LT5 and LT7 Worksheet

Preliminaries

For this worksheet, I would like you to conduct and interpret a 2 x 2 between-person ANOVA and logistic regression analyses using what you have learned from LT5 and LT7.

As always, please write your answers and code in an R notebook and submit it as a .pdf file in the Moodle submission point.

Before we start, lets load the necessary packages: "tidyverse", "car, and"readr".

```
library(tidyverse)
etc
```

2 x 2 Between Persons factorial ANOVA

To work on testing and interpeting 2 x 2 between persons ANOVA, we are going to use data from an experiment in which alertness level of male and female subjects was measured after they had been given one of two possible dosages of a drug called duluxetine. The dataset can be downloaded from here: https://www.dropbox.com/s/b5hh2bllsybtm0w/alert.csv?dl=0. The file is also avalaible in the PB130 R project on the github under the folder worksheets –> LT3 and LT7.

As a first step in the analysis, download the file and save it to a **logical** place where it can be easily located. I would recommend that you save it in your PBS folder under a subfolder called Worksheets. Now load the file by clicking on the dataframe name in the files window where you saved the protest data and click import data". A new screen will appear with the dataframe. Press"update" in the top right corner and then when it loads click import.You can also copy the R code produced in the bottom right corner and paste that code in an R chunk on your worksheet. For example:

```
alert <- read_csv("Worksheet 4/alert.csv")
alert
```

Study Description

This study investigated the impact of an antidepressent drug called Duloxetine on self-reproted alterness among clinical impatients. The researchers had a couple of aims. The first was understand whether the dose of the drug influenced alertness. Two doses were given to participants - a 'low' dose of 25mg and a 'high' dose of 40mg. The second aim was to understand whether the effect of the drug was influenced by gender - 'male' or 'female'.

Alongside these main effects, the researchers also wanted to know whether the effect of dose on alertness was moderated by gender.

In this research alterness was the outcome variable and scored on a continuous scale from 1 to 30 and measured at one time point only. Higher scores indicate higher altertness. Gender and dose were categorical predictor variables each with 2 levels. Thus, this is a 2X2 between persons ANOVA design with an interaction.

Activity

Research Question 1 - Is there a significant difference in alertness between patients who took a low or high dose of Duluextine?

Research Question 2 - Is there a significant difference in alertness between patients who took were male or female?

Research Question 3 - Is the effect of dose on alertness moderated by gender?

Null Hypothesis 1 - The alertness difference between patients who took a low or high dose of Duluextine will be zero.

Null Hypothesis 2 - The alertness difference between patients who were male or female will be zero.

Null Hypothesis - The interaction of dose and gender on alterness will be zero.

There are several steps needed to conduct this analysis:

Task 1

Lets build the moderated ANOVA model of alertness using the aov() function and save it as a new R object called "anova.model". The aov() function takes categorical not numerical input for the predictors and therefore alert dataframe has categorical predictor variables.

In the same chunk, request the ANOVA summary and then comment on the main effects for; (1) dose, (2) gender, and (3) the dose*gender interaction.

Task 2

Given we have main effects for dose and gender, we need to conduct post-hoc analyses that examine all possible group comparisons being tested using something called pairwise differences. These kinds of comparisons are often called simple effects, apparently referring to the fact they are just comparing means in a straight forward way.

In the anonva.model there are two specific significant mean comparisons being tested:

1. The difference in alertness between those who high a high dose of the drug and those who had a low does of the drug
2. The difference in alertness between males and females

We discussed several options for simple effects in the lecture and workshop. Of those options, a tool called Tukey's "Honestly Significant Difference" (or Tukey's HSD for short) is the most widely applied test for 2 x 2 factorial ANOVA. It constructs 95% simulataneous confidence intervals around each specific mean comparison. Simultaneous just means that there is a 95% probability *all* of these confidence intervals include the true value in the population. To all intents and purposes, it is interpreted in the same way we have been interpreting 95% confidence intervals to date.

In this R chunk, go ahead and request Tukey's HSD using the TukeyHSD() function and input the anova.model you have just built.

