STAT0032: INTRODUCTION TO STATISTICAL DATA SCIENCE 2019

Answer ALL questions. Section A carries 40% of the total marks and Section B carries 60%. The relative weights attached to each question are as follows: A1 (10), A2 (10), A3 (10), A4 (10), B1 (30), B2 (30). The numbers in square brackets indicate the relative weights attached to each part question. Marks will not only be given for final (numerical) answers but also for the accuracy and clarity of the reasoning.
Time allowed: Two hours.

Section A

A1 Consider $U \sim \text{Uniform}[0,1]$, the uniform distribution on the interval $[0,1]$ with density function f_U satisfying

$$f_U(u) = \begin{cases} 0 & \text{if } u < 0 \\ 1 & \text{if } 0 \leq u \leq 1 \\ 0 & \text{if } 1 < u \end{cases}$$

(a) Calculate an expression for the cumulative distribution function of U, $F_U(u) = P(U \leq u)$, for values of u in the interval $(-\infty, \infty)$ [3]

Solution

$$\begin{aligned} F_U(u) &= \int_{-\infty}^{u} f_U(x)dx \\ &= \begin{cases} 0 & \text{if } u < 0 \\ u & \text{if } 0 \leq u \leq 1 \\ 1 & \text{if } 1 < u \end{cases} \end{aligned}$$

(b) Calculate, showing your working, the expected value of U, $E[U]$ [3]

Solution

$$\begin{aligned} E[U] &= \int_{-\infty}^{\infty} u f_U(u)du \\ &= \int_0^1 u du \\ &= \frac{u^2}{2}\Big|_0^1 \\ &= 1/2 \end{aligned}$$

(c) Calculate, showing your working, the variance of U, $Var(U)$ [4]

CONTINUED

Solution

$$\begin{aligned} Var(U) &= E[U^2] - E[U]^2 \\ E[U^2] &= \int_0^1 u^2 du \\ &= \frac{u^3}{3}\Big|_0^1 \\ &= 1/3 \\ \Rightarrow Var(U) &= 1/3 - (1/2)^2 \\ &= 1/12 = 0.0833 \end{aligned}$$

A2 A number of hypothesis tests, n, are conducted where the null hypothesis is known to be true in every case. The probability of incorrectly rejecting the null hypothesis (a false positive) is the level of the test, α, independently for each of the n tests.

(a) Calculate an expression for the probability of no false positives across the n tests [3]

Solution

Probability of a false positive is α

Probability of avoiding a false positive is $1 - \alpha$

Probably of avoiding a false positive for each of the n (independent) tests is $(1 - \alpha)^n$

(b) Calculate an expression for the probability of at least one false positive across the n tests [2]

Solution

Probability of at least one false positive is one minus the probability of no false positives, $1 - (1 - \alpha)^n$

(c) In the case where $\alpha = 0.05$ determine the smallest value of n such that the probability of at least one false positive exceeds 0.5 [3]

Solution

$$\begin{aligned} 1 - (1 - 0.05)^n &\geq 0.5 \\ \Rightarrow 0.95^n &\leq 0.5 \\ \Rightarrow n &\geq \frac{\log(0.5)}{\log(0.95)} \\ \Rightarrow n &\geq 13.51 \\ \Rightarrow n &= 14 \end{aligned}$$

(d) Name and describe a correction to the level of each test such that the probability of at least one false positive is no more than α [2]

Solution

The Bonferroni correction is one such correction, conducting each test instead at the level α/n

A3 For each of n individuals p covariates are measured

$$x_{i,j} \qquad i = 1, 2, \ldots, n \qquad j = 1, 2, \ldots, p$$

in order to fit a linear regression to predict the outcome measurement for each individual, y_i.

