Firm Size and R&D Spending: Testing for Functional Form

I. Introduction

One of the most frequently investigated topics in the economics of technological change literature is the relationship between firm size and the corresponding level of R&D spending. Over more than three decades researchers have been examining numerous functional forms in an effort to determine whether larger firms spend more on R&D relative to their size than do smaller firms. Over time, a double-log regression model has evolved to be the "accepted" specification.[1] However, the statistical appropriateness of a particular functional form has received almost no attention.[2]

This paper tests for functional form using an extended Box-Cox model. We conclude from our analysis of firm-level industrial data that there is considerable evidence to suggest that the double-log specification is, in fact, the most appropriate within the class of models represented by the Box-Cox transformation.

II. The Analytical Framework

Consider the following model relating the ith firm's R&D expenditures, RD_i, to its size, S_i:

$$RD_i^{(\lambda_1)} = \alpha + \beta S_i^{(\lambda_2)} + \epsilon_i \qquad (1)$$

where $RD_i^{(\lambda_1)} = (RD_i^{\lambda_1} - 1)/\lambda_1$ if $\lambda_1 \neq 0$ and $RD_i^{(\lambda_1)} = \ln RD_i$ if $\lambda_1 = 0$, and the values of $S_i^{(\lambda_2)}$ are similarly defined. Such a specification is what Savin and White [6] refer to as the Box-Cox extended model to distinguish it from the original Box-Cox [2] paper which primarily dealt with transformations of the dependent variable.

If ϵ_i is assumed to be normally and independently distributed with zero mean and constant variance σ^2, then the log-likelihood function for equation (1) is:

$$
\begin{aligned}
L = {} & -(n/2)\ln 2\pi - (n/2)\ln \sigma^2 \\
& - (1/\sigma^2)\sum (RD_i^{(\lambda_1)} - \alpha - \beta S_i^{(\lambda_2)})^2 \\
& + (\lambda_1 - 1)\sum \ln RD_i.
\end{aligned}
\qquad (2)
$$

L is a complicated nonlinear function of α, β, λ_1, and λ_2. Spitzer [7] has recently developed a powerful algorithm for the numerical maximization of this specific function. Once the log

---

1. Much of this literature is reviewed in Kamien and Schwartz [3]. Two recent studies that posit a double-log regression model are by Mansfield [5] and Bound et al. [1].

2. The one exception is the study by Loeb and Lin [4]. They employ a set of tests due to Ramsey. Ramsey's tests identify model specification errors but they do not specifically identify a single functional form from within a class, as does the analysis below. Their data fit a quadratic specification best.

likelihood function is maximized, it is easy to perform likelihood ratio tests on four alternative functional forms using the statistic

$$\chi_2^2 = 2[L(\hat{\lambda}_1, \hat{\lambda}_2) - L(\lambda_1, \lambda_2)] \tag{3}$$

which has asymptotically a chi-square distribution with two degrees of freedom, $\hat{\lambda}_1$ and $\hat{\lambda}_2$ denote the maximum likelihood estimates, and λ_1 and λ_2 are values to be tested.

This procedure is attractive because equation (1) nests four common functional forms. Depending on the values of λ_1 and λ_2, we can have the double-log model, a linear model, or a model that is semilog in either the dependent or independent variables. Specifically we have:

$$\ln RD_i = \alpha + \beta \ln S_i + \epsilon_i \qquad \text{if } \lambda_1 = 0, \lambda_2 = 0; \tag{4}$$

$$\ln RD_i = \alpha + \beta S_i + \epsilon_i \qquad \text{if } \lambda_1 = 0; \lambda_2 = 1; \tag{5}$$

$$RD_i = \alpha + \beta \ln S_i + \epsilon_i \qquad \text{if } \lambda_1 = 1; \lambda_2 = 0; \tag{6}$$

$$RD_i = \alpha + \beta S_i + \epsilon \qquad \text{if } \lambda_1 = 1; \lambda_2 = 1. \tag{7}$$

III. The Statistical Results

Equation (1) was estimated by maximum likelihood for nine separate industry groupings using 1985 firm data on R&D expenditures reported in *Business Week's* "R&D Scoreboard." Firm size was measured as 1985 sales, as reported in *Business Week's* "Corporate Scoreboard."[3] The statistical results are shown in Table I.

In seven of the nine industry studied, we could not reject the null hypothesis that $\lambda_1 = \lambda_2 = 0$ when testing at either the .05 or .01 level. This result was found in electronics, chemicals, machinery, fuel, instruments, oil service, and miscellaneous manufacturing. In these industries, the double-log model is not sharply different from the optimal model within the class of models represented by equation (1). Further, the three alternative simple functional forms were almost always decisively rejected by the chi-square test, with only a semilog transformation of the dependent variable being competitive. This semilog model was not rejected at the .05 level in two cases (machinery and instruments) and in other cases was usually less vigorously rejected than the other two simple functional forms.

