Volume Title: R & D, Patents, and Productivity

Volume Author/Editor: Zvi Griliches, ed.

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-30884-7

Volume URL: http://www.nber.org/books/gril84-1

Publication Date: 1984

Chapter Title: Who Does R&D and Who Patents?

Chapter Author: John Bound, Clint Cummins, Zvi Griliches, Bronwyn H. Hall, Adam B. Jaffe

Chapter URL: http://www.nber.org/chapters/c10043

Chapter pages in book: (p. 21 - 54)

2 Who Does R & D and Who Patents?

John Bound, Clint Cummins, Zvi Griliches, Bronwyn H. Hall, and Adam Jaffe

2.1 Introduction

As part of an ongoing study of R & D, inventive output, and productivity change, the authors are assembling a large data set for a panel of U.S. firms with annual data from 1972 (or earlier) through 1978. This file will include financial variables, research and development expenditures, and data on patents. The goal is to have as complete a cross section as possible of U.S. firms in the manufacturing sector which existed in 1976, with time-series information on the same firms for the years before and after 1976. This paper presents a preliminary analysis of these data in the cross-sectional dimension, laying some groundwork for the future by exploring the characteristics of this sample and by describing the R & D and patenting behavior of the firms in it. This paper follows previous work on a smaller sample of 157 firms (see Pakes and Griliches 1980 and Pakes 1981).

We first describe the construction of our sample from the several data sources available to us. Then we discuss the reporting of our key variable,

John Bound, Clint Cummins, and Adam Jaffe are graduate students in the Department of Economics at Harvard University. Zvi Griliches is professor of economics at Harvard University, and program director, Productivity and Technical Change, at the National Bureau of Economic Research Research. Bronwyn H. Hall is a research analyst for the National Bureau of Economic Research.

This paper is a revision of an earlier draft presented at the NBER conference on R & D, Patents, and Productivity in Lenox, Massachusetts, October 1981. That version contained preliminary results on patenting in the drug and computer industries which have been replaced in this revision by a section on patenting in all manufacturing industries.

This work has been supported by the NBER Productivity and Technical Change Studies Program and NSF grants PR79–13740 and SOC79–04279. We are indebted to Sumanth Addanki and Elizabeth Stromberg for research assistance. The research reported here is part of NBER's research program in productivity. Any opinions expressed are those of the authors and not those of NBER.

research and development expenditures, and relate this variable to firm characteristics, such as industry, size, and capital intensity. An important issue is whether the fact that many firms do not report R & D expenditures will bias results based only on firms which do. We attempt to correct for this bias using the well-known Heckman (1976) procedure.

Section 2.4 describes the patenting behavior of the same large sample of firms. We attempt to quantify the relationship between patenting, R & D spending, and firm size, and to explore the interindustry differences in patenting in a preliminary way. Because of the many small firms in this data set, we pay considerable attention to the problem of estimation when our dependent variable, patents, takes on small integer values. The paper concludes with some suggestions for future work using this large and fairly rich data set.

2.2 Sample Description

The basic universe of the sample is the set of firms in the U.S. manufacturing sector which existed in 1976 on Standard and Poor's Compustat Annual Industrial Files. The sources of data for these tapes are company reports to the Securities and Exchange Commission (SEC), primarily the 10-K report, supplemented by market data from such sources as National Association of Securities Dealers Automated Quotations (NASDAQ) and occasionally by personal communication with the company involved. The manufacturing sector is defined to be firms in the Compustat SIC groups 2000–3999 and conglomerates (SIC 9997).[1]

Company data were taken from four Compustat tapes. The Industrial file includes the Standard and Poor 400 companies, plus all other companies traded on the New York and American Stock Exchanges. The Over the Counter (OTC) tape includes companies traded over the counter that command significant investor interest. The Research tape includes companies deleted from other files because of acquisition, merger, bankruptcy, and the like. Finally, the Full Coverage tape includes other companies which file 10-K's, including companies traded on regional exchanges, wholly owned subsidiaries, and privately held companies. From these tapes we obtained data on the capital stock, balance sheets, income statements including such expense items as research and development expenditures, stock valuation and dividends, and a few miscellaneous variables such as employment.

Unfortunately, our patent data do not come in a form which can be matched easily at the firm level. Owing to the computerization of the

1. This limitation is primarily for convenience; about 97 percent of company-sponsored R & D was performed in the manufacturing sector in 1976 (NSF 1979). It does, however, exclude a few large performers of R & D in the communications and computer service industries.

U.S. Patent Office in the late 1960s, we are able to obtain a file with data on each individual patent granted by the Patent Office from 1969 through 1979. For each such patent we have the year it was applied for, the Patent Office number of the organization to which it was granted, an assignment code telling whether the organization is foreign or domestic, corporate or individual, and some information on the product field and SIC of the patent. We also have a file listing the Patent Office organization numbers and the correspondent names of these organizations. The difficulty is that these patenting organizations, although frequently corporations in our sample, may also be subsidiaries of our firms or have a slightly different name from that given on the Compustat files ("Co." instead of "Inc." or "Incorporated" and other such changes or abbreviations).[2] Thus, the matching of the Patent Office file with the Compustat data is a major task in our sample creation.

To do the matching, we proceeded as follows: All firms in the final sample (about 2700) were looked up in the *Dictionary of Corporate Affiliations* (National Register 1976). Their names as well as the names of their subsidiaries were entered in a data file to be matched by a computer program to the names on the Patent Office organization file. This program had various techniques for accommodating differences in spelling and abbreviations. The matched list of names which it produced was checked for incorrect matches manually, and a final file was produced which related the Compustat identifying Committee on Uniform Securities Identification Procedures (CUSIP) number of each firm to one or more (in some cases, none) Patent Office organization numbers. Using this file, we aggregated the file with individual patent records to the firm level. As this paper is being written, we are engaged in a reverse check of the matching process which involves looking at the large patenting organizations which are recorded as domestic U.S. corporations, but which our matching program missed. The results of this check may further increase some of our patent totals.

In assembling this data set we have attempted to confine the sample to domestic corporations, since the focus of our research program is the interaction between research and development, technological innovation, and productivity growth within the United States. Inspection of the Compustat files reveals that at least a few large foreign firms, mostly Japanese, are traded on the New York Stock Exchange, and they consequently file 10-K's with the SEC and would be included in our sample, although presumably their R & D is primarily done abroad and their U.S. patents are recorded as foreign owned. To clean our sample of these firms we did several things: First, we were able to identify and delete all firms which Compustat records as traded on the Canadian Stock Exchange.

2. The vast majority of patents are owned by principal companies. In our earlier sample about 10 percent of total patents were accounted for by patents of subsidiaries.

Then we formed a ratio of foreign-held U.S. patents to total number of U.S. patents for each firm in our sample. For most of our sample, this ratio is less than 15 percent; the list of firms for which it is larger includes most of the American Deposit Receipts (ADR) firms on the New York Stock Exchange and several other firms clearly identifiable as foreign. After deleting these firms from the sample, as a final check we printed a list of the remaining firms with "ADR" or "LTD" in their names. There were eighteen such firms remaining, which we deleted from the sample.

The firms which were left still had a few foreign-owned patents (about 2 percent of the total number of patents in 1976) from joint ventures or foreign subsidiaries. Since their Compustat data are consolidated and include R & D done by these subsidiaries in the R & D figure, we added those patents to the domestic patents to produce a total successful patent application figure for the firm.

Our final 1976 cross section consists of data on sales, employment, book value in various forms, pre-tax income, market vaue, R & D expenditures, and patents applied for in 1976 for approximately 2600 firms in the manufacturing sector. The selection of these firms is summarized in table 2.1. Except for a few cases, firms without reported gross plant value in 1976 are firms which did not exist in 1976. Seventy-seven firms were deleted because they were either wholly owned subsidiaries of another company in our sample or duplicates in the Compustat files; another thirty-one had zero or missing sales or gross plant value. The final sample consists of 2595 firms, of which 1492 reported positive R & D in 1976. In section 2.3 we present some results on the R & D characteristics of these firms.

2.3 The Reporting of Research and Development Expenditures

In 1972 the SEC issued new requirements for reporting R & D expenditures on Form 10-K. These requirements mandate the disclosure of the

Table 2.1 **Creation of the 1976 Cross Section**

Compustat File	Manufacturing Firms on Compustat Tape	Gross Plant Reported in 1976	Positive Gross Plant & Sales in 1976[a]	Positive R & D
Industrial	1299	1294	1248	770
OTC	489	472	458	292
Research	414	138	132	83
Full coverage	1019	867	757	347
Total number of firms	3221	2771	2595	1492

[a]Duplicates, subsidiaries, or foreign not included.

estimated amount of R & D expenditures when (a) it was "material," (b) it exceeded 1 percent of sales, *or* (c) a policy of deferral or amortization of R & D expenses was pursued. Acting on these new requirements, the Financial Accounting Standards Board issued a new standard for reporting R & D expeditures in June 1974. Until this time, accepted accounting practices appear to have allowed the amortizing of R & D expenditures over a short time period as an alternative to simple expensing, but the new standard allows only expensing (San Miguel and Ansari 1975). Accordingly, we believe that by 1976 most of our firms were reporting R & D expense when it was "material" and that the expense reported had been incurred that year.

For the purpose of this paper, we make no distinction among firms whose R & D is reported by Compustat as "not available," "zero," or "not significant."[3] All such firms are treated as not reporting positive R & D because of both the nature of the SEC reporting requirements for R & D and the way Compustat handles company responses. As noted above, companies are supposed to report "material" R & D expenditures. If the companies and their accountants conclude that R & D expenditures were "not material" (possibly zero but not necessarily), they sometimes say this in the 10-K report, in which case Compustat records "zero."[4] Alternatively, a company may say nothing about R & D, in which case Compustat records "not available." It is also likely that companies reported as "not available" include some which are "randomly" missing, that is, a company performs "material" R & D but for some reason Compustat could not get the number for that year.[5]

Another source of data on aggregate R & D spending by U.S. industry is the National Science Foundation which reports total R & D spending in the United States every year, broken down into approximately thirty industry groupings. These data are obtained from a comprehensive survey of U.S. enterprises by the Industry Division of the U.S. Bureau of the Census, which covers larger firms completely and samples smaller firms. Although there are several important differences between these data and those reported by Compustat, it is interesting to compare the aggregate figures, which we show in table 2.2. The company R & D figures are the most directly comparable to our Compustat numbers, but we also show the figures for total R & D since NSF does not provide a breakdown between company-sponsored and federal-sponsored R & D expenditures for many of the industries (to avoid disclosing individual company data). There are several reasons for the discrepancies between the Compustat

3. The "not significant" code is a 1977 Compustat innovation which appears in 1976 data only for the Full Coverage tape companies.

