10 Statistics

10.12 Analysis of Performance Data (8 units)

Attendance at the Part II courses Principles of Statistics and Statistical Modelling is recommended. Familiarity with R may also be helpful, but is not required.

Part 1: Linking crime and unemployment

As part of a study on unemployment and crime, Farrington *et al.* (1986) use the data in Table 1 on 32 boys. These boys were selected from a total of nearly 400 boys in the study as those who

- committed at least one offence while in employment (Em) or in unemployment (Un), and,
- have had at least 0.25 years in each of Em and Un.

(For details on the provenance of the data see the paper by Farrington *et al.*, who also discuss at length the difficulties in assessing quantitatively the eventual effects unemployment might bear on the rate of offending.)

Boy	Years in Employment	Offences in that Time	Years in Unemployment	Offences in that Time
1	0.83	1	0.88	3
2	1.02	0	0.89	8
3	1.17	0	2.50	2
4	1.21	3	0.68	1
5	1.25	4	0.96	2
6	1.31	2	0.69	1
7	1.50	2	1.25	0
8	1.52	2	0.64	2
9	1.54	1	0.85	0
10	1.58	2	1.08	0
11	1.66	4	0.61	1
12	1.75	1	0.25	0
13	2.21	1	1.37	4
14	2.22	0	0.28	1
15	2.25	0	0.75	1
16	2.37	1	0.55	0
17	2.39	2	0.44	1
18	2.42	3	0.78	0
19	2.45	1	0.47	0
20	2.51	2	0.40	0
21	2.64	1	0.70	0
22	2.65	8	0.60	0
23	2.67	1	0.67	0
24	2.83	1	0.37	0
25	2.84	1	0.75	0
26	2.97	2	0.36	0
27	3.00	1	0.34	0
28	3.07	1	0.27	0
29	3.15	2	0.43	0
30	3.16	1	0.67	1
31	3.28	4	0.45	0
32	3.37	2	0.30	0

Table 1: Data set for Part 1. This data can be downloaded in the file `II-10-12-2020-01.csv` from the CATAM website.

Let n_{1i} and n_{2i} be the numbers of offences committed by the i^{th} boy in Employment and Unemployment respectively, in times t_{1i} and t_{2i} years. Assume that n_{1i} and n_{2i} are independent

Poisson variables, with parameters $\lambda_i t_{1i}^{\alpha}$ and $\phi\lambda_i t_{2i}^{\alpha}$ respectively. Here α and ϕ represent the parameters of interest with ϕ being the additional 'risk' (if any) of committing an offence while unemployed rather than while employed.

Question 1 Write down the likelihood of the data in terms of the unknown parameters ϕ, α and $(\lambda_1, \ldots, \lambda_{32})$.

Question 2 Show that the distribution of n_{2i} conditional on $n_{1i} + n_{2i} = m_i$, say, is Binomial with parameters m_i and ψ_i, where

$$\log \frac{\psi_i}{1-\psi_i} = \log \phi + \alpha \log \frac{t_{2i}}{t_{1i}} .$$

Hence write down the likelihood of the data L $(n_{2,1}, \ldots, n_{2,32})$ conditional on the observed marginal totals $(m_1, \ldots, m_{32})$ as a function of the unknown parameters ϕ and α.

Question 3 Find by Newton–Raphson iteration the maximum likelihood estimates (m.l.e.) $\hat{\phi}$ and $\hat{\alpha}$ and their corresponding standard errors. Taking the m.l.e. you obtained, plot the conditional log likelihood as a function of $\log \phi$. Comment on its shape.

Question 4 Do you think unemployment increases the rate of offending? Justify your answer.

Reference: D.P. Farrington *et al.* (1986) *Unemployment, school-leaving and crime.* British Journal of Criminology, **26**, 335–356.

Part 2: Academic College Tables

A newspaper publishes every year the equivalent of a league table ranking the colleges in a famous academic institution. It builds the table by allocating a score for the class of degree obtained by each graduating student. A first class degree yields 5 points, a II.1 (upper second) 3 points, a II.2 (lower second) 2 points, and a third 1 point. As the proportions of firsts and other classes vary by subject, these scores are adjusted such that the proportion of firsts become equal across all subjects. The resultant scores are then split by college and summed up to produce the college score in the table. This procedure is said to allow a fairer comparison between the colleges in the institution.

