**Workshop: Practice General Linear Models**

In this workshop, you have the opportunity to analyze up to 4 datasets, which cover themes from previous discussion of general linear models. If you complete this work as part of a class, we encourage you to work in small groups.

For each question, you should follow the general workflow for analyzing General Linear Models:

1. Familiarize yourself with the dataframe you’ve created. Note the column names and the names of variables within columns. Think of what types of variables you have.
2. Plot the data as you’ve been shown in lectures. Ask yourself: Do any treatment appear different from others? Do you think the data will meet the assumption of a normal distribution?
3. Think: How should these data be analyzed to accomplish the scientific question posed to you?
4. Construct a statistical model, and use it to address your scientific question.
   1. Check your assumptions before looking at the final results
   2. If your model does not meet the assumptions of your test, either: (i) transform the data until assumptions are met, or (ii) choose a new, and more appropriate statistical test.
   3. Interpret the outcome of your model, including any necessary post-hoc tests (or produce an interaction plot),to identify and understand any differences that lie in your data.
5. What do you conclude? Base your conclusions on both p-values and effect sizes.

1. Chauvet et al (2016; J Neuroendocrinology) studied the level of expression of 2 genes (CDH1 and CDH2) that code for cell-adhesion molecules in samples taken from pituitary human tumours. The tumours can be classified as invasive (“I”) vs. non invasive (“NI”). They were sampled from two different types of tumours: either from growth hormone adenomas (“GH”) or prolactinomas (“PRL”, tumours that overproduce growth hormone or prolactin). Hence, gene expression was examined for two types of “treatments”: Invasiveness (I vs NI) and tumor phenotype (GH vs PRL). All the tumours come from different patients, and the dataset contains at least 8 data points for each combination of Invasiveness and Phenotype categories. Note that these data are **unbalanced**.

The data are in the file, “pitTumours.csv”.

Analyse these data to test whether gene expression of CDH1 (***not*** CDH2, which is also in the dataset) depends on Invasiveness and Phenotype. The expression of CDH1 is in the column, “CDH1”.

1. Upload the data into R.
2. Follow the general workflow, outlined, above. *Hint: consider a square-root transformation.*
3. What do you conclude?

2. Cook et al. (1993) studied the reduction of Hippocampal volume (as a %) in 107 patients with various forms of drug-resistant epilepsy. The patients fell into one of 3 groups: a) a record of childhood febrile seizures (“CFS”); b) childhood non-febrile seizures (“no CFS”); and no childhood seizures (“no seizures”).

The data are in the file, “seizuresCFS.csv”.

Analyze these data to determine whether the loss in Hippocampal volume differs among the groups.

1. Upload the data into R.
2. Follow the general workflow, outlined, above. Data expressed as a % are often better analyzed with an arcsine square-root transformation, as given here: asin(sqrt((seiz$hippoVolumeRatio)/100))
3. What do you conclude?

3. Mutations that occur during sperm production can be passed on to a father’s offspring. Older human fathers have had more time to accumulate mutations in tissues that produce sperm cells. As a result, we might predict that, on average, older fathers pass on more mutations to their offspring than younger fathers do. Kong et al. (2012) used complete genome sequencing of 21 human father-offspring pairs to count the number of mutations given by fathers to their children. Kong et al. (2012) also recorded each father’s age.

These data are recorded in the file, “mutations.csv”.

Analyse these data to test whether the number of mutations passed on by a father depends on the father’s age.

1. Upload the data into R.
2. Follow the general workflow, outlined, above.
3. What do you conclude?

4. Mole rats are the only known mammals with distinct social classes. Like bees, a single queen and a small number of males are the only reproducing individuals in a colony. Remaining individuals, “workers”, do the work (gather food, defense care for the young, etc). Recent data made researchers question whether there were actually two types of workers that appeared to differ in the amount of work they did: “worker” and “lazy”. Scantlebury et al. (2006) wished to study the physiology of the two supposed worker types. They were interested in two questions. Firstly, they wished to examine the relationship between an individual’s mass and their daily energy expenditure. Secondly, they wished to know whether this relationship was the same for the two supposed worker classes. Weight and energy expenditure were evaluated for each mole rat, which was classified as either “worker” and “lazy”. Note that the data are **un-balanced**.

These data are available in the file, “moleRats.csv”.

Analyse these data to test whether: a) energy expenditure increases with body mass; b) whether this relationship differs between “worker” and “lazy” groups.

1. Upload the data into R.
2. Follow the general workflow, outlined, above.
   1. NOTE: the data have already been log-transformed, for convenience
   2. It is possible to produce a scatterplot with different colours that represent different treatments. To do this, we add the option “col” to a plot() function, and write a little function that uses different colours for different groups. For this question, try this:

plot(lnMass~lnEnergy, xlab="ln(Body mass)", ylab="ln(Energy expenditure)", pch=2, col=ifelse(caste=="worker", "red", "blue"),data=mr)

1. What do you conclude?

5. The figure, below, is found in the paper, Marland et al. (2020: <https://doi.org/10.1016/j.sbsr.2020.100375>). In it, the authors use 2 t-tests to interpret their results: one t-test at time zero, and another t-test at time 24 hours. In each case, they compare a measurement under two conditions: Control vs. BSA treatment. Although not explicitly stated in the text, they appear to use these two t-tests to demonstrate that the effect of BSA vs. control differs between the two time points, 0 vs. 24. Is their approach appropriate to make this conclusion? If not, how could they improve their approach?

