



*Due date: Week 13 Friday 5 Nov, 2021, 23:59 pm*

### Instructions

- For all the questions please provide the relevant mathematical derivations, the computer output (only using R software) and the plots.
- Please submit on iLearn a single PDF file containing all your work (code, computations, plots, etc.). Other file formats (e.g. Word, html) will NOT be accepted.
- We strongly recommend you to use `Rmarkdown` to typeset your solution. This is also the only way to achieve full mark in this assignment.
- Don't include any more R output than necessary and include only concise explanations.
- **There is a limit of 22 pages for your submission. There will be a 10% penalty if your submission is 23 pages long and another 10% penalty if your submission is 24 pages long. After that, there will be 1% penalty for each CHARACTER (space included) from page 25 onwards.**
- A late penalty applies if you submit your assignment late without a successful special consideration. See the unit guide for more details.
- *You should submit the assignment using the Turnitin tool on iLearn.*

**Rubric** This assignment is worth 30% of the unit marks. This is an assessment task that will test both, your statistical knowledge and technical skills.

- Question 1 [19 marks]. Tests your applied statistics skills
- Question 2 [41 marks]. Tests your applied statistics skills
- Question 3 [15 marks]. Tests your applied statistics skills
- Question 4 [10 marks]. Tests your RMarkdown/Mathematical Typesetting skills

Marking Guide/Rubric for Question 4:

- 6 marks are awarded for making your assignment submission reproducible.
  - Only 3 marks if the assignment file is compiled, eg. from RMD to HTML/Word before converting to PDF.
  - The full 6 marks if the assignment file is compiled from RMD to PDF ( $\text{\LaTeX}$ ) directly.
- The final 4 marks are awarded for typesetting all the relevant mathematical equation in  $\text{\LaTeX}$ .
  - You will receive 0 for this part if you hand-written it & take pictures of them or typesetting them in Word or any other format.

**Question 1 (19 marks)**

van den Broek (1995) describes a study of 98 HIV-infected men attending the Utrecht University Hospital. The research question of interest is whether the number of urinary tract infections (UTIs) experienced in the period of observation, was associated with the patient's immune status. Immune status is determined by measuring the CD4+ cell count, with high CD4+ indicating high or good immune status, and a low CD4+ count indicating a compromised immune system and progression of HIV. CD4+ is reported as the number of cells per cubic millimetre of blood, with a normal CD4+ count being in the range 500 to 1500. The dataset is available on iLearn as `uti.csv`.

Variable name	Description
episode	number of UTIs in the period of observation
months	number of months of observation
<code>sqrtd4</code>	square-rooted transformed of CD4+ count

You should:

- [2 marks] explain why we need to adjust for the number of months of observation when we model the number of UTIs. Also explain how we can carry out the adjustment in a count regression setting.
- [7 marks] investigate the data graphically (including some visualisations to check whether zero-inflated models should be considered).
- [3 marks] start by fitting a Poisson model with `sqrtd4` as your covariate and then try to find the most appropriate distribution for this dataset by consider alternative models. Use an appropriate model selection criterion to assist you.
  - Make sure you add in the adjustment from part a). The help menu can be of assistant if you need help.
  - You can drop/add `sqrtd4` as your covariate as part of your model building process.
  - You can assume any Negative Binomial model to have constant shape.
- [3 marks] write down the fitted model equation for your final model.
- [2 marks] obtain the quantile residuals of your final model and check its goodness of fit.
- [2 marks] interpret your model parameters.
- [2 marks] summarise your finding in a sentence or two.

**Question 2 (41 marks)**

We consider data on Body Mass Index (BMI) collected on individuals aged 49 and over in the Blue Mountains Eye Study (BMES).