(a) State a relationship between the number of observations, n, and the number of covariates, p, which must hold for the coefficients of the full model

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \epsilon_i$$

to be uniquely determined by minimising the residual sum of squares [2]

Solution

$n > p$

(b) Name an alternative statistics to the residual sum of squares for quantifying model fit which includes a penalisation for model complexity [2]

Solution

AIC (Akaike Information Criterion)/BIC (Bayesian Information Criterion)/Mallows C_p

(c) Briefly outline the steps involved in a forward stepwise selection methodology using the model fit statistic you specified in part (b) of this question [6]

Solution

Begin with the null model, including only the intercept, quantifying its fit using the appropriate statistic, eg. AIC.
Fit all models with a single covariate, quantifying fit using AIC.
Determine the best fitting model with a single covariate - the model with the smallest AIC.
Building on this model with a single further covariate, quantify

CONTINUED

model fit for all models with one additional covariate.
Determine the resulting best fitting model.
Repeat, adding an additional covariate at each step until the saturated model is reached.
Compare the AIC scores of the resulting $p+1$ models, reporting the one with the smallest AIC as the best model under forward stepwise selection.

A4 A bank provides loans of a fixed amount of money, £5 000, to each of n individuals for a period of one year. At the end of the year the bank records

$$y_i = \begin{cases} 1 & \text{if the amount was repaid in full} \\ 0 & \text{if the amount wasn't repaid in full} \end{cases}$$

$x_i =$ the individual's yearly earnings

A logistic regression model is fitted with $\mu = E[Y] = P(Y = 1) = p$ and

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \tag{1}$$

(a) Briefly describe why in this scenario a logistic regression model might be preferred to a linear regression model of the form

$$p = \beta_0' + \beta_1' x$$

[3]

Solution
Salaries can take a very wide range of values from zero upwards. In the case of the linear regression, there will be some salary values, x, for which the probability of repaying the loan, p, is greater than one or less than zero. This is not a concern for the logistic regression as the inverse of the logit function, the logistic function, transforms values in the range $(-\infty, \infty)$ to the range $[0, 1]$.

(b) State, with brief justification given the scenario in the question, whether you would expect β_1 to be greater than, equal to, or less than zero [3]

Solution
It is fair to assume the amount that an individual earns in a year

would influence their probability of paying back the loan, therefore we rule out the case of $\beta_1 = 0$. It is also reasonable to assume that the more an individual earns, the greater their probability of paying off the loan on schedule, therefore we settle on the case of $\beta_1 > 0$.

(c) Solve equation (1) for p, that is determine the function f of β_0, β_1 and x such that $p = f(\beta_0, \beta_1, x)$ [4]

Solution

$$\begin{aligned}
\log\left(\frac{p}{1-p}\right) &= \beta_0 + \beta_1 x \\
\Rightarrow \frac{p}{1-p} &= \exp(\beta_0 + \beta_1 x) \\
\Rightarrow p &= \exp(\beta_0 + \beta_1 x) - p\exp(\beta_0 + \beta_1 x) \\
\Rightarrow p(1 + \exp(\beta_0 + \beta_1 x)) &= \exp(\beta_0 + \beta_1 x) \\
\Rightarrow p &= \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}
\end{aligned}$$

Section B

B1 A lecture of n UCL students are used as a sample to investigate the relationship between an individual's height, weight and level of physical activity. An individual is recorded as being physically active if they engage in three or more hours of exercise a week, if they engage in less than 3 hours of exercise a week they are recorded as physically inactive. The recorded measurements for individual i are:

y_i = weight of student i in kilograms

$x_{1,i}$ = height of student i in metres

$$x_{2,i} = \begin{cases} 1 & \text{if student } i \text{ is physically active } (\geq 3 \text{ hours of exercise weekly}) \\ 0 & \text{otherwise} \end{cases}$$

$$x_{3,i} = \begin{cases} 1 & \text{if student } i \text{ is physically inactive } (< 3 \text{ hours of exercise weekly}) \\ 0 & \text{otherwise} \end{cases}$$

Consider first the normal linear model of the form

$$y_i = \beta_0 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon_i \qquad (2)$$

CONTINUED

(a) Calculate expressions for the expected weight of a physically active student and the expected weight of a physically inactive student [2]

Solution
Expected height for a physically active student is $E[Y|x_2 = 1, x_3 = 0] = \beta_0 + \beta_2$. Expected height for a physically inactive student is $E[Y|x_2 = 0, x_3 = 1] = \beta_0 + \beta_3$.