4. Or, more recently, "not significant." See note 3.

5. Also included in "missing" are companies that reported R & D but Compustat concluded that their definition of R & D did not conform.

and NSF totals. First, the industry assignment of a company is not necessarily the same across the two sets of data: the most striking difference is in the communications industry, which includes AT & T in the NSF/Census sample, while AT & T is assigned to SIC 4800 on the

Table 2.2 **Comparison of Aggregate R & D Spending Reported to Compustat and NSF for 1976 (dollars in millions)**

	NSF[a]			
Industry	Total	Federal	Company	Compustat
Food & kindred products	329	—	—	336
Textiles & apparel	82	—	—	92
Lumber, wood products & furniture	107	0	106	53
Paper & allied products	313	—	—	128
Chemicals & allied products	3017	266	2751	3173
Industrial chemicals	1323	249	1074	1604
Drugs & medicines	1091	—	—	1053
Other chemicals	602	—	—	516
Petroleum refining & extraction	767	52	715	908
Rubber products	502	—	—	346
Stone, clay & glass products	263	—	—	218
Primary metals	506	26	481	302
Ferrous metals & products	256	4	252	151
Nonferrous metals & products	250	22	229	151
Fabricated metal products	358	36	322	186
Machinery	3487	532	2955	2898
Office, computing, & accounting machines	2402	509	1893	2035
Electrical equipment & communication	5636	2555	3081	2543
Radio & TV receiving equipment	52	0	52	119
Electronic components	691	—	—	327
Communication equipment & communication	2511	1093	1418	231
Other electrical equipment	2382	—	—	866
Motor vehicles & motor vehicles equipment	2778	383	2395	2847
Other transportation equipment	94	—	—	54
Aircraft & missiles	6339	4930	1409	851
Professional & scientific instruments	1298	155	1144	1195
Scientific & mechanical measuring instruments	325	6	318	315
Optical, surgical, photographic & other instruments	974	148	826	880
Other manufacturing	217	5	212	93
Conglomerates	—	—	—	563
Total manufacturing	26093	9186	16906	15470

Note: Columns do not add up due to NSF suppression of cells with small numbers of firms.

[a]Source: *Research and Development in Industry, 1977.* Surveys of Science Resources Series, National Science Foundation, Publication no. 79–313.

Compustat files and is therefore not in our sample. Adding the 1976 R & D for AT & T and its subsidiary, Western Electric, to the Compustat communications total would raise it to about $1 billion, not enough to account for the difference.

There are also definitional differences between the Form 10-K R & D and that in the Census survey. The 10-K includes international and contracted out R & D, while these are entered on a separate line of the Census survey.[6] The total amount involved is about $1.7 billion in 1976. This is likely to explain why our industrial chemicals figure is too high, for example. Some firms include engineering or product testing on one survey but exclude it on the other, apparently because the Census survey is quite explicit about the definition of research and development, while the 10-K allows considerably more flexibility. Finally, the coverage of firms in the U.S. manufacturing sector by Compustat is less complete than by the Census for two reasons: (1) privately held firms are not required to file Form 10-K, and (2) some large firms which do file a 10-K record their R & D as not "material" even though a positive figure is reported to the Census Bureau. In spite of all these caveats, the Compustat and NSF numbers do seem to match fairly well across industries, and the total is within 15 percent after correcting for AT & T and the international and contracted out R & D.

Table 2.3 presents some summary statistics for the firms in the sample, broken down into twenty-one industry categories. The categories are based approximately on the NSF applied R & D categories shown in table 2.2, with some aggregation, and the separation of the lumber, wood, and paper, and consumer goods categories from miscellaneous manufacturing. The exact industry category assignment scheme which we used throughout this paper, based on SIC codes, is presented in the appendix. A few firms with exceptionally large or small R & D-to-sales ratios have been "trimmed" from the sample in this table (see below for an exact definition of the criterion used). As the table shows, the population of the industry categories and the fraction of firms reporting R & D varies greatly, from 20 percent for the miscellaneous category to above 80 percent for drugs and computers.

Table 2.4 shows the size distribution of firms in the sample. A large number of small firms are included; there are about seventy firms with less than $1 million in sales, and over six hundred with less than $10 million. These firms, however, account for less than 1 percent of total sales of firms in the sample. As might be expected, larger firms tend to report R & D more often even though they do about the same amount as

6. This comparison of the definitions in the two surveys is drawn from a letter detailing the differences, from Milton Eisen, Chief, Industry Division, U.S. Bureau of the Census, to Mr. William L. Stewart, R & D Economic Studies Section, Division of Science Resources Studies, National Science Foundation, in April 1978.

Table 2.3 **Statistics for the 1976 Cross Section: Trimmed Data**

Industry	NFIRMS	AVEPLANT	AVESALES
Food & kindred products	182	178.7	585.7
Textile & apparel	188	55.2	137.8
Chemicals, excl. drugs	121	503.2	693.6
Drugs & medical inst.	112	116.6	301.7
Petroleum refining & ex.	54	3200.1	4622.8
Rubber & misc. plastics	98	122.4	214.8
Stone, clay & glass	81	186.1	243.6
Primary metals	103	499.6	488.5
Fabric. metal products	196	57.8	131.0
Engines, farm & const. equip.	64	186.9	457.3
Office, comp. & acctg. eq.	106	288.2	352.9
Other machinery, not elec.	199	40.8	116.1
Elec. equip. & supplies	105	155.0	405.5
Communication equipment	258	31.8	89.9
Motor veh. & transport eq.	105	464.2	1233.6
Aircraft and aerospace	37	237.4	754.1
Professional & sci. equip.	139	73.4	130.5
Lumber, wood, and paper	163	204.2	260.4
Misc. consumer goods	100	81.6	232.5
Conglomerates	23	1174.3	2202.3
Misc. manuf., n.e.c.	148	36.3	89.3
All firms	2582	230.9	417.2

Note:
NFIRMS = Total number of firms in industry.
AVEPLANT = Average gross plant in millions of dollars.
AVESALES = Average sales in millions of dollars.
AVEEMP = Average employment in thousands.

a fraction of sales. This is shown graphically in figure 2.1. Up until about $100 million in sales, only about half the companies report R & D, but above $10 billion almost 90 percent do. Previous analysts have suggested that this may be because big companies are able to do their accounting more carefully (San Miguel and Ansari 1975), but it is surprising *how* big a company must be before it has a 75 percent probability of reporting R & D.

As we indicated above, the nature of SEC reporting rules results in ambiguity in the interpretation of firms' reporting zero R & D or not reporting R & D. This ambiguity has implications for the analysis of the subsample of firms that do report R & D ("the R & D sample"). Although we do not believe that the non–R & D sample firms all do zero R & D, it is likely that they do less than the firms that report it. Also, they possibly do less R & D than would be expected, given their other characteristics such as industry, size, and capital intensity. If so, then their exclusion from regressions of R & D on firm characteristics will

AVEEMP	NRNDFIRM	AVERND	AVERATIO	NPATFIRM	AVEPAT
8.9	62	5.4	0.005	46	5.8
4.3	49	1.9	0.018	33	5.9
9.1	92	18.6	0.021	67	39.0
6.8	96	14.4	0.045	64	28.2
20.0	26	34.9	0.005	25	72.0
5.3	59	5.9	0.016	35	12.2
5.3	31	7.0	0.019	26	22.4
8.6	39	7.7	0.013	44	14.6
2.6	102	1.8	0.011	77	5.4
8.8	51	10.2	0.016	42	25.7
8.3	94	21.6	0.061	42	39.0
2.8	149	2.3	0.021	111	5.8
10.7	77	11.2	0.023	56	34.3
2.5	199	3.4	0.040	110	13.3
22.2	59	49.2	0.012	48	25.0
15.6	26	32.7	0.042	17	39.0
3.3	118	8.0	0.051	65	16.0
4.7	64	2.8	0.007	49	6.9
5.2	44	1.8	0.013	41	5.2
50.1	13	43.3	0.014	20	37.3
2.1	29	0.7	0.027	16	2.1
6.8	1479	10.5	0.027	1034	19.1

NRNDFIRM = Number of firms with nonzero R & D.
AVERND = Average R & D expenditure in millions of dollars for firms with nonzero R & D.
AVERATIO = Average R & D to sales ratio for firms with nonzero R & D.
NPATFIRM = Number of firms with nonzero patents.
AVEPAT = Average number of patents for firms with nonzero patents.

result in biased estimates of the association of these characteristics with the firms' propensity to do R & D.

To shed light on this problem, the distribution of reported R & D was examined in several ways. First, if firms consider R & D expenditures to be immaterial if they fall below some absolute amount, then the distribution of R & D would be truncated from below. We find no evidence of such truncation in the R & D distribution. R & D may also be considered immaterial if it is small *relative* to firm size. This seems particularly likely because, in addition to the requirement to report material R & D expenditures in item 1(b)(6) of the 10-K, the SEC requires firms to report *all* expense categories that exceed 1 percent of sales. Figure 2.2 is a histogram of R & D as percent of sales; once again, no truncation is apparent. In fact, the mode of the distribution occurs at about .3 percent of sales.