Then, comment on the simple effects for dose and gender.

Task 3

Having conducted the ANOVA and simple effects, have a go at writing up the analysis using the template provided in the workshop and lecture:

Logistic Regression

Introduction

To work on testing and interpeting logistic regression, we are going to use data from the Premier League on penalty kicks to examine whether there is a relationship between the time a penalty was awarded and whether it was scored by the taker. The data file is named penalty and can be downloaded from here: https://www.dropbox.com/s/o3phzkfx6kgjla8/penalty.csv?dl=0. The file is also avalaible in the PB130 R project on the github under the folder worksheets –> LT5 and LT7.

As a first step in the analysis, download the file and save it to a **logical** place where it can be easily located. I would recommend that you save it in your PBS folder under a subfolder called Worksheets. Now load the file by clicking on the dataframe name in the files window where you saved the protest data and click import data". A new screen will appear with the dataframe. Press"update" in the top right corner and then when it loads click import.You can also copy the R code produced in the bottom right corner and paste that code in an R chunk on your worksheet. For example:

```
protest <- read_csv("Worksheet 4/penalty.csv")
protest
```

Data Description

In this data are 439 penalties awared to teams in the Premier League. The data frame contains a variable called scored, which is a categorical variable that measures whether the penalty was scored or not scored (scored = 1, not scored = 0). The data frame also includes a variables called time, which is a continuous predictor measuring the time in the game that the penalty was awared. The Premier League want to know if there is a relationship between the time a penulty was awared and whether it was scored or not. They suspect that penalties awared later in games would be scored less often that those awared earlier in games, but they do not know. This type of research question requires simple logistic regression as there is one categorical outcome variable - 'scored' - and one continuous predictor variable - 'time'.

Before we investigate this relationship, though, lets just clarify the research question and null hypothesis you will be testing:

Research Question - Is there a relationship between game time and penalties scored in Premier League matches?

Null Hypothesis - The relationship between game time and penalties scored will be zero

To test this research question, there are several tasks needed:

Task 1

To build the logistic regression model we will use glm() rather than lm(). This is becuase logistic regression is a special case of the linear model - one in which the outcome is estimated in terms of log(odds) rather than raw units. We need to do this becuase the outcome only goes from 0 to 1. But otherwise, the coding is exactly the same as we have been working with to date for linear regression.

For the simple logisitic regression model, lets create a new R object called penalty.model and tell R to run a logistic regression on the categorical outcome 'scored'. The code of the glm function is similar to that of lm, except that we must pass the argument family = binomial in order to tell R to run a logistic regression rather than some other type of generalized linear model.

When you have run this model, call the summary for it and then comment on the meaning of the slope estimate and whether it is statisticially significant.

```
penalty.model <- glm(??, data = penalty, family = "binomial")
summary(??)
```

Task 2

Now use the Boot() function to create 5,000 resamples with replacement, estimating the logistic regression model parameters on each occasion, and save them in a new R object called "penalty.boot". Then, use the confint() function to request the bootstrap confidence intervals for the estimates. Comment on the bootstrap confidence interval of the slopefrom the output.

```
penalty.boot <- Boot(??, f=coef, R = ??)
confint(??, level = ??, type="norm")
```

Task 3

To interperet the slope estimate, we need to exponentiate it so that we arrive at an estiamte expressed in terms of odds-ratio. Remember, to get the exponentiated coefficients, you tell R that you want to exponentiate (exp), and that the object you want to exponentiate is called coefficients and it is part of penalty.model (coef(penalty.model)).

Go ahead and calcuate the odds ration for the time estimate and comment on the meaning of this in terms of the relationship between time and scoring (remeber that this is a negatvie relationship).

```
exp(coef(??))
```

Task 4

Finally, use the anova() function to call the chi-square model fit for the the logistic regression model and comment on the meaning of the chi-sqaure and significance.

```
anova(??, test="Chisq")
```

Task 5

Having conducted the logistic regression, have a go at writing up the analysis using the template provided in the workshop and lecture: