STAT0032: INTRODUCTION TO STATISTICAL DATA SCIENCE - EXAM 2020-21

- Answer ALL questions.
- You have three hours to complete this paper. After the three hours has elapsed, you have one additional hour to upload your solutions.
- You may submit only one answer to each question.
- The relative weights attached to each question are: Question 1 [30 marks], Question 2 [30 marks].
- Marks are awarded not only for the final result but also for the clarity of your answer.

Administrative details

- This is an open-book exam. You may use your course materials to answer questions.
- You may not use any computer software.
- **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.
- The overall word limit for this exam has been set at 1500 words.
- Some part-questions require text-based answers; many of these will indicate the maximum number of sentences you may write. You must adhere to this or you will lose marks.
- Some questions may ask you to approach a problem in a particular way; please take note of this. Failure to do so may result in marks being deducted.

Formatting your solutions for submission

- 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.
- The exam consists of two questions. For each question, your answer should consist of maximum: (i) 750 words of typed text, and (ii) two A4 pages of mathematical derivations. Note that these are not target numbers to be reached: it is possible to obtain full marks with less and full marks will only be given to clear and concise answers.

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.
- 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.

Question 1

Suppose that you are big fan of sailing and are very excited about a forthcoming race. Before the event starts, you would like to use your newly developed skills in statistical data science to predict which team will win the race. To do so, you have acquired a dataset containing details of all international tournaments for the last ten years ($n = 734$). For each boat and race, you have access to a number of variables including the weight of the boat in kg (denoted x_1), the size of its sail in m^2 (denoted x_2), the length of the race in km (denoted x_3) and the amount of time it took to complete the race in hours (denoted y).

You decide to create a linear regression model for y in terms of the other variables. However, you are not certain that all of these explanatory variables are relevant and would like to automate the process with the automatic model selection tools you have learnt about in STAT0032.

Answer the following sub-questions describing your analysis of this problem through penalised regression and automatic model selection tools.

(i) A description (with equations) of the Gaussian linear regression model with all three explanatory variables, including details of all assumptions required. This should also describe (again with equations) how to fit the model using least-squares. [4 marks]

(ii) A description (with equations) of how to fit that same model through the LASSO, and a discussion in your own words (in maximum two sentences) of the main advantage of this approach compared to least-squares. Your discussion should mention how the choice of penalisation parameter impacts the resulting fit (again, in maximum two sentences). In particular, you should comment on the behaviour in the limiting settings where this penalisation parameter tends to zero or to infinity. [8 marks]

(iii) A discussion in your own words (and in maximum three sentences) of whether you think LASSO regression will be useful for modelling the number of hours required to complete the race for a given boat, and of how you would go about setting the penalisation parameter. [3 marks]

(iv) A detailed description (in the form of an algorithm, and using your own words) of how to perform forward-stepwise selection for linear regression with the Akaike Information Criterion (AIC). You should then also discuss how this differs from best subset selection and backward-stepwise selection. [6 marks]

(v) A derivation of the best model according to best subset selection, forward-stepwise regression and backward-stepwise regression based on Akaike's Information Criterion (AIC). To do so, you should use the table below and clearly justify your answer. [6 marks]

Model	Explanatory Variables	AIC	RSS
1	none (i.e. only an intercept)	1596.21	247.21
2	x_1	1564.67	213.69
3	x_2	1556.15	205.15
4	x_3	1533.83	182.83
5	x_1, x_2	1524.36	171.36
6	x_2, x_3	1524.78	171.78
7	x_1, x_3	1531.19	178.18
8	x_1, x_2, x_3	1534.85	179.84

Table 1: AIC for linear regression models of the number of hours in took to complete a race. Each row corresponds to a model indexed by a number given in the first column. Each model contains an intercept as well as a number of explanatory variables, given in the second column. For each model, the corresponding AIC and residual sum of squares (RSS) values are given in the third and fourth column respectively.

