STAT0032: INTRODUCTION TO STATISTICAL DATA SCIENCE 2019-2020

- Answer ALL questions.
- You may submit only one answer to each question.
- The exam has a total of 40 marks with relative weights attached to each question are Question 1 (20 marks) and Question 2 (20 marks).
- The 20 marks for each question are assigned according to the following mark scheme:
  - 8 marks for content, including detailed justification of the methods used, accurate presentation of these methods and discussion of advantages and disadvantages.
  - 4 marks for clear and detailed derivation of mathematical equations, and correct calculations using the data.
  - 8 marks for presentation, including structure, clarity of writing and lack of typographical errors.
- There are no uniquely "correct" answers to these questions. The questions aim to test your understanding of the concepts encountered in the course. They also aim to test your ability to use these concepts in a coherent way to solve an applied problem.
- Marks are awarded not only for the final result but also for the clarity of your answer. As a result, after your initial attempt at answering the questions, you might want to consider rewriting the answers in the clearest way possible.
- UCL requires that all 24-hour online exams have a specified overall word limit. The overall word limit for this exam has been set well in excess of the expected amount of work so that you do not need to worry about exceeding it. Therefore, we expect that solutions to the paper will be much shorter than the specified word limit. However, each question has a word or page limit. You must adhere to this or risk losing marks.

Administrative details

- This is an open-book exam. You may use your course materials to answer questions.
- This is a 24-hour exam, but you should expect to complete it in about two hours.

- You may not use any computer software. The only exception is to obtain quantiles of probability distributions.
- **You may not contact the course lecturer with any questions**, even if you want to clarify something or report an error on the paper. If you have any doubts about a question, make a note in your answer explaining the assumptions that you are making in answering it.

Formatting your solutions for submission

- The exam consists of two questions. For *each* question, your answer should consist of:
  1. Up to 500 words of *typed* text.
  2. Up to one A4 page of hand-written (or typed) mathematical derivations.
- The page of mathematical derivations should be scanned, then separated into sub-pictures which are inserted on the relevant part of the typed text (for example, using the Office Lens software).
- No text is allowed on the page of mathematical derivations and equations are not allowed in the typed text. **Failure to follow these rules may result in marks being deducted.**
- You should submit ONE document that contains your solutions for all questions/ part-questions. Please follow UCL's guidance on combining text and photographed/ scanned work.
- Make sure that your handwritten solutions are clear and are readable in the document you submit. You are encouraged to write out solutions neatly once you are happy with them.

Plagiarism and collusion

- You must work alone. In particular, *any discussion of the paper with anyone else is not acceptable*. You are encouraged to read the Department of Statistical Science's advice on collusion and plagiarism, which you can find here.
- Parts of your submission will be screened via Turnitin to check for plagiarism and collusion.
- If there is any doubt as to whether the solutions you submit are entirely your own work you may be required to participate in an investigatory viva to establish authorship.

CONTINUED

Question 1

Your neighbour and colleague Jon is currently commuting to work on foot, but he is wondering whether he should buy a bike to reduce his commute time. Jon's main concern about using a bike is safety. He decides to use statistical modelling to inform his decision. Jon obtains a dataset consisting of 332 individuals commuting by bike in the UK. For each individual, the following variables were recorded over the last year:

- *bikes*: Average number of other cyclists encountered by the individual per journey to work over the last year.
- *height*: Height of the individual at the start of the year (in cm)
- *cars*: Average number of cars encountered by the individual per journey to work over the last year.
- *weight*: Weight of the individual at the start of the year (in kg).
- *age*: Age of the individual at the start of the year.
- *number-of-accidents*: Number of bike accidents suffered by the individual in one year.
- *accident*: True or False (or equivalently 1 or 2 respectively). Indicates whether the individual had a bike accident over the last year.