In one of the nine industries none of the simple functional forms were acceptable at the .01 level (computers), and in another industry none was acceptable at the .05 level (food and beverage). However, even though all the four simple forms were rejected in these two industries, it is reassuring that the double-log model is again closest to acceptance in the sense of having the lowest chi-square value. If one were willing to accept some error in the functional form in return for a simple form of the model, the double-log model would again be the choice. Thus, in all of the nine cases it was found that the double-log model was either not rejected or was close to non-rejection.

One important feature of the results in Table I concerns the misleading results that could

---

3. Some firms listed in the "R&D Scoreboard" were not listed in the "Corporate Scoreboard." Rather than approximate the sales of those firms from the R&D-to-sales ratio in the "R&D Scoreboard" (carried there to only one decimal point), these few firms were eliminated.

Table I. Maximum Likelihood Estimates from Equation (1) (standard errors in parentheses)

	λ_1	λ_2	$L(\lambda_1, \lambda_2)$	χ_2^2	R^2
Electronics ($n = 41$)					
Optimal Model	0.166	0.179	-164.761	—	0.857
	(0.089)	(0.115)			
Double-log	0	0	-166.455	3.388	0.812
Semilog (RD)	0	1	-187.660	45.798**	0.471
Semilog (S)	1	0	-248.054	166.586**	0.544
Linear	1	1	-206.523	83.524**	0.940
Chemicals ($n = 32$)					
Optimal Model	0.198	0.226	-141.837	—	0.865
	(0.100)	(0.115)			
Double-log	0	0	-144.017	4.360	0.811
Semilog (RD)	0	1	-159.794	35.914**	0.494
Semilog (S)	1	0	-198.537	113.400**	0.601
Linear	1	1	-169.408	55.142**	0.935
Food and Beverage ($n = 23$)					
Optimal Model	0.384	0.772	-79.606	—	0.814
	(0.168)	(0.275)			
Double-log	0	0	-83.182	7.152*	0.727
Semilog (RD)	0	1	-84.034	8.856*	0.706
Semilog (S)	1	0	-97.685	36.158**	0.584
Linear	1	1	-87.968	16.724**	0.821
Computers ($n = 20$)					
Optimal Model	0.344	0.244	-100.157	—	0.958
	(0.104)	(0.113)			
Double-log	0	0	-106.147	11.980**	0.896
Semilog (RD)	0	1	-123.764	47.214**	0.396
Semilog (S)	1	0	-145.466	90.618**	0.588
Linear	1	1	-117.236	34.158**	0.976
Machinery ($n = 18$)					
Optimal Model	0.332	0.921	-55.658	—	0.769
	(0.194)	(0.398)			
Double-log	0	0	-57.927	4.538	0.611
Semilog (RD)	0	1	-57.577	3.836	0.626
Semilog (S)	1	0	-74.478	37.640**	0.582
Linear	1	1	-65.803	20.290**	0.841

be reached by a naive comparison of the R^2 values. Since the dependent variable is not constant across the four simple functional forms, the values of R^2 are not comparable, as is well known. Still, researchers often appeal to high values of R^2 as at least a rough measure of goodness-of-fit, and this criterion would be highly deceptive here.

For example, in the electronics industry, the linear model produces an R^2 of .940 but is decisively rejected by a chi-square value of 83.524. In contrast, the double-log model yields an R^2

Table I. Continued

	λ_1	λ_2	$L(\lambda_1,\lambda_2)$	χ_2^2	R^2
Fuel ($n = 17$)					
Optimal Model	0.220	0.308	−83.758	—	0.843
	(0.144)	(0.165)			
Double-log	0	0	−85.804	4.092	0.807
Semilog (RD)	0	1	−92.405	17.292**	0.581
Semilog (S)	1	0	−105.063	42.610**	0.467
Linear	1	1	−95.417	23.318**	0.829
Instruments ($n = 15$)					
Optimal Model	−0.024	0.328	−63.264	—	0.541
	(0.271)	(0.520)			
Double-log	0	0	−63.456	0.384	0.534
Semilog (RD)	0	1	−64.037	1.546	0.496
Semilog (S)	1	0	−70.414	14.300**	0.560
Linear	1	1	−71.373	16.218**	0.500
Oil Service and Supply ($n = 14$)					
Optimal Model	0.574	0.054	−49.887	—	0.805
	(0.260)	(0.320)			
Double-log	0	0	−52.414	5.053	0.740
Semilog (RD)	0	1	−55.855	11.935**	0.575
Semilog (S)	1	0	−51.469	3.163	0.798
Linear	1	1	−53.588	7.401*	0.726
Miscellaneous Manufacturing ($n = 42$)					
Optimal Model	−0.255	−0.041	−140.180	—	0.673
	(0.144)	(0.263)			
Double-log	0	0	−142.823	5.286	0.687
Semilog (RD)	0	1	−146.997	13.634**	0.618
Semilog (S)	1	0	−221.090	161.820**	0.420
Linear	1	1	−200.332	120.304**	0.784

*significant at the .05 level (> 5.99)
**significant at the .01 level (> 9.21)

of .812 yet its chi-square value of 3.388 suggests it is not distinguishable from the optimal model. Inspection of the table shows that almost invariably the R^2 criterion would lead to an erroneous model choice. (R^2 is defined as the squared correlation of the predicted and actual values.)

