Data analysis practice

On Blackboard there are five datasets in the folder called "Data Analysis Practice".

Your task is to analyse each one to get answers to some questions below. You already have the tools to analyse these data appropriately. The best analysis might be a t-test, or a regression, or perhaps a generalized linear model (GLM) with an appropriate error structure and link. It is up to you to decide how to analyse them.

I describe each dataset below. Use your judgement and perhaps look at plots of the data to decide which statistical tools to use. You can also find inspiration in the previous handouts.

Ask for help if you get stuck!

Dataset 1: Maze [maze.csv]
Adult and child subjects were tested on navigating a complex maze. The average number of mistakes they made was recorded after many trials.

> *Q1 – are children less good at navigating mazes than adults?*

Dataset 2: American football field goals [NFLfieldgoal.csv]
These data are a record of all of the field goal attempts in the US National Football League (American Football) season of 2003. A field goal is a means of scoring in American football. To score a field goal, the team in possession of the ball must kick the ball through the goal, i.e., between the uprights and over the crossbar, during a play.
The data has 2 columns: the distance in meters, and a 0/1 indicator of whether it was successful (1) or not (0).

> *Q1 – is there a significant correlation between the distance and the probability of scoring?*
> *Q2 – from a plot, what is the approximate probability of scoring from a distance of 40m.*

Dataset 3: Morphometry in humans [morphometry.csv]
This dataset is composed of accurate height, hand length and foot length data on men and women.

> *Q1 – Hand length and foot length are both highly correlated with height (obviously). But does sex change the correlation: i.e. does the effect of hand length/foot length depend on sex?*

Dataset 4: Species, biomass and soil acidity [species.csv]

These data give the count of number of plant species on quadrats of land that have different biomass and different soil pH. The pH has been categorized as being low, mid, or high.

> *Q1 – How does the number of species vary with increasing biomass?*
>
> *Q2 – What is the effect of soil pH, and does pH influence the effect of biomass (i.e. is there an interaction?).*

Dataset 5: Species richness on British islands [britainSpecies.csv]

This dataset is of the number of species on 42 British islands. The data include Area, Elevation, number of soil types, latitude, and distance from Britain.

> *Q1 – fit a suitable model to look at the relationship between area and the number of species. Remember to remove "Britain" itself from the data first.*
>
> *Q2 – fit a model with all the predictors (area, elevation, number of soil types, latitude and distance). Simplify the model as much as possible. Remember that you might want to create new logged versions of some of the variables.*