# Notes on random noise in the RoS tables

Minority inflow numbers are very low in many data zones and this may potentially reveal details about the residents in an area after data linkage. In order to protect individuals, each cell value was increased by a random amount. The random value was drawn from a Poisson distribution with $λ=3$. Researchers who wish to use the data for statistical analyses need to bear in mind the nature of the random noise in order to obtain unbiased results.

For descriptive statistics: mean values will be biased upwards by 3, median values upwards by approximately 1.27, and variances will be inflated by 3.

For regression models: the results from linear regression models will be unbiased except for the intercept which will be biased upwards by 3. Since the data is in the form of counts (with low count values for minority groups), a Poisson regression model (or log-linear model) is probably more appropriate for multivariate analyses. Researchers can obtain unbiased results from these models by modifying the log-likelihood function.

In a Poisson regression model the expected value of the outcome variable $Y$ given the predictors $X$ is:

$$E(Y|X)=e^{\sum\_{}^{}β\_{j}x\_{j}}$$

Where $x\_{j}$ is a predictor and $β\_{j}$ is a member of $β$ which denotes the regression parameters that we wish to find. The probability of observing any given value of $Y$ is given by the probability mass function (pmf) of the Poisson distribution with $λ= e^{\sum\_{}^{}β\_{j}x\_{j}}$. This pmf is then used in the log-likelihood function to obtain maximum likelihood estimates of $β$.

Taking $Y$ as a cell value in the RoS, since random noise has been added to the data, the probability of observing $Y$ is still given by the Poisson distribution but with $λ= e^{\sum\_{}^{}β\_{j}x\_{j}}+3$. This is version of the pmf is to be used in the log-likelihood function. Since a custom log-likelihood function is being used, a researcher would need to implement the regression model themselves. An example of such a routine is given in R.

An approachable guide to maximum likelihood estimation is given in Gelman and Hill (2007; Chapter 18). Researchers with experience of using maximum likelihood estimation (or MCMC) can use similar approaches to correct biases arising from the addition of random noise for other types of statistical models.

**References**

Gelman, A., & Jennifer, H. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press. https://doi.org/https://doi.org/10.1017/CBO9780511790942