UPPSALA UNIVERSITY
Department of Mathematics
Rolf Larsson

Exam in Mathematical Statistics
Analysis of Categorical Data 1MS370
2022–01–10

Time: 8.00-13.00. Limits for the credits 3, 4, 5 are 18, 25 and 32 points, respectively, including bonus from hand-in assignments. 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

$$p(y;\mu) = \left(1 - \frac{1}{\mu}\right)^{y-1} \frac{1}{\mu},$$

for $y = 0, 1, 2, ...$ and $\mu > 1$.

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

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

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

2. The numbers of deaths in motor vecicle accidents in the counties of Stockholm and Norrbotten for the years 1976 and 1996, men and women, are given in the following table. (Data from Statistics Sweden.)

year	gender	Stockholm	Norrbotten
1976	men	99	34
	women	53	15
1996	men	46	13
	women	21	7

   (a) For the 1976 data, estimate the odds ratio, calculate a 95% confidence interval for the odds ratio and interpret. (2p)

   (b) For the 1976 data, test if gender is independent of county with respect to tendency of dying in motor vehicle accidents. Use the Pearson X^2 test or the likelihood ratio test. (2p)

   (c) Use the data from both 1976 and 1996 to test if gender is independent of county with respect to tendency of dying in motor vehicle accidents, conditional on the year. (3p)

Please turn the page!

3. Let $P(Y=1|x) = \pi(x) = F(\alpha+\beta x) = 1-P(Y=0|x)$. Consider the following suggestions for the function F (in all cases, $-\infty < z < \infty$):

   (i) $F(z) = \frac{\exp(2z)}{1+\exp(z)+\exp(2z)}$,

   (ii) $F(z) = z^2$,

   (iii) $F(z) = (1+z^{-1})^{-1}$.

   (a) Which (if any) of the suggestions gives a suitable model, and which do not? Why or why not? (3p)

   (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 $\pi(x) = 1/2$? (3p)

Please turn the page!

4. The numbers of employed people (converted to full-time employees) in different teaching categories at Uppsala university in 2011, by gender, are given in the table below.

Categories are in Swedish: 'Professor' means full professor, 'lektor' is a lecturer with a PhD exam, 'adjunkt' is a lecturer without a PhD exam and 'fo-ass' is short for 'forskarassistent', which was a time-limited position (similar to today's 'biträdande lektor') with a high percentage of research time for persons who recently got their PhD.

	men	women
Professor	438	126
Lektor	248	249
Adjunkt	64	104
Fo-ass	108	73

(a) From the table, what is the proportion of lecturers ('lektor') that are women divided by the proportion of professors that are women? Also, calculate the proportion of lecturers that are women divided by the proportion of people in the category 'adjunkt' that are women. (1p)

(b) Consider the baseline-category model

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

where $j = 2$ for 'lektor', $j = 3$ for 'adjunkt' and $j = 4$ for 'fo-ass'. The baseline category, $j = 1$, corresponds to professor.

Moreover, $\pi_j(x) = P(Y = j|x)$ with $x = 0$ for men and 1 for women.

Such a model was estimated based on the data, where $j = 1, 2, 3, 4$ correspond to 'professor', lektor', 'adjunkt' and 'fo-ass', respectively.

The parameter estimates were $\hat{\alpha}_2 = -0.5688$, $\hat{\alpha}_3 = -1.9234$, $\hat{\alpha}_4 = -1.4001$, $\hat{\beta}_2 = 1.2500$, $\hat{\beta}_3 = 1.7315$, $\hat{\beta}_4 = 0.8543$.

Estimate the proportions in (a) as probability ratios from the model and comment. (2p)

(c) Consider a shift of the numbering so that the baseline category $j = 1$ is 'adjunkt', and we have $j = 2$ for 'lektor', $j = 3$ for 'professor' and $j = 4$ for 'fo-ass'.

What do you think $\hat{\alpha}_3$ would be, and why? (2p)

(d) If the teacher categories could be ordered in a natural way, suggest a suitable statistical model that takes such ordering into account. (But you don't need to estimate the model parameters.) Do not forget to impose parameter restrictions, if needed. (2p)

Please turn the page!