(b) Using your answer to part (a) name and provide a brief explanation of a problem experienced when attempting to estimate the coefficients of the normal linear model denoted in equation (2) [3]

Solution
The model with both dummy variables has three unknowns, but only two equations. As a result, there does not exist a unique solution – a constant could be added to β_0 and subtracted from both β_2 and β_3 and the result remain identical. This is because x_2 and x_3 are collinear.

Consider next the normal linear model of the form

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_3 x_{i,3} + \epsilon_i$$

This model is fitted in a computational statistics package and the output for β_1 includes a t-statistic value alongside a p-value.

(c) Write down the null and alternative hypotheses to which the p-value for β_1 relate [3]

Solution
$H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$ given the other terms in the model (physical activity)

(d) Write down three assumptions on the error terms, ϵ_i, which are assumed in order to calculate the p-value for β_1 [3]

Solution
The error terms are assumed to be independent and normally distributed with (mean zero and) constant variance

(e) In the case where the assumptions you described in part (d) are not satisfied a bootstrap procedure may be used to determine a 95% confidence interval for β_1. Name and describe, carefully and in detail, how you would conduct such a bootstrap procedure. [12]
You do not need to state the exact form of the estimator $\hat{\beta}_1$ for a given set of observations. You may use the result that for $Z \sim$

$N(0,1)$, *a standard normal random variable,* $P(-1.96 < Z < 1.96) = 0.95$.

Solution

In the case of the normal nonparametric bootstrap first produce m samples of size n of x's and y's by resampling with replacement data pairs from the original data set. For each sample estimate the corresponding $\hat{\beta}_{1,j}$. Calculate the mean and standard deviation of the collection of $\hat{\beta}_{1,j}$ via

$$\mu(\hat{\beta}_1) = \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_{1,j}$$

$$sd\left(\hat{\beta}_1\right) = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\beta}_{1,j} - \mu(\hat{\beta}_1)\right)^2}.$$

Construct the confidence interval as $[\mu(\hat{\beta}_1) - 1.96sd(\hat{\beta}_1), \mu(\hat{\beta}_1) + 1.96sd(\hat{\beta}_1)]$.

(f) Briefly describe how the 95% confidence interval obtained in part (e) could be used to test the hypotheses you stated in part (c) at the 5% significance level [3]

Solution

If zero lies within the confidence interval fail to reject H_0. If zero doesn't lie within the confidence interval reject H_0.

Consider finally the normal linear model of the form

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_3 x_{i,3} + \beta_4 x_{i,1} x_{i,3} + \epsilon_i$$

(g) Calculate an expression for the expected difference in weight for two physically active students who differ in height by 10 centimetres [2]

Solution

$\beta_0 + 0.1\beta_1$

(h) Calculate an expression for the expected difference in weight for two physically inactive students who differ in height by 10 centimetres [2]

Solution

$(\beta_0 + \beta_3) + 0.1(\beta_1 + \beta_4)$

B2 The relationship between two variables, x and y, is to be modelled by a smooth function f, $y = f(x) + \epsilon$. The data available to make inference on f is a collection of n distinct observations, $x_1, x_2, \ldots, x_n$, of x alongside corresponding paired observations, $y_1, y_2, \ldots, y_n$, of y.

Consider first the modelling of f via regression with higher order terms, via a function f of the form

$$f(x) = \beta_0 + \sum_{i=1}^{u} \beta_i x^i$$

(a) Write down the number of degrees of freedom in f [2]

Solution

There are $u+1$ coefficients and therefore $u+1$ degrees of freedom

(b) Briefly describe a potentially undesirable feature of regression models with higher order terms [3]

Solution

In a polynomial regression model every observation influences the coefficient estimates, which then strongly impact f across the whole range of x values. This may be undesirable when observations spread across wide or disjoint ranges of x values.

