Biases in approximating log production

Sun, Kai and Henderson, Daniel J. and Kumbhakar, Subal C.

State University of New York at Binghamton, Department of Economics

2010

Biases in Approximating Log Production*

Kai Sun, Daniel J. Henderson[†]and Subal C. Kumbhakar
Department of Economics
State University of New York at Binghamton

September 7, 2010

Abstract

Most empirical work in economic growth assumes either a Cobb-Douglas production function expressed in logs or a log-approximated constant elasticity of substitution specification. Estimates from each are likely biased due to logging the model and the latter can also suffer from approximation bias. We illustrate this with a successful replication of Masanjala and Papagerogiou (2004) and then estimate both models in levels to avoid these biases. Our estimation in levels gives results in line with conventional wisdom.

1 Introduction

There has been a long standing tradition of estimating production functions in logs when studying economic growth. For example, Mankiw, Romer and Weil (1992) derive the steady state level of output per worker and estimate the parameters of their model using a log-linearized version of the standard Cobb-Douglas (CD) production function. Mankiw et al. (1992), and other pioneering papers in this field, have literally been cited thousands of times and an overwhelming majority of these papers estimate the parameters of their models via ordinary least-squares (OLS), which is possible when the CD model is expressed in logarithmic form. Although the CD production function is relatively standard in this literature, other functional forms have been proposed. For example, Duffy and Papageorgiou (2000) suggest using a constant elasticity of substitution (CES) production function in place of the standard CD production function in order to capture nonlinearities in

*The authors would like to thank three anonymous referees, the editor Badi Baltagi, Chris Papageorgiou, João Santos Silva and Jon Temple for useful comments which led to an improved version of this article. All data and R code for this paper are available from the authors upon request.

[†]Daniel J. Henderson, Department of Economics, State University of New York, Binghamton, NY 13902-6000. Phone: 607-777-4480, Fax: 607-777-2681, E-mail: djhender@binghamton.edu.

the growth process and potential heterogeneity across countries. CES production functions are inherently nonlinear, but the authors take logs and linearize the CES production function using a Taylor series expansion as in Kmenta (1967) in order to employ OLS.[1] A wide range of papers are now using CES production functions to study growth (for example, Backus, Henriksen and Storesletten (2008), Duffy, Papageorgiou and Perez-Sebastian (2004), and Klump, McAdam and Willman (2007)).[2]

Although popular in economic growth, estimating CD and CES models in logs may lead to unnecessary biases. First, both the CD and CES models estimated in logs can produce biased estimates because the expected value of the logarithm of the error term generally depends upon the regressors. This point has been argued compellingly by Santos Silva and Tenreyro (2006) in the context of the gravity model of trade.[3] A second bias can also arise with CES models when they are approximated. For example, it is well known that the Kmenta approximation can suffer from omitted variable bias. Specifically, the Kmenta approximation does not guarantee that the underlying CES parameters are consistently estimated because the approximation is a Taylor series expansion and the remainder term becomes an omitted variable in the regression context. This point was laid out in Thursby and Lovell (1978) and by others. While these past papers have made an impact in the empirical international trade and productivity literatures, these potential biases appear to be ignored or are not considered to be too problematic by many growth economists.

Here we showcase each of these biases by replicating Masanjala and Papagerogiou (2004) who estimate both CD and CES models in logs. We then estimate their models in levels, both with and without using an approximation, via non-linear least-squares (NLLS) and Poisson pseudo-maximum likelihood estimation (PPML). The approximated models avoid the bias due to logging. When we estimate the models in levels without an approximation, we avoid both the bias due to estimating the model in logs as well as the approximation bias. Although we find minimial bias due to logging the model in the CD case, we find substantial bias in the log-approximated model in the CES case.

[1] The authors also estimate nonlinear versions of their models using non-linear least-squares. These models do not use Taylor series expansion, but in their paper they are estimated in logs.

[2] The *Journal of Macroeconomics* in 2008 devoted a special issue (Volume 30, Issue 2) to the CES production function and its impact on the theory and empirics of economic growth.

[3] See Papke and Wooldridge (1996) and Manning and Mullahy (2001) for related discussion.

On the other hand, we are unable to find substantial bias in the approximated CES model. When estimating the levels model without an approximation, we are able to consistently estimate the parameters of interest and these results are in line with conventional wisdom.

