

# The Economics of Innovation and Intellectual Property

## Chapter 6: Exercise 1

Bronwyn H. Hall & Christian Helmers  
2024

The data for this exercise are available in the files `chapter6_1_data.dta` (for Stata) and `chapter6_1_data.xlsx` (MS Excel).<sup>1</sup> There are 11 variables:

Variable	Description
<code>id</code>	firm ID number
<code>year</code>	Year of data
<code>sic2d</code>	2 digit SIC with a few 3-digits
<code>hmsect</code>	Industry at the quasi 2-digit level
<code>sales</code>	Real sales (\$M)
<code>emply</code>	Fulltime employment (1000s)
<code>netplt</code>	Real net P&E, capital stock, end of year (\$M)
<code>rstock</code>	Real R&D capital, end of year (\$M)
<code>rnd</code>	Real R&D expenditures (\$M)
<code>inv</code>	Real capital expenditures (\$M)
<code>shmat</code>	Materials/sales for 2-digit industry of firm

These data are for an unbalanced panel of 1,128 publicly traded U.S. manufacturing firms. `hmsect` is a broad industry classification useful for analyzing R&D spending by firms in the manufacturing sector:

---

<sup>1</sup>The data are a cleaned selection from the data used in Mairesse, J. and B. H. Hall (1996). Estimating the productivity of research and development in French and US. manufacturing firms: An exploration of simultaneity issues with GMM methods. K. Wagner and B. van Ark (eds.), *International Productivity Differences and Their Explanations..* Amsterdam, Elsevier-North Holland: 285-315.

hmsect	Industry
1	Paper & printing
2	Chemicals
3	Rubber & plastics
4	Wood & miscellaneous mfg
5	Primary metals
6	Fabricated metal
7	Machinery
8	Electric machinery
9	Autos
11	Textiles & leather
12	Pharmaceuticals
13	Food & beverage
14	Computers and instruments

The production function model including R&D capital is the following:

$$s_{it} = a_t + \alpha c_{it} + \beta l_{it} + \gamma k_{it} + \epsilon_{it} \quad (1)$$

where  $i = 1, \dots, N$  is the firm and  $t = 1, \dots, T$  indicates time.  $s$  is sales,  $c$  is capital stock,  $l$  is labor, and  $k$  is R&D capital, all in logs. Each year has a separate intercept, denoted  $a_t$ , which will be included in the equation as time or year dummies.

The challenge in estimating this equation is what we assume about the disturbance  $\epsilon_{it}$ . Consistency and efficiency of ordinary least squares (OLS) requires that  $\epsilon_{it}$  be homoskedastic and serially uncorrelated as well as unrelated to capital, labor, and R&D capital. All this is somewhat unlikely. For consistency alone, we merely require that  $\epsilon_{it}$  be uncorrelated with the right hand side variables. So we will suggest a number of ways to estimate the equation in this exercise. Proceed as follows: [**Note:** you can carry out the statistical analysis using your preferred statistical software (e.g. Stata, R, MS Excel, Python, etc.)]

1. Report sample statistics (number of observations, mean, median, standard deviation, min max) for the log and level variables.
2. Compute the ratio of **rnd** to sales for each firm, and the means and medians of this variable by **hmsect**, the industry classification, and by year. Do they differ across industries in the ways you expect? How does the R&D to sales ratio evolve over time?
3. Perform the following basic estimations of the model (1) above and report the results in a table:
  - (a) Ordinary least squares for equation (1). Don't forget to control for the year and industry, as we expect average productivity to differ across these. (Optional: cluster the standard errors by firm)

- (b) Rewrite the equation by subtracting `log empty` from `log sales`, `log netplt`, and `log rstock`, and re-estimate it. What is different and what is the same, compared to the results in (a)?
4. Describe how you would compute the elasticity of output with respect to R&D capital using the equation above and compute it from your estimates.
  5. Describe how you would compute the rate of return to R&D capital using the equation above and compute this quantity from your estimates. Note that there are multiple ways you might approximate this quantity using the data available.
  6. How would you expect the two quantities you computed in questions 4. and 5. to vary across industrial sector?