STAT3510 Winter 2020
Assignment 2

Due: Friday, February 28, 11:59pm to Crowdmark

NOTE: Only Questions 1 - 4 MUST be completed and submitted. Questions 5 - 7 are for practice/review only, and should NOT be submitted.

---

1. Download the data set `IQBRAIN.CSV` from Courselink, and import it into **R**. (You should notice that this data set has multiple variables, however for now we are only interested in two of them). Researchers in the neuroscience department are interested in a model that explains the relationship between the volume of a human brain (denoted VOL, and measured in cm^3) and its area (denoted AREA, and measured in cm^2). In particular, they would like to explain the volume of a brain for a based on the area of the brain.

   (a) Create a scatter plot of the data, and comment on the appearance of a potential linear relationship between the variables.

   (b) Conduct a simple linear regression. Include the R output as a figure in your report. Use the output to determine if there is evidence of a linear relationship between volume and area of brain. You should include the appropriate null and alternate hypotheses under investigation in your report, and use specific values from the output to justify your response.

   (c) Create the appropriate residual plots to investigate if the assumptions for linear regression have been violated. Include the plots and your commentary in your reports.

   (d) Assuming the relationship between volume and area is significant, and the residual plots indicate that the model assumptions are satisfied (we will continue with this question even if this is *not* the case), what would be the predicted volume for a brain that is $1900cm^2$ in area?

2. The data are measurements on breeding pairs of land-bird species collected from 16 islands around Britain over the course of several decades[1]. For each species, the data set contains the average time to extinction (`time`), the average number of nesting pairs (`pairs`), the size of the species (`size`, large (L) or small (S)), and the migratory status (`migration`, migrant (M) or resident (R)). The data is available in the file `Bird_Extinction_Data.csv` on Courselink.

   The researchers were interested in what effect the number of nesting pairs, size, and migratory status had (if any) on the time to extinction, and if the effect nesting pairs depended on the size of the bird.

   Note: In order to improve the linear relationship between time to extinction and number of nesting pairs, a natural log transformation of the response variable is required before the analysis is carried out. References to "time to extinction" below refer to the transformed variable.

   (a) Create a scatter plot of average time to extinction versus average number of nesting pairs, using different colours or symbols to identify species of different sizes *and* of different migratory status. Include this plot in your report.

[1]Data from *The Statistical Sleuth*, 3^{rd} ed. (2013). Ramsey F.L. and Schafer D.W. Brooks/Cole Publishing. Canada. Page 306.

(b) Fit an *additive* model to the data set, that only includes the main effects of nesting pairs, size, and migratory status.

i. Include the R output in your write-up. What can be concluded about the *significance* of the effect of number of nesting pairs, size, and migratory status on time to extinction?

ii. Write out the estimated regression equation for each type of bird: (1) small and migratory, (2) small and resident, (3) large and migratory, and (4) large and resident. Use these equations to interpret each of the regression coefficients in the model.

iii. If you were to hand-sketch this regression model, what would its geometry look like?

(c) Fit a second model that incorporates an interaction effect between the number of nesting pairs and size of the bird.

i. Include the R output in your write-up. Provide an interpretation of the coefficient for the interaction term.

ii. Create a hand-sketch of this model that illustrates its geometry for the four types of birds. Include this sketch in your write-up.

iii. What can be concluded about the *significance* of the interaction effect between number of nesting pairs and the size of the bird?

iv. Use R to conduct an Extra Sum of Squares F test comparing the model from part (b) to the interaction model. Include the output in your write-up, along with the appropriate hypotheses under investigation and a conclusion.

(d) For the preferred model, create the appropriate residual plots to determine if the assumptions for linear regression have been violated. Include the plots and your commentary in your report.

3. Consider the data set on success / failure of field goal attempts as a function of yardage, found at `http://users.stat.ufl.edu/~winner/data/fieldgoal.dat` (with a description of the data set found at `http://users.stat.ufl.edu/~winner/data/fieldgoal.txt`).