We should point out here that we are making several assumptions when claiming to consistently estimate the parameters of the model. First, we are assuming that the CD or CES model is the correct parametric specification. That being said, we perform functional form specification tests and are unable to reject the models estimated in levels. Second, we assume that there are no omitted variables which may bias our results. In other words, we are completely ignoring the issue of endogeneity of the regressors in this article. While our proposed estimation strategy may help with biases, estimating cross-country production functions in levels (as opposed to growth rates) potentially amplifies the endogeneity problem which is severe in this line of work. Third, although we allow for heterogeneity in the shares of physical and human capital in the CES model, we do not allow for parameter heterogeneity. There is good reason to believe that the parameters of the model may vary drastically across different groups of countries. We urge readers to consider the potential impact of each of these caveats, and perhaps others (for example, see Durlauf, Johnson and Temple (2005)), when interpreting our results. Even with these potential problems in our estimates, we still advocate for estimating models in levels, but suggest that practitioners be cautious of other potential biases in their own work.

The remainder of the paper is organized as follows: Section 2 shows the potential biases associated with estimating production functions. The third section describes estimation in logs while Section 4 estimates each of our production functions in levels. The final section concludes.

2 Biases in Growth Models

Approximating a function using a Taylor series expansion is well known to economists. What is often ignored in many applications is the bias associated with the remainder term. Since this term is omitted in estimation, the resulting estimators will likely suffer from omitted variable bias (similar to Heckman's selectivity bias). Although this approximation bias has been emphasized by several authors (for example, Byron and Bera (1983) and Thursby and Lovell (1978)), it is largely ignored in the growth literature (for an exception see Temple (2001), pages 914-915).

Consider the CES production function in Masanjala and Papagerogiou (2004). Given that the CES production function is nonlinear, Masanjala and Papagerogiou (2004) use the Kmenta (1967) approximation in order to estimate several of their models. This approximation is based on the elasticity of substitution (σ) being equal to unity (equations (4) and (6) in Masanjala and Papagerogiou (2004)). The approximation error here (the remainder term in the Taylor series expansion) is non-zero (function of x) and likely grows when σ deviates from 1.[4] This source of bias is addressed in Thursby and Lovell (1978) in the context of CES production functions. The other source of bias is due to the non-constant (function of x) mean of the error term when using logs. This is addressed by Santos Silva and Tenreyro (2006) in the context of gravity models, but it also applies here. They suggest estimating models in levels to avoid the potential bias and inconsistency associated with an error term whose conditional mean is a function of x.

Here we address biases in the estimates of CES production functions (although the arguments provided here are also applicable for other parametric functions) arising from both problems.[5] To distinguish between the models in level and logarithmic form, with and without approximation error, we write the CES production function in levels as

$$y = f(x, \theta) + u, \tag{1}$$

where y is a scalar output, x is a vector of inputs and θ is a vector of unknown parameters. Here we assume that $f(x, \theta)$ has a CES form and that $E(u|x) = 0$. We allow the error term u to be either homoskedastic or heteroskedastic. We can rewrite (1) as $y \equiv f(x, \theta)\,\eta$ and express it in logarithmic form, viz.,

$$\ln y = \ln f(x, \theta) + \ln \eta, \tag{2}$$

[4]The remainder term will also vary with the values of the regressors. See Thursby and Lovell (1978) for details on the bias associated with the Kmenta approximation.

[5]Noting that CD is a special case of CES when $\sigma = 1$, the CD model will not suffer from approximation bias as it does not require the Kmenta approximation. Logging both sides of the equation makes the CD production function linear in parameters.

where $\eta = (1 + u/f(x,\theta))$, and $E(\eta|x) = 1$. We can also use a linear or quadratic approximation of $f(x,\theta)$ at $\sigma = 1$ and write it as

$$y = f^0(x,\theta) + r(x,\theta) + u, \tag{3}$$

where $f^0(x,\theta)$ is a linear or quadratic approximation of $f(x,\theta)$ at $\sigma = 1$ and $r(x,\theta)$ is the approximation error (it contains higher order terms in the Taylor series expansion). The models in (1) and (3) are referred to as models in levels.

We can take a linear or quadratic approximation of (2) to get

$$\ln y = \ln f^0(x,\theta) + R(x,\theta) + \ln \eta, \tag{4}$$

where $\ln f^0(x,\theta)$ is a log-linear or log-quadratic approximation of $\ln f(x,\theta)$ at $\sigma = 1$ and $R(x,\theta)$ is the approximation error. We refer to the models in (2) and (4) as the log and log-approximated models, respectively.

