Introduction to Econometrics
Session 7 – Linear Regression: Inference for Clustered Data

September 2025

1 Problem

In their 2014 paper, Apicella et al. study the origins of the endowment effect, a bias relative to rational choice theory that leads individuals to overvalue what they own compared to what they do not own. The authors assess the universality of this effect by studying a small population of hunter-gatherers living in Tanzania, who are differentially exposed to modern society due to their geographic dispersion. The goal of this exercise is to replicate part of their results.

1. Use the `read_dta` function from the `haven` package to create an object called `endowment` from the file `endowment_data.dta`.

2. Each row of the dataset represents an individual from the Hadza population, living around Lake Eyasi, with the GPS coordinates of the camp where they live (variables `gpsx` and `gpsy`). The GPS coordinates of the village center of Mangola are (-3.519, 35.33). Use the `distm` function from the `geosphere` package to compute the great-circle distance between each individual's camp and the village of Mangola.

3. Each individual in the study participated in two experiments designed to measure their sensitivity to the endowment effect: one involving food items (cookies) and another involving lighters. Create a new table called `endowment_long` in which each row corresponds to the result of one experiment (thus two rows per individual), along with their distance to Mangola, the camp they live in, their age, and their sex.

4. Compute the frequency of experiments that displayed endowement effect for each camp, and plot this frequency against the distance to Mangola, with bubbles with size related to the number of surveyed individuals (Figure 2 panel B).

5. Create a variable `exposure` taking the value `"high"` for the four camps closest to Mangola, and `"low"` for all other camps.

6. Regress the variable measuring the presence of the endowment effect on the indicator for belonging to the highly exposed group due to proximity to Mangola (Column 1 of Table 1 in the paper).

7. Does the default variance–covariance matrix provided by R apply here? Why or why not? Can we rely solely on a heteroskedasticity-robust variance–covariance matrix?

8. Use the `vcovCL` function to compute the variance–covariance matrix with clusters defined at the camp level.

9. What criticism can be made regarding this way of estimating the precision of the regression?