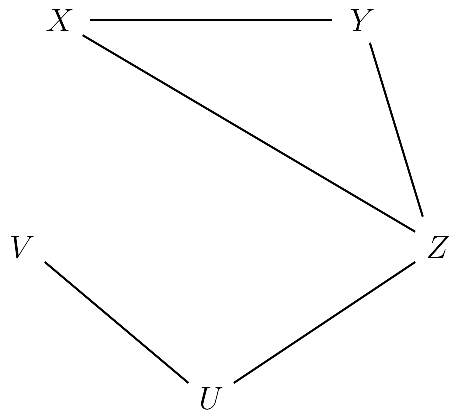

Figure 1: Model, problem 5.

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 U? Why or why not? (1p)

   (c) Is X conditionally independent of U given Z? Why or why not? (1p)

   (d) Is X conditionally independent of U given Y? Why or why not? (1p)

   (e) If we sum over the possible values that the variable V can take, which model do we get, and why? (2p)

Please turn the page!

6. We have a data set from a survey of workers in the US cotton industry, which records whether they were suffering from the lung desease byssinosis, as well as the values of three categorical variables: dustiness of the workplace, smoking status of the worker and the length of employment.

   Let the probability for subject i to suffer of the desease, $P(Y=1|X=i,Z=j,W=k)=\pi_{ijk}$, fulfill the model

$$\text{logit}(\pi_{ijk}) = \alpha + \beta_i^X + \beta_j^Z + \beta_k^W,$$

   where $i=1,2,3$ is the degree of dustiness (`dust`), $j=1,2$ correspond to smoker/non smoker (`smoke`), and $k=1,2,3$ correspond to length of employment (`emple`).

   The coefficients with index equal to one were set to zero. The R print from the estimation is given here.

```
> m=glm(yes/n~factor(dust)+factor(smoke)+factor(emple),family=binomial(link=logit),
weights=n);summary(m)

Call:
glm(formula = yes/n ~ factor(dust) + factor(smoke) + factor(emple),
    family = binomial(link = logit), weights = n)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6239  -0.8119  -0.2576   0.3307   1.6605

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.8336     0.1525 -12.026  < 2e-16 ***
factor(dust)2   -2.5493     0.2614  -9.753  < 2e-16 ***
factor(dust)3   -2.7175     0.1898 -14.314  < 2e-16 ***
factor(smoke)2  -0.6210     0.1908  -3.255 0.001133 **
factor(emple)2   0.5060     0.2490   2.032 0.042119 *
factor(emple)3   0.6728     0.1813   3.710 0.000207 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 322.527  on 64  degrees of freedom
Residual deviance:  43.882  on 59  degrees of freedom
  (7 observations deleted due to missingness)
AIC: 162.56

Number of Fisher Scoring iterations: 5
```

   A modified model, including an interaction effect between dustiness and smoking,

$$\text{logit}(\pi_{ijk}) = \alpha + \beta_i^X + \beta_j^Z + \beta_k^W + \beta_{ij}^{XZ},$$

   was also estimated. *Please turn the page!*

The coefficients with at least one index equal to one were set to zero. The R print from the estimation is given here.

```
> m=glm(yes/n~factor(dust)+factor(smoke)+factor(emple)+factor(dust):factor(smoke),
family=binomial(link=logit),weights=n);summary(m)

Call:
glm(formula = yes/n ~ factor(dust) + factor(smoke) + factor(emple) +
    factor(dust):factor(smoke), family = binomial(link = logit),
    weights = n)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4325  -0.7602  -0.2795   0.4335   1.3123

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -1.7573     0.1555 -11.301  < 2e-16 ***
factor(dust)2                -2.9576     0.3565  -8.295  < 2e-16 ***
factor(dust)3                -2.8325     0.2230 -12.701  < 2e-16 ***
factor(smoke)2               -0.9573     0.2751  -3.480 0.000502 ***
factor(emple)2                0.4990     0.2499   1.997 0.045869 *
factor(emple)3                0.6638     0.1819   3.649 0.000264 ***
factor(dust)2:factor(smoke)2  1.1807     0.5490   2.151 0.031497 *
factor(dust)3:factor(smoke)2  0.4864     0.4338   1.121 0.262201
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 322.527  on 64  degrees of freedom
Residual deviance:  39.031  on 57  degrees of freedom
  (7 observations deleted due to missingness)
AIC: 161.71

Number of Fisher Scoring iterations: 5
```

