EMAT30007 Applied Statistics 2022/23

Coursework

Nikolai Bode

(nikolai.bode@bristol.ac.uk)

Due: 26th April 2023, 01:00pm (week 23)

Submit on BLACKBOARD

General information:

Attempt to answer all questions. This coursework has three parts with one question each. The questions are released successively as the material is covered in the course.

Submit a Matlab script (.m file) file with your answers. This should run when copied into the same folder as the data files (see below) and should only use commands and Matlab packages used in the worksheets. Clearly annotate your code and include the required discussion of your findings directly in the script.

The limit for your submission is 900 lines of standard Matlab script with at most 100 characters per line, in addition to a restriction on the number of figures/plots for each question (see below). Just for indication: the model solution is 600 lines long with a lot of empty lines.

There are additional files available on Blackboard for this piece of coursework that contain data to be used in the questions. The contents of the files are described in more detail below.

The only way I will answer questions about the coursework is via the dedicated coursework discussion forum on Blackboard. This is to ensure that the entire class has access to the same information.

There are a total of 60 marks for the coursework, 20 marks for each part of the coursework.

PART I – (≤300 lines of code)

Question 1 (20 marks):

A research team in a pharmaceutical company wants to test the efficacy of a new drug. The drug is very expensive and preliminary tests have shown that due to an unknown physiological process, the active substance of the drug is not delivered to the bloodstream of all people. Fortunately, there is a separate, much cheaper screening test to determine with high accuracy whether the active substance of the drug will appear in the blood stream of individuals. The screening test produces a score, and higher scores correlate with a higher chance of the active substance appearing in the blood stream. The research team devise the following study. First, 100 participants are recruited through adverts and screened using the cheap test. Second, the 40 participants with the highest scores from the screening test are randomly divided into two groups of 20 participants each. One randomly chosen group receives a placebo, the other the real drug. The researchers and participants do not know during the trial who gets the placebo and who gets the drug. The trial concludes with the measurement of an appropriate physiological variable in the 40 participants. The research team hopes that the drug will increase the value of this variable and conduct an appropriate statistical test to assess this. They find that in the group receiving the drug, the variable is 20% higher on average, and they claim, based on their statistical test, that this finding is statistically significant. They write a report in which they recommend the introduction of the drug to market.

Due to a very expensive administrative error, the drug is actually also administered to half of the 60 participants not included in the second part of the study, with the other 30 participants receiving a placebo. This data is found years later by a student doing an internship at the company.

The files `drugs20.txt` and `placebo20.txt` contain the final measurements for the groups of 20 participants who received the drug or the placebo, respectively. The files `drugs30.txt` and `placebo30.txt` contain the data the intern finds.

(a) Conduct the statistical test the scientists performed and conduct the same test on the data the intern has found. State what you find.

(b) Extending the material covered in weeks 13-16, devise and briefly justify a method to approximate a 95% confidence interval for the percentage difference in the variable measured between the drug and the placebo group under the assumption that instead of the initial screening, 40 participants were selected at random from the initial participant pool for the second part of the study. State what you find and discuss why this can only be an approximation.

(c) Fit an appropriate probability distribution to the data for all 100 participants using maximum likelihood estimation. Note that a standard probability distribution may not be appropriate, so you may have to define your own. Show qualitatively that the fit of the probability density function of your chosen distribution to the data is good (1 figure).

(d) Write a statistical critique of the study described above, referring to your answers in questions 1(a)-(c). Provide your critique in bullet points, not continuous text. You should discuss positives, negatives, and possible improvements for the study.

PART II – (≤300 lines of code)

Question 2 (20 marks):

A random walk is a spatial random process that can be described as $x(t) = x(t - \delta t) + \vartheta$, where $x(t)$ denotes the position of the walk at time t and this position is updated in time increments of δt seconds. ϑ is drawn from an appropriate probability distribution and describes the step the random walk makes within one time increment.

The files `walk_data1.txt` and `walk_data2.txt` contain the movement paths for two different types of one-dimensional random walks. One-dimensional random walks describe movement on a line. The time increment is set to $\delta t = 1$ second.

(a) Plot the movement paths of the two random walks over time (1 figure).

(b) Fit Linear models that correctly capture the movement dynamics to the data. Justify that your models are appropriate (2 figures).

