Quantitative Economics

Tutorial 4: Applied Regression

1. **The Experimental Approach in Science**

   (a) Read section 1.1. (*The experimental approach in science*, pp. 106-117) in List and Rasul (2011), and briefly summarise the potential shortcomings of social experiments. Available here.

   (b) Read "How Do Randomized Experiments Contribute to Educational Research?" [Sandy Baum and Michael McPherson; June 5, 2012; *The Chronicle of Higher Education*], and comment on what we *can* and *cannot* learn from RCTs. Available here.

   (c) Economists (and those working in development economics in particular) are running thousands of RCTs around the world. While many emphasize their *pros*, others highlight their *cons*. Watch Angus Deaton's talk about the limits of RCTs. What are the main limitations of RCTs? Available here.

2. Two economists have run an expensive experiment in a large city to estimate the *income elasticity* of food consumption. The experiment consisted in making non-negative income transfers, from 0 to 10,000 monetary units in intervals of 100 monetary units each, to randomly selected households –*one and only one* individual in each household was allowed to receive an income transfer. The experiment randomly selected 10,000 households: households that were randomly assigned the treatment of a 0-income transfer constitute the *control* group; the other households are treated with different intensity, from 100 monetary units to 10,000 monetary units. After randomly selecting the households to receive each type of transfer, the two economists went to each household and collected information on the household head (age, years of completed education, height) and the household (household size, household income, money spent on food in the household in the last week). At the end of the survey, the household received the income transfer if positive. Otherwise, nothing happened. One week later, after the income transfer, they went back and asked about money spent on food in the household in the last week.

   (a) Table 2.1 displays the results of a regression of transferred income on household (and household head) characteristics, including the $F - statistic$ that all the slope coefficients are zero. What is the interpretation of the $F - statistic$? Explain.

   (b) Table 2.2 displays the results of a regression of food consumption after the income transfer on the income transfer *without* and *with* control variables. Construct a 95% confidence interval for the impact of transferred income on food consumption *without* control variables and interpret it.

   (c) Given your answer to question (a), why do you think our two economists included control variables in Table 2.2? Construct a 95% confidence interval for the impact of transferred income on food consumption *with* control variables. Compare this confidence interval with the one constructed in (b), and explain the intuition behind the conclusion from your comparison.

   (d) Our two economists are interested in the income elasticity of food consumption. A reviewer suggests to re-estimate the regression in column (1) of Table 2.2 in logs, that is using log(food consumption) as the outcome variable and log(income transfer) as the treatment variable. Is there any limitation with such an approach? Can you think of an alternative way to estimate the income elasticity of food consumption? Explain.

Table 2.1: Regression of transferred income on household characteristics

Age	0.000
	(0.009)
Years of education	0.001
	(0.008)
Household size	0.002
	(0.004)
Height	−0.001
	(0.005)
Income before the income transfer	0.020
	(0.021)
Food consumption before the income transfer	0.003
	(0.018)
$F - statistic$	0.07
Number of observations	10,000

Note: The regression includes a constant term, not reported in the table. Standard errors are reported in parentheses.

Table 2.2: Regression of food consumption after the income transfer on transferred income (and control variables)

	(1)	(2)
Income transfer	0.659	0.642
	(0.123)	(0.079)
Age		0.121
		(0.034)
Years of education		0.041
		(0.008)
Household size		0.101
		(0.026)
Height		0.051
		(0.019)
Food consumption before the income transfer		0.876
		(0.121)
$F - statistic$		1245.16
R^2	0.14	0.34
Number of observations	10,000	10,000

Note: Each regression includes a constant term, not reported in the table. Standard errors are reported in parentheses.