(a) Estimate the probability π_{221} for the two models, and compare to the empirical proportion of diseased for this category, which is $5/278 \approx 0.0180$. (3p)

(b) Test the first model (without interaction) vs the second model (including the interaction) at the 5% level and draw a conclusion. (2p)

(c) Residual analysis for the model including the interaction is given in figures 2-6 below. Looking at these figures, do you think that the model assumptions are satisfied? (2p)

GOOD LUCK!

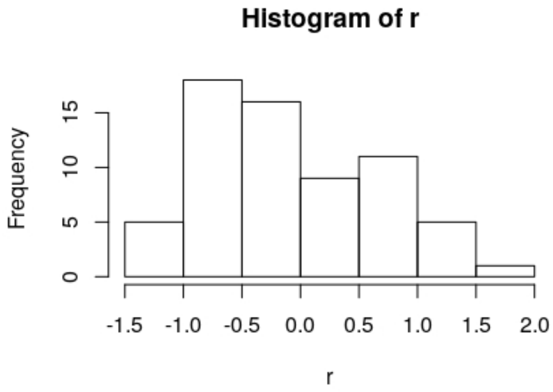

Figure 2: Histogram of standardized residuals.

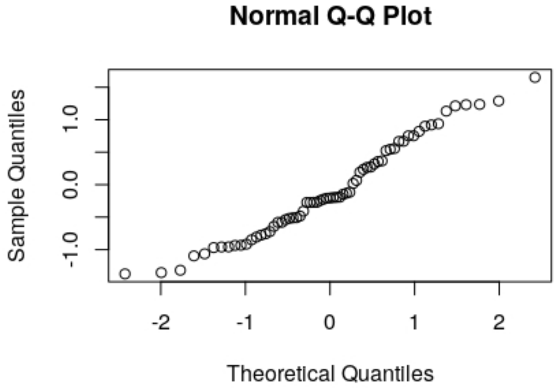

Figure 3: QQ plot of standardized residuals.

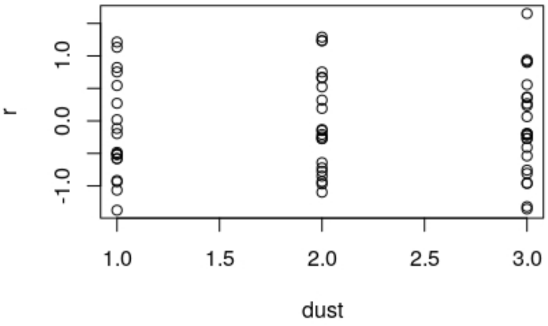

Figure 4: Plot of standardized residuals vs the dust variable.

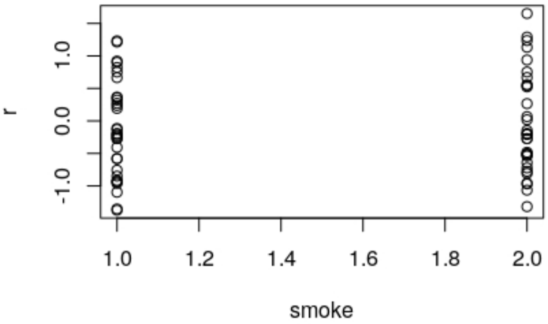

Figure 5: Plot of standardized residuals vs the smoke variable.

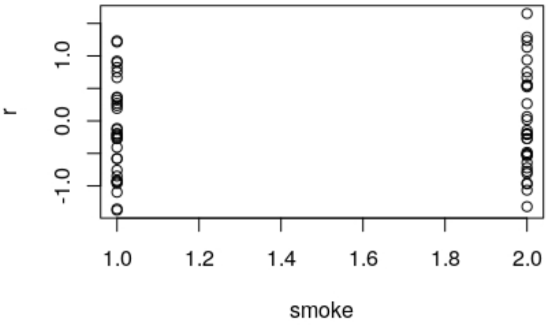

Figure 6: Plot of standardized residuals vs the length of employment variable.

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.