STATG006: Introduction to Statistical Data Science Mock Exam

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, given in total marks (the overall sum of marks is 100): A1 (15), A2 (25), B1 (25), B2 (20), B3 (15). The numbers in square brackets indicate the relative weight attached to each part question.

Section A

A1 Answer the following questions about probability:

(a) Two fair dice are thrown. Let X be the smallest of the two numbers obtained (or the common value if the same number is obtained on both dice). Find the probability mass function of X. Find $P(X > 3)$. [4]

(b) Let X be a random variable with expectation μ. Find the expectation of the random variable $Y = X - \mu$. [2]

(c) The proportion of time during a 40-hour week that an industrial robot is in operation is modelled by a random variable X with probability density function

$$f(x) = \begin{cases} cx, & \text{if } 0 \le x \le 1, \\ 0, & \text{otherwise,} \end{cases}$$

where c is a constant. Find c. Find $P(X < 12)$ and $P(X > 1/3 \mid X < 1/2)$. [4]

(d) Suppose that the number of distinct uranium deposits in a given area is a Poisson random variable with parameter $\mu = 10$. If, in a fixed period of time, each deposit is independently discovered with probability 1/50, find the probability that (i) exactly one, (ii) at least one and, (iii) at most one deposit is discovered during that time. The probability mass function of a Poisson with parameter μ is $p(x) = \mu^x e^{-\mu}/x!$, $x = 0, 1, 2, \ldots$. For a binomial (n, μ), $p(x) = \binom{n}{x}\mu^x(1-\mu)^{n-x}, x = 0, 1, 2, \ldots, n$. [5]

A2 Answer the following questions about general statistical methodology and probabilistic tools:

(a) If you toss a coin 1,000 times and observe 570 heads, explain briefly the steps you would take to assess the claim that the coin is fair. [5]

(b) Let X be a random variable whose distribution is parameterised by some θ. If $L(X)$ is a transformation of a random variable such that $P(L(X) \leq \theta) = 1 - \alpha_1$, and $U(X)$ is another transformation such that $P(U(X) \geq \theta) = 1 - \alpha_2$, and $L(x) \leq U(x)$ for all x, show that $P(L(X) \leq \theta \leq U(X)) = 1 - \alpha_1 - \alpha_2$. [4]

(c) Explain what a plot of residuals versus fitted values tells you about the fit of a linear regression model. [4]

(d) Assess whether the following is true or false, and explain why: The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k+1)-variable model identified by forward stepwise selection. [3]

(e) A pair of variables (X_1, X_2) follows a bivariate Gaussian with zero mean, unit variance and correlation coefficient of 0.4. A different pair of variables (Y_1, Y_2) follows a bivariate Gaussian distribution with zero mean, unit variance and correlation coefficient of zero. Is $P(0 \leq X_1 \leq 3$ and $0 \leq X_2 \leq 3)$ equal, less than or more than $P(0 \leq Y_1 \leq 3$ and $0 \leq Y_2 \leq 3)$? Explain your reasoning. [5]

(f) Suppose you are doing penalized curve fitting

$$\hat{g} = argmin_g \left(\sum_{i=1}^{n} (y^{(i)} - g(x^{(i)}))^2 + \lambda \int \left[g^{(m)}(x) \right]^2 \, dx \right),$$

where $g^{(m)}$ is the m-th derivative of g. Explain what happens when $\lambda = \infty$ under $m = 0$ and $m = 1$. [4]

Section B

B1 Consider a study of police stops in 75 different precincts of New York City in a particular year, broken down by the ethnicity of the persons being stopped. There is a record for every combination of precinct and ethnicity. More specifically, each record contains: the precinct code `precinct`; the ethnicity `eth`, encoded non-numerically as a set of categorical levels, `Black`, `Hispanic` and `White`; the population of the precinct, `pop`; and the total number of stops for that year, `stops`.

(a) Describe the likelihood function for a Poisson generalised linear model of the regression of `stops` on the other variables, justifying your choice of link function. [5]

(b) What would be the interpretation of the model if we fixed the coefficient for `pop` to 1? Would it make more sense to regress on `pop` or on the logarithm of `pop`? Explain. [5]

(c) Alternatively, what modelling issues would you have to face if the outcome variable was defined to be the ratio of `stops` by `pop`? [5]

(d) Figure 1 is a plot where the horizontal axis is the predicted value of $\hat{y}$ of `stops` for each data point, versus the residual $y - \hat{y}$ on the vertical axis, where each y is the observed value of `stops`. The bottom and top dashed lines represent an approximate confidence interval of 95% under the Poisson family. Explain what this figure tells you about the fit of the model and which recommended change to the model you would make. [5]

(e) The data has 225 rows, since each of the 75 precincts is combined once with each of the three ethnicities. Which unwanted implications this may have to your analysis, and what would you attempt to do to mitigate this? [5]

B2 Consider a regression problem where the outcome variable is out-of-state tuition per year, in dollars, for any given American college. Covariates include `private`, a binary variable indicating whether the college is public or private; `accept`, the number of applications accepted in the latest cohort; and `enroll`, the number of students enrolled in the last cohort.

