STATG006: Introduction to Statistical Data Science Mock Exam, Marking Sheet

Section A

A1 *Question A1 will always be about basic probability. Questions are meant the be simple and not to take much of your time. For this mock exam, I've chosen questions that are more time consuming on average than you might find in an actual exam.*

(a) 3 points for the pmf, 1 for the computation of $P(X > 3)$. This question assesses basic understanding on how to write down a random variable based on a problem specification.

(b) I would accept something even simpler, like $E[X - \mu] = E[X] - E[\mu] = \mu - \mu$ directly. But I would like to see the understanding that expectation carries over linear manipulations, which is the content being assessed here.

(c) 2 points for finding c, 1 point for each probability. This assesses an understanding of normalisation of pdfs, and the relationship between pdfs and probability.

(d) This is as hard as you could expect for Part A. Writing down $P(Y = k)$ as a function of $P(Y\ |N)$ and $P(N)$ is given 1 point, then 2 points for the calculation and 2 points for the probabilities (one point is given if one of the three is correct). This assess the concept of Law of Total Probability and how two random variables can be linked in a single model.

A2 *Question A2 will be miscellanea short questions about statistical methodology, with the odd question about manipulation of probabilities that is relevant to a statistical method (like item (b) here). Do NOT spend much time writing long answers. You might actually be penalized for answering to what was not asked even if your full answer includes the right points, in cases I interpret it as you not understanding what the question was about: shooting at every direction also means shooting your own foot.*

(a) 2 points for setting up the null and alternative, 2 points for correctly referring to the critical region and 1 point for the mentioning of the relevant distribution of the statistic. A very bog-standard question that can be answered in a very direct way.

(b) This is the hardest item of this question, conceptually. Essentially 2 points for setting up how $P(L(X) \leq \theta \leq U(X))$ would relate to $P(L(X) \leq \theta)$ and $P(\theta \leq U(X))$, and 2 points for the remaining calculations. This assesses understanding that the boundaries $L(X)$ and $U(X)$ in a confidence interval are random, not θ, and how these boundaries can be manipulated as in any standard probabilistic manipulation. This is not the easiest question, and overall Part A questions will be more straightforward than this.

(c) 2 points for mentioning what we expect from residuals, and 2 points for relating that to the visual aspect of the plot. Again, a very bog-standard piece of knowledge.

(d) 1 point for saying true or false, and 2 points for a very straightforward explanation. Some of the questions in Part A ask for very "obvious" explanations, but of course you need to know what you are talking about to begin with. In any case, this is a reminder that if something looks so straightforward to the point of looking tricky, this just means that the answer is actually straightforward instead of tricky!

(e) 2 points for mentioning how the contours of the distribution of (Y_1, Y_2) look like, 1 point for relating that to the event on the upper right quadrant, and 2 points to relate that to the distribution of (X_1, X_2). Purely conceptual question to assess your qualitative understanding of multivariate Gaussians, but which is on the hard end as it requires a bit of mental visualization.

(f) 2 points for each case. Only very basic understanding of a derivative is necessary here, no calculations required. This again assesses your conceptual understanding of what penalization means and how it relates to derivatives in a conceptual way.

Section B

B1 *A typical Part B question. Expect for sure at least one of these questions in the exam: interpret a model, criticize it in some way, suggest*

particular ways of improving it. Expect a graph or table to be posed to you, also with requests for interpretation.

(a) 2 points for the equation for the linear response, 2 points for a clear likelihood function and 1 point for justifying the link function. You should always expect a question along these lines in the exam. This assesses your basic modelling skills and how you represent the inputs and outputs of a regression problem.

(b) 2 points for the interpretation, 2 points for commenting on appropriateness and 1 point for commenting on the logarithm. A common mistake here is forgetting to mention that the coefficient changes $\log(\mu)$ (or η). It is NOT a change of Y, it is a change in the distribution of Y!

(c) 2 points for explaining how the interpretation changes, 1 point for mentioning the adequacy of the Poisson, 2 points for mentioning the possible issues of the proposed alternative.

(d) 1 points for mentioning points outside the interval, 2 points for explaining what this means, 2 points for suggesting an alternative.

(e) 1 point for pointing out the sharing of precinct information. 3 points for describing its implications. 1 point for a basic description of an alternative. This is a harder than average question and it is meant to assess how students understand the less obvious assumptions made in a model, such as independence of the data points.

B2 *There are commonalities between B2 and B1, something you should not see as unusual. At the very least, the methods covered will be different, and questions addressing particular properties of the method will arise.*

(a) 2 points for the equation for the linear response, 2 points for a clear likelihood function and 1 point for avoiding a pointless nonlinear function attributed to `predict`.

(b) 1 point for describing clearly what is changing as a function of the variable. 2 points for describing how it does not matter what happens to the other two.

(c) 1 point for yes/no. 1 point for initialization, 2 for the overall structure, 1 for the criterion for stopping the iterations. Notice that it is not important to be too formal here. I'm not asking you to write a computer program. The goal is to assess your understanding of the logical structure of backfitting.

(d) 2 points for setting up a comparison, 2 points for explaining what should be observed if additivity is adequate. The goal is to assess whether you understand the implications of additivity.

(e) 2 for explaining how the Gaussian fails to capture the intended shape of the distribution. 1 for an alternative. Other answers could be, for instance, using a Gamma regression model. *Just because this is not said explicitly it doesn't mean we cannot have more than one correct answer!*

B3 *A typical question on unsupervised learning. Again, interpretation is important, and more than one right answer is sometimes possible.*

(a) 3 points for clearly stating the meaning, particularly by comparing it to the total variance. 2 points for comments on the relevance. This question assesses basic understanding of meaning of principal component analysis.

(b) 3 points for explaining what the underlying rationale is, 2 points for a better approach. A harder than average question, it requires you to criticize a way in which a statistical method is used, deconstructing which faulty reasoning led to it.

(c) 3 points for setting up the simulation, 2 for the brief explanation on what will happen to the faulty approach in (b) instead.

End of Paper