STATG006: Introduction to Statistical Data Science Mock Exam, Solutions

Section A

A1 (a) Since the two dice are fair and independent, the probability of any combination of dice faces is $1/6 \times 1/6 = 1/36$. We will get $X = 1$ for outcomes $(1, 1)$, then $(x, 1)$ or $(1, x)$ for $x > 1$. That's 11 combinations. For $X = 2$ we have $(2, 2)$, then $(x, 2)$ or $(2, x)$ for $x > 2$. That's 9. Following the same reasoning for $X > 2$,

x	1	2	3	4	5	6
$P(X = x)$	11/36	9/36	7/36	5/36	3/36	1/36

So $P(X > 3) = P(X = 4)+P(X = 5)+P(X = 6) = 9/36 = 0.25$.

(b)

$$E[Y] = E[X - \mu] = \int (x - \mu)p(x)\ dx = \int xp(x) - \int \mu p(x)\ dx =$$
$$E[X] - \mu = 0.$$

(c) We must have

$$\int_{-\infty}^{\infty} p(x)\ dx = 1,$$

so $\int_0^1 cx\ dx - [cx^2/2]_0^1 = c/2 = 1$, so $c = 2$.
$P(X < 1/2) = \int_0^{1/2} 2x\ dx = [x^2]_0^{1/2} = 1/4$.
$P(X > 1/3 \mid X < 1/2) = P(1/3 < X < 1/2)/P(X < 1/2) = 4\int_{1/3}^{1/2} 2x\ dx = 4(1/3 - 1/9) = 5/9$.

(d) Let N be the number of deposits in the area, and let Y be the number of deposits discovered. So for $N = n$, $Y \sim Bin(n, 1/50)$. Hence for all integers $k \geq 0$,

$$\begin{aligned} P(Y = k) &= \textstyle\sum_{n=0}^{\infty} P(Y = k \mid N = n)P(N = n) = \\ &= \textstyle\sum_{n=k}^{\infty} P(Y = k \mid N = n)P(N = n) \\ &= \textstyle\sum_{n=k}^{\infty} \binom{n}{k}(1/50)^k(49/50)^{n-k} \times e^{-10}10^n/n! \\ &= \frac{(1/5)^k e^{-10}}{k!} \textstyle\sum_{n=k}^{\infty} \frac{(49/5)^{n-k}}{(n-k)!} \\ &= \frac{(1/5)^k e^{-10}}{k!} \textstyle\sum_{n=0}^{\infty} \frac{(49/5)^n}{n!} = \frac{(1/5)^k e^{-1/5}}{k!}. \end{aligned}$$

Notice this is the pmf of a Poisson with parameter $1/5$, so $Y \sim Poisson(1/5)$. We now have (i) $P(Y = 1) = 0.164$, (ii) $P(Y \geq 1) = 1 - P(Y = 0) = 1 - e^{-1/5} = 0.181$ and (iii) $P(Y \leq 1) = 0.983$.

A2 (a) I would test the null hypothesis $H_0 : \theta = 0.5$ against the alternative $H_1 : \theta \neq 0.5$, where θ is the probability of the coin showing heads in a single toss. I would reject the fairness assumption if the observed frequency of 570 heads out of 1000 falls in the level α critical region of the test, formed by the $\alpha/2$ and $1 - \alpha/2$ tails of a binomial $(1000, 0.5)$.

(b) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, so $P(\{L(X) \leq \theta\} \cup \{\theta \leq U(X)\}) = P(L(X) \leq \theta) + P(\theta \leq U(X)) - P(L(X) \leq \theta \leq U(X))$. But $P(\{L(X) \leq \theta\} \cup \{\theta \leq U(X)\})$ is 1, because if $L(X) \leq \theta$ is false, then $\theta \leq U(X)$ is true (and vice-versa) since $L(x) \leq U(x)$ for all x. So $1 = 1 - \alpha_1 + 1 - \alpha_2 - P(L(X) \leq \theta \leq U(X))$ and the result follows.

(c) In a linear regression model, we expect the error terms to be independent and identically distributed regardless of the mean of the outcome. A residual is just an estimate of the error, so the plot should show that as we look at the different possible fitted values, the distribution of the residuals should remain the same everywhere. If not, the linear regression model is not adequate.

(d) This is true, because forward regression proceeds by adding one variable at a time, keeping the existing ones, while improving on a particular objective function.

(e) More. The contours of the distribution of (Y_1, Y_2) are circles centered at zero, so for any a we have $P(-a \leq Y_1 \leq a \text{ and } -a \leq Y_2 \leq a)$ will be split equally over all four quadrants given by the $Y_1 = 0$ and $Y_2 = 0$ axes. In contrast, the elliptical contours given by the joint distribution of (X_1, X_2) will put more mass in the upper right and bottom left than in the upper left and bottom right. So the upper right quadrant will have more mass in the (X_1, X_2) case than in the (Y_1, Y_2) case.

(f) If $m = 0$ and the function is not identically zero, then $[g^{(0)}(x)]^2 = [g(x)]^2$ is positive and the penalization is infinite. Hence, the function will be zero everywhere[1]. If $m = 1$ and the function is not a constant, the reasoning is the same, and the method will boil down to the the average of the $y^{(i)}$.