Note that there are two sources of bias in estimating (4): (i) the deterministic terms $R(x,\theta)$ might not be zero (and might not converge to a constant as the sample size increases) and (ii) the expectation of the random term $E(\ln \eta)$ is almost always a function of x.[6] Each will cause bias and inconsistency if (4) is estimated using (linear or nonlinear) least-squares methods. That is, the least-squares estimator will suffer from bias and inconsistency arising from both sources. If the approximated non-log model in (3) is estimated using NLLS, it will obviously not suffer a bias due to logging the model, but it will suffer from omitted variable bias (due to excluding the $r(x,\theta)$ term). Using NLLS or PPML on (1) will avoid biases from both sources.[7]

[6] Expanding $\ln \eta = \ln [1 + u/f(x,\theta)] = \left[u/f(x,\theta) - \frac{1}{2}u^2/(f(x,\theta))^2 + \frac{1}{3}u^3/(f(x,\theta))^3 - ...\right]$ and then taking expectations of both sides shows that $E(\ln \eta)$ is a function of x. In the special case where u is heteroskedastic and the form is such that $u = f(x,\theta)v$ where v has zero mean and higher order moments that do not depend on x, then $E(\ln \eta)$ will be a constant. Otherwise, $E(\ln \eta)$ will be a function of x and we will refer to this as the bias associated with logging the model.

[7] Although the consistency of both estimators does not depend on the form of heteroskedasticity of u, the primary difference between the two is that PPML makes an assumption on the form of the conditional variance whereas the NLLS estimator does not. For more details see Cameron and Trivedi (2005) and Gourieroux, Monfort and Trognon (1984).

3 Estimation in Logs

Using the computing environment R (R Development Core Team (2009)) we were able to successfully replicate all of the tables and figures in Masanjala and Papagerogiou (2004). Given the specific interest of this paper, we solely present the replication results for the input shares obtained from the basic Solow growth model (without human capital) and its extended counterpart (with human capital).[8] In the CD version of the model, α and β are the actual shares of physical and human capital, respectively. In the CES case, α and β are distribution parameters which can be used to calculate the shares of physical and human capital.[9] In the CD model, the elasticity of substitution parameter σ is unity and this parameter is allowed to differ from unity in the CES specification.

Our Tables 1 and 2 present the results for the basic and extended Solow models, respectively. The replication results correspond to the rows which are listed as "OLS (Log)" or "NLLS (Log)." Specifically, we report the implied α and σ in Table 1 and the implied α, β and σ in Table 2. Each of our parameter estimates as well as our heteroskedastic robust standard errors for the log models are identical to those reported in Masanjala and Papagerogiou (2004).

By construction, the shares for the CD production function are constant across countries and their values are reported in the tables. On the other hand, the shares for the CES production function are heterogeneous. Masanjala and Papagerogiou (2004) report the shares for each country for each estimation method. To conserve space, we decided to simply plot the kernel densities of these shares in Figures 1 and 2 for the basic Solow and the extended Solow models, respectively. In Figure 1, the solid line represents the kernel density for the shares of physical capital from the log-approximated CES production function. The estimated density is relatively flat and roughly covers the range 0.3 to 1.0. The wide range likely represents heterogeneity across countries as the authors argue, but it may also be due to biases in their setup. In Figure 2 we see that the shares for physical capital have a mode near 0.26. Conventional wisdom suggests that these shares should be near one-third and this downward bias is perhaps due to the problems associated with logging the model and the Taylor series expansion as discussed above. For human capital, the log-approximated

[8]These correspond to the estimates of equations 1, 2, 4 and 6 in Table I of Masanjala and Papagerogiou (2004). All other replicated tables and figures are available upon request.

[9]See pages 177-179 in Masanjala and Papagerogiou (2004) for the exact form of the shares.

density extends from roughly 0.2 to 0.4. This partially overlaps the range of 1/3 to 1/2 which some authors consider to be "sensible" (for example, Mankiw et al. (1992)).

To get a feel of whether either model is correctly specified we performed RESET tests (Ramsey (1969)) on each logged model. In each case we reject the null that the proposed model is correctly specified. While the rejection of the RESET test does not point us in any particular direction, the results of this section suggest that there may be room for improvement and hence we now consider estimation of the CD and CES production functions in levels.