Position	College	Score	I	II.1	II.2	III
1	Agincourt	306.9	11	46	13	2
2	Resolution	305.0	27	50	33	6
3	Erin	304.8	23	75	30	4
4	Duke	304.8	25	61	26	10
5	Colingwood	299.5	25	69	45	4
6	Sovereign	296.9	18	54	26	6
7	Malaya	296.7	17	50	25	8
8	Elizabeth	296.4	22	70	37	5
9	Howe	289.5	18	62	40	4
10	Nelson	288.1	15	49	28	7
11	Fisher	286.9	9	41	18	4
12	Valiant	286.6	37	88	69	12
13	Queen Mary	286.5	29	74	55	13
14	Vanguard	284.0	16	60	40	9
15	Rodney	283.3	13	57	32	8
16	Prince	282.5	13	47	48	9
17	Anson	281.7	16	51	43	6
18	Barham	281.6	13	60	39	5
19	King George	281.1	20	53	41	8
20	Hood	280.8	12	53	32	5
21	Jellicoe	278.9	10	50	32	6
22	Beaty	274.0	17	77	49	12
23	Cunningham	270.2	4	48	32	2
24	Lord	269.6	14	34	49	6
25	Lewin	264.8	9	43	42	7
26	Mountbatten	247.1	5	35	32	5

Table 2: College Table 2007, in order of merit. This data can be downloaded in the file `II-10-12-2020-02.csv` from the CATAM website.

College	Score	I	II.1	II.2	III
Agincourt	288.7	11	31	28	1
Resolution	290.0	20	58	43	3
Erin	293.5	19	58	39	2
Duke	321.4	24	68	18	3
Colingwood	305.1	30	69	36	8
Sovereign	292.5	22	51	37	9
Malaya	303.5	15	53	24	2
Elizabeth	321.5	33	64	29	4
Howe	293.3	13	79	29	2
Nelson	287.6	11	50	22	3
Fisher	274.6	8	31	24	3
Valiant	303.5	49	97	49	16
Queen Mary	309.0	31	92	38	4
Vanguard	278.6	5	77	32	6
Rodney	299.0	22	60	32	7
Prince	295.3	20	61	45	6
Anson	285.8	15	64	34	6
Barham	289.1	13	78	26	8
King George	283.9	20	52	30	11
Hood	281.7	12	55	39	2
Jellicoe	273.3	11	43	49	2
Beaty	262.9	9	86	53	12
Cunningham	270.1	4	48	30	2
Lord	284.8	11	61	33	4
Lewin	292.8	19	57	45	7
Mountbatten	271.9	12	51	48	12

Table 3: College Table 2008, in order of merit. This data can be downloaded in the file `II-10-12-2020-03.csv` from the CATAM website.

Take x_i, y_i as the score per 100 students for the i^{th} college for 2007 and 2008 respectively, for $i = 1, \ldots, 26$.

Question 5 Fit the model

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2),$$

where α, β and σ^2 are unknown parameters. Comment on the fit thus obtained.

Question 6 Find a 95% confidence intervals for β, and $\alpha + \beta\bar{x}$ (i.e., the predicted 2008 score for a college with 2007 score equal to the mean of $x_1, \ldots, x_{26}$). How does this confidence interval change if $\bar{x}$ is replaced by 305.0?

Let n_{ij} be the frequency for college i, class j for 2007 ($i = 1, \ldots, 26$, $j = 1, \ldots, 4$). Assume (n_{ij}) independent, multinomial, parameters n_i and (p_{ij}), where $\sum_j p_{ij} = 1$ for each i and n_i is defined as $\sum_j n_{ij}$.

Question 7 Using the appropriate large sample result, test the hypothesis

$$H_0 : p_{ij} = \lambda_j \qquad \text{for each} \quad i, j$$

where (λ_j) is unknown, and $\sum_{j=1}^{4} \lambda_j = 1$. Interpret H_0, and then the result of your test.

Question 8 Comment critically on the presentation of these data and on the appropriateness of the two analyses performed above.