3. **Ordinary Least Squares: Reviewing its mechanics.** The file `mroz2016` (in `.dta`, `.RData` or `.xls`) contains information on wages, education, and age for women and men in the US, and the unemployment rate in their county of residence. The description of the variables is as follows:

`age`: Wife's age
`educ`: Wife's educational attainment, in years
`wage`: Wife's 1975 average hourly earnings, in 1975 dollars
`hage`: Husband's age
`heduc`: Husband's educational attainment, in years
`hwage`: Husband's wage, in 1975 dollars
`unemployment`: Unemployment rate in county of residence

Suppose that you run the following short regression

$$\texttt{lhwage} = \alpha^S + \beta^S \texttt{heduc} + e^S \tag{1}$$

where $\texttt{lhwage} = \log(\texttt{hwage})$. The OLS estimand of β^S is given by

$$\beta^S = \frac{Cov(\texttt{lhwage}, \texttt{heduc})}{Var(\texttt{heduc})}$$

If instead you run the long regression

$$\texttt{lhwage} = \alpha^L + \beta^L \texttt{heduc} + \gamma^L \texttt{hage} + e^L \tag{2}$$

then the OLS estimands of β^L and γ^L are given by

$$\beta^L = \frac{Cov(\texttt{lhwage}, \widetilde{\texttt{heduc}})}{Var(\widetilde{\texttt{heduc}})}$$

where $\texttt{heduc} = \pi_0 + \pi_1 \texttt{hage} + \widetilde{\texttt{heduc}}$, and

$$\gamma^L = \frac{Cov(\texttt{lhwage}, \widetilde{\texttt{hage}})}{Var(\widetilde{\texttt{hage}})}$$

where $\texttt{hage} = \delta_0 + \delta_1 \texttt{heduc} + \widetilde{\texttt{hage}}$.

(a) Run regression (1). What is the estimate you obtain for β^S? Interpret it. Construct a 99% confidence interval for β^S and interpret it.

(b) Run regression (2). What are the estimates you obtain for β^L and γ^L?

(c) Instead of running regression (2) to estimate β^L, proceed (with caution) as follows:

- First, run a regression of `heduc` on `hage` and construct the residual from this regression, call it `heducresid`, that is construct the following

$$\texttt{heducresid} = \texttt{heduc} - \widehat{\pi}_0 - \widehat{\pi}_1 \texttt{hage}$$

- Second, run a regression of `lhwage` on `heducresid`. What do you get? Compare your estimate with that of β^L in (b).

(d) Instead of running regression (2) to estimate γ^L, proceed (with caution) as follows:

- First, run a regression of `hage` on `heduc` and construct the residual from this regression, call it `hageresid`, that is construct the following

$$\texttt{hageresid} = \texttt{hage} - \widehat{\delta}_0 - \widehat{\delta}_1 \texttt{heduc}$$

- Second, run a regression of `lhwage` on `hageresid`. What do you get? Compare your estimate with that of γ^L in (b).

(e) What is the intuition behind your findings in (b), (c) and (d)?

4. Suppose that you want to estimate the parameters α and β of the following production function:

$$Y_i = AL_i^{\alpha} K_i^{\beta} \exp(u_i)$$

using a sample of n countries, $i = 1, ..., n$, where A is an unobserved measure of total factor productivity (constant across countries), Y_i is the annual total (real) GDP in country i, L_i is the total number of workers in country i, K_i is the number of machines in country i, and u_i captures random productivity differences across countries.

(a) How would you estimate α and β using OLS? Explain.

(b) How would you test whether the production function exhibits constant returns to scale? Explain.

(c) The file `production2015` (in `.dta`, `.RData`, or `.xls`) contains information on output (`Y2015`), capital (`K2015`) and labour (`L2015`) for 39 countries in 2015. Estimate α and β using these data and test whether the production function exhibits constant returns to scale.

(d) Suppose that A is not constant but varies across countries. Moreover, suppose that A_i is correlated with both L_i and K_i. Is this a concern for the OLS estimates of α and β? Explain.

(e) **[Extra.]** If you happened to have information on Y, L and K, for two years, say 2014 and 2015, for all these countries, how would you estimate this model using OLS? Explain.

5. **[Extra.]** Suppose that you sample observations on test scores, Y_{is}, where $i = 1, ..., n_s$ indexes students from each of $s = 1, ..., S$ schools. Assume that in the population of interest students were *randomly* allocated to schools, and your sample is representative of the population. Consider the following regression:

$$Y_{is} = \alpha + \beta \overline{Y_s} + e_{is}$$

where $\overline{Y_s} = \frac{1}{n_s} \sum_{i \in S} Y_{is}$ is the average test score in school s.

(a) Suppose that you actually observe the whole population. Write down the OLS estimand for β.

(b) Does it make sense to estimate "peer effects" on academic achievement in this way? Why? [Hint: What is β?]