Probability and Statistics Assignment
(Bioinformatics, MSc)

Instructions

- Write your **NAME** and **REGISTRATION NUMBER** on top of every page.
- There are **FIVE** questions. Attempt all.
- Show all your working clearly and neatly.
- Submit the hardcopy and softcopy of your work. Where R is used, the code will also have to be included.
- Submit your work by 24**th January** 2020 **(beofre** 16 : 31**)**.

Question One (Probability)

A bag contains 100 balls that are identical apart from colour. 50 balls are black, 25 red and 25 blue. Balls are to be drawn from the bag at random and without replacement. Find the probabilities of the following events:

(a) the first ball drawn is black; **[1 mark]**

(b) the first ball drawn is black and the second ball drawn is also black; **[2 marks]**

(c) there are no blue balls among the first three drawn. **[3 marks]**

Question Two (Probability)

A routine screening test is being made available, on which a person tests either positive or negative for a certain disease. 5% of the population has the disease. 90% of people with the disease will test positive. 95% of people who do not have the disease will test negative.

(a) What proportion of the whole population will test positive? **[4 marks]**

(b) What proportion of people who test positive will actually have the disease? **[2 marks]**

(c) Three people from the same village are tested, and all test positive. What is the probability that none of them actually has the disease? Discuss any assumption you have to make in order to calculate this probability. **[4 marks]**

Question Three (R and Hypothesis Testing)

The file *bp.txt*, which is made available to you, contains measurements of the blood pressure and heart rate of a patient over time. It contains the following columns.

hour	Hour at which the measurement was taken $(0-2)$
minute	Minute at which the measurement was taken $(0-59)$
bps	Systolic blood pressure in mmHg
bpd	Diastolic blood pressure in mmHg
hrt	Heart rate in beats per minute

(a) Read the data into R, making sure that you read in missing values correctly. **[3 marks]**

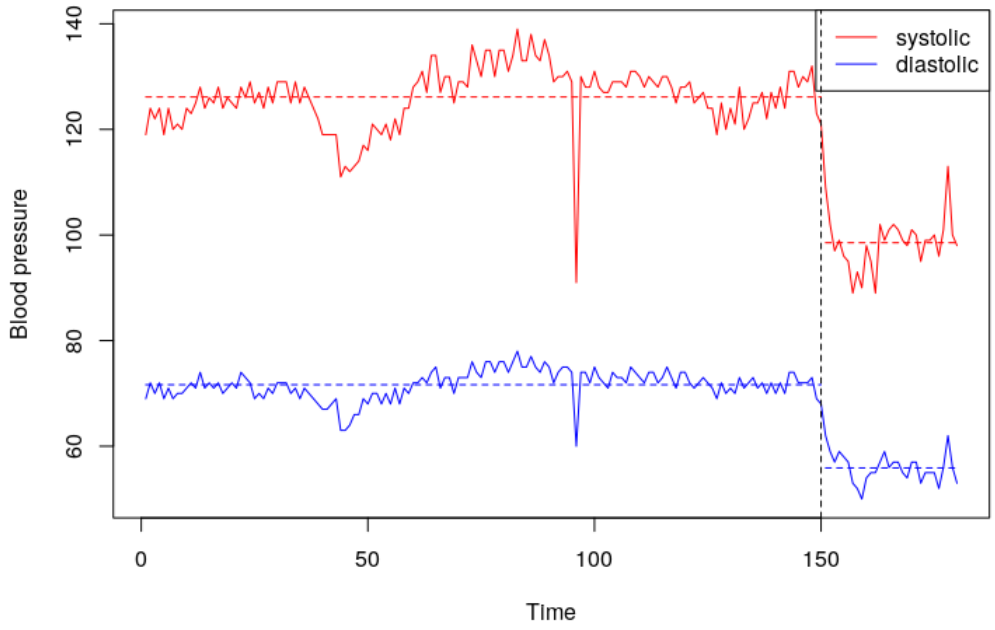
(b) Add a column $time = 60 \times hour + minute + 1$ to the data frame from part (a). **[2 marks]**

