UPPSALA UNIVERSITY
Department of Mathematics
Rolf Larsson

Exam in Mathematical Statistics
Analysis of Categorical Data 1MS370
2020–08–28

Time: 8.00-13.00. Limits for the credits 3, 4, 5 are 18, 25 and 32 points, respectively. The solutions should be well motivated.

Permitted aids: The course book or copies thereof. Hand-written sheet of formulae (two-sided is permitted). Pocket calculator. Dictionary. *No electronic device with internet connection.*

1. A discrete random variable Y has probability mass function

$$p(y;\pi) = \pi^{3y}(1-\pi^3)^{1-y},$$

for $y = 0, 1$. Assume that $0 < \pi < 1$.

   (a) Does this distribution belong to the exponential family, and in that case, why? (2p)

   (b) Suggest an appropriate link function $g(\pi)$. (2p)

   (c) Let x be an explanatory variable that can take any real value. Discuss if the GLM $g(\pi) = \alpha + \beta x$ is a suitable model. (2p)

2. In the 2000 general Society Syrvey in the US, a random sample of voters were asked about their political party identification. The results are given in the table below.

	Democrat	Independent	Republican
Females	762	327	468
Males	484	239	477

   (a) Test if political party identification was independent of gender. (3p)

   (b) Partition the G^2 statistic to see if the gender comparison as in (a) turns out different when considering only Democrat vs Independent and then Democrat or Independent vs Republican. (3p)

Please turn the page!

3. Let $P(Y = 1|x) = \pi(x) = F(\alpha + \beta x)$. Consider the following suggestions for the function F (below, $\pi \approx 3.14$, not a probability):

   (*i*) $F(z) = \begin{cases} 0 & \text{if } z < 0, \\ \sin\left(\frac{\pi}{2}z\right) & \text{if } 0 \leq z \leq 1, \\ 1 & \text{if } z > 1, \end{cases}$

   (*ii*) $F(z) = \frac{1}{\pi}\arctan(z) + \frac{1}{2}$,

   (*iii*) $F(z) = \frac{2}{\pi}\arctan(z)$.

   (a) Which (if any) of the suggestions gives a suitable model, and which do not? Why or why not? (2p)

   (b) Take your favourite choice of function F from above. For which x (as a function of α and β) is it true that the function $\pi(x) = 1/2$? (2p)

   (c) What is the rate of increase of the function $\pi(x)$ at this point? (2p)

4. The grades for those who passed the course "Applied Statistics" in March 2020 at the STS program in Uppsala, for year of birth -97 (up to 1997) and 98- (1998 and on) are given in the table below.

   All probabilities mentioned below should be interpreted as conditional probabilities given that the student has passed the exam.

	-97	98-
3	11	2
4	16	14
5	12	10

   (a) Based on the table, estimate the ratio of the probability to get a grade of 4 to the probability to get a grade of 3 for the two age groups. (1p)

   (b) Consider the baseline-category model

$$\log\left\{\frac{\pi_j(x)}{\pi_1(x)}\right\} = \alpha_j + \beta_j x,$$

   where $j = 2, 3, ..., J$ and $\pi_j(x) = P(Y = j|x)$ with $x = 0$ for born -97 and 1 for born 98-.

   Such a model was estimated based on the data, where $j = 1, 2, 3$ correspond to the grades 3, 4, 5, respectively.

   The parameter estimates were $\hat{\alpha}_2 = 0.37469$, $\hat{\alpha}_3 = 0.08701$, $\hat{\beta}_2 = 1.57122$, $\hat{\beta}_3 = 1.52243$.

   Calculate the corresponding estimates as in (a) and comment. (3p)

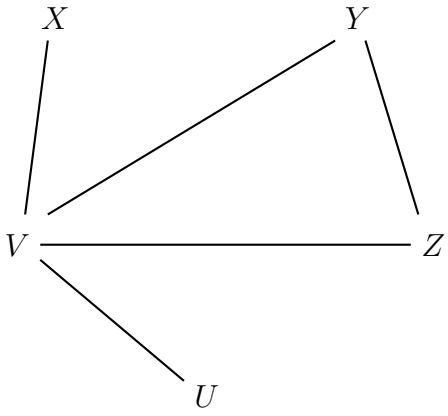

Figure 1: Model, problem 5.