Although no obvious truncation was visible, either in absolute magnitude or as a percent of sales, we cannot rule out the likelihood that a combination of cutoffs, both absolute and relative (as interpreted by a

Table 2.4 **Size Distribution of Firms**

Size Class (sales in 1976 dollars)	Number of Firms	Number of Firms Reporting R & D	Percent of Firms Reporting R & D	Percent of Total Sales	Percent of Total R & D
Less than 1 million	72	33	46	0.003	0.019
1 to 10 million	545	293	54	0.23	0.42
10 to 100 million	1097	575	53	4.1	3.4
100 million to 1 billion	663	412	62	19.1	14.8
1 to 10 billion	205	167	81	48.3	50.6
Over 10 billion	13	12	92	28.2	30.7

firm's accountants), are in effect, implying an indeterminate bias in the relationship of observed R & D to a firm's characteristics. Therefore, we attempt to quantify the reporting and not reporting of R & D with a probit equation after presenting results for the firms which do report R & D.

In figure 2.3 we show a plot of log R & D vesus log sales for the R & D sample, which summarizes the basic relationship between R & D and firm size in our data. It is apparent from this plot that the slope and degree of curvature of this relationship are likely to be influenced strongly by a few outlying points; some very small firms do large amounts of R & D, and a few firms in the intermediate size range do very little R & D. To test for the sensitivity of the results to these few points, the sample was trimmed by eliminating seven firms (.5 percent) with the lowest R & D/sales ratios, and seven firms with the highest. The firms removed are those outside the diagonal lines drawn on the plot. This reduces the mean ratio of R & D to sales from 4.1 percent to 2.7 percent and the standard deviation from 35 percent to 3.8 percent. The effects on the log distribution are much less dramatic. The smallest ratio that was deleted from the upper tail was .716; the largest from the lower was .0002. These are beyond three standard deviations of even the untrimmed distribution, whether it is viewed as normal or (more plausibly) lognormal. Since the results with trimmed data were not strikingly different from those with untrimmed data, we present only one set of results for our regressions, using the trimmed data throughout.

The first question we investigated in this sample was the nature of industry variation in R & D performance and the relationship between R & D and firm size. Equations of the form

$$\log R = \alpha + \beta \log S + \epsilon, \tag{1}$$

where R is R & D and S is sales, were estimated separately for the twenty-one industries in table 2.3. Except for the textile industry and miscellaneous manufacturing, the estimated betas were not significantly

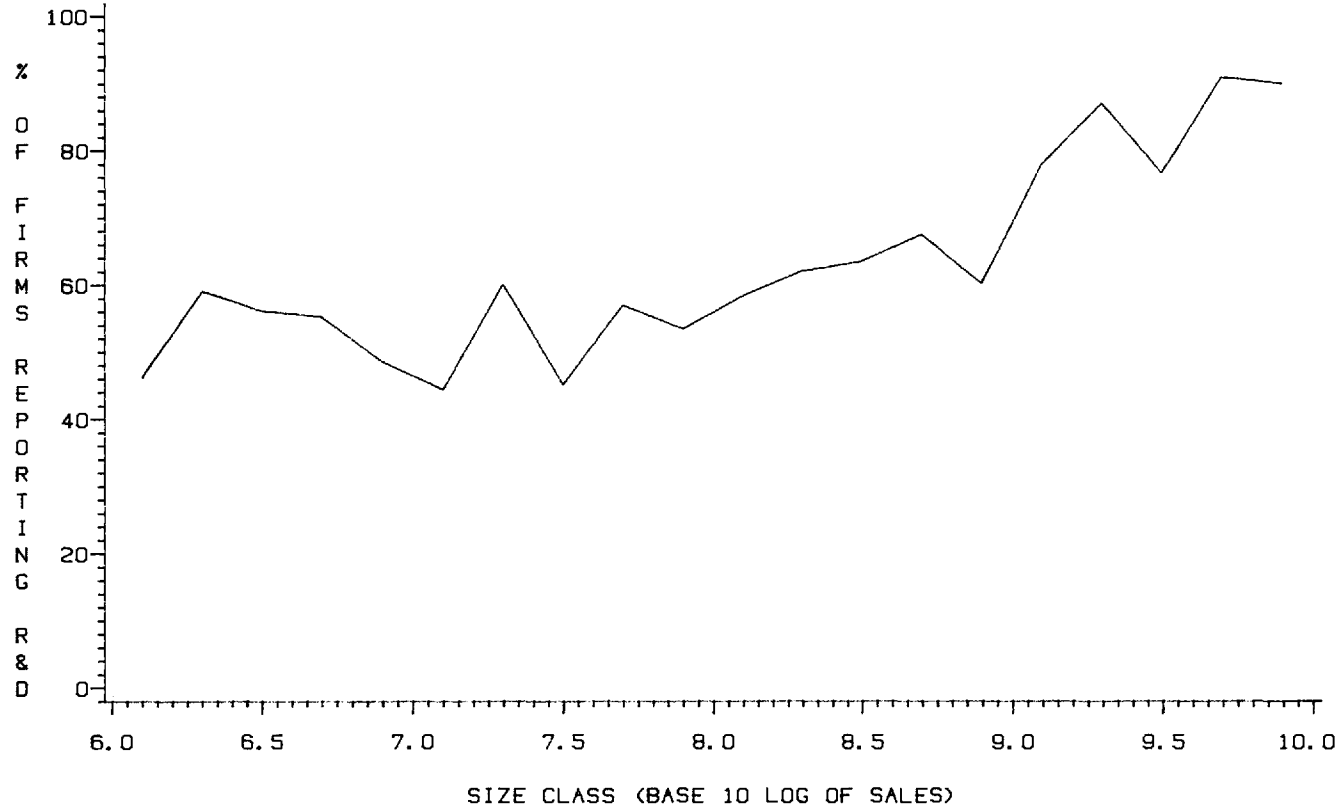

Fig. 2.1 Fraction of firms reporting R & D by size class. Firms with less than $1 million in sales were added to the smallest size class, and those with more than $10 billion were added to the largest.

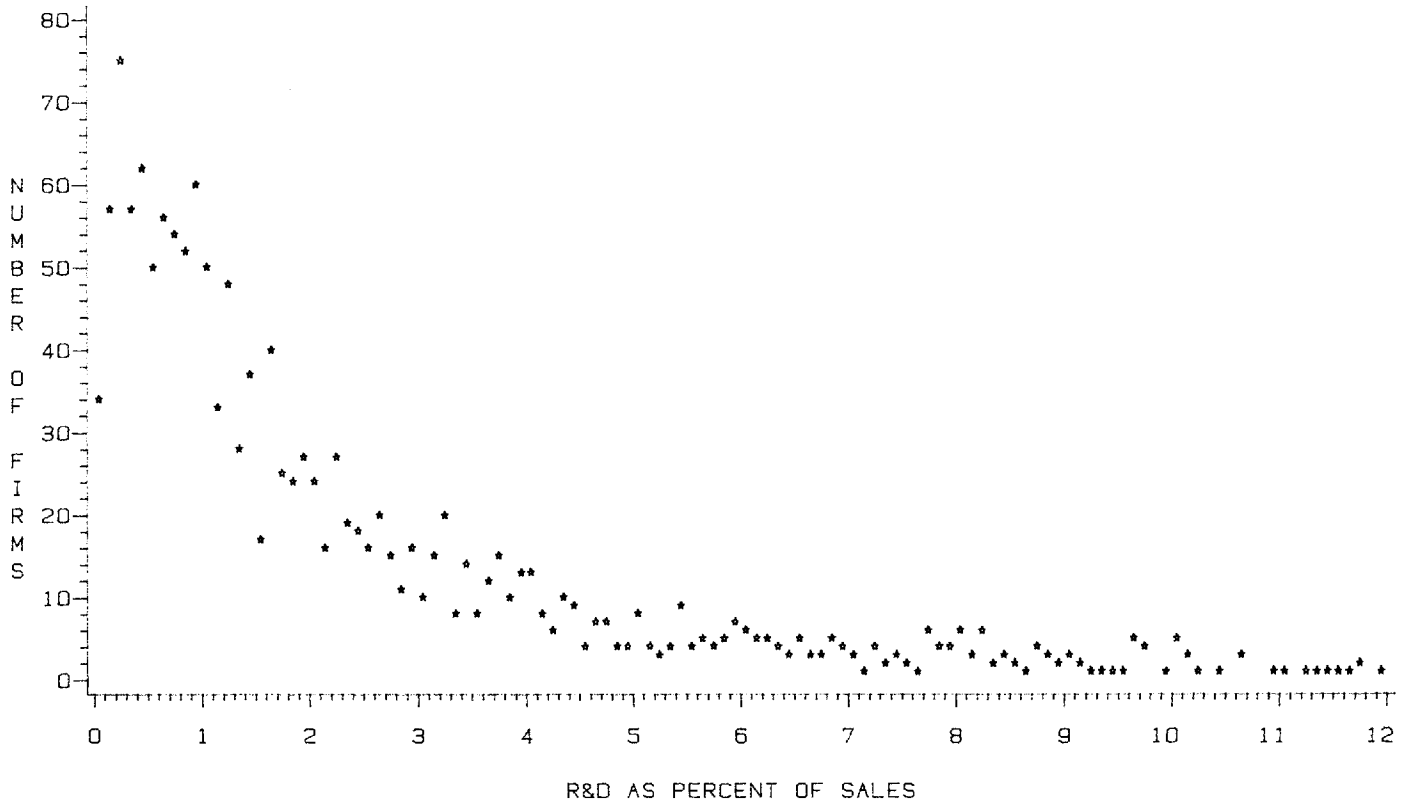

Fig. 2.2 Distribution of R & D as percent of sales for firms reporting R & D. Observations with R & D/sales percentage greater than 12 are not shown.