Consider next the modelling of f via cubic splines where the regime changes at each of v knot locations, $k_1 < k_2 < \ldots < k_v$, via a function f of the form

$$f(x) = \begin{cases} \beta_{0,0} + \beta_{1,0}x + \beta_{2,0}x^2 + \beta_{3,0}x^3 & \text{for } x < k_1 \\ \beta_{0,1} + \beta_{1,1}x + \beta_{2,1}x^2 + \beta_{3,1}x^3 & \text{for } k_1 \le x < k_2 \\ \quad \vdots & \\ \beta_{0,v-1} + \beta_{1,v-1}x + \beta_{2,v-1}x^2 + \beta_{3,v-1}x^3 & \text{for } k_{v-1} \le x < k_v \\ \beta_{0,v} + \beta_{1,v}x + \beta_{2,v}x^2 + \beta_{3,v}x^3 & \text{for } k_v \le x \end{cases}$$

(c) Specify, with brief details, suitable knot locations, $k_1, k_2, \ldots, k_v$, with reference to the observations, $x_1, x_2, \ldots, x_n$ [3]

Solution

The knots could be relatively evenly spaced according to the $a/(v+1)$ quantiles of the x_i for $a = 1, 2, \ldots, v$

(d) Constraints on the $\beta_{i,j}$ can be derived to ensure continuity in up to and including the second derivative of f. Calculate expressions

for the required constraints to maintain this level of continuity at the first knot, k_1. [8]

Solution

$$\beta_{0,0} + \beta_{1,0}k_1 + \beta_{2,0}k_1^2 + \beta_{3,0}k_1^3 = \beta_{0,1} + \beta_{1,1}k_1 + \beta_{2,1}k_1^2 + \beta_{3,1}k_1^3$$
$$\beta_{1,0} + 2\beta_{2,0}k_1 + 3\beta_{3,0}k_1^2 = \beta_{1,1} + 2\beta_{2,1}k_1 + 3\beta_{3,1}k_1^2$$
$$2\beta_{2,0} + 6\beta_{3,0}k_1 = 2\beta_{2,1} + 6\beta_{3,1}k_1$$

(e) Calculate an expression for the number of degrees of freedom in f when continuity in up to and including the second derivative of f is enforced [3]

Solution

There are a total of $4(v+1)$ coefficients and $3v$ constraints, leaving $v + 4$ degrees of freedom

An equivalent to cubic splines is to represent f as the sum of basis functions $b_i(x)$ with

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 b_1(x) + \beta_5 b_2(x) + \ldots + \beta_{v+3} b_v(x)$$
$$b_i(x) = (x - k_i)_+^3$$
$$= \begin{cases} (x - k_i)^3 & \text{for } x - k_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

(f) Briefly describe a computational advantage to modelling f by the sum of basis functions rather than by cubic splines [3]

Solution

The coefficients of the cubic splines are determined via numerical optimisation under constraints. In the case of the basis functions numerical optimisation is used without the need for constraints. Numerical optimisation is easier/faster when there are no constraints.

Consider finally the modelling of f as any function which is continuous in up to and including the second derivative which for $\lambda \geq 0$ minimises

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_{\min(x_i)}^{\max(x_i)} \left(\frac{d^2 f}{dx^2}(x) \right)^2 dx \qquad (3)$$

(g) Comment on the form of equation (3) and why the function f which minimises this expression might be a desirable representation of the relationship between x and y [5]

Solution

The first term is a measure of the fit of f to the observed data. The second term penalises the rate at which the function fluctuates by taking the integral of the square of the second derivative (the square being taken to ensure that the contributions to the integral are all positive). Minimising the expression could lead to a function f which fits better to general data x and y rather than just the specifically observed x and y (bias/variance tradeoff).

(h) Describe the form of the function f which minimises equation (3) as $\lambda \to \infty$ [3]

Solution

As λ gets very large the expression is dominated by reducing the penalty term. The penalty term is zero in the case where f has zero second derivative across the whole range of the x_i. This is satisfied when f is a linear function, returning to the method of simple linear regression.

End of Paper