4 Estimation in Levels

Given the possible biases due to logging the model, we decided to estimate both the CD and CES models in levels. For the CD model, there is no need for an approximation as the elasticity of substitution is equal to unity. In the CES case, we will also consider the model with the Taylor series expansion around $\sigma = 1$. Our presumption is that NLLS and PPML estimation will be consistent in the CD case. On the other hand, the estimators from equation (3) will suffer from approximation bias in the CES framework whereas the estimators of equation (1) will be immune to both biases.

4.1 Cobb-Douglas

The top panels of Tables 1 and 2 give the results for the CD specification. We see that with the basic Solow model (Table 1), the implied share of physical capital (α) is relatively constant across estimation procedures. Our initial expectation was that the NLLS and PPML estimates would be significantly different from the log-linear estimates. This does not appear to be the case in either table. Specifically, this was our expectation because, as argued in Section 2, $E(\ln \eta | x)$ is a function of x.

This unexpected result likely occurs because either $E(\ln \eta | x)$ is close to a constant, the heteroskedasticity of u is close to the specific form mentioned above or some combination thereof. Given that the parameter estimates from the logged model are similar to those from the NLLS and

PPML estimators, some may take this to mean that the CD model estimated in logs is appropriate. We do not oppose this possibility, but note that it would be impossible to determine this in practice unless the level model is also estimated. Further, formal tests should also be performed before making this conclusion. For example, we also performed the RESET test on the CD model estimated via NLLS and PPML and were unable to reject the null in either case. This evidence suggests that the CD model is correctly specified and that we should estimate the model in levels. However, we choose to reserve judgement until we view the results from the CES specification.

4.2 Constant Elasticity of Substitution

The estimates for the CES production functions can also be found in Tables 1 and 2. The bottom panel in first table gives the results for the CES specification without human capital and the second table gives the estimates incorporating human capital. For the basic Solow model (Table 1), we see large differences between the logged and level methods. Recall that the log model is subject to two biases (logging and approximation). The levels estimators (NLLS and PPML) with approximation are subject to approximation bias whereas the remaining estimators should not suffer from either. The implied value of α for the logged estimator is much smaller than that of any other estimation method. Hence, it appears that the combination of logging the model and the Taylor series expansion cause biases in these estimates. This is in contrast to the CD production function, where logging the function did not make major differences in the parameter estimates.

Figure 1 clearly shows the difference in each of the estimators. Again, this figure plots the kernel densities for the shares of physical capital. The logged estimator potentially suffers from two separate biases whereas the estimator from equation (3) suffers from approximation bias. The bias due to logging appears to be obvious here and the approximation bias appears to be minimal.[10] The lack of a large approximation bias is likely due to the estimated value of σ being near unity. Thursby and Lovell (1978) show that the bias will grow as the estimated value of σ deviates from unity (the PPML estimate of σ is 1.037).

As compared to the CD case, we see larger differences between the NLLS and PPML estimators

[10] Note that although the parameter estimates don't differ significantly, there is a substantial decrease in the variability of the physical capital shares when we switch from the NLLS (Approximated) to the NLLS estimator.

(approximated or not) in the CES case. However, none of these parameter estimates (across estimation methods) are statistically different from one another. Another interesting result is that we fail to reject the null that σ is different from unity in nearly each model (NLLS approximated being the exception). In other words, each of our models point to CD as the preferred functional form. The most compelling results come from the PPML estimator. The PPML CD and CES implied α and σ are not statistically different from one another and are nearly identical.

The results for the basic Solow CES model are interesting, but may suffer from an omitted variable bias due to the exclusion of human capital. The bottom panel of Table 2 gives the results for the extended Solow CES model. Again, we see substantial differences between the implied estimates of α, β and σ between the logged and level models. We again conclude that the logged model is likely biased. This is supported by a rejection of the RESET test.

Again, the models estimated in levels, whether or approximated or not, do not have significant differences in their parameter estimates. The only qualitative difference is that the NLLS estimators reject the null that $\sigma = 1$ and the PPML estimators do not. Thus, the evidence in favor of CD is mixed in the extended CES case. The RESET test does not help us pick between level estimators as it again fails to reject any of the models estimated in levels.[11]