(vi) A discussion of the algorithm you would use for model selection with the sailing dataset (out of best-subset, forward-stepwise or backward stepwise selection), and a brief discussion (in maximum 2 sentences) of whether the Bayesian Information Criterion (BIC) should be preferred over AIC for this problem. [3 marks]

Sketch Solution:

(i) [bookwork] A Gaussian linear regression model can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i \qquad \forall i = 1, \ldots, n.$$

Here y_i is the value of the dependent variable for the i^{th} data point, and x_{i1}, x_{i2}, x_{i3} are the corresponding explanatory variables. The parameters $\beta_0, \ldots, \beta_3$ are coefficients taking values in $\mathbb{R}$. $\{\epsilon_i\}_{i=1}^n$ are iid unobserved noise terms following a Gaussian distribution with mean zero.

To fit this model (i.e. find the optimal parameter values) using least squares, one needs to solve the following problem:

$$\arg\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} - y_i)^2$$

Sketch Solution:

(ii) [bookwork/seen] To fit this model with the LASSO, one needs to solve the following problem:

$$\arg\min_{\beta_0,\dots,\beta_p} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} - y_i)^2 + \lambda(|\beta_0| + |\beta_1| + |\beta_2| + |\beta_3|)$$

where $\lambda > 0$ is a regularisation/penalisation parameter which has to be set. The main advantage of the LASSO over least squares is that it leads to sparsity: if λ is large enough, some of the parameter values will be set to zero, with only a small number of non-zero entries remaining. When λ goes to zero, we recover the least-squares problem described in (i). When λ goes to infinity, the optimal parameters will all be zero.

Sketch Solution:

(iii) [unseen] The main advantage of the LASSO is that the penalty shrinks the parameters to zero and encourages sparsity. In this particular situation, we already have a fairly small number of parameters, so it is unclear why sparsity would be useful. As a result, it would be reasonable to make λ small, or even completely remove the penalisation. (Note: Any answer will be accepted as long as it is properly justified)

Sketch Solution:

(iv) [bookwork/seen] Forward-stepwise selection works as follows. Let $\mathcal{M}_0$ denote the model with an intercept only (i.e. model 1). Then, for $k = 0, 1, 2, 3$, define $\mathcal{M}_k$ to be the the model with the smallest residual sum of squares (RSS) containing k predictors, including the predictors from $\mathcal{M}_{k-1}$. Finally, return the model with the smallest AIC out of $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$.

For backward-stepwise selection, the procedure works in reverse, starting with $\mathcal{M}_3$ and removing one variable at a time, then selecting the best model in terms of AIC. Finally, best subset selection is similar to both procedure, but at each iteration searches over all possible models (rather than models with +1 or -1 variable compared to the previous iteration).

Sketch Solution:

(v) [unseen] *Forward stepwise selection* would pick

- Model 4 (x_3) as the best one-variable model (since it has the lowest RSS out of all one variable models), then
- Model 6 (x_2, x_3) as the best two-variable model (since it has the lowest RSS out of all two-variable models containing x_3).
- *Model 6* (x_2, x_3) has the lowest AIC out of Model 1 (best zero variable model), 4, 6 and 8 (best three variable model), so this will be the model returned by the algorithm.

Backward stepwise selection would pick

- Model 5 (x_1, x_2) as the best two-variable model (since it has the lowest RSS out of all two-variable models), then
- Model 3 (x_2) as the best one-variable model (since it has the lowest RSS out of all one-variable models containing x_1 or x_2).
- *Model 5* (x_1, x_2) has the lowest AIC out of Model 1 (best zero variable model), 3, 5 and 8 (best three variable model), so this will be the model returned by the algorithm.

Best subset selection would pick

- Model 4 (x_3) as the best one-variable model (since it has the lowest RSS out of all one variable models), then
- Model 5 (x_1, x_2) as the best two-variable model (since it has the lowest RSS out of all two-variable models).
- *Model 5* (x_1, x_2) has the lowest AIC out of Model 1 (best zero variable model), 4, 5 and 8 (best three variable model), so this will be the model returned by the algorithm.