Jon considers the use of logistic regression to model the variable *accident*. The first step of his analysis is to summarise the data which will be used for this model:

```
> summary(dataset)
     bikes            height           cars            weight           age        accident
 Min.   : 4.795   Min.   :132.5   Min.   : 38.80   Min.   : 43.40   Min.   :21.00   No :223
 1st Qu.: 5.987   1st Qu.:148.0   1st Qu.: 56.35   1st Qu.: 50.64   1st Qu.:23.00   Yes:109
 Median : 7.137   Median :156.0   Median : 65.80   Median : 57.60   Median :27.00
 Mean   : 8.480   Mean   :159.6   Mean   : 66.48   Mean   : 61.14   Mean   :31.32
 3rd Qu.:10.064   3rd Qu.:168.1   3rd Qu.: 74.40   3rd Qu.: 67.17   3rd Qu.:37.00
 Max.   :22.084   Max.   :198.5   Max.   :134.20   Max.   :136.80   Max.   :81.00
```

Following this initial exploration, he obtains the following fit:

```
Call:
glm(formula = accident ~ ., family = binomial(link = "logit"),
    data = dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9192  -0.6362  -0.3693   0.6158   2.5532

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -19.11043    2.14705  -8.901  < 2e-16 ***
bikes         0.14448    0.05911   2.445  0.01451 *
height        0.07433    0.01099   6.763 1.35e-11 ***
cars          0.04248    0.01099   3.867  0.00011 ***
weight        0.02823    0.01114   2.533  0.01130 *
age           0.01493    0.01750   0.853  0.39370
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 420.30  on 331  degrees of freedom
Residual deviance: 286.78  on 326  degrees of freedom
AIC: 298.78

Number of Fisher Scoring iterations: 5
```

Unfortunately, Jon does not know much about statistical modelling, and as a result does not know how to interpret these results. However, he knows that you have taken a MSc-level class on “Introduction to Statistical Data Science” at UCL, and asks for your help.

Prepare a report of up to 500 words (supplemented with up to one A4 page of calculations if required) describing the results of the analysis performed by your neighbour, and making suggestions as to how to refine this analysis. Your report should include the following:

(i) A description of the fitted model, how it fits within the generalised linear models framework, its latent variable representation, and an interpretation of the coefficients.

(ii) Under this model, give *both* the estimated probability and odds that Jon will have an accident in the next year. You may use the fact that Jon is 32 years old, 1.78m and 100kg. You may also use the fact that the average number of cars encountered on your daily commute is 169.4, whilst the average number of bikes encountered is 49.7. Since Jon is both your neighbour and colleague, his journey to work is identical to yours.

(iii) A discussion of the assumptions Jon would need to make in order to use this model, and for the prediction in (ii) to be reliable. This should lead to the conclusions of your analysis, and a discussion of whether this analysis should be used to decide whether Jon should buy a bike.

(iv) Suggestions as to how Jon could refine his current model. This should include further data collection, alternative models and a discussion of automatic model choice. In each case, you may want to discuss advantages and disadvantages.

Sketch Solution:

CONTINUED

- Denote by o the odds of having an accident, and p the probability of having an accident. The fitted model is a logistic regression model, which can be written in latent variable form as follows:

$$\begin{aligned} \textit{accident} &= \begin{cases} 1 \text{ (or True)}, & \text{if } z \geq 0 \\ 0 \text{ (or False)}, & \text{if } z \leq 0 \end{cases} \\ z &= \beta_0 + \beta_1 \text{bikes} + \beta_2 \text{height} + \beta_3 \text{cars} + \beta_4 \text{weight} + \beta_5 \text{age} + \epsilon \end{aligned}$$

  This is a model with 5 explanatory variables (bikes, height, cars, weight and age) and one response variable. The noise ϵ is IID realisations from a logistic distribution.