(a) Write the likelihood function for a generalised additive model for the regression of tuition on `private`, `accept` and `enroll`, explaining your notation. Assume a Gaussian outcome model. The

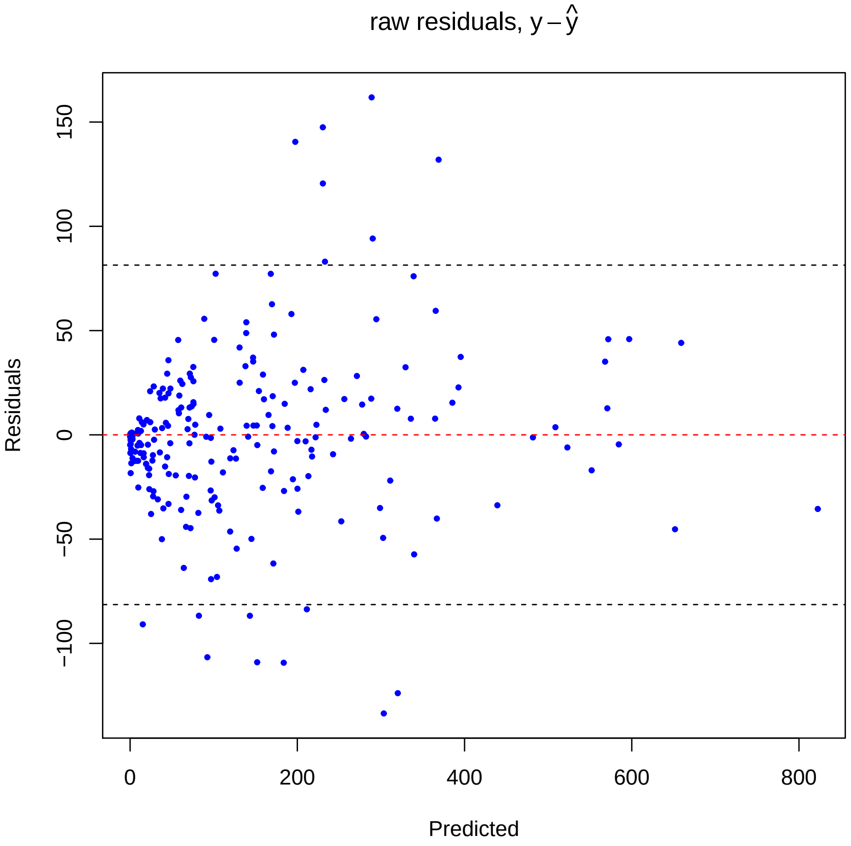

Figure 1: Figure used in question B1(d).

Gaussian density function is $p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$ for parameters μ and σ^2. [5]

(b) Explain very briefly why it is enough to plot each additive component separately to understand its contribution to the model. [3]

(c) Could the backfitting algorithm be used to fit this model by maximum likelihood? Explain why and how, concisely. [5]

(d) Alternatively, you could create two models, one for the subset of the data corresponding to private universities, and one corresponding to the subset formed by public universities. What can this tell you about the adequacy of the additivity assumption for the covariate `private` in the model of part (a)? [4]

(e) Criticize the assumption of Gaussianity for this data, given that we should expect that the distribution of tuition should be uneven in the sense a few colleges will charge much more than the average as expected by the Gaussian model. Which modification to the model would you suggest and why? [3]

B3 A researcher collects expression measurements for 1,000 genes in 100 tissue samples. The data can be written as a 1,000 by 100 matrix, which we call $\mathbf{X}$, in which each row represents a gene and each column a tissue sample. Each tissue sample was processed on a different day, and the columns of $\mathbf{X}$ are ordered so that the samples that were processed earliest are on the left, and the samples that were processed later are on the right. The tissue samples belong to two groups: control (C) and treatment (T). The C and T samples were processed in a random order across the days. The researcher wishes to determine whether each gene's expression measurements differ between the treatment and control groups.

As a pre-analysis (before comparing T versus C), the researcher performs a principal component analysis of the data, and finds that the first principal component (a vector of length 100) has a strong linear trend from left to right, and explains 10% of the variation. The researcher now remembers that each patient sample was run on one of two machines, A and B, and machine A was used more often in the earlier times while B was used more often later. The researcher has a record of which sample was run on which machine.

(a) Explain what it means that the first principal component "explains 10% of the variation", and how relevant this component might be. [5]

(b) The researcher decides to replace the (i, j)th element of $\mathbf{X}$ with

$$x_j^{(i)} - z_1^{(i)}\phi_{ji}$$

where $z_1^{(i)}$ is the ith score, and ϕ_{ji} is the jth loading, for the first principal component. He will then perform a two-sample t-test on each gene in this new data set in order to determine whether its expression differs between the two conditions. Critique this idea, and suggest a better approach. [5]

(c) Describe a hypothetical small simulation experiment to demonstrate the superiority of your idea, explaining its motivation. [5]

End of Paper