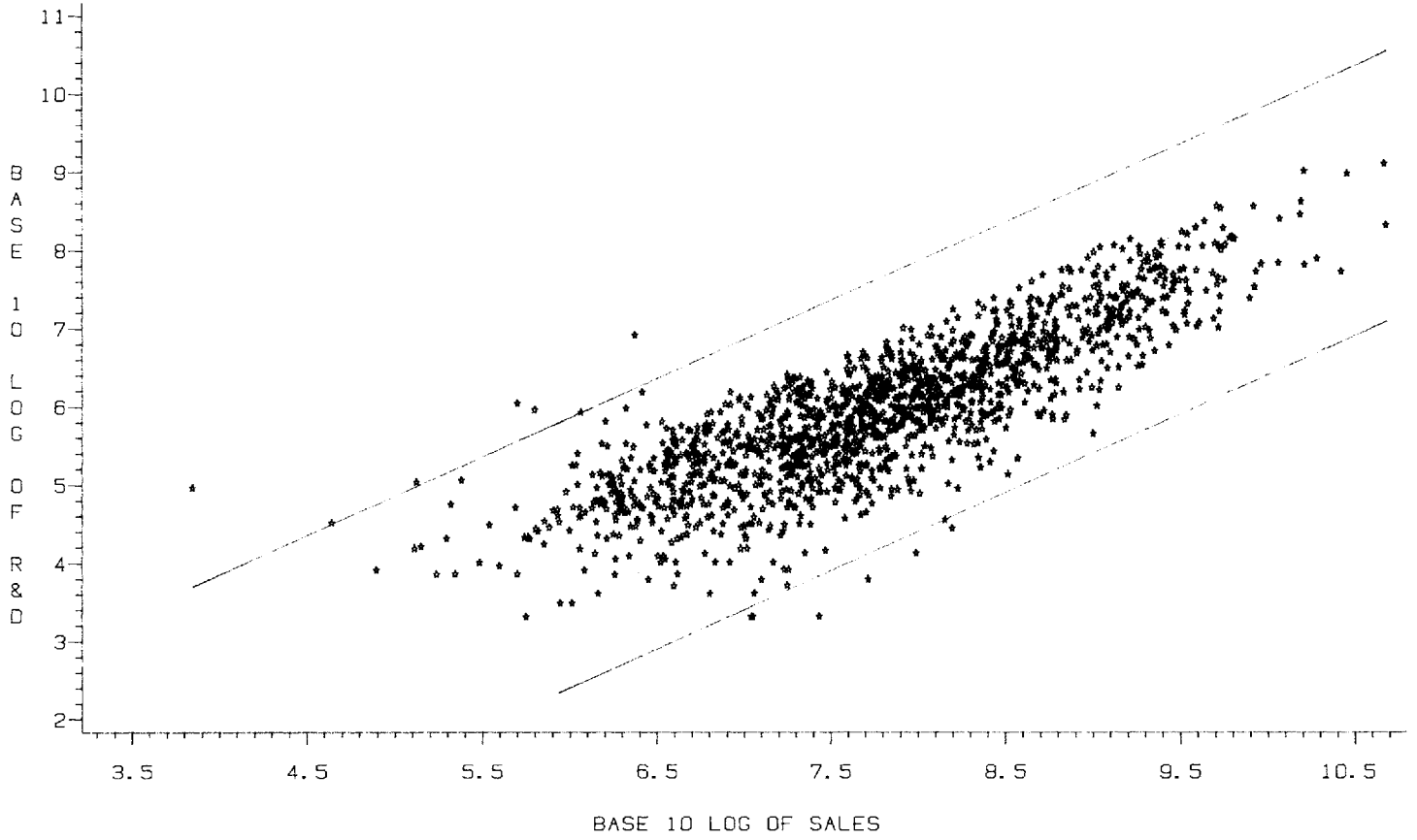

Fig. 2.3 Log(R & D) versus log(sales) for 1976 cross section.

different from one another statistically, and the R-squares were above .65. The remainder of the analysis was performed using uniform slope coefficients, while allowing for different industry intercepts by using industry dummies. This was done primarily for convenience, but it is not inconsistent with the individual industry results. While such aggregation is rejected by a conventional F-test for the simple regression of log R & D on log sales ($F[20, 1437] = 3.34$), given the size of our sample one should really use a much higher critical value (about 8), in which case one need not reject it.[7]

After accepting the hypothesis of equality of the slope coefficients, we estimated equations of the form

$$\log R = \beta_1 \log S + \beta_2 \log A + \beta_3 (\log S)^2 + \gamma_i + \epsilon, \tag{2}$$

where R and S are as previously defined, A is gross plant, and γ_i is a set of industry intercepts. Simple statistics on the regression variables are shown in table 2.5 and basic regression results in table 2.6.

The first column in table 2.6 gives the results of the simplest regression. Although we know that this story is incomplete, this equation indicates almost no fall in R & D intensity with increasing firm size. An analysis of variance using this equation and restrictions on it is also interesting. Log sales explains 73 percent of the total variance in log R & D and 79 percent of the variance remaining after we control for the variations in industry means. Looked at the other way, the industry dummies explain 10 percent of the total variance and 30 percent of the variance remaining after we control for log sales.

The second column shows the effect of capital intensity on R & D intensity. If we interpret this equation in terms of the equivalent regression of log R & D on log sales and log of the capital-sales ratio, we find it implies a sales coefficient of .95, almost identical to that of the first column, and a complementarity between capital intensity and R & D intensity (coefficient of .24 for log [gross plant/sales]). While this effect is highly significant, its additional contribution to the fit is small.

The third and fourth columns in table 2.6 indicate significant nonlinearity in the relationship between log R & D and log sales. These estimates imply that the elasticity of R & D with respect to sales varies from .7 at sales of $1 million to 1.2 at sales of $1 billion. This nonlinearity is also apparant in the scatter plot of log R & D and log sales presented in figure 2.3. While a fairly linear relationship may exist for large firms, it clearly breaks down for smaller firms. This may be a result, at least in part, of the selection bias discussed above; more will be said about this below.

7. Leamer (1978) suggests using critical values for this F-test based on Bayesian analysis with a diffuse prior as a solution to the old problem of almost certain rejection of the null hypothesis with a sufficiently large sample. Using his formula (p. 114), the 5 percent level for this F-test is 7.8, implying that we would accept the hypothesis of equal slopes in these data.

Table 2.5 **Key Variables for the R & D Sample (number of observations = 1479)**

Variable	Mean	Standard Deviation	Minimum[a] (thousands)	Maximum[a] (billions)
Log R & D	−0.15	2.19	$30	$1.3
Log sales	4.10	2.19	$79	$49
Log gross plant	2.99	2.43	$37	$30
R & D/sales	0.026	0.038	0.00024	0.57

[a]The antilogs of the extreme are shown for the first three variables.

Table 2.6 **Log R & D Regression Estimates (observations = 1479)**

				4		
Variable	1	2	3	All Firms	Small Firms	Large Firms
Log sales	.965	.713	.684	.519	.576	.641
	(.013)	(.043)	(.036)	(.050)	(.105)	(.101)
Log gross plant	—	.240	—	.187	.113	.187
		(.039)		(.039)	(.074)	(.046)
(Log sales)^2	—	—	.035	.031	.044	.020
			(.004)	(.004)	(.052)	(.008)
Standard error	.954	.942	.932	.925	.910	
R^2	.813	.818	.821	.824	.832	

Note: All regressions include twenty-one industry dummies, except that for small firms, in which the primary metals and conglomerate dummies were dropped because of lack of firms. There are 319 small firms (less than $10 million in sales) and 1160 large firms.

In the last two columns of table 2.6 we present the results for the fourth regression estimated separately for small firms (up to $10 million in sales) and large firms (all others). The fit is improved slightly; the F ratio for aggregation of the two subsamples is 3.29 (22, 1433). Allowing for differences in the slopes of log sales and log gross plant together diminishes the significance of the log sales squared term, particularly for the small firms.

Our measurement of the contemporary relationship between R & D and sales may be a biased estimate of the true long-run relationship because of the transitory component and measurement error in this year's sales, particularly if we are interpreting sales as a measure of firm size. To correct for these errors in variables bias, we obtained instrumental variable estimates of a regression of log R & D on log sales, log sales squared, and the industry dummies using log gross plant and its square as instruments for the sale variables. The estimated coefficients were .755 (.042) and .028 (.005) for log sales and its square, implying an elasticity of R & D with respect to sales of .985 at the sample mean. This compares to an

elasticity of .972 for equation (3) in table 2.6 and suggests that the errors in variables bias, although probably present, are not very large in magnitude.

As a first step in our attempts to correct for possible bias from nonreporting of R & D, we estimated a probit equation whose dependent variable was one when R & D was reported and zero otherwise. The model underlying this equation is the following: The true regression model for R & D is

$$\log R_i = X_i\beta + \epsilon_{i1}, \tag{3}$$

where X_i is a vector of firm characteristics such as industry and size, and ϵ_{i1} is a disturbance. We observe R when it is larger than some (noisy) threshold value C_i, different for each firm. This model is a variation of the generalized Tobit model, described by many authors; this particular version is in Nelson (1977) and is equivalent to a model described by Griliches, Hall, and Hausman (1978). C_i contains the 1 percent of sales rule and anything else the firm uses to decide whether R & D is "material," plus a stochastic piece, ϵ_2, which describes our inability to predict exactly when a firm will report:

$$C_i = Z_i\delta + \epsilon_{i2}. \tag{4}$$

In this framework, the probability of observing R & D may be expressed as Prob $(\epsilon_1 - \epsilon_2 > Z_i\delta - X_i\beta \mid Z_i, X_i)$. If we assume ϵ_1 and ϵ_2 are distributed jointly as multivariate normal, we get the standard probit model

$$\text{Prob } (R_i \text{ observed}) = 1 - F[(Z_i\delta - X_i\beta)/\sigma], \tag{5}$$

where σ is the variance of $\epsilon_1 - \epsilon_2$, and $F(\cdot)$ is the cumulative normal probability function. Since the probit model is only identified up to a scale factor, we can only estimate δ/σ and β/σ. Deriving the model in this way also reveals what it is we are estimating when we run a probit on this data: presumably Z_i and X_i include many, if not all, of the same variables. For example, if the Z_i were only log sales and the 1 percent rule was being followed, the coefficient δ would be unity, and if the true elasticity of R & D with respect to sales were also unity, the probit equation would yield a sales coefficient of zero. However, if reporting depended only on the absolute amount of R & D performed, then C_i would be a constant, and predicting large R & D would be equivalent to predicting high reporting probability; this hypothesis implies that the coefficients in the probit should be the same as those in the R & D regression (up to a scale factor). Finally, if reporting depends in a more complex way on industry and size of the firm, then no obvious relationship is needed between the coefficients of the probit model and those of the regression.

