# Department of Statistics
# STATS 330: Statistical Modelling
# Assignment 3
# Summer School, 2022

Total: 100 marks Due: 14:00, 3 February 2022

## Notes:

(i) Write your assignment using R Markdown. Knit your report to either a Word or PDF document.

(ii) Create a section for each question. Include all relevant code and output in the final document.

(iii) Marks may be deducted for poor style. Please keep your code and plots neat.

(iv) These assignments do require you to write R code, but we appreciate this course is not specifically about programming. If you are struggling with the programming aspects of this assignment, please ask for help at our labs or office hours. If you can describe specifically what you want your code to do, then we can point you in the right direction.

(v) Please remember to upload your R Markdown file before the deadline, too. If the markers identify an error in your work, being able to run the code you have written can help determine what you did wrong.

## Introduction

The occurrence of prime numbers in a sequence of consecutive numbers is seemingly random, despite having a precise definition and deterministic ways to search for them. We are interested in modelling the number of primes inside a random sequence of consecutive numbers and the proportion of primes amongst a fixed number of consecutive numbers.

## Question 1

The data set `prime.poisson.txt` contains the following variables:

- `y` is the number of primes in the sequence of consecutive numbers $\{x_1, \ldots, x_n\}$.
- `size` is the total number of the elements in the sequence.
- `center` is the median of the elements in the sequence.
- `variance` is the variance of the elements in the sequence.
- `even` whether the sequence contains an even number of elements.
- `cramer` whether the difference $x_n - x_1$ is less than or equal to $(\ln x_1)^2$.
- `ratio` is the ratio $x_n/x_1$.
- `PNT` is the difference $x_n/\ln(x_n + |\epsilon|) - x_1/\ln(x_1 + |\epsilon|)$ where $\epsilon \sim N(0, 1)$

which are generated using `prime.poisson.data.generator.R`.

(a) Write down equations that correspond to the following model.

[2 marks]

```
poisson.fit = glm(y ~ center + ratio, family = "poisson",
                  offset = log(size), data = prime.poisson.df)
```

(b) Fit the model as it is given in part (a), then run the following three hypothesis tests. Write down the null hypothesis for each test. State the conclusion for each test based on the corresponding p-value.

[7 marks]

```
# Hypothesis test 1 & 2 -----
anova(poisson.fit, test = "Chisq") # two p-values
# Hypothesis test 3 -----
1-pchisq(poisson.fit$deviance, poisson.fit$df.residual) # one p-value
```

(c) Use the package `mgcv` to fit a generalized additive model, then use the corresponding GAM plots to investigate whether we need to transform `center` or `ratio` for the model in part (a). Comment the GAM plots.

[6 marks]

(d) Run the following lines of R code to generate two boxplots and a histogram. Comment on what the three plots collectively show us in terms of the skewness of `ratio`.

[2 marks]

```r
boxplot(prime.poisson.df$ratio, ylab = "ratio")

boxplot(prime.poisson.df$ratio, ylab = "ratio", outline=FALSE)

# New variable dl.ratio
prime.poisson.df$dl.ratio = log(log(prime.poisson.df$ratio))

hist(prime.poisson.df$dl.ratio, xlab = "ratio")
```

(e) Modify the generalized additive model in part (c) to include `dl.ratio` instead of `ratio`. Use the corresponding GAM plots to investigate whether we need to transform `center` or `dl.ratio` further. Comment the GAM plots.

[3 marks]

(f) Use the package `VGAM` to fit a generalized additive model, then use its `summary` output to decide whether we need to transform `center` or `dl.ratio`. **Hint:** detach `mgcv`.

(g) Based on the GAM plots of the model in part (f), propose and fit a new generalized linear model using `vglm` to capture the nonlinear effect. Produce a plot to compare and comment how close this new GLM is to the GAM in part (f).

[5 marks]

(h) Use the deviance goodness of fit (GOF) test and a residual plot to discuss whether the model you created in part (g) is adequate.

[6 marks]

(i) Use the parametric bootstrap to estimate the sampling distribution of the deviance of your model in part (g). Plot an appropriate histogram of the bootstrap deviance values and overlay that with the asymptotic $\chi^2_{df}$ approximation density curve. Use this plot to comment on the adequacy of the GOF test in part (h).

[7 marks]

(j) Find the parametric bootstrap GOF test p-value of the model.

[4 marks]

(k) Based on your model in part (g), compute the 95% inverted parametric bootstrap confidence interval for the intercept of your model in part (g).

[6 marks]

(l) Based on your model in part (g), compute the 95% inverted parametric bootstrap confidence interval for the mean number of primes in the sequence of consecutive numbers {100, 101, . . . ,500}, for which the corresponding explanatory variables are given by `new.poisson.df.txt`.

[6 marks]

(m) Based on your model in part (g), compute the 95% parametric bootstrap prediction interval for the number of primes in the sequence of consecutive numbers {100, 101, . . . ,500}, for which the corresponding explanatory variables are given by `new.poisson.df.txt`.

[6 marks]

## Question 2

The data set `prime.logistic.txt` contains the following variables:

- `y.binomial` is the number of primes in the sequence of consecutive numbers $\{x_1, \ldots, x_n\}$.
- `n.binomial` is the total number of elements $n$ in the sequence.
- `center` is the median of the elements in the sequence
- `cramer` whether the difference $x_n - x_1$ is less than or equal to $(\ln x_1)^2$.
- `ratio` is the ratio $x_n/x_1$.
- `PNT` is the difference $x_n/\ln(x_n + |\epsilon|) - x_1/\ln(x_1 + |\epsilon|)$ where $\epsilon \sim N(0, 1)$

which are generated using `prime.logistic.data.generator.R`.

(a) Use the package `gam` to fit a generalized additive model, then use its `summary` output to decide whether a nonlinear function of `center` might improve the following model.

[3 marks]

```
quasibinomial.fit=glm(cbind(y.binomial,n.binomial-y.binomial)~cramer*center,
  family=quasibinomial, data=prime.logistic.df)
```

(b) Construct a plot displaying the estimated relationship between the explanatory variables and the response variable for your GAM as well as the GLM in part (a). Comment the plot.

[5 marks]

(c) Use the non-parametric bootstrap to estimate the sampling distribution of the intercept of your GAM model in part (a). Plot a histogram of the bootstrap intercepts and put intercept of the GAM model on the same plot. Discuss whether we could use parametric bootstrap instead of non-parametric bootstrap in this case.

[6 marks]

(d) Based on your GAM model in part (a), compute 95% inverted non-parametric bootstrap confidence interval for the proportion of primes in the sequence of consecutive numbers {10, 101, ...,100}, for which the corresponding explanatory variables are given by `new.logistic.df.txt`.

[6 marks]

(e) There is no easy way to compute prediction intervals for a quasi-binomial model. Let us pretend your GAM model in part (a) as a binomial model rather than quasi-binomial, and compute the 95% non-parametric bootstrap prediction interval for the number of primes in the sequence of consecutive numbers {10, 11, ...,100}, for which the corresponding explanatory variables are given by `new.logistic.df.txt`.

[6 marks]

(f) Discuss why the prediction interval in part (e) is likely to be wider than the true prediction interval.

[4 marks]