Sketch Solution:

(vi) [unseen] Here, the number of possible models is fairly small since the number of explanatory variables is only 3. As a result, there is no reason to prefer backward stepwise or forward stepwise over best subset selection. If we had more variables, or if n was so large that each model was prohibitively expensive to fit, then it would be best to use either backward stepwise or forward stepwise selection.

Then BIC criterion penalises the number of parameters more in cases where a large number of data points are available. However, the number of parameters is already relatively low so this is not really necessary for this problem and AIC might be preferable as a result.

Question 2

You would now like to refine your study of the sailing race dataset in Question 1 in order to make a prediction for a forthcoming race. To do so, you decide to also develop a logistic regression model. Once again, the data has size $n = 734$ and each entry corresponds to the characteristic of a boat for a given race. You once again have access to the weight of the boat in kg (denoted x_1), the size of its sail in m^2 (denoted x_2) and the length of the race in km (denoted x_3). However, the response variable is now a variable describing whether or not the boat won the race it was in (denoted y', with $y' = 1$ if the boat won, and $y' = 0$ otherwise).

You would like to use this model to determine whether your favourite boat will win the forthcoming race. This boat weighs 300kg and has a sail of 10.71m^2. The race is 300km long.

You fit an initial logistic regression model of y' in terms of x_1, x_2, x_3 and obtain the following parameter estimates: $\beta_0 = -3298.78$, $\beta_1 = -68.39$, $\beta_2 = 706.97$ and $\beta_3 = 54.20$ where β_0 is an intercept and β_i the coefficient corresponding to x_i for $i = 1, 2, 3$.

Answer the following sub-questions describing your analysis of this problem through logistic regression.

(i) A description (with equations) of the logistic regression model with these three explanatory variables. You should include a discussion in your own words (and maximum three sentences) of all assumptions made, and of whether these are likely to hold for your model. [4 marks]

(ii) An explanation (in words and equations) of why this model falls within the framework of generalised linear models, and an alternative formulation based on latent variables. You should also discuss in your own words how the latent variable should be interpreted in the context of your model of the boat race (maximum 1 sentence)? [5 marks]

(iii) An estimate of the probability that your favourite team will win the race. You should provide a detailed derivation (using equations) of how you arrive at this result. Since you would like to bet on your team, you should also explain how to obtain the odds of the team winning from this probability, and calculate such odds. You should also comment (using words and equations) on how your answer would have differed if the boat had a sail of 10.5m^2. [5 marks]

(iv) A description (with equations) of how to add an interaction term between x_2 and x_3. Explain in your own words (and a maximum of 2 sentences) how such an interaction term could be justified from the point of view of a sailing fan, and describe the impact on the odds of increasing x_3 by 1km. [6 marks]

(v) A discussion of possible non-linear transformations which could be used for x_1 in your model. In particular, your discussion should focus on a comparison of cubic regression and of cubic splines with K knots for x_1. You should write down both models (in equations) and discuss the advantages and disadvantages of each approach in your own words (and in a maximum of two sentences). [6 marks]

(vi) A discussion in your own words of the advantages and disadvantages of the linear regression models obtained in Question 1 compared to the logistic regression models considered in Question 2. You should also comment on which approach you would recommend (maximum 4 sentences). [4 marks]

Sketch Solution:

(i) [bookwork/unseen] The logistic regression model can be expressed as a linear model on log-odds

$$\log\left(\frac{P(Y_i'=1|x_i)}{P(Y_i'=0|x_i)}\right)=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}\qquad\forall i=1,\dots,n.$$

Here, x_{i1},x_{i2},x_{i3} are the explanatory variables for the i^{th} boat. The parameters $\beta_0,\dots,\beta_3$ are coefficients taking values in $\mathbb{R}$.

Note that we are assuming the data for each boat is independent, but this is unlikely to be the case here. Indeed, in a given race, there can only be one winner, meaning that the outcome variable will take value 1 only once per race.

