**ST417 Bayesian Modelling 2022**

***The Commuting Crisis:***

***Relationship between Time, Distance and Cost to Commute for University of Galway Students***

***Fionn McGlacken - Hugo Mead - Alex Horkan***

We are in an accommodation crisis. Third-level students without access to accommodation near their universities commute longer hours, over greater distances, at higher costs. In this project, we will attempt to investigate the **time**, **distance**, and **cost** of **travel** for University of Galway Students. We are also interested in the effects of potentially confounding variables, such as weather, day, time, traffic, age, city or town, and mode of transport.

In order to explore the effects of the different variables associated with commuting, we have set a list of aims we wish to explore throughout the project:

1. Investigate the relationship between time, distance, and commuting costs.
2. Investigate the effects of confounding variables, such as weather, day, time, traffic, age, and mode of transport.
3. Calculate point and interval estimates of time, distance, and cost to travel.
4. Investigate if the average time, distance, and cost of travel for students have increased over previous years.
5. Investigate the most expensive city or town to travel from, on average.
6. Investigate prevailing weather, the most common mode of transport, and traffic patterns.
7. Estimate the carbon footprint of travel for University of Galway students.
8. Compare differences between year groups and see which year has the greatest travel times.

The data we used for the purposes of this analysis was collected via an online questionnaire which was distributed amongst University of Galway students. The questionnaire primarily consisted of questions pertaining to travel time, travel distance and travel cost. Additionally, respondents were asked about the mode of transport, the origin of the commute as well as the year of study. We received a total of 74 responses comprised mostly of 3rd and 4th-year students, with the remaining being made up of 1st and 2nd-year as well as postgraduate students.

Before we began analysis, the raw data .csv file was read into R where we applied some custom functions to the data in order to make it interpretable within the R environment.

The first step of our bayesian analysis was to generate prior and posterior distributions for commute time, commute distance and commute cost. The priors were based on data collected from respondents who were asked to estimate the total time, distance and cost of commuting for the average University of Galway student. The priors generated from this data were updated with the actual time, distance and cost of commute of each respondent in order to generate posterior distributions.

From these updated posterior distributions, we were able to simulate new values in order to form point estimates as well as 95% credible intervals.

The 95% credible intervals suggest that:

* The average time of a commute should be between 19.6 and 30.6 minutes, with the point estimate suggesting a commute time of 25.19 minutes
* The average distance of a commute should be between 5.3 and 13.14 kilometres, with the point estimate suggesting a commute length of 9.2 kilometres.
* The average cost of a commute should be between €0.91 and €2.50, with the point estimate suggesting a cost of €1.71 for a student's commute to the University of Galway.

## RJAGS MCMC

Using RJAGS we defined a Markov Chain Monte Carlo (MCMC) model in order to generate a set of samples from the posterior distributions. After defining the values of all variables in the model definition, we compiled the model and the data, including the starting values as well as defining the number of chains at 3.

After a burn-in period of 1,000 iterations, we simulated 10,000 new values from the posterior data. From these simulations, we could plot trace and density plots of each feature.

The trace plots all showed good stability and mixing throughout the iterations, as well as not showing any clear patterns throughout, this suggests that the model has reached convergence. The density plots appear to be smooth distributions, suggesting that the simulations were representative of all possible values in the posterior distributions.

The MCMC model produced similar values to the 95% credible intervals and point estimates of the original posterior distribution, suggesting that the data we based our analysis on was probably a good predictor of reality.

A Gelman-Rubin statistic, where the estimated variance of the parameter is a weighted sum of the between and within chain variances, of R≅1 suggests that the chains have converged.

The parameters all have values of R=1 and the plots show convergence around 1 so this suggests that the 3 Markov chains in the model have converged.

Autocorrelation plots showed very minimal autocorrelation across all features in the model.

The 95% credible interval plots show that the average student takes between 19.5 and 30.8 minutes to travel between 5.2 and 13.2 kilometres at a cost of between €0.89 and €2.50. This is very similar to the predictions made from the original posterior distribution.