We argue that logging a model or using approximations can lead to biases, but the question should be, what type of difference does correcting these biases make in terms of explaining the economics. The most striking economic evidence can be found in Figure 2. In the first panel we see the kernel density estimates for the shares of physical capital. Conventional wisdom suggests that the share devoted to physical capital should be near one-third. In the panel we see that the density of estimated shares obtained from the log-linearized method has a mode less than 1/3 and the approximated NLLS estimator gives physical capital shares larger than one-third. On the other hand, the mode for the shares of physical capital from the level NLLS estimator are essentially centered around what is hypothesized. The level and approximated PPML physical capital shares are slightly less than 1/3 and show much less variation. It is unclear which level estimator is preferable in this situation without examining mean squared error, but it should be

[11]It is worth noting that the PPML estimator has the largest p-value for the RESET test.

obvious that both level estimators are intuitively more plausible than the logged model. This figure shows the biases due to logging and approximation, but it also shows how avoiding these biases validates conventional thought. In contrast, the shares of human capital show a substantial amount of overlap amongst estimators.

In conclusion, for the CD case, for this particular sample of data, given the estimation strategies that we have employed, we recommend estimating the CD model in levels with human capital included in the model. In the CES case, the log-approximated model introduces bias. Estimation of CES should, where possible, use a level estimator. We have mixed evidence that the elasticity of substitution is equal to unity here. If it is not different from one, we should switch to the CD model. It would be interesting to see how these results translate to a larger sample with more up to date data.

We feel it imperative that in applied work, authors should estimate the model which is consistent under the most general conditions and then make attempts to identify a more parsimonious, efficient model. In this particular scenario, there is some evidence to suggest that the CD model is the preferred specification. However, in practice, this is not necessarily going to hold true for all datasets.

5 Conclusion

In this paper we were able to successfully replicate Masanjala and Papagerogiou (2004). Along the way we pointed out known potential biases due to logging the model and using Taylor series expansions to approximate functional forms. To avoid these biases, we estimated their models in levels. We showed that with this particular paper, elimination of these biases led to conclusions which were in line with conventional wisdom.

We again note that logging a model and approximating a functional form need not lead to significant biases in practice. However, given the computing speed and canned statistical packages available today, the cost of estimating models in levels has been greatly reduced. Therefore, we advocate estimating cross-country production functions in levels as a way to avoid potential and unnecessary biases.

Again, as noted in the introduction, our results depend upon several assumptions. When estimating models in levels authors should still be concerned with correct functional form specification, omitted variable bias, parameter heterogeneity as well as other possible biases. Controlling for each of these could potentially change the results of this experiment.

References

Backus, D., Henriksen, E. and Storesletten, K. (2008), 'Taxes and the global allocation of capital', *Journal of Monetary Economics* **55**, 48–61.

Byron, R. and Bera, A. (1983), 'Least squares approximate to unknown regression functions: A comment', *International Economic Review* **24**, 255–260.

Cameron, A. and Trivedi, P. (2005), *Microeconometrics: Methods and Applications*, Cambridge University Press, New York.

Duffy, J. and Papageorgiou, C. (2000), 'A cross-country empirical investigation of the aggregate production function specification', *Journal of Economic Growth* **5**, 87–120.

Duffy, J., Papageorgiou, C. and Perez-Sebastian, F. (2004), 'Capital-skill complementarity? evidence from a panel of countries', *Review of Economics and Statistics* **86**, 327–344.

Durlauf, S., Johnson, P. and Temple, J. (2005), *Handbook of Economic Growth*, Vol. 1, Elsevier.

Gourieroux, C., Monfort, A. and Trognon, A. (1984), 'Pseudo maximum likelihood methods: Applications to poisson models', *Econometrica* **52**, 701–720.

Klump, R., McAdam, P. and Willman, A. (2007), 'Factor substitution and factor-augmenting technical progress in the United States: A normalized supply-side system approach', *Review of Economics and Statistics* **89**, 183–192.

Kmenta, J. (1967), 'On estimation of the CES production function', *International Economic Review* **8**, 180–189.

Mankiw, G., Romer, D. and Weil, N. (1992), 'A contribution to the empirics of economic growth', *Quarterly Journal of Economics* **107**, 407–437.

Manning, W. and Mullahy, J. (2001), 'Estimating log models: To transform or not to transform?', *Journal of Health Economics* **20**, 461–494.

Masanjala, W. and Papagerogiou, C. (2004), 'The Solow model with CES technology: Nonlinearities and parameter heterogeneity', *Journal of Applied Econometrics* **19**, 171–201.

Papke, L. and Wooldridge, J. (1996), 'Econometric methods for fractional response variables with an application to 401(k) plan participation rates', *Journal of Applied Econometrics* **11**, 619–632.

