STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5006
EXAMINATION
February 16 2022

Categorical Data Analysis – Examination

February 16, 2022, 8.00-13.00

Examination by: Ola Hössjer, ph. 070 671 12 18, `ola@math.su.se`
Allowed to use: Miniräknare/pocket calculator and tables at the appendix of this exam.
Återlämning/Return of exam: Will be communicated on the course homepage and by email upon request.

Each correct solution to an exercise yields 10 points.
Limits for grade: A, B, C, D, and E are 45, 40, 35, 30, and 25 points of 60 possible points (including bonus of 0-10 points from computer assignments).

Reasoning and notation should be clear. You might answer in Swedish or English.

Read first through the whole exam. Exercises need not to be ordered from simpler to harder.

Problem 1

Let Y equal 1 or 0 depending on whether a person of age x years ever had coronary heart disease (CHD) symptoms or not.

a. Formulate a logistic regression model for this problem. (2p)

b. In a certain study (Hosmer and Lemeshow, 1989), 100 subjects participated and reported their age x_i and CHD status y_i, $i = 1, \ldots, 100$. Parameter estimates where $\hat{\alpha} = -5.310$ (intercept) and $\hat{\beta} = 0.111$ (effect parameter). Use this to compute a prediction of the probability $\pi(60)$ that a 60 year old person in the study population had evidenced CHD symptoms. (3p).

c. The parameter estimates are approximately normally distributed with an estimated covariance matrix

$$\begin{pmatrix} \widehat{\mathrm{Var}}(\hat{\alpha}) & \widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta}) \\ \widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta}) & \widehat{\mathrm{Var}}(\hat{\beta}) \end{pmatrix} = \begin{pmatrix} 1.2852 & -0.0267 \\ -0.0267 & 0.0006 \end{pmatrix}.$$

Use this information to compute a 95% confidence interval for $\pi(60)$. (5p)

Problem 2

a. Use a Wald test for the data set of Problem 1 in order to show that age has a significant effect on CHD symptoms at level 0.05. (2p)

b. The deviance is $G^2(M_1) = 107.35$ for the logistic regression model M_1 with age included as a predictor, and $G^2(M_0) = 136.66$ for the submodel M_0 without age as a predictor. Show by means of a likelihood ratio test that age has a significant effect on CHD symptoms at level 0.05. (2p)

c. Derive the likelihood score equations for α and β for the data set $\{(x_i, y_i); i = 1, \ldots, 100\}$ of Problem 1. (4p)

d. Suppose age is divided into I levels $x^1, \ldots, x^I$, and that the total number of subjects (observed value 100) is Poisson distributed. Find a contingency table for the data set and a loglinear model corresponding to M_1. (2p)

Problem 3

Consider a 2×2 table with independent binomial sampling for the two rows, so that

$$\begin{aligned} N_{11} &\sim \mathrm{Bin}(n_1, \pi_1), \\ N_{21} &\sim \mathrm{Bin}(n_2, \pi_2), \end{aligned}$$

where $n_1 = n_{11} + n_{12}$ and $n_2 = n_{21} + n_{22}$ are the two row sums of the table, and n_{ij} is an observation of the random variable N_{ij}.

a. Define the relative risk r. (1p)

b. Find the maximum likelihood (ML) estimate $\hat{r}$ of r. Hint: Find first (without proof) the ML estimates of π_1 and π_2. (2p)

c. It can be shown that $\log(\hat{r})$ is approximately normally distributed when n_1 and n_2 are both large, with mean $\log(r)$ and variance

$$\mathrm{Var}\left(\log(\hat{r})\right) = \frac{1-\pi_1}{n_1\pi_1} + \frac{1-\pi_2}{n_2\pi_2}. \tag{1}$$

Use (1) in order to express the standard error

$$\mathrm{SE} = \sqrt{\widehat{\mathrm{Var}}\left(\log(\hat{r})\right)}$$

as a function of $n_{11}, n_{12}, n_{21}, n_{22}$. (2p)

d. An automobile insurance company investigated whether the accidents rates at two geographic regions A and B were different. They registered how many of their customers in each region that had car accidents or not during a one-year period, with the following result:

Regions	Accident? Yes	No	Total
A	200	19 800	20 000
B	360	39 640	40 000

Compute a 95% confidence interval for the relative risk r. Is there a significant difference in accident rates between the two regions? (5p)

Problem 4

The 2x2x2 contingency table below contains data for 576184 car accidents in Florida (Agresti, 2013). Three categorical variables were observed for each accident; whether the driver used safety belt (S), was ejected (E) and had a fatal injury (I) or not.

