School of Computing Sciences

Module: **CMP-7008A Applied Statistics**
Assignment: **Course Work 2**

Set by: Chris Greenman (email: C.Greenman@uea.ac.uk)
Date set: 12th December 2023
Value: 50%
Date due: 18th January 2024 3pm
Returned by: 9th February 2024
Submission: Blackboard
Checked by: KH

Learning outcomes

To be able to apply ANOVA and Survival Analysis models. To be able to communicate results effectively.

Specification

Overview

This assessment covers the theory and application of ANOVAs and Survival Analysis.

Description

Answer all THREE questions (see below for details). Present solutions to all problems. Provide full working to demonstrate the steps required to obtain the correct solution, and the principles used. The logic of the solution is important. All questions should include any R code and output used as part of the explanation.

Deliverables

The assignment should be submitted on blackboard as ONE file, in PDF format (multiple files are not acceptable - phone photos can be placed in a single word document and saved as pdf), including your student number.

Resources

Lectures and seminar notes.

Marking Scheme

Accuracy of answers and code. Understanding of subject displayed. Clarity of working and logical argument.

Plagiarism, collusion, and contract cheating

The University takes academic integrity very seriously. You must not commit plagiarism, collusion, or contract cheating in your submitted work. Our Policy on Plagiarism, Collusion, and Contract Cheating explains:

- what is meant by the terms 'plagiarism', 'collusion', and 'contract cheating',
- how to avoid plagiarism, collusion, and contract cheating,
- using a proof reader,
- what will happen if we suspect that you have breached the policy.

It is essential that you read this policy and you undertake (or refresh your memory of) our school's training on this. You can find the policy and related guidance here:

https://my.uea.ac.uk/departments/learning-and-teaching/students/academic-cycle/regulations-and-discipline/plagiarism-awareness

The policy allows us to make some rules specific to this assessment. Note that in this assessment working with others is not permitted. All aspects of your submission, including but not limited to: research, design, development and writing, must be your own work according to your own understanding of topics. Please pay careful attention to the definitions of contract cheating, plagiarism and collusion in the policy and ask your module organiser if you are unsure about anything.

If you have any queries try my office (BIO2.18) or email me: C.Greenman@uea.ac.uk

Applied Statistics Assignment 2 Questions

1. An experiment analysing the acidity of balsamic vinegar in four new factories (locations London, Birmingham, Sydney and New York) along with an established factory (in Paris), measures the pH level from 36 samples in total. The acidities were collated and can be found in the file *vinegar.csv* on blackboard. Answer the following six questions.

   (a) Load the data into R, and display the acidities graphically for the different types. Provide a description of what you observe. *[marks 8]*

   (b) Implement a one way ANOVA to determine if there are any significant effects, stating any null or alternative hypotheses that may apply. *[marks 8]*

   (c) State any assumptions made in the ANOVA and investigate whether these assumptions are satisfied. *[marks 8]*

   (d) Determine whether any new factories have significantly different vinegar acidity to the established factory. Determine whether the new factories have consistent pH levels. *[marks 8]*

   (e) Implement (i) Holm and (ii) Bonferonni corrections for multiple testing on all p values analyzed above and comment on any changes to the conclusions. *[marks 8]*

   (f) Comment on the results of the experiment and how it could be improved (max 6 lines). *[marks 5]*

2. The data in the file $VenomYield.csv$ contains the venom yield (mg) from some male funnel-web spiders. The *Body Length* (cm) of the spiders are also known, which are also grouped into a *Body Class* variable (small or large). The *Expression* of a gene (thought to be related to venom yield) is also known for each snake (low or high). You may presume the assumptions for ANOVAs and ANCOVAs are satisfied in the following questions.

   (a) Visualize the data and interpret what you see. *[marks 9]*

   (b) Run a two way ANOVA to look for significant effects, using *Body Class* and *Expression* as factors, and provide an interpretation of what you discover. *[marks 9]*

   (c) Run an ANCOVA using *Expression* factor and *Body Length* covariate to look for significant effects, and provide an interpretation of what you discover. *[marks 9]*

   (d) Comment with reasons on whether *Body Length* is a suitable covariate for the ANCOVA. *[marks 5]*

   (e) Comment with reasons on which of (b) or (c) is the better approach. *[marks 5]*

3. The lifetimes of industrial air conditioning units are monitored over a 10 year period.

   (a) It is thought the hazard function (in years) for the units is $h(t) = 1 - e^{-t}$. Determine the survival function $S(t)$ and the failure probability density function $f(t)$ and plot the results (either manually or with R). *[marks 9]*

   (b) A population of 100 such units are annually monitored and failures recorded, along with censoring events (where units where removed for other reasons (e.g. stolen, no longer required). The failures over the decade are $1, 0, 3, 4, 11, 8, 8, 15, 17, 10$. The censoring events are $0, 0, 2, 0, 0, 5, 3, 4, 2, 1$. Produce a lifetable for the failures and plot the functions $S(t)$, $h(t)$ and $f(t)$. Comment on whether the model in (a) looks accurate. *[marks 9]*