(a) Fit a simple logistic regression model to the data set. Include the R output as a figure in your report.

(b) Comment if there appears to be evidence of a relationship between yardage and a successful field goal attempt, using values from the R output to justify your response.

(c) Provide an interpretation of the coefficient for yardage on both the $log(odds)$ and $odds$ scale.

(d) Based on your knowledge of NFL field goal attempts, what do you believe is the upper limit for yardage at which a Kicker will still have a successful field goal attempt? (Note: you can do some casual searching on Google if you are not familiar with the NFL!). Now, use the estimated logistic regression equation to determine the probability of a successful field goal attempt at the yardage you specified. Include your speculated yardage, your predicted probability, and a brief commentary on each value in your report.

4. Data on the stability of pillars that form the supporting structure of coal mines is found in the file `Mine_Stability.csv` on Courselink, and a brief description of the data is available at `http://users.stat.ufl.edu/~winner/data/pillar_stability.txt`. In brief, the stability of 29 coal mine pillars (stable = 1, not stable = 0) was measured along with a series of potential explanatory variables. We will investigate which explanatory variables may be important for predicting coal mine pillar stability in this question.

(a) Fit an additive multiple logistic regression equation to the data set, using the variables `Depth`, `Width`, `Height`, and `Uniaxial Compression Strength` as the explanatory variables. Include the *relevant* R output in your solutions as a figure.

(b) Based on the results of your logistic regression analysis, comment on the statistical significance of each of the explanatory variables on the stability of a coal mine pillar. Use specific values from your output to justify your conclusions.

(c) Fit a second additive multiple logistic regression equation to the data set, this time only including the explanatory variables `Width` and `Height`. Include the relevant R output in your solutions as a figure.

(d) Based on this reduced model, provide an interpretation of the estimated regression coefficients for each of the explanatory variables on the *odds* scale.

(e) Perform a Drop in Deviance test to compare the first and second model you fit to the data. Be sure to state the null and alternate hypothesis under investigation, the test statistic, p-value, and appropriate conclusion. You can do this test (1) using R, and be sure to include your output, or (2) by hand, and be sure to show your calculations. (NOTE: Make sure you are comfortable with both approaches!).

(f) Using the model you identified as the preferred model, demonstrate how the model can be used to predict the probability that a coal mine pillar will be stable (this will require you to select reasonable values for the explanatory variables).

5. A toxicity experiment of copper on flathead minnows involved exposing a fixed number of flathead minnows to a given dose of copper and observing how many minnows died. The data from the experiment is show below[2]:

Dosage (mg/L)	Number Exposed	Number Dead
5	25	3
10	30	9
20	30	15
40	25	21
80	40	38

(a) Plot the relationship between the proportion of flathead minnows that died and the $\log_{10}$(Dose).

- Note that the use of the $\log_{10}$(Dose) on the horizontal axis is to create a more condensed scale on which to visualize the relationship between proportion of deaths and dose of chemical.
- There are two ways to plot $\log_{10}$(Dose): (1) Create a new variable of $\log_{10}$(Dose) and plot this with the proportion of deaths, or (2) Use the option `log = ‘‘x’’` in the plot command, which will create a semi-log graph with the X-axis on the $\log_{10}$ scale. Both with give you the same shape of the curve, but with slightly different scales.

(b) Fit a logistic regression model to the data set (note: the model should include dose as the explanatory variable, not $\log_{10}$(dose)). Comment on the significance, magnitude, and direction of the effect of dose on the *odds* of a flathead minnow dying. Include the relevant R output from the fitted model in your final solutions as a figure.

[2]Adapted from Question 5, page 272 in *Environmental Risk Assessment*. (2003). Hubert, J.J. Department of Mathematics and Statistics, University of Guelph.