  This corresponds to a generalised linear model with Bernoulli distribution and a logistic link function. In particular, it can be expressed as:

$$\begin{aligned} \mathbb{P}[\textit{accident} = y] &= p^y (1-p)^y \quad (y \in \{0, 1\}) \\ p &= g^{-1}(\beta_0 + \beta_1 \text{bikes} + \beta_2 \text{height} + \beta_3 \text{cars} + \beta_4 \text{weight} + \beta_5 \text{age}) \\ g^{-1}(x) &= \frac{1}{1 + \exp(-x)} \end{aligned}$$

  It is possible to obtain p from o and vice-versa using the following expressions:

$$\frac{p}{1-p} = o \qquad p = \frac{1}{1 + \exp(-o)}$$

  The coefficients where fitted by maximum likelihood. This is usually done numerically, as highlighted by the "Number of Fisher Scoring Iterations". The parameter estimates obtained were $\beta_0 = -19.11043, \beta_1 = 0.14448, \beta_2 = 0.07433, \beta_3 = 0.04248, \beta_4 = 0.02823, \beta_5 = 0.01493$.

  The easiest way to interpret these parameters is in terms of how they impact the expected log-odds. In particular, note that:

$$\log o = \beta_0 + \beta_1 \text{bikes} + \beta_2 \text{height} + \beta_3 \text{cars} + \beta_4 \text{weight} + \beta_5 \text{age}$$

  Therefore, an increase of 1 bike (with all other variables staying constant) leads to an increase of the log-odds by $\beta_1 = 0.14448$, or equivalently by an increase of the odds by $\exp(\beta_1) = \exp(0.14448) \approx 1.155$. Similar interpretations can be obtained for all other coefficients.

- To make a prediction for Jon, we can simply plug-in the data provided in the model. For example, the log-odds become:

$$\begin{aligned} \log o &= \beta_0 + \beta_1 \text{bikes} + \beta_2 \text{height} + \beta_3 \text{cars} + \beta_4 \text{weight} + \beta_5 \text{age} \\ &= -19.11043 + 0.14448 * 49.7 + 0.07433 * 178 + 0.04248 * 169.4 \\ &\quad + 0.02823 * 100 + 0.01493 * 32 \\ &\approx 11.79 \end{aligned}$$

  which implies odds of $o = \exp(11.79) \approx 131926.47$. This implies a probability of

$$p = \frac{1}{1 + \exp(-o)} \approx 1$$

- There are a large number of assumptions needed. These include:

  1. The noise is IID and follows a logistic distribution. This may not be a realistic assumption for a large number of reasons.
  2. The data for the UK should be representative. This is very unlikely to hold since we do not know where Jon lives, and there would for example be a very large difference between cities and more rural areas which may not be well explained by our variables.
  3. We need to assume that the numerical method has converged to the global maximum of the likelihood. Otherwise it is very difficult to know the impact on predictions.

  As a result of all these assumptions which are likely to be violated, it is unclear whether this would actually be a useful result. One additional argument against is that the prediction in (ii) actually corresponds to a data point far outside of the range of data points used for fitting the model. Take for example the variable bikes: the value of 49.7 is outside of the range of data points (the largest value is 22.084). Similarly, the variable cars takes a value of 169.4 (the largest value in the training data is 134.4).

  There are also many other concerns which might come into account. For example, we have only collected data about bike accidents, but we do not know anything about whether walking to work is safe.

- There are many acceptable answers for how to refine the analysis. This could include:

  – The most important improvement to the model would be to collect more data about journeys which are similar to those Jon will be taking in terms of number of cars and bikes encountered.

  – Taking the response variable to be *number-of-accidents* rather than *accidents*, in order to get a more detailed picture of each case. This would require using an alternative generalised linear model, for example based on the Poisson or Negative-binomial distributions. There is then the question of whether these are reasonable models for this data.

  – It might be interesting to consider interactions of some of these variables or transformations. For example, interactions weight-height or car-bike might give a clearer picture of the problem. One might think that the level of danger for a journey with a lot of bikes and cars is very different to that of a journey with only a lot of bikes. Similarly, someone which is both very heavy and tall is usually physiologically different to someone which is very heavy but short.