(c) Explain what the model fits tell you about the dynamics of the two random walks.

(d) Consider a random walk with $\vartheta \sim N(\mu, \sigma)$, where $\sigma = 0.1$ m and

$$\mu = 0 \text{ m} \qquad \text{if} \qquad |x(t - \delta t)| < 1$$

$$\mu = -0.05 * x(t - \delta t) \text{ m} \qquad \text{if} \qquad |x(t - \delta t)| \geq 1$$

<u>Simulate*</u> 10,000 time increments ($\delta t = 1$ s) for this random walk, plot the movement path against time and fit the appropriate Linear model to your simulated data (1 figure). Note that $|x|$ denotes the absolute value of x here.

(e) Formulate a random walk that can be captured by a Linear model but violates one of the assumptions of these models and demonstrate this appropriately (1 figure).

(f) Discuss, whether/how data from random walks in two dimensions (movement on the plane) could be explained using statistical models covered in the course.

*<u>Simulations</u> in this context are computer-generated movement paths.

PART III – (≤300 lines of code)

Question 3 (20 marks):

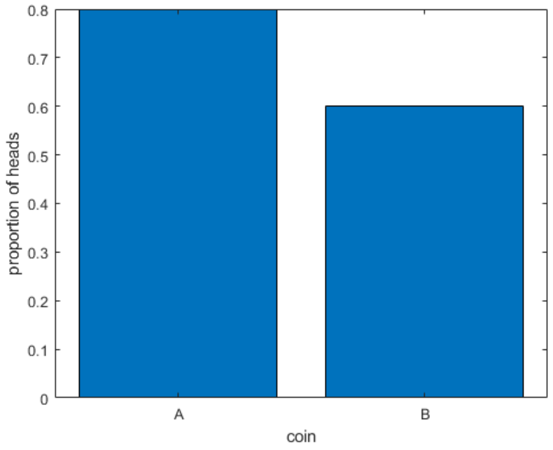
Consider the following three examples of reporting on statistical analysis.

(i) A linear model was fit to data, found to be appropriate, and gave the following results.

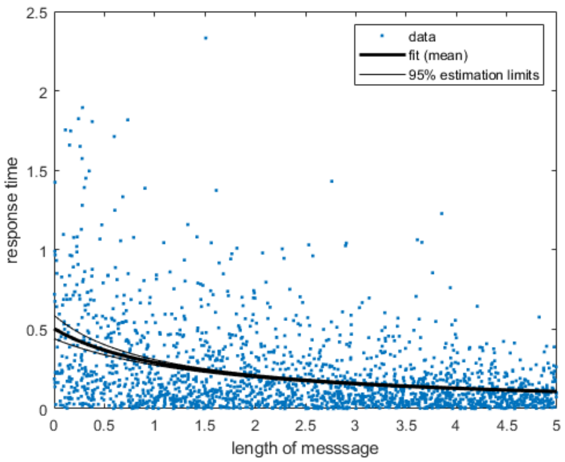
	Estimate	**p-value**
Intercept	8x10^{-5}	p>0.05
x	0.2	p<0.05
y	0.5	p<0.05
Interaction term	0.1	p<0.05

Root Mean Squared Error: 0.1

(ii) Experimental data on two biased coins is compared using a two sample Z-test for proportions. The proportion of heads for coin A in n_A=100 coin flips was 0.8 and for coin B in n_B=50 flips it was 0.6 (p<0.01).



(iii) A statistical analysis looked at the exponentially distributed waiting times between responses to claims sent out by an insurance company depending on the length of the text messages accompanying the claims. An appropriate statistical model was fit to the data (n=2,000), producing the plot below. It is claimed that the plot can be used for predictions.



For each of the three examples, simulate a toy data set*, recreate the analysis, and use this to critically discuss the reporting of statistical results above, covering good practice, shortcomings, and any improvements (which you should implement). You are allowed one figure for each example in your answer. Give your answer in bullet points.

*A toy data set is intended to illustrate something. When creating it you can choose numbers and parameters ***that are not given in the question (directly or indirectly)*** in such a way that the demonstration is clear. For example, in worksheet 7, we created toy data sets to illustrate how different types of predictors work in Linear Models.