(c) Plot the fitted model on the same graph as your observed data points. There are a few ways to do this:

- First, you will need to create a vector of doses, ranging from some selected lower limit to some selected upper limit. Use the `seq` command to create this vector.
- Second, you will need to estimate the probability of dying for each value in your dose vector. You can do this by:
  - (1) Manually calculating p for each value of dose, using the relationship $p = \frac{\exp(\beta_0+\beta_1*Dose)}{1+\exp(\beta_0+\beta_1*Dose)}$
  - (2) Using the function `ilogit` in the `faraway` library, which essentially just does the above calculation for you.

(d) Use the function `dose.p` in the `MASS` library to estimate the LD_{50} and LD_{90} values. Add these points onto your plot of proportion of deaths versus log_{10}(Dose). Include this plot (with observed proportions, estimated regression line, and LD_X values incorporated) in your final solutions. Comment on how well the regression line seems to fit your data.

(e) Create 95% confidence intervals for the estimated LD_X values in part (d). Include these confidence intervals in your final solutions, and provide an appropriate interpretation of each interval.

6. A study was conducted to investigate the effect of caffeine consumption on fertility. The risk factor was defined as consuming more than 300 mg per day, and the outcome of interest was if conception was delayed (for example, if it took more than a year to conceive). The data are displayed in the table below[3]:

		Caffeine Consumption	
		Yes	No
Delayed Conception	Yes	650	560
	No	350	440

(a) Determine the relative risk of experiencing delayed conception in the caffeine consumption group versus those that do not consume excessive caffeine. Show your work, or include your R output if you use R. Provide an interpretation/conclusion from this value.

The coffee industry felt that this analysis oversimplified the problem, and believed that alcohol consumption could be a contributing factor to fertility issues. They insisted that information on alcohol consumption also be recorded as part of the study. The data is displayed in the following table:

Alcohol Consumption:		YES		NO	
Caffeine Consumption:		YES	NO	YES	NO
Delayed Conception	YES	560	160	90	400
	NO	240	40	110	400

(b) Determine the relative risk of experiencing delayed conception in the caffeine consumption group versus those that do not consume excessive caffeine, within each of the alcohol consumption groups. Show your work, or include your R output if you use R. Provide an interpretation/conclusion for the values you calculate.

[3](Adapted from Example 2, page 93 in *Environmental Risk Assessment*. (2003). Hubert, J.J. Department of Mathematics and Statistics, University of Guelph.)

(c) In your own words, explain the apparent contradiction you are seeing between parts (a) and (b).

(d) Which analysis, (a) or (b), do you believe to be more useful, and why?

(e) Draw a carefully labelled "Detailed Confounding Diagram with Probabilities", as shown in class. Please draw to scale (use software if you can, but not required), and describe the plot in a few brief sentences.

(f) If you were a representative of the alcohol industry, what might be your response to this study (in only 2 - 3 sentences).

7. For this question, you will need to access the article:

   Page, L. Savage, D.A., and Torgler, B. Variation in Risk Seeking Behaviour Following Large Losses, A Natural Experiment. (2014). *European Economic Review*. Vol. 71, pp. 121-131.

   This article can be accessed through the University of Guelph Library website. For all of the following questions, make sure you provide an explanation/justification of your answers.

   (a) Would you describe this study as an observational study or an experiment?

   (b) The primary variables of interest can be described as *affected by the flood* (yes/no) and *engage in risky behaviour* (yes - lottery ticket, no - $10 cash). Which do you believe is the "exposure", and which is the "outcome"?

   (c) Would you describe this study as a cross-sectional, cohort, or case-control study?

   (d) Focussing on the data presented in Table 2, the Authors' used a chi-square test to determine if there was an association between the two variables. Using an alternative method (appropriate for the type of study you selected in part (c)), confirm the results of the paper that there is a relationship between these two variables. Comment on the magnitude, direction, and significance of this relationship in your conclusion.

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 #1. 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.