  – Automatic model choice could be done using shrinkage penalties (e.g. ridge or lasso). The advantage of ridge is that it can reduce the number of variables under consideration, but it leads to non-convex optimisation. On the other hand ridge regression is a convex problem but does not reduce the overall number of variables. It does how reduce the impact of certain variables.

CONTINUED

Question 2

Suppose you are about to have an important job interview which will take place via video-conference. Your flat is currently subscribing to an internet provider for a package including 15Mbps upload speed. However, your flatmates are avid users of the internet and regularly watch movies or have to make video calls for their work. This has meant that your upload speed is usually significantly lower than it should be, and you are worried about having technical difficulties during your interview. Your strong suspicion is that the average upload speed μ is only $\mu_0 = 8.5$ Mbps, and you know that the video-conference software would not work with a speed less than or equal to $\mu_- = 8.4$ Mbps.
You decide to test the upload speed during the week prior to your interview. To do so, you use a website which claims to give measurements that are normally distributed with a mean equal to the actual upload speed, and with a standard deviation of $\sigma = 0.05$. Suppose $n = 10$ independent measurements yielded the following upload speeds (all expressed in Mbps):

$$x_1 = 8.30, x_2 = 8.21, x_3 = 8.54, x_4 = 8.32, x_5 = 8.43,$$
$$x_6 = 8.48, x_7 = 8.42, x_8 = 8.46, x_9 = 8.57, x_{10} = 8.42$$

Prepare a report of up to 500 words (supplemented with up to one A4 page of calculations if required) describing how you could use the framework of hypothesis testing to decide whether you should be worried about the internet speed for your interview. Your report should include the following:

(i) A hypothesis test for your suspected value of the mean upload speed μ_0 at 5% significance level. You should clearly define the hypothesis being tested, the test statistic, the conclusion of the test and how it was obtained.

(ii) A hypothesis test to check whether your mean upload speed is below μ_- at 5% significance level. You should clearly define the hypothesis being tested, the test statistic, the conclusion of the test and how it was obtained.

(iii) A clear statement of any assumption underlying (i) and (ii), and a discussion of whether such assumptions are likely to hold in practice.

(iv) A discussion of the implications of your analysis for your upcoming interview, and a discussion of further analysis you could consider to gain more insight into this problem.

Sketch Solution: A good solution to the problem would likely include the following content:

- A hypothesis test for the mean upload speed which makes use of the Gaussianity of the observations. A possible solution is highlighted below.
  - The null hypothesis is $H_0 : \mu = \mu_0$ and the alternative hypothesis is $H_1 : \mu \neq \mu_0$, where $\mu_0 = 8.5$.
  - The level of the test is $\alpha = 0.05$, and we consider a two-sided test.

– Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$. A reasonable statistic would be

$$Z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}$$

The test statistic follows a $\mathcal{N}(0,1)$ under H_0.

– We therefore expect the value of the test statistic to fall within the following interval with probability $(1-\alpha/2)$ under H_0:

$$Z \in \left[-z_{\alpha/2}, z_{\alpha/2}\right]$$

where $z_{\alpha/2} =$ is the $(1-\alpha/2)\%$ quantile of a standard normal distribution. Plugging in the value of $z_{\alpha/2} = 1.96$ when $\alpha = 0.05$, we get:

$$Z \in [-1.96, 1.96]$$

Equivalently, we would expect the sample mean to be within the following interval with probability $(1-\alpha/2)$ under H_0:

$$\bar{X} \in \left[\mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

which simplifies to

$$\bar{X} \in \left[\mu_0 - 1.96\frac{0.05}{\sqrt{10}}, \mu_0 + 1.96\frac{0.05}{\sqrt{10}}\right] = [\mu_0 - 0.031, \mu_0 + 0.031]$$

when $\alpha = 0.05$.