Safety belt?	Ejected?	Injury Nonfatal	Fatal
Yes	Yes	1105	14
	No	411 111	483
No	Yes	4 624	497
	No	157 342	1008

We assume that all these numbers are observations of independent Poisson distributed random variables N_{sei} for $1 \leq s, e, i \leq 2$. The deviances $G^2(M)$ for a number of fitted loglinear models M are as follows:

Model	G^2	p	df
(S,E,I)	11 444		
(SE,I)	3 568		
(SI,E)	9 557		
(S,EI)	9 021		
(SE,EI)	1 145		
(SE,SI)	1 681		
(SE,SI,EI)	2.85		
(SEI)	0		

a. Fill out the remaining two columns for the number of parameters p and degrees of freedom df of each model. (Motivate your answer, but you don't need to specify the parameters of the models.) (2p)

b. Which model is selected by the AIC criterion? (Hint: Part of the solution is to motivate why AIC(M) need not be computed for any model M.) (2p)

c. Which of the two models (SE,SI,EI) or (SEI) is selected by a likelihood ratio test at significance level 0.05? (3p)

d. Specify the loglinear parameters of model (SE,SI,EI) and write down the distribution of N_{sei} in terms of these parameters. Which parameters are constrained to equal 0 if 2 is the baseline level for S, E and I? (3p)

Problem 5

Let n_{ij}, $1 \leq i \leq I$, $1 \leq j \leq J$ be the observed counts for all cells of a two-way $I \times J$ contingency table. Assuming that all n_{ij} are observations of random variables N_{ij}:

a. Determine the joint probability function $P(N_{11} = n_{11}, \ldots, N_{IJ} = n_{IJ})$ under Poisson sampling in terms of all $\mu_{ij} = E(N_{ij})$. (2p)

b. Determine the joint probability function $P(N_{11} = n_{11}, \ldots, N_{IJ} = n_{IJ})$ under multinomial sampling with a total of $n = \sum_{ij} n_{ij}$ observations and cell probabilities π_{ij}. (3p)

c. Which of the three studies; cohort, case-control and clinical trial, is typically based on multinomial sampling? (2p)

d. Show that multinomial sampling can be obtained from Poisson sampling by conditioning on the value of $N = \sum_{ij} N_{ij}$. (Hint: The cell probabilities are $\pi_{ij} = \mu_{ij}/\mu$, where $\mu = \sum_{ij} \mu_{ij}$.) (3p)

Good luck!

Appendix A - Table for chi-square distribution

Table 1: Quantiles of the chi-square distribution with df $= 1, 2, \ldots, 12$ degrees of freedom

	degrees of freedom											
prob	1	2	3	4	5	6	7	8	9	10	11	12
0.8000	1.64	3.22	4.64	5.99	7.29	8.56	9.80	11.03	12.24	13.44	14.63	15.81
0.9000	2.71	4.61	6.25	7.78	9.24	10.64	12.02	13.36	14.68	15.99	17.28	18.55
0.9500	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31	19.68	21.03
0.9750	5.02	7.38	9.35	11.14	12.83	14.45	16.01	17.53	19.02	20.48	21.92	23.34
0.9800	5.41	7.82	9.84	11.67	13.39	15.03	16.62	18.17	19.68	21.16	22.62	24.05
0.9850	5.92	8.40	10.47	12.34	14.10	15.78	17.40	18.97	20.51	22.02	23.50	24.96
0.9900	6.63	9.21	11.34	13.28	15.09	16.81	18.48	20.09	21.67	23.21	24.72	26.22
0.9910	6.82	9.42	11.57	13.52	15.34	17.08	18.75	20.38	21.96	23.51	25.04	26.54
0.9920	7.03	9.66	11.83	13.79	15.63	17.37	19.06	20.70	22.29	23.85	25.39	26.90
0.9930	7.27	9.92	12.11	14.09	15.95	17.71	19.41	21.06	22.66	24.24	25.78	27.30
0.9940	7.55	10.23	12.45	14.45	16.31	18.09	19.81	21.47	23.09	24.67	26.23	27.76
0.9950	7.88	10.60	12.84	14.86	16.75	18.55	20.28	21.95	23.59	25.19	26.76	28.30
0.9960	8.28	11.04	13.32	15.37	17.28	19.10	20.85	22.55	24.20	25.81	27.40	28.96
0.9970	8.81	11.62	13.93	16.01	17.96	19.80	21.58	23.30	24.97	26.61	28.22	29.79
0.9980	9.55	12.43	14.80	16.92	18.91	20.79	22.60	24.35	26.06	27.72	29.35	30.96
0.9990	10.83	13.82	16.27	18.47	20.52	22.46	24.32	26.12	27.88	29.59	31.26	32.91
0.9991	11.02	14.03	16.49	18.70	20.76	22.71	24.58	26.39	28.15	29.87	31.55	33.20
0.9992	11.24	14.26	16.74	18.96	21.03	22.99	24.87	26.69	28.46	30.18	31.87	33.53
0.9993	11.49	14.53	17.02	19.26	21.34	23.31	25.20	27.02	28.80	30.53	32.23	33.90
0.9994	11.78	14.84	17.35	19.60	21.69	23.67	25.57	27.41	29.20	30.94	32.65	34.32
0.9995	12.12	15.20	17.73	20.00	22.11	24.10	26.02	27.87	29.67	31.42	33.14	34.82
0.9996	12.53	15.65	18.20	20.49	22.61	24.63	26.56	28.42	30.24	32.00	33.73	35.43
0.9997	13.07	16.22	18.80	21.12	23.27	25.30	27.25	29.14	30.97	32.75	34.50	36.21
0.9998	13.83	17.03	19.66	22.00	24.19	26.25	28.23	30.14	31.99	33.80	35.56	37.30
0.9999	15.14	18.42	21.11	23.51	25.74	27.86	29.88	31.83	33.72	35.56	37.37	39.13