STAT3510 Winter 2020
Assignment 3

Due: Friday, April 3, 11:59pm to Crowdmark

NOTE: All questions must be completed and submitted.

---

1. A study on the number of successful matings and ages (recorded at the beginning of the study) of 41 male elephants is available in the file `Elephant_Data.csv` on Courselink[1]. All elephants included in the study were observed for the same 8-year period, and the purpose of this study was to determine the effect of age on mating success.

   (a) Fit a Poisson regression model to the data set. Comment on the significance, magnitude, and direction of the effect of age on the *mean* number of successful matings. Include relevant R output in your final solutions.

   (b) Calculate and interpret a 95% confidence interval for β_1. Include both the interval and your interpretation in your final solutions.

   (c) Conduct a Goodness of Fit test for this model. Be sure to include the null and alternate hypothesis, test statistic, p-value, and conclusion.

   (d) Fit a linear regression model to the data set, where the number of matings is regressed on age. Include the relevant R output in your final solutions. How does the interpretation of $\hat{\beta}_1$ in the linear model compare to the interpretation of $\hat{\beta}_1$ in the Poisson model?

   (e) Create residual plots from the linear model (normal qq plot and a plot of residuals versus fitted values). Comment on your conclusions from these plots. Include both the plots and your comments in your final solutions.

   (f) Create a scatter plot of number of matings versus age. Superimpose both your linear model and Poisson model on the plot. Comment on which model visually seems to fit the data better.

   (g) Using the model that (visually) seems to fit the data better, what is the predicted mean number of mating pairs for elephants that are 37 years old?

2. Rates of risky driving among novice adolescents and at least one of their parents over the first 18 months of the adolescents licensure were investigated. Parents drove the same vehicles over the same period and geographic area as their newly licensed adolescents. Data-recording systems installed in participants vehicles provided information on driving performance (including crashes and near-crashes) of 42 newly licensed adolescent drivers and their parents (55 parents total). The response variable of interest was number of crashes or near-crashes (`ncrash`). The explanatory variables recorded were whether or not the driver at the time was an adolescent (`adol`, 1 if adolescent, 0 otherwise), whether or not the driver at the time had taken a driver's education course (`DE`, 1 if yes, 0 if no), and the mileage driven (`expos`, measured in 1000 miles).[2]

---

[1] Adapted from Case Study 22.1.1 in *The Statistical Sleuth*, 3^{rd} ed. (2013). Ramsey F.L. and Schafer D.W. Brooks/Cole, Canada

[2] Data simulated based off of the article Bruce G. Simons-Morton, Marie Claude Ouimet, Zhiwei Zhang, Sheila E. Klauer, Suzanne E. Lee, Jing Wang, Paul S. Albert, and Thomas A. Dingus. Crash and Risky Driving Involvement Among Novice Adolescent Drivers and Their Parents. American Journal of Public Health: December 2011, Vol. 101, No. 12, pp. 2362-2367. (Am J Public Health. 2011;101:23622367)

The data is available in the file `Driver_data.csv` available on Courselink. Use this data for the following analysis:

(a) Fit a Poisson regression model to the data set, including mileage driven as an offset term. Include the relevant R output in your solutions. (Hint: remember that the offset term is included as the natural log of the variable being used as the offset).

(b) Provide a brief explanation as to why mileage driven was included as an offset (that is, explain why it is reasonable to need to adjust the number of crashes/near-crashes by miles driven).

(c) Provide an interpretation of the estimated intercept, $\hat{\beta}_0$, on the both the log(mean) and mean (i.e. original) scale. (Hint: Remember that the mileage was measured in units of 1000 miles)

(d) Provide an interpretation of the coefficient for driver's education, on both the log(mean) and mean scale. Be sure to comment on significance, magnitude, and direction.

(e) Create a plot of deviance residuals versus linear predictor, as well as a half normal plot of the deviance residuals. Include the plots in your report, and comment on whether or not there appears to be any significant violation of the model assumptions.

- To get the deviance residuals, use the `residuals` function and specify "deviance" for the `type` option. To get the linear predictor (which is essentially the predicted *function* on the left-side of the GLM equation), use the `predict` function and specify "link" for the `type` option. The half normal plot is available in the `faraway` library, under the function name `halfnorm`.

(f) Regardless of significance of the explanatory variables, what would be the estimated mean number of crashes/near-crashes (per 1000 miles) for an adolescent who did not take driver's education?

(g) If an adolescent who did not take driver's education drove 7000 miles, how many crashes/near-crashes would we expect to occur?

3. This question comes from Dr. Prosser's notes, with adaptations by me:

(a) Construct an acute SSD for chloride in R using the CCME data set for **either** the vertebrates in the data set (fish and amphibians) or the invertebrates. The data is available in Table 3 of the Canadian Water Quality Guidelines document on Courselink.