Table 2.7 **Log R & D Regression Corrected for Selectivity Bias**

Variable	Probit Estimates[a]	Log R & D Regression	
		Uncorrected	Corrected
Log sales	.016 (.051)	.519 (.050)	.536 (.050)
(Log sales)^2	.0018 (.0050)	.031 (.004)	.032 (.004)
Log gross plant	.140 (.039)	.186 (.039)	.246 (.044)
Mills ratio	—	—	.933 (.326)
Standard error	—	.925	.923
R^2	—	.824	.825

Note: All models contain industry dummies.

[a]These are the maximum likelihood estimates of the coefficients in equation (5), the probability of R & D reporting. There are 2582 observations and 1479 report R & D. The χ^2 for the three variables besides the industry dummies is 233.

The results of the probit estimation are presented in the first column of table 2.7. The coefficient on log sales is .016 (.05) compared to .52 (.05) in the comparable ordinary least squares (OLS) equation for log R & D. At the mean of log sales for the whole sample, the coefficient is .077. The coefficient on log gross plant is reduced somewhat from OLS estimates. These results suggest that the first of our two hypotheses above is closer to the truth: R & D reporting depends primarily on R & D intensity and not on the absolute level of R & D spending, with perhaps a smaller effect from firm size.

If it is true that the nonreporting firms are characterized only by lower than average R & D as percent of sales, the OLS estimates of elasticities presented earlier are not necessarily biased, although the constant term and industry dummy coefficients could be. Since it is also true, however, that the nonreporting firms are smaller on average,[8] the OLS elasticity estimates may be biased downward. This possibility was investigated using the procedure popularized by Heckman (1976). For each observation with R & D reported, the "inverse Mills ratio" was calculated as:

$$\text{M} = \frac{f(\hat{u})}{F(\hat{u})}, \tag{6}$$

where $\hat{u}$ is the argument of the probit equation $(Z_i\delta - X_i\beta)/\sigma$ evaluated for this observation's data and the estimated probit coefficients, and $f(\cdot)$ and $F(\cdot)$ are the standard normal density and cumulative distribution functions, respectively. When M is added to the OLS estimations, it "corrects" for selectivity bias.

A regression including the Mills ratio variable is presented in the third

8. Average sales for reporting firms is $620 million, for nonreporting firms, $240 million.

column of table 2.7, together with the "uncorrected" estimates for comparison. The coefficients on the Mills ratio is positive and significant, indicating the presence of selectivity bias. There is only a slight rise in the sales coefficients, however, and the nonlinearity is about the same. The largest increase is in the log gross plant coefficient, which was also the best predictor of R & D reporting. Thus we would underestimate the complementarity of capital intensity and R & D intensity if we did not take into account the fact that non–capital-intensive firms also tend to be those which do not report R & D expenditures.

It should be emphasized that in this application of the Heckman technique the Mills ratios are nonlinear functions of all the other independent variables in the equation, because we have no variables that predict reporting but not quantity of R & D. For this reason, the incremental explanatory power of the M variable is caused solely by the nonlinearity of its relationship to the other variables in the model. We know, however, that the dependence of R & D on these variables is likely to be nonlinear to begin with. In the absence of a reporting predictor that is excluded from the quantity equation, it is impossible to distinguish selectivity bias and "true" nonlinearity in the R & D-size relationship. This makes it impossible to draw a definitive conclusion regarding the possibility of bias in the OLS estimates.

2.4 Patenting

The matching project described in the section 2.1 yielded 4,553 patenting entities which were matched to the companies in our sample. Of our 2582 companies, 1754 were granted at least one patent during the 1965–79 period, but only about 60 percent of that number applied for a patent in 1976. Firms with R & D programs are far more likely to apply for patents: about 20 percent of the firms with zero or missing R & D have at least one patent in 1976, but this fraction rises rapidly with size of R & D program until well over 90 percent of firms with R & D larger than $10 million have patents in 1976.

If we look at the size of the firm rather than the R & D program, 28 percent of the small firms (less than $10 million in sales) applied for a patent in contrast to the 53 percent which reported R & D, but this difference results primarily from the integer nature of the patents data: When we consider all years rather than just 1976, the percentage who patent rises to sixty. These same small firms account for 4.3 percent of sales, 3.8 percent of R & D, but 5.7 percent of patent in our sample. However, the latter number may be an overestimate since we know that approximately one-third of all domestic corporate patents remain unmatched in 1976 in our sample, and it is likely that some of these belong to

subsidiaries of our larger companies which we have overlooked. Further checking of these patents is being done.

In table 2.3 we show the mean number of patents and number of firms which have one or more patents for each of our twenty-one industry classes. As we expect, patenting is higher in the science-based or technological industries in terms of both the fraction of firms which patent and the average number of patents taken out by the patenting firms. The industries with more than twenty-five patents per firm are chemicals, drugs, petroleum, engines, computers, electrical equipment, motor vehicles, aircraft and aerospace, and conglomerates. Presumably petroleum, motor vehicles, and conglomerates appear on this list partly because of the average size of the firms in those industries. On the other hand, the scientific instrument and the machinery industries have a large number of patents per R & D dollar but are composed of relatively small firms.

Earlier studies by Pakes and Griliches (1980) on a sample of 157 large U.S. manufacturing firms show a strong contemporaneous relationship between patent applications and R & D expenditures across firms in several industries, and they suggest that patents are a fairly good indicator of the inventive output of the research department of a firm. We consider the relationship again in figure 2.4. Because of the large size range of our firms, the patents–R & D relationship will be obscured by the simple correlation between number of patents and size of firm. Therefore, we plot the log of patents normalized by gross plant versus the log of R & D normalized by the same quantity for the firms which both do R & D and patent. The plot shows a strong correlation between patenting and R & D for those firms with a slope slightly greater than one and a hint of nonlinearity in the relationship (increasing slope for higher R & D). There is considerable variance: the range of patents per million dollars of R & D for the firms which patent is from about one-seventh of a patent to ninety patents. The typical firm has a ratio of about two, that is, half a million dollars of R & D per patent.[9]

This picture is slightly misleading, however, since it covers only one-third of our sample. Accordingly, when we turn to modeling the relationship, we want to include the zero observations on both patents and R & D in our estimation. We attempt to solve this problem in two ways: First we set log patents to zero for all zero patent observations and allow those firms to have a separate intercept (PATDUM) in our regressions, as suggested by Pakes and Griliches (1980). It should be emphasized that there are about 1700 such observations, which suggest that the significance level of our estimates needs to be interpreted with caution. The

9. Scherer (1981), using data on 443 large industrial corporations comprising 59 percent of corporate patenting activity in the United States, found an R & D cost per patent of $588,000 for the period of June 1976 through March 1977 (adjusted to annual basis).

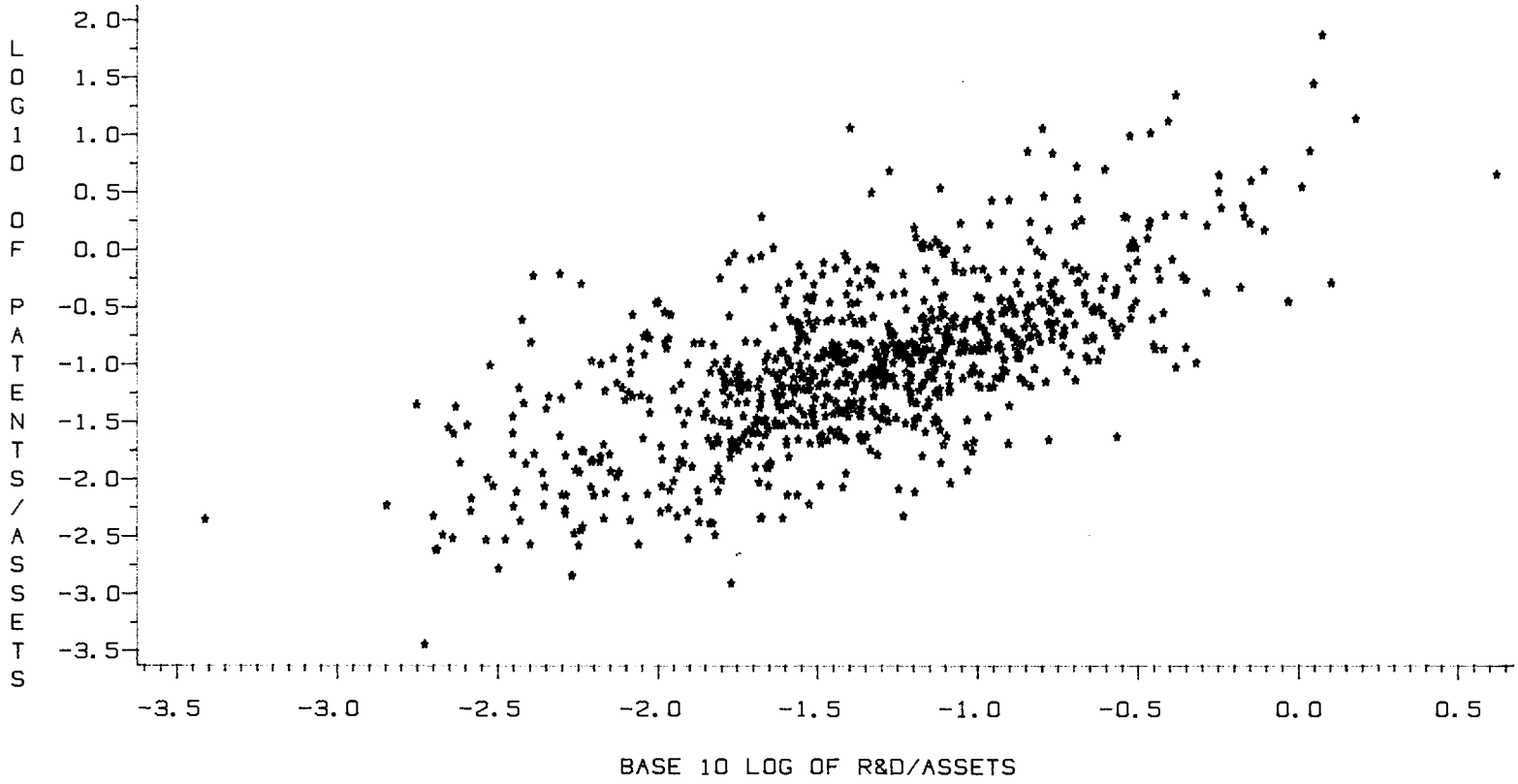

Fig. 2.4 Plots of log(patents/assets) versus log(R & D/assets) for 1976 cross section.

estimates we obtain imply that the observations with no patents have an expected value of about one-half of a patent. Second, we model the patents properly as a counts (Poisson) variable, taking on values 0, 1, 2, etc., as suggested by Hausman, Hall, and Griliches (1984). In this case, with our many small and few very large observations, the Poisson model turns out to give quite different results from the logarithmic OLS model.