The industry estimates in Table I can be compared easily across different functional forms if we employ an elasticity. Following Savin and White [6], the elasticities were computed at the mean value of both the dependent and independent variables, and are reported in Table II.[4] For the double-log model, the calculated values range from 0.752 to 1.288. The average for all nine industries (weighted by sample size) is 1.050, extremely close to unity. For the optimal model, the range is 0.675 to 1.464, with a weighted mean of 1.074.

---

4. Savin and White [6] note the elasticity for equation (1) is given by $\beta(\overline{S}^{\lambda_2}/\overline{RD}^{\lambda_1})$. In the double-log model where $\lambda_1 = \lambda_2 = 0$, this gives the familiar result that β is the elasticity. When $\lambda_1 = \lambda_2 = 1$, the elasticity is β times the ratio of the means.

Table II. Calculated Elasticities from Table I

Industry	Optimal	Double-log	Semilog (RD)	Semilog (S)	Linear
Electronics	0.994	1.005	0.443	1.267	0.964
Chemicals	1.053	1.068	0.562	1.150	0.969
Food and Beverage	1.045	0.940	1.138	0.738	1.075
Computers	0.992	1.090	0.491	0.979	0.854
Machinery	1.153	1.288	1.103	1.542	1.569
Fuel	0.998	0.893	0.989	0.588	1.023
Instruments	0.757	0.752	0.591	0.870	0.670
Oil Service and Supply	0.675	0.797	0.765	0.628	0.652
Miscellaneous Manufacturing	1.464	1.275*	0.856	1.857	1.797

*significantly different from unity at the .05 level

Finally, we tested the null hypothesis that the elasticity equals unity using the values of $\hat{\beta}$ from the double-log model. In eight of nine cases, as noted in Table II, we could not reject the null hypotheses of unitary elasticity when testing at the .05 level. Testing hypotheses about the elasticity is of course easy for the double-log model because standard errors are available. To our knowledge it is not similarly possible to test hypotheses about elasticities for the other three functional forms since a standard error is not directly available.[5]

IV. Conclusions

While conventional wisdom is not always the best guide, in the topic investigated here the common practice appears to be quite accurate. For more than three decades researchers have explored the R&D-to-size relationship. Many now simply posit a double-log regression model and, after testing, conclude that the estimated elasticity is close to unity. Here, we have systematically investigated the appropriateness of alternative function forms, and we find evidence to suggest that the customary practice is not inappropriate.

Albert N. Link
Terry G. Seaks
Sabrina R. Woodbery
University of North Carolina at Greensboro
Greensboro, North Carolina

5. One should be cautioned in attempting to compute the standard errors for the elasticities associated with equations (5), (6), and (7) by allowing only for the randomness in $\hat{\beta}$. For example, the elasticity for equation (7) is estimated by $\hat{\beta}(\overline{S}/\overline{RD})$, but a confidence interval for this elasticity cannot be based upon $(\overline{S}/\overline{RD})$ times the two endpoints of the interval $(\hat{\beta} \pm t s_{\hat{\beta}})$ because this captures only the sampling error in $\hat{\beta}$ while ignoring the sampling error in $(\overline{S}/\overline{RD})$.

References

1. Bound, John, Clint Cummings, Zvi Griliches, Bronwyn Hall, and Adam Jaffe. "Who Does R&D and Who Patents?" in *R&D, Patents, and Productivity*, edited by Z. Griliches. Chicago: University of Chicago Press, 1984.
2. Box, G. E. P. and D. R. Cox, "An Analysis of Transformations." *Journal of the Royal Statistical Society*, Series B 26, 1964, 211–43.

3. Kamien, Morton I. and Nancy L., Schwartz, "Market Structure and Innovation: A Survey." *Journal of Economic Literature*, March 1975, 1–37.

4. Loeb, Peter D. and Vincent Lin, "Research and Development in the Pharmaceutical Industry—A Specification Error Approach." *Journal of Industrial Economics*, September 1977, 45–51.

5. Mansfield, Edwin, "Composition of R&D Expenditures: Relationship to Size of Firm, Concentration, and Innovative Output." *Review of Economics and Statistics*, November 1981, 610–15.

6. Savin, N. E. and Kenneth J. White, "Estimation and Testing for Functional Form and Autocorrelation." *Journal of Econometrics*, August 1978, 1–12.

7. Spitzer, John J., "A Fast and Efficient Algorithm for the Estimation of Parameters in Models with the Box-and-Cox Transformation." *Journal of the American Statistical Association*, December 1982, 760–66.