**BIFX 503: Biostatistics in R**

**Final Examination**

**December 8, 2021**

**J. Jedediah Smith**

*Instructions:*

*This is an open-book, open-note exam … books being the course textbook (Hothorn & Everitt) and notes being your personal notes taken for class, homework assignments, and course slides.*

*Permitted technology for calculations is limited to R, Excel, or a calculator. No internet use (other than Blackboard) or smartphone use is permitted during the exam.*

*Type your answers into this Word document, and paste in results from R as needed. You may use RMarkdown if you prefer. When you finish the exam, submit it via Blackboard. The exam is due at 11:59pm on Wednesday December 8.*

*The old cliché is true – you miss 100% of the shots you don’t take ☺ Please don’t leave any questions blank! Partial credit will be given for incomplete responses.*

*Good luck!*

Otitis media is a common infection in young children. A clinical study evaluated whether encouraging treatment adherence improved treatment outcome in 50 children with a history of otitis media. The dataset **bacteria.csv** contains the data from this study and consists of the following variables: ID (participant ID), week (week of the study), bac\_pos (tested positive for bacteria, 1=yes 0=no), tx (1=active, 0=placebo), hilo (level of adherence encouraged hi or lo). The children were randomized to receive either active drug or placebo (variable tx), and also randomized to receive either high or low encouragement to adhere with treatment (variable hilo). The presence of the bacterial strain *H influenzae* was assessed at four time points (weeks 0, 2, 4, and 11).

1. Run a repeated-measures logistic regression model that consists of bac\_pos as the dependent variable, and the following independent variables: week, tx, hilo, and the interaction tx\*hilo. Report the results as odds ratios with 95% confidence intervals for each of the four terms in the model. You can use either geeglm() or glmer().

For week, the odds ratio was 0.890 and the interval was 0.815 to 0.97. For tx, the odds ratio was 0.401 and the interval was 0.130 to 1.11. For hilo, the odds ratio was 0.539 and the interval was 0.153 to 1.89. For tx:hilo, the odds ratio was 1.176 and the interval was 0.260 to 5.29.

Text

Description automatically generated with medium confidence

1. Are any terms statistically significant? What is the direction of the association and what does that mean?

The only term under the 0.05 threshold for significance is week with a value of 0.0084. We can tell from its odds ratio of 0.890 that it has a negative association with bac\_pos because it is under 1. Therefore, a unit increase in week makes the bacteria positive outcome 0.890 times less likely.

A picture containing table

Description automatically generated

1. What do the results tell you about the relationship between the two treatments, time, and the chances of a child in the study being bacteria-positive?

These results seem to indicate that the different treatments were not significant factors in determining whether a child is bacteria positive. The only significant factor appears to be time. As time goes on, the child is less likely to be bacteria positive.

The dataset **CPUs.csv** contains data on the relative performance of 209 CPUs. The variables include manufacturer/model, cycle time, minimum and maximum main memory, cache size, minimum and maximum number of channels, and log10 of performance.

1. First, examine the distributions of each of the seven continuous variables in this dataset using histograms. (Note: you are excluding the variable Manufacturer\_model). Which ones look normal? For the skewed distributions, choose a transformation that will normalize them, and regenerate the plots for the transformed variables. *Note: to save time, don’t try multiple transformations. Just choose one reasonable one, and go with it.*

Initial histograms. Only log\_perf is normal, others need to be transformed.

Chart, histogram

Description automatically generated

Transformed histograms with sqrt. Slightly more normal. Further transformations not performed as per question instructions.

Chart, histogram

Description automatically generated

1. Perform a hierarchical cluster analysis on the 7 continuous variables, using transformed versions of skewed variables, and plot the dendrogram. How many clusters are apparent? Draw rectangles around the clusters you identified using rect.hclust().

There appeared to be 3 apparent clusters. The graph is below.

Diagram, schematic

Description automatically generated

1. Cluster analysis requires that data have no missing observations. How can deleting all records with missing values prior to analysis compromise the validity of your results?

There could be good reasons for why the data is missing. Deleting it will bias the results in favor of the subjects that, for whatever reason, don’t have missing data. Furthermore, if a significant portion of the data has missing values, it may no longer accurately represent the majority of the population being studied.

The dataset **VA** (from the MASS package) contains results from a clinical trial of treatment for lung cancer. The time-to-event variable is stime (in days), the censoring variable is status (0=censored, 1=dead), treatment variable is treat (standard or test). Use ?VA for a complete description of the variables.

1. Visualize the relationship between survival and treatment using a Kaplan-Meier plot.

There were two treatment types. Standard (treat=1) is represented by the solid line. Test (treat=2) is represented by the dotted line. As show in the graph below, the test group participants survived for almost twice as long as the standard group ones at low survival probability levels.

Chart, histogram

Description automatically generated

1. Calculate the median survival time for each treatment group and compare. Does treatment appear to prolong survival in these lung cancer patients?

Median survival time is 103 for standard (treat=1) and 52.5 for test (treat=2). This appears to indicate that the test treatment actually reduced the survival time for patients.

Text

Description automatically generated

1. Use a log-rank test to test the hypothesis that treatment affects survival.

The p-value of 0.9 is far above the 0.05 threshold for significance. This would seem to indicate that the differences observed are not significant.

Text

Description automatically generated with low confidence

1. Run a Cox proportional hazards model that includes treatment, age, and Karnofsky score. What are the results? What does this tell you about the determinants of survival in this clinical trial?

The results show that only the Karnofsky score has a p-value below the 0.05 threshold of significance. Therefore, it appears that Karnofsky score is a strong determinant of survival, while treatment and age are not.

Text

Description automatically generated with medium confidence