The first column of table 2.8 displays the results of a regression of log patents on log R & D expenditures, dummies for zero or missing R & D and patents, and our twenty-one industry dummies. The estimate of the log R & D coefficient is considerably lower than the comparable estimates by Pakes and Griliches (1980), .61 (.08), or by Hausman, Hall, and Griliches (1984), .81 (.02). The difference could be attributed to the size range of firms in our sample which is far greater than in the earlier work and also to the large number of zeroes in our variables. For comparison, the coefficient of log R & D is .59 (.02) when we use only firms with nonzero patents and R & D. The overlap of this last sample of firms with the Pakes and Griliches sample is about 100 firms out of 831, consisting primarily of the larger firms from the complete sample. We will return to the question of how to handle the enormous size range of our complete sample after we discuss the Poisson and negative binomial results for this model.

The industry dummies from the regression in the first column of table 2.8 are a measure of the average propensity to patent in the particular industry, holding R & D expenditures constant. Relative to the overall mean, the industries with significantly higher than average patenting propensity are chemicals, drugs, petroleum, engines, farm and construction machinery, electrical equipment, aircraft, and the conglomerates. Several industries which are highly technology based, such as communications equipment and computers, do not seem to patent any more

Table 2.8 **Log Patents Regressions (number of observations = 2582)**

Variable	1	2	3	4
Log R & D	.38 (.01)	.37 (.008)	.37 (.008)	.32 (.010)
Log gross plant	—	—	—	.064 (.008)
Log R & D squared	—	.083 (.002)	.084 (.002)	.081 (.002)
PATDUM	−.79 (.04)	−.82 (.03)	−.85 (.03)	−.76 (.03)
Other variables included	R&DDUM, industry dummies	R&DDUM, industry dummies	R&DDUM, intercept	R&DDUM industry dummies
Standard error	.713	.589	.595	.583
R^2	.653	.763	.756	.768
Test for industry dummies		$F_{20,2557} = 3.5$		

than the average: in fact, a firm in the computer industry has 85 percent of the patents of an average firm doing the same amount of R & D.

To allow for possible nonlinearity in the patenting-R & D relationship, we add the log of R & D expenditures squared to the regression in column two of table 2.8. This coefficient is highly significant and implies a substantially higher propensity to patent for firms with larger R & D programs, with an elasticity of .25 at R & D of half a million, rising to over unity at R & D expenditures of $100 million. The *F*-test for the industry dummies is now $F(20, 2557) = 3.5$, implying very little difference in the average propensity to patent across industries once we allow R & D to have a variable coefficient. This is a bit surprising and probably reflects the nonhomogeneity of the firms in our industry classes and the problems associated with assigning each firm to one and only one industry. The industries which have coefficients significantly different from the average are the petroleum industry (patenting 30 percent higher on average), engines, farm and construction machinery (28 percent), conglomerates (76 percent), and computers (20 percent less on average). We reestimated the equation with no industry dummies (column three of table 2.8) and found that the slopes hardly changed; this result held true for several different specifications of the model, including one with only the log of R & D in the equation.[10] Although we believe that there are significant differences in the relationship of R & D and patenting at the detailed industry level from inspection of the distribution of the two variables by industry, these differences do not affect the basic results of this aggregate study. We have therefore omitted the industry dummies for the sake of simplicity in what follows.

In the fourth column of table 2.8 we add the log of gross plant value to the regression to control for firm size independently of R & D expenditures. Larger firms may patent more often simply because they are bigger and employ patent lawyers and other personnel solely for this purpose. The coefficient estimate for log gross plant lends some support to this hypothesis. However, one should be careful in interpreting the estimated size (assets) effects. To a significant extent they may be just compensating for transitory and timing errors in our R & D measure. The equation estimated assumes that this year's patents applied for depends only on this year's R & D expenditures. We know that this is not exactly correct (see Pakes and Griliches, this volume). Some of the patents applied for are the result of R & D expenditures in years past, while not all of the R & D expenditures in this year will result in patents, even in subsequent

10. We also looked at this question for two different size classes of firms: above and below $100 million in gross plant. We found that the smaller firms had a lower R & D coefficient (.26 in contrast to .36) and slightly less curvature. For the smaller firms, the industry dummies were completely insignificant, whereas they remained at about the same level for the large firms.

years. In this sense, the R & D variable is subject to significant error which will be exacerbated once we control for size, thereby reducing the signal-to-noise ratio. This may explain both the reduction of the R & D coefficient when assets are introduced as a separate variable and the rather large estimated pure size effect. We cannot do much about this in this paper, but we shall return to this topic when we turn to the panel aspects of this data set in later work.

We now turn to the Poisson formulation of the patents model. This model treats the patents for each firm as arising from a Poisson distribution whose underlying mean is given by $\exp(X\beta)$, where $X\beta$ is a regression function of the independent variables in our model. Coefficients estimated for this model are directly comparable to those from a log patents regression; we have merely taken account of the fact that the dependent variable is nonnegative counts rather than continuous. However, for our data we might expect the Poisson formulation of the model to give quite different answers from a simple log patents regression for two reasons: First, over half of our observations on patents are zero, and many are quite small. Second, the Poisson objective function tends to give the largest observations more weight than least squares on log patents, therefore these observations will have more influence on the results. This is what we find in our results, which are shown in the second column of table 2.9, together with the OLS estimates for comparison. The OLS estimates imply an elasticity of patenting with respect to R & D which rises from zero at \$100,000 of R & D to well above one at \$1 billion. For the Poisson model, on the other hand, the elasticity is one at \$4 million of R & D and falls to one-half at \$1 billion. This is because the very largest firms do less patenting per R & D dollar than would be predicted by a linear regression of log patents on log R & D, and they are having more influence on the Poisson estimates than the OLS. We show this graphically in figure 2.5: What is plotted is the predicted logarithm of patents versus the logarithm of R & D expenditures, superposed on the actual data. Clearly the differences in fit of the models are most pronounced in the tails of the distribution.

As was pointed out by Hausman, Hall, and Griliches (1984), the Poisson model is highly restrictive, since it imposes a distribution on the data whose mean is equal to its variance. This property arises from the independence assumed for the Poisson arrival of "events" (patent applications) and is unlikely to be true, even approximately, of our data. One way out of this problem is the negative binomial model in which the Poisson parameter is drawn from a gama distribution with parameters $\exp(X\beta)$ and δ. We estimated such a model in the third column of table 2.9 and found that the results, although qualitatively closer to the OLS estimates than to the Poisson, produce quite different predictions over the range of the data and imply a lower and less varying elasticity of

Table 2.9 **Comparison of Patents Models (number of observations = 2582)**

Variable	OLS Log $P = X\beta + u$	Poisson	Negative Binomial	Nonlinear Least Squares $P = \exp(X\beta) + \epsilon$
Log R & D	.37 (.008)	1.13 (.010)	.58 (.018)	2.18 (.10)
Log R & D squared	.084 (.002)	−.047 (.002)	.012 (.003)	−.16 (.009)
Dummy (R & D = 0)	−.11 (.03)	−.43 (.01)	−1.37 (.08)	1.85 (..58)
Constant	.97 (.02)	.61 (.02)	−1.33 (.04)	−1.67 (.26)
D(patents = 0)	−.85 (.03)	—	—	—
$\hat{\delta}$	—	—	.059 (.0016)	—
Standard error	.595	—	—	20.57
Log likelihood	—	56,171.	63,558.	—