[1]Note from lecturer: this is not quite right, see marking sheet document.

CONTINUED

Section B

B1 (a) I choose to code `eth` with two binary variables `eth_1` and `eth_2` so that $(0,0),(0,1),(1,0)$ correspond, respectively, to the levels `Black`, `White`, `Hispanic`. In the same way, I use 74 binary variables `precint_i`, $i = 1, 2, \ldots, 74$ to represent the 75 precints. For a generic data point, I define

$$\eta = \beta_0 + \sum_{i=1}^{74} \beta_i \texttt{precint\_i} + \beta_{75}\texttt{eth}_1 + \beta_{76}\texttt{eth\_2} + \beta_{77}\texttt{pop}.$$

Given $\eta^{(i)}$ for each data point $i = 1, 2, \ldots, 75 \times 3$, the likelihood function for $\beta_0, \ldots \beta_{77}$ is given by

$$L(\beta_0, \ldots, \beta_{77}) = \prod_{i=1}^{225} \mu_i^{y^{(i)}} e^{-\mu_i} / y^{(i)}!$$

where $y^{(i)}$ is the corresponding number of stops, and $\mu_i = \exp(\eta^{(i)})$. That is, I used the log link function to make sure that the mean of each Poisson is positive.

(b) $\beta_{77} = 1$ means that there is an implied change of 1 unit of $\log(\mu)$ per extra person observed in the precint. This implies an exponential increase of μ per units of increase of `pop`. This does not seen adequate, so using the logarithm of `pop` instead as a covariate sounds more reasonable.

(c) If our outcome Y is now given by `stops/pop`, this would keep the attractive interpretation of coefficients modelling directly the rate of stop per population. This adopts a default intuitive relationship between the two variables, avoiding the need to interpret what β_{77} might mean otherwise. However, this is now a continuous outcome variable. The Poisson model cannot be used anymore. A Gaussian model with independent error terms might be a problem if for some reason we need the variance to increase with the mean of the ratio `stops/pop`.

(d) There seems to be a small but non-neglible number of points beyond what would be expected for a 95% interval. One possibility is that the variance implied by the Poisson is not large enough to model this data. Hence, I would recommend other model for this regression problem, such as the negative binomial, which allows for large variance/mean relationships.

(e) The model assumes that data points are independent, but since some points share information from the same sources (such as precints), the assumption of independence might be unrealistic. This leads, for instance, to optimistic confidence intervals (there is actually less information in the data than the independence assumption implies). One way of mitigating this is to use a model that allows for dependencies across some data points.

B2 (a) Let `private` take values in $\{0, 1\}$. For a generic data point, I define

$$\mu = \beta_0 + \beta_1 \texttt{private} + f_2(\texttt{accept}) + f_3(\texttt{enroll}),$$

where $f_2(\cdot), f_3(\cdot)$ are functions to be determined by a particular method such as spline regression. Given μ_i for each data point $i = 1, 2, \ldots, n$, where n is the sample size, the likelihood function for $\beta_0, f_j(\cdot)$ and a variance parameter σ^2 is given by

$$L(\beta_0, \{f_j(\cdot)\}, \sigma^2) = \prod_{i=1}^{n} e^{-(y^{(i)} - \mu_i)^2/(2\sigma^2)}/\sigma^2,$$

where $y^{(i)}$ is the corresponding tuition.

(b) If we fix all variables but, say, `accept`, I can assess how μ changes as `accept` takes different values. This change is given entirely by f_2, which assumes the same value for a given `accept` regardless of the values of `private` and `enroll`.

(c) Yes. Maximum likelihood here is the same as least-squares for the coefficients and function parameters. We do the usual initalization of backfitting by setting $\hat{\beta}_0$ to be the average of tuition, and $\hat{\beta}_1 = 0$, $\hat{f}_2 = 0$ and $\hat{f}_3 = 0$ for all points. We can then just iterate over the three inputs, calculating the residual given by using the other two variables and fitting a single-input regression method (linear for `private`, non-linear for the others) until $\hat{\beta}_1$ and $\hat{f}_2, \hat{f}_3$ do not change noticeably.

(d) Suppose we compare pairs of universities (say, chosen from the training set or otherwise hypothetical) that have approximately same `accept` and `enroll` values but differ only on whether they are private or not. If across all pairs the difference in the expected outcome is approximately the same, then the additivity assumption for `private` would be sensible. If not, then the original model is not good in this respect.

CONTINUED

(e) Gaussianity assumes some symmetric distribution of variability, which is not the case. Taking the logarithm of the output might be a sensible choice, although the interpretation of the model changes.

B3 (a) "10%" here is merely the proportion of total variance in the sample projected into the first principal component, compared to the total variance of the original sample. This low number seems to indicate that some of the signal dependency across tissues is captured by the first component, although its source is unclear.

(b) This assumes that the principal component is the contribution of the type of machine to the dependency of the tissue measurements, which may not be. Since we do know which tissues were measured by which machines, we can split the data by machine, and do tests within each subset of data.

(c) Simulated data can be generated as four independent datasets, one for each machine and condition, to be added to the distribution of each gene. In this case, there is no reason to expect that PCA as in (b) should remove the machine contribution, while my suggestion will still work.

End of Paper