- When fitting a distribution to the SSD, use a log-normal distribution. Obtain the summary of this fitted model, and include this output in your final report.
- What is the estimated HC5 and 95% parametric bootstrapped confidence interval?
- According to the Canadian Water Quality Guidelines document, what was the HC5 across *all* species (vertebrates and invertebrates combined)? What species fell below the accepted HC5 value?
- Using your model, adapt the commands given by Dr. Prosser to create a Species Sensitivity Distribution plot, with the fitted model and bootstrapped confidence intervals included. Include this plot in your final report.

(b) Construct an EED for chloride in R using PWQMN data. The data is available in the file `ONCHLOR.csv` on Courselink.

- Note that the concentrations in the spreadsheet are already listed in ascending order.
- Use the log-normal distribution to fit a model to this data. Include the relevant R summaries in your final report.

- Based on your fitted model, what is the 95^{th}-centile and 95% parametric bootstrapped confidence interval?
- Adapt the commands given by Dr. Prosser to create an Environmental Exposure Distribution plot, with the fitted model and bootstrapped confidence intervals included. Include this plot in your final report.

(c) How does your SSD HC5 compare to your EED 95^{th}-centile?

- Create a plot with both the species sensitivity and environmental exposure data included. Include this plot in your final report.
- This plot should have concentration on the x-axis. The x-axis should be on the log_{10} scale, and the proportion of species affected/proportion of sites with each concentration should be on the y-axis.
- Include points for your HC5 and 95^{th}-centile values on the plot.
- Comment on the overlap (or lack thereof) of the HC5 and 95^{th}-centile values. Provide an appropriate interpretation of what this overlap (or lack thereof) suggests.

Formatting Details and Submission Instructions

Please read the following formatting information and submission instructions carefully. Projects that do not follow these formatting requirements, and/or that are submitted incorrectly, may not receive full marks.

FORMATTING DETAILS

- You do not need a cover page for your assignment, but instead you must include your name and student ID# as a header on every page.
- Your assignment will already be separated into separate files for each question, but please clearly label parts of each question (i.e. (a), (b), etc.).
- As much as possible, please write up your answers in software (Microsoft Word, LaTeX, other). If you are doing parts by hand, you can scan or take pictures of the work and embed the images in your Word document. If you are doing an entire question by hand, you can scan (preferred over pictures) your work and upload it.
- R output should be clearly labelled, and scaled so as not to take up more space than required. Please review the overall appearance of your work before it is submitted.
- There is no restriction on page limits for answering questions, and no specifications for style, format, etc., however work that is unnecessarily long (or too short!), has a poor overall appearance, is difficult to read or follow, or generally does not give the impression of an overall good effort will not receive full marks.

SUBMISSION INSTRUCTIONS

- You will receive an email from Crowdmark in the next day or so. **Do not delete this email, or forward it to anyone else**. This is your personalized link for submitting your Assignment #3. If you do not receive it, first check your junk mail folder. If you are sure it is not there, contact me as soon as possible.

- If you are completing the assignment with a partner, **SET YOUR GROUP UP FIRST**. To do this:
  - One partner can access their link on Crowdmark, and click on *Add Group Member*.
  - Search for the classmate you want to add to your group. Follow the instructions to complete your group registration.
  - **WARNING**: When you submit your report, if you have not created a group Crowdmark will ask you to confirm that you are not part of a group. If you make this confirmation, only the submitting partner will receive the grade. Make sure you identify your partner to avoid them receiving a grade of 0!
  - If you are completing the project as an individual, you do not need to create a group. You only need to upload and submit your report when ready.
- When you are ready to submit your assignment, you can click on the link provided in the email. This will open up a window where you can upload your files.
- You should see a space to upload your files for each question separately. YOU MUST HAVE A SEPARATE FILE FOR EACH QUESTION.
- Questions that have been answered in word processing software must be saved and uploaded as PDFs (or JPGs, for pictures). No other file formats are accepted by the system. A caution of submitting a photo of your work: if we cannot read your work, you will not receive full marks. It is YOUR responsibility to ensure your files are legible.
- Upload your files for each question in the appropriate space. **It is your responsibility to ensure the correct files have been uploaded to the correct space**. Answers that have been uploaded to the wrong location will receive a mark of 0.
- Review your assignment to ensure all pages have uploaded correctly. Once you are satisfied, you can submit your assignment.
- You can change and re-upload your assignment files up to the project deadline. After the project deadline, files already uploaded will be "locked in", and you will not be able to change them. Any new projects uploaded after the deadline will be flagged for a 100% late penalty.
- Help for uploading assignments can be found at:
  `https://crowdmark.desk.com/customer/portal/articles/1639407-completing-and-submitting-an-assignment`

Please review the University of Guelph's policies on Academic Misconduct, as mentioned in the course outline and detailed in the University of Guelph Undergraduate Calendar. It is your responsibility to know what constitutes academic misconduct. Students found in violation of any of the University policies on academic integrity will be charged with academic misconduct, and penalized accordingly.