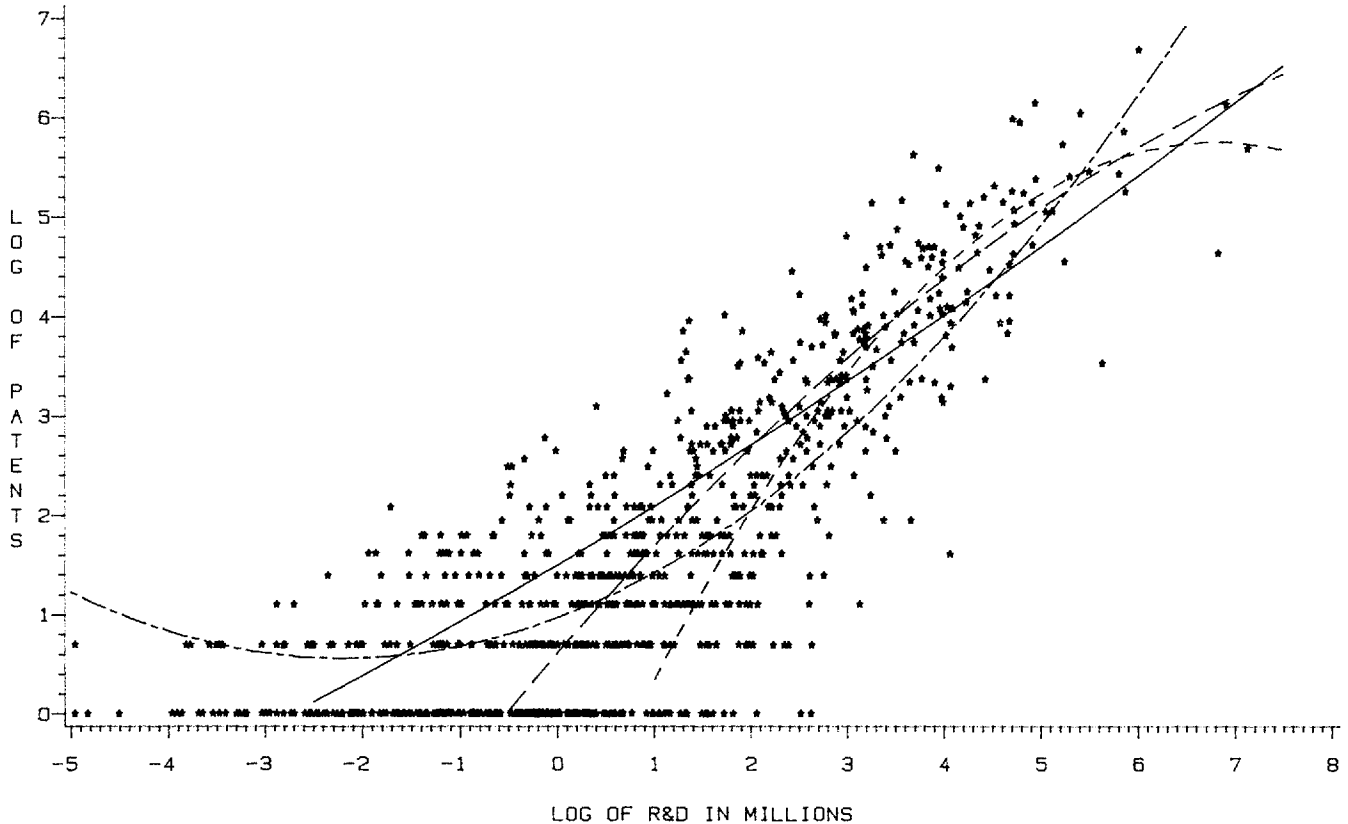

Fig. 2.5 Predictions for models with log R, $(\log R)^2$, $D(R = 0)$, no industry dummies, 831 observations plotted (total = 2.582)(*** = data, — = NB, --- = NLLS, ———— = OLS, —— = Pois).

patenting with respect to R & D. The range of elasticities is now .55 at $100,000 in R & D to .66 at $1 billion. A typical firm with zero or missing R & D is predicted to have applied for 1.3 patents in 1976, as opposed to 2.4 under the OLS model.

A defect of the negative binomial model is that it imposes a specific distribution, namely gamma, on the multiplicative disturbance. Unlike the least squares case, if this distribution is wrongly specified, the resulting maximum likelihood estimates may be inconsistent. For this reason and because of the large swings in our estimates under the models we tried, we also estimated our model with nonlinear least squares using patents as the dependent variable, which was proved by Gourieroux, Monfort, and Trognon (1981) to be consistent for a wide class of Poisson-type models. This produced the result shown in the last column of table 2.9. The discrepancies between these estimates and those of the Poisson model are a kind of "specification" test, since both are consistent estimates of a large class of count models with additive or multiplicative disturbances. Our data, however, have one feature which violates the assumptions of most of these models: not only is the residual variance of patents larger than the mean, but the ratio increases as the magnitude of the exogenous variables (R & D) increases (see Hausman, Hall, and Griliches 1984). This implies a correlation between the X's in the model and the disturbance which can lead to inconsistent estimates of the slope parameters. Figure 2.5, which displays the nonzero portion of the data distribution with the predictions for our various specifications superimposed, reveals that in trying to impose a quadratic on our data to look for scale effects we may mislead ourselves seriously because of the very large range of the data and the peculiar distribution of the dependent variable. It appears that the form we choose for the error distribution of the patents variable will have a considerable effect on the results. It should be emphasized that this result does not depend only on the large number of zero observations in the data: we obtained qualitatively the same results when we reestimated, including only those firms with both nonzero patents and R & D.

Because of the increasing variance with R & D and the difficulty of choosing a proper functional form for both tails of the distribution simultaneously, we chose to look at the interesting questions in this data (the existence of a patenting threshold and the measurement of returns to scale at the upper end of the R & D distribution) by dividing the sample into two parts, using R & D as the selection variable. To do this we first plotted the patents–R & D ratio for firms with both patents and R & D grouped by R & D size class, as shown in figure 2.6. This plot is consistent with a patenting elasticity of considerably less than one up to about $1 or $2 million of R & D and an elasticity of about one after that, with a hint of

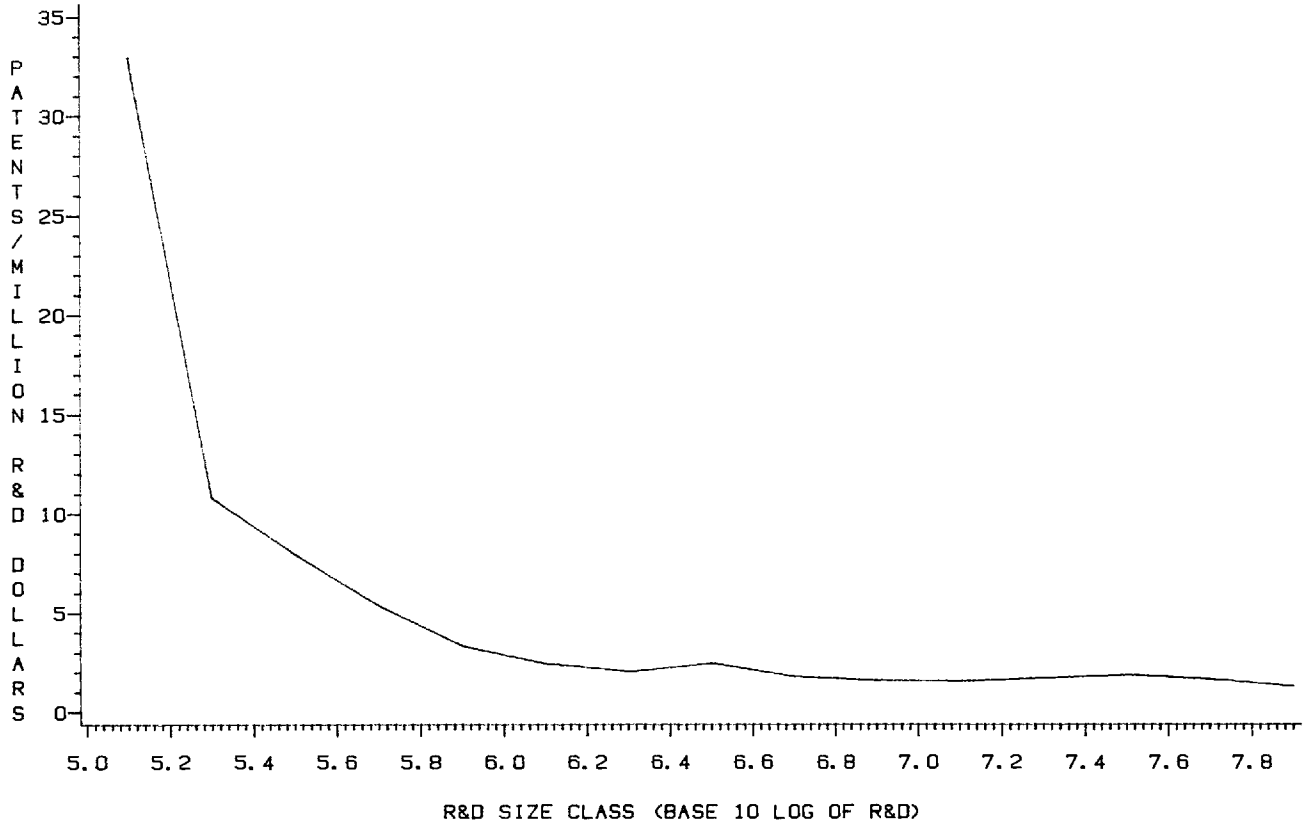

Fig. 2.6 Patents per million R & D dollars by R & D size class for firms with both R & D and patents.

downturn at the upper end (above $100 million). Accordingly, we divided our sample into two groups: those with R & D greater than $2 million and those with R & D less than $2 million or missing.

The coefficients of interest from estimates on the two groups of firms are shown in table 2.10 and the differences between them are striking. The small firms show both the features we might have expected: the Poisson-type models all are quite different from OLS on log patents, since most of these firms have less than five patents, and the estimates are all much closer to each other, since the problem of inconsistency arising from the increasing variance of patents is considerably mitigated. Substantively, there is no real evidence of curvature in the relationship of R & D and patents at this end of the distribution, and the elasticity of patenting with respect to R & D is close to the earlier estimates for large firms, albeit not very well determined.

Turning to the larger firms, as we might expect, since the range of R & D is about ten times that of the smaller firms, there is considerably more variation in the estimates. The log patents regression estimates are much closer to the others, since the integer nature of the patents data is not much of a problem here. However, there does seem to be some evidence of a decrease in the elasticity of patenting with respect to R & D for the largest firms. The Poisson and nonlinear least squares estimates exhibit increasing returns up to about $20–40 million of R & D and then start declining, whereas the OLS and negative binomial estimates show decreasing returns with a slightly higher elasticity than the smaller firms throughout. It is clear, however, that we have not really solved the specification problem for these large firms. The predicted values from these estimates exhibit nearly the same sensitivity to exactly how we weight the observations as did those from the whole sample. Our tentative conclusion is that there are nearly constant returns to scale in patenting throughout the range of R & D above $2 million, with decreasing returns setting in some place above $100 million.

