data, this model can yield a particular set of production frontier estimates for a particular, say, firm. Therefore, the potential heterogeneity of technology can be captured by this model. Since the environmental factors enter most parameters in the model nonparametrically and the elimination of these factors reduces the semiparametric model to

its parametric counterpart, a testing procedure for the relevance of these factors is proposed. An empirical example using real data is presented and the advantages of the semiparametric approach over standard parametric approach are further revealed. A possible extension of this paper could be to relax the exponential functional form of the variance of the inefficiency term. This means, however, more work should be done to impose positivity constraint on the variance estimates.

References

Aigner, D. J., Lovell, C. A. K. and Schmidt, P. (1977), 'Formulation and estimation of stochastic frontier production functions', *Journal of Econometrics* **6**(1), 21–37.

Alvarez, A., Amsler, C., Orea, L. and Schmidt, P. (2006), 'Interpreting and testing the scaling property in models where inefficiency depends on firm characteristics', *Journal of Productivity Analysis* **25**(3), 201–212.

Battese, G. E. and Coelli, T. J. (1988), 'Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data', *Journal of Econometrics* **38**, 387–399.

Battese, G. E. and Coelli, T. J. (1992), 'Frontier production functions, technical efficiency and panel data: With applications to paddy farmers in India', *Journal of Productivity Analysis* **3**, 153–169.

Battese, G. E. and Coelli, T. J. (1995), 'A model for technical inefficiency effects in a stochastic frontier production function for panel data', *Empirical Economics* **20**, 325–32.

Battese, G. E., Prasada Rao, D. S. and O'Donnell, C. (2004), 'A metafrontier production function for estimation of technical efficiencies and technology gaps for firms operating under different technologies', *Journal of Productivity Analysis* **21**, 91–103.

Huang, C. J. and Liu, J.-T. (1994), 'Estimation of a non-neutral stochastic frontier production function', *Journal of Productivity Analysis* **5**, 171–180.

Kumbhakar, S. C. (1990), 'Production frontiers, panel data, and time-varying technical efficiency', *Journal of Econometrics* **46**, 201–12.

Kumbhakar, S. C., Ghosh, S. and McGuckin, J. T. (1991), 'A generalized production frontier approach for estimating determinants of inefficiency in US dairy farms', *Journal of Business and Economic Statistics* **9**, 279–86.

Kumbhakar, S. C. and Lovell, C. A. K. (2000), *Stochastic Frontier Analysis*, Cambridge Univeristy Press.

Li, Q., Huang, C., Li, D. and Fu, T. (2002), 'Semiparametric smooth coefficient models', *Journal of Business and Economic Statistics* **20**(3), 412–422.

Li, Q. and Racine, J. S. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Li, Q. and Racine, J. S. (2010), 'Smooth varying-coefficient estimation and inference for qualitative and quantitative data', *Econometric Theory* **26**, 1607–1637.

Lien, G., Størdal, S. and Baardsen, S. (2007), 'Technical efficiency in timber production and effects of other income sources', *Small-scale Forestry* **6**, 65–78.

Meeusen, W. and van den Broeck, J. (1977), 'Efficiency estimation from Cobb-Douglas production functios with composed error', *International Economic Review* **18**(2), 435–44.

O'Donnell, C., Prasada Rao, D. S. and Battese, G. E. (2008), 'Metafrontier frameworks for the study of firm-level efficiencies and technology ratios', *Empirical Economics* **34**, 231–55.

Simar, L., Lovell, C. A. K. and van den Eeckaut, P. (1994), Stochastic frontiers incorporating exogenous influences on efficiency. Discussion Paper No.9403, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Wang, H.-J. and Schmidt, P. (2002), 'One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels', *Journal of Productivity Analysis* **18**, 129–44.

Figure 1: **Technical Efficiency**

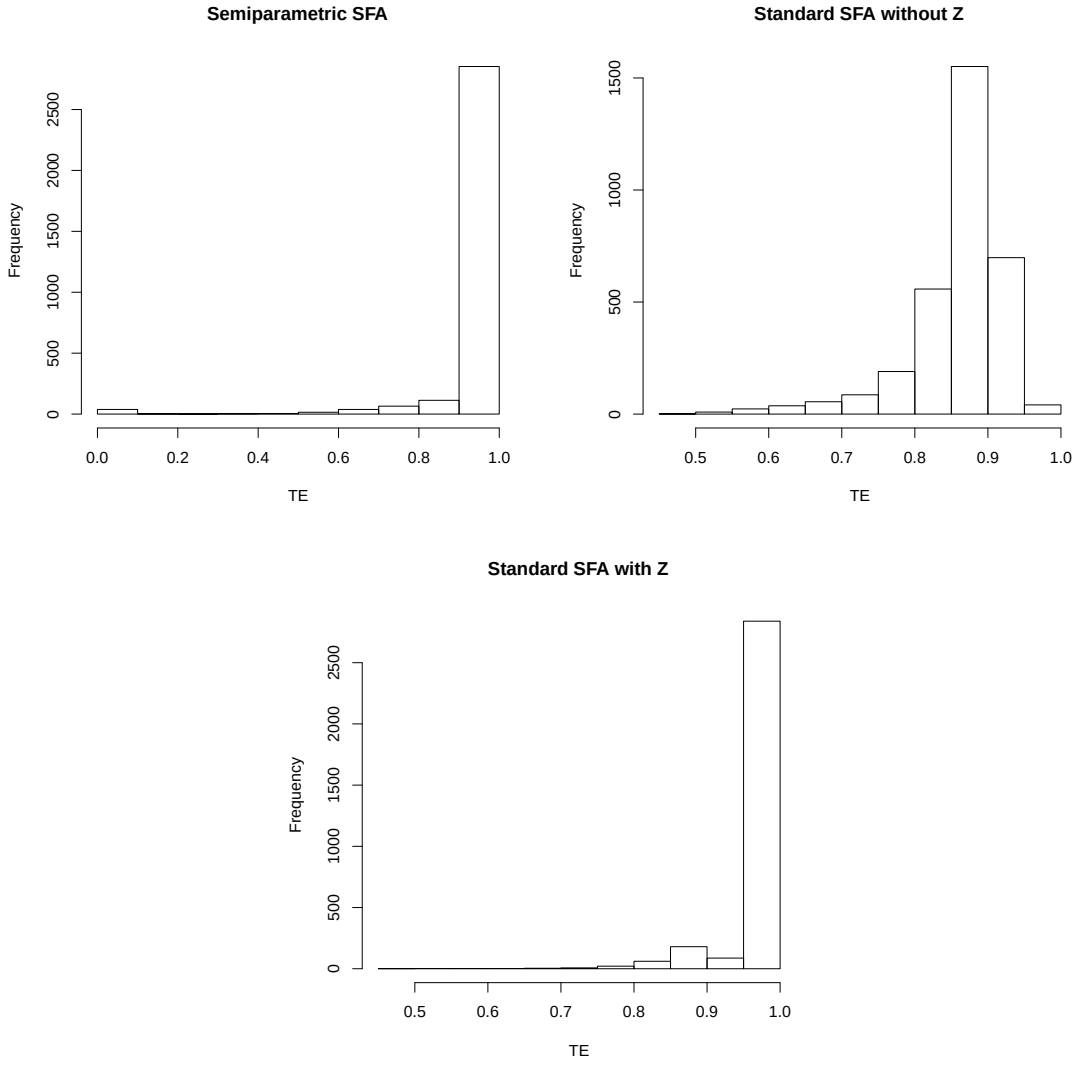

Figure 2: **Regression Coefficients**