(c) Consider the cumulative logit model

$$\text{logit}\{P(Y \leq j|x)\} = \alpha_j + \beta x,$$

where $j = 1, 2, ..., J-1$, $\alpha_1 \leq \alpha_2 \leq ... \leq \alpha_{J-1}$.
This model was estimated on data (js corresponding to grades as in (a)), resulting in the estimates $\hat{\beta} = -0.6891$, $\hat{\alpha}_1 = -1.1252$, $\hat{\alpha}_2 = 0.9843$.
Interpret the sign on $\hat{\beta}$, calculate the corresponding estimates as in (a) and comment. (3p)

5. Suppose we have variables X, Y, Z, U, V and a loglinear model described by the graph in figure 1.

(a) Give the name (symbol) of a model that may be described by this graph expressed in a form like (X, Y, YZ) (which is not the model under question here). Also, write down the model equation (in a form like $\log \mu_{ijklm} = \lambda + \lambda_i^X + ...$). (2p)

(b) Is X independent of Z? (2p)

(c) Is X conditionally independent of Z given V? (2p)

(d) Is X conditionally independent of Z given Y? (2p)

6. Aggregate data on applicants to graduate school at Berkeley for the six largest departments (d) in 1973 classified by admission (a) and gender (g) was analyzed. The admission variable is 1 for admitted and 0 for rejected. The gender variable is 1 for male and 0 for female.

   Let μ_{ijk} be the expected count in cell (i, j, k), $i, j = 1, 2$, $k = 1, 2, ..., 6$, where i corresponds to admission, j corresponds to gender and k corresponds to department.

   A log linear Poisson model was fit to this data, including all main effects and all two-way interactions. The coefficients with at least one index equal to one were set to zero. The R print from the model estimation is given below.

```
> m=glm(n~a+g+factor(d)+a:g+a:factor(d)+g:factor(d));summary(m)

Call:
glm(formula = n ~ a + g + factor(d) + a:g + a:factor(d) + g:factor(d))

Deviance Residuals:
     1       2       3       4       5       6       7       8       9
 14.50  -14.50  -14.50   14.50   16.50  -16.50  -16.50   16.50    8.25
    10      11      12      13      14      15      16      17      18
 -8.25   -8.25    8.25  -24.75   24.75   24.75  -24.75   12.25  -12.25
    19      20      21      22      23      24
-12.25   12.25  -26.75   26.75   26.75  -26.75

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    327.50      35.84   9.137 0.000263 ***
a              170.00      43.51   3.907 0.011332 *
g             -323.00      43.51  -7.423 0.000699 ***
factor(d)2    -104.00      49.34  -2.108 0.088870 .
factor(d)3    -114.25      49.34  -2.316 0.068433 .
factor(d)4     -73.25      49.34  -1.485 0.197781
factor(d)5    -177.25      49.34  -3.592 0.015669 *
factor(d)6      -3.25      49.34  -0.066 0.950035
a:g            -71.00      32.89  -2.158 0.083349 .
a:factor(d)2   -57.00      56.97  -1.000 0.363015
a:factor(d)3  -271.50      56.97  -4.765 0.005035 **
a:factor(d)4  -261.50      56.97  -4.590 0.005894 **
a:factor(d)5  -279.50      56.97  -4.906 0.004452 **
a:factor(d)6  -445.50      56.97  -7.819 0.000549 ***
g:factor(d)2    91.00      56.97   1.597 0.171105
g:factor(d)3   492.50      56.97   8.644 0.000342 ***
g:factor(d)4   337.50      56.97   5.924 0.001955 **
g:factor(d)5   459.50      56.97   8.065 0.000475 ***
g:factor(d)6   342.50      56.97   6.012 0.001830 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 1623)

    Null deviance: 451210  on 23  degrees of freedom
Residual deviance:   8115  on  5  degrees of freedom
AIC: 247.87

Number of Fisher Scoring iterations: 2
```

(a) Explain why the number of degrees of freedom for residual deviance is 5. (1p)

(b) Test the model vs the saturated model and interpret the result. (2p)

(c) Which logit model for the probability of admittance does this loglinear model correspond to? Based on the R estimation of the loglinear model above, which are the estimated paramaters of this logit model? What is the corresponding number of degrees of freedom? Explain! (4p)

GOOD LUCK!

APPENDIX B

Chi-Squared Distribution Values

	Right-Tailed Probability						
df	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.96	26.12
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	23.83	28.41	31.41	34.17	37.57	40.00	45.32
25	29.34	34.38	37.65	40.65	44.31	46.93	52.62
30	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	45.62	51.80	55.76	59.34	63.69	66.77	73.40
50	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	77.58	85.53	90.53	95.02	100.4	104.2	112.3
80	88.13	96.58	101.8	106.6	112.3	116.3	124.8
90	98.65	107.6	113.1	118.1	124.1	128.3	137.2
100	109.1	118.5	124.3	129.6	135.8	140.2	149.5

Categorical Data Analysis, Third Edition. Alan Agresti.