2.5 Conclusion

We began this paper with a question: Who does R & D and who patents? We can now provide at least a partial answer. We have seen that research and development is done across all manufacturing industries with much higher intensities in such technologically progressive industries as chemicals, drugs, computing equipment, communication equipment, and professional and scientific instruments. We have found an elasticity of R & D with respect to sales of close to unity, but we also found significant nonlinearity in the relationship, implying that both very small and very large firms are more R & D intensive than average-size

Table 2.10 **Estimates for Two R & D Size Classes**

Variable	OLS $\log P = X\beta + u$	Poisson	Negative Binomial	Nonlinear Least Squares $P = \exp(X\beta) + \epsilon$
		Small Firms (N = 2102)		
Log R & D	.10 (.03)	.62 (.06)	.49 (.10)	.58 (.32)
Log R & D squared	.017 (.006)	.014 (.020)	.026 (.030)	−.004 (.193)
Elasticity (R & D = \$100k)	.02	.56	.37	.60
Elasticity (R & D = (\$2M)	.12	.64	.53	.57
		Large Firms (N = 480)		
Log R & D	.90 (.12)	1.57 (.02)	.90 (.08)	2.19 (.22)
Log R & D squared	−.003 (.02)	−.098 (.003)	−.034 (.009)	−.16 (.02)
Elasticity (R & D = \$2M)	.90	1.43	.85	1.97
Elasticity (R & D = \$100M)	.65	.67	.59	.72

firms.[11] This effect remained after an attempt to account for the (possibly) nonrandom selection of the dependent variable, although the lack of an exclusion restriction in this procedure casts some doubt on the completeness of this correction. We also found evidence of complementarity between capital intensity and R & D intensity, which was increased when we corrected for the selectivity of R & D.

These results are contrary to the preponderance of previous work on the size–R & D intensity relationship.[12] Hamberg (1964) and Comanor (1967) found a weakly decreasing relationship between R & D intensity and firm size. Scherer (1965a) found that R & D intensity increased with firm size up to an intermediate level, and then decreased (except in the chemical and petroleum industries, in which it increased throughout). This has been interpreted to imply, for most industries, a threshold size necessary before R & D is performed, presumably because of fixed costs in performing R & D (Kamien and Schwartz 1975). As noted above, our results suggest the opposite, though the selectivity issue precludes a definitive conclusion. In any case, these data cast strong doubt on the existence of any significant R & D threshold.[13]

There are several possible reasons for these conflicting results. First, earlier studies were based on small samples of larger companies of the *Fortune* 500 variety. An attempt was made to approximate these samples by estimating equation (3) of table 2.6 on those firms with sales of $500 million or more (256 observations). This regression indicates that this sample difference is not the source of the discrepancy; the relationship was close to linear with an implied elasticity of R & D with respect to sales of 1.23 at sales of $1 billion.[14]

In addition, our R & D variable is an expenditure variable, whereas much of the previous work used the number of R & D employees. If R & D expenditures per research employee rise fast enough with increasing firm size, perhaps because of greater capital intensity of R & D, we would expect the observed difference in the results. It is not possible, with these data, to test this hypothesis.

Finally, it is possible that the size–R & D intensity relationship has changed since the earlier work was done.[15] Because that work did not look at small firms at all, it would be sufficient to postulate increased

11. It should be emphasized, however, that our finding of increasing R & D intensity as firm size rises does not necessarily imply returns to scale in R & D unless one assumes homogeneity of some degree in the R & D production function (see Fisher and Temin 1973, 1979).

12. See Kamien and Schwartz (1975) for a summary.

13. These data also do not support the existence of a peculiar size–R & D intensity relationship in the chemical or petroleum industries.

14. The coefficients (standard errors) were: log sales: 1.29 (.61); log sales squared: −.008 (.038).

15. Hamberg used 1960 data; Comanor used 1955 and 1960 data; Scherer used 1955 data.

relative R & D intensity by the largest firms to reconcile their results with ours. We hope that our examination of the time-series component of this data set will shed some light on this question.

Turning to the second question in our title, we have found that some, but not all, of the firms which do R & D also patent, and that there is a strong relationship between the two activities throughout our sample. The small firms which do R & D tend to patent more per R & D dollar than larger firms, and firms with R & D programs larger than about $1 or $2 million have a nearly constant ratio of patenting to R & D throughout the sample, except for the firms with the very largest R & D programs.

Previous research on the relationship of R & D and patenting, in particular Scherer (1965b), has tended to focus on the largest U.S. corporations. Scherer found an elasticity of patenting with respect to R & D *employment* of unity with a hint of diminishing returns at the highest R & D input intensity. Our data do not contradict this result, but they do suggest that for these larger firms the elasticity of patenting with respect to R & D may have fallen slightly between 1955 and 1976. However, measurement issues cloud this conclusion since we are relating contemporary R & D expenditures and successful patent applications, while Scherer looks at patents granted and the number of R & D employees (lagged by four years). It is not easy to say a priori which relationship will be most free of noise, and we must wait for time-series studies to give us a better reading on the precise relationship of the two variables. Work thus far (Pakes and Griliches 1980; Pakes 1981) has shown a strong contemporaneous relationship of R & D and patent applications, but it has also found a total elasticity closer to one when lagged R & D is included.

These data also confirm and extend what others, including Scherer, have observed: a higher output of patents per R & D dollar for smaller firms. However, our results are for many more smaller firms than previously, and they show much sharper decreasing returns both in the measured elasticity and in the basic patents-to-R & D ratio. We also found that for this sample it mattered very much whether we used a model and estimation method which allowed for zero-valued observations.

In looking at these results on smaller firms, however, it is important to emphasize that although we include all manufacturing firms in our sample, whether or not they do R & D or patent, another kind of selectivity is at work: for a smaller firm, whether or not it appears on the Compustat file in the first place is a sign of success of some sort, or of a need for capital. The basic definition which gets a firm into the sample (if it is not automatically included as a result of being traded on a major stock exchange) is that it "commands sufficient investor interest." One of the likely causes of interest is a successful R & D program, and hence some

patent applications. Thus we tend to observe small firms only when they have become "successful," whereas almost all large firms are publicly traded and will appear in our sample whether or not they have been particularly successful recently in research or innovation. We find it difficult to argue purely from this data that small firms have a higher return to R & D when we have reason to believe that only those which are successful at R & D are likely to be in our sample in the first place.

This is our first exploration of this rather large and rich data set. We hope to focus in the near future on the time-series characteristics of these data. We expect to be able to construct a consistent set of data for at least seven years (1972–78) for over a thousand firms. This should allow us to investigate more thoroughly some of these same questions and also many other aspects of R & D and patenting behavior.

Appendix

Composition of Industry Classes

Industry	Included SIC Groups
Food and kindred products	20
Textiles & apparel	22, 23
Chemicals, excluding drugs	28, excluding 2830, 2844
Drugs & medical instruments	2830, 2844, 3841, 3843
Petroleum refining & extraction	29
Rubber & misc. plastics	30
Stone, clay, and glass	32
Primary metals	33
Fabricated metal products	34, excluding 3480
Engines, farm & construction equipment	3510–3536
Office, computers, & accounting equipment	3570, 3573
Other machinery, not electric	35, excluding 3510–3536, 357
Electric equipment & supplies	36, excluding 3650–3679
Communication equipment	3650–3679
Motor vehicles & transportation equipment	37, excluding 3720–3729, 3760
Aircraft & aerospace	3720–3729, 3760
Professional & scientific equipment	38, excluding 3841, 3843
Lumber, wood & paper	24, 25, 26
Miscellaneous consumer goods	21, 31, 3480, 3900–3989
Miscellaneous manufacturers, n.e.c.	27, 3990

References

Comanor, W. S. 1967. Market structure, product differentiation, and industrial research. *Quarterly Journal of Economics* 81, no. 4:639–57.

Fisher, Franklin M., and Peter Temin. 1973. Returns to scale in research and development: What does the Schumpeterian hypothesis imply? *Journal of Political Economy* 81, no. 1:56–70.

———. 1979. The Schumpeterian hypothesis: Reply. *Journal of Political Economy* 87, no. 2:386–89.

Gourieroux, C., A. Monfort, and A. Trognon. 1981. Pseudo maximum likelihood methods: Applications to Poisson models. Université Paris IX, CEPREMAP, and ENSAE, December.

Griliches, Zvi, Bronwyn H. Hall, and Jerry A. Hausman. 1978. Missing data and self-selection in large panels. *Annales de l'INSEE* 30–31 (April–September): 137–76.

Hamberg, D. 1964. Size of firm, oligopoly, and research: the evidence. *Canadian Journal of Economics* 30:62–75.

———. 1967. Size of enterprise and technical change. *Antitrust Law and Economics* 1, no. 1:43–51.

Hausman, Jerry A., Bronwyn H. Hall, and Zvi Griliches. 1984. Econometric models for count data with an application to the patents–R & D relationship. *Econometrica*, forthcoming.

Heckman, James J. 1976. The common structure of statistical models of truncation, sample selection, and limited dependent variables and a sample estimator for such models. *Annals of Economic and Social Measurement* 5: 475–92.

Kamien, Morton I., and Nancy L. Schwartz. 1975. Market structure and innovation: A survey. *Journal of Economic Literature* 13:1–37.

Leamer, Edward E. 1978. *Specification searches: Ad hoc inferences with nonexperimental data*. New York: Wiley.

National Register Publishing Company. 1972. 1976. *Dictionary of corporate affiliations*. Skokie, Illinois.

National Science Foundation. 1979. *Research and Development in Industry, 1977*. Surveys of Science Resources Series, Publication no. 79–313.

Nelson, Forrest D. 1977. Censored regression models with unobserved stochastic censoring thresholds. *Journal of Econometrics* 6:309–22.

Pakes, Ariel. 1981. Patents, R & D, and the one period rate of return. NBER Discussion Paper no. 786.

Pakes, Ariel, and Zvi Griliches. 1980. Patents and R & D at the firm level: A first report. *Economics Letters* 5:377–81.

San Miguel, Joseph G., and Shahid L. Ansari. 1975. Accounting by business firms for investment in R & D. New York University, August.

Scherer, F. M. 1965a. Size of firm, oligopoly, and research: A comment. *Canadian Journal of Economics* 31, no. 2:256–66.

———. 1965b. Firm size, market structure, opportunity, and the output of patented inventions. *American Economic Review* 55, no. 5:1097–1125.

———. 1981. Research and development, patenting, and the microstructure of productivity growth. Final report, National Science Foundation, grant no. PRA-7826526.

Standard and Poor's Compustat Service, Inc. 1980. *Compustat II.* Englewood, Colorado.