The first BMES sample was conducted in 1992. We have  $n = 3499$  subjects in the dataset `bmes1`, with the following variables:

Variable	Description
age	age in years
sex	1 = female; 2 = male
race	1 = White; 2 = Aboriginal; 3 = Negroid; 4 = Hispanic; 5 = Oceanian; 6 = Asian; 7 = Indian;
bmi	body mass index ( $\text{kg}/\text{m}^2$ )

- [2 marks] Investigate the variable `race` and provide a reason why this variable *is not* going to be that helpful in predicting the subject's BMI, especially for the minority groups?

- b) [26 marks] How does BMI change with age in this population of older people, and is this the same for both sexes? Investigate this question, using a model with normal response distribution. You should:
- check the normality of the response variable and transform it if necessary;
  - you only need to consider **age** and **sex** as covariates in this part;
  - investigate the data graphically (or use a table if appropriate)
  - if the age–BMI relationship appears to be non-linear, investigate using a generalized additive model (GAM) and then refit the model using polynomial regression;
  - there is no need to consider interaction effect here;
  - write out the fitted model for your final model;
  - obtain a residual vs fitted values plot and comment; (normally you will need to do many more model diagnostics but we assessed that previously and so we are skipping those steps here);
  - provide a conclusion of your final model and suggest a reason for your conclusion.
- c) [13 marks] The World Health Organisation classifies BMI as

BMI	Classification
< 18.5	underweight
[18.5, 25)	normla weight
[25, 30)	overwieght
≥ 30	obese

Develop an appropriate statistical model for the BMI classification using **sex** and **age** as covariates. You should

- create the classification and show the first few entries of the dataset (for the marker(s) to check this has done properly) and you may find functions such as `dplyr::case_when()`, `facnycut::fancycut()` or `forcats::fct_recode()` useful here.
- write down the model equation(s);
- justify whether the model assumption(s) is satisfied;
- comment on whether your conclusion here is consistent with those you obtained in part (b).

### Question 3 (15 marks)

We are analysing the results of a study to access whether barn owl nestlings (i.e. baby owls) beg more intensely in the presence of their mother than in the presence of their father. 27 broods/group, containing on average 4 (range 2–7) nestlings, were used to collect a total of 599 samples during the study period. The dataset is called `Owl` and the variables are given below:

Variable	Description
SiblingNegotiation	a numeric vector giving the number of calls from a group of nestlings
SexParent	a factor describing the sex of parent: Female or Male that brought the food
FoodTreatment	a factor describing food treatment: Deprived or Satiated (i.e. Satisfy)
BroodSize	brood/group size

```
Owl <- read_csv("Owl.csv") %>%
  mutate(Nest = factor(Nest))
```

After conducting a cross-sectional study, it was determined the model with the following covariates is appropriate

```
model_cs <- glm.nb(SiblingNegotiation ~ FoodTreatment + SexParent +  
  offset(log(BroodSize)), data= Owl)
```

- a) [1 marks] Explain why we should consider GLMM or GEE models for this dataset.
- b) [9 marks] In this part, we will consider a random intercept model.
  - i. Fit the “equivalent” random intercept model;
  - ii. Write out the fitted model equation for the random intercept model. Remember to not only to state the fixed effect, you also need to state the distribution assumption and the random intercept component;
  - iii. Interpret the coefficient for **SexParent**. You will need to provide us the numerical interpretation whether the variable is significant or not.
  - iv. Do we actually need a random intercept model here? Use an appropriate model selection criterion to justify your answer.
- c) [5 marks] In this part, we will consider a marginal GEE model.
  - i. Provide a reason, without fitting any models, why an **exchangeable** correlation structure is more appropriate than an **AR(1)** structure here.
  - ii. Fit the “equivalent” GEE model with an **exchangeable** correlation structure;
  - iii. Interpret of the coefficient for **FoodTreatment**. You will need to provide us the numerical interpretation whether the variable is significant or not.

## References

van den Broek, Jan. 1995. “A Score Test for Zero Inflation in a Poisson Distribution.” *Biometrics* 51 (2): 738–43. <https://doi.org/10.2307/2532959>.