(c) Create a plot of the blood pressure measurements against time. Both systolic and diastolic blood pressure should be shown in the same plot, using different colours. The label of the x-axis should be *Time* and the label of the y-axis should be *Blood pressure*. Add a legend to the plot. **[5 marks]**

(d) It is believed that the patient became unstable 150 minutes after measurements had begun (i.e. $time > 150$). Compute the average systolic and diastolic blood pressure separately for the first 150 minutes and the remaining time. **[2 marks]**

(e) Add a vertical dashed line at $time = 150$. **[1 mark]**

(f) Add dotted horizontal lines to the plot indicating the average systolic and diastolic blood pressure for the first 150 minutes and the remaining time. The horizontal lines should not cross the vertical line at 150. **[3 marks]**
Your final plot should look similar to the plot below.



(g) Write a function *twosample.t.test* which takes two vectors x and y as arguments and which computes the test statistic t of the two-sample t-test using the following formula:

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{xy}^2}}$$

where n_1 is the number of observations $x_1, x_2, x_3, ..., x_{n_1}$; n_2 is the number of observations $y_1, y_2, y_3, ..., y_{n_2}$, and

$$S_{xy}^2 = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (x_i - \overline{x})^2 + \sum_{i=1}^{n_2} (y_i - \overline{y})^2 \right)$$

with $\overline{x} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$ and $\overline{y} = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i$. **[5 marks]**

(h) Using the function *twosample.t.test*, compute the test statistic for the two-sample t-test of the null hypothesis that the average systolic blood pressure has changed after $time = 150$. (Test this hypothesis using $\alpha = 0.05$.)

Denote by $x_1, x_2, x_3, ..., x_{150}$ the systolic blood pressure measurements at times 1 to 150 and denote by $y_1, y_2, y_3, ..., y_{30}$ the systolic blood pressure measurements at times 151 to 180.

Hint: You can check your result using the built-in function t.test using the option var.equal = TRUE. **[2 marks]**

Question Four (LM and GLM)

You are given the following data:

$$x = (-6, -6, -4, -1, 0.5, 2, 8, 8, 11, 11.5)^T$$
$$y = (-3.7, -4.3, -3.9, -4.6, 0.5, -6.9, 10.2, 16.1, 6, 19.5)^T$$

(a) Fit a linear regression model to these data and show the model output. **[2 marks]**

(b) Describe the resulting regression line:

i. What is the relationship between variables X and Y? **[2 marks]**

ii. How much (on average) does Y change when X changes by 1? **[1 mark]**

iii. What value does Y take (on average) when $X = 0$? **[1 mark]**

(c) Compute the coefficient of determination R^2, the adjusted R^2, the likelihood and the AIC. Which of these tell you how good your model fits the data? **[4 marks]**

(d) Compute the residuals $r_i = y_i - \hat{y}_i$ and do a normal distribution QQ plot. **[4 marks]**

(e) What other diagnostic check(s) could you do? Do this and explain whether you think this is a good model. **[2 marks]**

(f) Re-fit the model, but now including a term for X^2: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$. Check and discuss the resulting model and compare it to the previous one. Which model would you recommend for this dataset? **[4 marks]**

Question Five (LM and GLM)

Download (from GitHub) and load the dataset *cuse.csv*.

This is a dataset on contraceptive use. *using*, *notUsing* lists how many people in each group implied by combinations of *age*, *education*, *wantsMore* are currently using contraceptives. *age*, *education* are self-explanatory. *wantsMore* lists whether individuals want more children or not.

(a) Model the binary variable specified by the 2 columns *using*, *notUsing* in terms of *age*, *education*, *wantsMore*. **[5 marks]**

(b) Discuss your results. **[4 marks]**

(c) What can you say about the deviance? Does it look like this is a good model? **[3 marks]**

(d) What happens if you include an interaction term between the age variable and the desire for more children variable? **[4 marks]**