Sketch Solution:

(ii) [bookwork/unseen] Alternatively, the model can be described as:

$$\begin{aligned}&y_i'\sim\text{Bernoulli}(\theta_i),\\&\theta_i=g^{-1}(\nu_i)=\frac{1}{1+\exp(-\nu_i)}\\&\nu_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+\epsilon_i\qquad\forall i=1,\dots,n.\end{aligned}$$

where $\{\epsilon_i\}_{i=1}^n$ are iid and follow a logistic(0,1) distribution. From this formulation, we can see that it is a GLM with Bernoulli distribution and logit link function. Furthermore, ν_i can be thought of as a latent variable representing the strength of boat i. It is called latent because the noise terms are unknown, and therefore ν_i is also unobserved. The so-called latent-variable formulation is equivalent to the above and is usually written as:

$$y_i'=\begin{cases}1 & \text{if } \beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+\epsilon_i>0\\0 & \text{else.}\end{cases}$$

Sketch Solution:

(iii) [similar to seen] The log-odds can be obtained directly by plugging in the data into the expression given in the solution to (i):

$$\log o = -3298.78 - 68.39 \times 300 + 706.97 \times 10.71 + 54.20 \times 300 = 15.87$$

so we can simply take the exponential to obtain odds of $o = 7792716$. Now given some odds o, we can calculate the probability of winning (denoted p) by inverting the formula:

$$o = \frac{p}{1-p}$$

This gives:

$$p = \frac{o}{1+o} = 0.99$$

With a sail of 10.5m^2, $\log(o) = -132.59$, $o \approx 0$ and $p \approx 0$. A small change in the size of the sail can therefore have a drastic impact on the performance of a boat, at least according to our model.

Sketch Solution:

(iv) [bookwork/unseen] The new model would take the form:

$$\log\left(\frac{P(Y_i' = 1|x_i)}{P(Y_i' = 0|x_i)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2} x_{i3} \qquad \forall i = 1, \ldots, n.$$

where β_4 is also real valued.

Such an interaction term could be used to model the fact that having a larger sail over a long distance can lead to a larger advantage when it comes to winning the race.

In this case, given a fixed value of x_{i2}, an increase in x_{i3} by 1km would increase the log-odds by $\beta_3 + \beta_4 x_{i2}$. Equivalently, the odds would increase by $\exp(\beta_3 + \beta_4 x_{i2})$.

Sketch Solution:

(v) [bookwork/unseen] In the case of cubic polynomial regression, the latent variable would be:

$$\nu_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \beta_4 x_2 + \beta_5 x_3 + \epsilon_i$$

In the case of cubic splines, given the knot c, the model would be:

$$\nu_i = \begin{cases} \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \beta_4 x_2 + \beta_5 x_3 + \epsilon_i & \text{if } x < c \\ \beta_0 + \beta_1' x_1 + \beta_2' x_1^2 + \beta_3' x_1^3 + \beta_4 x_2 + \beta_5 x_3 + \epsilon_i & \text{if } x \geq c \end{cases}$$

where the coefficients would be selected so that derivatives match up to order 2 at the point c. Alternatively, this can be written as:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \beta_4 (x_1 - c)_+^3 + \beta_5 x_2 + \beta_6 x_3 + \epsilon_i$$

where $(x_1 - c)_+^3 = (x_1 - c)^3$ if $x_1 > c$ and 0 else. In the case of K knots, the model would simply include further basis functions of the form $(x_1 - c^i)^3$ for $i = 1, ..., K$.

The advantage of the cubic regression model is that it has less parameters and can hence be fit faster. On the other hand, the cubic spline is more flexible and can hence lead to a better fit, but could lead to overfitting if not used appropriately. One other difficulty with the cubic spline is the need to pick the number of knots and their locations.

Sketch Solution:

(vi) [unseen] On the one hand the logistic regression model is more appropriate as it models directly the quantity of interest, and we obtain a probability of a win/loss which tells us how uncertain we are about the result of the race. On the other hand, the logistic regression does not seem very appropriate for tackling this problem since we cannot assume that the data is iid as only one boat can win a given race. This will create some issues when fitting the parameters, and might make the model unreliable. On top of this a small change in x_2 can have a significant impact on the odds obtained (due to the magnitude of the coefficients). As a result, the linear regression of Question 1 should probably be preferred.

[END OF PAPER]