R Development Core Team (2009), 'The R foundation for statistical computing', *http://cran.r-project.org/* .

Ramsey, J. (1969), 'Tests for specification errors in classical linear least squares regression analysis', *Journal of the Royal Statistical Society, Series B* **31**, 350–371.

Santos Silva, J. and Tenreyro, S. (2006), 'The log of gravity', *Review of Economics and Statistics* **88**, 641–658.

Temple, J. (2001), 'Generalizations that aren't? Evidence on education and growth', *European Economic Review* **45**, 905–918.

Thursby, J. and Lovell, C. (1978), 'An investigation of the Kmenta approximation to the CES function', *International Economic Review* **19**, 363–377.

Table 1: **Growth Model with Physical Capital**

	Implied α	Implied σ
Basic CD		
OLS (Log)	0.5981^{***}	1
	(0.0170)	
NLLS	0.5907^{***}	1
	(0.0254)	
PPML	0.6057^{***}	1
	(0.0165)	
Basic CES		
OLS (Log)	0.4984^{***}	1.5425
	(0.0500)	(0.5574)
NLLS (Approximated)	0.7009^{***}	$0.9499^{\star\star\star}$
	(0.0229)	(0.0055)
PPML (Approximated)	0.6376^{***}	0.9623
	(0.0936)	(0.0736)
NLLS	0.7486^{***}	0.8354
	(0.1192)	(0.1122)
PPML	0.6172^{***}	0.9813
	(0.0457)	(0.0745)

1. The numbers in the parentheses are heteroscedastic-robust standard errors.
2. $***(\star\star\star)$: statistically different from 0(1) at the 1% level.

Table 2: **Growth Model with Physical and Human Capital**

	Implied α	Implied β	Implied σ
Extended CD			
OLS (Log)	0.3082^{***}	0.2743^{***}	1
	(0.0465)	(0.0356)	
NLLS	0.2845^{***}	0.3091^{***}	1
	(0.0831)	(0.0751)	
PPML	0.3028^{***}	0.3005^{***}	1
	(0.0545)	(0.0464)	
Extended CES			
NLLS (Log)	0.2395^{***}	0.3582^{***}	$1.1894^{\star\star\star}$
	(0.0484)	(0.0512)	(0.0337)
NLLS (Approximated)	0.4295^{***}	0.2773^{***}	$0.9219^{\star\star\star}$
	(0.0693)	(0.0611)	(0.0138)
PPML (Approximated)	0.2796^{***}	0.3140^{***}	1.0590
	(0.0649)	(0.0564)	(0.0918)
NLLS	0.3928^{***}	0.2851^{***}	$0.8381^{\star}$
	(0.1185)	(0.0705)	(0.0962)
PPML	0.2851^{***}	0.3105^{***}	1.0369
	(0.0650)	(0.0551)	(0.0582)

1. The numbers in the parentheses are heteroscedastic-robust standard error.
2. $*(\star)$,$**(\star\star)$,$***(\star\star\star)$: statistically different from 0(1) at the 10%, 5%, and 1% level, respectively.

Figure 1: **Physical Capital Shares in the Basic CES Model**

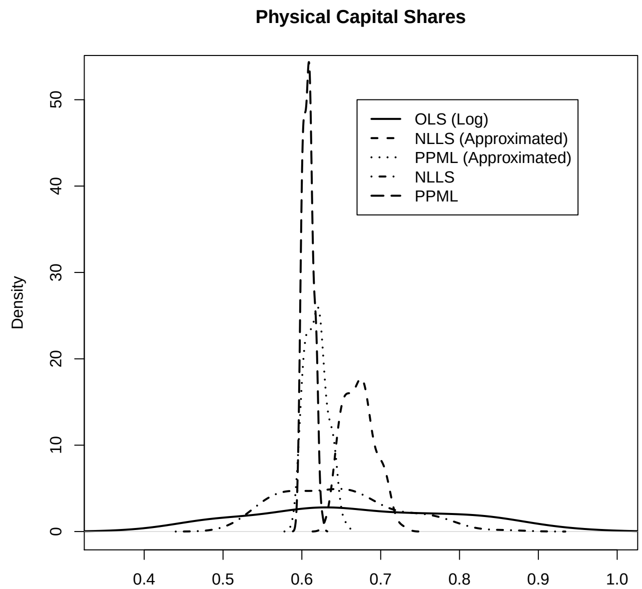

Figure 2: **Physical and Human Capital Shares in the Extended CES Model**