In-semester Exam

Week 8 2021

Name:

SID:

Instructions

- The exam will run for 2 hours.
- Use the pre-filled rmarkdown file from **Canvas** to help arrange your answers. You are welcome to use or ignore the code chunks I have inserted or add additional ones.
- Submit a final html file to **turnitin** in Canvas to be marked.
- There are 3 questions. The first two are worth 40% each, the third is worth 20%.
- The exam question sheet is 6 pages long.

Question 1 (40%)

Question 1 has 3 parts.

The Department of Plastic Surgery, University Hospital of Odense, Denmark performed a study on patients with malignant melanoma during the period 1962 to 1977. Each patient had their tumour removed by surgery. The surgery consisted of complete removal of the tumour together with about 2.5cm of the surrounding skin. Among the measurements taken were the thickness of the tumour and whether it was ulcerated or not. These are thought to be important prognostic variables in that patients with a thick and/or ulcerated tumour have an increased chance of death from melanoma.

Variable	Description
time	Survival time in days since the operation, possibly censored.
status	The patients status at the end of the study. 1 indicates that they had died from melanoma, 2 indicates that they were still alive and 3 indicates that they had died from causes unrelated to their melanoma.
sex	The patients sex; 1=male, 0=female.
age	Age in years at the time of the operation.
year	Year of operation.
thickness	Tumour thickness in mm.
ulcer	Indicator of ulceration; 1=present, 0=absent.
location	Anatomical location of first identified tumour.

Part 1

```
melanoma = read.csv("https://wimr-genomics.vip.sydney.edu.au/AMED3002/data/e21/melanoma2021.csv")
```

a. How many variables and observations are in the dataset?

b. Comment on the class of these variables and how they are stored in R.

c. Is there any missing data in this dataset?

Part 2

The researchers would like to understand the relationships between tumour thickness and the other variables.

d. Test to see if tumour thickness is different between people with and without ulceration. Check your assumptions and make a conclusion using a signficance threshold of 0.05.

e. Test to see if tumour thickness is higher in females than males. Check your assumptions and make a conclusion using a signficance threshold of 0.05.

f. Fit a two-way ANOVA model without an interaction effect to assess if ulceration and sex both have a relationship with tumour thickness when included in the same model. Be sure to comment on the assumptions needed for this test.

g. Create an interaction plot with sex and ulceration as factors and thickness as the response. From this plot, is there evidence that there is an interaction effect? Why?

h. Include an interaction term in the two-way ANOVA model to assess if ulceration and sex both have a relationship with tumour thickness. What is your conclusion?

Part 3

The researchers would like to test if age has an effect on tumour thickness.

i. If there was no relationship between tumour thinkess and age, what value should the slope of an appropriate regression model be?

j. Fit a regression model and then assess and comment on the model assumptions.

k. What would you conclude from the test?

l. What proportion of variation in tumour thickness can be explained by age?

Question 2 (40%)

Question 2 has 2 parts.

The Cleveland Heart Clinic in Ohio, USA performed a study to understand the risk factors associated with developing heart disease. Between May 1981 and September 1984 they recorded clinical and noninvasive test results for 303 consecutive patients referred for coronary angiography. No patient had a history or electrocardiographic evidence of prior myocardial infarction or known valvular or cardiomyopathic disease. The information they recorded is listed in the following table.

Variable name	Short desciption
age	Age of patient
sex	Sex (0 for male)
cp	chest pain
trestbps (measured in mmHg)	resting blood pressure
chol	serum cholesterol
fbs	fasting blood sugar larger 120mg/dl (1 true)
restecg	resting electroc. result (1 anomality)
thalach	maximum heart rate achieved (measured in in bpm)
exang	exercise induced angina (1 true)
oldpeak	ST depression induc. ex.
slope	slope of peak exercise ST
ca	number of major vessel
disease	diagnosis of heart disease (1 true)

Part 1

```
heart = read.csv("https://wimr-genomics.vip.sydney.edu.au/AMED3002/data/e21/heartDiseaseV1.csv")
```

From clinical experience the researchers believe that there may be a relationship between sex and whether an individual had heart disease.

a. What types of variables should sex and whether they had heart disease be?

b. What is an appropriate statistical test that could be used by the researchers to test this question?

c. What is the corresponding null and alternate hypothesis?

d. Construct a contingency table using the variables *sex* and *disease*.

e. Perform the appropriate test.

f. Using a siginficance threshold of 0.05 what would you conclude from this test?

g. What were the assumptions for this test? Comment on them in the context of the observed data.

h. Report and interpret the corresponding Odds Ratio.

Part 2

Use logistic regression to model the chance of having heart disease. For the following do not check model assumptions:

i. Fit a logistic regression model that uses all of the variables to model heart disease. Comment on which variables appear to be informative.

j. Use backwards step-wise variable selection to fit a simpler model. How many variables are included in the model?

k. When the step-wise variable selection was performed, which model fit criteria was used to decide whether a variable should be included or not? Feel free to use an acronym.

l. From either one of the logistic regression models, calculate odds ratios and comment on the relationship between heart disease and the maximum heart rate achieved.

Question 3 (20%)

Question 3 has 2 parts

In a manuscript by Andrews, et al. (1985), 79 urine specimens were analyzed in an effort to determine if certain physical characteristics of the urine might be related to the formation of calcium oxalate crystals. They recorded the following variables.

Variable	Description
crystals	Indicator of the presence of calcium oxalate crystals.
gravity	The specific gravity of the urine.
ph	The pH reading of the urine.
osmo	The osmolarity of the urine. Osmolarity is proportional to the concentration of molecules in solution.
cond	The conductivity of the urine. Conductivity is proportional to the concentration of charged ions in solution.
urea	The urea concentration in millimoles per litre.
calc	The calcium concentration in millimoles per litre.

We will start by reading in the data and exploring the properties and structure of the data. The data can be read in as follows:

Part 1

```
urine = read.csv("https://wimr-genomics.vip.sydney.edu.au/AMED3002/data/e21/UrineDataV1.csv")
```

a. How many variables and observations are in the dataset?

b. Comment on the class of these variables and how they are stored in R.

c. Use a visualization to check if the dataset has any missing data?

d. In a sentence, explain why you would conclude that the data is either MCAR, MAR or MNAR?

e. Perform case deletion.

Part 2

f. Use hierarchical clustering to cluster the variables. Hint: you made need to use `t()`.

g. Comment on why you did or did not decide to *scale* the data when performing the analysis.

h. How does this clustering inform the researchers' primary question?

i. Use k-means clustering to cluster the observations in the dataset. Use all of the variables except for `crystals` to cluster, and, set a seed of `51773` before you cluster.

j. Why is it advisable to set a seed?

k. Is there any evidence of a relationship between this k-means clustering and the formation of crystals?

l. How does this clustering inform the researchers' primary question?