The Health Department is conducting a survey to gain insight into the general health of the population. To reduce the burden on survey participants and to increase the likelihood that they will complete the survey, each survey participant is only required to provide two answers on the survey: how many days they exercise each week (i.e., their number of "weekly exercise days") and their weight (in kilograms). Given the budget available to conduct this survey, the Health Department decided to run this survey each month for the first five months of the year, where for each month a new random sample of participants was selected. The data are stored in the file AssignmentData.RData in the data frame Q1.df. The data frame contains two columns for each month, one for the participants' number of weekly exercise days and one for their weights. For example, for the survey conducted in January, there are two columns named ExerciseJan and WeightJan.

For parts (a) to (d), you will be analysing data from the survey conducted in February.

(a)The Health Department's budget only allows them to collect one random sample each month. A data analyst at the Health Department suggests that the sample of participants selected for the survey conducted in February could be reused by "resampling" these participants. Specifically, she suggests randomly selecting some participants from the original sample of participants selected for the survey conducted in February with replacement and using this "new" sample of participants to perform further statistical analyses. Resampling in this manner is sometimes referred to as m-out-of-n bootstrap resampling. For a new sample of 15 participants selected in this manner, find the probability that more than 10 participants exercised more than 2 days each week.

(b) Another data analyst at the Health Department is opposed to the m-outof-n bootstrap resampling suggested by the previous data analyst. Instead, he thinks it would be better to randomly select some participants from the original sample of participants selected for the survey conducted in February without replacement. Resampling in this manner is sometimes referred to as subsampling. For a new sample of 10 participants selected in this manner, find the probability that either 2 or 3 participants exercised more than 4 days each week.

(c) In February, test whether the population proportion of people who weigh more than 99 kilograms is more than 0.015. Clearly state your hypotheses, making sure to define any parameters, and use a significance level of $\alpha = 2.5\%$. Comment on your conclusion. Do not use any R functions that are designed to perform hypothesis tests.

(d) Test whether the population proportion of people who exercise either 3 or 6 days each week was greater in March than in February. Clearly state your hypotheses, making sure to define any parameters, and use a significance level of $\alpha = 10\%$. Do not use any R functions that are designed to perform hypothesis tests.

PART 2

The Education Department conducted a study to compare undergraduate student performance at tertiary institutions across five cities in Australia. For each city, a random sample of tertiary students who had completed their undergraduate degree in the previous year was selected and for each student, their average grade across all their courses was recorded. The study was repeated each year from 2001 to 2005, where new random samples of students were selected in each year. The data are stored in the file AssignmentData.RData in the data frame Q2.df. The data frame contains two columns for each year, one for the students' cities and one for their average grades. For example, for the study conducted in 2001, there are two columns named City2001 and AvgGrade2001. For this question you will be analysing data from the study conducted in 2002.

A) Test whether the population mean average grade for students from either Canberra or Sydney is greater than 70. Clearly state your hypotheses and use a significance level of $\alpha = 5\%$. Do not use any R functions that are designed to perform hypothesis tests.

b) Test whether the population mean average grade for students from either Canberra or Sydney is greater than 70. Clearly state your hypotheses and use a significance level of $\alpha = 5\%$. Do not use any R functions that are designed to perform hypothesis tests.

For the remaining parts of this question, assume that the population variance of average grades is unknown for all cities.

c) Test whether the population mean average grade for students from either Canberra or Sydney is greater than 70. Clearly state your hypotheses and use a significance level of $\alpha = 5\%$. Do not use any R functions that are designed to perform hypothesis tests.

d) Test whether the population mean average grade for students from Perth is greater than for students from Melbourne by more than 1.2. Clearly state your hypotheses and use a significance level of $\alpha = 10\%$. Do not use any R functions that are designed to perform hypothesis tests.

You will now conduct a one-way ANOVA on the average grades from the study conducted in 2002 with city as the factor.

e) Test whether the population mean average grade for students is the same for all five cities. Clearly state your hypotheses and use a significance level of $\alpha = 5\%$. Do not use any R functions that are designed to perform hypothesis tests or to perform, analyse or interpret an ANOVA.

f) Discuss whether the assumptions for a one-way ANOVA hold for this data. You do not need to conduct any hypothesis tests, but make sure to provide clear justifications for your answer.

PART 3

People over 50 are generally recommended to regularly monitor their cholesterol levels, as they can increase with age. A long-term study was conducted to investigate the effect of ageing on cholesterol levels. A random sample of 200 people was selected and for each person, their cholesterol level (in mmol/L) was measured each year on their birthday from age 50 to age 55. The data are stored in the file AssignmentData.RData in the data frame Q3.df. The data frame contains columns for the cholesterol levels measured at each age (i.e., Cholesterol50 to Cholesterol55). For this question, you will be analysing the cholesterol levels at ages 50 and 52.

(a) Create a scatter plot of the cholesterol level at age 52 against the cholesterol level at age 50. Make sure to give your plot an appropriate title and appropriate labels for the x and y axes. Describe the relationship between these two variables.

(b) Test whether the population mean cholesterol level at age 52 is greater than the population mean cholesterol level at age 50 by more than 0.44. Clearly state your hypotheses and use a significance level of $\alpha = 0.5\%$. Do not use any R functions that are designed to perform hypothesis tests.

(c) Fit a simple linear regression model with the cholesterol level at age 52 as the dependent variable and the cholesterol level at age 50 as the independent variable. Write down the estimated regression model.

(d) Discuss whether the assumptions for a simple linear regression model hold for the model you fitted in part (c), making sure to provide clear justifications for your answer. (e) [3 marks] For the model you fitted in part (c), test whether the intercept is less than 4.1. Clearly state your hypotheses and use a significance level of $\alpha = 2.5\%$. Do not use any R functions that are designed to perform hypothesis tests.

(f) Using the model you fitted in part (c), calculate a 90% confidence interval for someone's expected cholesterol level at age 52 if their cholesterol level at age 50 was 3.7. Comment on the accuracy of this inference. Do not use any R functions that are designed to calculate any predictions, confidence intervals or predictions intervals.