# Generation of Simulated Data

This document describes the generation of simulated data for the PhD thesis "Application of Small Area Estimation Methods to Labour Force Statistics in Ireland" by Jillian Delaney.

In this thesis, simulated data is generated in a series of synthetic population simulations and used for theoretical comparison in Chapters 4 and 5. These simulations are designed to compare the standard Population-Covariate (PC) model and the proposed Sample-Covariate (SC) model under both unconditional and finite population perspectives.

- **Geographic Structure:** The simulated population is divided into **50 areas**.
- **Population Sizes ($N_i$):** For each area, the population size is generated by drawing 50 observations from a $\chi^2$ distribution with a noncentrality parameter of 20,000 and 3 degrees of freedom.
- **Auxiliary Variables ($\mathbf{x}_{ij}$):** Two covariates are generated for every unit $j$ in area $i$
  - $x_1$: Simulated from a **Normal distribution** $N(2, 1)$.
  - $x_2$: Simulated from a $\chi^2$ distribution with 3 degrees of freedom.
- **Random Area Effects ($v_i$):** These are selected for each area from a normal distribution $N(0, \sigma_v^2)$, with $\sigma_v^2$ taking values of **0.2, 0.4, or 0.8** to test various levels of area-level variation.
- **Dependent Variable ($y_{ij}$):** The outcome is generated using the model: $y_{ij} = \mathbf{x}_{ij}^T\beta + v_i + \epsilon_{ij}$.
- **Simulation Configurations:** Four different configurations are used to vary the regression coefficients ($\beta$) and the nature of the residual terms ($\epsilon_{ij}$)
  - $\beta$ **Values:** Simulations 1 and 3 use "larger" coefficients $\beta^T = (3, 2, 4)$. Simulations 2 and 4 use "smaller" coefficients $\beta^T = (3, 1, 0.5)$ to reduce the variability of the synthetic estimates.
  - **Residual Terms ($\epsilon_{ij}$):** In Simulations 1 and 2, residuals are randomly selected from $N(0, \nu_i)$, where the mean variance $\bar{\nu}_i = \lambda\sigma_v^2$ and $\lambda \in \{0.1, 1, 10\}$. In Simulations 3 and 4, residuals are generated specifically to force a **negative correlation** between the sample synthetic estimates and the area residual means.
- **Sampling Procedure:** From each simulated population, a **1% sample** is selected using Simple Random Sampling (SRS). This process is repeated for **1,000 Monte Carlo iterations**.