– We now calculate the actual values of these quantities:

$$\bar{X} = 8.415$$

$$Z = \sqrt{10}\frac{(8.415 - 8.5)}{0.05} \approx 3.162\frac{(8.415 - 8.5)}{0.05} \approx -5.38$$

Since the value of the test statistic Z does not fall within the interval, we reject the null hypothesis (or equivalently, since $\bar{X}$ does not fall within the interval).

– Of course, one could also do a t-test, which would use the same test statistic but with a different distribution under H_0. This is only justified if we have reason to believe that the σ given is unreliable. Otherwise, we are using a worse test and discarding useful information. In this case, the sample standard deviation is:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2} \approx \frac{1}{3}\sqrt{\sum_{i=1}^{n}(x_i - 8.415)^2}$$
$$= \frac{\sqrt{0.11}}{3} \approx 0.111$$

This sample standard deviation is significantly larger than the one advertised by the website, so it is reasonable to question the accuracy of the standard deviation advertised.

CONTINUED

- A hypothesis test to check whether the mean upload speed is below μ_-. A possible solution is highlighted below.

  - $H_0 : \mu \leq \mu_-$, $H_1 : \mu > \mu_-$, where $\mu_- = 8.4$
  - This time, it makes more sense to do a one-sided test due to the asymmetry of the problem. Indeed, we would only want to reject the null hypothesis when the sample size is much larger than μ_- (and not when it is much lower). We do so again at level $\alpha = 0.05$.
  - We may keep the same test statistic Z, which once again follows a standard Gaussian distribution. This time, under H_0, we would expect to find our statistic within the following interval with probability $(1 - \alpha)$:

    $$Z \in (-\infty, z_\alpha]$$

    where $z_\alpha =$ is the $(1 - \alpha)\%$ quantile of a standard normal distribution. Equivalently, we would expect:

    $$\bar{X} \in \left(-\infty, \mu_- + \frac{\sigma}{\sqrt{n}} z_\alpha\right]$$

  - We can now compute the test statistic. The sample mean is unchanged and takes value $\bar{X} = 8.415$, but the test statistic now takes value (since we change the mean being tested):

    $$Z = \sqrt{10}\frac{(8.415 - 8.5)}{0.05} \approx 3.162\frac{(8.415 - 8.4)}{0.05} \approx 0.949$$

  - Plugging in the actual values from the data, we get that $z_\alpha = 1.645$ so that we would expect (under H_0 that):

    $$Z \in (-\infty, 1.645]$$

    $$\bar{X} \in \left(-\infty, 8.4 + \frac{0.05}{\sqrt{10}} 1.645\right] \approx (-\infty, 8.426]$$

  - Since our test statistic is within this interval, we do not have enough evidence to reject the null hypothesis.

- A few remarks on the assumptions underlying the tests. For example:

  1. The tests rely on the assumption of Gaussianity of the measurements. If this assumption is false, then the tests will not be reliable.
  2. The measurements are likely not be Gaussian since we would never expect negative measurements of upload speed. However, the standard deviation is very small, so this choice will place very little probability on negative numbers, and the approximations is hence justified in practice.

3. The tests assume that the measurements are IID. This may not be true in practice since we are measuring at different times, and it is likely that the upload speed depends on the time of the day or day of the week at which the measurement was made.
4. A comment on the value of σ, as discussed above.

- An interpretation of the results of these tests which is stated carefully.

  For example, for the first test, a conclusion along the lines of "Our test implies the mean upload speed is not 8.5Mbps" is not an acceptable conclusion from this test. A preferable conclusion would be "Our test implies that the mean upload speed is unlikely to be 8.5Mbps at our significance level. However, we cannot discard this completely."

  The overall conclusion should be that we should be worried about the quality of the internet connection. Suggestions could include collecting more data, to reduce our uncertainty about the mean upload speed.

  Alternatively, one could collect new data at times where the flatmates are not using the internet. If the conclusion of the second test is to reject the null hypothesis, this could suggest a possible solution to the problem of the mean upload speed.