

Practice Set 1.2: Simulation and the Regression Model

1. Helping children of age three to acquire cognitive skills is likely to have large and persistent returns. Suppose that a research institution sets up a social experiment consisting in giving three extra hours of schooling to children. The research institution randomizes the admission to this program: from all applicants, they only accept 30% randomly chosen children. Consider the following data generation process for the kids' educational achievements:

$$grades = 6 + 0.07 * program + 0.25 * family + u$$

where

$$u \sim uniform(0, 2)$$

$$family \sim uniform(0, 1)$$

$$program = \begin{cases} 1 & \text{with probability 0.3} \\ 0 & \text{with probability 0.7} \end{cases}$$

Set the seed to 457 for replicability. Set population size as 5000 observations. Set the number of replications as 500. For each replication:

- (a) Generate, in this order, u , $family$, $program$, $grades$.
- (b) For samples of 100 and 5000 observations, estimate $cov(program, family)$ and $cov(program, u)$.
- (c) For samples of 100 and 5000 observations, obtain the ols estimates from model

$$grade = \gamma_0 + \gamma_1 program + u$$

Describe the distribution of $\hat{\gamma}_1$ for each sample size. Comment the results in view of the result obtained in (b).

2. In this exercise we are going to simulate the residuals obtained from OLS estimation in a simple regression model. Suppose that we doubt whether the t -ratio from the coefficient follows a t or normal distribution for a small sample. We simulate the model by resampling the residuals from the initial OLS and re-estimate the model. The technique of simulating from actual samples and estimates is usually referred to as “bootstrapping” (for a good discussion of simulation-based tests and bootstrapping, see Davidson and MacKinnon 2004, chapter 4). In this exercise, we are going to carry out the so-called “parametric bootstrap”: we simulate the estimated residuals and add to them the fitted model under the null. In `gretl`, the function `resample()` conducts resampling, with replacement, of a series: given an original data series `x`, the command `genr xr = resample(x)` creates a new series each of whose elements is drawn at random from the elements of `x`. To carry out the bootstrapping, we are going to use the file `foodchild.gdt`, which contains data for a sample of households in a developing country. Economists are interested in the relationship between the household's total food expenditure and the number of children in the household, so that we run a regression of food expenditure (measured in dollars) on the number of children computing robust standard errors:

$$food_exp_i = \beta_0 + \beta_1 n_child_i + u_i$$

- (a) Using robust standard errors, report the t -ratio of the slope and its p -value under the null that $\beta_1 = 0$. Comment the results.

- (b) Using the OLS results, simulate 500 times the residuals of the model. Using these residuals and the estimated coefficients, construct simulated values for $food_exp_i$ under the null that $\beta_1 = 0$. In each simulation, run the regression and save the estimate of the slope and the t -ratio.
- (c) Compute the standard error of the 500 simulations of the slope. Compare the result with the robust standard error obtained in part (a).
- (d) What is the proportion of cases in which the simulated t ratio is larger, in absolute value, from the t -ratio reported in part (a)? Comment the result.